Fungal genome evolution and software for the life sciences

By

Jacob Lucas Steenwyk

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

For the degree of

DOCTOR OF PHILOSOPHY

In

Biological Sciences

May 13, 2022

Nashville, Tennessee

Approved:

John. A Capra, Ph.D.

Brandt F. Eichman, Ph.D.

Nicole Creanza, Ph.D.

John G. Gibbons, Ph.D.

Antonis Rokas, Ph.D.

Dedication

To my mother, Mercy T. Steenwyk, who gave me the gift and taught me the importance of education. To my father, Howard H. Steenwyk, who instilled in me the value of dedicated and passionate work. To my siblings, Emily L. Steenwyk and Nina C. Steenwyk, who shared the joys of the world with me.

Acknowledgements

This work was made possible thanks to financial support from the Howard Hughes Medical Institute James H. Gilliam Fellowships for Advanced Study program and the Mosig Endowment for Biological Sciences at Vanderbilt University. For their patience, support, and willingness to bestow their wisdom onto me, I would also like to thank my advisor, Dr. Antonis Rokas, as well as current and former members of his laboratory including Drs. Xiaofan Zhou, Jennifer H. Wisecaver, Haley R. Eidem, Xing-Xing Shen, Matthew E. Mead, Abigail L. Labella, Yuanning Li; collaborators such as Drs. Gustavo G. Goldman, Nicholas H. Oberlies, Chris Todd Hittinger, Dana Opulente, Judith Berman, Huzefa A. Raja, Sonja L. Knowles, Brand F. Eichman, Noah P. Bradley; and members of my thesis committee—Drs. John A. Capra, Brandt F. Eichman, Nicole Creanza, John G. Gibbons, and Antonis Rokas. I also thank non-academic collaborators—Thomas J. Buida and Dayna C. Goltz—for their assistance in software engineering and development.

I want to expand upon my gratitude and thanks for my advisor, Dr. Antonis Rokas. As a teacher and mentor, he taught more than words can contain. He always fostered my curiosity, enabled me to explore, was patient with my shortcomings, and showed me how to use my strengths. By example, he taught me not only how to be a better scientist, but also how a person—both professional and personal—should be. I would similarly like to thank persons who did not agree to be my mentor but naturally developed a mentor-trainee relationship with me—such as Drs. Judith Berman and David Hibbett. I also thank Dr. John G. Gibbons who fostered my scientific curiosity and creativity during a master's program at Clark University. During this time, Dr. John G. Gibbons gifted me a robust technical and scientific foundation to conduct comparative

genomic analyses. I also would like to thank Dr. Xing-Xing Shen who taught me much of what I know about the field of phylogenetics and phylogenomics.

Last, and certainly not least, I thank my family. I consider myself incredibly lucky to have been born into an inspiring, kind, and loving family. Through the years, they have supported me in ways that these few words—nor essays, books, or volumes for that matter—could accurately describe. My love for you, as your support has been for me, will remain steadfast.

Table of Contents

DEDICATION ..... **iii**

ACKNOWLEDGEMENTS ..... **iv-v**

LIST OF FIGURES ..... **vii**

LIST OF TABLES ..... **xii**

LIST OF ABBREVIATIONS ..... **xiii**

CHAPTER 1: INTRODUCTION **1**

Billions of base pairs: deep sequencing of the fungal kingdom 1

Fungi, champions of ecological and genomic diversity 2

Bioinformatics, a young field with growing pains 3

Insights into fungal genome evolution and new software for the life sciences 5

CHAPTER 2: A ROBUST PHYLOGENOMIC TIMETREE FOR BIOTECHNOLOGICALLY AND MEDICALLY IMPORTANT FUNGI IN THE GENERA *ASPERGILLUS* AND *PENICILLIUM* **9**

INTRODUCTION 9

MATERIALS AND METHODS 14

Genome sequencing and assembly 14

Data collection and quality assessment 15

Phylogenomic data matrix construction 16

Maximum likelihood phylogenetic analyses 18

Evaluating topological support 21

Estimating divergence times 28

Statistical analysis and figure making 31

Data availability 31

RESULTS 32

The examined genomes have nearly complete gene sets 32

The generated data matrices exhibit very high taxon occupancy 32

A genome-scale phylogeny for the family Aspergillaceae 33

Examination of the Aspergillaceae phylogeny reveals 14 incongruent bipartitions ... 33

Incongruence in the Aspergillaceae phylogeny 44

Topology tests 47

A geological timeline for the evolutionary diversification of the .....

Aspergillaceae family 47

DISCUSSION 49

CHAPTER 3: AN EVOLUTIONARY GENOMIC APPROACH REVEALS BOTH CONSERVED AND SPECIES-SPECIFIC GENETIC ELEMENTS RELATED TO HUMAN DISEASE IN CLOSELY RELATED *ASPERGILLUS* FUNGI 55

INTRODUCTION 55

MATERIALS AND METHODS 58

Genome procurement, assembly, and annotation 58

Inference of Gene Families 59

Phylogenomic data matrix construction and analyses 60

Gene family history 62

Gene Ontology Enrichment Analyses 63

Gene Family Expansions and Contractions 64

Estimating rates of molecular evolution 65

Amoeba Predation Assays 68

Virulence assays in the great wax moth (Galleria mellonella) model of fungal disease 68

Data Availability 69

RESULTS 70

A genome-scale phylogeny of Aspergillus section Fumigati 70

Broad conservation of genes and gene families, including those related to virulence, across section Fumigati 70

The distributions of few gene families are associated with pathogenicity 73

Many genes experienced faster rates of evolution in pathogenic species 75

Transcription factors with pathogenicity-related patterns of evolution have diverse effects on virulence 76

DISCUSSION 81

CHAPTER 4: VARIATION AMONG BIOSYNTHETIC GENE CLUSTERS, SECONDARY METABOLITE PROFILES, AND CARDS OF VIRULENCE ACROSS *ASPERGILLUS* SPECIES **88**

INTRODUCTION 88

MATERIALS AND METHODS 92

Strain acquisition, DNA extraction, and sequencing 92

Genome assembly, quality assessment, and annotation 92

Maximum likelihood phylogenetics and Bayesian estimation of divergence times 95

Identification of gene families and analyses of putative biosynthetic gene clusters ... 97

Identification and characterization of secondary metabolite production 99

Data Availability 104

RESULTS 105

Conservation and diversity of biosynthetic gene clusters within and between species 105

Within and between species variation in secondary metabolite profiles of A. fumigatus and its closest relatives 110

Conservation and divergence among biosynthetic gene clusters implicated in A. fumigatus pathogenicity 114

DISCUSSION 120

CHAPTER 5: GENOMIC AND PHENOTYPIC ANALYSIS OF COVID-19-ASSOCIATED PULMONARY ASPERGILLOSIS ISOLATES OF *ASPERGILLUS FUMIGATUS* **126**

INTRODUCTION 126

MATERIALS AND METHODS 129

Patient information and ethics approval 129

DNA quality control, library preparation, and sequencing 129

Genome assembly and annotations 130

Polymorphism identification 131

Maximum likelihood molecular phylogenetics 132

Biosynthetic gene cluster prediction 135

Characterization of biosynthesized secondary metabolites 136

Infection of Galleria mellonella 138

Growth assays 139

Data Availability 139

RESULTS 140

CAPA isolates belong to A. fumigatus and are closely related to reference strains Af293 and A1163 140

CAPA isolate genomes contain polymorphisms in genetic determinants of virulence and biosynthetic gene clusters 142

CAPA isolates display strain heterogeneity in virulence and in a few virulence-related traits 147

DISCUSSION 150

CHAPTER 6: EXTENSIVE COPY NUMBER VARIATION IN FERMENTATION-RELATED GENES AMONG *SACCHAROMYCES CEREVISIAE* WINE STRAINS 152

INTRODUCTION 152

MATERIALS AND METHODS 155

Data Mining, Quality Control and Mapping 155

CN Variant Identification 156

Diversity in CN Variation and GO Enrichment 158

Identifying Loci Absent in the Reference Strain 159

RESULTS 159

Descriptive Statistics of CN variation 159

CN Diversity in Subtelomeres 161

GO Enrichment of CN Diverse Genes 162

Genic CN Diversity 164

Functional Implications CN Variable Genes 169

Identifying loci absent from CN variation analysis 173

DISCUSSION 173

CHAPTER 7: EXTENSIVE LOSS OF CELL CYCLE AND DNA REPAIR GENES IN AN ANCIENT LINEAGE OF BIPOLAR BUDDING YEASTS 178

INTRODUCTION 178

MATERIALS AND METHODS 184

DNA sequencing 184

Phenotyping 184

Genome assembly and annotation 184

Assembly assessment and identification of orthologs 187

Phylogenomic analyses 189

Gene presence and absence analysis 192

Ploidy estimation 194

Molecular evolution and mutation analysis 195

RESULTS 200

An exceptionally high evolutionary rate in the FEL stem branch 200

The genomes of FEL species have lost substantial numbers of genes 201

FEL gene losses are associated with accelerated sequence evolution 212

Greater sequence instability in the FEL and signatures of endogenous and exogenous DNA damage 217

DISCUSSION 220

CHAPTER 8: EXAMINATION OF GENE LOSS IN THE DNA MISMATCH REPAIR PATHWAY AND ITS MUTATIONAL CONSEQUENCES IN A FUNGAL PHYLUM 227

INTRODUCTION 227

MATERIALS AND METHODS 233

Curation of the set of DNA mismatch repair pathway genes 233

MMR gene conservation analysis 234

Microsatellite identification and characterization 235

Estimation of mutational bias and rate of sequence evolution 236

Data availability 237

RESULTS 237

MMR genes are highly conserved across the fungal phylum Ascomycota 237

Extensive loss of MMR genes in a lineage of powdery mildews 238

Higher MMR gene loss taxa show increased number and length of microsatellites 241

Higher loss taxa show mutational biases 242

Higher loss taxa have experienced accelerated rates of sequence evolution 244

DISCUSSION 244

CHAPTER 9: PATHOGENIC ALLODIPLOID HYBRIDS OF *ASPERGILLUS* FUNGI 251

INTRODUCTION 251

MATERIALS AND METHODS 254

Fluorescence-assisted cell sorting for DNA content determination 255

Asexual spore size measurements 255

DNA extraction and sequencing 255

Genome assembly and annotation 256

Prediction of secondary metabolic gene clusters 256

Assigning genes in hybrid genomes to parents of origin 256

Maximum likelihood phylogenetic and phylogenomic analyses 258

Examination of loss of heterozygosity using copy number variation analysis 260

Macrophage isolation 261

In vitro phagocytosis by macrophages 262

Viability of Aspergillus hyphae 262

NETosis assays 263

Growth in the presence of different stresses 265

Hydrogen peroxide tolerance 265

Antifungal susceptibility assays 266

Data availability 267

RESULTS 267

Six clinical isolates previously characterized as A. nidulans are diploid 267

Diploid clinical isolates are Aspergillus latus, a species of hybrid origin 270

The genomes of the A. latus allodiploid hybrid isolates are stable 275

Hybrids exhibit wide variation for infection-relevant traits 277

DISCUSSION 280

CHAPTER 10: BIOKIT: A VERSATILE TOOLKIT FOR PROCESSING AND ANALYZING DIVERSE TYPES OF SEQUENCE DATA.....287

INTRODUCTION.....287

MATERIALS AND METHODS.....289

Genome assembly quality assessment.....290

Processing and assessing the properties of multiple sequence alignments.....291

Examining features of coding sequences including relative synonymous codon usage.....293

Implementing high standards of software development.....294

RESULTS.....295

Genome assembly quality and characteristics among 901 eukaryotic genomes.....295

Properties of multiple sequence alignment from 10 phylogenomic studies.....297

Relative synonymous codon usage in 107 budding yeast and filamentous fungi.....299

Patterns of gene-wise codon usage bias can be used to assess codon optimization and predict steady-state gene expression levels.....300

DISCUSSION.....303

CHAPTER 11: PHYKIT: A BROADLY APPLICABLE UNIX SHELL TOOLKIT FOR PROCESSING AND ANALYZING PHYLOGENOMIC DATA 304

INTRODUCTION.....304

MATERIALS AND METHODS.....306

Evaluating information content and biases in phylogenomic datasets.....307

Calculating gene-gene evolutionary rate covariation or coevolution.....311

Identifying polytomies in phylogenomic data.....312

Data availability 313

RESULTS 314

Summarizing information content and biases in phylogenomic data 314

A network of gene-gene covariation reveals neighborhoods of genes with shared function 316

Identifying polytomies in phylogenomic datasets 319

DISCUSSION 321

CHAPTER 12: CLIPKIT: A MULTIPLE SEQUENCE ALIGNMENT-TRIMMING SOFTWARE FOR ACCURATE PHYLOGENOMIC INFERENCE 323

INTRODUCTION 323

MATERIALS AND METHODS 325

ClipKIT availability and usage 325

Practical considerations when using ClipKIT 331

Dataset acquisition and generation 332

Measuring accuracy and support of phylogenetic inferences 334

Software availability 336

Data availability 336

RESULTS 337

DISCUSSION 342

CHAPTER 13: ORTHOSNAP: A TREE SPLITTING AND PRUNING ALGORITHM FOR RETRIEVING SINGLE-COPY ORTHOLOGS FROM GENE FAMILY TREES 344

INTRODUCTION 344

MATERIALS AND METHODS 348

orthoSNAP availability and documentation 348

orthoSNAP algorithm description and usage 348

Development practices and design principles to ensure long-term software stability 350

Dataset generation 351

Measuring and comparing information content among SC-OGs and SNAP-OGs ... 352

Data Availability 354

RESULTS 355

SC-OGs and SNAP-OGs have similar information content 355

SC-OGs and SNAP-OGs have similar patterns of support in a contentious branch in the tree of life 358

DISCUSSION 360

CHAPTER 14: ORTHOFISHER: A BROADLY APPLICABLE TOOL FOR AUTOMATED GENE IDENTIFICATION AND RETRIEVAL 363

INTRODUCTION 363

MATERIALS AND METHODS 365

RESULTS 368

orthofisher and BUSCO obtain similar results 368

orthofisher and BUSCO perform similarly to OrthoFinder 370

orthofisher is helpful for estimating the number of members in a gene family 372

DISCUSSION 372

CHAPTER 15: TREEHOUSE: A USER-FRIENDLY APPLICATION TO OBTAIN SUBTREES FROM LARGE PHYLOGENIES 374

INTRODUCTION 374

MATERIALS AND METHODS 375

Data acquisition 375

Description of the software 375

RESULTS 376

A three-step workflow to obtain subtrees 376

DISCUSSION 378

CHAPTER 16: GGPUBFIGS: COLORBLIND-FRIENDLY COLOR PALETTES AND GGPLOT2 GRAPHIC SYSTEM EXTENSIONS FOR PUBLICATION-QUALITY SCIENTIFIC FIGURES 379

INTRODUCTION 379

MATERIALS AND METHODS 380

RESULTS 382

DISCUSSION 382

CHAPTER 17: CONCLUDING DISCUSSION AND FUTURE DIRECTIONS 383

REFERENCES 387

List of Figures

Figure 1. A robust genome-scale phylogeny for the fungal family *Aspergillaceae* 22

Figure 2. Topological similarity between the 36 phylogenies constructed using 6 different data matrices, 2 different sequence types, and 3 analytical schemes 35

Figure 3. The eight internodes not recovered in all 36 phylogenies 37

Figure 4. The three internodes recovered in all 36 phylogenies but that exhibit very low internode certainty values 39

Figure 5. A visual comparison of the differences between the phylogeny reported in this study and the phylogeny reported in the work of Kocsubé et al. 40

Figure 6. A molecular time tree for the family *Aspergillaceae* 42

Figure 7. Genome-scale phylogeny and evolution of net gene gains or losses across *Aspergillus* section Fumigati. 61

Figure 8. Gene families are largely conserved across section Fumigati, regardless of pathogenicity level 63

Figure 9. Genes in section Fumigati exhibit both pathogen- and species-specific rates of evolution. 66

Figure 10. Multiple transcription factors whose evolution varies with respect to *Aspergillus* pathogenicity affect the survival of *A. fumigatus* during amoeba predation. 77

Figure 11. Multiple transcription factors in *A. fumigatus* whose evolution differs with respect to pathogenicity affect virulence in the greater wax moth model of disease 79

Figure 12. Diverse genetic repertoire of biosynthetic gene clusters and extensive presence and absence polymorphisms between and within species 106

Figure 13. *Aspergillus oerlinghausenensis* shares more gene families and BGCs with *A. fischeri* than *A. fumigatus* 109

Figure 14. *A. oerlinghausenensis* and *A. fischeri* have more similar secondary metabolite profiles than *A. fumigatus* 112

Figure 15. Conservation in the gliotoxin BGC correlates with conserved production of gliotoxin analogs in A. fumigatus and nonpathogenic close relatives 116

Figure 16. Conservation and divergence in the locus encoding the fumitremorgin and intertwined fumagillin/pseurotin BGCs 119

Figure 17. Secondary metabolism-associated "cards" of virulence among A. fumigatus and close relatives 124

Figure 18. Inhalation of *Aspergillus* spores can result in fungal infection. 127

Figure 19. Phylogenomics confirms that COVID-19-associated pulmonary aspergillosis (CAPA) isolates are *Aspergillus fumigatus* 141

Figure 20. Mutational spectra among genetic determinants of virulence 143

Figure 21. COVID-19-associated pulmonary aspergillosis (CAPA) isolates of *Aspergillus fumigatus* have biosynthetic gene clusters (BGCs) that encode the toxic small molecule gliotoxin 146

Figure 22. Strain heterogeneity among COVID-19-associated pulmonary aspergillosis (CAPA) isolates of Aspergillus fumigatus 148

Figure 23. Size distribution and location of CN variable loci 160

Figure 24. GO enriched terms for high CN diverse genes 163

Figure 25. CN variation of genes and gene families 165

Figure 26. Model summary of CN variable genes in wine yeast strains and their cellular functions 172

Figure 27. The evolutionary history, rate, and timeline of *Hanseniaspora* diversification 191

Figure 28. dN/dS (ω) analyses support a historical burst of accelerated evolution in the FEL 197

Figure 29. Gene presence and absence analyses reflect phenotype and reveal disrupted pathways 203

Figure 30. Gene presence and absence in the budding yeast cell cycle 208

Figure 31. A panoply of genome-maintenance and DNA repair genes are absent among *Hanseniaspora*, especially in the FEL 211

Figure 32. Analyses of base substitutions and indels reveal a higher mutational load in the FEL compared to the SEL 214

Figure 33. The DNA Mismatch Repair (MMR) pathway corrects mismatched bases produced during DNA replication and prevents instability in microsatellites 228

Figure 34. Conservation of mismatch repair (MMR) pathway genes across the fungal phylum Ascomycota 231

Figure 35. The powdery mildews *Erysiphe* and *Blumeria* have lost many more mismatch repair (MMR) pathway genes than closely related species 232

Figure 36. Genomes of higher loss taxa (HLT; blue bars) show a proliferation of mononucleotide runs and an increase in their microsatellite lengths compared to lower loss taxa (LLT; grey bars) 242

Figure 37. Higher loss taxa (HLT) show diverse types of mutational bias compared to lower loss taxa (LLT) 243

Figure 38. Powdery mildew higher loss taxa (HLT) show accelerated rates of evolution 245

Figure 39. Six Clinical Isolates Previously Characterized as *Aspergillus nidulans* and the Type Strain of *Aspergillus latus* Are Diploids 268

Figure 40. The 6 Clinical Diploids Belong to *A. latus*, an Allodiploid Species Formed via Hybridization of *A. spinulosporus* and a Close Relative of *A. quadrilineatus* 271

Figure 41. *A. latus* Hybrids Exhibit Strain Heterogeneity and Differ from Parental Species, *A. quadrilineatus*, and *A. nidulans* in Infection-Relevant Phenotypes 278

Figure 42. Proposed Model for the Evolution of *A. latus* via Allodiploid Hybridization 284

Figure 43. Summary of genome assembly metrics across 901 genomes from three eukaryotic classes 296

Figure 44. Summary metrics among multiple sequence alignments from phylogenomic studies 298

Figure 45. Relative synonymous codon usage across 171 fungal genomes 299

Figure 46. Mean gene-wise relative synonymous codon usage accurately estimates codon optimization 301

Figure 47. Summary of information content in four empirical phylogenomic datasets 315

Figure 48. Gene–gene covariation network inferred from ~550 million years of evolution across 1107 fungi 318

Figure 49. Identifying polytomies from phylogenomic data 320

Figure 50. The 14 alignment trimming strategies tested differ in resulting MSAs and metrics of phylogenetic tree accuracy and support 338

Figure 51. ClipKIT is a top-performing software for trimming MSAs 340

Figure 52. Cartoon depiction of three classes of paralogs: outparalogs, inparalogs, and coorthologs 345

Figure 53. Cartoon depiction of orthoSNAP workflow 347

Figure 54. SC-OGs and SNAP-OGs display similar patterns of support in a contentious branch concerning deep evolutionary relationships among placental mammals 355

Figure 55. SC-OGs and SNAP-OGs have similar phylogenetic information content 357

Figure 56. Workflow overview for orthofisher 366

Figure 57. A simple three-step workflow for using *treehouse* 377

Figure 58. Examples of ggplot2 extensions and color palettes available in ggpubfigs 381

List of Tables

Table 1. Topology tests reject the sister group relationship of genus *Penicillium* and *Aspergillus* section *Nidulantes* as well as the monophyly of narrow *Aspergillus*. 63

Table 2. Species and strains used in the present study. 111

Table 3. Table 3. Select A. fumigatus secondary metabolites implicated in modulating host biology 130

nucleotide (NT) and amino acid (AA); Credible Interval (CI); Millimolar (mM); Ethylenediaminetetraacetic acid (EDTA); relative composition variability (RCV); ultrafast bootstrap approximation approach (UFBoot); internode certainty (IC); gene-wise log-likelihood scores (GLS); gene support frequencies (GSF); difference in GLS (ΔGLS); GSF for NT (GSF_{NT}) and AA (GSF_{AA}); resampling estimated log-likelihood (RELL); Markov chain Monte Carlo (MCMC); degree of violation of a molecular clock (DVMC); rate of nonsynonymous substitutions (dN); rate of synonymous substitutions (dS); null hypothesis (H_O); alternative hypothesis (H_A); Czapek-Dox medium (CZD); transcription factor-encoding (TF); biosynthetic gene clusters (BGCs); Northern Regional Research Laboratory (NRRL); yeast extract soy peptone dextrose (YESD); photodiode array detector (PDA); potato dextrose agar (PDA); ultraperformance liquid chromatography-photodiode array-electrospray ionization high resolution tandem mass spectrometry (UPLC-PDA-HRMS-MS/MS); Principal component analysis (PCA); hexadehydroastechrome (HAS); Really Interesting New Gene (RING); moderate to severe respiratory distress syndrome (ARDS); loss of function (LOF); single-strand circle DNA (ssCir DNA); single nucleotide polymorphisms (SNPs); insertion-deletion polymorphisms (indels); copy number (CN); Burrows-Wheeler Aligner (BWA); Phosphate buffered saline (PBS); COVID-19-associated pulmonary aspergillosis (CAPA); ATP binding cassette (ABC); base pairs (bp); Polymorphic Index Content (PIC); Gene ontology (GO); megabases (Mb); kilobase (kb); CN variable regions (CNVRs); wild-type (WT); Acireductone Dioxygenase (ADI); Anaphase-Promoting Complex (APC); AROmatic amino-acid requiring (ARO); Associated with Spindles and Kinetochores 1 (ASK1); Australian Wine Research Institute 3580 (AWRI3580); Branched-chain Amino-acid Transaminase (BAT); Centraalbureau voor Schimmelcultures 314 (CBS 314);

Cell Division Cycle 13 (CDC13); credible interval (CI); Death Upon Overproduction 1 And Death Upon Overproduction 1 and MonoPolar Spindle 1 interacting (DAD); Duo1 and MonoPolar Spindle 1 (DAM1); Dam1 Complex (DASH); rate of nonsynonymous substitutions (dN); rate of synonymous substitutions (dS); Daughter-Specific Expression 2 (DSE2); Dutch State Mines 2768; DSN1 (DSM2768); Death Upon Overproduction (DUO); E2 promoter binding Factor (E2F); Eocene (Eo); EXOnuclease 1 (EXO1); Fructose-1,6-BisPhosphatase 1 (FBP1); faster-evolving lineage (FEL); GALactose metabolism (GAL); Guanine–Cytosine (GC); Glutamate DeHydrogenase (GDH); genomic DNA (gDNA); Glycogen 7-Interacting Protein 1 (GIP1); GLuTamate synthase (GLT); Hidden Markov Model (HMM); Helper of ASK1 3 (HSK3); IsoMAltase (IMA); Inducer of MEiosis 1 (IME1); likelihood ratio test (LRT); Mitotic Arrest-Deficient 1 (MAD1); 3-MethylAdenine DNA Glycosylase 1 (MAG1); MALtose fermentation (MAL); MALtose fermentation locus 1 or 3 (MALx); Mini-Chromosome Maintenance (MCM); Methylthioribulose-1-phosphate Dehydratase (MDE); Mitosis Entry Checkpoint 3 (MEC3); Multicopy Enhancer of Upstream activation site (MEU); Mis TWelve-like 1 protein Including Necessary for Nuclear Function 1 protein,Nnf1 Synthetic Lethal 1 protein, Dosage Suppressor of Necessary for Nuclear Function 1 protein complex (MIND); Miocene (Mio.); Meiotic Nuclear Divisions 2 (MND2); Mediator of the Replication Checkpoint 1 (MRC1); MethylthioRibose-1-phosphate Isomerase (MRI); Mis TWelve-like 1 (MTW1); million years ago (mya); Necessary for Nuclear Function 1 (NNF1); Northern Regional Research Laboratory (NRRL); Nnf1 Synthetic Lethal 1 (NSL1); orthologous gene (OG); Oligocene (Oligo.); Origin Recognition complex (ORC); Paleocene (Paleo.); Pbp1p Binding Protein 2 (PBP2); Peroxisomal Coenzyme A Diphosphatase 1 (PCD1); Phosphoenolpyruvate CarboxyKinase 1 (PCK1); Pho85 CycLin 1 (PCL1); Precocious Dissociation of Sisters 1 (PDS1); PHOsphate metabolism (PHO);

PHotoreactivation Repair deficient (PHR1); Pleistocene (Pleisto.); Pliocene (Plio.); POLymerase (POL); Quaternary (Quat); RADiation sensitive 9 (RAD9); Replication Factor A (RFA3); Regulatory Factor X 1 (RFX1); Repressor/activator site binding protein-Interacting Factor 1 (RIF1); S-AdenosylMethionine requiring (SAM); Switching deficient 4/6 cell-cycle box-binding factor (SBF); slower-evolving lineage (SEL); Slow Growth Suppressor 1 (SGS1); Substrate/Subunit Inhibitor of Cyclin-dependent protein kinase 1 (SIC1); SNooze proximal Open reading frame (SNO); Spindle Pole Component (SPC); SPErmidine auxotroph (SPE); SPOrulation 12 (SPO12); Sporulation-specific protein 1 (SSP1); SUCrose 2 (SUC2); Spore Wall Maturation 1 (SWM1); Tyrosyl-DNA Phosphodiesterase 1 (TDP1); THIamine regulon (THI); University of Trás-os-Montes and Alto Douro 222 (UTAD222); Unidentified Transcript (UTR); WHIskey 5 (WHI5); Yeast KU protein 70 (YKU70); mismatch repair (MMR); microsatellite instability (MSI); higher loss taxa (HLT); lower loss taxa (LLT); Kyoto Encyclopedia of Genes and Genomes (KEGG); *Schizosaccharomyces pombe* database (PomBase); *Saccharomyces* Genome Database (SGD); profile Hidden Markov Models (pHMMs); Interactive Tree of Life (iTOL); Microsatellite Identification (MISA); higher loss taxa (HLT); Honest Significant Differences (HSD); transition to transversion (Ts/Tv); repeat-induced point (RIP); classical nonhomologous end joining (C-NHEJ); chronic granulomatous disease (CGD); Secondary metabolic gene clusters (SMGCs); false discovery rate (FDR); false positive rate (FPR); polymorphonuclear cells (PMN); phorbol 12-myristate 13-acetate (PMA); Forward Scatter Height (FSC-H); Forward Scatter Area (FSC-A); Side Scatter Area (SSC-A); minimal inhibitory concentration (MIC); *A. latus* (Alat); *A. spinulosporus* (Aspi); *A. quadrilineatus* (Aqua); and *A. nidulans* (Anid); near-universally single-copy orthologous (BUSCO) genes; guanine-cytosine (GC); gene-wise relative synonymous codon usage (gw-RSCU); Relative synonymous codon usage (RSCU); Multiple sequence alignments

(MSAs); average bipartition support (ABS); Block Mapping and Gathering with Entropy (BMGE); continuous integration (CI); general time reversible (GTR); normalized Robinson–Foulds (nRF); Whelan and Goldman (WAG); single-copy orthologs (SC-OGs); splitting and pruning orthologs (SNAP-OGs); Robinson-Foulds (RF); parsimony informative (PI); alignment length (Aln. len.)

CHAPTER 1

Introduction

Billions of base pairs: deep sequencing of the fungal kingdom

Sequencing of nucleotide molecules has advanced diverse biology disciplines including evolutionary biology (Rokas and Abbot, 2009; Houldcroft et al., 2017; Manolio et al., 2021). In particular, genome sequencing has shed light on grand challenges in the field of evolutionary biology such as the genetic underpinnings of aging, human population history, and the tempo and mode of evolution across diverse lineages (An integrated map of genetic variation from 1,092 human genomes, 2012; Jarvis et al., 2014; Feng et al., 2017; Shen et al., 2018; Choin et al., 2021; Kolora et al., 2021; Li et al., 2021).

Among eukaryotic lineages, nuclear genomes of species from the Kingdom Fungi, an ancient and diverse lineage estimated to contain approximately 2-5 million species (Blackwell, 2011; Hawksworth and Lücking, 2017), were among the first to be sequenced. In fact, *Saccharomyces cerevisiae*, the model baker's or brewer's yeast, was the first eukaryotic nuclear genome to be sequenced (Goffeau et al., 1996). As of January 7^{th,} 2022, nearly 10,000 fungal genomes are available (https://www.ncbi.nlm.nih.gov/). These rich genomic resources have enabled researchers to shed light on the dynamics of genome evolution in numerous fungal lineages such as species of yeast, filamentous fungi, mycorrhizal fungi, mushroom-forming fungi, and others as well as within species, such as *S. cerevisiae* and *Schizosaccharomyces pombe* (Nagy et al., 2014; Jeffares et al., 2015a; Gallone et al., 2016; Nagy et al., 2016; Jeffares et al., 2017; Peter et al., 2018; Shen et al., 2018; Kjærbølling et al., 2020; Mead et al., 2020; Miyauchi et al., 2020).

Among other findings, these studies have underscored the importance of evolution by gene duplication, loss, and retention, horizontal gene transfer, and hybridization/introgression.

Fungi, champions of ecological and genomic diversity

Notwithstanding these discoveries, the fungal kingdom remains largely unexplored and numerous outstanding questions remain unresolved. For example, one major branch of research aims to determine what makes some fungi pathogenic whereas others are harmless or even beneficial to human welfare (Fedorova et al., 2008; Butler et al., 2009; Moran et al., 2011; Shang et al., 2016; Rokas et al., 2020a; Singh-Babak et al., 2021). The spectrum of pathogenic-to-beneficial-to-human-welfare observed in fungi is particularly striking among the sister genera *Aspergillus* and *Penicillium* fungi wherein some species are major pathogens of humans (e.g., *Aspergillus fumigatus*, *Aspergillus flavus*) or of plants (e.g., *Penicillium digitatum*, *Penicillium citrinum*) whereas other are used in the production of fermented foods (e.g., *Aspergillus oryzae*, *Aspergillus sojae*, *Penicillium roqueforti*, *Penicillium nalgiovense*) or diverse biomolecules (e.g., *Aspergillus niger*, *Aspergillus nidulans*, *Penicillium decumbens*) (Houbraken and Samson, 2011; Houbraken et al., 2014; Samson et al., 2014; Visagie et al., 2014; Tsang et al., 2018; Steenwyk et al., 2019c; Rokas et al., 2020a). As a result, fungi from these lineages serve as valuable models to study the evolution of diverse fungal lifestyles that have equally diverse impact on humans.

The diversity of fungal lifestyles, even among closely related species, is in part caused by the rapid tempo of evolution among fungi (Fedorova et al., 2008; Shen et al., 2018, 2020b). Determining what causes variation in evolutionary rates can shed light on important topics such as pathogen microevolution, which can contribute to the emergence of antibiotic resistance and

recurrent pathogen infection (Davies and Davies, 2010; Billmyre et al., 2017; Rhodes et al., 2017a; Steenwyk, 2021a), as well as the domestication of fungi (Gibbons and Rinker, 2015; Gallone et al., 2016; Bodinaku et al., 2019). For example, the fungal pathogen *Candida glabrata* has a diminished response to DNA damage compared to *S. cerevisiae*, which is thought to contribute to antifungal resistance (Shor et al., 2020), an important component of pathogenicity. Hybridization, the genetic crossing of distinct lineages, can also be a rapid driver of evolution among fungal pathogens (Neafsey et al., 2010; Stukenbrock, 2016; Mixão and Gabaldón, 2020). Signatures of rapid evolution can also be observed within populations of a single species. For example, studies investigating copy number variants (duplicated or deleted loci in a population) have revealed a mutation rate 100-1,000 times that of single nucleotide polymorphisms (Zhang et al., 2009; Sener, 2014; Steenwyk and Rokas, 2018). A comprehensive map of copy number variants in populations can shed light on their evolutionary dynamics and reveal regions of the genome that are more (or less) likely to harbor this type of variation. In summary, studies investigating the drivers of rapid mutation as well as their mutational landscape across the genome will shed light on diverse aspects of fungal biology including pathogenesis, domestication, and, more broadly, fungal ecology.

Bioinformatics, a young field with growing pains

The aforementioned studies have relied on an unprecedented amount of genomic data. To keep pace with data generation, technical and methodological advances—which often require interdisciplinary teams of software engineers, biologists, and others—occurred concomitantly (Muir et al., 2016). For example, numerous software generate and assess the quality and completeness of genome assemblies and gene annotations, a key first step for many studies

(Stanke and Waack, 2003; Korf, 2004; Holt and Yandell, 2011; Bankevich et al., 2012; Gurevich et al., 2013; Waterhouse et al., 2018a). Other software aims to make use of genomes and gene annotations by conducting ortholog identification, multiple sequence alignment and alignment trimming, or reconstructing the evolutionary histories of nucleotide and amino acid sequences. The output files from these pieces of software can be used to infer diverse kinds of evolutionary events such as evolutionary radiations of species lineages and positive selection of individual codons in a gene's coding region (Lanyon, 1988; Phillips and Penny, 2003; Yang, 2007; Capella-Gutierrez et al., 2009; Katoh and Standley, 2013; Salichos and Rokas, 2013; Stamatakis, 2014a; Struck, 2014; Liu et al., 2017; Waterhouse et al., 2018a; Zhang et al., 2018; Emms and Kelly, 2019; Minh et al., 2020).

As described above, the output files from one software are often the input files for another resulting in bioinformatic workflows that rely on numerous pieces of software, custom scripts, and manual examination or processing of input/output files. This can lead to complex and difficult-to-maintain bioinformatic pipelines that threaten scientific reproducibility (Mangul et al., 2019a, 2019b). Case in point, a recent study found that approximately 28% of bioinformatic software are no longer supported by developers and fail to install due to implementation errors (Mangul et al., 2019b). A non-exhaustive list of reasons contributing to this issue may include: that the academe rewards publications more than maintenance of the tools that led to these publications; trainees who lead the project may move onto different jobs or fields; and inadequate training for biologists. Although some issues may only be resolved on a case-by-case basis, there are some principles that can be implemented by software engineers to help ensure long-term software stability. For example, developers can utilize integration and unit testing

coupled to continuous integration pipelines, an often overlooked component of software development that automatically tests building, packaging, installation, and faithful functionality of the software.

Another outstanding issue is that analyses may require a combination of web-server applications, scripts available through repositories and supplemental materials, standalone software, and/or extensive programming in diverse languages. As a result, bioinformatic pipelines are not only difficult to maintain, but their accessibility to the scientific community is stymied. One approach to addressing this issue is to develop unified toolkits, software that conduct diverse analyses. However, engineering and developing unified toolkits often require large collaborations (or even consortiums) resulting in software that is difficult to coordinate, maintain, and deploy. In summary, despite methodologic advances, there are still numerous areas ripe for improvement in the field of bioinformatics.

Insights into fungal genome evolution and new software for the life sciences

In this thesis, I describe work that aims to address these shortcomings: elucidating the evolutionary dynamics of fungal genomes (chapters one through seven) and methods/software development to enable scientific discovery (chapters eight through 14). Studies of fungal genome evolution focus primarily on medically and technologically relevant fungi (e.g., pathogens and wine-associated yeast). Software described herein focus primarily on methods for evolutionary genomic studies.

In chapters two through nine, I describe dynamics of fungal genome evolution across and within species from the phylum Ascomycota, a diverse phylum with at least 83,000 known species (James et al., 2020; Shen et al., 2020b). In chapter two, I describe a robust workflow for investigating the evolutionary relationships among fungi, a prerequisite for understanding genome evolution, using species from the biomedically and technologically significant *Aspergillus* and *Penicillium* genera (Steenwyk et al., 2019c). In chapters three and four, I highlight how evolutionary-guided approaches, which leverage workflows and findings from chapter two, can be used to shed light on the evolutionary makings of a fungal pathogen (Steenwyk et al., 2020d; Mead et al., 2021). To briefly foreshadow their findings, these two chapters are surprising in that numerous genetic determinants of virulence are found in nonpathogenic fungi, which raises an important question—*"what makes a fungal pathogen?"* In chapter five, I highlight an international genome sequencing and phenotyping effort to characterize coronavirus disease 2019 (COVID-19) associated pulmonary isolates of *A. fumigatus* and raise awareness of the clinical importance of superinfections (Steenwyk et al., 2021d).

In chapters six through nine, I describe rapid evolutionary processes among fungi. In chapter six, I discuss how copy number variants—duplicated and deleted loci in a population—can contribute to substantial variation in a population despite low genetic variation among single nucleotide polymorphisms (Steenwyk and Rokas, 2017). In chapters seven and eight, I discuss how losses of DNA repair and cell cycle genes, which collectively contribute to genome stability, can lead to punctuated sequence evolution (Steenwyk et al., 2019a; Phillips et al., 2021). In chapter nine, I discuss how allodiploid hybridization—the combining of whole

genomes from distinct parental species—contributes to the evolution of a pathogenic fungus, *Aspergillus latus* (Steenwyk et al., 2020c).

The remaining seven chapters are dedicated to software that facilitate the processing, analysis, or plotting of biological data, such as sequence and phylogenomic data. In chapters 10 and 11, I describe BioKIT and PhyKIT—two pieces of software that, among other things, can be used to assess the information content in multiple sequence alignments and phylogenetic trees as well as infer evolutionary events such as rapid radiations (Steenwyk et al., 2021a, 2021b). In chapter 12, I describe a novel approach to multiple sequence alignment trimming that focuses on retaining phylogenetically informative sites (Steenwyk et al., 2020b). In chapter 13, I describe orthoSNAP, a tree-splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees, which can facilitate creating larger data matrices for molecular evolutionary studies (Steenwyk et al., 2021c). In chapter 14, I describe orthofisher, a toolkit for putative ortholog identification and retrieval, a common first step in many bioinformatic workflows (Steenwyk and Rokas, 2021b). In chapters 15 and 16, I describe relatively simple pieces of software that aim to increase the accessibility of scientific findings using two different approaches. In chapter 15, I describe treehouse, a graphical user-interface software that allows users to obtain subtrees from phylogenies, which enables researchers not familiar with phylogenetic software to quickly obtain the evolutionary relationships of their species of interest (Steenwyk and Rokas, 2019). In chapter 16, I describe ggpubfigs, an R package with ggplot2 extensions (Wickham, 2009), that facilitate creating publication quality figures that are also colorblind friendly (Steenwyk and Rokas, 2021a).

Thereafter, I discuss future avenues of research that may build upon these findings or software. A fair summary of this section is that there is an abundance of exciting research to be done and, to date, my humble contributions, enabled by a supportive network of collaborators, are minimal. Nonetheless, it is my sincere hope that the research described herein—which has been an immense honor and privilege to conduct—inspires, informs, and enables future biological research.

CHAPTER 2

A robust phylogenomic timetree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*[1]

Introduction

The vast majority of the 1,062 described species from the family Aspergillaceae (phylum Ascomycota, class Eurotiomycetes, order Eurotiales) (Houbraken et al., 2014) belong to the genera *Aspergillus* (42.5%; 451 / 1,062) and *Penicillium* (51.6%; 549 / 1,062) (Benson et al., 2007; Sayers et al., 2009). Fungi from Aspergillaceae exhibit diverse ecologies; for example, *Penicillium verrucosum* is widespread in cold climates but has yet to be isolated in the tropics (Pitt, 2002), whereas *Aspergillus nidulans* is able to grow at a wide range of temperatures but favors warmer ones (Ogundero, 1983). Several representative species in the family are exploited by humans, while a number of others are harmful to humans or their activities (Gibbons and Rokas, 2013). Examples of useful-to-humans organisms among *Aspergillus* species include *Aspergillus oryzae*, which is used in the production of traditional Japanese foods including soy sauce, sake, and vinegar (Machida et al. 2008; Gibbons et al. 2012) as well as of amylases and proteases (Kobayashi et al., 2007) and *Aspergillus terreus*, which produces mevinolin (lovastatin), the potent cholesterol-lowering drug (Albert et al., 1980). Examples of useful-to-humans *Penicillium* species include *Penicillium camemberti* and *Penicillium roqueforti*, which contribute to cheese production (Nelson 1970; Lessard et al. 2012), and *Penicillium citrinum*,

[1]This work is published in: Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H., and Rokas, A. (2019). A Robust Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the Genera *Aspergillus* and *Penicillium*. MBio 10. doi:10.1128/mBio.00925-19.

which produces the cholesterol lowering drug mevastatin, the world's first statin (Endo 2010). In contrast, examples of harmful-to-humans organisms include the pathogen, allergen, and mycotoxin-producing species *Aspergillus fumigatus* and *Aspergillus flavus* (Nierman et al., 2005; Hedayati et al., 2007) and the post-harvest pathogens of citrus fruits, stored grains, and other cereal crops *Penicillium expansum*, *Penicillium digitatum*, and *Penicillium italicum* (Marcet-Houben et al., 2012; Ballester et al., 2015; Li et al., 2015).

Much of the ubiquity, ecological diversity, and wide impact on human affairs that Aspergillaceae exhibit is reflected in their phenotypic diversity, including their extremotolerance (e.g., ability to withstand osmotic stress and wide temperature range) (Magan and Lacey, 1984; Marín et al., 1998; Pitt and Hocking, 2009; Vinnere Pettersson and Leong, 2011) and ability to grown on various carbon sources (Pitt and Hocking, 2009; de Vries et al., 2017). Fungi from Aspergillaceae are also well known for their ability to produce a remarkable diversity of secondary metabolites, small molecules that function as toxins, signaling molecules, and pigments (Pitt, 1994; Keller et al., 2005; Frisvad and Larsen, 2015; Macheleidt et al., 2016; Rokas et al., 2018). Secondary metabolites likely play key roles in fungal ecology (Rohlfs et al., 2007; Fox and Howlett, 2008; Stierle and Stierle, 2015), but these small molecules often have biological activities that are either harmful or beneficial to human welfare. For example, the *A. fumigatus*-produced secondary metabolite gliotoxin is a potent virulence factor in cases of systemic mycosis in vertebrates (Rohlfs and Churchill, 2011), and the *A. flavus*-produced secondary metabolite aflatoxin is among the most toxic and carcinogenic naturally occurring compounds (Squire, 1981; Keller et al., 2005). In contrast, other secondary metabolites are mainstay antibiotics and pharmaceuticals; for example, the *Penicillium chrysogenum*-produced

penicillin is among the world's most widely used antibiotics (Chain et al., 1940; Fleming, 1980; Aminov, 2010) and the *P. citrinum*-produced cholesterol lowering statins are consistently among the world's blockbuster drugs (Endo, 2010).

Understanding the evolution of the diverse ecological lifestyles exhibited by Aspergillaceae members as well as the family's remarkable chemodiversity requires a robust phylogenetic framework. To date, most molecular phylogenies of the family Aspergillaceae are derived from single or few genes and have yielded conflicting results. For example, it is debated whether the genus *Aspergillus* is monophyletic or if it includes species from other genera such as *Penicillium* (Pitt and Taylor, 2014; Samson et al., 2014). Furthermore, studies using genome-scale amounts of data, which could have the power to resolve evolutionary relationships and identify underlying causes of conflict (Rokas et al., 2003; Salichos and Rokas, 2013), have so far tended to use a small subset of fungi from either *Aspergillus* or *Penicillium* (de Vries et al., 2017; Nielsen et al., 2017; Kjærbølling et al., 2018). Additionally, these genome-scale studies do not typically examine the robustness of the produced phylogeny; rather, based on the high clade support values (e.g., bootstrap values) obtained, these studies infer that the topology obtained is highly accurate (Yang et al., 2016; de Vries et al., 2017; Nielsen et al., 2017; Kjærbølling et al., 2018).

In recent years, several phylogenomic analyses have shown that high clade support values can be misleading (Phillips et al., 2004; Kumar et al., 2012; Salichos and Rokas, 2013), that incongruence, the presence of topological conflict between different data sets or analyses, is widespread (Hess and Goldman, 2011; Song et al., 2012; Salichos and Rokas, 2013; Zhong et al., 2013), and that certain branches of the tree of life can be very challenging to resolve, even with

genome-scale amounts of data (Shen et al. 2016; Suh 2016; Arcila et al. 2017; King and Rokas 2017; Shen et al. 2017). Comparison of the topologies inferred in previous phylogenomic studies in Aspergillaceae (Yang et al., 2016; de Vries et al., 2017; Nielsen et al., 2017; Kjærbølling et al., 2018) suggests the presence of incongruence (Figure S1 from Steenwyk et al., 2019c). For example, some studies have reported section *Nidulantes* to be the sister group to section *Nigri* (de Vries et al., 2017), whereas other studies have placed it as the sister group to *Ochraceorosei* (Kjærbølling et al., 2018) (Figure S1 from Steenwyk et al., 2019c).

A robust phylogeny of Aspergillaceae is also key to establishing a robust taxonomic nomenclature for the family. In recent years, the taxonomy of *Aspergillus* and *Penicillium* has been a point of contention due to two key differences among inferred topologies based on analyses of a few genes (Kocsubé et al., 2016; Taylor et al., 2016). The first key difference concerns the placement of the genus *Penicillium*. One set of analyses places the genus as a sister group to *Aspergillus* section *Nidulantes*, which would imply that *Penicillium* is a section within the genus *Aspergillus* (Taylor et al., 2016), whereas a different set of analyses suggests that the genera *Penicillium* and *Aspergillus* are reciprocally monophyletic (Kocsubé et al., 2016). The second key difference concerns whether sections *Nigri*, *Ochraceorosei*, *Flavi*, *Circumdati*, *Candidi*, and *Terrei*, which are collectively referred to as "narrow *Aspergillus*", form a monophyletic group (Taylor et al., 2016) or not (Kocsubé et al., 2016). Both of these differences are based on analyses of a few genes (4 loci, Taylor et al. 2016 and 9 loci, Kocsubé et al. 2016) and the resulting phylogenies typically exhibit low support values for deep internodes, including for the ones relevant to this debate.

To shed light on relationships among these fungi, we employed a genome-scale approach to infer the evolutionary history among *Aspergillus, Penicillium*, and other fungal genera from the family Aspergillaceae. More specifically, we used the genome sequences of 81 fungi from Aspergillaceae spanning 5 genera, 25 sections within *Aspergillus* and *Penicillium*, and 12 outgroup fungi to construct nucleotide (NT) and amino acid (AA) versions of a data matrix comprised of 1,668 orthologous genes. Using three different maximum likelihood schemes (i.e., gene-partitioned, unpartitioned, and coalescence), we inferred phylogenies from the 1,668-gene data matrix as well as from five additional 834-gene data matrices derived from the top 50% of genes harboring strong phylogenetic signal according to five different criteria (alignment length, average bootstrap value, taxon completeness, treeness / relative composition variability, and number of variable sites). Using the same schemes, we also inferred phylogenies of the 1,668-gene data matrix using different alignment trimming methods as well as of a reduced 1,331-gene data matrix that was filtered for potential hidden paralogs. Comparisons of these phylogenies coupled with complementary measures of internode certainty (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016) identified 14 / 78 (17.9%) incongruent bipartitions in the phylogeny of Aspergillaceae. These cases of incongruence can be grouped into three categories: (i) 2 shallow bipartitions with low levels of incongruence likely driven by incomplete lineage sorting, (ii) 4 shallow bipartitions with high levels of incongruence likely driven by hybridization or introgression (or very high levels of incomplete lineage sorting), and (iii) 8 deeper bipartitions with varying levels of incongruence likely driven by reconstruction artifacts likely linked with poor taxon sampling. We also estimated divergence times across Aspergillaceae using relaxed molecular clock analyses. Our results suggest Aspergillaceae originated in the lower Cretaceous, 117.4 (95% Credible Interval (CI): 141.5 - 96.9) million

years ago (mya), and that *Aspergillus* and *Penicillium* originated 81.7 mya (95% CI: 87.5 - 72.9) and 73.6 mya (95% CI: 84.8 - 60.7), respectively. We believe this phylogeny and timetree are highly informative with respect to the ongoing debate on *Aspergillus* systematics and taxonomy, and provide a state-of-the-art platform for comparative genomic, ecological, and chemodiversity studies in this ecologically diverse and biotechnologically and medically significant family of filamentous fungi.

Materials and Methods

Genome sequencing and assembly.

Mycelia were grown on potato dextrose agar for 72 hours before lyophilization. Lyophilized mycelia were lysed by grinding in liquid nitrogen and suspension in extraction buffer (100 mM Tris-HCl pH 8, 250 mM NaCl, 50 mM EDTA, and 1% SDS). Genomic DNA was isolated from the lysate with a phenol/chloroform extraction followed by an ethanol precipitation.

DNA was sequenced with both paired-end and mate-pair strategies to generate a high-quality genome assembly. Paired-end libraries and Mate-pair libraries were constructed at the Genomics Services Lab at HudsonAlpha (Huntsville, Alabama) and sequenced on an Illumina HiSeq X sequencer. Paired-end libraries were constructed with the Illumina TruSeq DNA kit, and mate-pair libraries were constructed with the Illumina Nextera Mate Pair Library kit targeting an insert size of 4 Kb. In total, 63 million paired-end reads and 105 million mate-pair reads, each of which was 150 bp in length, were generated.

The *A. spinulosporus* genome was assembled using the iWGS pipeline (Zhou et al., 2016). Paired-end and mate-pair reads were assembled with SPADES, version 3.6.2 (Bankevich et al., 2012), using optimal k-mer lengths chosen using KMERGENIE, version 1.6982 (Chikhi and Medvedev, 2014) and evaluated with QUAST, version 3.2 (Gurevich et al., 2013). The resulting assembly is 33.8 MB in size with an N50 of 939 Kb.

Data collection and quality assessment.

To collect a comprehensive set of genomes representative of Aspergillaceae, we used "Aspergillaceae" as a search term in NCBI's Taxonomy Browser and downloaded a representative genome from every species that had a sequenced genome as of February 5th 2018. We next confirmed that each species belonged to Aspergillaceae according to previous literature reports (Houbraken and Samson, 2011; de Vries et al., 2017). Altogether, 80 publicly available genomes and 1 newly sequenced genome spanning 5 genera (45 *Aspergillus* species; 33 *Penicillium* species; one *Xeromyces* species; one *Monascus* species; and one *Penicilliopsis* species) from the family Aspergillaceae were collected (File S1 from Steenwyk et al., 2019c). We also retrieved an additional 12 fungal genomes from representative species in the order Eurotiales but outside the family Aspergillaceae to use as outgroups.

To determine if the genomes contained gene sets of sufficient quality for use in phylogenomic analyses, we examined their gene set completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO), version 2.0.1 (Waterhouse et al., 2018a) (Figure S2 from Steenwyk et al., 2019c). In brief, BUSCO uses a consensus sequence built from hidden Markov models derived from 50 different fungal species using HMMER, version 3.1b2 (Eddy, 2011) as a query in

TBLASTN (Camacho et al., 2009; Madden, 2013) to search an individual genome for 3,156 predefined orthologs (referred to as BUSCO genes) from the Pezizomycotina database (creation date: 02-13-2016) available from ORTHODB, version 9 (Waterhouse et al., 2013). To determine the copy number and completeness of each BUSCO gene in a genome, gene structure is predicted using AUGUSTUS, version 2.5.5 (Stanke and Waack, 2003), with default parameters, from the nucleotide coordinates of putative genes identified using BLAST and then aligned to the HMM alignment of the same BUSCO gene. Genes are considered "single copy" if there is only one complete predicted gene present in the genome, "duplicated" if there are two or more complete predicted genes for one BUSCO gene, "fragmented" if the predicted gene is shorter than 95% of the aligned sequence lengths from the 50 different fungal species, and "missing" if there is no predicted gene.

Phylogenomic data matrix construction.

In addition to their utility as a measure of genome completeness, BUSCO genes have also proven to be useful markers for phylogenomic inference (Waterhouse et al., 2018a), and have been successfully used in phylogenomic studies of clades spanning the tree of life, such as insects (Ioannidis et al., 2017) and budding yeasts (Shen et al. 2016). To infer evolutionary relationships, we constructed nucleotide (NT) and amino acid (AA) versions of a data matrix comprised of the aligned and trimmed sequences of numerous BUSCO genes (Figure S3 from Steenwyk et al., 2019c). To construct this data matrix, we first used the BUSCO output summary files to identify orthologous single copy BUSCO genes with > 50% taxon-occupancy (i.e., greater than 47 / 93 taxa have the BUSCO gene present in their genome); 3,138 (99.4%) BUSCO genes met this criterion. For each BUSCO gene, we next created individual AA fasta files by

combining sequences across all taxa that have the BUSCO gene present. For each gene individually, we aligned the sequences in the AA fasta file using MAFFT, version 7.294b (Katoh and Standley, 2013), with the BLOSUM62 matrix of substitutions (Mount, 2008), a gap penalty of 1.0, 1,000 maximum iterations, and the "genafpair" parameter. To create a codon-based alignment, we used a custom PYTHON, version 3.5.2 (https://www.python.org/), script using BIOPYTHON, version 1.7 (Cock et al., 2009a), to thread codons onto the AA alignment. The NT and AA sequences were then individually trimmed using TRIMAL, version 1.4 (Capella-Gutierrez et al., 2009), with the "automated1" parameter. To remove potentially spuriously aligned sequences, we removed BUSCO genes whose sequence lengths were less than 50% of the untrimmed length in either the NT or AA sequences resulting in 1,773 (56.2%) BUSCO genes. Lastly, we removed BUSCO genes whose trimmed sequence lengths were too short (defined as genes whose alignment length was less than or equal to 167 AAs and 501 NTs), resulting in 1,668 (52.9%) BUSCO genes. The NT and AA alignments of these 1,668 BUSCO genes were then concatenated into the full 1,668-gene NT and AA versions of the phylogenomic data matrix.

To examine the stability of inferred relationships across all taxa, we constructed additional NT and AA data matrices by subsampling genes from the 1,668-gene data matrix that harbor signatures of strong phylogenetic signal. More specifically, we used 5 measures associated with strong phylogenetic signal (Shen et al. 2016) to create 5 additional data matrices (1 data matrix per measure) comprised of the top scoring 834 (50%) genes for NTs and AAs (Figure S4 from Steenwyk et al., 2019c). These five measures were: alignment length, average bootstrap value, taxon completeness, treeness / relative composition variability (RCV) (Phillips and Penny,

2003), and the number of variable sites. We calculated each measure with custom PYTHON scripts using BIOPYTHON. Treeness / RCV was calculated using the following formula:

$$\frac{Treeness}{RCV} = \frac{\sum_{u=1}^{b} l_u \Big/ l_t}{\sum_{i=1}^{c} \sum_{j=1}^{n} \frac{|c_{ij} - \overline{c_i}|}{s \bullet n}}$$

where l_u refers to the internal branch length of the *u*th branch (of *b* internal branches), l_t refers to total tree length, *c* is the number of different characters per sequence type (4 for nucleotides and 20 for amino acids), *n* is the number of taxa in the alignment, c_{ij} refers to the number of *i*th *c* characters for the *j*th taxon, $\overline{c_i}$ refers to the average number of the *i*th *c* character across *n* taxa, and *s* refers to the total number of sites in the alignment. Altogether, we constructed a total of 12 data matrices (one 1,668-gene NT data matrix, one 1,668-gene AA data matrix, five NT subsample data matrices, and five AA subsample data matrices).

Maximum likelihood phylogenetic analyses.

We implemented a maximum likelihood framework to infer evolutionary relationships among taxa for each of the 1,668 single genes and each of the 12 data matrices separately. For inferences made using either the 1,668- or 834-gene data matrices, we used three different analytical schemes: concatenation with gene-based partitioning, concatenation without partitioning, and gene-based coalescence (Felsenstein, 1981; Rokas et al., 2003; Edwards, 2009; Mirarab and Warnow, 2015). All phylogenetic trees were built using IQ-TREE, version 1.6.1 (Nguyen et al., 2015). In each case, we determined the best model for each single gene or partition using the "-m TEST" and "-mset raxml" parameters, which automatically estimate the best fitting model of substitutions according to their Bayesian Information Criterion values for

either NTs or AAs (Kalyaanamoorthy et al., 2017) for those models shared by RAxML (Stamatakis, 2014a) and IQ-TREE.

We first examined the inferred best fitting models across all single gene trees. Among NT genes, the best fitting model for 1,643 genes was a general time reversible model with unequal rates and unequal base frequencies with discrete gamma models, "GTR+G4" (Tavaré, 1986; Yang, 1994, 1996), and for the remaining 25 genes was a general time reversible model with invariable sites plus discrete gamma models, "GTR+I+G4" (Tavaré, 1986; Vinet and Zhedanov, 2011) (Figure S5a from Steenwyk et al., 2019c). Among AA genes, the best fitting model for 643 genes was the JTT model with invariable sites plus discrete gamma models, "JTT+I+G4" (Jones et al., 1992; Vinet and Zhedanov, 2011), for 362 genes was the LG model with invariable sites and discrete gamma models, "LG+I+G4" (Le and Gascuel, 2008; Vinet and Zhedanov, 2011), for 225 genes was the JTT model with invariable sites, empirical AA frequencies, and discrete gamma models "JTT+F+I+G4" (Jones et al., 1992; Vinet and Zhedanov, 2011), and for 153 genes was the JTTDCMut model with invariable sites and discrete gamma models, "JTTDCMut+I+G4" (Kosiol and Goldman, 2005; Vinet and Zhedanov, 2011) (Figure S5b from Steenwyk et al., 2019c). We used IQ-TREE for downstream analysis because a recent study using diverse empirical phylogenomic data matrices showed that it is a top-performing software (Zhou et al., 2018).

To reconstruct the phylogeny of Aspergillaceae using a partitioned scheme where each gene has its own model of sequence substitution and rate heterogeneity across sites parameters for any given data matrix, we created an additional input file describing these and gene boundary

parameters. More specifically, we created a nexus-format partition file that was used as input with the "-spp" parameter, which allows each gene partition in the data matrix to have its set of evolutionary rates (Chernomor et al., 2016). To increase the number of candidate trees used during maximum likelihood search, we changed the "-nbest" parameter from the default value of 5 to 10. Lastly, we conducted 5 independent searches for the maximum likelihood topology using 5 distinct seeds specified with the "-seed" parameter and chose the search with the best log-likelihood score. We used the phylogeny inferred using a partitioned scheme on the full NT data matrix as the reference one for all subsequent comparisons (Figure 1).

To infer the phylogeny of Aspergillaceae using a non-partitioned scheme, we used a single model of sequence substitution and rate heterogeneity across sites for the entire matrix. To save computation time, the most appropriate single model was determined by counting which best fitting model was most commonly observed across single gene trees. The most commonly observed model was "GTR+F+I+G4" (Waddell and Steel, 1997; Vinet and Zhedanov, 2011), which was favored in 1,643 / 1,668 (98.5%) of single genes, and "JTT+I+G4" (Jones et al., 1992; Vinet and Zhedanov, 2011), which was favored in 643 / 1,668 (38.5%) of single genes, for NTs and AAs, respectively, (Figure S5 from Steenwyk et al., 2019c). In each analysis, the chosen model was specified using the "-m" parameter.

To reconstruct the phylogeny of Aspergillaceae using coalescence, a method that estimates species phylogeny from single gene trees under the multi-species coalescent (Edwards, 2009), we combined all NEWICK (Felsenstein, 1986, 1996) formatted single gene trees inferred using

their best fitting models into a single file. The resulting file was used as input to ASTRAL-II, version 4.10.12 (Mirarab and Warnow, 2015) with default parameters.

To evaluate support for single gene trees and for the reference phylogeny (Figure 1), we used the ultrafast bootstrap approximation approach (UFBoot) (Hoang et al., 2018), an accurate and faster alternative to the classic bootstrap approach. To implement UFBoot for the NT 1,668-gene data matrix and single gene trees, we used the "-bb" option in IQ-TREE with 5,000 and 2,000 ultrafast bootstrap replicates, respectively.

Evaluating topological support.

To identify and quantify incongruence, we used two approaches. In the first approach, we compared the 36 topologies inferred from the full 1,668-gene NT and AA data matrices and five additional 834-gene data matrices (constructed by selecting the genes that have the highest scores in five measures previously shown to be associated with strong phylogenetic signal; see above) using three different maximum likelihood schemes (i.e., gene partitioned, non-partitioned, coalescence) and identified all incongruent bipartitions between the reference phylogeny (Figure 1) and the other 35. In the second approach, we scrutinized each bipartition in the reference phylogeny using measures of internode certainty (IC) measures for complete and partial single gene trees (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016). To better understand single gene support among conflicting bipartitions, we calculated gene-wise log-likelihood scores (GLS) (Shen et al., 2017) and gene support frequencies (GSF) for the reference and alternative topologies at conflicting bipartitions.

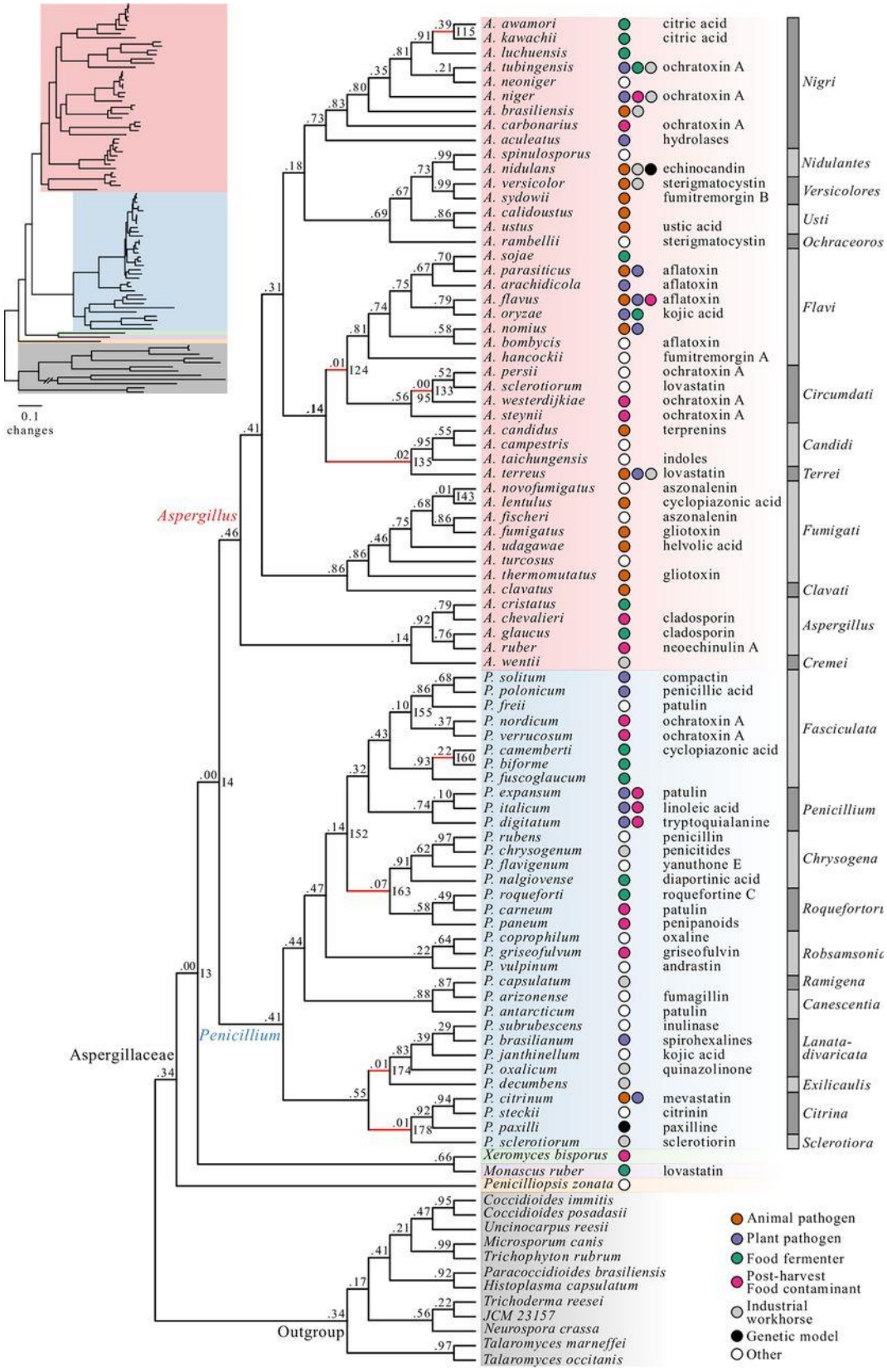
0.1 changes
Aspergillus
Penicillium
Aspergillaceae
Outgroup
A. awamori
A. kawachii
A. luchuensis
A. tubingensis
A. neoniger
A. niger
A. brasiliensis
A. carbonarius
A. aculeatus
A. spinulosporus
A. nidulans
A. versicolor
A. sydowii
A. calidoustus
A. ustus
A. rambellii
A. sojae
A. parasiticus
A. arachidicola
A. flavus
A. oryzae
A. nomius
A. bombycis
A. hancockii
A. persii
A. sclerotiorum
A. westerdijkiae
A. steynii
A. candidus
A. campestris
A. taichungensis
A. terreus
A. novofumigatus
A. lentulus
A. fischeri
A. fumigatus
A. udagawae
A. turcosus
A. thermomutatus
A. clavatus
A. cristatus
A. chevalieri
A. glaucus
A. ruber
A. wentii
P. solitum
P. polonicum
P. freii
P. nordicum
P. verrucosum
P. camemberti
P. biforme
P. fuscoglaucum
P. expansum
P. italicum
P. digitatum
P. rubens
P. chrysogenum
P. flavigenum
P. nalgiovense
P. roqueforti
P. carneum
P. paneum
P. coprophilum
P. griseofulvum
P. vulpinum
P. capsulatum
P. arizonense
P. antarcticum
P. subrubescens
P. brasilianum
P. janthinellum
P. oxalicum
P. decumbens
P. citrinum
P. steckii
P. paxilli
P. sclerotiorum
Xeromyces bisporus
Monascus ruber
Penicilliopsis zonata
Coccidioides immitis
Coccidioides posadasii
Uncinocarpus reesii
Microsporum canis
Trichophyton rubrum
Paracoccidioides brasiliensis
Histoplasma capsulatum
Trichoderma reesei
JCM 23157
Neurospora crassa
Talaromyces marneffei
Talaromyces occitanis
citric acid
citric acid
ochratoxin A
ochratoxin A
ochratoxin A
hydrolases
echinocandin
sterigmatocystin
fumitremorgin B
ustic acid
sterigmatocystin
aflatoxin
aflatoxin
aflatoxin
kojic acid
aflatoxin
fumitremorgin A
ochratoxin A
lovastatin
ochratoxin A
ochratoxin A
terprenins
indoles
lovastatin
aszonalenin
cyclopiazonic acid
aszonalenin
gliotoxin
helvolic acid
gliotoxin
cladosporin
cladosporin
neoechinulin A
compactin
penicillic acid
patulin
ochratoxin A
ochratoxin A
cyclopiazonic acid
patulin
linoleic acid
tryptoquialanine
penicillin
penicitides
yanuthone E
diaportinic acid
roquefortine C
patulin
penipanoids
oxaline
griseofulvin
andrastin
fumagillin
patulin
inulinase
spirohexalines
kojic acid
quinazolinone
mevastatin
citrinin
paxilline
sclerotiorin
lovastatin
Nigri
Nidulantes
Versicolores
Usti
Ochraceoros
Flavi
Circumdati
Candidi
Terrei
Fumigati
Clavati
Aspergillus
Cremei
Fasciculata
Penicillium
Chrysogena
Roquefortori
Robsamsonic
Ramigena
Canescentia
Lanata-divaricata
Exilicaulis
Citrina
Sclerotiora
Animal pathogen
Plant pathogen
Food fermenter
Post-harvest Food contaminant
Industrial workhorse
Genetic model
Other

Figure 1. A robust genome-scale phylogeny for the fungal family Aspergillaceae. Different genera are depicted using different-colored boxes: *Aspergillus* is shown in red, *Penicillium* in blue, *Xeromyces* in green, *Monascus* in purple, and *Penicilliopsis* in orange. Different sections within *Aspergillus* and *Penicillium* are depicted with alternating dark gray and gray bars. Internode certainty values are shown below each internode, and bootstrap values are shown above each internode (only bootstrap values lower than 100% are shown). Internode certainty values were calculated using the 1,668 maximum likelihood single-gene trees. Five thousand ultrafast bootstrap replicates were used to determine internode support. Internodes were considered unresolved if they were not present in one or more of the other 35 phylogenies represented in Figure 2—the branches of these unresolved internodes are drawn in red. Additional incongruent internodes were identified using calculations of IC. The inset depicts the phylogeny with branch lengths corresponding to estimated nucleotide substitutions per site. Colored circles next to species names indicate the lifestyle or utility of the species (i.e., animal pathogen, dark orange; plant pathogen, purple; food fermenter, green; postharvest food contaminant, pink; industrial workhorse, gray; genetic model, black; other, white). Exemplary secondary metabolites produced by different *Aspergillaceae* species are written to the right of the colored circles.

Identifying internodes with conflict across subsampled data matrices.

To identify incongruent bipartitions between the reference phylogeny and the other 35 phylogenies, we first included the 36 generated phylogenetic trees into a single file. We next evaluated the support of all bipartitions in the reference topology among the other 35 phylogenies using the "-z" option in RAxML. Any bipartition in the reference phylogeny that was not present in the rest was considered incongruent; each conflicting bipartition was identified through manual examination of the conflicting phylogenies. To determine if sequence type, subsampling method, or maximum likelihood scheme was contributing to differences in observed topologies among conflicting internodes, we conducted multiple correspondence analysis of these features among the 36 phylogenies and visualized results using the R, version 3.3.2 (R Development Core Team, 2008), packages FACTOMINER, version 1.40 (Lê et al., 2008) and FACTOEXTRA, version 1.0.5 (Kassambara and Mundt, 2017).

Identifying internodes with conflict across the 1,668 gene trees.

To examine the presence and degree of support of conflicting bipartitions, we calculated the internode certainty (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016; Zhou et al., 2017) of all internodes in the reference phylogeny (Figure 1) using the 1,668 gene trees as input. In general, IC scores near 0 indicate that there is near-equal support for an alternative, conflicting bipartition among a set of trees compared to a given bipartition present in the reference topology, which is indicative of high conflict. Therefore, we investigated incongruence in all internodes in the reference phylogeny (Figure 1) that exhibited IC scores lower than 0.1. To calculate IC values for each bipartition for the reference phylogeny, we created a file with all 1,668 complete and partial single gene trees. The resulting file of gene trees, specified with the "-z" parameter in RAxML, were used to calculate IC values using the "-f i" argument. The topology was specified with the "-t" parameter. Lastly, we used the Lossless corrected IC scoring scheme, which corrects for variation in taxon number across single gene trees (Kobert et al., 2016). We also used these IC values to inform which data type (NT or AA) provided the strongest signal for the given set of taxa and sequences. We observed that NTs consistently exhibited higher IC scores than AAs (hence our decision to use the topology inferred from the full NT data matrix using a gene-partitioned scheme – shown in Figure 1 – as the "reference" topology in all downstream analyses).

Examining gene-wise log-likelihood scores for incongruent internodes.

To determine the per gene distribution of phylogenetic signal supporting a bipartition in the reference phylogeny or a conflicting bipartition, we calculated gene-wise log-likelihood scores (GLS) (Shen et al., 2017) using the NT data matrix. We chose to calculate GLS using the NT

data matrix because distributions of IC values from phylogenies inferred using NTs had consistently higher IC values across schemes and data matrices (Figure S6 from Steenwyk et al., 2019c). To do so, we used functions available in IQ-TREE. More specifically, we inputted a phylogeny with the reference or alternative topology using the "-te" parameter and informed IQ-TREE of gene boundaries, their corresponding models, and optimal rate heterogeneity parameters in the full 1,668-gene data matrix using the "-spp" parameter. Lastly, we specified that partition log-likelihoods be outputted using the "-wpl" parameter. To determine if a gene provided greater support for the reference or alternative bipartition, we calculated the difference in GLS (ΔGLS) using the following formula:

$$\Delta\mathrm{GLS}_i = \ln L(G_i)_{ref} - \ln L(G_i)_{alt}$$

where $ln\ L(G_i)_{ref}$ and $ln\ L(G_i)_{alt}$ represent the log-likelihood values for the reference and alternative topologies for gene G_i. Thus, values greater than 0 reflect genes in favor of the reference bipartition, values lower than 0 reflect genes in favor of the alternative bipartition, and values of 0 reflect equal support between the reference and alternative bipartitions.

Calculating gene support frequencies for reference and conflicting bipartitions.

We next examined support for bipartitions in the reference topology as well as for their most prevalent conflicting bipartitions by calculating their gene support frequencies (GSF). GSF refers to the fraction of single gene trees that recover a particular bipartition. Currently, RAxML can only calculate GSF for trees with full taxon representation. Since our dataset contained partial gene trees, we conducted customs tests for determining GSF. To calculate GSF for NT (GSF_{NT}) and AA (GSF_{AA}) single gene trees, we extracted subtrees for the taxa of interest in individual

single gene trees and counted the occurrence of various topologies. For example, consider there are three taxa represented as A, B, and C, the reference rooted topology is "((A,B),C);" and the alternative rooted topology is "((A,C),B);". We counted how many single gene trees supported "(A,B)," or "(A, C),". For reference and alternative topologies involving more than three taxa or sections, we conducted similar tests. For example, if the reference rooted topology is "(((A,B),C),D);" and the alternative rooted topology is "((A,B),(C,D));", we counted how many single gene phylogenies supported "((A,B),C)," as sister to D and how many single gene phylogenies supported "(A,B)," and "(C,D)," as pairs of sister clades. For conflicting bipartitions at shallow depths in the phylogeny (i.e., among closely related species), we required all taxa to be present in a single gene tree; for conflicting bipartitions near the base of the phylogeny (i.e., typically involving multiple sections), we required at least one species to be present from each section of interest. Scripts to determine GSF were written using functions provided in NEWICK UTILITIES, version 1.6 (Junier and Zdobnov, 2010).

Filtering potential hidden paralogs.

Potential hidden paralogs among individual groups of orthologous genes can be identified by examining their ability to recover well established monophyletic clades (Rodríguez-Ezpeleta et al., 2007; Philippe et al., 2009; Salichos and Rokas, 2013). To filter genes containing potential hidden paralogs among the 1,668 NT orthologs, we removed single genes that did not recover six well established clades among *Aspergillus* and *Penicillium* species (Kocsubé et al., 2016; Yang et al., 2016; Nielsen et al., 2017; Kjærbølling et al., 2018). More specifically, we examined the 1,668 NT gene trees for monophyly of three *Aspergillus* clades (1: *Nigri*, 2: *Fumigati* and *Clavati*, and 3: *Aspergillus*) and three *Penicillium* clades (1: *Lanata-divaricata*, 2: *Chrysogena*,

and 3: *Citrina*). We identified 337 NT gene trees that did not recover these six clades. Removal of these 337 NT genes resulted in data matrix containing 1,331 NT genes. Using these 1,331 genes, we recalculated IC across the phylogeny and GSF at poorly supported bipartitions.

Alternative trimming methods.

Alignment trimming methodologies can have a drastic effect on inferred phylogenies (Tan et al., 2015). To examine if our inferences were robust to different trimming methods, we also trimmed single gene alignments using an entropy-based approach implemented in BMGE, version 1.12 (Criscuolo and Gribaldo, 2010). We used two different maximum entropy thresholds of 0.5 and 0.7, which we hereafter refer to as $BMGE_{0.5}$ and $BMGE_{0.7}$, respectively. To examine the influence of this entropy-based alignment trimming approach, we used these additional datasets to re-infer single-gene phylogenies, species-level phylogenies, calculate IC values, and examine GSF at incongruent bipartitions using both the full 1,668-gene data matrix and the potential hidden paralog-filtered 1,331-gene data matrix.

Topology tests.

To test the previously reported hypotheses of a) the genus *Penicillium* being the sister group to *Aspergillus* section *Nidulantes* and b) monophyly of narrow *Aspergillus* (sections *Nigri*, *Ochraceorosei*, *Flavi*, *Circumdati*, *Candidi*, *Terrei*) (Taylor et al., 2016), we conducted a series of tree topology tests using the 1,668-gene nucleotide data matrix using IQ-TREE (Nguyen et al., 2015). More specifically, we used the "GTR+F+I+G4" model and conducted the Shimodaira-Hasegawa (Shimodaira and Hasegawa, 1999) and the approximately unbiased tests (Shimodaira, 2002) as specified with the "-au" parameter. These tests were conducted using 10,000

resamplings using the resampling estimated log-likelihood (RELL) method (Kishino et al., 1990) as specified by the "-zb" parameter. We tested each hypothesis separately by generating the maximum likelihood topology under the constraint that the hypothesis is correct (specified using the "-z" parameter) and comparing its likelihood score to the score of the unconstrained maximum likelihood topology.

Estimating divergence times.

To estimate the divergence times for the phylogeny of the Aspergillaceae, we analyzed our NT data matrix used the Bayesian method implemented in MCMCTREE from the PAML package, version 4.9d (Yang, 2007). To do so, we conducted four analyses: we (i) identified genes evolving in a "clock-like" manner from the full data matrix, (ii) estimated the substitution rate across these genes, (iii) estimated the gradient and Hessian (Dos Reis and Yang 2013) at the maximum likelihood estimates of branch lengths, and (iv) estimated divergence times by Markov chain Monte Carlo (MCMC) analysis.

(i) Identifying "clock-like" genes.

Currently, large phylogenomic data matrices that contain hundreds to thousands of genes and many dozens of taxa are intractable for Bayesian inference of divergence times; thus, we identified and used only those genes that appear to have evolved in a "clock-like" manner in the inference of divergence times. To identify genes evolving in a "clock-like" manner, we calculated the degree of violation of a molecular clock (DVMC) (Liu et al., 2017) for single gene trees. DVMC is the standard deviation of root to tip distances in a phylogeny and is calculated using the following formula:

$$\mathrm{DVMC} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2}$$

where t_i represents the distance between the root and species *i* across *n* species. Using this method, genes with low DVMC values evolve in a "clock-like" manner compared to those with higher values. We took the top scoring 834 (50%) genes to estimate divergence times.

(ii) Estimating substitution rate.

To estimate the substitution rate across the 834 genes, we used BASEML from the PAML package, version 4.9d (Yang, 2007). We estimated substitution rate using a "GTR+G" model of substitutions (model = 7) and a strict clock model (clock = 1). Additionally, we point calibrated the root of the tree to 96 million years ago (mya) according to TIMETREE (Hedges et al., 2006), which is based on several previous estimates (Berbee and Taylor 2001: 50.0 mya; Vijaykrishna et al. 2006: 96.1 mya; Sharpton et al. 2009: 146.1 mya). We estimated a substitution rate of 0.04 substitutions per 10 million years.

(iii) Estimation of the gradient and Hessian.

To save computing time, the likelihood of the alignment was approximated using a gradient and Hessian matrix. The gradient and Hessian refer to the first and second derivatives of the log-likelihood function at the maximum likelihood estimates of branch lengths (Dos Reis and Yang 2013), and collectively describe the curvature of the log-likelihood surface. Estimating gradient and Hessian requires an input tree with specified time constraints. For time constraints, we used

the *Aspergillus flavus – Aspergillus oryzae* split (3.68-3.99 mya: Sharpton et al. 2009; Da Lage et al. 2013), the *Aspergillus fumigatus – Aspergillus clavatus* split (35-59 mya: Sharpton et al. 2009; Da Lage et al. 2013), the origin of the genus *Aspergillus* (43-85 mya: Kensche et al. 2008; Sharpton et al. 2009; Beimforde et al. 2014; Fan et al. 2015; Gaya et al. 2015), and the origin of Aspergillaceae (50-146 mya: Berbee and Taylor 2001; Vijaykrishna et al. 2006; Sharpton et al. 2009) as obtained from TIMETREE (Hedges et al., 2006).

(iv) Estimating divergence times using MCMC analysis.

To estimate divergence times using a relaxed molecular clock (clock = 2), we used the resulting gradient and Hessian results from the previous step for use in MCMC analysis using MCMCTREE (Yang, 2007) and the topology inferred using the gene partitioned approach and the 834-gene NT matrix from the top scoring DVMC genes. To do so, a gamma distribution prior shape and scale must be specified. The gamma distribution shape and scale is determined from the substitution rate determined in step ii where shape is $a=(s/s)^2$ and scale is $b=s/s^2$ and s is the substitution rate. Therefore, $a=1$ and $b=25$ and the "rgene_gamma" parameter was set to "1 25." We also set the "sigma2_gamma" parameter to "1 4.5." To minimize the effect of initial values on the posterior inference, we discarded the first 100,000 results. Thereafter, we sampled every 500 iterations until 10,000 samples were gathered. Altogether, we ran 5.1 million iterations (100,000 + 500 x 10,000), which is 510 times greater than the recommended minimum for MCMC analysis (Raftery and Lewis, 1995). Lastly, we set the "finetune" parameter to 1.

Statistical analysis and figure making.

All statistical analyses were conducted in R, version 3.3.2 (R Development Core Team, 2008). Spearman rank correlation analyses (Sedgwick, 2014) were conducted using the "rcorr" function in the package HMISC, version 4.1-1 (Harrell Jr, 2015). Stacked barplots, barplots, histograms, scatterplots, and boxplots were made using GGPLOT2, version 2.2.1 (Wickham, 2009). Intersection plots (also known as UpSet plots), were made using UPSETR, version 1.3.3 (Conway et al., 2017). The topological similarity heatmap and hierarchical clustering were done using PHEATMAP, version 1.0.8 (Kolde, 2012). Phylogenetic trees were visualized using FIGTREE, version 1.4.3 (Rambaut, 2009). The phylogenetic tree with the geological time scale was visualized using STRAP, version 1.4 (Bell and Lloyd, 2015). Artistic features of figures (e.g., font size, font style, etc.) were minimally edited using the graphic design software Affinity Designer (https://affinity.serif.com/en-us/).

Data availability

All data matrices, species-level and single-gene phylogenies are available through the figshare repository https://figshare.com/s/3098a7f59afa071a5c28 (doi: 10.6084/m9.figshare.6465011). The provided link is a private link for review purposes only. The genome sequence and raw reads of *Aspergillus spinulosporus* have been uploaded to GenBank as BioProject PRJNA481010.

Results

The examined genomes have nearly complete gene sets

Assessment of individual gene set completeness showed that most of the 93 genomes (81 in the ingroup and 12 in the outgroup) used in our study contain nearly complete gene sets and that all 93 genomes are appropriate for phylogenomic analyses. Specifically, the average percentage of BUSCO single-copy genes from the Pezizomycotina database (Waterhouse et al., 2013) present was 96.2 ± 2.6% (minimum: 81.1%; maximum: 98.9%; Figure S2 from Steenwyk et al., 2019c). Across the 93 genomes, only 3 (3.2%) genomes had < 90% of the BUSCO genes present in single-copy (*Penicillium carneum*: 88.6%; *Penicillium verrucosum*: 86.1%; and *Histoplasma capsulatum*: 81.1%).

The generated data matrices exhibit very high taxon occupancy

The NT and AA alignments of the 1,668-gene data matrix were comprised of 3,163,258 and 1,054,025 sites, respectively. The data matrix exhibited very high taxon occupancy (average gene taxon occupancy: 97.2 ± 0.1%; minimum: 52.7%; maximum: 100%; Figure S7a, b from Steenwyk et al., 2019c; File S2 from (Steenwyk et al., 2019c)). 417 genes had 100% taxon-occupancy, 1,176 genes had taxon-occupancy in the 90% to 99.9% range, and only 75 genes had taxon occupancy lower than 90%. Assessment of the 1,668 genes for five criteria associated with strong phylogenetic signal (gene-wise alignment length, average bootstrap value, completeness, treeness / RCV, and the number of variable sites) facilitated the construction of five subsampled matrices derived from 50% of the top scoring genes (Figure S7 from Steenwyk et al., 2019c; File S2 from Steenwyk et al., 2019c).

Examination of the gene content differences between the 5 NT subsampled data matrices as well as between the 5 AA data matrices revealed that they are comprised of variable sets of genes (Figure S8 from Steenwyk et al., 2019c). For example, the largest intersection among NT data matrices comprised of 207 genes that were shared between all NT matrices except the completeness-based one; similarly, the largest intersection among AA data matrices was 228 genes and was shared between all AA matrices except the completeness-based one (Figure S8a, b from Steenwyk et al., 2019c). Examination of the number of gene overlap between the NT and AA data matrices for each criterion (Figure S8c from Steenwyk et al., 2019c) showed that three criteria yielded identical or nearly identical NT and AA gene sets. These were completeness (834 / 834; 100% shared genes; $r_s = 1.00$, $p < 0.01$; Figure S7c from Steenwyk et al., 2019c), alignment length (829 / 834; 99.4% shared genes; $r_s = 1.00$, $p < 0.01$; Figure S7f from Steenwyk et al., 2019c), and the number of variable sites (798 / 834; 95.7% shared genes; $r_s = 0.99$, $p < 0.01$; Figure S7i from Steenwyk et al., 2019c). The other two criteria showed greater differences between NT and AA data matrices (average bootstrap value: 667 / 834; 80.0% shared genes; $r_s = 0.78$, $p < 0.01$; Figure S7l from Steenwyk et al., 2019c; treeness / RCV: 644 / 834; 77.2% shared genes; $r_s = 0.72$, $p < 0.01$; Figure S7o from Steenwyk et al., 2019c).

A genome-scale phylogeny for the family Aspergillaceae

NT and AA phylogenomic analyses of the full data matrix and the five subsampled data matrices under three analytical schemes recovered a broadly consistent set of relationships (Figure 1, 2, 3, 4). Across all 36 species-level phylogenies, we observed high levels of topological similarity (average topological similarity: 97.2 ± 2.5%; minimum: 92.2%; maximum: 100%) (Figure 2), with both major genera (*Aspergillus* and *Penicillium*) as well as all sections in *Aspergillus* and

Penicillium (Houbraken and Samson, 2011; Kocsubé et al., 2016) recovered as monophyletic (Figures 1, 3, and 4). Additionally, all but one internodes exhibited absolute UFBoot scores (Hoang et al., 2018); the sole exception was internode 33 (I33), which received 95 UFBoot support (Figure 1 and S9 from Steenwyk et al., 2019c).

Surprisingly, one taxon previously reported to be part of Aspergillaceae, *Basipetospora chlamydospora*, was consistently placed among outgroup species (Figure 1) and may represent a misidentified isolate. To identify the isolate's true identity, we blasted the nucleotide sequence of *tef1* from the isolate against the "nucleotide collection (nr/nt)" database using MEGABlast (Morgulis et al., 2008) on NCBI's webserver. We found the top three hits were to *Podospora anserina* (Class Sordariomycetes, PODANS_1_19720; e-value: 0.0, max score: 1753, percent identity: 91%), *Scedosporium apiospermum* (Class Sordariomycetes, SAPIO_CDS5137; e-value: 0.0, max score: 1742, percent identity: 92%), and *Isaria fumosorosea* (Class Sordariomycetes, ISF_05984; e-value: 0.0, max score: 1724, percent identity: 90%). These results make it difficult to ascribe the genome of the misidentified isolate to a specific genus and species but confirm its placement outside of Aspergillaceae; we refer to the isolate by its strain identifier, JCM 23157.

Examination of the Aspergillaceae phylogeny reveals 14 incongruent bipartitions

Examination of all 36 species-level phylogenies revealed the existence of 8 (8 / 78; 10.3%) incongruent bipartitions. Complementary examination of IC, a bipartition-based measure of incongruence, revealed an additional 3 / 78 (3.8%) bipartitions that displayed very high levels of incongruence at the gene level. Examination of the stability of robustly inferred bipartitions to

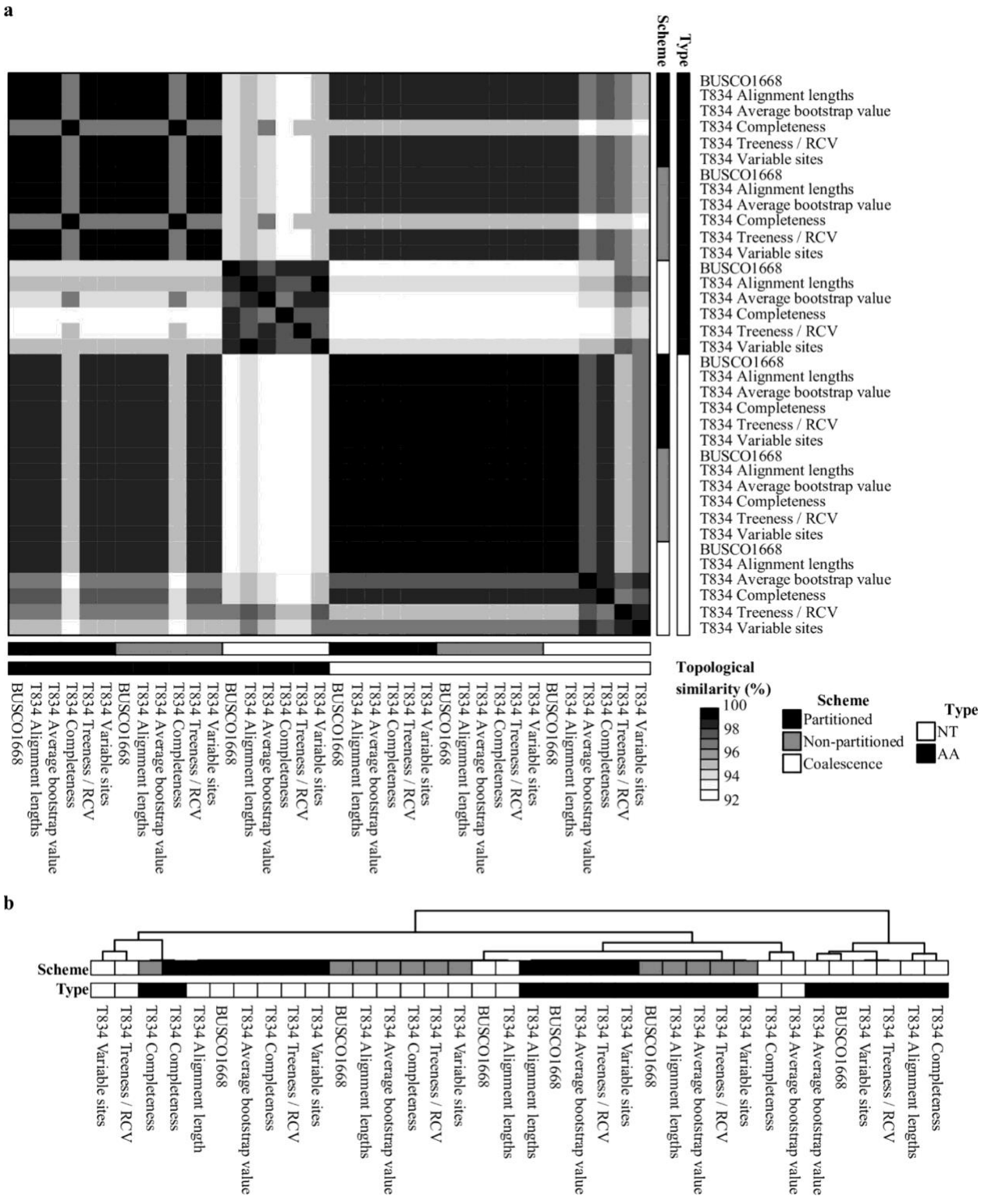
a
Scheme
Type
BUSCO1668
T834 Alignment lengths
T834 Average bootstrap value
T834 Completeness
T834 Treeness / RCV
T834 Variable sites
Topological similarity (%)
100
98
96
94
92
Scheme
Partitioned
Non-partitioned
Coalescence
Type
NT
AA
b
Scheme
Type
T834 Variable sites
T834 Treeness / RCV
T834 Completeness
T834 Alignment length
BUSCO1668
T834 Average bootstrap value
T834 Alignment lengths

Figure 2. Topological similarity between the 36 phylogenies constructed using 6 different data matrices, 2 different sequence types, and 3 analytical schemes.
(a) A heat map depiction of topological similarity between the 36 phylogenies constructed in this study. The 36 phylogenies were inferred from analyses of 2 different sequence types (i.e., protein, depicted in black; nucleotide, depicted in white), 3 different analytical schemes (i.e., partitioned, depicted in black; nonpartitioned, depicted in gray; coalescence, depicted in white), and 6 different matrices (full data matrix, "BUSCO1668," and 5 subsampled ones, all starting with "T834"; depending on the subsampling strategy, they are identified as "T834 Alignment lengths," "T834 Average bootstrap value," "T834 Completeness," "T834 Treeness/RCV," and "T834 Variable sites"). (b) Hierarchical clustering based on topological similarity values among the 36 phylogenies.

---

alternative alignment trimming approaches revealed an additional 3 / 78 (3.8%) bipartitions with high levels of incongruence at the gene level raising the total number of incongruent bipartitions to 14 (14 / 78; 17.9%).

Examination of the eight conflicting bipartitions stemming from the comparison of the 36 phylogenies showed that they were very often associated with data type (NT or AA) and scheme employed (concatenation or coalescence). For example, the first instance of incongruence concerns the identity of the sister species to *Penicillium biforme* (I60; Figure 1 and 3a); this species is *P. camemberti* in the reference phylogeny but analyses of the full and two subsampled AA data matrices with coalescence recover instead *Penicillium fuscoglaucum*. The data type and analytical scheme employed also appear to underlie the second and third instances of incongruence, which concern the placement of sections *Exilicaulis* and *Sclerotiora* (I74 and I78; Figures 1 and 3b), the fourth and fifth instances, which concern relationships among *Aspergillus* sections (I24 and I35; Figures 1 and 3c), as well as the sixth instance, which concerns relationships among the sections *Digitata*, *Chrysogena*, and *Roquefortorum* (I63; Figure 1 and 3d). The seventh instance is also associated with data type, but not with the scheme employed; while the reference as well as most subsampled NT matrices support the *Aspergillus persii* and

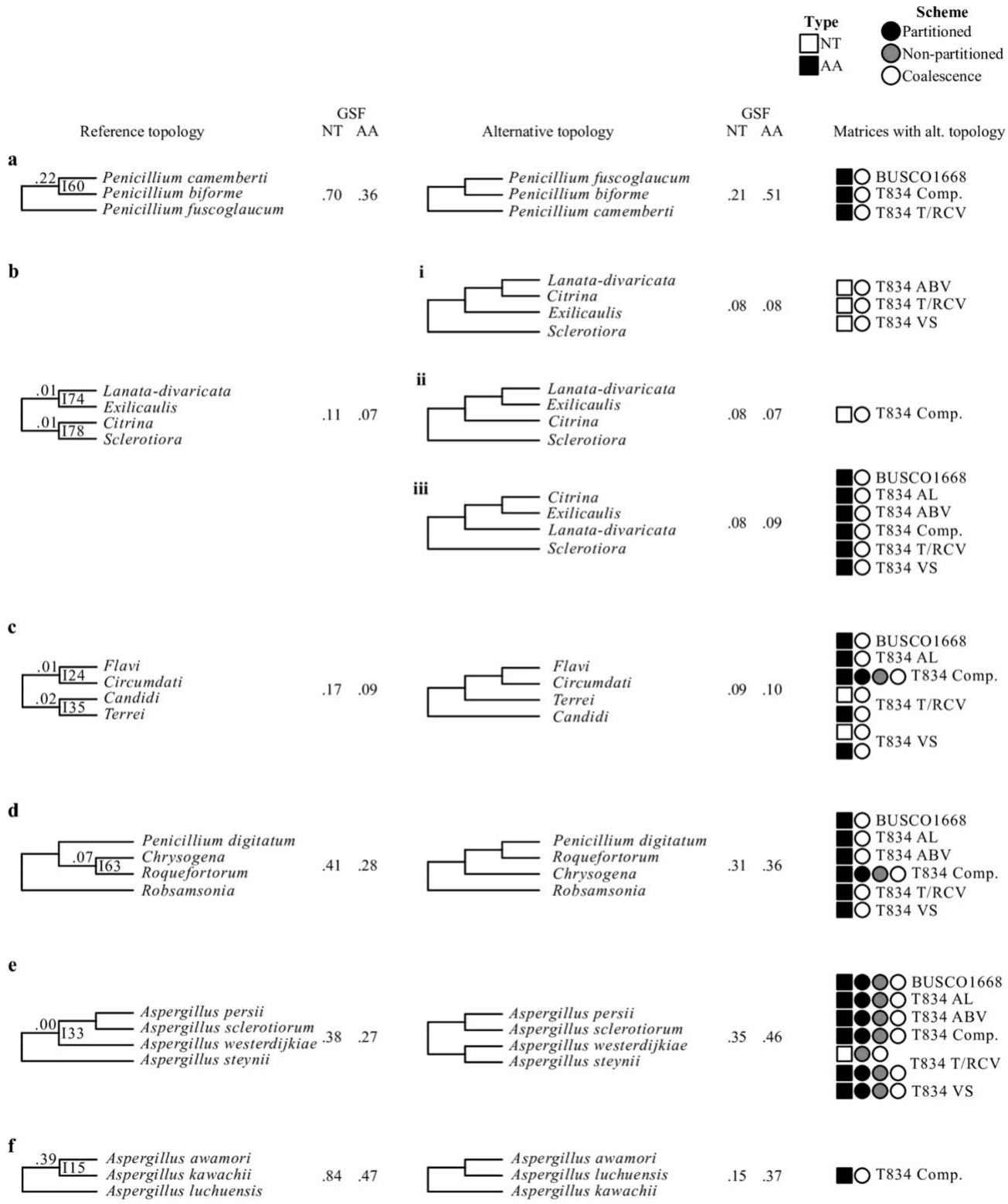
Type
NT
AA
Scheme
Partitioned
Non-partitioned
Coalescence
Reference topology
GSF NT AA
Alternative topology
GSF NT AA
Matrices with alt. topology
a
.22 I60 Penicillium camemberti
Penicillium biforme
Penicillium fuscoglaucum
.70 .36
Penicillium fuscoglaucum
Penicillium biforme
Penicillium camemberti
.21 .51
BUSCO1668
T834 Comp.
T834 T/RCV
b
i
Lanata-divaricata
Citrina
Exilicaulis
Sclerotiora
.08 .08
T834 ABV
T834 T/RCV
T834 VS
.01 I74 Lanata-divaricata
Exilicaulis
.01 I78 Citrina
Sclerotiora
.11 .07
ii
Lanata-divaricata
Exilicaulis
Citrina
Sclerotiora
.08 .07
T834 Comp.
iii
Citrina
Exilicaulis
Lanata-divaricata
Sclerotiora
.08 .09
BUSCO1668
T834 AL
T834 ABV
T834 Comp.
T834 T/RCV
T834 VS
c
.01 I24 Flavi
Circumdati
.02 I35 Candidi
Terrei
.17 .09
Flavi
Circumdati
Terrei
Candidi
.09 .10
BUSCO1668
T834 AL
T834 Comp.
T834 T/RCV
T834 VS
d
Penicillium digitatum
.07 I63 Chrysogena
Roquefortorum
Robsamsonia
.41 .28
Penicillium digitatum
Roquefortorum
Chrysogena
Robsamsonia
.31 .36
BUSCO1668
T834 AL
T834 ABV
T834 Comp.
T834 T/RCV
T834 VS
e
.00 I33 Aspergillus persii
Aspergillus sclerotiorum
Aspergillus westerdijkiae
Aspergillus steynii
.38 .27
Aspergillus persii
Aspergillus sclerotiorum
Aspergillus westerdijkiae
Aspergillus steynii
.35 .46
BUSCO1668
T834 AL
T834 ABV
T834 Comp.
T834 T/RCV
T834 VS
f
.39 I15 Aspergillus awamori
Aspergillus kawachii
Aspergillus luchuensis
.84 .47
Aspergillus awamori
Aspergillus luchuensis
Aspergillus kawachii
.15 .37
T834 Comp.

Figure 3. The eight internodes not recovered in all 36 phylogenies.
Internode numbers refer to internodes that have at least one conflicting topology among the 36 phylogenetic trees inferred from the full and five subsampled data matrices across three different schemes and two data types. The internode recovered from the analysis of the 1,668-gene nucleotide matrix (Fig. 1) is shown on the left and the conflicting internode(s) on the right. Next to each of the internodes, the nucleotide (NT) and amino acid (AA) gene support frequency (GSF) values are shown. On the far right, the sequence type, scheme, and data matrix characteristics of the phylogenies that support the conflicting internodes are shown. NT and AA sequence types are represented using white and black squares, respectively; partitioned concatenation, nonpartitioned concatenation, and coalescence analytical schemes are depicted as black, gray, or white circles, respectively; and the matrix subset is written next to the symbols.

Aspergillus sclerotiorum clade as sister to *Aspergillus westerdijkiae* (I33; Figure 1 and 3e), most AA data matrices recover a conflicting bipartition where *A. steynii* is the sister group of *A. westerdijkiae.* The final instance of incongruence was the least well supported, as 35 / 36 (97.2%) phylogenies supported *Aspergillus kawachii* as the sister group to *Aspergillus awamori* (I15, Figure 1 and 3f), but analysis of one AA subsampled data matrix with coalescence instead recovered *Aspergillus luchuensis* as the sister group.

For each of these bipartitions (Figure 3), we examined clustering patterns using multiple correspondence analysis of matrix features (i.e., sequence type and subsampling method) and analysis scheme among trees that support the reference and alternative topologies (Figure S10 from Steenwyk et al., 2019c). Distinct clustering patterns were observed for I74, I78, and I33 (Figure 3 and S10 from Steenwyk et al., 2019c). For I74 and I78, there are three alternative, conflicting topologies, with the first two clustering separately from the third (Figure 3b and S10b from Steenwyk et al., 2019c). For I33, phylogenies that support the reference and alternative topologies formed distinct clusters (Figure 3e). Examination of the contribution of variables along the second dimension, which is the one that differentiated variables that supported each

topology, revealed that the distinct clustering patterns were driven by sequence type (Figure S10g and h from Steenwyk et al., 2019c).

Examination of IC values revealed three additional bipartitions with strong signatures for incongruence at the gene level, defined as IC score lower than 0.10. The first instance concerns the sister taxon to the *Aspergillus* and *Penicillium* clade. Although all 36 phylogenies recover a clade comprised of *Xeromyces bisporus* and *Monascus ruber* as the sister group, the IC score for this bipartition is 0.00 (I3; Figure 4a); the most prevalent, conflicting bipartition supports *Penicilliopsis zonata* as sister to *Aspergillus* and *Penicillium* (Figure 4a). Similarly, although all 36 phylogenies recover *Penicillium* as sister to *Aspergillus*, the IC score for this bipartition is also 0.00 (I4; Figure 4b); the most prevalent, conflicting bipartition supports *X. bisporus* and *M.*

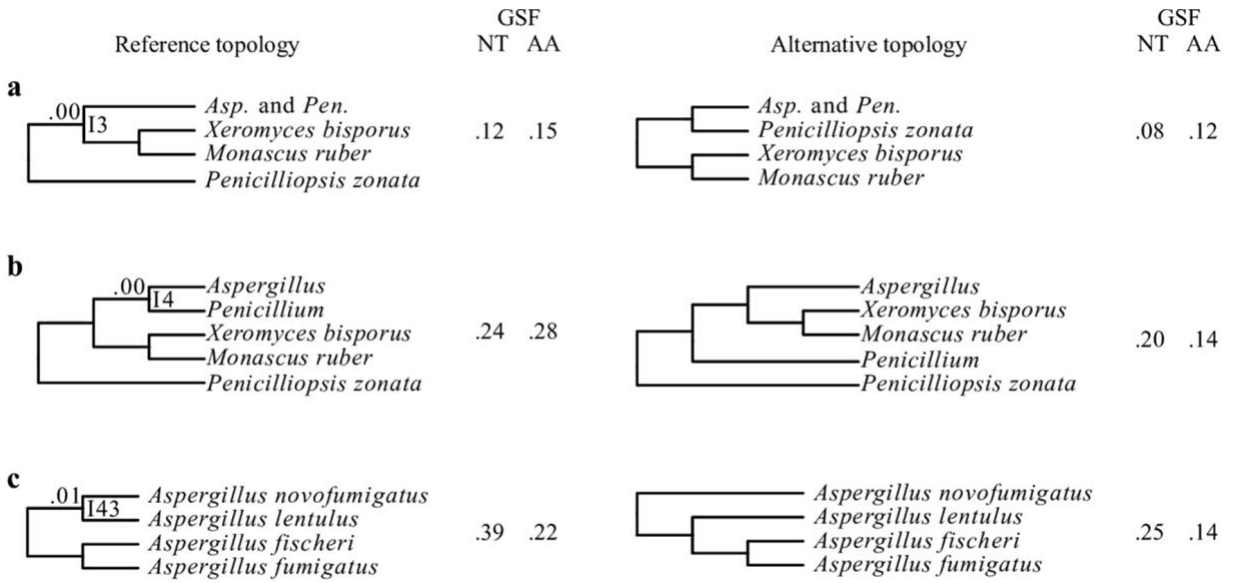

Figure 4. The three internodes recovered in all 36 phylogenies but that exhibit very low internode certainty values.
Three bipartitions were recovered by all 36 phylogenies but had internode certainty values below 0.10 (a to c). The internode recovered from the analysis of all 36 phylogenies, including of the 1,668-gene nucleotide matrix (Figure 1b), is shown on the left and the most prevalent,

conflicting internode on the right. Next to each of the internodes, the nucleotide (NT) and amino acid (AA) gene support frequency (GSF) values are shown.

ruber as the sister clade to *Aspergillus* (Figure 4b). In the third instance, all 36 phylogenies support *Aspergillus novofumigatus* and *Aspergillus lentulus* as sister species, but the IC score of this bipartition is 0.01 (I43; Figure 4c); the most prevalent, conflicting bipartition recovers *A. lentulus* as the sister species to a clade comprised of *Aspergillus fumigatus* and *Aspergillus fischeri* (Figure 4c).

To examine the underlying individual gene support to the resolution of these 11 bipartitions, we examined the phylogenetic signal contributed by each individual gene in the full NT data matrix. In all 11 bipartitions, we found that inferences were robust to single gene outliers with strong

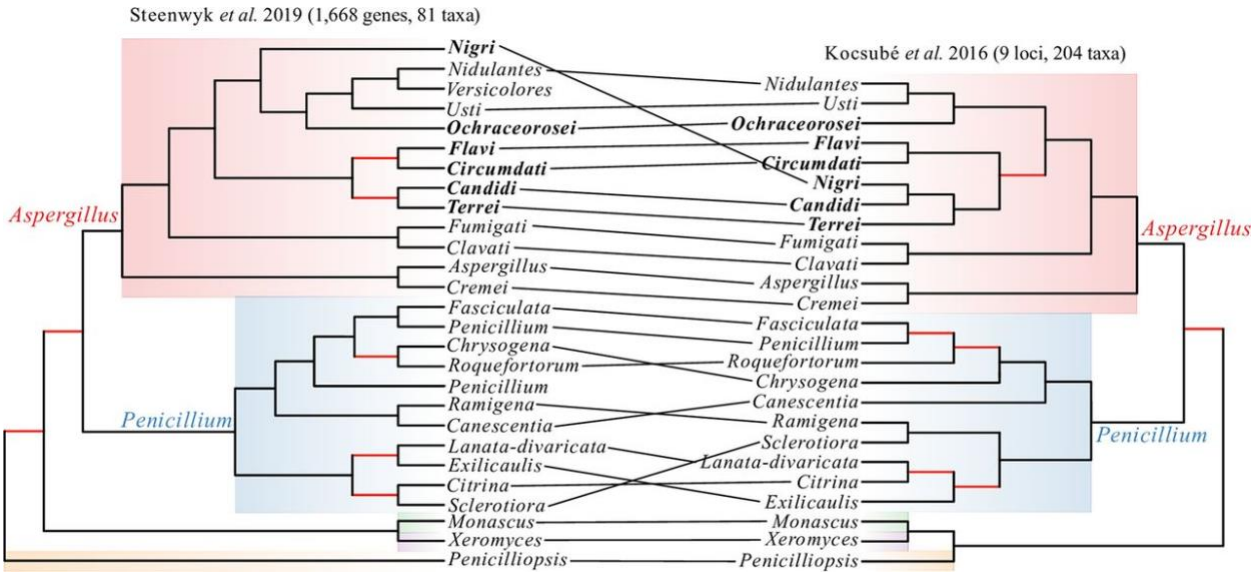

Figure 5. A visual comparison of the differences between the phylogeny reported in this study and the phylogeny reported in the work of Kocsubé et al..
Tanglegram between the section-level phylogeny presented in this study (left) and the section-level phylogeny presented by Kocsubé et al. (right). The key differences between the two phylogenies lie in the placements of sections *Nigri*, *Ramigena*, and *Canescentia*. Species in bold

belong to narrow *Aspergillus*, and red branches represent bipartitions that are not robustly supported in each study.

---

phylogenetic signal (Figure S11 from Steenwyk et al., 2019c; File S4 from Steenwyk et al., 2019c).

To determine if robustly identified internodes were sensitive to potential hidden paralogs, we reevaluated IC in a set of 1,331 genes that passed our hidden paralogy filter. We observed that measurements of IC were very similar between the 1,668 and 1,331 NT datasets (r_s = 0.98, p < 0.01; Figure S12 from Steenwyk et al., 2019c). Notably, we did not identify any additional internodes with evidence of incongruence. In contrast, examination of IC in the 1,331 gene tree set showed reduced levels of incongruence at I63 (Figure 3d; IC value using the 1,668-gene data matrix = 0.07, IC value using the 1,331-gene data matrix = 0.10).

To determine if our estimates of incongruence were robust to various trimming methods, we recalculated IC scores using gene trees whose alignments were trimmed with BMGE (Criscuolo and Gribaldo, 2010) using maximum entropy cut-off values of 0.5 or 0.7 (or $BMGE_{0.5}$ or $BMGE_{0.7}$, respectively). These analyses were done for both the set of 1,668 genes and the set of potential hidden paralogy-filtered 1,331 genes (File S7 from Steenwyk et al., 2019c). Altogether, we used four additional data sets ([two BMGE trimming parameters X two data sets of 1,668 and 1,331 genes) to further evaluate bipartition support. We observed that IC values were very similar between all data sets (r_s ranged from 0.97-0.99, p < 0.01 for all tests; Figure S13 from Steenwyk et al., 2019c). Furthermore, we noted that among incongruent internodes, GSF values were similar regardless of trimming method or whether or not potential hidden paralogs had been

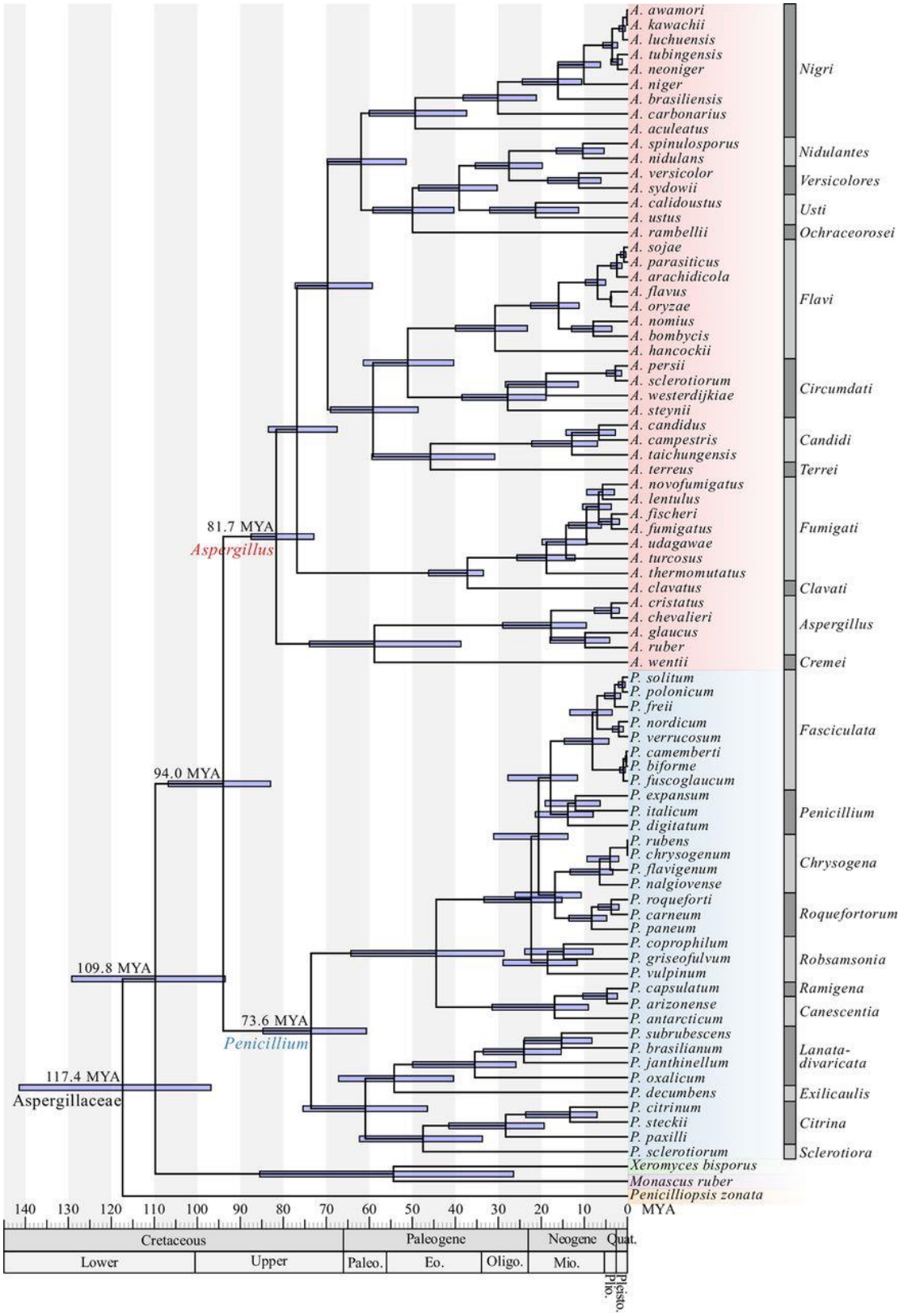
A. awamori
A. kawachii
A. luchuensis
A. tubingensis
A. neoniger
A. niger
A. brasiliensis
A. carbonarius
A. aculeatus
A. spinulosporus
A. nidulans
A. versicolor
A. sydowii
A. calidoustus
A. ustus
A. rambellii
A. sojae
A. parasiticus
A. arachidicola
A. flavus
A. oryzae
A. nomius
A. bombycis
A. hancockii
A. persii
A. sclerotiorum
A. westerdijkiae
A. steynii
A. candidus
A. campestris
A. taichungensis
A. terreus
A. novofumigatus
A. lentulus
A. fischeri
A. fumigatus
A. udagawae
A. turcosus
A. thermomutatus
A. clavatus
A. cristatus
A. chevalieri
A. glaucus
A. ruber
A. wentii
P. solitum
P. polonicum
P. freii
P. nordicum
P. verrucosum
P. camemberti
P. biforme
P. fuscoglaucum
P. expansum
P. italicum
P. digitatum
P. rubens
P. chrysogenum
P. flavigenum
P. nalgiovense
P. roqueforti
P. carneum
P. paneum
P. coprophilum
P. griseofulvum
P. vulpinum
P. capsulatum
P. arizonense
P. antarcticum
P. subrubescens
P. brasilianum
P. janthinellum
P. oxalicum
P. decumbens
P. citrinum
P. steckii
P. paxilli
P. sclerotiorum
Xeromyces bisporus
Monascus ruber
Penicilliopsis zonata
Nigri
Nidulantes
Versicolores
Usti
Ochraceorosei
Flavi
Circumdati
Candidi
Terrei
Fumigati
Clavati
Aspergillus
Cremei
Fasciculata
Penicillium
Chrysogena
Roquefortorum
Robsamsonia
Ramigena
Canescentia
Lanata-
divaricata
Exilicaulis
Citrina
Sclerotiora
81.7 MYA
Aspergillus
94.0 MYA
109.8 MYA
73.6 MYA
Penicillium
117.4 MYA
Aspergillaceae
140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 MYA
Cretaceous
Paleogene
Neogene
Quat.
Lower
Upper
Paleo.
Eo.
Oligo.
Mio.
Plio.
Pleisto.

Figure 6. A molecular time tree for the family *Aspergillaceae*.
Blue boxes around each internode correspond to 95% divergence time confidence intervals for each branch of the *Aspergillaceae* phylogeny. For reference, the geologic time scale is shown right below the phylogeny. Different genera are depicted using different-colored boxes; *Aspergillus* is shown in red, *Penicillium* in blue, *Xeromyces* in green, *Monascus* in purple, and *Penicilliopsis* in orange. Different sections within *Aspergillus* and *Penicillium* are depicted with alternating dark gray and gray bars. Dating estimates were calibrated using the following constraints: origin of *Aspergillaceae* (I2; 50 to 146 million years ago [mya]), origin of *Aspergillus* (I5; 43 to 85 mya), the *A. flavus* and *A. oryzae* split (I30; 3.68 to 3.99 mya), and the *A. fumigatus* and *A. clavatus* split (I38; 35 to 39 mya); all constraints were obtained from TimeTree.

---

removed (r_s ranged from 0.97-1.00, $p < 0.01$ for all tests; Figure S14 from Steenwyk et al., 2019c).

Through these recalculations of IC, we identified an additional three incongruent internodes that had values below 0.10 in one or more of these data sets. All three cases concerned relationships within the genus *Penicillium*, namely internodes I55, I52, and I62 (Figure 4d-f). I55 had IC values of 0.09 and 0.08 in the 1,668- and 1,331-gene data sets trimmed by $BMGE_{0.5}$, respectively; I52 had an IC value of 0.07 in the 1,668-gene data set trimmed by $BMGE_{0.5}$; I62 had IC values of 0.09 and 0.09 in the 1,668- and 1,331-gene data sets trimmed by $BMGE_{0.5}$.

To determine if removal of potential hidden paralogs and the use of different alignment trimming methods influenced inference of the species phylogeny, we re-inferred species trees using the three different maximum likelihood approaches across the five datasets resulting in 25 additional phylogenies ([two sequence types X two BMGE trimming approaches X three maximum likelihood schemes X two gene datasets of size 1,668 and 1,331] + 1,331-gene dataset trimmed using TRIMAL). Neither the removal of potential hidden paralogs nor the use of different trimming methods altered the topology of the species phylogeny in 21 of the 25 (84%) cases. In

the remaining four cases, the topologies recovered conflicted with the species phylogeny in Figure 1 with respect to an already identified conflict (Figures 3 and 4). Specifically, the species phylogeny inferred using coalescence with the 1,668-NT gene matrix trimmed using $BMGE_{0.7}$ inferred the topology discussed in Figure 3biii; the 1,668-AA gene matrix trimmed using $BMGE_{0.5}$ and $BMGE_{0.7}$ and the 1,331-NT gene matrix trimmed using $BMGE_{0.7}$ (all analyzed using coalescence) inferred the topology discussed in Figure 3f.

Incongruence in the Aspergillaceae phylogeny

Examination of the 14 incongruent bipartitions with respect to their placement on the phylogeny (shallow, i.e., near the tips of the phylogeny or deeper, i.e., away from the tips and toward the base of the phylogeny) and the amount of conflict (quantified using IC and GSF) allowed us to group them into three categories: (i) shallow bipartitions (I15 and I60) with low levels of incongruence, (ii) shallow bipartitions (I33, I43, I55, and I62) with high levels of incongruence, and (iii) deeper bipartitions (I3, I4, I24, I35, I52, I63, I74, and I78) with varying levels of incongruence and typically associated with single taxon long branches.

(i) Shallow bipartitions with low levels of incongruence.

The two bipartitions that fell into this category, I60 (Figure 3a) and I15 (Figure 3f), exhibited low levels of incongruence among closely related taxa. For I60, the reference bipartition was observed in 33 / 36 phylogenies, had an IC score of 0.22, and GSF_{NT} and GSF_{AA} scores of 0.70 and 0.21, respectively. Similarly, the reference bipartition for I15 was observed in 35 / 36 phylogenies, had an IC score of 0.39, and GSF_{NT} and GSF_{AA} scores of 0.84 and 0.47,

respectively. Notably, the GSF_{NT} scores were substantially higher for the reference bipartitions in both of these cases.

(ii) Shallow bipartitions with high levels of incongruence.

The four shallow bipartitions, I33 (Figure 3e), I43 (Figure 4c), I55 (Figure 4d), and I62 (Figure 4f), in this category exhibited high levels of incongruence among closely related taxa. For I33, the reference bipartition was observed in 16 / 36 (44.4%), had an IC score of 0.00, and GSF_{NT} and GSF_{AA} scores of 0.38 and 0.27, respectively. The reference bipartition for I43 was observed in all 36 phylogenies, had an IC score of 0.01 and GSF_{NT} and GSF_{AA} scores of 0.39 and 0.22, respectively. Similarly, the reference bipartition I55 was observed in all 36 phylogenies, had an IC score of 0.10 using the 1,668-gene data set trimmed using TRIMAL, but an IC score of 0.09 and 0.08 in the 1,668- and 1,331-gene data sets trimmed by $BMGE_{0.5}$, respectively. I55 had GSF_{NT} and GSF_{AA} scores of 0.51 and 0.31, respectively. Lastly, reference bipartition I62 was observed in all 36 phylogenies, had an IC score of 0.10 using the 1,668-gene data set trimmed using TRIMAL, but an IC score of 0.09 in both the 1,668- and 1,331-gene data sets trimmed by $BMGE_{0.5}$. I62 had GSF_{NT} and GSF_{AA} scores of 0.55 and 0.19, respectively. Notably, in all four cases, substantial fractions of genes supported both the reference and the conflicting bipartitions, with both the GSF_{NT} and GSF_{AA} scores of each pair of bipartitions being almost always higher than 0.2.

(iii) Deeper bipartitions often associated with single taxon long branches.

The seven bipartitions in this category were I74 and I78 (Figure 3b), I24 and I35 (Figure 3c), I63 (Figure 3d), I3 (Figure 4a), I4 (Figure 4b), and I52 (Figure 4e). All of them are located deeper in

the tree and most involve single taxa with long terminal branches (Figure 1). The reference bipartitions for internodes I74 and I78, which concern relationships among the sections *Lanata-divaricata*, *Exilicaulis*, *Citrina*, and *Sclerotiora* were observed in 26 / 36 (72.2%) phylogenies; the remaining 10 / 36 (27.8%) phylogenies recovered three alternative, conflicting bipartitions. Both reference bipartitions had IC scores of 0.01, and GSF_{NT} and GSF_{AA} scores of 0.11 and 0.07, respectively. The reference bipartitions for internodes I24 and I35, which concern the placement of *Aspergillus terreus*, the single taxon representative of section *Terrei*, were observed in 27 / 36 (75.0%) phylogenies, had IC scores of 0.01 and 0.02, and GSF_{NT} and GSF_{AA} scores of 0.17 and 0.09, respectively. The reference bipartition I63, which involved the placement of the *Penicillium digitatum*, the sole representative of section *Digitata*, was observed in 28 / 36 (77.8%), had an IC score of 0.07, and GSF_{NT} and GSF_{AA} scores of 0.41 and 0.28, respectively. Interestingly, the IC score for this bipartition in the hidden paralogy-filtered 1,331-gene data set increased to 0.10, suggesting that hidden paralogy may be a contributing factor to the observed incongruence at this internode. Finally, the reference bipartitions I3 and I4 (Figure 4), which concern the identity of the sister taxon of *Aspergillus* and *Penicillium* (I3) and the identity of the sister taxon of *Aspergillus* (I4), were found in all 36 phylogenies but both had IC values of 0.00. For I3, GSF_{NT} and GSF_{AA} scores were 0.12 and 0.15, respectively. For I4, GSF_{NT} and GSF_{AA} scores were 0.24 and 0.28, respectively. Lastly, the reference bipartition I52 was observed in all 36 phylogenies, had an IC score of 0.14 using the 1,668-gene data set trimmed using TRIMAL, but an IC score of 0.07 in the 1,668-gene data set trimmed by $BMGE_{0.5}$. For I52, GSF_{NT} and GSF_{AA} scores were 0.09 and 0.05, respectively. Notably, this is the only instance of incongruence not associated with a single taxon branch.

Topology tests

The phylogeny of the genera *Aspergillus* and *Penicillium* has been a topic of debate. Our topology supports the reciprocal monophyly of *Aspergillus* and *Penicillium* and rejects the monophyly of narrow *Aspergillus*. Both of these results are consistent with some previous studies (Kocsubé et al., 2016) (Figure 6) but in contrast to other previous studies, which recovered a topology where *Penicillium* is sister to section *Nidulantes* within *Aspergillus* and narrow *Aspergillus* (sections *Nigri*, *Ochraceorosei*, *Flavi*, *Circumdati*, *Candidi*, *Terrei*) was monophyletic (Pitt and Taylor, 2014; Taylor et al., 2016). To further evaluate both of these hypotheses, we conducted separate topology constraint analyses using the Shimodaira-Hasegawa (Shimodaira and Hasegawa, 1999) and the approximately unbiased tests (Shimodaira, 2002). Both tests rejected the constrained topologies (Table 1; p-value < 0.001 for all tests), providing further support that *Aspergillus* and *Penicillium* are reciprocally monophyletic and that narrow *Aspergillus* is not monophyletic (Figure 6).

Table 1. Topology tests reject the sister group relationship of genus *Penicillium* and *Aspergillus* section *Nidulantes* as well as the monophyly of narrow *Aspergillus.*

Constrained topology	Likelihood of unconstrained tree	Likelihood of constrained tree	Difference in log likelihood	Shimodaira-Hasegawa test *p* value	Approximately unbiased test *p* value
Sister group relationship of genus *Penicillium* and *Aspergillus* section *Nidulantes*	-99617175.719	-99767653.909	150478.190	<0.001	<0.001
Monophyly of narrow *Aspergillus*	-99617175.719	-99730789.937	113614.218	<0.001	<0.001

A geological timeline for the evolutionary diversification of the Aspergillaceae family

To estimate the evolutionary diversification among *Aspergillaceae*, we subsampled the 1,668-gene matrix for high-quality genes with "clock-like" rates of evolution by examining DVMC (Liu et al., 2017) values among single gene trees. Examination of the DVMC values facilitated the identification of a tractable set of high-quality genes for relaxed molecular clock analyses (Figure S15 from Steenwyk et al., 2019c). We found that *Aspergillaceae* originated 117.4 (95% CI: 141.5 - 96.9) mya during the Cretaceous period (Figure 5). We found that the common ancestor of *Aspergillus* and *Penicillium* split from the *X. bisporus* and *M. ruber* clade shortly thereafter, approximately 109.8 (95% CI: 129.3 - 93.5) mya. We also found that the genera *Aspergillus* and *Penicillium* split 94.0 (95% CI: 106.8 - 83.0) mya, with the last common ancestor of *Aspergillus* originating approximately 81.7 mya (95% CI: 87.5 - 72.9) and the last common ancestor of *Penicillium* originating approximately 73.6 mya (95% CI: 84.8 - 60.7).

Among *Aspergillus* sections, section *Nigri*, which includes the industrial workhorse *A. niger*, originated 49.4 (95% CI: 60.1 - 37.4) mya. Section *Flavi*, which includes the food fermenters *A. oryzae* and *A. sojae* and the toxin-producing, post-harvest food contaminant, and opportunistic animal and plant pathogen *A. flavus*, originated 30.8 (95% CI: 40.0 - 23.3) mya. Additionally, section *Fumigati*, which includes the opportunistic human pathogen *A. fumigatus*, originated 18.8 (95% CI: 25.7 - 12.2) mya. Among *Penicillium* sections, section *Fasciculata*, which contains Camembert and Brie cheese producer *P. camemberti* and the ochratoxin A producer, *P. verrucosum*, originated 8.1 (95% CI: 14.7 - 4.3) mya. Section *Chrysogena*, which includes the antibiotic penicillin producing species *P. chrysogenum*, originated 6.5 (95% CI: 13.3 - 3.4) mya.

Additionally, section *Citrina*, which contains *P. citrinum*, which the first statin was isolated from and is commonly associated with moldy citrus fruits (Endo et al. 1976), originated 28.3 (95% CI: 41.5 - 19.3) mya.

Finally, our analysis also provides estimates of the origins of various iconic pairs of species within *Aspergillus* and *Penicillium*. For example, among *Aspergillus* species pairs, we estimate that *A. fumigatus* and the closest relative with a sequenced genome, *A. fischeri* (Mead et al., 2018), diverged 3.7 (95% CI: 6.7 – 1.9) mya and *Aspergillus flavus* and the domesticated counterpart, *A. oryzae* (Gibbons et al., 2012), 3.8 (95% CI: 4.0 – 3.7) mya. Among *Penicillium* species pairs, we estimate *P. camemberti*, which contributes to cheese production to have diverged from its sister species and cheese contaminant *P. biforme* (Ropars et al., 2012) approximately 0.3 (95% CI: 0.5 – 0.1) mya. Finally, we estimate that *P. roqueforti*, another species that contributes to cheese production, diverged from its close relative *P. carneum* (Ropars et al., 2012) 3.8 (95% CI: 6.8 – 2.0) mya.

Discussion

Our analyses provide a robust evaluation of the evolutionary relationships and diversification among Aspergillaceae, a family of biotechnologically and medically significant fungi. We scrutinized our proposed reference phylogeny (Figure 1) against 35 other phylogenies recovered using all possible combinations of six multi-gene data matrices (full or subsamples thereof), three maximum likelihood schemes, and two sequence types and complemented this analysis with bi-partitioned based measures of support (Figures 1 and 2). We also examined the robustness of our proposed reference phylogeny to different sequence alignment trimming

methods and the removal of potential hidden paralogs. Through these analyses, we found that 14 / 78 (17.9%) bipartitions were incongruent (Figure 3 and 4) and explored the characteristics as well as sources of these instances of incongruence. Finally, we placed the evolution and diversification of Aspergillaceae in the context of geological time.

Comparison of our 81-taxon, 1,668-gene phylogeny to a previous one based on a maximum likelihood analysis of 9 loci for 204 Aspergillaceae species (Kocsubé et al., 2016), suggests that our analyses identified and strongly supported several new relationships and resolved previously poorly supported bipartitions (Figure 1, Figure 6). The robust resolution of our phylogeny is likely due to the very large size of our data matrix, both in terms of genes as well as in terms of taxa. For example, the placement of *Aspergillus* section *Nigri* has been unstable in previous phylogenomic analyses (Figure S1 from Steenwyk et al., 2019c) (Yang et al., 2016; de Vries et al., 2017; Kjærbølling et al., 2018), but our denser sampling of taxa in this section as well as inclusion of representative taxa from sections *Nidulantes*, *Versicolores*, *Usti*, and *Ochraceorosei* now provides strong support for the sister relationship of the *Aspergillus* section *Nigri* to sections *Nidulantes*, *Versicolores*, *Usti*, and *Ochraceorosei* (Figure 1).

However, our analysis also identified several relationships that exhibit high levels of incongruence (Figures 3 and 4). In general, gene tree incongruence can stem from biological or analytical factors (Rokas et al., 2003; Shen et al., 2017). Biological processes such as incomplete lineage-sorting (ILS) (Degnan and Salter, 2005), hybridization (Sang and Zhong, 2000), gene duplication and subsequent loss (Hallett et al., 2004), horizontal gene transfer (Doolittle and Bapteste, 2007) and natural selection (Castoe et al., 2009; Li et al., 2010b), can cause the

histories of genes to differ from one another and from the species phylogeny. Importantly, although the expected patterns of incongruence will be different for each factor and depend on a number of parameters, the observed patterns of conflict in each of the 14 cases of incongruence in the Aspergillaceae phylogeny can yield insights and allow the formation of hypotheses about the potential drivers in each case. For example, ILS often results in relatively low levels of incongruence; for instance, examination of the human, chimp, and gorilla genomes has showed that 20-25% of the gene histories differ from the species phylogeny (Patterson et al., 2006; Hobolth et al., 2007). In contrast, recent hybridization is expected to typically produce much higher levels of incongruence due to rampant sequence similarity among large amounts of genomic content; for instance, examination of *Heliconius* butterfly genomes revealed incongruence levels higher than 40% (Martin et al., 2013).

Additionally, analytical factors such as model choice (Phillips et al., 2004), taxon sampling (Rokas and Carroll, 2005; Nabhan and Sarkar, 2012), hidden paralogy (Rodríguez-Ezpeleta et al., 2007; Philippe et al., 2009), and alignment strategy (Tan et al., 2015) can lead to erroneous inference of gene histories. Perhaps the most well-known instance of incongruence stemming from analytical factors is what is known as “long branch attraction”, namely the situation where highly divergent taxa, i.e., the ones with the longest branches in the phylogeny, will often artifactually group with other long branches (Gribaldo and Philippe, 2002). Examination of the effects of removal of potential hidden paralogs and different alignment trimming strategies showed that these analytical factors did not substantially contribute to the observed incongruence (Figure S12-S14 from Steenwyk et al., 2019c). Using an aggressive trimming strategy, we did

identify three additional instances of incongruence, but these concerned internodes that exhibit very low IC scores in all additional analyses.

Examination of the patterns of incongruence in the Aspergillaceae phylogeny allows us to not only group the 14 incongruent internodes with respect to their patterns of conflict but also to postulate putative drivers of the observed incongruence. For example, both I15 and I60 are shallow internodes exhibiting low levels of incongruence, suggesting that one likely driver of the observed incongruence is ILS. In contrast, the shallow internodes I33, I43, I55, and I62 exhibit much higher levels of incongruence that are most likely to be the end result of processes, such as hybridization or repeated introgression. Finally, the remaining eight incongruent internodes (I3, I4, I24, I35, I52, I63, I74, and I78) exhibit varying levels of incongruence and are typically associated with single taxon long branches (Figures 1, 3, and 4), implicating taxon sampling as a likely driver of the observed incongruence. Given that inclusion of additional taxa robustly resolved the previously ambiguous placement of the long-branched *Aspergillus* section *Nigri* (see discussion above) as well as of other contentious branches of the fungal tree of life, such as the placement of the budding yeast family Ascoideaceae (Shen et al., 2017, 2018), we predict that additional sampling of taxa that break up the long branches associated with these seven internodes will lead to their robust resolution. Lastly, the IC value of internode I63 following removal of hidden paralogs marginally increased, suggesting that incongruence at this internode may also be associated with hidden paralogs.

Notably, the topology of our phylogeny was able to resolve two contentious issues that emerged from analyses of data matrices containing a few genes (Kocsubé et al., 2016; Taylor et al., 2016)

and that are important for taxonomic relationships within the family. Specifically, our phylogenetic analyses rejected the sister group relationship of genus *Penicillium* and *Aspergillus* section *Nidulantes* as well as the monophyly of a group of *Aspergillus* sections that are referred to as narrow *Aspergillus* (Table 1, p-value < 0.001 for all tests). Instead, our phylogeny shows that the genera *Aspergillus* and *Penicillium* are reciprocally monophyletic. These results are consistent with the current nomenclature proposed by the International Commission of *Penicillium* and *Aspergillus* (https://www.aspergilluspenicillium.org/), and inconsistent with the phylogenetic arguments put forward in proposals for taxonomic revision (Taylor et al., 2016). However, it should be noted that our study did not include representatives of the genera *Phialosimplex* and *Polypaecilum*, which lack known asexual stages, and appear to be placed within the genus *Aspergillus* (Kocsubé et al., 2016; Taylor et al., 2016). *Basipetospora* species also lack known asexual stages and are also placed within *Aspergillus* (Kocsubé et al., 2016; Taylor et al., 2016); unfortunately, the sole genome sequenced from this genus, JCM 23157, appears to be a contaminant from the class Sordariomycetes (Figure 1).

Finally, our relaxed molecular clock analysis of the Aspergillaceae phylogeny provides a robust and comprehensive time-scale for the evolution of Aspergillaceae and its two large genera, *Aspergillus* and *Penicillium* (Figure 5), filling a gap in the literature. Previous molecular clock studies provided estimates for only four internodes, mostly within the genus *Aspergillus* (Berbee and Taylor, 2001; Hedges et al., 2006; Vijaykrishna et al., 2006; Kensche et al., 2008; Sharpton et al., 2009; Da Lage et al., 2013; Beimforde et al., 2014; Fan et al., 2015; Gaya et al., 2015) and yielded much broader time intervals. For example, the previous estimate for the origin of Aspergillaceae spanned nearly 100 mya (50-146 mya: Berbee and Taylor 2001; Vijaykrishna et

al. 2006; Sharpton et al. 2009) while our dataset and analysis provided a much narrower range of 44.7 mya (mean: 117.4; 95% CI: 141.5 - 96.9). Notably, the estimated origins of genera *Aspergillus* (~81.7 mya) and *Penicillium* (~73.6 mya) appear to be comparable to those of other well-known filamentous fungal genera, such as *Fusarium*, whose date of origin has been estimated at ~91.3 mya (Ma et al., 2013; O'Donnell et al., 2013).

Fungi from Aspergillaceae have diverse ecologies and play significant roles in biotechnology and medicine. Although most of the 81 genomes from Aspergillaceae are skewed towards two iconic genera, *Aspergillus* and *Penicillium*, and do not fully reflect the diversity of the family, they do provide a unique opportunity to examine the evolutionary history of these important fungi using a phylogenomic approach. Our scrutiny of the Aspergillaceae phylogeny, from the Cretaceous to the present, provides strong support for most relationships within the family as well as identifies a few that deserve further examination. Our results suggest that the observed incongruence is likely associated with diverse processes such as incomplete lineage sorting, hybridization and introgression, as well as with analytical issues associated with poor taxon sampling. Our elucidation of the tempo and pattern of the evolutionary history of Aspergillaceae aids efforts to develop a robust taxonomic nomenclature for the family and provides a robust phylogenetic and temporal framework for investigation the evolution of pathogenesis, secondary metabolism, and ecology of this diverse and important fungal family.

CHAPTER 3

An evolutionary genomic approach reveals both conserved and species-specific genetic elements related to human disease in closely related *Aspergillus* fungi[2]

Introduction

The ability of a microbe to cause disease is a multifactorial trait that is dependent upon diverse genomic loci, including genes and non-coding regulatory elements. For opportunistic fungal pathogens, whose "accidental" infections of humans are not a part of their normal life cycle (Casadevall and Pirofski, 2007), the evolution of genomic loci contributing to virulence is thought to have been shaped by diverse evolutionary and ecological pressures, such as avoiding predation from soil-dwelling amoebae and surviving in warm and stressful environmental niches similar to those found inside human hosts (Tekaia and Latgé, 2005; Nielsen et al., 2007; Hillmann et al., 2015). However, the genetic differences between fungal pathogens and their non-pathogenic relatives have only recently begun to be understood (Fedorova et al., 2008; Butler et al., 2009; Sharpton et al., 2009; Moran et al., 2011; Taylor, 2015; Gabaldón et al., 2016; Gupta et al., 2020; Rokas et al., 2020a). This is especially true for filamentous fungi in the genus *Aspergillus*, which infect hundreds of thousands of humans each year (Brown et al., 2012; Bongomin et al., 2017).

Aspergillosis, the spectrum of diseases caused by fungi in the genus *Aspergillus*, afflicts a broad

[2]This work is published in: Mead, M. E., Steenwyk, J. L., Silva, L. P., de Castro, P. A., Saeed, N., Hillmann, F., et al. (2021). An evolutionary genomic approach reveals both conserved and species-specific genetic elements related to human disease in closely related *Aspergillus* fungi. Genetics 218. doi:10.1093/genetics/iyab066.

range of animals, including humans (Seyedmousavi et al., 2015). In humans, aspergillosis is a major global health issue and primarily affects individuals with compromised immune systems or who have other lung diseases or conditions (Barrs et al., 2013; Gregg and Kauffman, 2015; Frisvad and Larsen, 2016). Approximately 70% of aspergillosis patients are infected with *Aspergillus fumigatus*, but other members of the genus cause the rest of the infections, with no individual species responsible for a disproportionately large amount of cases (Alastruey-Izquierdo et al., 2014; Perlin et al., 2017; Latgé and Chamilos, 2019). Some of these pathogenic species are very closely related to *A. fumigatus* and belong to the same taxonomic section, section *Fumigati* (Balajee et al., 2005; Alastruey-Izquierdo et al., 2013; Houbraken et al., 2016). In contrast, most of the approximately 60 species in section *Fumigati* do not cause disease or have rarely been found in the clinic, suggesting that the ability to cause disease in humans evolved multiple times independently (or convergently) within *Aspergillus* (Rokas et al., 2020a). For example, *Aspergillus oerlinghausenensis* and *Aspergillus fischeri*, the two closest relatives of *A. fumigatus* are both considered non-pathogenic (Houbraken et al., 2016; Mead et al., 2019a; Steenwyk et al., 2020d).

Why some *Aspergillus* species routinely infect humans whereas their very close relatives never or rarely do remains an open question (Rokas et al., 2020a). To date, studies addressing this question have focused on comparing *A. fumigatus* to one or a few closely related (usually pathogenic) species (Fedorova et al., 2008; Wiedner et al., 2013; Sugui et al., 2014, 2015; Mead et al., 2019a; dos Santos et al., 2020b; Knowles et al., 2020; Steenwyk et al., 2020d) or to larger numbers of very distantly related species (Kjærbølling et al., 2018). Many individual genes and pathways are known to contribute to *A. fumigatus* virulence (Abad et al., 2010; Bignell et al.,

2016; Brown and Goldman, 2016; Steenwyk et al., 2021d), but if they are present or function in the same manner in other section *Fumigati* species, including in non-pathogens, has rarely been studied (Knowles et al., 2020; Steenwyk et al., 2020d). Genomic loci (i.e., genes and non-coding regulatory elements, and their variants) associated with pathogenicity could be shared or absent amongst all pathogens, including *A. fumigatus*, following a "conserved pathogenicity" model, or be uniquely present (or absent) in each pathogen ("species-specific pathogenicity" model) (Rokas et al., 2020a). The two models are not mutually exclusive, and the limited evidence available suggests that some genomic loci likely follow the conserved pathogenicity model (Fedorova et al., 2008; Kjærbølling et al., 2018; Knowles et al., 2020; Steenwyk et al., 2020d), whereas others follow the species-specific pathogenicity model (Fedorova et al., 2008; Kowalski et al., 2019; Mead et al., 2019a; Steenwyk et al., 2020d).

To study the signatures of the repeated evolution of human pathogenicity we conducted diverse evolutionary analyses on the genomes of 18 *Aspergillus* strains representing 13 species, including both pathogenic and non-pathogenic species from section *Fumigati*. Our results show that previously identified virulence-related genes are largely conserved throughout section *Fumigati* and outgroups. Consistent with the species-specific pathogenicity model, we found dozens of gene families that were present only in a given pathogen as well as dozens of genes whose evolutionary rates differed between a given pathogen and the rest of the taxa. For example, we identified 72 *A. fumigatus*-specific gene families and 34 genes whose evolutionary rate was uniquely different in *A. fumigatus*. Consistent with the conserved pathogenicity model, we identified over 1,700 genes that showed pathogen-specific evolutionary rates; however, we did not identify any gene families that were shared only by pathogenic taxa. To test whether our

approach could identify loci that contribute to *Aspergillus* disease-related traits, we carried out functional assays of deletion mutants of 17 transcription factor-encoding genes identified in our bioinformatic analyses as consistent with either pathogenicity model. We found that eight genes (four consistent with the conserved model and four consistent with the species-specific model) significantly affected pathogenicity-related traits, suggesting that the evolution of *Aspergillus* pathogenicity involved both conserved and species-specific genetic contributors. More broadly, our study shows that an evolutionary genomic approach is a useful framework for gaining insights into the molecular mechanisms by which *Aspergillus* species impact human health.

Materials and Methods

Genome procurement, assembly, and annotation

Genomes and annotations for *A. fumigatus* strains Af293 and A1163, along with all non-*A. fumigatus*, publicly available (as of July 2019) annotated genomes from section *Fumigati* were downloaded for analyses (see Table S1 from Mead et al., 2021 for NCBI accession numbers). We also obtained genomes and annotations for four outgroup species to facilitate phylogenetic analyses and comparisons. To expand the number of genomes analyzed, we assembled and/or annotated five additional *Aspergillus* genomes. More specifically, raw genomic reads for *A. fumigatus* strains F16311 and 12-7505446 were downloaded from NCBI for genome assembly and annotation. These strains were chosen because they, together with *A. fumigatus* strains Af293 and A1163, span the known diversity of *A. fumigatus* (Lind et al., 2017); additionally, available genomes for *A. cejpii* FS110, *A. neoellipticus* NRRL 5109, and *A. viridinutans* FRR 0576 were downloaded from NCBI and annotated (Abdolrasouli et al., 2015b; Li et al., 2018; Urquhart et al., 2019) (Table S1 from Mead et al., 2021). To quality trim sequence reads, we

used Trimmomatic, version 0.36 (Bolger et al., 2014) using parameters described elsewhere (Steenwyk and Rokas, 2017). The resulting high-quality reads were used as input to the genome assembly software SPAdes, version 3.8.1 (Bankevich et al., 2012), with the 'careful' parameter to reduce mismatches and short indels and the 'cov-cutoff' parameter set to 'auto.' Partial and complete gene models were predicted using Augustus, version 2.5.5 (Stanke and Waack, 2003), with the 'minexonintronprob' and 'minmeanexonintronprob' parameters set to 0.1 and 0.4, respectively. Genome annotation quality was assessed using BUSCO, version 2.0.1 (Waterhouse et al., 2018a), with the Pezizomycotina database of orthologs from OrthoDB, version 9 (Waterhouse et al., 2013). Genome annotation quality was similar between publicly available genomes and genomes assembled and/or annotated in the present project. For example, the publicly available assembly and annotation for *A. fumigatus* strain A1163 had 94.0% of BUSCO genes present in single copy while the assembled and annotated genome for *A. fumigatus* strain F16311 had 93.9% of BUSCO genes present in a single copy.

Inference of Gene Families

We first identified orthologous genes by clustering genes with high sequence similarity into orthologous groups using Markov clustering (van Dongen, 2000) as implemented in OrthoMCL, version 1.4 (Li et al., 2003), with an inflation parameter of 2.8. Gene sequence similarity was determined using a blastp "all-vs-all" using NCBI's Blast+, version 2.3.0 (Camacho et al., 2009) with an e-value cutoff of 1e-10, a 30% identity cutoff, and a 70% match cutoff. In subsequent analyses, these 14,294 orthogroups were used as proxies for gene families. 3,601 of the 14,294 orthogroups had all 18 taxa represented by a single sequence and are hereafter referred to as single-copy orthologous genes. Finally, 16 orthogroups had the same number of family members

in each taxon but in more than one copy, and 10,677 orthogroups had a different number of family members in at least one taxon.

Phylogenomic data matrix construction and analyses

To construct a phylogenomic data matrix, we retrieved the protein sequences of the 3,601 single-copy orthologous genes and individually aligned them with Mafft, version 7.402 (Katoh and Standley, 2013), using the same parameters as described elsewhere (Steenwyk et al., 2019c). Nucleotide sequences were threaded onto the protein alignments using the thread_dna function in PhyKIT, version 0.1 (Steenwyk et al., 2021b). The codon-based sequences were subsequently trimmed using trimAl, version 1.2rev59 (Capella-Gutierrez et al., 2009), using the 'automated1' parameter. The resulting single-gene alignments were concatenated into a single data matrix using the create_concat function in PhyKIT, version 0.1 (Steenwyk et al., 2021b).

To infer the evolutionary history of *Aspergillus* species in section *Fumigati* and the outgroup taxa, we used concatenation without gene-based partitioning, concatenation with gene-based partitioning, and gene-based coalescence in a maximum likelihood framework (Felsenstein, 1981; Rokas et al., 2003; Edwards, 2009; Zhang et al., 2018). For concatenation without gene-based partitioning, we used the 3,601-gene matrix as input to IQ-TREE (Nguyen et al., 2015) and inferred the best-fitting model of substitutions according to Bayesian information criterion values using the "-m TEST" parameter. The best-fitting model was determined to be a general time-reversal model with invariable sites, empirical nucleotide frequencies, and a discrete gamma model with four rate categories or "GTR+F+I+G4" (Tavaré, 1986; Yang, 1994; Gu et al., 1995). Lastly, we increased the number of candidate trees used during maximum likelihood

search by setting the "-nbest" parameter to 10. Bipartition support was assessed using 5,000 ultrafast bootstrap approximations (Hoang et al., 2018). We refer to the tree inferred using this method as the reference tree topology depicted in Figure 7.

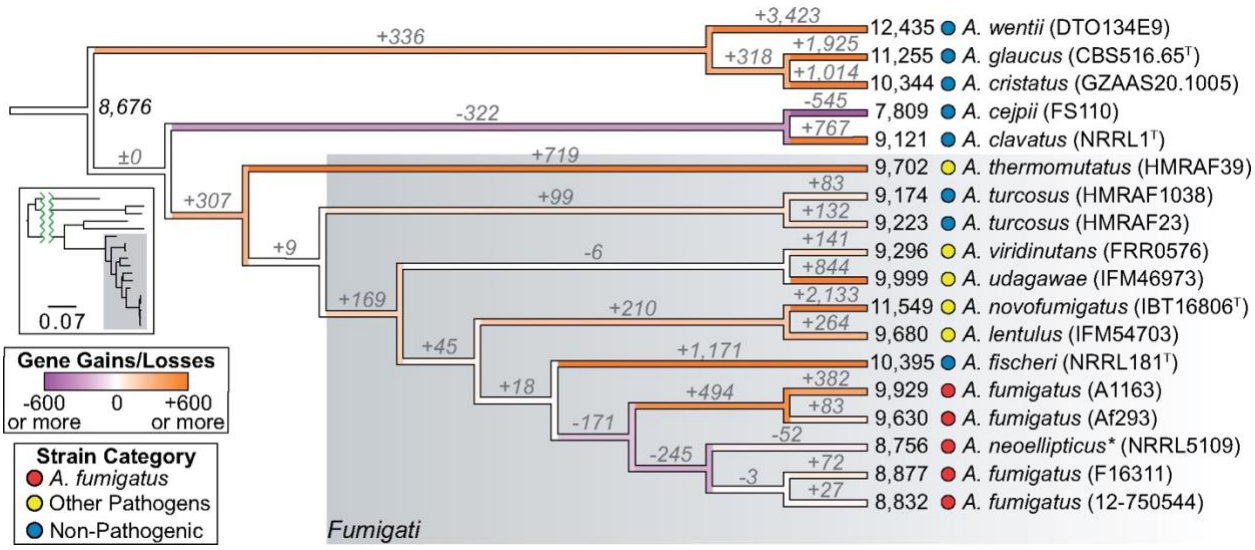

Figure 7. Genome-scale phylogeny and evolution of net gene gains or losses across *Aspergillus* section *Fumigati*.
Relationships among taxa in section *Fumigati* inferred from a concatenation-based, maximum likelihood approach. The asterisk next to *A. neoellipticus* denotes that in our coalescence approach this taxon was placed sister to all *A. fumigatus* strains. Gene gains and losses were calculated based on a maximum likelihood framework implemented in DupliPHY-ML (Ames *et al.* 2012) that utilized the 14,294 orthogroups we constructed as part of our phylogenomic analyses as proxies for gene families. Branches are colored based on the number of net gene gains or losses, and 8,676 genes were inferred at the last common ancestor of all taxa studied. Numbers at branch tips represent the total number of genes in that genome. Strain designations are in parenthesis next to species names and type strains are denoted by a superscript "T" next to their strain designations. Insert shows the phylogeny with branch lengths reflective of the estimated number of nucleotide substitutions per site (scale bar is 0.07 substitutions/site); taxa are in the same order as the larger cladogram. The number of gene gains, losses, and the net gain or loss are shown in Supplementary Table S2 from Mead et al., 2021.

To infer the evolutionary history of *Aspergillus* species in section *Fumigati* and the outgroup strains using concatenation with gene-based partitioning and coalescence, we first determined the best-fitting model of substitution using the "-m TEST" parameter and reconstructed the

phylogeny of the 3,601 single-copy orthologous genes individually using default IQ-TREE parameters (Nguyen et al., 2015). For concatenation with gene-based partitioning, we created a nexus-format partition file that describes gene boundaries in the 3,601-gene matrix and the best-fitting model of substitutions for each partition. We used the nexus-format partition file as input using the "-spp" parameter along with the concatenated 3,601-gene matrix to reconstruct the Fumigati phylogeny. Bipartition support was assessed using 5,000 ultrafast bootstrap approximations (Hoang et al., 2018). For coalescence, we first collapsed lowly supported bipartitions in all single-gene trees defined as less than 80% ultrafast bootstrap approximation support to reduce signal from poorly supported bipartitions. To do so, we assessed bipartition support using 5,000 ultrafast bootstrap approximations for individual single-gene trees (Hoang et al., 2018). To infer a coalescence-based phylogeny, we combined all single-gene trees with collapsed bipartitions into a single file and used it as input to ASTRAL-III, version 5.6.3 (Zhang et al., 2018), with default parameters. Bipartition support was assessed using posterior probabilities.

Gene family history

To determine the evolutionary history of the 14,294 gene families across section *Fumigati* species and outgroups, we implemented a maximum likelihood framework with a birth-death innovation model and gamma-distributed rates across families as implemented in DupliPHY-ML (Ames et al., 2012). DupliPHY-ML takes as input a matrix of gene family copy number and a phylogeny. To construct a matrix of gene family copy number, we used all orthologous groups of genes constructed as part of our phylogenetic analyses as proxies for gene families and used the

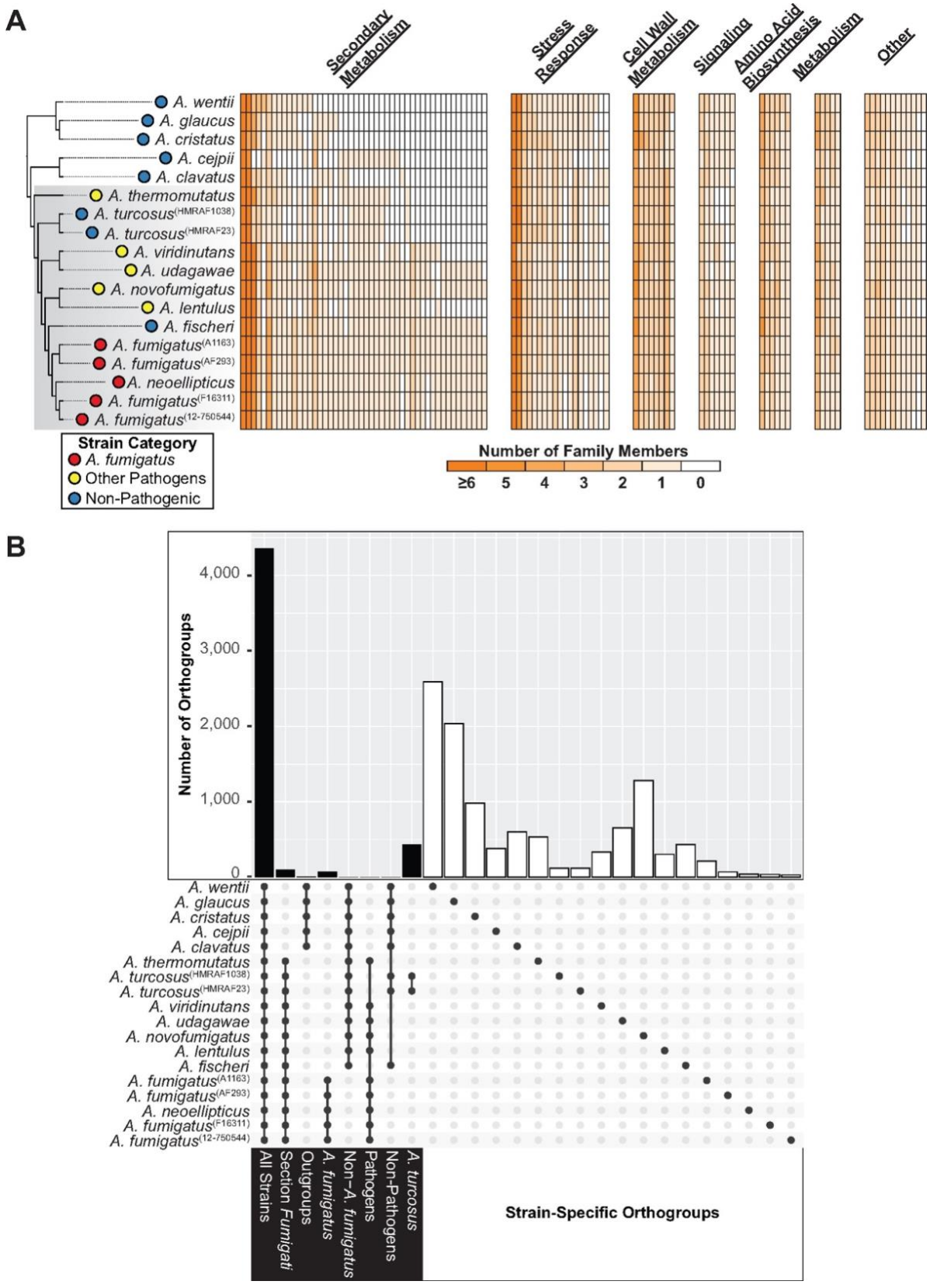
A
Secondary Metabolism
Stress Response
Cell Wall Metabolism
Signaling
Amino Acid Biosynthesis
Metabolism
Other
A. wentii
A. glaucus
A. cristatus
A. cejpii
A. clavatus
A. thermomutatus
A. turcosus(HMRAF1038)
A. turcosus(HMRAF23)
A. viridinutans
A. udagawae
A. novofumigatus
A. lentulus
A. fischeri
A. fumigatus(A1163)
A. fumigatus(AF293)
A. neoellipticus
A. fumigatus(F16311)
A. fumigatus(12-750544)
Strain Category
A. fumigatus
Other Pathogens
Non-Pathogenic
Number of Family Members
≥6 5 4 3 2 1 0
B
Number of Orthogroups
4,000
3,000
2,000
1,000
0
All Strains
Section Fumigati
Outgroups
A. fumigatus
Non-A. fumigatus
Pathogens
Non-Pathogens
A. turcosus
Strain-Specific Orthogroups

Figure 8. Gene families are largely conserved across section Fumigati, regardless of pathogenicity level.
(A) Some virulence-related genes have different presence/absence patterns across strains in section Fumigati. Left, cladogram from Figure 1 showing the relationships amongst the 18 genomes studied. Gray box indicates taxa belonging to section Fumigati. Right, heatmaps of the 105/189 gene families related to virulence that exhibited at least one gene presence/absence change in at least one taxon, split into groups based on their general biological functions. Gene family labels can be found in Supplementary Table S3 from Mead et al., 2021 in the same order presented here from left to right. (B) Gene families with representatives from all strains are the most prevalent. Upset plot (Conway et al. 2017) showing the number of all gene families present or absent in specific sets of strains. Black bars, gene family sets with members in more than one strain. White bars, strain-specific gene family sets.

---

number of gene sequences for a given species as the copy number information per gene family.

For the phylogeny, we used the reference phylogeny described previously.

Gene Ontology Enrichment Analyses

To determine if lists of genes of interest contained enriched Gene Ontology terms, we used GOATOOLs version 0.9.7 (Klopfenstein et al., 2018). Annotations for the *A. fumigatus* Af293 genome were downloaded from version 45 of FungiDB (Basenko et al., 2018), and the 2019-07-01 version of the basic Gene Ontology (Ashburner et al., 2000; GeneOntologyConsortium, 2004) was used for all analyses. A term was considered enriched if it had an adjusted p-value (using the Benjamini-Hochberg method) less than 0.05.

Gene Family Expansions and Contractions

To study if the number of gene family members is expanded or contracted in classes of strains (pathogens or non-pathogens) or in specific strains, we carried out a phylogenetically-informed analysis of variance with the phylANOVA function located within version 0.7-70 of the phytools package (Revell, 2012) with the 10,677 orthogroups that had a different number of family members in at least one taxon. Taxon relationships were provided from the phylogenetic tree that

resulted from the concatenation without gene based partitioning approach. The simulation-based ANOVA was performed for each gene family and run with 10,000 simulations in order to derive a p-value reflecting if the average number of genes was different in the three groups of strains: *A. fumigatus* strains, other pathogens, and non-pathogens (see the Results section for how these groups were defined). P-values were then corrected using the Benjamini-Hochberg method found within the "p.adjust" function in R. Gene families were considered significantly different if their adjusted p-values were less than 0.05. Tukey's range post-hoc test from the Python module "statsmodels" version 0.10.0 (Seabold and Perktold, 2010) was then carried out on significantly different gene families in order to determine if the average number of gene family members differed in any of the pairwise comparison (ex. the number of genes in *A. fumigatus* vs non-pathogenic species).

Estimating rates of molecular evolution

To determine the rate of sequence evolution across the evolutionary history of *Fumigati* species on a per gene basis, we used measures of the rate of nonsynonymous substitutions (dN) over the rate of synonymous substitutions (dS) (hereafter referred to as dN/dS or ω) using an approach described elsewhere (Steenwyk et al., 2019a). To do so, we used untrimmed codon-based alignments generated during the construction of the 3,601-gene matrix used for phylogenomic analyses. For each of the 3,601 genes, we calculated ω using PAML, version 4.9 (Yang, 2007), under two hypotheses: a null hypothesis (H_O) and an alternative hypothesis (H_A). For H_O, we allowed a single ω value to represent the rate of sequence evolution across the reference phylogeny. For the first H_A, we tested if different groups (*A. fumigatus*, other pathogens, or non-

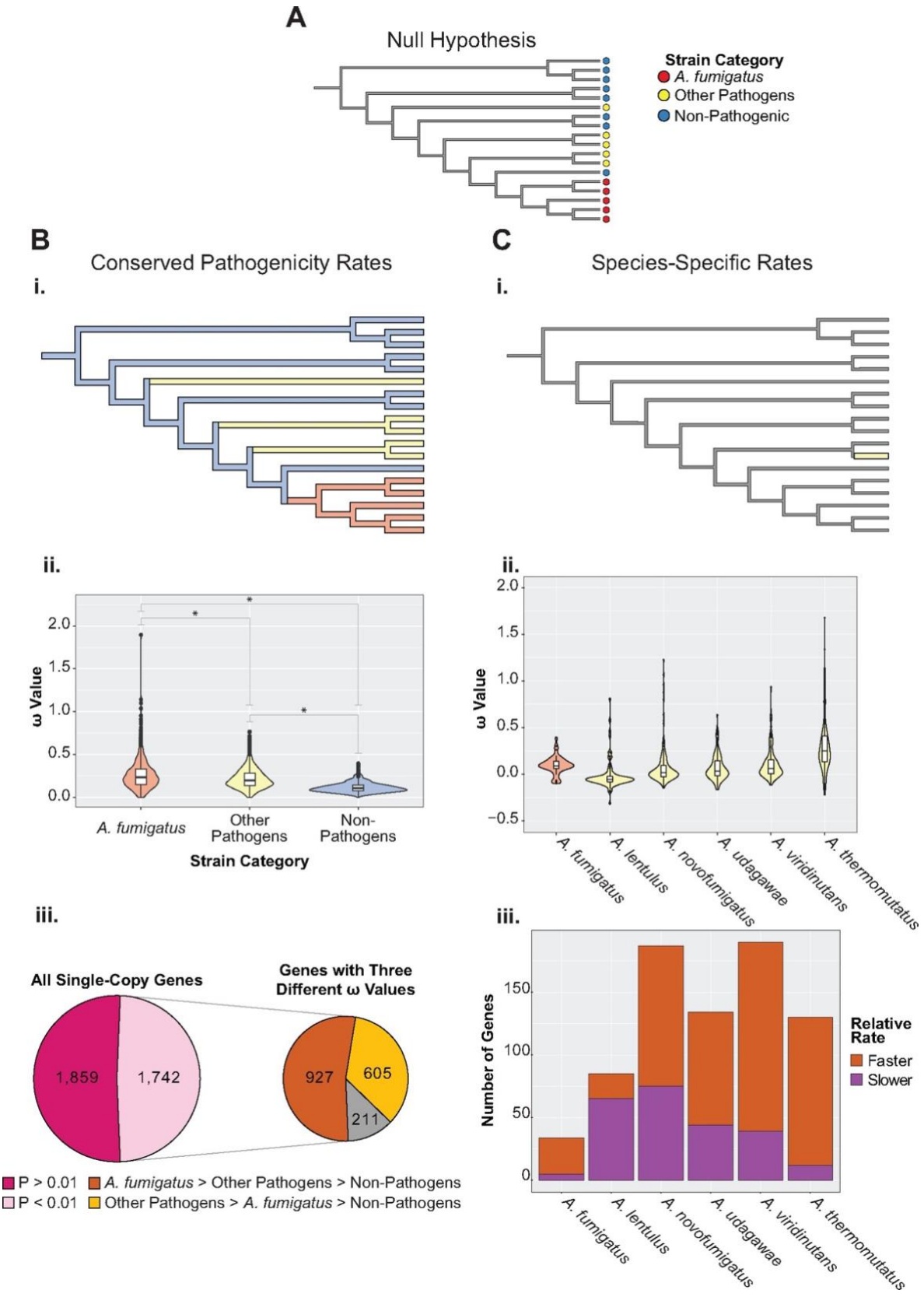
A
Null Hypothesis
Strain Category
A. fumigatus
Other Pathogens
Non-Pathogenic
B
Conserved Pathogenicity Rates
i.
ii.
ω Value
2.0
1.5
1.0
0.5
0.0
A. fumigatus
Other Pathogens
Non-Pathogens
Strain Category
iii.
All Single-Copy Genes
Genes with Three Different ω Values
1,859
1,742
927
605
211
P > 0.01
P < 0.01
A. fumigatus > Other Pathogens > Non-Pathogens
Other Pathogens > A. fumigatus > Non-Pathogens
C
Species-Specific Rates
i.
ii.
ω Value
2.0
1.5
1.0
0.5
0.0
−0.5
A. fumigatus
A. lentulus
A. novofumigatus
A. udagawae
A. viridinutans
A. thermomutatus
iii.
Number of Genes
150
100
50
0
Relative Rate
Faster
Slower
A. fumigatus
A. lentulus
A. novofumigatus
A. udagawae
A. viridinutans
A. thermomutatus

Figure 9. Genes in section Fumigati exhibit both pathogen- and species-specific rates of evolution.
(A) The null hypothesis that all branches in the phylogeny have the same ω value. (Bi) Alternative hypothesis examining the conserved pathogenicity model of gene evolution where genes are evolving at rates that correspond to their pathogenicity level. Branches are colored based on their strain category (*A. fumigatus*—most pathogenic, other pathogens, and nonpathogens—least pathogenic) and evolutionary rate. (Bii) Violin and box plots showing the ω values for each of 1,742 genes (49% of all single-copy genes) that exhibited different ω values (P-value < 0.01) in the three groups of strains (*A. fumigatus*, other pathogens, and nonpathogens). Ten genes had ω values > 0.8 in *A. fumigatus* strains and are not shown here (see Supplementary Table S4 from Mead et al., 2021). *adjusted P-value < 0.0001 in a Paired Wilcoxon Signed-Rank test. (Biii) Left, pie chart showing the number of genes that either did (light pink) or did not (dark pink) have different ω values in strains with different pathogenicity levels. Right, pie chart further detailing the relative magnitude of ω values of genes in strains with different pathogenicity levels. (Ci) A second alternative hypothesis examining the species-specific pathogenicity model of gene evolution where genes are evolving at one rate in one taxon and a different rate in all other taxa. Shown here is the model that was tested for *A. lentulus* where genes in *A. lentulus* experienced one rate of evolution (yellow), whereas their counterparts in all other taxa exhibited a different rate (gray). (Cii) Violin and box plots showing the difference between ω values of genes that had one ω in the pathogen, a different ω value in other taxa analyzed (P < 0.01), and were not found also to have pathogen-specific rates (i.e., were not included in the set of 1,742 genes shown in B). The numbers of genes used to construct these plots were 34 for *A. fumigatus*, 85 for *A. lentulus*, 187 for *A. novofumigatus*, 130 for *A. thermomutatus*, 134 for *A. udagawae*, and 190 for *A. viridinutans*. ω value differences between the pathogen of interest and all other strains that were greater than two are not shown and constituted only 13/643 comparisons that exhibited P < 0.01 and were not found in the conserved pathogenicity analyses (B). Strains are colored the same as in (B). (Ciii) Bar chart showing the number of genes where the gene in one pathogen is evolving faster or slower than its counterparts in all other strains. “Faster,” genes evolving faster in the pathogen compared with all other taxa. “Slower,” genes evolving slower in the pathogen compared with all other taxa. Taxa in all phylogenies are in the same order as in Figure 7.

---

pathogens) were associated with different rates of sequence evolution. For the second H_A, we tested if each gene was evolving at a unique rate in each pathogen, relative to the other branches in the tree (Figure 9Ci). For each comparison, to determine if H_A significantly differed from H_O, we used a likelihood ratio test ($\alpha = 0.01$).

Amoeba Predation Assays

To test whether our evolutionary genomic analyses could identify loci that contribute to resistance to phagocytosis, we conducted amoeba predation assays. Asexual spores (conidia) of *A. fumigatus* (either transcription factor mutants previously described (Furukawa et al., 2020) and obtained as described in (Zhao et al., 2019a) or the background strain CEA17) were incubated 4 h at 37°C in Czapek-Dox medium (CZD, Sigma-Aldrich Chemie, Munich, Germany) to induce swelling, and confronted with *Protostelium aurantium* at a prey-predator ratio of 10:1 (10^5 conidia and 10^4 trophozoites of *P. aurantium*) for 18 h at 22°C. Mutants were chosen based on the traits their genes possessed in the lists of genes identified in the evolutionary genomic analyses (ex. only in *A. fumigatus* or fast-evolving in pathogens). After confrontation, the assay plate was incubated for 1 h at 37°C to inactivate the amoebae. Subsequently, 0.002% [w/v] resazurin (Sigma-Aldrich, Taufkirchen, Germany) was added and metabolic rates were calculated from the time dependent reduction of resazurin to the fluorescent resorufin over 3 h at 37°C using an Infinite M200 Pro fluorescence plate reader (Tecan, Männedorf, Switzerland). Survival was determined from the difference in the metabolic rates of the fungus after amoeba confrontation and amoeba-free controls. These controls were also used to determine the fitness of each strain in CZD-medium. Essentially the same assay was carried out to determine the survival of germlings of *A. fumigatus*, except that conidia of *A. fumigatus* were pre-grown to germlings for 10 h at 37°C in CZD medium before the addition of trophozoites of *P. aurantium*.

Virulence assays in the great wax moth (Galleria mellonella) model of fungal disease

To test whether our evolutionary genomic analyses could identify loci that contribute to fungal disease, we conducted virulence assays using the greater wax moth (*Galleria mellonella*) model

of fungal disease. *Galleria mellonella* larvae were obtained by breeding adult larvae (Fuchs et al., 2010) weighing 275-330 mg in starvation conditions in petri dishes at 37°C in the dark for 24 hours prior to infection. All selected larvae were in the final stage of larval (sixth) stage development. Fresh asexual spores of each strain of *A. fumigatus* were obtained. For each strain, spores were counted using a hemocytometer and the initial concentration of the spore suspensions for the infections were 2×10^8 spores/ml. A total of 5 µl (1×10^6 spores) of each suspension was inoculated per larva. The control group was composed of larvae inoculated with 5 µl of PBS to observe death by physical trauma. The inoculation was performed using a Hamilton syringe (7000.5KH) through the last left proleg. After infection, the larvae were kept in petri dishes at 37°C in the dark and were scored daily. Larvae were considered dead when a lack of movement was observed in response to touch. The viability of the inoculum administered was determined by plating a serial dilution of the asexual spores in 37% YAG medium. The statistical significance of the comparative survival values was calculated using the log rank analysis of Mantel-Cox and Gehan-Brestow-Wilcoxon found in the statistical analysis package Prism.

Data Availability

All supplementary material and their descriptions can be found on figshare at https://doi.org/10.6084/m9.figshare.14424386.

Results

A genome-scale phylogeny of *Aspergillus* section *Fumigati*

Phylogenetic relationships among taxa in section *Fumigati* (Table S1 from Mead et al., 2021) were examined using three maximum likelihood approaches – concatenation without gene-based partitioning, concatenation with gene-based partitioning, and coalescence – using 3,601 single-copy orthologous genes. These 3,601 genes were the subset of the 14,294 groups of orthologous genes inferred for these 18 taxa (see Methods). Both concatenation approaches yielded the same topology, recovering *A. neoellipticus* nested within *A. fumigatus* (Figure 7). All bipartitions received full support except the split between *A. neoellipticus* and *A. fumigatus* strains F16311 and 12-750544, which received 98% ultrafast bootstrap approximation support. The coalescence approach inferred a fully supported alternative topology that placed *A. neoellipticus* sister to the four *A. fumigatus* strains. Whether *A. neoellipticus* is conspecific with *A. fumigatus* or a distinct species has been previously discussed in the literature (Li et al., 2014) and our genome-scale analyses reflect this debate. Given the close evolutionary relationship of the two species, we choose to refer to *A. neoellipticus* as a strain of *A. fumigatus* rather than a distinct species.

Broad conservation of genes and gene families, including those related to virulence, across section *Fumigati*

To understand variation in the distribution of genes, including ones known to be involved in *A. fumigatus* virulence (Abad et al., 2010; Bignell et al., 2016; Kjærbølling et al., 2018; Mead et al., 2019a; Urban et al., 2019; Steenwyk et al., 2021d), we inferred gene and gene family gains and losses for every branch on the phylogeny (Figures 7 and S1 from Mead et al., 2021). The number of gene family members at each node of the tree was estimated based on a maximum likelihood

framework implemented in DupliPHY-ML (Ames et al., 2012) that utilized the 14,294 orthogroups we constructed as part of our phylogenomic analyses as proxies for gene families. The same dataset was used in our gene family analyses; in these analyses, we did not use the numbers of family members in each taxon but whether a specific gene family was present / absent in a given taxon. We inferred a net gain of 307 genes in the last common ancestor of section *Fumigati*. In addition, we found a net loss of 171 genes in the last common ancestor of *A. fumigatus* strains (Figure 7 and Table S2 from Mead et al., 2021). An estimated net gain of 494 genes occurred in the last common ancestor of the two *A. fumigatus* reference strains, A1163 and Af293. The same general patterns of genome expansion and contraction were observed when gene family gain and loss were estimated (Figure S1 from Mead et al., 2021).

To identify genes and gene families whose evolution was consistent with the conserved pathogenicity and species-specific pathogenicity models, we searched the 18 *Aspergillus* genomes for genes that were conserved across pathogens or specific to individual pathogens using both candidate and unbiased approaches. The candidate approach consisted of inferring the presence or absence pattern of 206 virulence-related genes (Steenwyk et al., 2021d) in each of the 18 genomes. Our gene family analysis placed the 206 virulence-related genes into 189 gene families. The largest virulence-related gene family (containing the transporter *abcC* – Afu1g14330 (Paul et al., 2013)) had 259 family members spread across all 18 genomes, and the smallest gene family (containing the terpene cyclase *fma-TC* – Afu8g00520 – from the fumagillin biosynthetic gene cluster (Guruceaga et al., 2018)) had six family members spread across only six of the genomes we analyzed. The same number of family members was found in every genome for 84/189 (~44%) of the virulence-related gene families, including for 81 families

with one gene family member in each strain. Of those virulence-related gene families that differed in family member size across the 18 genomes, there were no virulence-related gene families with only members in pathogens, section *Fumigati* species, or *A. fumigatus* (Figure 2A).

We inferred that 164/189 virulence-related gene families were already present in the last common ancestor of all section *Fumigati* species. Similarly, we estimated that on average, 12 genes have been lost from virulence-related gene families during the evolution of *A. fumigatus* strains 12-750544, F16311, and *A. neoellipticus* and 2 genes have been gained from virulence-related gene families during the evolution of *A. fumigatus* strains Af293 and A1163 compared to the *A. fumigatus* last common ancestor. The finding that many virulence-related genes are conserved across both pathogens and non-pathogens in section *Fumigati* suggests that most known genetic determinants of virulence likely evolved for functions other than causing disease in humans and have been instead recruited into performing roles important for pathogenicity in certain species.

Our unbiased approach consisted of analyzing all 14,294 gene families that resulted from constructing orthologous groups of genes from all 18 *Aspergillus* genomes. Similar to what we observed with virulence-related genes, we found that 4,361/14,294 gene families (~31%) had family members in each of the 18 strains analyzed, and no gene families were present only in pathogens (Figure 2B). However, we found 98 gene families that were specific to section *Fumigati* (Figure 2B and Table S4 from Mead et al., 2021). While the 98 gene families were not enriched for any Gene Ontology biological processes, molecular functions, or cellular compartments, the group contained genes associated with previously identified virulence-related

traits, such as: a gene encoding a dimethylallyl tryptophan synthase (*cdpNPT* – Afu8g00620) located near the fumitremorgin-fumagillin-pseurotin supercluster (Yin et al., 2007; Wiemann et al., 2013), a major facilitator type transporter (*mdr3* – Afu3g03500) whose gene is highly expressed in *A. fumigatus* strains resistant to drugs (Nascimento et al., 2003; Da Silva Ferreira et al., 2004), and a homolog of *mgtC* (Afu7g05060), a bacterial virulence factor required for survival in macrophages (Blanc-Potard and Groisman, 1997; Gastebois et al., 2011).

We found 72 gene families that were uniquely present in *A. fumigatus* (Figure 2B and Table S4 from Mead et al., 2021). These *A. fumigatus*-specific genes were not enriched for any GO terms and have not previously been tested for roles in virulence-related traits. The number of uniquely present gene families in other pathogens ranged from 1,280 in *A. novofumigatus* to 303 in *A. lentulus*. We also found two gene families (predicted glucose-methanol-choline oxidoreductase family members - NFIA_036190/NFIA_036210, and a membrane dipeptidase - NFIA_057190) that had members in all other taxa except *A. fumigatus* and one gene family found only in non-pathogenic taxa (a hypothetical protein with no identifiable domains – NFIA_057720). In summary, gene families are largely conserved across pathogens and non-pathogens in section *Fumigati*, but 72 gene families were found only in *A. fumigatus*; while these gene families have yet to be investigated for their potential roles in virulence, they represent candidates for the species-specific pathogenicity model.

The distributions of few gene families are associated with pathogenicity

Our analyses did not identify gene families whose presence/absence patterns were conserved in all pathogens found in section *Fumigati*. An alternative hypothesis is that the number of gene

family members in a given taxon could reflect the organisms' ability to cause disease. To test this hypothesis we carried out a phylogenetically informed ANOVA (Revell, 2012) on all 10,677 gene families that displayed a different number of gene family members in at least one taxon. For this analysis we split the 18 *Aspergillus* taxa into three groups based on their pathogenicity levels (i.e., how frequently they are generally found in the clinic): *A. fumigatus* (most pathogenic), other pathogens (that are not *A. fumigatus*), and non-pathogens (least pathogenic). While we focus here on identifying pathogenicity-related genes, this approach will likely also identify genes important for *A. fumigatus*-specific traits unrelated to pathogenicity as *A. fumigatus* is the only species in its category.

We found 83 gene families that had statistically significant differences in the number of members between groups (Figure S2 from Mead et al., 2021). After conducting Tukey's post-hoc test on all 83 gene families, we observed that 72/83 gene families had more copies in *A. fumigatus* and were in fact those previously identified as "*A. fumigatus*-specific" in our strict gene presence/absence analysis (Table S4 from Mead et al., 2021). One of the remaining 11 gene families was the membrane dipeptidase (NFIA_057190) found during our gene presence/absence analysis in all genomes other than *A. fumigatus*. The hypothetical gene family (NFIA_057720) found only in non-pathogenic species with the same gene family presence/absence analysis (Figure 8B) was also identified via the phylogenetically informed ANOVA. The remaining nine gene families had the same number of genes in *A. fumigatus* and non-pathogens but a different number of family members in the other pathogens. Three (P174DRAFT_459701, P174DRAFT_448681, and P174DRAFT_440824) possessed no conserved domains, and the other six had a carbohydrate binding domain (P174DRAFT_502341 - PF09362), three ankyrin

repeat domains (P174DRAFT_501662 – PF12796), a major facilitator superfamily domain (P174DRAFT_432793 – PF07690), a domain of unknown function (P174DRAFT_509497 – PF11905), a sulfur-carrier domain (P174DRAFT_347437 – PF03473), and an aldehyde dehydrogenase family domain (P174DRAFT_378794 – PF00171), respectively. Together, these data show that few gene families exhibit significant variation in their numbers across section *Fumigati* with respect to pathogenicity.

Many genes experienced faster rates of evolution in pathogenic species

Another way in which genomes evolve that can affect pathogenicity is through changes in the evolutionary rates of their constituent genes (Yang and Bielawski, 2000). We carried out two evolutionary rate analyses to test whether our set of 3,601 single-copy orthologous genes exhibited different rates of evolution in pathogens compared to non-pathogens. For both analyses, our null hypothesis was that for a given single-copy gene, a single rate (ω) represented the rate of sequence evolution for each gene in every strain, regardless of the pathogenicity level of the organisms examined (Figure 9A). In the first analysis, our alternative hypothesis was that each gene evolved at a unique rate in each of our three groups (*A. fumigatus* strains, other pathogens, and non-pathogens) (Figure 9Bi). We observed that 49% of genes tested (1,742/3,601) rejected the null hypothesis, suggesting that the evolutionary rate of these genes differs among the three groups. Of the 1,742 genes with three different ω values, 88% (1,532/1,742) had faster rates in pathogenic organisms (Figure 9Biii) and 10 had relatively high ω values (> 0.8) in *A. fumigatus* (Table S4 from Mead et al., 2021). None of these 10 fastest-evolving genes have previously been studied and contain a variety of domains likely involved in diverse functions ranging from RNA binding to catalyzing oxidation/reduction reactions. Each

group also exhibited its own statistically different distribution of ω values (Figure 9Bii). Of the 81/189 virulence-related gene families that were present in a single copy in each *Aspergillus* genome, 56% (45/81) exhibited different rates of evolution.

In the second analysis, we tested if each gene was evolving at a unique rate in each pathogen, relative to the other species we analyzed (Figure 9Ci). We found that, on average, 127 single-copy genes exhibited a different rate in the pathogen of interest than in the rest of species and were not found also to have pathogen-specific rates, with most evolving faster in the pathogens (Figure 9Cii and 9Ciii). *A. fumigatus* had the smallest number of genes whose evolutionary rates differed from the rest of the species (34), while *A. viridinutans* had the most (190) (Table S5 from Mead et al., 2021). Overall, our data show that genes in pathogens are evolving faster than in non-pathogens, both in a conserved and species-specific manner.

Transcription factors with pathogenicity-related patterns of evolution have diverse effects on virulence

To test if any of the genes whose evolutionary signatures differed between pathogens and non-pathogens directly affected either fungal or host survival, we tested 17 knockout strains of transcription factor-encoding (TF) genes (Furukawa et al., 2020) in two virulence-related assays. One TF (Afu7g00210) was found only in *A. fumigatus* (Figure 8B), one (Afu6g08540) was identified as being fast-evolving in *A. fumigatus*, four (Afu2g17895, Afu3g02160, Afu7g04890, and *gliZ* - Afu6g09630) were members of gene families with a statistically significant higher number of family members in pathogens in a preliminary one-way ANOVA (but not in our phylogenetically-informed ANOVA), five (Afu1g11000, Afu2g00470, Afu6g11750,

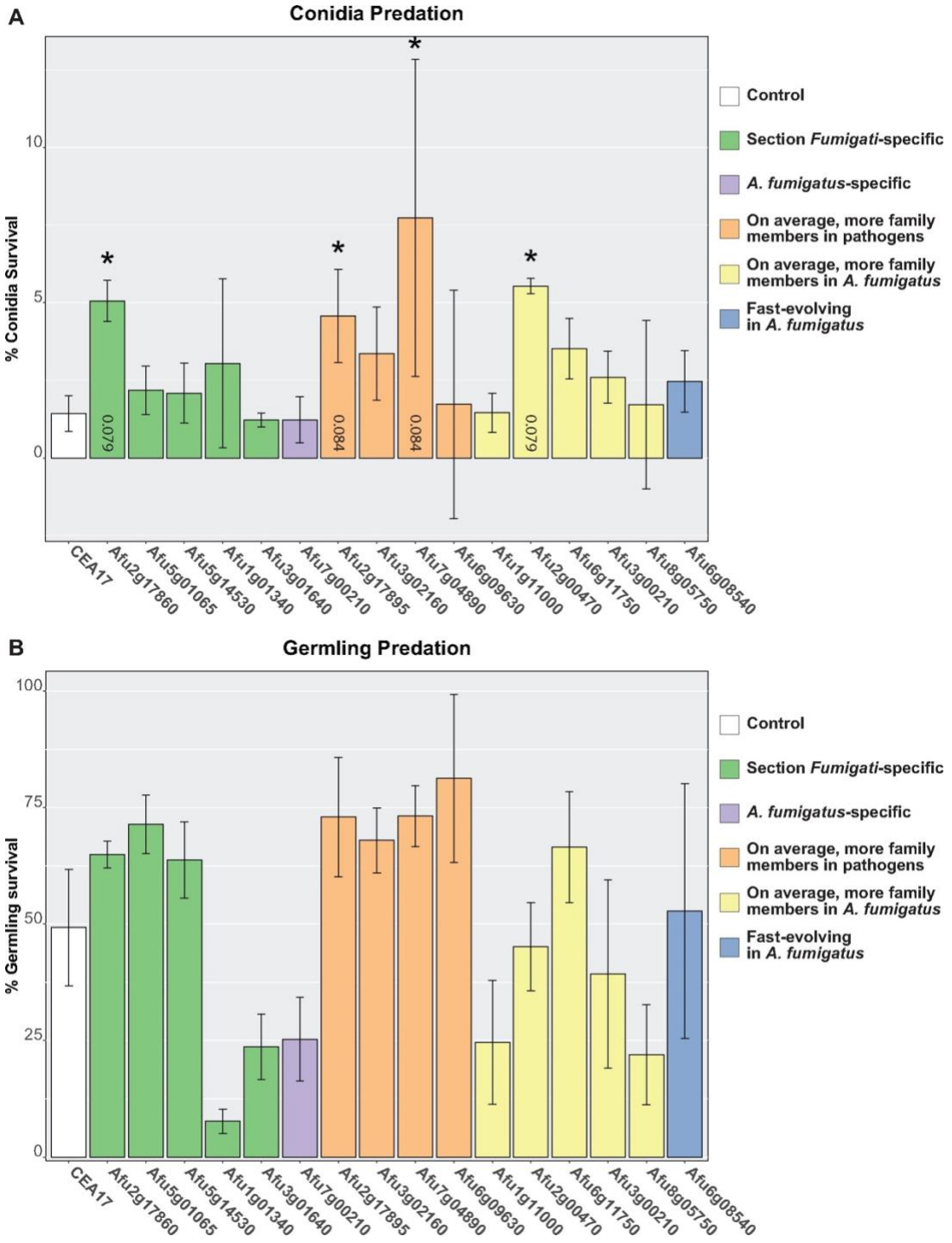

Figure 10. Multiple transcription factors whose evolution varies with respect to *Aspergillus* pathogenicity affect the survival of *A. fumigatus* during amoeba predation. (A) Survival of swollen *A. fumigatus* asexual spores (conidia) after interaction with *P. aurantium*. Spores of *A. fumigatus* were incubated 4 h at 37°C in CZD medium and confronted with *P. aurantium* at a prey–predator ratio of 10:1 (10^5 spores and 10^4 trophozoites of *P.*

aurantium). Survival is expressed as the relative reduction in the metabolic rate of the fungus in comparison to amoeba-free controls over 3 h. Data represent the mean and SD of three biological replicates. $*P < 0.1$ in an adjusted Dunn Test comparing the survival of the mutant strain to the parental strain CEA17. (B) Survival of *A. fumigatus* germlings after interaction with *P. aurantium*. Asexual spores of *A. fumigatus* were pre-grown to germlings for 10 h at 37°C in CZD medium and confronted with *P. aurantium*. All other assay parameters are the same as in (A). No mutant strain exhibited statistically significant difference in survival relative to CEA17 in an adjusted Dunn test. Both asexual spores and germling confrontation assays were confirmed to have significant *P*-values (<0.05) in Kruskal–Wallis tests before carrying out the post-hoc test and mutant strains did not exhibit large growth phenotypes in the absence of amoeba (Supplementary Figure S4 from Mead et al., 2021). Mutants are color-coded based on genomic traits related to pathogenicity that they possess. For the "On average, more family members in pathogens" and "On average, more family members in *A. fumigatus*" groups, the family members did not exhibit a statistically significant different number of family members in our phylogenetically informed ANOVA (Supplementary Figure S2 from Mead et al., 2021). Knockout mutants were constructed in the CEA17 background (Furukawa *et al.* 2020), but *A. fumigatus* strain Af293 gene ids for the corresponding orthologous genes are shown here and in Figure 5. Gene absence patterns were confirmed with tblastn. A potential, low-confidence ortholog of Afu5g01065 was found in *A. wentii*; however, all other *A. fumigatus* genes were found missing in the species listed. The mutant of Afu2g16310 could not be assayed due to technical reasons.

Afu3g00210, and Afu8g05750) were members of gene families with a statistically significant higher number of family members in *A. fumigatus* in the same preliminary one-way ANOVA (but also not in our phylogenetically-informed ANOVA), and six (Afu2g17860, Afu5g01065, Afu5g14530, Afu1g01340, Afu3g01640, and Afu2g16310) were found only section *Fumigati*. None of the TF mutants exhibited a growth defect compared to their parent strain (CEA17) when grown in conventional lab conditions (Figure S3 from Mead et al., 2021).

In the first assay, asexual spores (conidia) or germlings from either a background strain of *A. fumigatus* (CEA17) or one of the *A. fumigatus* knockout mutants of transcription factors were incubated with *Protostelium aurantium*, a fungivorous amoeba used to study how fungi may have evolved the ability to evade or survive phagocytosis by human immune cells (Radosa et al., 2019). We found that mutant asexual spores of four genes (Afu2g17860, Afu2g17895,

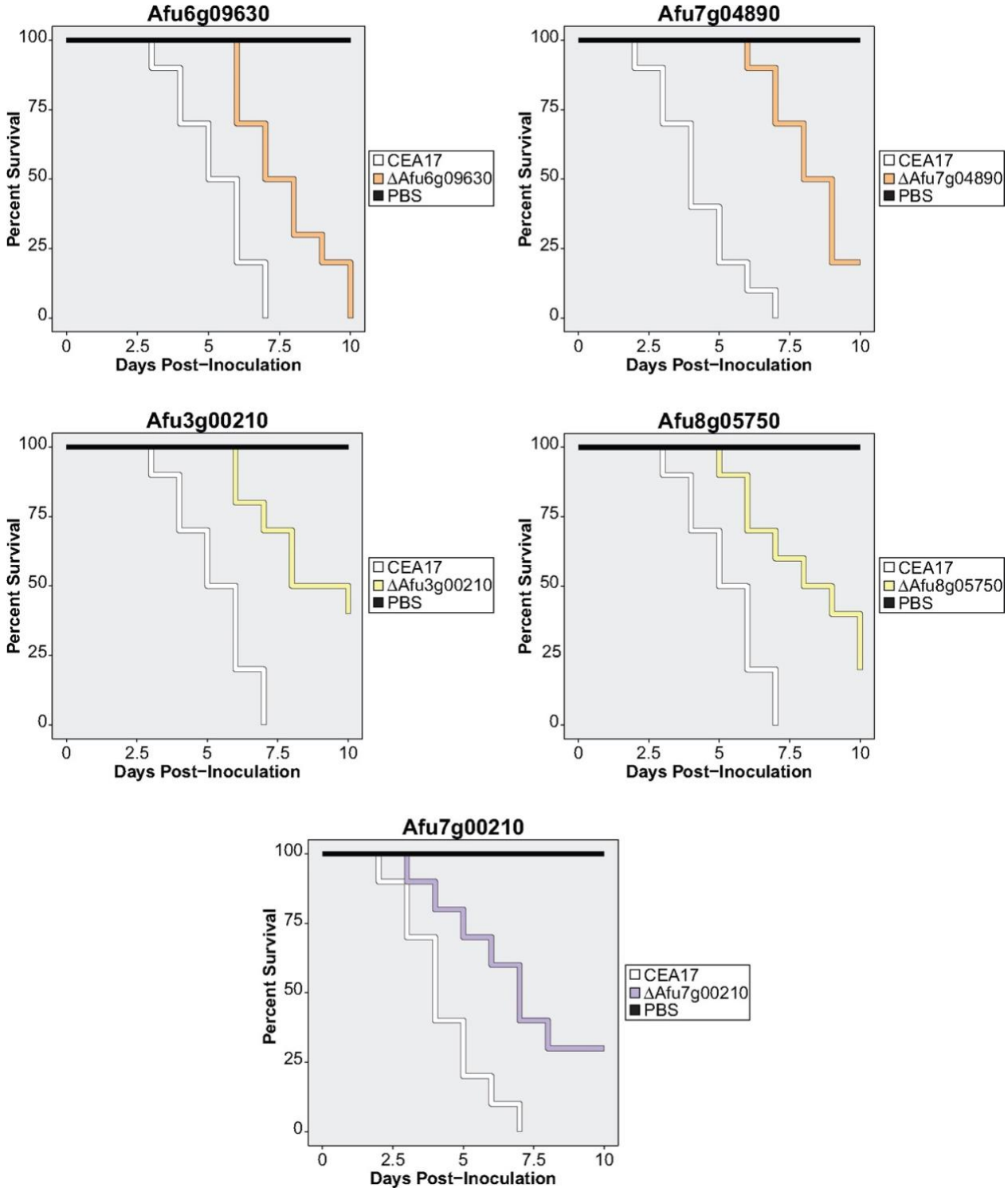
Afu6g09630
Percent Survival
Days Post−Inoculation
CEA17
ΔAfu6g09630
PBS
Afu7g04890
Percent Survival
Days Post−Inoculation
CEA17
ΔAfu7g04890
PBS
Afu3g00210
Percent Survival
Days Post−Inoculation
CEA17
ΔAfu3g00210
PBS
Afu8g05750
Percent Survival
Days Post−Inoculation
CEA17
ΔAfu8g05750
PBS
Afu7g00210
Percent Survival
Days Post−Inoculation
CEA17
ΔAfu7g00210
PBS

Figure 11. Multiple transcription factors in *A. fumigatus* whose evolution differs with respect to pathogenicity affect virulence in the greater wax moth model of disease.
Cumulative survival of *G. mellonella* larvae inoculated with phosphate buffered saline (control; black), asexual spores of the parental strain CEA17 (white), and asexual spores from null mutants of TFs whose evolution is associated with the observed differences in *Aspergillus* pathogenicity (various colors). Ten larvae were used per inoculation in all assays. Color scheme is the same as in Figure 10. All mutant survival curves shown here were statistically different ($P < 0.008$ in a Log-rank test) from the CEA17 survival curve. Mutants whose survival curves are shown in orange on average have more gene family members in pathogens, those in yellow have on average more family members in *A. fumigatus*, and the mutant in purple is *A. fumigatus*-specific. Note that the results of the mutants in orange support the conserved pathogenicity model, whereas those of the mutants in yellow and purple support the species-specific pathogenicity model and that the mutants with orange and yellow survival curves did not exhibit a statistically significant different number of family members in our phylogenetically informed ANOVA (Supplementary Figure S2 from Mead et al., 2021). The Afu6g09630 gene is *gliZ*, a regulator of the biosynthesis of the secondary metabolite gliotoxin, a known modulator of host biology.

Afu7g04890, and Afu2g00470) exhibited an increase in survival relative to CEA17 (Figure 10A). Given that overall spore killing of all strains tested was greater than 90%, to independently confirm the increase in viability of these four mutants, we conducted predation assays with germlings. While germlings of many mutants showed a qualitative difference in survival relative to CEA17 (including three out of the four with statistically significant increases in viability during the spore predation assays), those differences were not statistically significant (Dunn's Test adjusted p-value > 0.1).

In the second assay, we measured virulence in the greater wax moth (*Galleria mellonella*) model of *A. fumigatus* disease. We found that almost one third of all knockout mutants tested (5/17) exhibited a statistically significant decrease in virulence (Figures 5 and S5 from Mead et al., 2021). One of the transcription factor mutants that resulted in a significant decrease in larval killing was that of *gliZ* (Afu6g09630), a regulator of gliotoxin production (Bok et al., 2006) that we observed was found in all pathogenic taxa but missing in all non-pathogenic ones except *A.*

cejpii and *A. fischeri*. To our knowledge, this is the first time *gliZ* has been tested in the greater wax moth model of fungal disease or shown to contribute to fungal pathogenesis. None of the other transcription factors whose mutants exhibited a decrease in virulence have previously been studied, and for three (Afu3g00210, Afu8g05750, and Afu7g00210) the only PFAM domain they contained was a "Fungal Zn(2)-Cys(6) binuclear cluster domain" (PF00172) while the other predicted transcription factor (Afu7g04890) contained both the binuclear cluster domain and a "Fungal specific transcription factor domain" (PF04082). In summary, the data from both functional assays suggest that genes whose evolution differs between pathogens and non-pathogens are likely to contribute to disease-related traits.

Discussion

Our examination of *Aspergillus* genes whose evolution is associated with the observed differences in pathogenicity among section *Fumigati* taxa identified candidate genes that support the conserved pathogenicity model as well as candidates that support the species-specific pathogenicity model. Our results also show that previously described virulence-related genes are largely present in both pathogens and non-pathogens (Figure 8A), suggesting that most known genetic determinants of virulence are not likely to explain the observed pathogenicity differences between *Aspergillus* species.

Multiple transcription factors we identified as having virulence-related genomic traits also displayed roles in different virulence-related assays (Figures 10 and 11). All four spore mutants that were significant in the amoeba predation assays showed increased viability compared to the control strain. Survival differences in the presence of a phagocytic predator can occur at several

levels, such as recognition, uptake, or intracellular fate. It is possible that a mutant that has acquired an advantage against a phagocyte has a trade off in the complex environment of the host. For example, modified cell surface components may allow escape from recognition by a phagocyte, but also result in better adhesion to surface structures in the host. Mutants of *Cryptococcus*, another fungal pathogen of humans, can undergo filamentation and then survive better against amoebae, but in the host the increase in filamentation results in a decrease in virulence due to a reduction of fungal dissemination (Magditch et al., 2012). It is possible that some of the *A. fumigatus* mutants have a slightly extended resting stage and thus escape the predator over the limited time of the assay (dormant wildtype spores are inert to the amoebae (Ferling et al., 2020)). We observed no major differences in growth between the mutants and the wildtype strain (Figure S3 from Mead et al., 2021), but these resting stage effects may be very small, and much is still unknown regarding how they may impact pathogenicity. In addition, our amoeba predation assays measure one or a few events during disease progression, namely phagocytic cell interactions, but the greater wax moth model tested the entirety of disease progression in a susceptible host; those results showed the expected outcome of decreased virulence in the knockout mutants.

Of the eight transcription factors whose null mutants exhibited at least one phenotype in our two assays, half of them (including one of the genes, Afu2g17860, whose mutant increased viability of spores in the amoeba predation assay) were downregulated and none were upregulated during the switch to human body temperature (Lind et al., 2016). Additionally, *gliZ*, a regulator of gliotoxin production whose gene family we found to be largely pathogen-specific and whose mutant was less virulent than the WT strain, was heavily upregulated in *A. fumigatus* germlings

that were extracted 12-14 hours after mouse infection (McDonagh et al., 2008) while the seven other TFs we studied were not differentially regulated during the early events of mouse infection. Taken together, our functional assays show that our evolutionary genomic approach is useful for uncovering the molecular mechanisms underpinning the evolution of pathogenicity and also has the power to identify genes both previously connected to *A. fumigatus* virulence in addition to novel ones.

We analyzed all sequenced species in section *Fumigati* (as of July 2019) and a representative sampling of strains from *A. fumigatus*, carried out a diverse set of evolutionary genomic analyses, and functionally tested our identified genes in multiple assays, thus building on previous studies that used smaller numbers of section *Fumigati* species and close relatives in *Aspergillus* and focused on strict gene presence/absence (Fedorova et al., 2008; Mead et al., 2019a). Previous work also compared *A. novofumigatus*, one of the section *Fumigati* species we considered here, to its relative *A. fumigatus* (Kjærbølling et al., 2018), and while that study used a broader and less stringent list of virulence-related genes that also included allergens, they also saw high levels of gene conservation between the two species. This previous work and our own support the hypothesis that *A. novofumigatus* could be nearly as pathogenic as *A. fumigatus* due to this conservation of almost all virulence-related genes. In section *Flavi*, another taxonomic section in genus *Aspergillus* that contains the human and plant pathogen *Aspergillus flavus*, it has been hypothesized that transcription factors may be linked to pathogenicity (Kjærbølling et al., 2020), and similarly, we saw that one of the 72 *A. fumigatus*-specific genes we identified is a transcription factor whose deletion reduces virulence in our invertebrate model of fungal disease (Figure 11).

As more genomes from strains and species in section *Fumigati* become available, our power to detect quantitative differences will increase, and allow us to more robustly test the conserved pathogenicity model and expand our species-specific pathogenicity model to include "strain-specific" elements. This will be especially important considering the continued and growing appreciation for strain-specific traits and differences in *Aspergillus* genomes and pathogenicity (Keller, 2017; Ries et al., 2019; Bastos et al., 2020b; dos Santos et al., 2020b; Kjærbølling et al., 2020; Steenwyk et al., 2020d, 2020c; Kowalski et al., 2021). Similarly, we recognize that it is unlikely that all of the *A. fumigatus*-specific genes (Figure 2B) or genomic attributes (Figure 3) we discovered may be directly connected to pathogenicity but may instead be connected to other *A. fumigatus*-specific traits. This caveat notwithstanding, these *A. fumigatus*-specific genes constitute a useful list of targets for beginning to understand why *A. fumigatus* evolved to be pathogenic whereas its closest relatives did not. Future studies will also place *A. oerlinghausenensis*, another species closely related to *A. fumigatus* (Houbraken et al., 2016), within this evolutionary framework of pathogenicity, but based on our recent genome-wide phylogenomic analyses of *A. oerlinghausenensis*, *A. fischeri*, and *A. fumigatus* (Steenwyk et al., 2020d), we do not anticipate that inclusion of *A. oerlinghausenensis* will drastically change our findings.

The strains whose genomes we analyzed were isolated from both environmental and clinical locations (Table S1 from Mead et al., 2021), and based on published literature, we do not anticipate isolate setting to play a large role or confound our results analyzing pathogens and non-pathogens. For example, previous work (Ashu et al., 2017) reported that the ecological niche

of *A. fumigatus* strains (including whether or not they were isolated from the clinic or environment) contributed a very small but statistically significant amount to the overall amounts of observed diversity between 2,026 isolates. This suggests that the environmental and clinical strains used in our study are likely representative of *A. fumigatus* strain diversity. Furthermore, consistent with our results (Figure 2A), Puértolas-Balint et al. (Puértolas-Balint et al., 2019) reported, using their own set of virulence-related genes, that both clinical and environmental strains of *A. fumigatus* have similar "virulence genetic content".

In general, it appears that clinical isolates of *A. fumigatus* are slightly more pathogenic than environmental isolates (Mondon et al., 1996; Alshareef and Robson, 2014), perhaps due to within-host microevolution of clinical isolates, but this issue is still under active investigation in the field. For example, Kowalski et al. (Kowalski et al., 2016) showed that on average, clinical strains are indeed slightly more pathogenic in their sample of *A. fumigatus* clinical and environmental strains, but the difference is relatively small compared to the heterogeneity of pathogenicity observed between *A. fumigatus* strains. We are unaware if the genomic traits and pathogenicity of environmental and clinical strains have been compared in species other than *A. fumigatus*. We are currently designing these experiments with these and other strains.

Evolutionary studies have also been caried out in fungal pathogens outside of the genus *Aspergillus*, and when our results are placed in the context of this literature, a diverse set of mechanisms have driven the evolution of fungal pathogenicity (Taylor, 2015). The ability to infect humans has also evolved multiple times in *Candida* species found within the fungal subphylum Saccharomycotina (Gabaldón et al., 2016); however, gene family expansion and

interspecies hybridization were much more important for the evolution of pathogenicity in that clade (Butler et al., 2009) compared to the results we present here in section *Fumigati* where there was little evidence of dramatic changes in gene family member number between pathogens and non-pathogens (Figure S2 from Mead et al., 2021). Similarly, gene family size was hypothesized to be an important factor in the evolution of *Coccidioides* pathogens, and just as we only saw 84 genes with *A. fumigatus*-specific evolutionary rates, this group of pathogens had a relatively small number of genes with species-specific evolutionary rates (Sharpton et al., 2009). In pathogenic *Cryptococcus* species, mating-type loci and the switch from a tetrapolar to bipolar mating system have been suggested as being key in producing the genomic environment necessary for pathogenicity to evolve (Sun et al., 2019), but in *A. fumigatus*, mating-type loci do not appear to contribute to virulence (Losada et al., 2015) and the contribution of mating across section *Fumigati* has only rarely been studied (Rydholm et al., 2007).

Worldwide mortality rates for aspergillosis infections are estimated to range from as high as 95% to as low as 30%, and drug resistance is a frequent worry for clinicians (Brown et al., 2012). To combat this global health issue, more must be understood about *Aspergillus* biology and evolution. Here, we showed how an evolutionary approach can guide the identification of pathogenicity-associated genetic elements in *Aspergillus* fungi, presented many promising, novel candidates for future study, and placed them within an evolutionary context that will also guide their study with relation to non-*A. fumigatus* pathogenic species found within section *Fumigati*. Our data also provide clues on how *Aspergillus* pathogenicity evolved through the contribution of genetic elements that fit both the conserved pathogenicity and species-specific pathogenicity models. Furthermore, genes that fit the conserved pathogenicity model may be useful as targets

for the treatment of disease caused by all section *Fumigati* species, whereas genes that fit the species-specific pathogenicity model may be useful for species-specific treatment strategies. More generally, this work provides the basis for an evolutionary framework that can inform multiple aspects of the study of both *Aspergillus* species and the diseases they cause.

CHAPTER 4

Variation Among Biosynthetic Gene Clusters, Secondary Metabolite Profiles, and Cards of Virulence Across *Aspergillus* Species[3]

Introduction

Fungal diseases impose a clinical, economic, and social burden on humans (Drgona et al., 2014; Vallabhaneni et al., 2016; Benedict et al., 2019). Fungi from the genus *Aspergillus* are responsible for a considerable fraction of this burden, accounting for more than 250,000 infections annually with high mortality rates (Bongomin et al., 2017). *Aspergillus* infections often result in pulmonary and invasive diseases that are collectively termed aspergillosis. Among *Aspergillus* species, *Aspergillus fumigatus* is the primary etiological agent of aspergillosis (Latgé and Chamilos, 2019).

Even though *A. fumigatus* is a major pathogen, its closest relatives are not considered pathogenic (Mead et al., 2019a; Steenwyk et al., 2019c; Rokas et al., 2020a). Numerous studies have identified genetic determinants that contribute to *A. fumigatus* pathogenicity, such as the organism's ability to grow well at higher temperatures and in hypoxic conditions (Kamei and Watanabe, 2005; Tekaia and Latgé, 2005; Abad et al., 2010; Grahl et al., 2012). Genetic determinants that contribute to pathogenicity could be conceived as analogous to individual "cards" of a "hand" (set of cards) in a card game – that is, individual determinants are typically

[3]This work is published in: Steenwyk, J. L., Mead, M. E., Knowles, S. L., Raja, H. A., Roberts, C. D., Bader, O., et al. (2020). Variation Among Biosynthetic Gene Clusters, Secondary Metabolite Profiles, and Cards of Virulence Across Aspergillus Species. Genetics, genetics.303549.2020. doi:10.1534/genetics.120.303549.

insufficient to cause disease but can collectively do so (Casadevall, 2007).

Aspergillus fumigatus biosynthesizes a cadre of secondary metabolites and several metabolites could be conceived as "cards" of virulence because of their involvement in impairing the host immune system, protecting the fungus from host immune cell attacks, or acquiring key nutrients (Shwab et al., 2007; Losada et al., 2009; Yin et al., 2013; Wiemann et al., 2014; Bignell et al., 2016; Knox et al., 2016; Raffa and Keller, 2019; Blachowicz et al., 2020). For example, the secondary metabolite gliotoxin has been shown in *A. fumigatus* to inhibit the host immune response (Sugui et al., 2007; Spikes et al., 2008). Other secondary metabolites implicated in virulence include: fumitremorgin, which inhibits the activity of the breast cancer resistance protein (González-Lobato et al., 2010); verruculogen, which modulates the electrophysical properties of human nasal epithelial cells (Khoufache et al., 2007); trypacidin, which is cytotoxic to lung cells and inhibits phagocytosis (Gauthier et al., 2012; Mattern et al., 2015); pseurotin, which inhibits immunoglobulin E (Ishikawa et al., 2009); and fumagillin which causes epithelial cell damage (Guruceaga et al., 2018) and impairs the function of neutrophils (Fallon et al., 2010, 2011).

By extension, the metabolic pathways responsible for the biosynthesis of secondary metabolites could also be conceived as components of these secondary metabolism-associated "cards" of virulence. Genes in these pathways are typically organized in contiguous sets termed biosynthetic gene clusters (BGCs) (Keller, 2019). BGCs are known to evolve rapidly, and their composition can differ substantially across species and strains (Lind et al., 2015, 2017; de Vries

et al., 2017; Kjærbølling et al., 2018, 2020; Rokas et al., 2018, 2020b; Vesth et al., 2018). For example, even though *A. fumigatus* contains 33 BGCs and *A. fischeri* contains 48 BGCs, only 10 of those BGCs appear to be shared between the two species (Mead et al., 2019a). Interestingly, one of the BGCs that is conserved between *A. fumigatus* and *A. fischeri* is the gliotoxin BGC and both species have been shown to biosynthesize the secondary metabolite, albeit at different amounts (Knowles et al., 2020). These results suggest that the gliotoxin "card" is part of a winning "hand" that facilitates virulence only in the background of the major pathogen *A. fumigatus* and not in that of the nonpathogen *A. fischeri* (Knowles et al., 2020).

To date, such comparisons of BGCs and secondary metabolite profiles among *A. fumigatus* and closely related nonpathogenic species have been few and restricted to single strains (Mead et al., 2019a; Knowles et al., 2020). However, genetic and phenotypic heterogeneity among strains of a single species is an important consideration when studying *Aspergillus* pathogenicity (Kowalski et al., 2016; Keller, 2017; Kowalski et al., 2019; Ries et al., 2019; Bastos et al., 2020b; Blachowicz et al., 2020; dos Santos et al., 2020b; Drott et al., 2020; Steenwyk et al., 2020c). Examination of multiple strains of *A. fumigatus* and close relatives—including the recently described closest known relative of *A. fumigatus*, *A. oerlinghausenensis*, whose virulence has yet to be examined but which is not thought to be a human pathogen (Houbraken et al., 2016) and has never been associated with human infections—will increase our understanding of the *A. fumigatus* secondary metabolism-associated "cards" of virulence.

To gain insight into the genomic and chemical similarities and differences in secondary metabolism among *A. fumigatus* and nonpathogenic close relatives, we characterized variation in

BGCs and secondary metabolites produced by *A. fumigatus* and nonpathogenic close relatives. To do so, we first sequenced and assembled *A. oerlinghausenensis* CBS 139183^{T} as well as *A. fischeri* strains NRRL 4585 and NRRL 4161 and analyzed them together with four *A. fumigatus* and three additional *A. fischeri* publicly available genomes. We also characterized the secondary metabolite profiles of three *A. fumigatus*, one *A. oerlinghausenensis*, and three *A. fischeri* strains. We observed both variation and conservation among species- and strain-level BGCs and secondary metabolites. We found that the biosynthesis of the secondary metabolites gliotoxin and fumitremorgin, which are both known to interact with mammalian cells (Yamada et al., 2000; González-Lobato et al., 2010; Li et al., 2012; Raffa and Keller, 2019), as well as their BGCs, were conserved among pathogenic and nonpathogenic strains. Interestingly, we found only *A. fischeri* strains, but not *A. fumigatus* strains, biosynthesized verruculogen, which changes the electrophysical properties of human nasal epithelial cells (Khoufache et al., 2007). Similarly, we found that both *A. fumigatus* and *A. oerlinghausenensis* biosynthesized fumagillin and trypacidin, whose effects include broad suppression of the immune response system and lung cell damage (Ishikawa et al., 2009; Fallon et al., 2010, 2011; Gauthier et al., 2012), but *A. fischeri* did not. Taken together, these results reveal that nonpathogenic close relatives of *A. fumigatus* also produce some, but not all, of the secondary metabolism-associated cards of virulence known in *A. fumigatus*. Further investigation of the similarities and differences among *A. fumigatus* and close nonpathogenic relatives may provide additional insight into the “hand of cards” that enabled *A. fumigatus* to evolve into a deadly pathogen.

Materials and Methods

Strain acquisition, DNA extraction, and sequencing

Two strains of *Aspergillus fischeri* (NRRL 4161 and NRRL 4585) were acquired from the Northern Regional Research Laboratory (NRRL) at the National Center for Agricultural Utilization Research in Peoria, Illinois, while one strain of *Aspergillus oerlinghausenensis* (CBS 139183^{T}) was acquired from the Westerdijk Fungal Biodiversity Institute, Utrecht, The Netherlands. These strains were grown in 50 ml of liquid yeast extract soy peptone dextrose (YESD) medium. After approximately seven days of growth on an orbital shaker (100 rpm) at room temperature, the mycelium was harvested by filtering the liquid media through a Corning®, 150 ml bottle top, 0.22µm sterile filter and washed with autoclaved distilled water. All subsequent steps of DNA extraction from the mycelium were performed following protocols outlined previously (Mead et al., 2019b). The genomic DNA from these three strains was sequenced using a NovaSeq S4 at the Vanderbilt Technologies for Advanced Genomes facility (Nashville, Tennessee, US) using paired-end sequencing (150 bp) strategy with the Illumina TruSeq library kit.

Genome assembly, quality assessment, and annotation

To assemble and annotate the three newly sequenced genomes, we first quality-trimmed raw sequence reads using Trimmomatic, v0.36 (Bolger et al., 2014) using parameters described elsewhere (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10, leading:10, trailing:10, slidingwindow:4:20, minlen:50) (Steenwyk and Rokas, 2017). The resulting paired and unpaired quality-trimmed reads were used as input to the SPAdes, v3.11.1 (Bankevich et al., 2012), genome assembly algorithm with the ‘careful’ parameter and the ‘cov-cutoff’ set to ‘auto’.

We evaluated the quality of our newly assembled genomes, using metrics based on continuity of assembly and gene-content completeness. To evaluate genome assemblies by scaffold size, we calculated the N50 of each assembly (or the shortest contig among the longest contigs that account for 50% of the genome assembly's length) (Yandell and Ence, 2012). To determine gene-content completeness, we implemented the BUSCO, v2.0.1 (Waterhouse et al., 2018a), pipeline using the 'genome' mode. In this mode, the BUSCO pipeline examines assembly contigs for the presence of near-universally single copy orthologous genes (hereafter referred to as BUSCO genes) using a predetermined database of orthologous genes from the OrthoDB, v9 (Waterhouse et al., 2013). We used the OrthoDB database for Pezizomycotina (3,156 BUSCO genes). Each BUSCO gene is determined to be present in a single copy, as duplicate sequences, fragmented, or missing. Our analyses indicate the newly sequenced and assembled genomes have high gene-content completeness and assembly continuity (average percent presence of BUSCO genes: 98.80 ± 0.10%; average N50: 451,294.67 ± 9,696.11; Fig. S1 from Steenwyk et al., 2020d). These metrics suggest these genomes are suitable for comparative genomic analyses.

To predict gene boundaries in the three newly sequenced genomes, we used the MAKER, v2.31.10, pipeline (Holt and Yandell, 2011) which, creates consensus predictions from the collective evidence of multiple *ab initio* gene prediction software. Specifically, we created consensus predictions from SNAP, v2006-07-28 (Korf, 2004), and AUGUSTUS, v3.3.2 (Stanke and Waack, 2003), after training each algorithm individually on each genome. To do so, we first ran MAKER using protein evidence clues from five different publicly available annotations of *Aspergillus* fungi from section *Fumigati*. Specifically, we used protein homology clues from *A.*

fischeri NRRL 181 (GenBank accession: GCA_000149645.2), *A. fumigatus* Af293 (GenBank accession: GCA_000002655.1), *Aspergillus lentulus* IFM 54703 (GenBank accession: GCA_001445615.1), *Aspergillus novofumigatus* IBT 16806 (GenBank accession: GCA_002847465.1), and *Aspergillus udagawae* IFM 46973 (GenBank accession: GCA_001078395.1). The resulting gene predictions were used to train SNAP. MAKER was then rerun using the resulting training results. Using the SNAP trained gene predictions, we trained AUGUSTUS. A final set of gene boundary predictions were obtained by rerunning MAKER with the training results from both SNAP and AUGUSTUS.

To supplement our data set of newly sequenced genomes, we obtained publicly available ones. Specifically, we obtained genomes and annotations for *A. fumigatus* Af293 (GenBank accession: GCA_000002655.1), *A. fumigatus* CEA10 (strain synonym: CBS 144.89 / FGSC A1163; GenBank accession: GCA_000150145.1), *A. fumigatus* HMR AF 270 GenBank accession: GCA_002234955.1), *A. fumigatus* Z5 (GenBank accession: GCA_001029325.1), *A. fischeri* NRRL 181 (GenBank accession: GCA_000149645.2). We also obtained assemblies of the recently published *A. fischeri* genomes for strains IBT 3003 and IBT 3007 (Zhao et al., 2019b) which, lacked annotations. We annotated the genome of each strain individually using MAKER with the SNAP and AUGUSTUS training results from a close relative of both strains, *A. fischeri* NRRL 4161. Altogether, our final data set contained a total of ten genome from three species: four *A. fumigatus* strains, one *A. oerlinghausenensis* strain, and five *A. fischeri* strains (Table 2).

Table 2. Species and strains used in the present study.

Genus and species	Strain	Environmental/ Clinical	Genomic analysis	Secondary metabolite profiling	Reference
Aspergillus oerlinghausenensis	CBS 139183^{T}	Environmental	+	+	This study
Aspergillus fischeri	NRRL 4585	Environmental	+	+	This study
Aspergillus fischeri	NRRL 4161	Unknown	+	+	This study
Aspergillus fischeri	NRRL 181	Environmental	+	+	(Fedorova et al., 2008)
Aspergillus fischeri	IBT 3007	Environmental	+	-	(Zhao et al., 2019b)
Aspergillus fischeri	IBT 3003	Environmental	+	-	(Zhao et al., 2019b)
Aspergillus fumigatus	Af293	Clinical	+	+	(Nierman et al., 2005)
Aspergillus fumigatus	CEA10 / CEA17	Clinical	+	+	(Fedorova et al., 2008)
Aspergillus fumigatus	HMR AF 270	Clinical	+	-	BioSample: SAMN07177964
Aspergillus fumigatus	Z5	Environmental	+	-	(Miao et al., 2015)

'+' and '-' indicate if BGCs and secondary metabolite profiling was conducted on a particular strain. More specifically '+' indicates the strain was analyzed whereas '-' indicates that the strain was not analyzed.

Maximum likelihood phylogenetics and Bayesian estimation of divergence times

To reconstruct the evolutionary history among the ten *Aspergillus* genomes, we implemented a recently developed pipeline (Steenwyk et al., 2019c), which relies on the concatenation-approach to phylogenomics (Rokas et al., 2003) and has been successfully used in reconstructing species-level relationships among *Aspergillus* and *Penicillium* fungi (Bodinaku et al., 2019; Steenwyk et al., 2019c). The first step in the pipeline is to identify single copy orthologous genes in the

genomes of interest which, are ultimately concatenated into a larger phylogenomic data matrix. To identify single copy BUSCO genes across all ten *Aspergillus* genomes, we used the BUSCO pipeline with the Pezizomycotina database as described above. We identified 3,041 BUSCO genes present at a single copy in all ten *Aspergillus* genomes and created multi-FASTA files for each BUSCO gene that contained the protein sequences for all ten taxa. The protein sequences of each BUSCO gene were individually aligned using Mafft, v7.4.02 (Katoh and Standley, 2013), with the same parameters as described elsewhere (Steenwyk et al., 2019c). Nucleotide sequences were then mapped onto the protein sequence alignments using a custom Python, v3.5.2 (https://www.python.org/), script with BioPython, v1.7 (Cock et al., 2009a). The resulting codon-based alignments were trimmed using trimAl, v1.2.rev59 (Capella-Gutierrez et al., 2009), with the 'gappyout' parameter. The resulting trimmed nucleotide alignments were concatenated into a single matrix of 5,602,272 sites and was used as input into IQ-TREE, v1.6.11 (Nguyen et al., 2015). The best-fitting model of substitutions for the entire matrix was determined using Bayesian information criterion values (Kalyaanamoorthy et al., 2017). The best-fitting model was a general time-reversible model with empirical base frequencies that allowed for a proportion of invariable sites and a discrete Gamma model with four rate categories (GTR+I+F+G4) (Tavaré, 1986; Yang, 1994, 1996; Vinet and Zhedanov, 2011). To evaluate bipartition support, we used 5,000 ultrafast bootstrap approximations (Hoang et al., 2018).

To estimate divergence times among the ten *Aspergillus* genomes, we used the concatenated data matrix and the resulting maximum likelihood phylogeny from the previous steps as input to Bayesian approach implemented in MCMCTree from the PAML package, v4.9d (Yang, 2007). First, we estimated the substitution rate across the data matrix using a "GTR+G" model of

substitutions (model = 7), a strict clock model, and the maximum likelihood phylogeny rooted on the clade of *A. fischeri* strains. We imposed a root age of 3.69 million years ago according to results from recent divergence time estimates of the split between *A. fischeri* and *A. fumigatus* (Steenwyk et al., 2019c). We estimated the substitution rate to be 0.005 substitutions per one million years. Next, the likelihood of the alignment was approximated using a gradient and Hessian matrix. To do so, we used previously established time constraints for the split between *A. fischeri* and *A. fumigatus* (1.85 to 6.74 million years ago) (Steenwyk et al., 2019c). Lastly, we used the resulting gradient and Hessian matrix, the rooted maximum likelihood phylogeny, and the concatenated data matrix to estimate divergence times using a relaxed molecular clock (model = 2). We specified the substitution rate prior based on the estimated substitution rate (rgene_gamma = 1 186.63). The 'sigma2_gamma' and 'finetune' parameters were set to '1 4.5' and '1', respectively. To collect a high-quality posterior probability distribution, we ran a total of 5.1 million iterations during MCMC analysis which, is 510 times greater than the minimum recommendations (Raftery and Lewis, 1995). Our sampling strategy across the 5.1 million iterations was to discard the first 100,000 results followed by collecting a sample every 500th iteration until a total of 10,000 samples were collected.

Identification of gene families and analyses of putative biosynthetic gene clusters

To identify gene families across the ten *Aspergillus* genomes, we used a Markov clustering approach. Specifically, we used OrthoFinder, v2.3.8 (Emms and Kelly, 2019). OrthoFinder first conducts a blast all-vs-all using the protein sequences of all ten *Aspergillus* genomes and NCBI's Blast+, v2.3.0 (Camacho et al., 2009), software. After normalizing blast bit scores, genes are clustered into discrete orthogroups using a Markov clustering approach (van Dongen, 2000). We

clustered genes using an inflation parameter of 1.5. The resulting orthogroups were used proxies for gene families.

To identify putative biosynthetic gene clusters (BGCs), we used the gene boundaries predictions from the MAKER software as input into antiSMASH, v4.1.0 (Weber et al., 2015). To identify homologous BGCs across the ten *Aspergillus* genomes, we used the software BiG-SCAPE, v20181005 (Navarro-Muñoz et al., 2020). Based on the Jaccard Index of domain types, sequence similarity among domains, and domain adjacency, BiG-SCAPE calculates a similarity metric between pairwise combinations of clusters where smaller values indicate greater BGC similarity. BiG-SCAPE's similarity metric can then be used as an edge-length in network analyses of cluster similarity. We evaluated networks using an edge-length cutoff from 0.1-0.9 with a step of 0.1 (Fig. S3 from Steenwyk et al., 2020d). We found networks with an edge-length cutoff of 0.4-0.6 to be similar and based further analyses on a cutoff of 0.5. Because BiG-SCAPE inexplicably split the gliotoxin BGC of the *A. fumigatus* Af293 strain into two cluster families even though the BGC was highly similar to the gliotoxin BGCs of all other strains, we supplemented BiG-SCAPE's approach to identifying homologous BGCs with visualize inspection of microsyteny and blast-based analyses using NCBI's BLAST+, v2.3.0 (Camacho et al., 2009) for BGCs of interest. Similar sequences in microsynteny analyses were defined as at least 100 bp in length, at least 30 percent similarity, and an expectation value threshold of 0.01. Lastly, to determine if any BGCs have been previously linked to secondary metabolites, we cross referenced BGCs and BGC families with those found in the MIBiG database (Kautsar et al., 2019) as well as previously published *A. fumigatus* BGCs (Table S2 from Steenwyk et al., 2020d). BGCs not

associated with secondary metabolites were considered to likely encode for unknown compounds.

Identification and characterization of secondary metabolite production

General experimental procedures

The ^{1}H NMR data were collected using a JOEL ECS-400 spectrometer, which was equipped with a JOEL normal geometry broadband Royal probe, and a 24-slot autosampler, and operated at 400 MHz. HRESIMS experiments utilized either a Thermo LTQ Orbitrap XL mass spectrometer or a Thermo Q Exactive Plus (Thermo Fisher Scientific); both were equipped with an electrospray ionization source. A Waters Acquity UPLC (Waters Corp.) was utilized for both mass spectrometers, using a BEH C_{18} column (1.7 μm; 50 mm x 2.1 mm) set to a temperature of 40^{o}C and a flow rate of 0.3 ml/min. The mobile phase consisted of a linear gradient of CH_3CN-H_2O (both acidified with 0.1% formic acid), starting at 15% CH_3CN and increasing linearly to 100% CH_3CN over 8 min, with a 1.5 min hold before returning to the starting condition. The HPLC separations were performed with Atlantis T3 C_{18} semi-preparative (5 μm; 10 x 250 mm) and preparative (5 μm; 19 x 250 mm) columns, at a flow rate of 4.6 ml/min and 16.9 ml/min, respectively, with a Varian Prostar HPLC system equipped with a Prostar 210 pumps and a Prostar 335 photodiode array detector (PDA), with the collection and analysis of data using Galaxie Chromatography Workstation software. Flash chromatography was performed on a Teledyne ISCO Combiflash Rf 200 and monitored by both ELSD and PDA detectors.

Chemical characterization

To identify the secondary metabolites that were biosynthesized by *A. fumigatus*, *A.*

oerlinghausenensis, and *A. fischeri*, these strains were grown as large-scale fermentations to isolate and characterize the secondary metabolites. To inoculate oatmeal cereal media (Old fashioned breakfast Quaker oats), agar plugs from fungal stains grown on potato dextrose agar; difco (PDA) were excised from the edge of the Petri dish culture and transferred to separate liquid seed media that contained 10 ml YESD broth (2% soy peptone, 2% dextrose, and 1% yeast extract; 5 g of yeast extract, 10 g of soy peptone, and 10 g of D-glucose in 500 ml of deionized H_2O) and allowed to grow at 23˚C with agitation at 100 rpm for three days. The YESD seed cultures of the fungi were subsequently used to inoculate solid-state oatmeal fermentation cultures, which were either grown at room temperature (approximately 23˚C under 12h light/dark cycles for 14 days), 30˚C, or 37˚C; all growths at the latter two temperatures were carried out in an incubator (VWR International) in the dark over four days. The oatmeal cultures were prepared in 250 ml Erlenmeyer flasks that contained 10 g of autoclaved oatmeal (10 g of oatmeal with 17 ml of deionized H_2O and sterilized for 15–20 minutes at 121°C). For all fungal strains three flasks of oatmeal cultures were grown at all three temperatures, except for *A. oerlinghausenensis* (CBS 139183^T) at room temperature and *A. fumigatus* (Af293) at 37˚C. For CBS 139183^T, the fungal cultures were grown in four flasks, while for Af293 eight flasks were grown in total. The growths of these two strains were performed differently from the rest because larger amounts of extract were required in order to perform detailed chemical characterization.

The cultures were extracted by adding 60 ml of (1:1) MeOH-$CHCl_3$ to each 250 ml flask, chopping thoroughly with a spatula, and shaking overnight (~ 16 h) at ~ 100 rpm at room temperature. The culture was filtered *in vacuo*, and 90 ml $CHCl_3$ and 150 ml H_2O were added to the filtrate. The mixture was stirred for 30 min and then transferred to a separatory funnel. The

organic layer ($CHCl_3$) was drawn off and evaporated to dryness *in vacuo*. The dried organic layer was reconstituted in 100 ml of (1:1) MeOH–CH_3CN and 100 ml of hexanes, transferred to a separatory funnel, and shaken vigorously. The defatted organic layer (MeOH–CH_3CN) was evaporated to dryness *in vacuo*.

To isolate compounds, the defatted extract was dissolved in $CHCl_3$, absorbed onto Celite 545 (Acros Organics), and fractioned by normal phase flash chromatography using a gradient of hexane-$CHCl_3$-MeOH. *Aspergillus fischeri* strain NRRL 181 was chemically characterized previously (Knowles et al., 2019; Mead et al., 2019a). *A. fumigatus* strain Af293, grown at 37°C, was subjected to a 12g column at a flow rate of 30 ml/min and 61.0 column volumes, which yielded four fractions. Fraction 2 was further purified via preparative HPLC using a gradient system of 30:70 to 100:0 of CH_3CN-H_2O with 0.1% formic acid over 40 min at a flow rate of 16.9 ml/min to yield six subfractions. Subfractions 1, 2 and 5, yielded cyclo(L-Pro-L-Leu) (Li et al., 2008) (0.89 mg), cyclo(L-Pro-L-Phe) (Campbell et al., 2009) (0.71 mg), and monomethylsulochrin (Ma et al., 2004) (2.04 mg), which eluted at approximately 5.7, 6.3, and 10.7 min, respectively. Fraction 3 was further purified via preparative HPLC using a gradient system of 40:60 to 65:35 of CH_3CN-H_2O with 0.1% formic acid over 30 min at a flow rate of 16.9 ml/min to yield four subfractions. Subfractions 1 and 2 yielded pseurotin A (Wang et al., 2011) (12.50 mg) and bisdethiobis(methylthio)gliotoxin (Afiyatullov et al., 2005) (13.99 mg), which eluted at approximately 7.5 and 8.0 min, respectively.

A. fumigatus strain CEA10, grown at 37°C, was subjected to a 4g column at a flow rate of 18 ml/min and 90.0 column volumes, which yielded five fractions. Fraction 1 was purified via

preparative HPLC using a gradient system of 50:50 to 100:0 of CH_3CN-H_2O with 0.1% formic acid over 45 min at a flow rate of 16.9 ml/min to yield eight subfractions. Subfraction 1, yielded fumagillin (Halász et al., 2000) (1.69 mg), which eluted at approximately 18.5 min. Fraction 2 was purified via semi-preparative HPLC using a gradient system of 35:65 to 80:20 of CH_3CN-H_2O with 0.1% formic acid over 30 min at a flow rate of 4.6 ml/min to yield 10 subfractions. Subfraction 5 yielded fumitremorgin C (Kato et al., 2009) (0.25 mg), which eluted at approximately 15.5 min. Fraction 3 was purified via preparative HPLC using a gradient system of 40:60 to 100:0 of CH_3CN-H_2O with 0.1% formic acid over 30 min at a flow rate of 16.9 ml/min to yield nine subfractions. Subfraction 2 yielded pseurotin A (1.64 mg), which eluted at approximately 7.3 min.

Aspergillus oerlinghausenensis strain CBS 139183^T, grown at RT, was subjected to a 4g column at a flow rate of 18 ml/min and 90 column volumes, which yielded 4 fractions. Fraction 3 was further purified via preparative HPLC using a gradient system of 35:65 to 70:30 of CH_3CN-H_2O with 0.1% formic acid over 40 min at a flow rate of 16.9 ml/min to yield 11 subfractions. Subfractions 3 and 10 yielded spiro [5H,10H-dipyrrolo[1,2-a:1′,2′-d]pyrazine-2-(3H),2′-[2H]indole]-3′,5,10(1′H)-trione (Wang et al., 2008) (0.64 mg) and helvolic acid (Zhao et al., 2010) (1.03 mg), which eluted at approximately 11.5 and 39.3 min, respectively. (see NMR supporting information; figshare: 10.6084/m9.figshare.12055503).

Metabolite profiling by mass spectrometry

The metabolite profiling by mass spectrometry, also known as dereplication, was performed as stated previously (El-Elimat et al., 2013). Briefly, ultraperformance liquid chromatography-

photodiode array-electrospray ionization high resolution tandem mass spectrometry (UPLC-PDA-HRMS-MS/MS) was utilized to monitor for secondary metabolites across all strains (Af293, CEA10, CEA17, CBS 139183^{T}, NRRL 181, NRRL 4161, and NRRL 4585). Utilizing positive-ionization mode, ACD MS Manager with add-in software IntelliXtract (Advanced Chemistry Development, Inc.; Toronto, Canada) was used for the primary analysis of the UPLC-MS chromatograms. The data from 19 secondary metabolites are provided in the Supporting Information (see Dereplication table; figshare: 10.6084/m9.figshare.12055503), which for each secondary metabolite lists: molecular formula, retention time, UV-absorption maxima, high-resolution full-scan mass spectra, and MS-MS data (top 10 most intense peaks).

Metabolomics analyses

Principal component analysis (PCA) analysis was performed on the UPLC-MS data. Untargeted UPLC-MS datasets for each sample were individually aligned, filtered, and analyzed using MZmine 2.20 software (https://sourceforge.net/projects/mzmine/) (Pluskal et al., 2010). Peak detection was achieved using the following parameters, *A. fumigatus* at (Af293, CEA10, and CEA17): noise level (absolute value), 1×10^{6}; minimum peak duration, 0.05 min; *m/z* variation tolerance, 0.05; and *m/z* intensity variation, 20%; *A. fischeri* (NRRL 181, NRRL 4161, and NRRL 4585): noise level (absolute value), 1×10^{6}; minimum peak duration, 0.05 min; *m/z* variation tolerance, 0.05; and *m/z* intensity variation, 20%; and all strains (Af293, CEA10, CEA17, CBS 139183^{T}, NRRL 181, NRRL 4161, and NRRL 4585): noise level (absolute value), 7×10^{5}; minimum peak duration, 0.05 min; *m/z* variation tolerance, 0.05; and *m/z* intensity variation, 20%. Peak list filtering and retention time alignment algorithms were used to refine peak detection. The join algorithm integrated all sample profiles into a data matrix using the

following parameters: *m/z* and retention time balance set at 10.0 each, *m/z* tolerance set at 0.001, and RT tolerance set at 0.5 mins. The resulting data matrix was exported to Excel (Microsoft) for analysis as a set of *m/z* – retention time pairs with individual peak areas detected in triplicate analyses. Samples that did not possess detectable quantities of a given marker ion were assigned a peak area of zero to maintain the same number of variables for all sample sets. Ions that did not elute between 2 and 8 minutes and/or had an *m/z* ratio less than 200 or greater than 800 Da were removed from analysis. Relative standard deviation was used to understand the quantity of variance between the technical replicate injections, which may differ slightly based on instrument variance. A cutoff of 1.0 was used at any given *m/z* – retention time pair across the technical replicate injections of one biological replicate, and if the variance was greater than the cutoff, it was assigned a peak area of zero. Final chemometric analysis, data filtering (Caesar et al., 2018) and PCA was conducted using Sirius, v10.0 (Pattern Recognition Systems AS) (Kvalheim et al., 2011), and dendrograms were created with Python. The PCA scores plots were generated using data from either the three individual biological replicates or the averaged biological replicates of the fermentations. Each biological replicate was plotted using averaged peak areas obtained across four replicate injections (technical replicates).

Data Availability

Sequence reads and associated genome assemblies generated in this project are available in NCBI's GenBank database under the BioProject PRJNA577646. Additional descriptions of the genomes including predicted gene boundaries are available through Figshare (doi: 10.6084/m9.figshare.12055503). The Figshare repository is also populated with other data generated from genomic and natural products analysis. Among genomic analyses, we provide

information about predicted BGCs, results associated with network-based clustering of BGCs into cluster families, phylogenomic data matrices, and trees. Among natural products analysis, we provide information that supports methods and results, including NMR spectra.

Results

Conservation and diversity of biosynthetic gene clusters within and between species

We sequenced and assembled *A. oerlinghausenensis* CBS 139183^{T} and *A. fischeri* strains NRRL 4585 and NRRL 4161. Together with publicly available genomes, we analyzed 10 *Aspergillus* genomes (five *A. fischeri* strains; four *A. fumigatus* strains; one *A. oerlinghausenensis* strain; see Methods). We found that the newly added genomes were of similar quality to other publicly available draft genomes (average percent presence of BUSCO genes: 98.80 ± 0.10%; average N50: 451,294.67 ± 9,696.11; Fig. S1 from Steenwyk et al., 2020d). We predicted that *A. oerlinghausenensis* CBS 139183^{T}, *A. fischeri* NRRL 4585, and *A. fischeri* NRRL 4161 have 10,044, 11,152 and 10,940 genes, respectively, numbers similar to publicly available genomes. Lastly, we inferred the evolutionary history of the 10 *Aspergillus* genomes using a concatenated matrix of 3,041 genes (5,602,272 sites) and recapitulated species-level relationships as previously reported (Houbraken et al., 2016). Relaxed molecular clock analyses suggested that *A. oerlinghausenensis* CBS 139183^{T} diverged from *A. fumigatus* approximately 3.9 (6.4 – 1.3) million years ago and that *A. oerlinghausenensis* and *A. fumigatus* split from *A. fischeri* approximately 4.5 (6.8 – 1.7) million years ago (Fig. 12A; Fig. S2 from Steenwyk et al., 2020d).

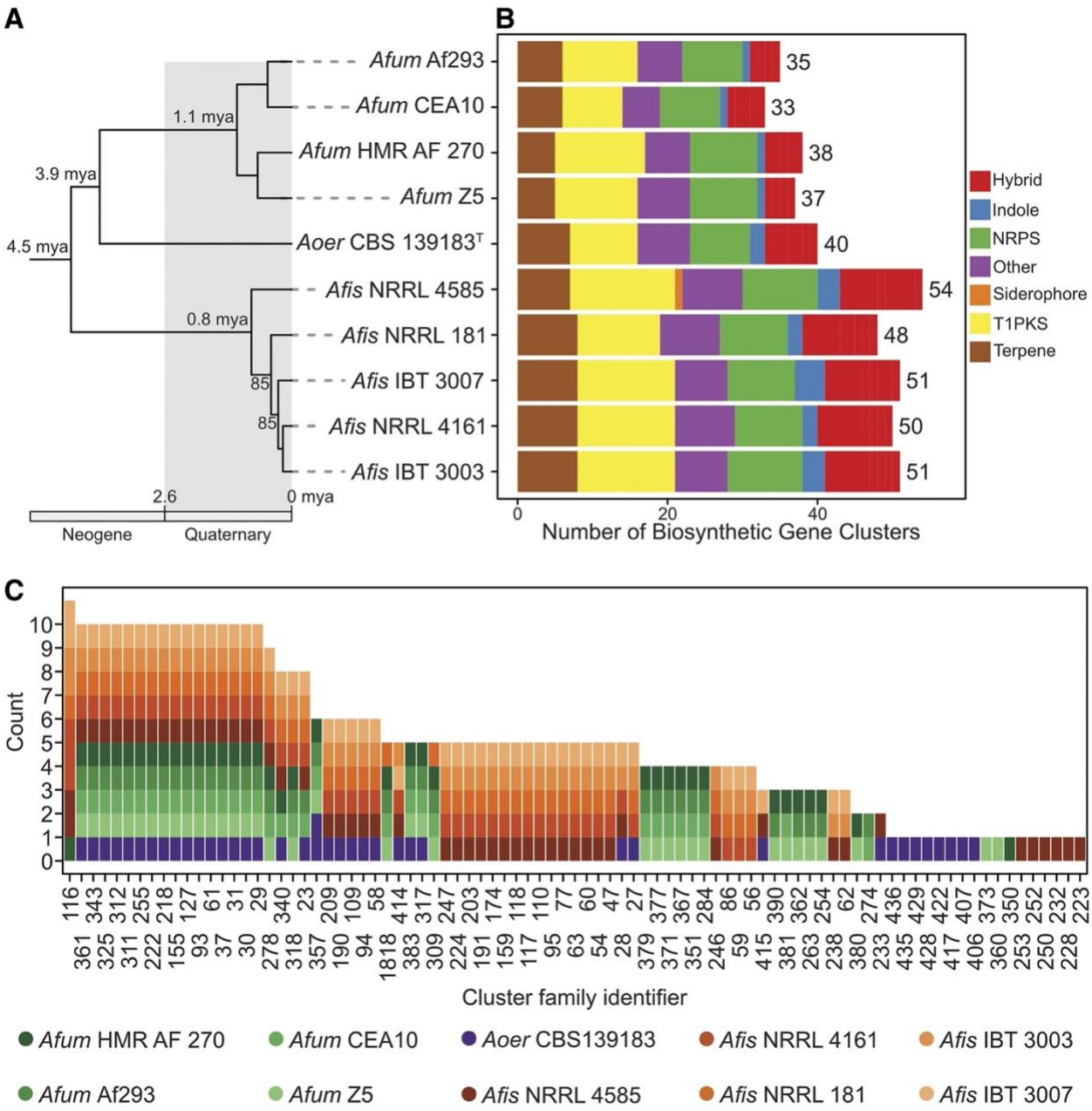

Figure 12. Diverse genetic repertoire of biosynthetic gene clusters and extensive presence and absence polymorphisms between and within species.

(A) Genome-scale phylogenomic analysis confirms *A. oerlinghausenensis* is the closest relative to *A. fumigatus*. Relaxed molecular clock analyses suggest *A. fumigatus*, *A. oerlinghausenensis*, and *A. fischeri* diverged from one another during the Neogene geologic period. Bipartition support is depicted for internodes that did not have full support. (B) *A. fumigatus* harbors the lowest number of BGCs compared to its two closest relatives. (C) Network-based clustering of BGCs into cluster families reveal extensive cluster presence and absence polymorphisms between species and strains. Cluster family identifiers are depicted on the *x*-axis; the number of strains represented in a cluster family are shown on the *y*-axis; the colors refer to a single strain from each species. Genus and species names are written using the following abbreviations: *Afum*, *A. fumigatus*; *Aoer*, *A. oerlinghausenensis*; *Afis*, *A. fischeri*. Classes of

BGCs are written using the following abbreviations: NRPS, nonribosomal peptide synthetase; T1PKS, type I polyketide synthase; Hybrid, a combination of multiple BGC classes.

Examination of the total number of predicted BGCs revealed that *A. fischeri* has the largest BGC count. Among *A. fumigatus*, *A. oerlinghausenensis*, and *A. fischeri*, we predicted an average of 35.75 ± 2.22, 40, 50.80 ± 2.17 BGCs, respectively, and found they spanned diverse biosynthetic classes (e.g., polyketides, non-ribosomal peptides, terpenes, etc.) (Fig. 12B). Network-based clustering of BGCs into cluster families (or groups of homologous BGCs) resulted in qualitatively similar networks when we used moderate similarity thresholds (or edge cut-off values; Fig. S3A from Steenwyk et al., 2020d). Using a (moderate) similarity threshold of 0.5, we inferred 88 cluster families of putatively homologous BGCs (Fig. 12C).

Examination of BGCs revealed extensive presence and absence polymorphisms within and between species. We identified 17 BGCs that were present in all 10 *Aspergillus* genomes including the hexadehydroastechrome (HAS) BGC (cluster family 311 or CF311), the neosartoricin BGC (CF61), and other putative BGCs likely encoding unknown products (Fig. S3B from Steenwyk et al., 2020d; Table S1 from Steenwyk et al., 2020d; data available from figshare, doi: 10.6084/m9.figshare.12055503). In contrast, we identified 18 BGCs found in single strains, which likely encode unknown products. Between species, similar patterns of broadly present and species-specific BGCs were observed. For example, we identified 18 BGCs that were present in at least one strain across all species; in contrast, *A. fumigatus*, *A. oerlinghausenensis*, and *A. fischeri* had 16, 8, and 27 BGCs present in at least one strain but

absent from the other species, respectively. These results suggest each species has a largely distinct repertoire of BGCs.

Examination of shared BGCs across species revealed *A. oerlinghausenensis* CBS139183^{T} and *A. fischeri* shared more BGCs with each other than either did with *A. fumigatus*. Surprisingly, we found ten homologous BGCs between *A. oerlinghausenensis* CBS 139183^{T} and *A. fischeri* but only three homologous BGCs shared between *A. fumigatus* and *A. oerlinghausenensis* CBS 139183^{T} (Fig. 13A; Fig. S3C) even though *A. oerlinghausenensis* is more closely related to *A. fumigatus* than to *A. fischeri* (Fig. 12A). BGCs shared by *A. oerlinghausenensis* CBS 139183^{T} and *A. fischeri* were uncharacterized while BGCs present in both *A. fumigatus* and *A. oerlinghausenensis* CBS 139183^{T} included those that encode fumigaclavine and fumagillin/pseurotin. Lastly, to associate each BGC with a secondary metabolite in *A. fumigatus* Af293, we cross referenced our list with a publicly available one (Table S2 from Steenwyk et al., 2020d) (Lind et al., 2017). Importantly, all known *A. fumigatus* Af293 BGCs were represented in our analyses.

At the level of gene families, there were few species-specific gene families in *A. oerlinghausenensis* (Fig. 2B). *A. oerlinghausenensis* CBS 139183^{T} has only eight species-specific gene families, whereas *A. fischeri* and *A. fumigatus* have 1,487 and 548 species-specific

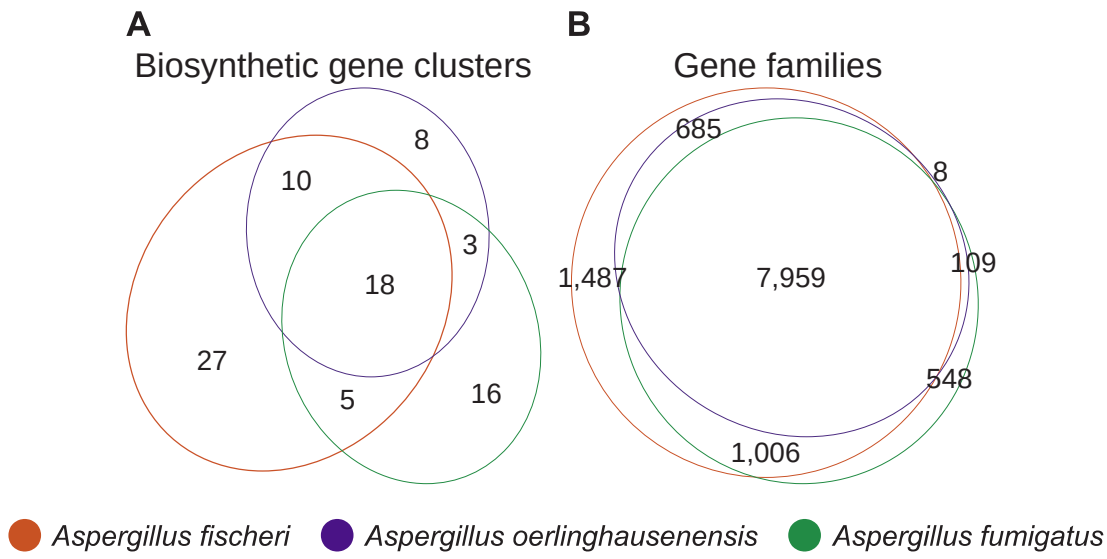

Figure 13. *Aspergillus oerlinghausenensis* shares more gene families and BGCs with *A. fischeri* than *A. fumigatus*.
(A) Euler diagram showing species-level shared BGCs. (B) Euler diagram showing species-level shared gene families. In both diagrams, *A. oerlinghausenensis* shares more gene families or BGCs with *A. fischeri* than *A. fumigatus* despite a closer evolutionary relationship. The Euler diagrams show the results for the species-level comparisons, which may be influenced by the unequal numbers of strains used for the three species; strain-level comparisons of BGCs and gene families can be found in Figures 12C and S4 from Steenwyk et al., 2020d, respectively.

---

gene families, respectively. Examination of the best BLAST hits of the eight species-specific gene families suggest that most are hypothetical or uncharacterized fungal genes. To determine if the eight *A. oerlinghausenensis* CBS 139183^{T} specific gene families were an artifact of using a single representative strain, we conducted and additional ortholog clustering analysis using a single strain of *A. fischeri* (NRRL 181), a single strain of *A. fumigatus* (Af293), or a single strain of each species (CBS 139183, NRRL 181, Af293). When using a single strain of *A. fischeri* or *A. fumigatus*, there were 23 or six gene families unique to each species, respectively. Therefore, the low number of *A. oerlinghausenensis*-specific gene families likely stems from our use of the genome of a single strain.

Despite a closer evolutionary relationship between *A. oerlinghausenensis* and *A. fumigatus*, we found *A. oerlinghausenensis* shares more gene families with *A. fischeri* than with *A. fumigatus* (685 and 109, respectively) suggestive of extensive gene loss in the *A. fumigatus* stem lineage. Lastly, we observed strain heterogeneity in gene family presence and absence within both *A. fumigatus* and *A. fischeri* (Fig. S4 from Steenwyk et al., 2020d). For example, the largest intersection that does not include all *A. fischeri* strains is 493 gene families, which were found in all but one strain, NRRL 181. For *A. fumigatus*, the largest intersection that does not include all strains is 233 gene families, which were shared by strains Af293 and CEA10.

Within and between species variation in secondary metabolite profiles of *A. fumigatus* and its closest relatives

To gain insight into variation in secondary metabolite profiles within and between species, we profiled *A. fumigatus* strains Af293, CEA10, and CEA17 (a *pyrG1/URA3* derivative of CEA10), *A. fischeri* strains NRRL 181, NRRL 4585, and NRRL 4161, and *A. oerlinghausenensis* CBS 139183^{T} for secondary metabolites. Specifically, we used three different procedures, including the isolation and structure elucidation of metabolites, where possible, followed by two different metabolite profiling procedures that use mass spectrometry techniques. Altogether, we isolated and characterized 19 secondary metabolites; seven from *A. fumigatus*, two from *A. oerlinghausenensis*, and ten from *A. fischeri* (Fig. S5 from Steenwyk et al., 2020d). These products encompassed a wide diversity of secondary metabolite classes, such as those derived from polyketide synthases, non-ribosomal peptide-synthetases, terpene synthases and mixed biosynthesis enzymes.

To characterize the secondary metabolites biosynthesized that were not produced in high enough quantity for structural identification through traditional isolation methods, we employed "dereplication" mass spectrometry protocols specific to natural products research on all tested strains at both 30°C and 37°C (see supporting information, dereplication example; figshare: 10.6084/m9.figshare.12055503) (El-Elimat et al., 2013; Ito and Masubuchi, 2014; Gaudêncio and Pereira, 2015; Hubert et al., 2017). We found that most secondary metabolites were present across strains of the same species (Table S3 from Steenwyk et al., 2020d); for example, monomethylsulochrin was isolated from *A. fumigatus* Af293, but through metabolite profiling, its spectral features were noted also in *A. fumigatus* strains CEA10 and CEA17. We identified metabolites that were biosynthesized by only one species; for example, pseurotin A was solely present in *A. fumigatus* strains. Finally, we found several secondary metabolites that were biosynthesized across species, such as fumagillin, which was biosynthesized by *A. fumigatus* and *A. oerlinghausenensis*, and fumitremorgin B, which was biosynthesized by strains of both *A. oerlinghausenensis* and *A. fischeri.* Together, these analyses suggest that closely related *Aspergillus* species and strains exhibit variation both within as well as between species in the secondary metabolites produced.

To further facilitate comparisons of secondary metabolite profiles within and between species, we used the 1,920 features (i.e., unique *m/z* – retention time pairs) that were identified from all strains at all temperatures (Fig. 14A), to perform hierarchical clustering (Fig. 14B) and Principal Components Analysis (PCA) (Fig. S6 from Steenwyk et al., 2020d). Hierarchical clustering at 37°C and 30°C indicated the chromatogram of *A. oerlinghausenensis* CBS 139183^{T} is more similar to the chromatogram of *A. fischeri* than to that of *A. fumigatus*. PCA results were broadly

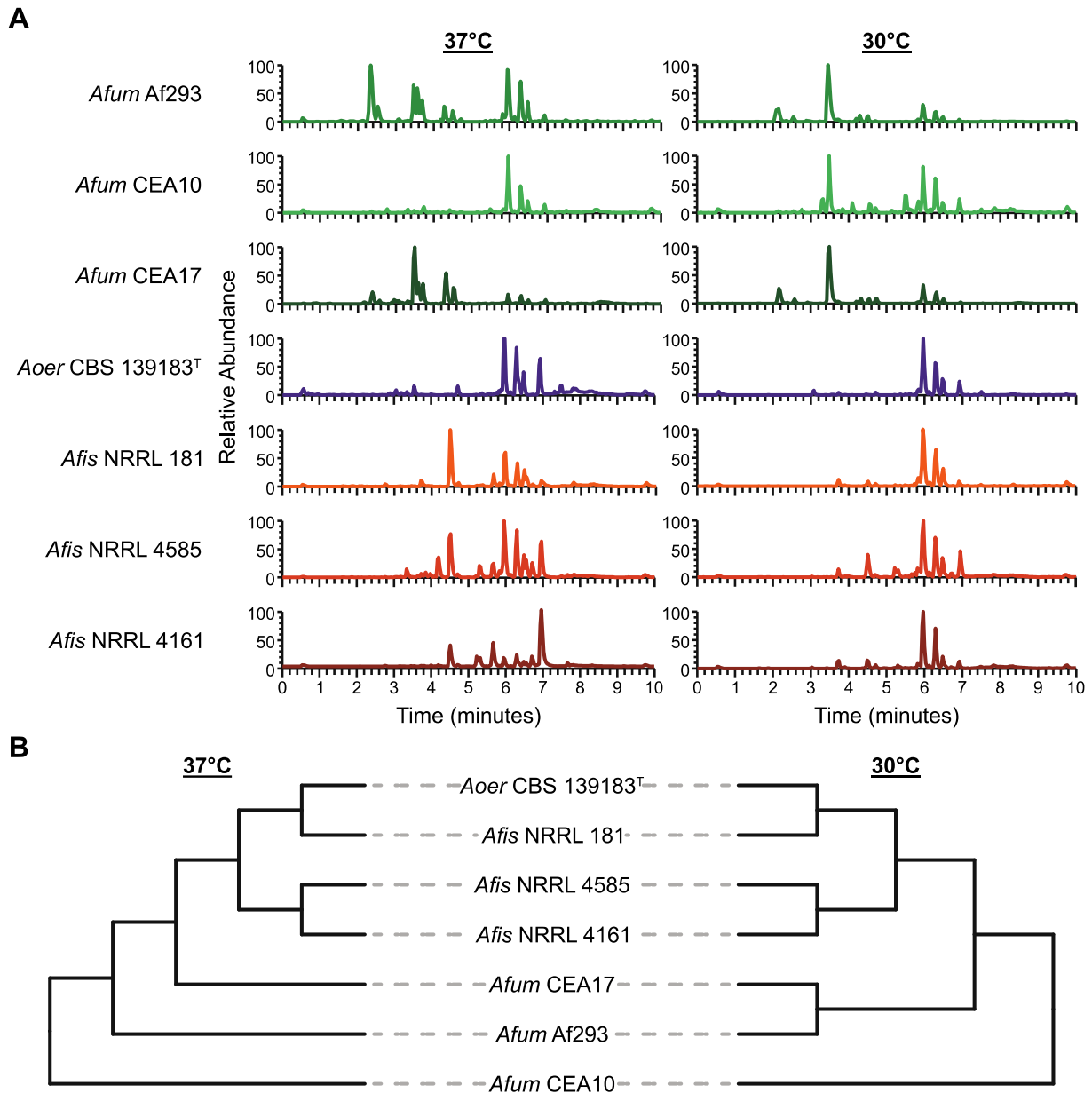

Figure 14. *A. oerlinghausenensis* and *A. fischeri* have more similar secondary metabolite profiles than *A. fumigatus*.
(A) UPLC-MS chromatograms of secondary metabolite profiles of *A. fumigatus* and its closest relatives, *A. oerlinghausenensis* and *A. fischeri* at 37°C and 30°C (left and right, respectively). (B) Hierarchical clustering of chromatograms (1,920 total features) reveals *A. oerlinghausenensis* clusters with *A. fischeri* and not its closest relative, *A. fumigatus* at 37°C and 30°C (left and right, respectively).

consistent with the clustering results, but suggested that *A. oerlinghausenensis* was just as

similar to *A. fischeri* strains as it was to *A. fumigatus* strains. This difference likely stems from the fact that hierarchical clustering is a total-evidence approach whereas PCA captures most but not all variance in the data (e.g., the two principal components in Fig. S6B and S6C from Steenwyk et al., 2020d capture 84.6% of the total variance). PCA analysis revealed greater variation in secondary metabolite production at 30°C compared to 37°C (Fig. S6 from Steenwyk et al., 2020d), suggesting there is a more varied response in how BGCs are being utilized at 30°C. PCA at both 37°C and 30°C showed that variation between *A. oerlinghausenensis* CBS 139183^{T} and *A. fischeri* strains was largely captured along the second principal component; in contrast, the differences between *A. oerlinghausenensis* CBS 139183^{T} and *A. fumigatus* strains are captured along the first principal component (Fig. S6D-E from Steenwyk et al., 2020d). Taken together, these results suggest that the three *A. fischeri* strains and *A. oerlinghausenensis* were the most chemically similar to each other.

In summary, even though *A. oerlinghausenensis* is phylogenetically more closely related to *A. fumigatus* than to *A. fischeri* (Fig. 12A), our chemical analyses suggest that the secondary metabolite profile of *A. oerlinghausenensis* is more similar to the profile of *A. fischeri* than it is to the profile of *A. fumigatus* (Fig. 14B and S6B-E from Steenwyk et al., 2020d). The similarity of secondary metabolite profiles of *A. oerlinghausenensis* and *A. fischeri* is consistent with our finding that the genome of *A. oerlinghausenensis* shares higher numbers of BGCs and gene families with *A. fischeri* than with *A. fumigatus* (Fig. 13). The broad clustering patterns in secondary metabolite-based plots (Fig. S6B-E from Steenwyk et al., 2020d) are less robust than, but consistent with, those of BGC-based plots (Fig. S6A from Steenwyk et al., 2020d),

suggesting that the observed similarities in the secondary metabolism-associated genotypes of *A. oerlinghausenensis* and *A. fischeri* are likely reflected in their chemotypes.

Conservation and divergence among biosynthetic gene clusters implicated in A. fumigatus pathogenicity

Secondary metabolites are known to play a role in *A. fumigatus* virulence (Raffa and Keller, 2019). We therefore conducted a focused examination of specific *A. fumigatus* BGCs and secondary metabolites that have been previously implicated in the organism's ability to cause human disease (Table 3). We found varying degrees of conservation and divergence that were

Table 3. Select *A. fumigatus* secondary metabolites implicated in modulating host biology

	Function	**Reference(s)**	**Evidence of biosynthetic gene cluster / secondary metabolite**						
			A. fumigatus			*A. oerlinghausenensis*	*A. fischeri*		
			Af293	CEA10	CEA17	CBS 139183^{T}	NRRL 181	NRRL 4585	NRRL 4161
Gliotoxin	Inhibits host immune response	(Sugui et al., 2007)	+/+	+/+	+/+	+/+	+/+	+/+	+/+
Fumitremorgin	Inhibits the breast cancer resistance protein	(González-Lobato et al., 2010)	+/-	+/+	+/-	+/+	+/+	+/+	+/+
Verruculogen	Changes electrophysical properties of human nasal epithelial cells	(Khoufache et al., 2007)	+/-	+/+	+/-	+/+	+/+	+/+	+/+
Trypacidin	Damages lung cell tissues	(Gauthier et al., 2012)	+/+	+/+	+/-	+/+	+/-	-/-	-/-
Pseurotin	Inhibits immunoglobulin E	(Ishikawa et al., 2009)	+/+	+/+	+/+	+/+	-/-	-/-	-/-

Fumagillin	Inhibits neutrophil function	(Fallon et al., 2010, 2011)	+/+	+/+	+/+	+/+	-/-	-/-	-/-

A list of select secondary metabolites implicated in human disease and their functional role are described here. All secondary metabolites listed or analogs thereof were identified during secondary metabolite profiling. Plus (+) and minus (-) signs indicate the presence or absence of the BGC and secondary metabolite, respectively. For example, +/+ indicates both BGC presence and evidence of secondary metabolite production, whereas +/- indicates BGC presence but no evidence of secondary metabolite production. '+/+' cells are colored orange; '-/-' cells are colored blue; '+/-' and '-/+' cells are colored green.

---

associated with the absence or presence of a secondary metabolite. Among conserved BGCs that were also associated with conserved secondary metabolite production, we highlight the mycotoxins gliotoxin and fumitremorgin. Interestingly, we note that only *A. fischeri* strains synthesized verruculogen, a secondary metabolite that is implicated in human disease and is encoded by the fumitremorgin BGC (Khoufache et al., 2007; Kautsar et al., 2019). Among BGCs that exhibited varying degrees of sequence divergence and divergence in their production of the corresponding secondary metabolites, we highlight those associated with the production of the trypacidin and fumagillin/pseurotin secondary metabolites. We found that nonpathogenic close relatives of *A. fumigatus* produced some but not all mycotoxins, which provides novel insight into the unique cocktail of secondary metabolites biosynthesized by *A. fumigatus*.

(i) Gliotoxin

Gliotoxin is a highly toxic compound and known virulence factor in *A. fumigatus* (Sugui et al., 2007). Nearly identical BGCs encoding gliotoxin are present in all pathogenic (*A. fumigatus*) and nonpathogenic (*A. oerlinghausenensis* and *A. fischeri*) strains examined (Fig. 15). Additionally, we found that all examined strains synthesized bisdethiobis(methylthio)gliotoxin a derivative from dithiogliotoxin, involved in the down-regulation of gliotoxin biosynthesis (Dolan et al., 2014), one of the main mechanisms of gliotoxin resistance in *A. fumigatus* (Kautsar et al., 2019).

(ii) Fumitremorgin and Verruculogen

Similarly, there is a high degree of conservation in the BGC that encodes fumitremorgin across all strains (Fig. 5). Fumitremorgins have known antifungal activity, are lethal to brine shrimp,

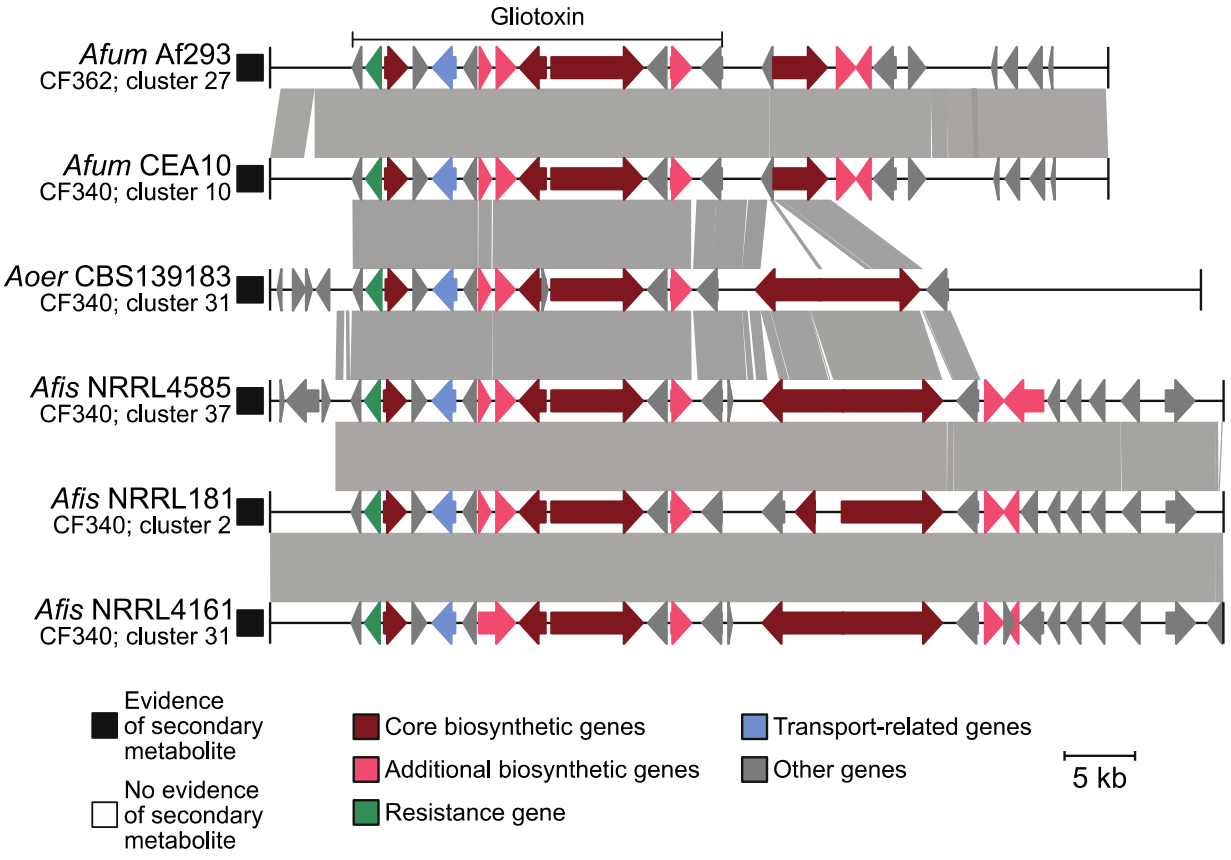

Figure 15. Conservation in the gliotoxin BGC correlates with conserved production of gliotoxin analogs in *A. fumigatus* and nonpathogenic close relatives.
Microsynteny analysis reveals a high degree of conservation in the BGC encoding gliotoxin across all isolates. The known gliotoxin gene cluster boundary is indicated above the *A. fumigatus* Af293 BGC. Black and white squares correspond to evidence or absence of evidence of secondary metabolite production, respectively. Genes are drawn as arrows with orientation indicated by the direction of the arrow. Gene function is indicated by gene color. Grey boxes between gene clusters indicate BLAST-based similarity of nucleotide sequences defined as being at least 100 bp in length, share at least 30% sequence similarity, and have an expectation value threshold of 0.01. Genus and species names are written using the following abbreviations: *Afum*: *A. fumigatus*; *Aoer*: *A. oerlinghausenensis*; *Afis*: *A. fischeri*. Below each genus and species abbreviation is the cluster family each BGC belongs to and their cluster number.

and are implicated in inhibiting mammalian proteins responsible for resistance to anticancer drugs in mammalian cells (Raffa and Keller, 2019). We found that conservation in the fumitremorgin BGC is associated with the production of fumitremorgins in all isolates examined. The fumitremorgin BGC is also responsible for the production of verruculogen, which is implicated to aid in *A. fumigatus* pathogenicity by changing the electrophysical properties of human nasal epithelial cells (Khoufache et al., 2007). Interestingly, we found that only *A. fischeri* strains produced verruculogen under the conditions we analyzed.

(iii) Trypacidin

Examination of the trypacidin BGC, which encodes a spore-borne and cytotoxic secondary metabolite, revealed a conserved cluster found in four pathogenic and nonpathogenic strains: *A. fumigatus* Af293, *A. fumigatus* CEA10, *A. oerlinghausenensis* CBS 139183^{T}, and *A. fischeri* NRRL 181 (Fig. S7 from Steenwyk et al., 2020d). Furthermore, we found that three of these four isolates (except *A. fischeri* NRRL 181) biosynthesized a trypacidin analog, monomethylsulochrin. Examination of the microsynteny of the trypacidin BGC revealed that it was conserved across all four genomes with the exception *A. fischeri* NRRL l81, which lacked a RING (Really Interesting New Gene) finger gene. Interestingly, RING finger proteins can mediate gene transcription (Poukka et al., 2000). We confirmed the absence of the RING finger protein by performing a sequence similarity search with the *A. fumigatus* Af293 RING finger protein (AFUA_4G14620; EAL89333.1) against the *A. fischeri* NRRL 181 genome. In the homologous locus in *A. fischeri*, we found no significant BLAST hit for the first 23 nucleotides of the RING finger gene suggestive of pseudogenization. Taken together, we hypothesize that

presence/absence polymorphisms or a small degree of sequence divergence between otherwise homologous BGCs may be responsible for the presence or absence of a toxic secondary metabolite in *A. fischeri* NRRL 181. Furthermore, inter- and intra-species patterns of trypacidin presence and absence highlight the importance of strain heterogeneity when examining BGCs.

(iv) Fumagillin/pseurotin

Examination of the intertwined fumagillin/pseurotin BGCs revealed that fumagillin has undergone substantial sequence divergence and that pseurotin is absent from strains of *A. fischeri*. The fumagillin/pseurotin BGCs are under the same regulatory control (Wiemann et al., 2013) and biosynthesize secondary metabolites that cause cellular damage during host infection (fumagillin (Guruceaga et al., 2019)) and inhibit immunoglobulin E production (pseurotin (Ishikawa et al., 2009)). Microsynteny of the fumagillin BGC reveals high sequence conservation between *A. fumigatus* and *A. oerlinghausenensis*; however, sequence divergence was observed between *A. oerlinghausenensis* and *A. fischeri* (Fig. 16). Accordingly, fumagillin production was only observed in *A. fumigatus* and *A. oerlinghausenensis* and not in *A. fischeri*. Similarly, the pseurotin BGC is conserved between *A. fumigatus* and *A. oerlinghausenensis*. Rather than sequence divergence, no sequence similarity was observed in the region of the pseurotin cluster in *A. fischeri*, which may be due to an indel event. Accordingly, no pseurotin production was observed among *A. fischeri* strains. Despite sequence conservation between *A. fumigatus* and *A. oerlinghausenensis*, no evidence of pseurotin biosynthesis was observed in *A. oerlinghausenensis*, which suggests regulatory decoupling of the intertwined

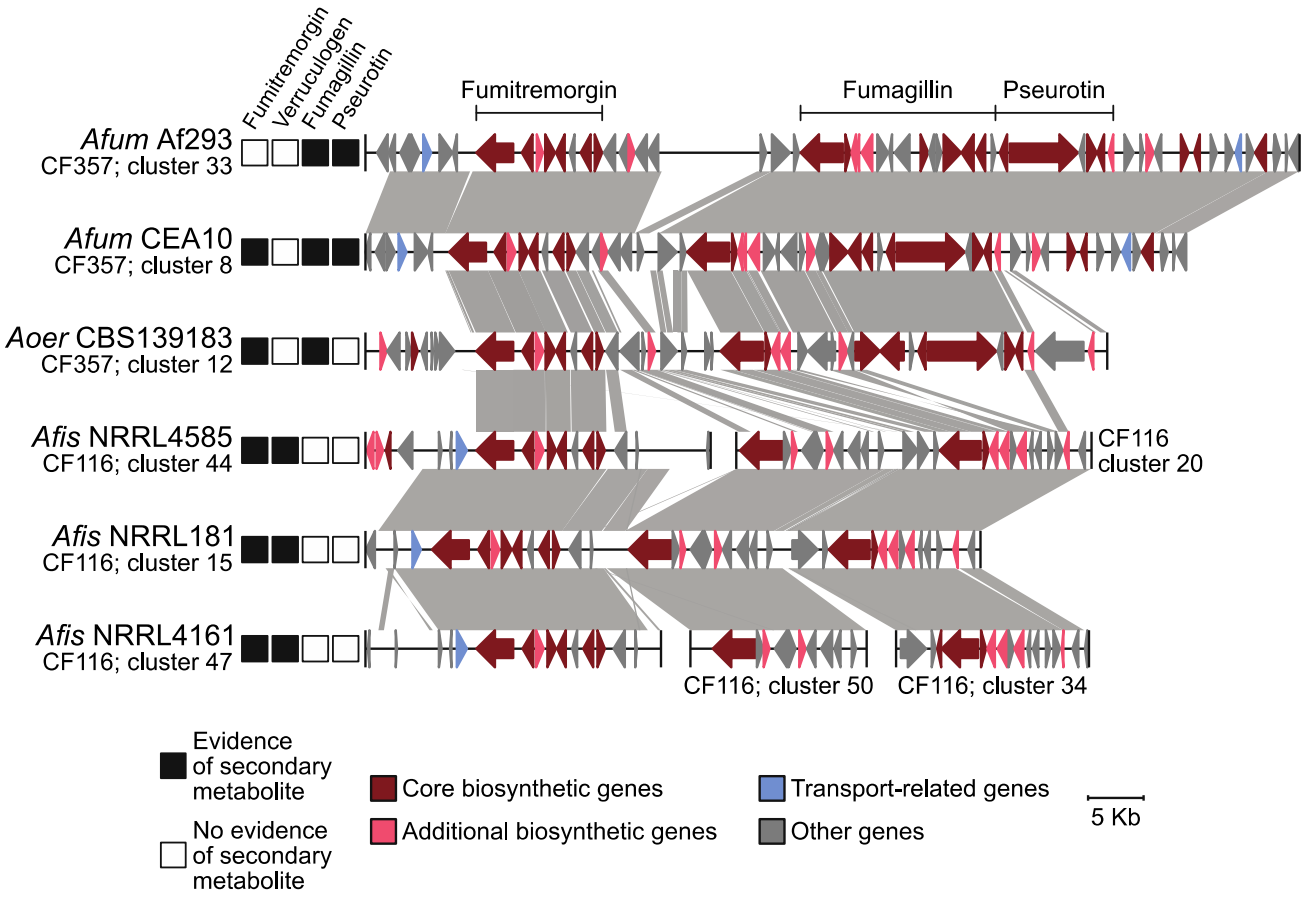

Figure 16. Conservation and divergence in the locus encoding the fumitremorgin and intertwined fumagillin/pseurotin BGCs.
Microsynteny analysis reveals conservation in the fumitremorgin BGC across all isolates. Interestingly, only *A. fischeri* strains synthesize verruculogen, a secondary metabolite also biosynthesized by the fumitremorgin BGC. In contrast, the intertwined fumagillin/pseurotin BGCs are conserved between *A. fumigatus* and *A. oerlinghausenensis* but divergent in *A. fischeri*. BGC conservation and divergence is associated with the presence and absence of a secondary metabolite, respectively. The same convention used in Fig. 4 is used to depict evidence of a secondary metabolite, represent genes and broad gene function, BGC sequence similarity, genus and species abbreviations, and BGC cluster families and cluster numbers.

---

fumagillin/pseurotin BGC. Alternatively, the genes downstream of the *A. fumigatus* pseurotin BGC, which are absent from the *A. oerlinghausenensis* locus, may contribute to BGC production and could explain the lack of pseurotin production in *A. oerlinghausenensis*. Altogether, these results show a striking correlation between sequence divergence and the production (or absence)

of secondary metabolites implicated in human disease among *A. fumigatus* and nonpathogenic closest relatives.

Discussion

Aspergillus fumigatus is a major fungal pathogen nested within a clade (known as section *Fumigati*) of at least 60 other species, the vast majority of which are nonpathogenic (Steenwyk et al., 2019c; Rokas et al., 2020a). Currently, it is thought that the ability to cause human disease evolved multiple times among species in section *Fumigati* (Rokas et al., 2020a). Secondary metabolites contribute to the success of the major human pathogen *A. fumigatus* in the host environment (Raffa and Keller, 2019) and can therefore be thought of as "cards" of virulence (Casadevall, 2007; Knowles et al., 2020). However, whether the closest relatives of *A. fumigatus*, *A. oerlinghausenensis* and *A. fischeri*, both of which are nonpathogenic, biosynthesize secondary metabolites implicated in the ability of *A. fumigatus* to cause human disease remained largely unknown. By examining genomic and chemical variation between and within *A. fumigatus* and its closest nonpathogenic relatives, we identified both conservation and divergence (including within species heterogeneity) in BGCs and secondary metabolite profiles (Fig. 12-16, S3, S5-8 from Steenwyk et al., 2020d; Table 3, S1, S3 from Steenwyk et al., 2020d). Examples of conserved BGCs and secondary metabolites include the major virulence factor, gliotoxin (Fig. 15), as well as several others (Fig. 16, S7 from Steenwyk et al., 2020d; Table 3, S1, S3 from Steenwyk et al., 2020d); examples of BGC and secondary metabolite heterogeneity or divergence include pseurotin, fumagillin, and several others (Fig. 16; Table 3, S1, S3 from Steenwyk et al., 2020d). Lastly, we found that the fumitremorgin BGC, which biosynthesizes

fumitremorgin in all three species, is also associated with verruculogen biosynthesis in *A. fischeri* strains (Fig. 16).

One of the surprising findings of our study was that although *A. oerlinghausenensis* and *A. fumigatus* are evolutionarily more closely related to each other than to *A. fischeri* (Fig. 12), *A. oerlinghausenensis* and *A. fischeri* appear to be more similar to each other than to *A. fumigatus* in BGC composition, gene family content, and secondary metabolite profiles. The power of pathogen-nonpathogen comparative genomics is best utilized when examining closely related species (Fedorova et al., 2008; Jackson et al., 2011; Moran et al., 2011; Mead et al., 2019a; Rokas et al., 2020a). Genomes from additional strains from the closest known nonpathogenic relatives of *A. fumigatus*, including from the closest species relative *A. oerlinghausenensis*, *A. fischeri*, and other nonpathogenic species in section *Fumigati* will be key for understanding the evolution of *A. fumigatus* pathogenicity.

Our finding that *A. oerlinghausenensis* and *A. fischeri* shares more gene families and BGCs with each other than they do with *A. fumigatus* (Fig. 12C, 13, S3, S4, S8 from Steenwyk et al., 2020d) suggests that the evolutionary trajectory of the *A. fumigatus* ancestor was marked by gene loss. We hypothesize that there were two rounds of gene family and BGC loss in the *A. fumigatus* stem lineage: (1) gene families and BGCs were lost in the common ancestor of *A. fumigatus* and *A. oerlinghausenensis* and (2) additional losses occurred in the *A. fumigatus* ancestor. In addition to losses, we note that 548 gene families and 16 BGCs are unique to *A. fumigatus*, which may have resulted from genetic innovation (e.g., *de novo* gene formation) or unique gene family and BGC retention (Fig. 13, S8 from Steenwyk et al., 2020d). In line with the larger number of

shared BGCs between *A. oerlinghausenensis* and *A. fischeri*, we found their secondary metabolite profiles were also more similar (Fig. 14, S6 from Steenwyk et al., 2020d). Notably, the evolutionary rate of the internal branch leading to the *A. fumigatus* common ancestor is much higher than those in the rest of the branches in our genome-scale phylogeny (Fig. S2B from Steenwyk et al., 2020d), suggesting that the observed gene loss and gene gain / retention events specific to *A. fumigatus* may be part of a wider set of evolutionary changes in the *A. fumigatus* genome. Analyses with a greater number of strains and species will help further test the validity of this hypothesis. More broadly, these results suggest that comparisons of the pathogen *A. fumigatus* against either the non-pathogen *A. oerlinghausenensis* (this manuscript) or the non-pathogen *A. fischeri* ((Mead et al., 2019a; Knowles et al., 2020) and this manuscript) will both be instructive in understanding the evolution of *A. fumigatus* pathogenicity.

When studying *Aspergillus* pathogenicity, it is important to consider any genetic and phenotypic heterogeneity between strains of a single species (Knox et al., 2016; Kowalski et al., 2016; Keller, 2017; Kowalski et al., 2019; Ries et al., 2019; Bastos et al., 2020b; Blachowicz et al., 2020; dos Santos et al., 2020b; Drott et al., 2020; Steenwyk et al., 2020c). Our finding of strain heterogeneity among gene families, BGCs, and secondary metabolites in *A. fumigatus* and *A. fischeri* (Fig. 12-14, S3, S4, S6, S8 from Steenwyk et al., 2020d) suggests considerable strain-level diversity in each species. For example, we found secondary metabolite profile strain heterogeneity was greater in *A. fumigatus* than *A. fischeri* (Fig. S6B-E from Steenwyk et al., 2020d). These results suggest that strain-specific secondary metabolite profiles may play a role in variation of pathogenicity among *A. fumigatus* strains. In support of this hypothesis, differential secondary metabolite production has been associated with differences in virulence

among isolates of *A. fumigatus* (Blachowicz et al., 2020). More broadly, our finding supports the hypothesis that strain-level diversity is an important parameter when studying pathogenicity (Kowalski et al., 2016; Keller, 2017; Kowalski et al., 2019; Ries et al., 2019; Bastos et al., 2020b; Blachowicz et al., 2020; dos Santos et al., 2020b; Drott et al., 2020; Steenwyk et al., 2020c).

Secondary metabolites contribute to *A. fumigatus* virulence through diverse processes including suppressing the human immune system and damaging tissues (Table 3). Interestingly, we found that the nonpathogens *A. oerlinghausenensis* and *A. fischeri* produced several secondary metabolites implicated in the ability of *A. fumigatus* human disease, such gliotoxin, trypacidin, verruculogen, and others (Fig. 15, 16, S7 from Steenwyk et al., 2020d; Table 3, S3 from Steenwyk et al., 2020d). Importantly, our work positively identified secondary metabolites for many structural classes implicated in a previous taxonomic study (Samson et al., 2007). These results suggest that several of the secondary metabolism-associated cards of virulence present in *A. fumigatus* are conserved in closely related nonpathogens (summarized in Fig. 17) as well as in closely related pathogenic species, such as *A. novofumigatus* (Kjærbølling et al., 2018).

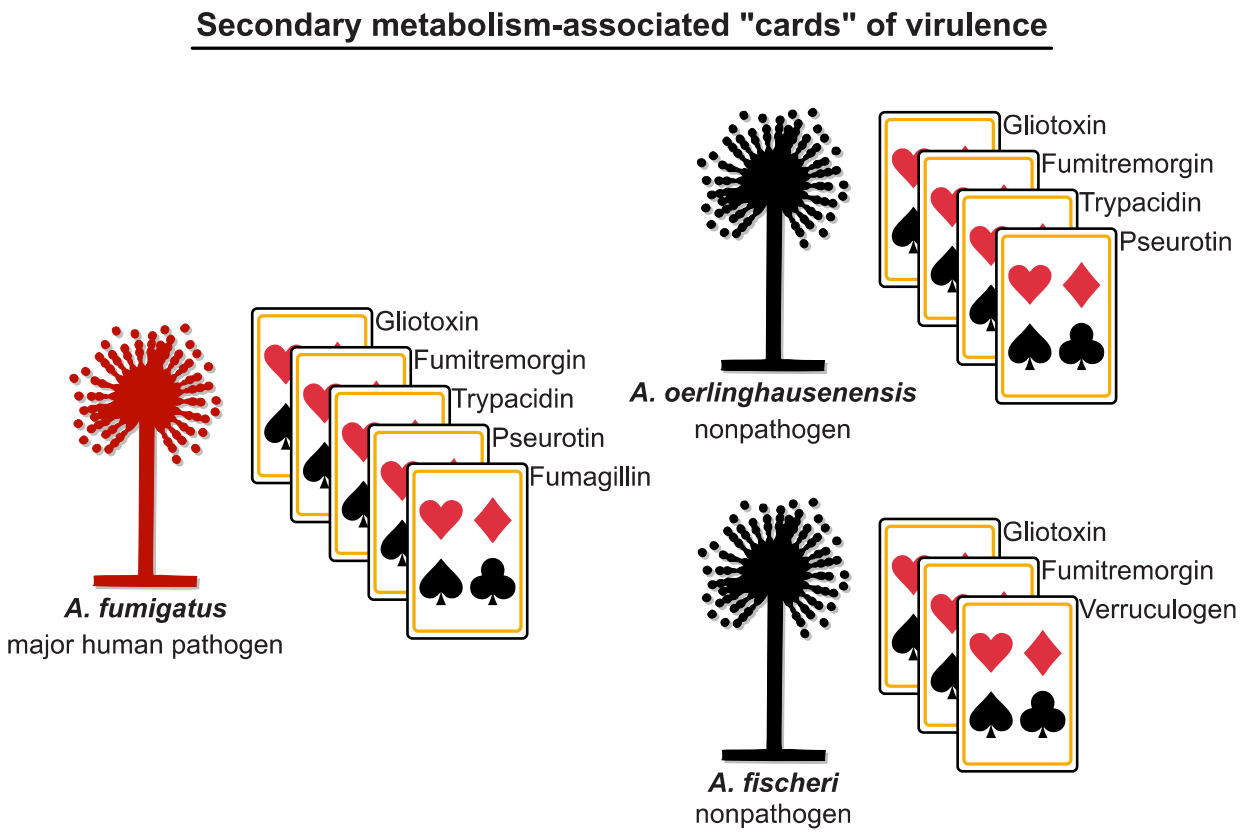

Figure 17. Secondary metabolism-associated "cards" of virulence among *A. fumigatus* and close relatives.
Secondary metabolites contribute to the "hand of cards"' that enable *A. fumigatus* to cause disease. Here, we show that the nonpathogenic closest relatives of *A. fumigatus* possess a subset of the *A. fumigatus* secondary metabolism-associated cards of virulence. We hypothesize that the unique combination of cards of *A. fumigatus* contributes to its pathogenicity and that the cards in *A. oerlinghausenensis* and *A. fischeri* (perhaps in combination with other non-secondary-metabolism-associated cards, such as thermotolerance) are insufficient to cause disease. Pathogenic and nonpathogenic species are shown in red and black, respectively. Cartoons of *Aspergillus* species were obtained from WikiMedia Commons (source: M. Piepenbring) and modified in accordance with the Creative Commons Attribution-Share Alike 3.0 Unported license (https://creativecommons.org/licenses/by-sa/3.0/deed.en).

Interestingly, disrupting the ability of *A. fumigatus* to biosynthesize gliotoxin attenuates but does not abolish virulence (Sugui et al., 2007; Dagenais and Keller, 2009; Keller, 2017), whereas disruption of the ability of *A. fischeri* NRRL 181 to biosynthesize secondary metabolites, including gliotoxin, does not appear to influence virulence (Knowles et al., 2020). Our findings,

together with previous studies, support the hypothesis that individual secondary metabolites are "cards" of virulence in a larger "hand" that *A. fumigatus* possesses.

CHAPTER 5

Genomic and phenotypic analysis of COVID-19-associated pulmonary aspergillosis isolates of *Aspergillus fumigatus*[4]

Introduction

On March 11, 2020, the World Health Organization declared the ongoing pandemic caused by SARS-CoV-2, which causes COVID-19, a global emergency (Sohrabi et al., 2020). Similar to other viral infections, patients may be more susceptible to microbial secondary infections, which can complicate disease management strategies and result in adverse patient outcomes (Brüggemann et al., 2020; Cox et al., 2020). For example, approximately one quarter of patients infected with the H1N1 influenza virus during the 2009 pandemic were also infected with bacteria or fungi (MacIntyre et al., 2018; Zhou et al., 2020). Among COVID-19 patients, one study found that ~17% of individuals also have bacterial infections (Langford et al., 2020) and another that ~40% of patients with severe COVID-19 pneumonia were also infected with filamentous fungi from the genus *Aspergillus* (Nasir et al., 2020). A third study reported that ~26% of patients with acute respiratory distress syndrome-associated COVID-19 were also infected with *Aspergillus fumigatus* and had high rates of mortality (Koehler et al., 2020). Other studies from around the world have also reported high incidences of *Aspergillus* infections among patients with COVID-19 (Alanio et al., 2020; Chen et al., 2020; Rutsaert et al., 2020; van Arkel et al., 2020). Taken together, these findings have prompted some to suggest routine

[4]This work is published in: Steenwyk, J. L., Mead, M. E., de Castro, P. A., Valero, C., Damasio, A., dos Santos, R. A. C., et al. (2021). Genomic and Phenotypic Analysis of COVID-19-Associated Pulmonary Aspergillosis Isolates of Aspergillus fumigatus. Microbiol. Spectr. 9. doi:10.1128/Spectrum.00010-21.

clinical testing for secondary infections of *Aspergillus* fungi among COVID-19 patients (Armstrong-James et al., 2020; Gangneux et al., 2020). Despite the prevalence microbial infections and their association with adverse patient outcomes, these secondary infections are only beginning to be understood.

Invasive pulmonary aspergillosis is caused by tissue infiltration of *Aspergillus* species after inhalation of their asexual spores (Fig. 18); more than 250,000 aspergillosis infections are estimated to occur annually and have high mortality rates (Bongomin et al., 2017). The major

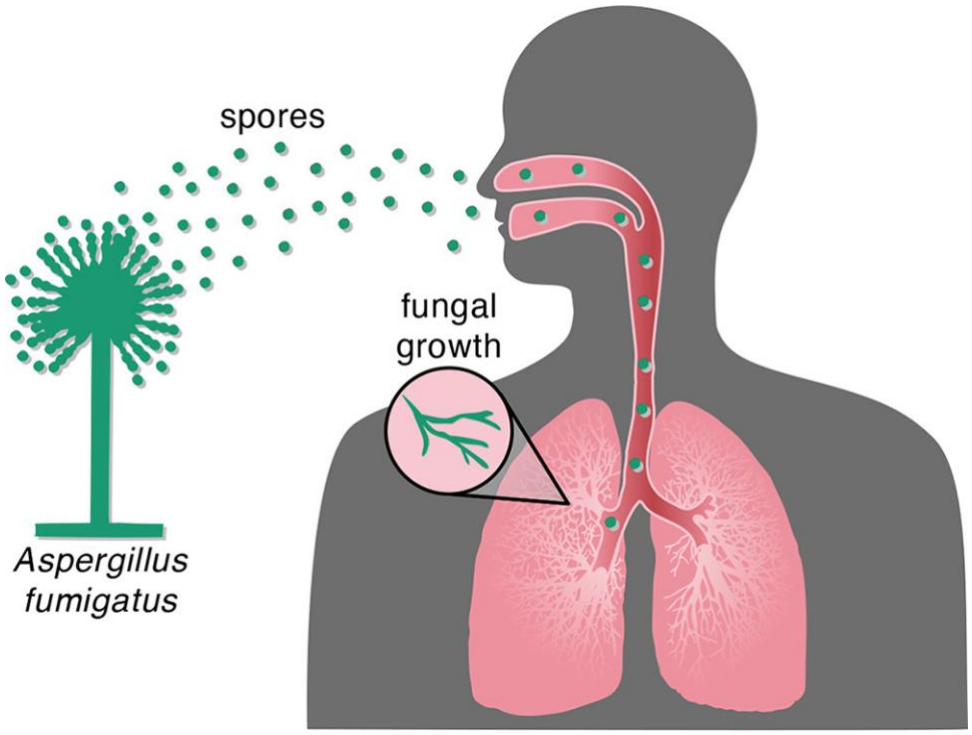

Figure 18. Inhalation of *Aspergillus* spores can result in fungal infection.
Inhalation of *Aspergillus* spores from the environment can travel to the lung and then grow vegetatively and spread to other parts of the body.

etiological agent of aspergillosis is *A. fumigatus* (Latgé and Chamilos, 2019), although a few other *Aspergillus* species are also known to cause aspergillosis (Bastos et al., 2020b; dos Santos et al., 2020b; Rokas et al., 2020a; Steenwyk et al., 2020c). Numerous factors are known to be

associated with *A. fumigatus* pathogenicity, including its ability to grow at the human body temperature (37°C) and withstand oxidative stress (Kamei and Watanabe, 2005; Tekaia and Latgé, 2005; Shwab et al., 2007; Losada et al., 2009; Abad et al., 2010; Grahl et al., 2012; Yin et al., 2013; Wiemann et al., 2014; Knox et al., 2016; Kowalski et al., 2019; Raffa and Keller, 2019; Blachowicz et al., 2020). Disease management of *A. fumigatus* is further complicated by resistance to antifungal drugs among strains (Chamilos and Kontoyiannis, 2005; Howard and Arendrup, 2011; Chowdhary et al., 2014; Sewell et al., 2019) Additionally, *A. fumigatus* strains have been previously shown to exhibit strain heterogeneity with respect to virulence and pathogenicity-associated traits (Kowalski et al., 2016; Keller, 2017; Kowalski et al., 2019; Ries et al., 2019; dos Santos et al., 2020b; Steenwyk et al., 2020d). However, it remains unclear whether the genomic and pathogenicity-related phenotypic characteristics of CAPA isolates are similar to or distinct from those of previously studied clinical strains of *A. fumigatus*.

To address this question and gain insight into the pathobiology of *A. fumigatus* CAPA isolates, we examined the genomic and phenotypic characteristics of four CAPA isolates obtained from four critically ill patients of two different centers in Cologne, Germany (Koehler et al., 2020) (Table 1 from Steenwyk et al. 2021d). All patients were submitted to intensive care units due to moderate to severe respiratory distress syndrome (ARDS). Genome-scale phylogenetic (or phylogenomic) analyses revealed CAPA isolates formed a monophyletic group closely related to reference strains Af293 and A1163. Examination of the mutational spectra of 206 genes known to modulate virulence in *A. fumigatus* (which are hereafter referred to as genetic determinants of virulence) revealed several putative loss of function (LOF) mutations. Notably, CAPA isolate D

had the most putative LOF mutations among genes whose null mutants are known to increase virulence. The

profiles of pathogenicity-related traits and of secondary metabolites of the CAPA isolates were similar to those of reference *A. fumigatus* strains Af293 and CEA17 or CEA10, which are parental strains of A1163 (Bertuzzi et al., 2020). One notable exception was that CAPA isolate D was significantly more virulent than other strains in an invertebrate model of disease, but on par with two other clinical strains of *A. fumigatus*. These results suggest that the genomes of *A. fumigatus* CAPA isolates contain nearly complete and intact repertoires of genetic determinants of virulence and have phenotypic profiles that are broadly expected for *A. fumigatus* clinical isolates. However, we did find evidence for genetic and phenotypic strain heterogeneity. These results suggest the CAPA isolates show similar phenotypic profiles as *A. fumigatus* clinical strains Af293 and A1163 and expand our understanding of CAPA.

Materials and Methods

Patient information and ethics approval

Patients were included into the FungiScope® global registry for emerging invasive fungal infections (www.ClinicalTrials.gov, NCT 01731353). The clinical trial is approved by the Ethics Committee of the University of Cologne, Cologne, Germany (Study ID: 05-102) (Seidel et al., 2017). Since 2019, patients with invasive aspergillosis are also included.

DNA quality control, library preparation, and sequencing

Sample DNA concentration was measured by Qubit fluorometer and DNA integrity and purity by agarose gel electrophoresis. For each sample, 1-1.5μg genomic DNA was randomly

fragmented by Covaris and fragments with average size of 200-400bp were selected by Agencourt AMPure XP-Medium kit. The selected fragments were end-repaired, 3' adenylated, adapters-ligated, and amplified by PCR. Double-stranded PCR products were recovered by the AxyPrep Mag PCR clean up Kit, and then heat denatured and circularized by using the splint oligo sequence. The single-strand circle DNA (ssCir DNA) products were formatted as the final library and went through further QC procedures. The libraries were sequenced on the MGISEQ2000 platform.

Genome assembly and annotations

Short-read sequencing data of each sample were assembled using MaSuRCA, v3.4.1 (Zimin et al., 2013). Each *de novo* genome assembly was annotated using the MAKER genome annotation pipeline, v2.31.11 (Holt and Yandell, 2011), which integrates three *ab initio* gene predictors: AUGUSTUS, v3.3.3 (Stanke and Waack, 2003), GeneMark-ES, v4.59 (Besemer and Borodovsky, 2005), and SNAP, v2013-11-29 (Korf, 2004). Fungal protein sequences in the SwissProt database (release 2020_02) were used as homology evidence for the genome annotation. The MAKER annotation process occurs in an iterative manner as described previously (Shen et al., 2018). In brief, for each genome, repeats were first soft-masked using RepeatMasker v4.1.0 (http://www.repeatmasker.org) with the library Repbase library release-20181026 and the "-species" parameter set to "Aspergillus fumigatus". GeneMark-ES was then trained on the masked genome sequence using the self-training option ("--ES") and the branch model algorithm ("--fungus"), which is optimal for fungal genome annotation. On the other hand, an initial MAKER analysis was carried out where gene annotations were generated directly from homology evidence, and the resulting gene models were used to train both AUGUSTUS

and SNAP. Once trained, the ab initio predictors were used together with homology evidence to conduct a first round of full MAKER analysis. Resulting gene models supported by homology evidence were used to re-train AUGUSTUS and SNAP. A second round of MAKER analysis was conducted using the newly trained AUGUSTUS and SNAP parameters, and once again the resulting gene models with homology supports were used to re-train AUGUSTUS and SNAP. Finally, a third round of MAKER analysis was performed using the new AUGUSTUS and SNAP parameters to generate the final set of annotations for the genome. The completeness of *de novo* genome assemblies and *ab initio* gene predictions was assessed using BUSCO, v4.1.2 (Waterhouse et al., 2018a) using 4,191 pre-selected 'nearly' universally single-copy orthologous genes from the Eurotiales database (eurotials_odb10.2019-11-20) in OrthoDB, v10.1 (Waterhouse et al., 2013).

Polymorphism identification

To characterize and examine the putative impact of polymorphisms in the genomes of the CAPA isolates, we identified single nucleotide polymorphisms (SNPs), insertion-deletion polymorphisms (indels), and copy number (CN) polymorphisms. To do so, reads were first quality-trimmed and mapped to the genome of *A. fumigatus* Af293 (RefSeq assembly accession: GCF_000002655.1) following a previously established protocol (Steenwyk and Rokas, 2017). Specifically, reads were first quality-trimmed with Trimmomatic, v0.36 (Bolger et al., 2014), using the parameters leading:10, trailing:10, slidingwindow:4:20, minlen:50. The resulting quality-trimmed reads were mapped to the *A. fumigatus* Af293 genome using the Burrows-Wheeler Aligner (BWA), v0.7.17 (Li, 2013), with the mem parameter. Thereafter, mapped reads

were converted to a sorted bam and mpileup format for polymorphism identification using SAMtools, v.1.3.1 (Li et al., 2009a).

To identify SNPs and indels, mpileup files were used as input into VarScan, v2.3.9 (Koboldt et al., 2012), with the mpileup2snp and mpileup2indel functions, respectively. To ensure only confident SNPs and indels were identified, a Fischer's Exact test p-value threshold of 0.05 and minimum variant allele frequency of 0.75 were used. The resulting Variant Call Format files were used as input to snpEff, v.4.3t (Cingolani et al., 2012), which predicted their functional impacts on gene function as high, moderate, or low. To identify CN variants, the sorted bam files were used as input into Control-FREEC, v9.1 (Boeva et al., 2011, 2012). The coefficientOfVariation parameter was set to 0.062 and window size was automatically determined by Control-FREEC. To ensure high-confidence in CN variant identification, a p-value threshold of 0.05 was used for both Wilcoxon Rank Sum and Kolmogorov Smirnov tests.

To identify evidence of putative pseudogenization between reference strains A1163 and Af293, we used a previously established approach (Ortiz-Merino et al., 2017; Steenwyk et al., 2020c). More specifically, we compared lengths of gene pairs as a proxy for pseudogenization. A gene was considered a putative pseudogene in one of the strains if the gene was 70% the length of its reciprocal best blast hit in the other strain.

Maximum likelihood molecular phylogenetics

To taxonomically identify the species of *Aspergillus* sequenced, we conducted molecular phylogenetic analysis of two different loci and two different datasets. In the first analysis, the

nucleotide sequence of the alpha subunit of translation elongation factor EF-1, *tef1* (NCBI Accession: XM_745295.2), from the genome of *A. fumigatus* Af293 was used to extract other fungal *tef1* sequences from NCBI's fungal nucleotide reference sequence database (downloaded July 2020) using the blastn function from NCBI's BLAST+, v2.3.0 (Camacho et al., 2009). *Tef1* sequences were extracted from the CAPA isolates by identifying their best BLAST hit. Sequences from the top 100 best BLAST hits in the fungal nucleotide reference sequence database and the four *tef1* sequences from the CAPA isolates were aligned using MAFFT, v7.402 (Katoh and Standley, 2013) using previously described parameters (Steenwyk et al., 2019c) with slight modifications. Specifically, the following parameters were used: --op 1.0 --maxiterate 1000 --retree 1 --genafpair. The resulting alignment was trimmed using ClipKIT, v0.1 (Steenwyk et al., 2020c), with default 'gappy' mode. The trimmed alignment was then used to infer the evolutionary history of *tef1* sequences using IQ-TREE2 (Minh et al., 2020). The best fitting substitution model—TIM3 with empirical base frequencies, allowing for a proportion of invariable sites, and a discrete Gamma model (Yang, 1994; Gu et al., 1995) with four rate categories (TIM3+F+I+G4)—was determined using Bayesian Information Criterion. In the second analysis, the same process was used to conduct molecular phylogenetic analysis using calmodulin nucleotide sequences from *Aspergillus* section *Fumigati* species and *Aspergillus clavatus*, an outgroup taxon, using sequences from NCBI that were made available elsewhere (dos Santos et al., 2020a). For calmodulin sequences, the best fitting substitution model was TNe (Tamura and Nei, 1993) with a discrete Gamma model with four rate categories (TNe+G4). Bipartition support was assessed using 5,000 ultrafast bootstrap support approximations (Hoang et al., 2018).

To determine what strains of *A. fumigatus* the CAPA isolates were most similar to, we conducted phylogenomic analyses using the 50 *Aspergillus* proteomes. To do so, we first identified orthologous groups of genes across all 50 *Aspergillus* using OrthoFinder, 2.3.8 (Emms and Kelly, 2019). OrthoFinder takes as input the proteome sequence files from multiple genomes and conducts all-vs-all sequence similarity searches using DIAMOND, v0.9.24.125 (Buchfink et al., 2015). Our input included 50 total proteomes: 47 were *A. fumigatus*, two were *A. fischeri*, and one was *A. oerlinghausenensis* (Fedorova et al., 2008; Lind et al., 2017; Steenwyk et al., 2020d). OrthoFinder then clusters sequences into orthologous groups of genes using the graph-based Markov Clustering Algorithm (van Dongen, 2000). To maximize the number of single-copy orthologous groups of genes found across all input genomes, clustering granularity was explored by running 41 iterations of OrthoFinder that differed in their inflation parameter. Specifically, iterations of OrthoFinder inflation parameters were set to 1.0-5.0 with a step of 0.1. The lowest number of single-copy orthologous groups of genes was 3,399 when using an inflation parameter of 1.0; the highest number was 4,525 when using inflation parameter values of 3.8 and 4.1. We used the groups inferred using an inflation parameter of 3.8.

Next, we built the phylogenomic data matrix and reconstructed evolutionary relationships among the 50 *Aspergillus* genomes. To do so, the protein sequences from 4,525 single-copy orthologous groups of genes were aligned using MAFFT, v7.402 (Katoh and Standley, 2013), with the following parameters: --bl 62 --op 1.0 --maxiterate 1000 --retree 1 --genafpair. Next, nucleotide sequences were threaded onto the protein alignments using function thread_dna in PhyKIT, v0.0.1 (Steenwyk et al., 2021b). The resulting codon-based alignments were then trimmed using ClipKIT, v0.1 (Steenwyk et al., 2020b), using the gappy mode. The resulting aligned and

trimmed alignments were then concatenated into a single matrix with 7,133,367 sites using the PhyKIT function create_concat. To reconstruct the evolutionary history of the 50 *Aspergillus* genomes, a single best-fitting model of sequence substitution and rate heterogeneity was estimated across the entire matrix using IQ-TREE2, v.2.0.6 (Minh et al., 2020). The best-fitting model was determined to be a general time reversible model with empirical base frequencies and invariable sites with a discrete Gamma model with four rate categories (GTR+F+I+G4) (Tavaré, 1986; Gu et al., 1995; Waddell and Steel, 1997; Vinet and Zhedanov, 2011) using Bayesian Information Criterion. During tree search, the number of candidate trees maintained during maximum likelihood tree search was increased from five to ten. Five independent searches were conducted and the tree with the best log-likelihood score was chosen as the 'best' phylogeny. Bipartition support was evaluated using 5,000 ultrafast bootstrap approximations (Hoang et al., 2018).

Biosynthetic gene cluster prediction

To predict BGCs in the genomes of *A. fumigatus* strains Af293 and the CAPA isolates, gene boundaries inferred by MAKER were used as input into antiSMASH, v4.1.0 (Weber et al., 2015). Using a previously published list of genes known to encode BGCs in the genome of *A. fumigatus* Af293 (Lind et al., 2017), BLAST-based searches using an expectation value threshold of 1×10^{-10} were used to identify BGCs implicated in modulating host biology using NCBI's BLAST+, v2.3.0 (Camacho et al., 2009). Among predicted BGCs that did not match the previously published list, we further examined their evolutionary history if at least 50% of genes showed similarity to species outside of the genus *Aspergillus*, which is information provided in

the antiSMASH output. Using these criteria, no evidence suggestive of horizontally acquired BGCs from distant relatives was detected.

Characterization of biosynthesized secondary metabolites

General Experimental Procedures

HRESIMS experiments utilized a Thermo LTQ Orbitrap XL mass spectrometer equipped with an electrospray ionization source. A Waters Acquity UPLC (Waters Corp.) was utilized using a BEH C_{18} column (1.7 μm; 50 mm x 2.1 mm) set to a temperature of 40°C and a flow rate of 0.3 ml/min. The mobile phase consisted of a linear gradient of CH_3CN-H_2O (both acidified with 0.1% formic acid), starting at 15% CH_3CN and increasing linearly to 100% CH_3CN over 8 min, with a 1.5 min hold before returning to the starting condition.

Growth and extraction of fungal cultures

To identify the chemical differences between the various *A. fumigatus* strains and isolates (Af293, CEA10, CAPA isolates A, B, C, and D), they were grown in a clinically relevant growth condition (37˚C) and extracted for chemometric analysis. Czapek Dox Agar (Sigma Aldrich) Petri plates were inoculated from the asexual spores of each strain in biological triplicates. Subsequently, the plates were incubated at 37°C in the dark for three days. The cultures were extracted by chopping and transferring the agar to 20 mL scintillation vials, adding of 10 mL of acetone, thoroughly shaking, and then letting the samples sit for 4 hours. Lastly, the cultures were filtered and evaporated to dryness under Nitrogen gas.

Metabolomics Analyses

Principal component analysis (PCA) was performed on the UPLC-MS data. Untargeted UPLC-MS datasets for each sample were individually aligned, filtered, and analyzed using MZmine 2.20 software (https://sourceforge.net/projects/mzmine/) (Pluskal et al., 2010). Peak detection was achieved using the following parameters: noise level (absolute value), 1.5×10^5; minimum peak duration, 0.05 min; *m/z* variation tolerance, 0.05; and *m/z* intensity variation, 20%. Peak list filtering and retention time alignment algorithms were used to refine peak detection. The join algorithm integrated all sample profiles into a data matrix using the following parameters: *m/z* and retention time balance set at 10.0 each, *m/z* tolerance set at 0.001, and RT tolerance set at 0.5 mins. The resulting data matrix was exported to Excel (Microsoft) for analysis as a set of *m/z* – retention time pairs with individual peak areas detected in quadruplicate analyses. Samples that did not possess detectable quantities of a given marker ion were assigned a peak area of zero to maintain the same number of variables for all sample sets. Ions that did not elute between 2 and 8 minutes and/or had an *m/z* ratio less than 100 or greater than 1,200 Da were removed from analysis. Relative standard deviation was used to quantify variance between the technical replicate injections, which may differ slightly based on instrument variance. A cutoff of 1.0 was used at any given *m/z* – retention time pair across the technical replicate injections of one biological replicate, and if the variance was greater than the cutoff, it was assigned a peak area of zero (Caesar et al., 2018). Final chemometric analysis, including data filtering and PCA were conducted using Python. The PCA plots were generated using data from the averaged biological replicates from the Petri dish cultures. Each biological replicate was plotted using averaged peak areas obtained across four replicate injections (technical replicates). The principal components (PC) were generated and processed via Scikit Learn decomposition and Pandas, v0.25.3, Python

libraries. The PCA data were plotted using Altair, v4.1.0, Python graphing libraries. These data were converted into a dataframe via Pandas, and the PCs were created from the dataframe using Scikit Learn decomposition. The PCA scores and loadings plots were then plotted using the PCs dataframe that was generated from Scikit Learn.

Infection of *Galleria mellonella*

Survival curves (n≥20/strain) of *Galleria mellonella* infected with CAPA isolates A, B, C, and D. Phosphate buffered saline (PBS) without asexual spores (conidia) was administered as a negative control. A log-rank test was used to examine strain heterogeneity followed by pairwise comparisons with Benjamini-Hochberg multi-test correction (Benjamini and Hochberg, 1995). All the selected larvae of *Galleria mellonella* were in the final (sixth) instar larval stage of development, weighing 275–330 milligram. Fresh conidia from each strain were harvested from minimal media (MM) plates in PBS solution and filtered through a Miracloth (Calbiochem). For each strain, the spores were counted using a hemocytometer and the stock suspension was done at 2×10^8 conidia/milliliter. The viability of the administered inoculum was determined by plating a serial dilution of the conidia on MM medium at 37°C. A total of 5 microliters (1×10^6 conidia/larva) from each stock suspension was inoculated per larva. The control group was composed of larvae inoculated with 5 microliters of PBS to observe the killing due to physical trauma. The inoculum was performed by using Hamilton syringe (7000.5KH) via the last left proleg. After infection, the larvae were maintained in petri dishes at 37°C in the dark and were scored daily. Larvae were considered dead by presenting the absence of movement in response to touch.

Growth assays

To examine growth conditions of the CAPA isolates and reference strains Af293 and CEA17, plates were inoculated with 10^4 spores per strain and allowed to grow for five days on solid MM or MM supplemented with various concentrations of osmotic (sorbitol, NaCl), cell wall (congo red, calcofluor white and caspofungin), and oxidative stress agents (menadione and t-butyl) at 37°C. MM had 1% (weight / volume) glucose, original high nitrate salts, trace elements, and a pH of 6.5; trace elements, vitamins, and nitrate salts compositions follow standards described elsewhere (Käfer, 1977). To correct for strain heterogeneity in growth rates, radial growth in centimeters in the presence of stressors was divided by radial growth in centimeters in the absence of the stressor.

To determine the minimal inhibitory concentrations of antifungal drugs in the CAPA isolates and reference strains Af293 and CEA17, strains were grown in 96-well plates at a concentration of 10^4 spores / well in 200 μl of RPMI-1640 supplemented with increasing concentrations of Amphotericin B, Voriconazole, Itraconazole and Posaconazole, according to the protocol elaborated by the Clinical and Laboratory Standards Institute (CLSI, 2008). Minimal inhibitory concentration was defined as the lowest concentration of drugs that visually inhibited 100% fungal growth. Three independent experiments were carried out for each antifungal drug.

Data Availability

Newly sequenced genomes assemblies, annotations, and raw short reads have been deposited to NCBI's GenBank database under BioProject accession PRJNA673120. Supplementary tables, figures, and files; additional copies of genome assemblies, annotations, and gene coordinates;

raw data including the genome assembly and annotations of all analyzed *Aspergillus* genomes; the aligned and trimmed phylogenetic and phylogenomic data matrices; polymorphisms identified in the present project; and predicted BGCs have been uploaded to figshare (doi: 10.6084/m9.figshare.13118549).

Results

CAPA isolates belong to *A. fumigatus* and are closely related to reference strains Af293 and A1163

To confirm that the CAPA isolates belong to *A. fumigatus*, we sequenced, assembled, and annotated their genomes (Table S1 from Steenwyk et al. 2021d; all supplementary tables are posted at figshare, doi: 10.6084/m9.figshare.13118549). Phylogenetic analyses conducted using *tef1* (Fig. S1 from Steenwyk et al. 2021d; all supplementary figures are posted at figshare, doi: 10.6084/m9.figshare.13118549) and calmodulin (Fig. S2 from Steenwyk et al. 2021d) sequences suggested that all CAPA isolates are *A. fumigatus*. Phylogenomic analysis using 50 *Aspergillus* genomes (the four CAPA isolates, 43 *A. fumigatus* genomes that span the known diversity of the species including strains Af293 and A1163 (Nierman et al., 2005; Fedorova et al., 2008; Liu et al., 2011; Abdolrasouli et al., 2015a; Knox et al., 2016; Lind et al., 2017; Paul et al., 2017; dos Santos et al., 2020b), *A. fischeri* strains NRRL 181 and NRRL 4585, and *A. oerlinghausenensis* strain CBS 139183^{T} (Fedorova et al., 2008; Steenwyk et al., 2020d)) confirmed that all CAPA isolates are *A. fumigatus* (Fig. 19). Phylogenomic analyses also revealed the CAPA isolates formed a monophyletic group closely related to reference strains Af293 and A1163. CAPA isolates are inferred to be closely related, which may be due to the fact that they are all from the same geographic area.

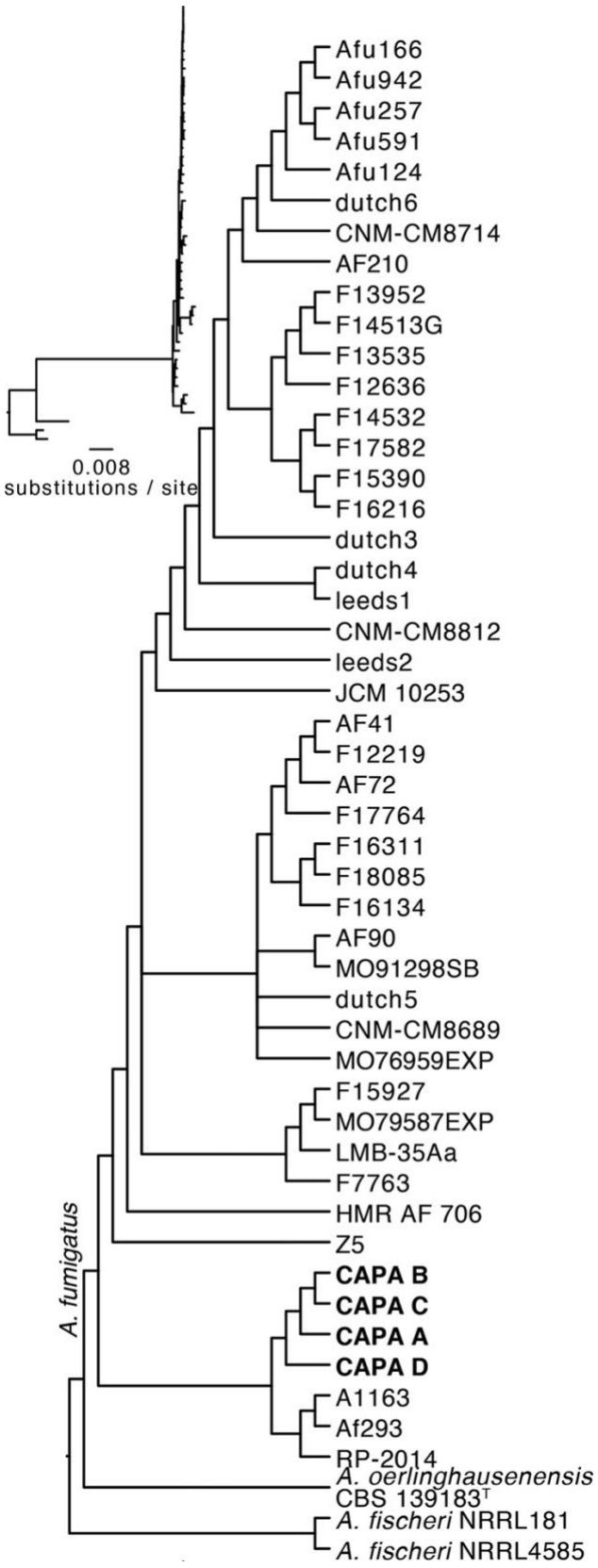
0.008
substitutions / site
Afu166
Afu942
Afu257
Afu591
Afu124
dutch6
CNM-CM8714
AF210
F13952
F14513G
F13535
F12636
F14532
F17582
F15390
F16216
dutch3
dutch4
leeds1
CNM-CM8812
leeds2
JCM 10253
AF41
F12219
AF72
F17764
F16311
F18085
F16134
AF90
MO91298SB
dutch5
CNM-CM8689
MO76959EXP
F15927
MO79587EXP
LMB-35Aa
F7763
HMR AF 706
Z5
CAPA B
CAPA C
CAPA A
CAPA D
A1163
Af293
RP-2014
A. fumigatus
A. oerlinghausenensis
CBS 139183T
A. fischeri NRRL181
A. fischeri NRRL4585

Figure 19. Phylogenomics confirms that COVID-19-associated pulmonary aspergillosis (CAPA) isolates are *Aspergillus fumigatus*.
Phylogenomic analysis of a concatenated matrix of 4,525 single-copy orthologous groups genes (sites: 7,133,367) confirmed CAPA isolates are *A. fumigatus*. Furthermore, CAPA isolates are closely related to reference strains A1163 and Af293. Bipartitions with less than 85% ultrafast bootstrap approximation support were collapsed.

CAPA isolate genomes contain polymorphisms in genetic determinants of virulence and biosynthetic gene clusters

An extensive literature and database search identified 206 genetic determinants of virulence (File S1; all supplementary files are posted at figshare, doi: 10.6084/m9.figshare.13118549) (Abad et al., 2010; Bignell et al., 2016; Kjærbølling et al., 2018; Mead et al., 2019a; Urban et al., 2019). We define genetic determinants of virulence as genes that alter virulence in an animal model of disease when deleted or are required for biosynthesis of secondary metabolites known to affect virulence. This definition resulted in a list of genes distinct from those previously published, which include genes that contribute to allergy-related phenotypes, genes that are computationally predicted to contribute to virulence but have yet to be validated in an animal model of fungal disease (Tomee and Kauffman, 2000; Askew, 2008; Puértolas-Balint et al., 2019; Pennerman et al., 2020).

To determine if the 206 genetic determinants of virulence are conserved in CAPA isolates, we conducted sequence similarity searches of gene sequences in the genomes of the CAPA isolates. We found that all 206 genes were present in the genomes of the CAPA isolates. Furthermore, none of the 206 genetic determinants of virulence showed any copy number variation among CAPA isolates. Examination of single nucleotide polymorphisms (SNPs) and insertion/deletion (indel) polymorphisms coupled with variant effect prediction in these 206 genes (Fig. 20; File S2

from Steenwyk et al. 2021d) showed that all CAPA isolates shared multiple polymorphisms resulting in early stop codons or frameshift mutations suggestive of loss of function (LOF) in *NRPS8* (AFUA_5G12730), a nonribosomal peptide synthetase gene that encodes an unknown secondary metabolite (Lind et al., 2017). LOF mutations in *NRPS8* are known to result in

---

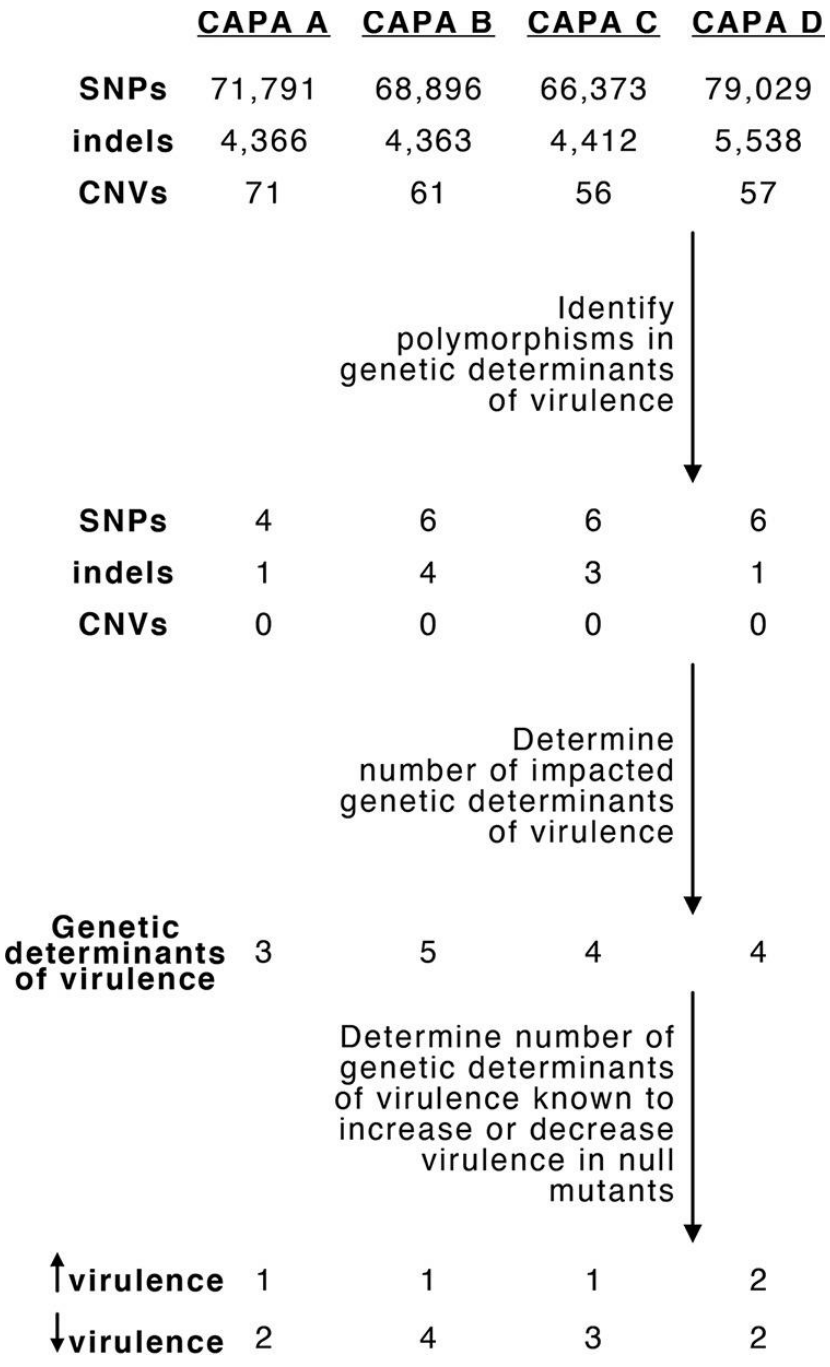

Figure 20. Mutational spectra among genetic determinants of virulence.
Genome-wide SNPs, indels, and CN variants were filtered for those present in genetic determinants of virulence. Then, the number of genetic determinants of virulence with high-impact polymorphisms was identified. The number known to increase or decrease virulence in null mutants was determined thereafter.

---

increased virulence in a mouse model of disease (O'Hanlon et al., 2011). Putative LOF mutations were also observed in genes whose null mutants decreased virulence. For example, all CAPA isolates shared the same SNPs resulting in early stop codons that likely result in LOF or partial LOF in *pptA* (AFUA_2G08590), a putative 4'-phosphopantetheinyl transferase, whose deletion results in reduced virulence in a mouse model of disease (Johns et al., 2017). In light of the close evolutionary relationships among CAPA isolates, we hypothesize that these shared mutations likely occurred in the genome of their most recent common ancestor.

In addition to shared polymorphisms, analyses of CAPA isolate genomes also revealed isolate-specific polymorphisms affecting genetic determinants of virulence (File S2 from Steenwyk et al. 2021d). For example, SNPs resulting in early stop codons, which likely lead to LOF, were observed in *CYP5081A1* (AFUA_4G14780), a putative cytochrome P450 monooxygenase, in CAPA isolates B and C. *CYP5081A1* LOF is associated with reduced virulence of *A. fumigatus* (Mitsuguchi et al., 2009). Other SNPs are found only in single isolates. CAPA isolate B has a frameshift mutation in a putative fatty acid oxygenase (AFUA_4G00180). CAPA isolate D has a mutation resulting in the loss of the start codon in *fleA* (AFUA_5G14740), a gene that encodes an L-fucose-specific lectin. Mice infected with *FLEA* null mutants have more severe pneumonia and invasive aspergillosis than wild-type strains. *FLEA* null mutants cause more severe disease because FleA binds to macrophages and therefore is critical for host recognition, clearance, and macrophage killing (Kerr et al., 2016). The only evidence of pseudogenization among the genetic

determinants of virulence in the reference strains was observed in *mybA* (AFUA_3G07070), a transcription factor involved in conidiation and conidial viability (Valsecchi et al., 2017), in strain A1163. *MybA* null mutants have reduced virulence compared to wild type strains (Valsecchi et al., 2017).

Examination of three additional clinical strains of *A. fumigatus* (IFM61407, CN-CM7555, and Afs35) revealed that CAPA isolates shared some, but not all, polymorphisms present in the 206 genetic determinants of virulence. For example, similar putative LOF mutations were observed in *NRPS8* (AFUA_5G12730). Polymorphisms that were not shared between CAPA isolates and the three clinical strains include the loss of a stop codon in *aspA*, a septin (Vargas-Muñiz et al., 2015), in Afs35; an early stop codon in *noc3*, a nuclear export protein (Hu et al., 2007), in CN-CM7555; and a lost stop codon in *cat2*, a bifunctional catalase-peroxidase (Paris et al., 2003), in IFM61407. A complete list of high impact polymorphisms in the three clinical strains are available in File S2 from Steenwyk et al. 2021d.

Examination of the presence of biosynthetic gene clusters (BGCs) revealed that all CAPA isolates had BGCs that encode secondary metabolites known to modulate host biology (Table 2 from Steenwyk et al. 2021d). For example, all CAPA isolates had BGCs encoding the toxic secondary metabolite gliotoxin (Fig. 21). Other intact BGCs in the genomes of the CAPA isolates include fumitremorgin, trypacidin, pseurotin, and fumagillin, which are known to modulate host biology (Ishikawa et al., 2009; González-Lobato et al., 2010; Gauthier et al., 2012); for example, fumagillin is known to inhibit neutrophil function (Fallon et al., 2010, 2011).

More broadly, all CAPA isolates had similar numbers and classes of BGCs (Fig. S3 from Steenwyk et al. 2021d).

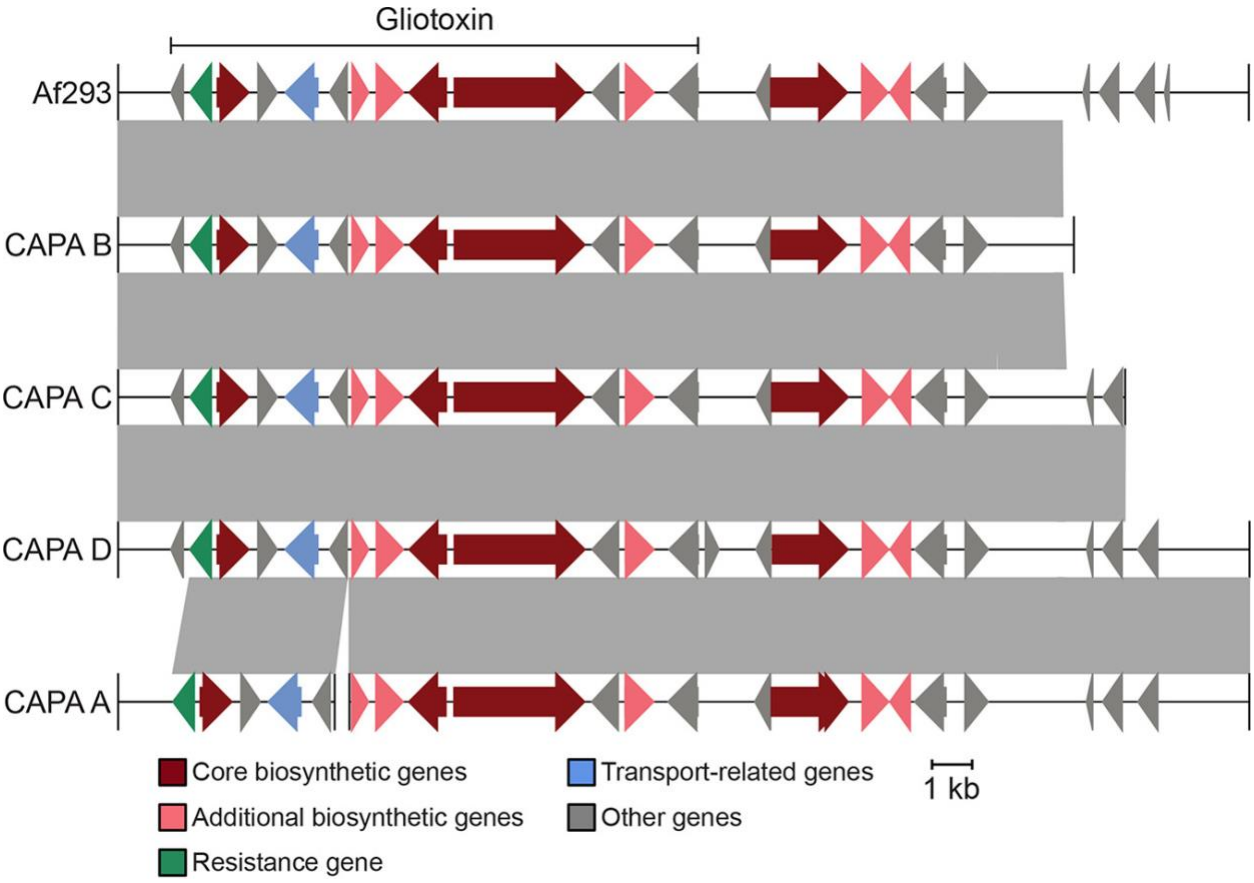

Figure 21. COVID-19-associated pulmonary aspergillosis (CAPA) isolates of *Aspergillus fumigatus* have biosynthetic gene clusters (BGCs) that encode the toxic small molecule gliotoxin.
Gliotoxin is known to contribute to virulence of *A. fumigatus*. The genomes of CAPA isolates of *A. fumigatus* contain biosynthetic gene clusters known to encode gliotoxin. Note, the BGC of CAPA isolate A was split between two contigs and, therefore, the BGC is hypothesized to be present.

In summary, we found that CAPA isolates were closely related to one another and had largely intact genetic determinants of virulence and BGCs. However, we observed strain-specific

polymorphisms in known genetic determinants of virulence in CAPA isolate genomes, which raises the hypothesis that CAPA isolates differ in their virulence profiles.

CAPA isolates display strain heterogeneity in virulence and in a few virulence-related traits

Examination of virulence and virulence-related traits revealed the CAPA isolates often, but not always, had similar phenotypic profiles compared to reference *A. fumigatus* strains Af293 and a CEA17 Δ*akuB*^{KU80} pyrG+ derivative of CEA17 *akuB*$^{KU80+}$, pyrG- (which is hereafter referred to as CEA17 for simplicity (Bertuzzi et al., 2020)). For example, virulence in the *Galleria* moth model of fungal disease revealed strain heterogeneity among CAPA isolates, Af293, CEA17, and a panel of three clinical strains of *A. fumigatus*, namely Afs35, IFM61407, and CN-CM7555 (Takahashi-Nakaguchi et al., 2015; Garcia-Rubio et al., 2018; Bertuzzi et al., 2020) ($p < 0.001$; log-rank test; Fig. 22A). Pairwise examination revealed that the observed strain heterogeneity

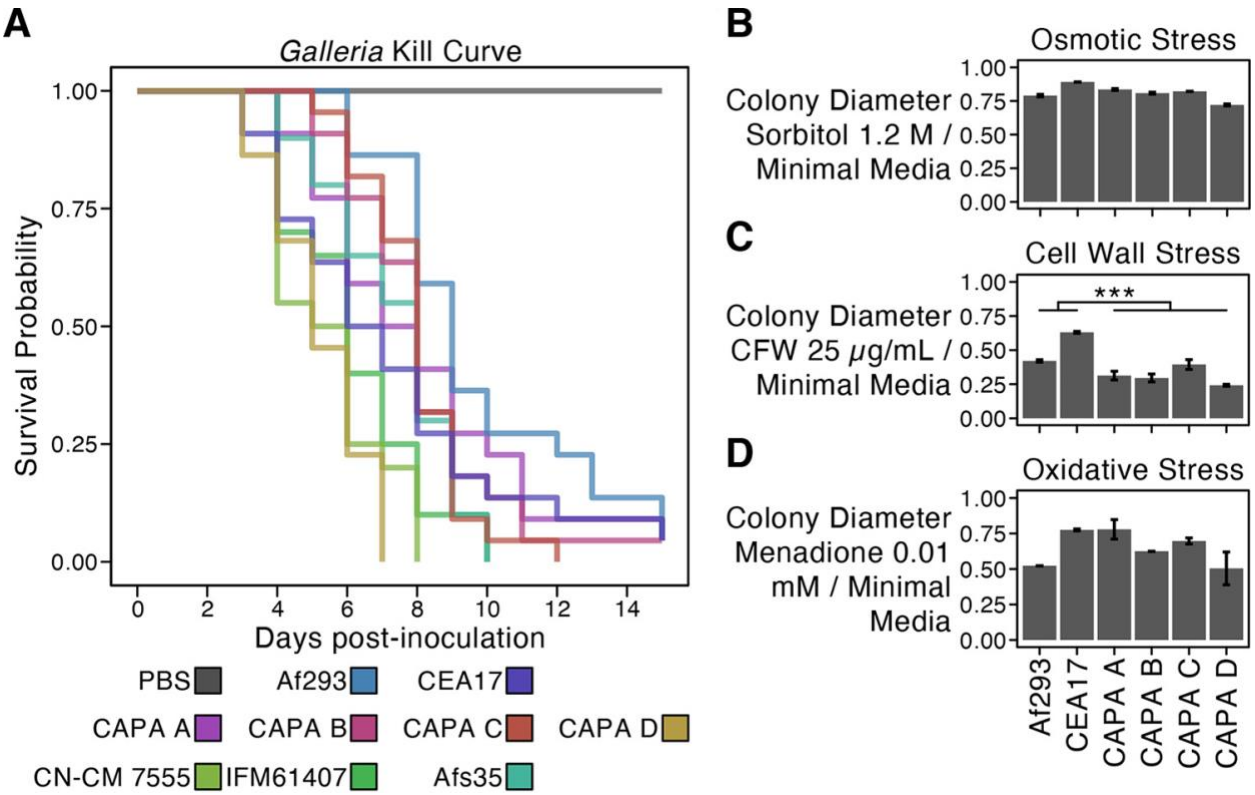

Figure 22. Strain heterogeneity among COVID-19-associated pulmonary aspergillosis (CAPA) isolates of *Aspergillus fumigatus*.

(A) The virulence of the CAPA isolates, reference strains Af293 and CEA17, and clinical strains Afs35, CN-CM7555, and IFM61407 significantly varied in the *Galleria* moth model of disease ($P < 0.001$; log-rank test; ≥20 replicates per strain). Pairwise examinations revealed CAPA isolate D was significantly more virulent than all other strains (Benjamini-Hochberg adjusted $P < 0.01$ when comparing CAPA isolate D to another isolate; log-rank test) with the exception of clinical strains IFM61407 and CN-CM 7555 ($P = 0.085$ and $P = 0.386$, respectively). Growth of CAPA isolates and references strains Af293 and CEA17 in the presence of osmotic (B), cell wall (C), and oxidative stressors (D). Growth differences between CAPA isolates and reference strains Af293 and CEA17 were observed across all growth conditions ($P < 0.001$; multifactor ANOVA). Pairwise differences were assessed using the *post hoc* Tukey's honestly significant difference (HSD) test and were only observed for growth in the presence of CFW at 25 μg/ml ($P < 0.001$; Tukey HSD test) in which the CAPA isolates did not grow as well as the reference isolates. To correct for strain heterogeneity in growth rates, radial growth in centimeters in the presence of stressors was divided by radial growth in centimeters in the absence of the stressor (MM only). Abbreviations of cell wall stressors are as follows: CFW, calcofluor white; CR, Congo red; CSP, caspofungin. Growth in the presence of other stressors is summarized in Fig. S4. Error bars in panels B to D represent the average of ± one standard deviation across three replicates.

---

was primarily driven by CAPA isolate D, which was significantly more virulent than all other

CAPA isolates, reference strains Af293 and CEA17, and clinical strain Afs35 (Benjamini-

Hochberg adjusted p-value < 0.05 when comparing CAPA isolate D to another isolate; log-rank test; File S3 from Steenwyk et al. 2021d). However, the virulence of the CAPA isolate D was on par to those of clinical strains IFM61407 and CN-CM 7555 ($p = 0.085$ and 0.386, respectively) (Fig. 22A). These results reveal that the CAPA isolates have generally similar virulence profiles compared to the reference strains Af293 and CEA17 with the exception of the more virulent CAPA isolate D. Furthermore, the virulence profiles of all CAPA isolates are within the known range of *A. fumigatus* clinical strains. Determining the association between virulence and the genetic polymorphisms described in the section above in addition to other polymorphisms identified in this study (Fig 20) is an important future direction.

Examination of growth in the presence of osmotic, cell wall, and oxidative stressors revealed that the phenotypic profiles of CAPA isolates were similar to the profiles of Af293 and CEA17 strains (Fig. 22B-D and Fig. S4 from Steenwyk et al. 2021d). The sole exception was growth in the presence of calcofluor white, where we observed that the CAPA isolates were more sensitive than reference strains Af293 and CEA17 ($p < 0.001$; Tukey's Honest Significant Difference test; Fig. 22C). Lastly, antifungal drug susceptibility profiles for amphotericin B, voriconazole, itraconazole, and posaconazole were similar between the CAPA isolates and reference strains Af293 and CEA17 (Table 3 from Steenwyk et al. 2021d). Following the guidelines of the Clinical and Laboratory Standards Institute (CLSI, 2008), the CAPA isolates are not considered multidrug resistant.

Secondary metabolites can impact host biology and virulence (Sugui et al., 2007; Raffa and Keller, 2019). Examination of secondary metabolite production revealed strain heterogeneity

among CAPA isolates and reference strains Af293 and CEA10, a pyrG+ and *akuB*$^{KU80+}$ strain that CEA17 is derived from (Bertuzzi et al., 2020). For example, principal component analysis of chromatogram features revealed that CAPA isolate A was substantially different from other CAPA isolates along the first and second principal components, which capture 82.47% of the total variance, whereas the CAPA isolate D was substantially different from other CAPA isolates along the second and third principal components, which capture 38.64% of total variance (Fig. S5 from Steenwyk et al. 2021d). Examination of the loadings plot, which identifies the individual secondary metabolites that contribute to the observed variation across strains, revealed gliotoxin and fumitremorgin as the largest contributors (Fig. S6 from Steenwyk et al. 2021d). Measurement of relative abundance of biosynthesized gliotoxin and fumitremorgin, two secondary metabolites known to modulate host biology (Raffa and Keller, 2019), showed that the largest amount of fumitremorgin was biosynthesized by the CAPA isolate A, and the largest amount of gliotoxin was biosynthesized by the Af293 strain followed by the CAPA isolate C (Fig. S6 from Steenwyk et al. 2021d; Table 2 from Steenwyk et al. 2021d).

In summary, we found that the CAPA isolates have similar phenotypic profiles with the exception of growth in the presence of calcofluor white and secondary metabolite biosynthesis compared to reference strains, and virulence on par with the known range of *A. fumigatus* clinical strains.

Discussion

The effects of secondary fungal infections in COVID-19 patients are only beginning to be understood. Our results revealed that CAPA isolates are generally, but not always, similar to *A.*

fumigatus clinical reference strains. Notably, CAPA isolate D was significantly more virulent than the other three CAPA isolates and two reference strains examined, but on par with other clinical strains. Taken together, these results are important to consider in the management of fungal infections among patients with COVID-19, especially those infected with *A. fumigatus*, and broaden our understanding of CAPA.

CHAPTER 6

Extensive copy number variation in fermentation-related genes among *Saccharomyces cerevisiae* wine strains[5]

Introduction

Saccharomyces cerevisiae, commonly known as baker's or brewer's yeast, has been utilized by humans for the production of fermented beverages since at least 1,350 B.C.E. but may go as far back as the Neolithic period 7,000 years ago (Mortimer, 2000; Cavalieri et al., 2003). Phylogenetic analyses and archaeological evidence suggest wine strains originated from Mesopotamia (Bisson, 2012) and were domesticated in a single event around the same time as the domestication of grapes (Schacherer et al., 2009; Sicard and Legras, 2011). Further phylogenetic, population structure and identity-by-state analyses of single nucleotide polymorphism (SNP) data reveal close affinity and low genetic diversity among wine yeast strains across the globe, consistent with a domestication-driven population bottleneck (Liti et al., 2009; Schacherer et al., 2009; Sicard and Legras, 2011; Cromie et al., 2013; Borneman et al., 2016). These low levels of genetic diversity have led some to suggest that further wine strain development should be focused on introducing new variation into wine yeasts rather than exploiting their standing variation (Borneman et al. 2016).

Many wine strains have characteristic variants that have presumably been favored in the wine-

[5]This work is published in: Steenwyk, J., and Rokas, A. (2017). Extensive Copy Number Variation in Fermentation-Related Genes Among *Saccharomyces cerevisiae* Wine Strains. G3 Genes, Genomes, Genet. 7.

making environment (Marsit and Dequin, 2015). For example, adaptive point mutations, deletions and rearrangements in the promoter and coding sequence of *FLO11* contribute to flocculation and floating thereby increasing yeast cells' ability to obtain oxygen in the hypoxic environment of liquid fermentations (Fidalgo et al., 2006). Similarly, duplications of *CUP1* are strongly associated with resistance to copper (Warringer et al., 2011), which at high concentrations can cause stuck fermentations, and *THI5*, a gene involved in thiamine metabolism whose expression is associated with an undesirable rotten-egg sensory perception in wine, is absent or down regulated among wine strains and their derivatives (Bartra et al., 2010; Brion et al., 2014). As these examples illustrate, the mutations underlying these, as well as many other, presumably adaptive traits are not only single nucleotide polymorphisms (SNPs), but also genomic structural variants, such as duplications, insertions, inversions, and translocations (Pretorius, 2000; Marsit and Dequin, 2015).

Copy number (CN) variants, a class of structural variants defined as duplicated or deleted loci ranging from 50 bp to whole chromosomes (Zhang et al., 2009; Arlt et al., 2014), have recently started receiving considerable attention due to their widespread occurrence (Sudmant et al., 2010; Bickhart et al., 2012; Axelsson et al., 2013; Pezer et al., 2015) as well as their influence on gene expression and phenotypic diversity (Freeman et al., 2006; Henrichsen et al., 2009). Mechanisms of CN variant evolution include non-allelic homologous recombination (Lupski and Stankiewicz, 2005) and retrotransposition (Kaessmann et al., 2009). CN variants are well studied in various mammals, including humans (*Homo sapiens*; Sudmant et al. 2015), cattle (*Bos taurus*; Bickhart et al. 2012), the house mouse (*Mus musculus*; Pezer et al. 2015), and the domestic dog (*Canis lupus familiaris*; Axelsson et al. 2013), where they are important contributors to genetic

and phenotypic diversity.

Relatively few studies have investigated whole-genome CN profiles in fungi (Hu et al., 2011; Farrer et al., 2013; Steenwyk et al., 2016). For example, the observed CN variation of chromosome 1 in the human pathogen *Cryptococcus neoformans* results in the duplications of *ERG11*, a lanosterol-14-α-demethylase and target of the triazole antifungal drug fluconazole (Lupetti et al., 2002), and *AFR1*, an ATP binding cassette (ABC) transporter (Sanguinetti et al., 2006), leading to increased fluconazole resistance (Sionov et al., 2010). Similarly, resistance to itraconazole, a triazole antifungal drug, is attributed to the duplication of cytochrome P-450-depdendent C-14 lanosterol α-demethylase (*pdmA*) – a gene whose product is essential for ergosterol biosynthesis – in the human pathogen *Aspergillus fumigatus* (Osherov et al., 2001). Finally, in the animal pathogen *Batrachochytrium dendrobatidis*, the duplication of Supercontig V is associated with increased fitness in the presence of resistance to an antimicrobial peptide, although the underlying genetic elements involved remain elusive (Farrer et al., 2013).

Similarly understudied is the contribution of CN variation to fungal domestication (Gibbons and Rinker 2015; Gallone et al. 2016). Notable examples of gene duplication being associated with microbial domestication include those of α-amylase in *Aspergillus oryzae*, which is instrumental in starch saccharification during the production of sake (Hunter et al., 2011; Gibbons et al., 2012), and of the *MAL1* and *MAL3* loci in beer associated strains of *S. cerevisiae*, which metabolize maltose, the most abundant sugar in the beer wort (Gallone et al., 2016; Gonçalves et al., 2016). Beer strains of *S. cerevisiae* often contain additional duplicated genes associated with maltose metabolism, including *MPH2* and *MPH3*, two maltose permeases, and the putative

maltose-responsive transcription factor, *YPR196W* (Gonçalves et al., 2016). Adaptive gene duplication in *S. cerevisiae* has also been detected in experimentally evolved populations (Dunham et al., 2002; Gresham et al., 2008; Dunn et al., 2012). Specifically, duplication of the locus containing the high affinity glucose transporters *HXT6* and *HXT7* has been observed in adaptively evolved asexual strains (Kao and Sherlock, 2008) as well as in populations grown in a glucose-limited environment (Brown et al., 1998; Dunham et al., 2002; Gresham et al., 2008). Altogether, these studies suggest that CN variation is a significant contributor to *S. cerevisiae* evolution and adaptation.

To determine the contribution of CN variation to genome evolution in wine strains of *S. cerevisiae*, we characterized patterns of CN variation across the genomes of 132 wine strains and determined the functional impact of CN variable genes in environments reflective of wine-making. Our results suggest that there is substantial CN variation among wine yeast strains, including in gene families (such as *CUP, FLO, HXT* and *MAL*) known to be associated with adaptation in the fermentation environment. More generally, it raises the hypothesis that CN variation is an important contributor to adaptation during microbial domestication.

Materials and Methods

Data Mining, Quality Control and Mapping

Raw sequence data for 132 *Saccharomyces cerevisiae* wine strains were obtained from three studies (Borneman et al. 2016, 127 strains, Bioproject ID: PRJNA303109; Dunn et al. 2012, 2 strains, Bioproject ID: SRA049752; Skelly et al. 2013, 3 strains, Bioproject ID: PRJNA186707) (Figure S1 from Steenwyk and Rokas, 2017, File S1 from Steenwyk and Rokas, 2017).

Altogether, these 132 strains represent a diverse set of commercial and non-commercial isolates from the 'wine' yeast clade (Borneman et al., 2016).

Sequence reads were quality-trimmed using TRIMMOMATIC, version 0.36 (Bolger et al., 2014) with the following parameters and values: leading:10, trailing:10, slidingwindow:4:20, minlen:50. Reads were then mapped to the genome sequence of the *S. cerevisiae* strain S288c (annotation release: R64.2.1; http://www.yeastgenome.org/) using BOWTIE2, version 1.1.2 (Langmead and Salzberg, 2012) with the 'sensitive' parameter on. For each sample, mapped reads were converted to the bam format, sorted and merged using SAMTOOLS, version 1.3.1. Sample depth of coverage was obtained using the SAMTOOLS depth function (Li et al., 2009a).

CN Variant Identification

To facilitate the identification of single nucleotide polymorphisms (SNPs), we first generated mpileup files for each strain using SAMTOOLS, version 1.3.1 (Li et al., 2009a). Using the mpileup files as input to VARSCAN, version 2.3.9 (Koboldt et al., 2009, 2012), we next identified all statistically significant SNPs (Fisher's Exact test; $p < 0.05$) present in the 132 strains that had a read frequency of at least 0.75 and minimum coverage of 8X. This step enabled us to identify 149,782 SNPs. By considering only SNPs that harbored a minor allele frequency of at least 10%, we retained 43,370 SNPs. These SNPs were used to confirm the evolutionary relationships among the strains using Neighbor-Net phylogenetic network analyses in SPLITSTREE, version 4.14.1 (Huson, 1998) as well as the previously reported low levels of SNP diversity (Figure S2 from Steenwyk and Rokas, 2017; Borneman *et al.* 2016).

To detect and quantify CN variants we used CONTROL-FREEC, version 9.1 (Boeva et al., 2011, 2012), which we chose because of its low false positive rate and high true positive rate (Duan et al., 2013). Importantly, the average depth of coverage or read depth of the 132 strains was 30.1 ± 14.7X (minimum: 13.0X, maximum: 104.5X; Figure S3 from Steenwyk and Rokas, 2017), which is considered sufficient for robust CNV calling (Sims et al., 2014).

CONTROL-FREEC uses LOESS modeling for GC-bias correction and a LASSO-based algorithm for segmentation. Implemented CONTROL-FREEC parameters included window = 250, minExpectedGC = 0.35, maxExpectedGC = 0.55 and telocentromeric = 7000. To identify statistically significant CN variable loci ($p < 0.05$), we used the Wilcoxon Rank Sum test. The same CONTROL-FREEC parameters, but with a window size of 25 base pairs (bp), were used to examine CN variation within the intragenic Serine/Threonine-rich sequences of *FLO11* (Lo and Dranginis, 1996). BEDTOOLS, version 2.25 (Quinlan and Hall, 2010) was used to identify duplicated or deleted genic loci (i.e., CN variable loci) that overlapped with genes by at least one nucleotide. The CN of each gene (genic CN) was then calculated as the average CN of the 250 bp windows that overlapped with the gene's location coordinates in the genome. The same method was used to determine non-genic CN for loci that did not overlap with genes (ie., non-genic CN variable loci). To identify statistically significant differences between CN variable loci that were duplicated versus those that were deleted, we employed the Mann-Whitney *U* test (Wilcoxon rank-sum test) with continuity correction (Wallace, 2004).

Diversity in CN Variation and GO Enrichment

To identify CN diverse loci we used two different measures. The first measure calculates the statistical variance (s^2) for each locus where CN variants were identified in one or more strains. s^2 values were subsequently $\log_{10}$ normalized. $\text{Log}_{10}(s^2)$ accounts for diversity in raw CN values but not for diversity in CN allele frequencies. Thus, we also employed a second measure based on the Polymorphic Index Content (PIC) algorithm, which has previously been used to identify informative microsatellite markers for linkage analyses by taking into account both the number of alleles present and their frequencies (Keith et al., 1990; Risch, 1990). PIC has also been used to quantify population-level diversity of simple sequence repeat loci and restriction fragment length polymorphisms in maize (Smith et al., 1997). PIC values were calculated for each locus harboring at least one CN variant based on the following formula:

$$PIC = 1 - \sum_{i=a}^{z} i^2$$

where i^2 is the squared frequency of *a* to *z* CN values (Smith et al. 1997). PIC values may range from 0 (no CN diversity) to 1 (all CN alleles are unique).

To create a list of loci exhibiting high CN diversity for downstream analyses, we retained only those loci that fell within the 50th percentile of $\log_{10}(s^2)$ values (min = -2.12, median = -1.02, and max = 2.40) or the 50th percentile of PIC values (min = 0.02, median = 0.14, and max = 0.96).

Genes overlapping with loci exhibiting high CN diversity were used for Gene ontology (GO) enrichment analysis with AMIGO2, version 2.4.24 (Carbon et al., 2009) using the PANTHER Overrepresentation Test (release 20160715) with default settings. This test uses the PANTHER Gene Ontology database, version 11.0 (Thomas et al. 2003; release date 2016-07-15) which is

directly imported from the GO Ontology database, version 1.2 (GeneOntologyConsortium 2004; release date 2016-10-27), a reference gene list from *S. cerevisiae*, and a Mann-Whitney *U* test (Wilcoxon rank-sum test) with Bonferroni multi-test corrected *p*-values to identify over- and under-represented GO terms (Mi et al., 2013). Statistical analyses and figures were created using PHEATMAP, version 1.0.8 (Kolde, 2012), GPLOTS, version 3.0.1, GGPLOT2 (Wickham, 2009) or standard functions in R, version 3.2.2 (R Development Core Team, 2008).

Identifying Loci Absent in the Reference Strain

To identify loci absent from the reference strain but present in other strains, we assembled unmapped reads from the 20 strains with the lowest percentage of mapped reads. The percentage of mapped reads was determined using SAMTOOLS (Li et al., 2009a); its average across strains was 96% (min = 70.5% and max = 99%; Figure S4 from Steenwyk and Rokas, 2017). Unmapped reads from the 20 strains with the lowest percentage of mapped reads were assembled using SPADES, version 3.8.1 (Bankevich et al., 2012). The identity of scaffolds longer than the average length of a *S. cerevisiae's* gene (~1,400 bp) was determined using blastx from NCBI's BLAST, version 2.3.0 (Madden, 2013) against a local copy of the GenBank non-redundant protein database (downloaded on January 5, 2017).

Results

Descriptive Statistics of CN variation

To examine CN variation across wine yeasts, we generated whole genome CN profiles for 132 strains (Figure S5 from Steenwyk and Rokas, 2017, File S2 from Steenwyk and Rokas, 2017). Across all strains, we identified a total of 2,820 CNVRs that overlapped with 2,061 genes and

spanned 3.7 megabases (Mb). The size distribution of CNVRs was skewed toward CN variants that were shorter than 1 kilobase (kb) in length (Figure 23A, Figure S6A from Steenwyk and Rokas, 2017 & Table S1 from Steenwyk and Rokas, 2017). Strains had an average of 97.8 ± 9.5 CNVRs (median = 86) (Figure S6B from Steenwyk and Rokas, 2017) that affected an average of 4.3% ± 0.1% of the genome (median = 4.1%) (Figure S6C from Steenwyk and Rokas, 2017).

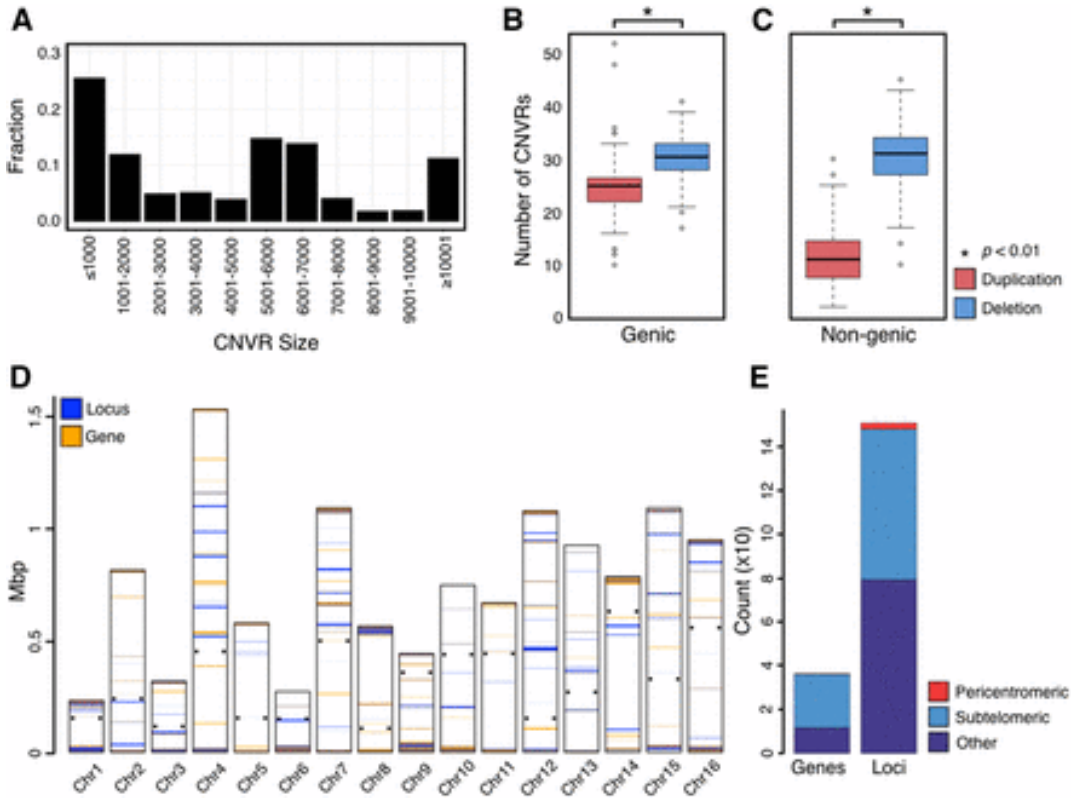

Figure 23. Size distribution and location of CN variable loci.
(A) The fraction of CNVRs (*y*-axis) for a given size range. Most CNVRs are ≤1000 bp. (B, C) Deleted genic (B) and nongenic (C) CNVRs are more prevalent than duplicated ones ($P < 0.01$ for both comparisons). Note that 26.23% of genic duplications occurred in multiples of three. (D) Location of CN variable loci across the 16 yeast chromosomes. The small, black squares on either side of each chromosome denote centromere location. Chromosomes are oriented with the start of the chromosome on the bottom and the end on top. Loci (blue bars) and genes (orange bars) harboring high $\log_{10}(s^2)$ or PIC values are shown. (E) A total of 684 of the 1502 CN diverse loci and 243 of the 363 CN diverse genes reside in subtelomeric regions of the yeast genome; in contrast, very few are found in pericentromeric regions (28 loci and three genes).

Due to the known influence of CN variable genes (Henrichsen et al., 2009; Orozco et al., 2009), we next quantified the number of genic and non-genic CNVRs (Figure 23B and C). We found statistically significant differences in the number of duplicated and deleted loci that are genic or non-genic (Mann-Whitney U test; $p < 0.01$ for both genic and non-genic comparisons) revealing that there were significantly more deleted genic and non-genic CNVRs than duplicated ones.

CN Diversity in Subtelomeres

To identify loci that exhibited high CN diversity, we retained only those loci that fell within the 50th percentile of at least one of our two different measures ($\log_{10}(s^2)$ and PIC) across the 132 strains. The distributions of the two measures (Figure S7 from Steenwyk and Rokas, 2017) were similar, with 1,326 loci (Figure S7C from Steenwyk and Rokas, 2017) and 291 genes (Figure S7D from Steenwyk and Rokas, 2017) identified in the top 50% of CN diverse genes by both measures.

In addition, the $\log_{10}(s^2)$ measure identified an additional 85 loci and 54 genes in its set of top 50% genes, and PIC an additional 85 loci and 18 genes. In total, our analyses identified 1,502 loci and 363 genes showing high CN diversity. Among the genes harboring the highest $\log_{10}(s^2)$ and PIC values were *YLR154C-G* (PIC = 0.96; $\log_{10}(s^2) = 2.16$), *YLR154W-A* (PIC = 0.96; $\log_{10}(s^2) = 2.16$), *YLR154W-B* (PIC = 0.96; $\log_{10}(s^2) = 2.16$), *YLR154W-C* (PIC = 0.96; $\log_{10}(s^2) = 2.16$), *YLR154W-E* (PIC = 0.96; $\log_{10}(s^2) = 2.16$), *YLR154W-F* (PIC = 0.96; $\log_{10}(s^2) = 2.16$) and *YLR154C-H* (PIC = 0.93; $\log_{10}(s^2) = 2.40$); these genes are all encoded within the 25S rDNA or 35S rDNA locus. The rDNA locus is known to be highly CN diverse (Gibbons et al., 2015) thereby demonstrating the utility and efficacy of our CN calling protocol as well as our two

measures of CN diversity. We next generated CN diversity maps for all 16 *S. cerevisiae* chromosomes (Figure 23D; Figure S8 from Steenwyk and Rokas, 2017). CN diversity was higher in loci and genes located in subtelomeres (defined as the 25 kb of DNA immediately adjacent to the chromosome ends; Barton *et al.* 2003). Specifically, 684 / 1,502 (45.5%) of CN diverse loci and 243 / 363 (66.9%) CN diverse genes were located in the subtelomeric regions. Conducting the same analysis using an alternative definition of subtelomere (defined as the DNA between the chromosome's end to the first essential gene (Winzeler et al., 1999)) showed similar results. Specifically, 721 / 1,502 (48%) of CN diverse loci and 233 / 363 (64.2%) of CN diverse genes were located in the subtelomeric regions.

GO Enrichment of CN Diverse Genes

To determine the functional categories over- and under-represented in the 363 genes showing high CN diversity, we performed GO enrichment analysis. The majority of enriched GO terms were associated with metabolic functions such as α-GLUCOSIDASE ACTIVITY ($p < 0.01$) and CARBOHYDRATE TRANSPORTER ACTIVITY ($p < 0.01$) (Figure 24 and File S3 from Steenwyk and Rokas, 2017).

Genes associated with these GO terms include *SUC2* (*YIL162W*, involved in hydrolyzing sucrose), all six members from the *MAL* gene family (involved in the fermentation of maltose and other carbohydrates) and all five members of the *IMA* gene family (involved in isomaltose, sucrose and turanose metabolism). Other enriched categories were associated with multi-cellular processes such as the FLOCCULATION ($p < 0.01$) and AGGREGATION OF UNICELLULAR ORGANISMS ($p = 0.03$). All members of the *FLO* gene family (involved in flocculation) and *YHR213W* (a

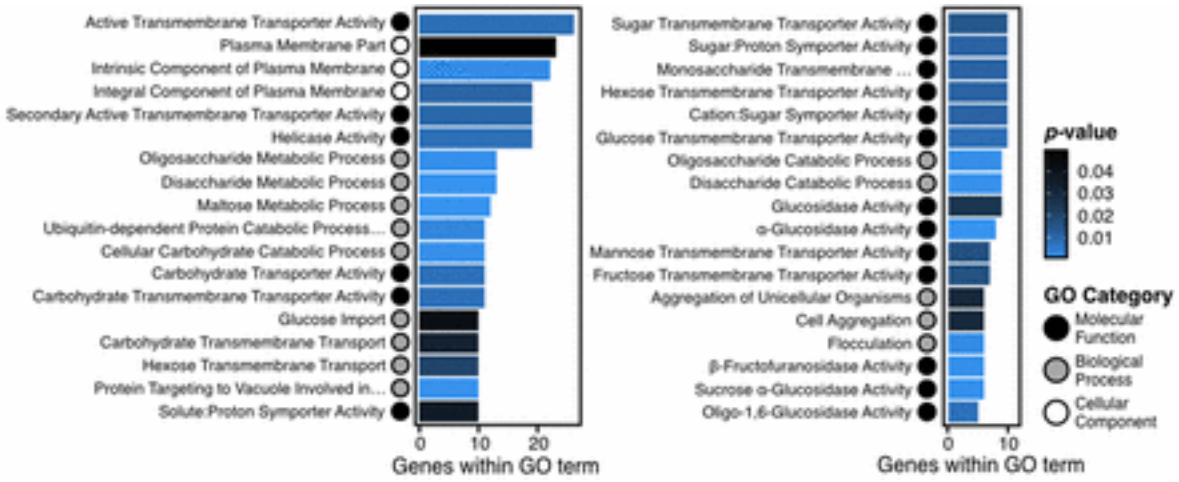

Figure 24. GO enriched terms from high CN diverse genes.
Molecular function (black), biological process (gray), and cellular component (white) GO categories are represented by circles, and are enriched among the 363 genes that overlap with CN diverse loci. Enriched terms are primarily related to metabolic function, such as α-GLUCOSIDASE ACTIVITY ($P < 0.01$), CARBOHYDRATE TRANSPORTER ACTIVITY ($P < 0.01$) and FLOCCULATION ($P < 0.01$).

flocculin-like gene) were associated with these GO enriched terms.

Contrary to overrepresented GO terms, underrepresented terms were associated with genes whose protein products are part of the interactome or protein-protein interactions such as PROTEIN COMPLEX ($p < 0.01$), MACROMOLECULAR COMPLEX ASSEMBLY ($p = 0.03$), TRANSFERASE COMPLEX ($p < 0.01$) and RIBONUCLEOPROTEIN COMPLEX BIOGENESIS ($p = 0.04$). Our finding of underrepresented GO terms being associated with multi-unit protein complexes supports the gene balance hypothesis, which states that the stoichiometry of genes contributing to multi-subunit complexes must be maintained to conserve kinetics and assembly properties (Birchler and Veitia, 2010, 2012). Thus, genes associated with multi-unit protein complexes are unlikely to exhibit CN variation.

Genic CN Diversity

To further understand the structure of CN variation in highly diverse CN genes, we first calculated the absolute CN of 23 genes associated with GO enriched terms related to wine fermentation processes (e.g., metabolic functions; Figure 24 and File S3 from Steenwyk and Rokas, 2017) as well as 57 genes with the highest PIC or $\log_{10}(s^2)$ values (Figure S9 from Steenwyk and Rokas, 2017 and File S4 from Steenwyk and Rokas, 2017; 69 total unique genes). Among these 69 genes, gene CN ranged from 0 to 92; both the highest CN diversity and absolute CN values were observed in segments of the rDNA locus (mentioned above).

Importantly, 35 of the 69 genes have also been reported to have functional roles in fermentation-related processes. For example, the CNs of *PAU3* (*YCR104W*), a gene active during alcoholic fermentation, and its gene neighbor *ADH7* (*YCR105W*), an alcohol dehydrogenase, both varied between 0 and 3. Similarly, the absolute CN of the locus containing both *CUP1-1* (*YHR053C*; PIC = 0.868) and its paralog *CUP1-2* (*YHR055C*; PIC = 0.879) ranged from 0-14 (Figure 25; File S4 from Steenwyk and Rokas, 2017), with 90 strains (68.2%) showing duplications (i.e., a CN greater than 1) and another 11 strains (8.3%) a deletion (i.e., a CN of 0). Interestingly, multiple copies of *CUP1* confer copper resistance to wine strains of *S. cerevisiae*, with CN variation at this locus thought to be associated with domestication (Warringer et al., 2011; Marsit and Dequin, 2015).

The expression of *SNO* family members is induced just prior to or after the diauxic shift as a response to nutrient limitation and is associated with vitamin B acquisition (Padilla et al., 1998; Rodríguez-Navarro et al., 2002). We found that *SNO2* (*YNL334C*) and *SNO3* (*YFL060C*) were

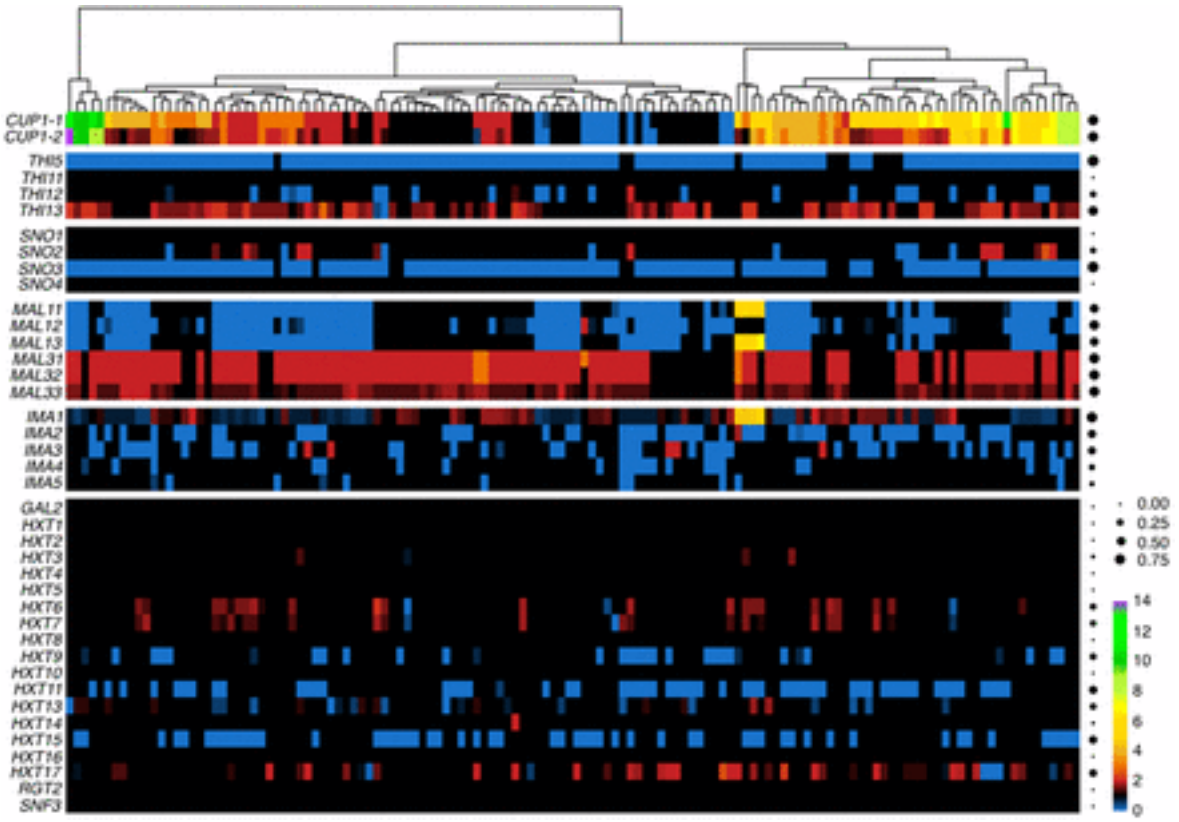

Figure 25. CN variation of genes and gene families.
Heat map of the CN profiles the *CUP*, *THI*, *SNO*, *MAL*, *IMA*, and *HXT* gene families; rows correspond to genes and columns to strains. Blue-colored cells correspond to deletions, black-colored cells to no CN variation and red-to-purple-colored cells to duplications (ranging from 2 to 14). Dots on the right side of the figure represent the proportion of individual strains that harbor CN variation in that gene—the larger the dot, the greater the proportion of the strains that is CN variable for that gene.

among the 363 genes with highest CN diversity. *SNO2* was duplicated in 14 strains (10.6%) and deleted in 9 strains (6.8%), while *SNO3* was deleted in 117 strains (88.6%) (Figure 25). The other two members of the SNO gene family, *SNO1* (*YMR095C*) and *SNO4* (*YMR322C*), both showed a CN of 1 in all strains.

Another gene family whose members show high CN diversity is the *THI* gene family, which is responsible for thiamine metabolism and is activated at the end of the growth phase during

fermentation (Brion et al., 2014). Specifically, *THI13* (*YDL244W;* PIC = 0.759) was among the 57 genes with the highest CN diversity (File S4 from Steenwyk and Rokas, 2017), and *THI5* (*YFL058W*) and *THI12* (*YNL332W*) among the 363 most CN diverse genes (File S3 from Steenwyk and Rokas, 2017). *THI13* was duplicated in 82 strains (62.1%) and deleted in 2 strains (1.5%) (Figure 25). In contrast, *THI5* was deleted in 121 strains (91.67%), whereas *THI12* was deleted in 23 strains (17.42%) and duplicated in only 3 strains (2.27%). Lastly, the CN of the last *THI* gene family member, *THI11* (*YJR156C*), did not exhibit CN variation.

In addition to the high CN diversity observed in all six members of the *MAL1* and *MAL3* loci responsible for maltose metabolism and growth on sucrose (Stambuk et al., 2000; Gallone et al., 2016), *MAL13* (*YGR288W*; PIC = 0.53) was among the 57 genes with the highest CN diversity (File S4). Evaluation of the absolute CN of all *MAL1* locus genes (Figure 25) showed that *MAL11*, *MAL12* (*YGR292W*), and *MAL13* were deleted in 65 (49.2%), 86 (65.2%), and 61 strains (46.2%), respectively. In contrast, the *MAL3* locus genes *MAL31* (*YBR298C*), *MAL32* (*YBR299W*), and *MAL33* (*YBR297W*) were duplicated in 100 (75.8%), 99 (75%), and 98 strains (74.2%), respectively. Interestingly, we did not observe any deletions in any of the *MAL3* locus genes across the 132 strains. When considering all members of the *MAL* gene family, we found that the 132 strains differed widely in their degree to which the locus had undergone expansion or contraction (Figure S10 from Steenwyk and Rokas, 2017).

All members of the *IMA* gene family, composed of genes aiding in sugar fermentation (Teste et al., 2010), were among the 363 genes with high CN diversity (File S3 from Steenwyk and Rokas, 2017) and *IMA1* (*YGR287C*; PIC = 0.87) was among the top 57 genes with the highest CN

diversity (Files S4 from Steenwyk and Rokas, 2017). *IMA1* was deleted in 54 strains (40.9%) and duplicated in 50 strains (37.9%) (Figure 25). Although many duplications or deletions did not span the entirety of *IMA1,* there were 4 strains that harbored high CNs between 4 and 6. These same four strains also had similar and unique duplications of *MAL11* and *MAL13*, suggesting that *IMA1*, *MAL11,* and *MAL13*, which are adjacent to each other in the genome, may have been duplicated as one locus. The other isomaltases (*IMA2-5*; *YOL157C*, *YIL172C*, *YJL221C* and *YJL216C*) were deleted in at least 11 strains (8.3%) and at most 55 strains (41.7%). No duplications in *IMA2*-5 were detected and only rarely in *IMA3* (5 strains, 3.8%). Altogether, the 132 strains exhibited both expansions and contractions of the *IMA* gene family (Figure S10 from Steenwyk and Rokas, 2017).

We identified 7 members of the *HXT* gene family (*HXT6*/*YDR343C*, *HXT7*/*YDR342C*, *HXT9/YJL219W*, *HXT11/YOL156W*, *HXT13/YEL069C*, *HXT15*/*YDL245C*, and *HXT17*/*YNR072W*), which is involved in sugar transport, that were among the 363 CN diverse genes (File S3 from Steenwyk and Rokas, 2017). Members of the *HXT* gene family were duplicated, deleted or had mosaic absolute CN values across the 132 strains. For example, *HXT6* and *HXT7* were primarily duplicated in 25 (18.9%) and 22 strains (16.7%), respectively, while only 3 strains (2.3%) had deletions in either gene (Figure 25). *HXT9*, *HXT11*, *HXT15* were deleted in 32 (24.2%), 57 (43.2%) and 53 strains (40.2%), respectively, while no strains had duplications. Finally, *HXT13* was duplicated in 12 strains (9.1%) and deleted in 17 strains (12.9%), and *HXT17* was duplicated in 37 strains (28%) and deleted in 9 strains (6.8%).

As expansions in the *HXT* gene family are positively correlated with aerobic fermentation in

Saccharomyces paradoxus and *S. cerevisiae* (Lin and Li, 2011), we also examined the absolute CN of all other 10 members (*GAL2/YLR081W, HXT1/YHR094C, HXT2/YMR011W, HXT4/YHR092C, HXT5/YHR096C, HXT8/YJL214W, HXT10/YFL011W, HXT16/YJR158W, RGT2/YDL138W,* and *SNF3/YDL194W*) of the *HXT* gene family (Figure 25). Interestingly, all remaining 10 members of the *HXT* gene family were not CN variable. Altogether, examination of the *HXT* family CN diversity patterns across the 132 strains suggests that wine yeast strains typically exhibit minor contractions (i.e., *HXT* gene deletions exceed those of duplications) relative to the S288c reference strain (Figure S10 from Steenwyk and Rokas, 2017).

All five members of the *FLO* gene family, which is responsible for flocculation (Govender et al., 2008), a trait shown to aid in the escape of oxygen limited environments during liquid fermentation (Fidalgo et al., 2006; Govender et al., 2008), were found to be among the 363 most CN diverse genes. Furthermore, *FLO5* (*YHR211W*; PIC = 0.82) and *FLO11* (*YIR019C*; PIC = 0.88) were among the 57 genes with the highest CN diversity (File S4 from Steenwyk and Rokas, 2017). Due to the importance of site directed CN variation in *FLO* family genes (Fidalgo et al., 2006), we modified our representation of CN variation to display intragenic CN variation using a 250 bp window (Figure S11 from Steenwyk and Rokas, 2017). *FLO5* was partially duplicated in 57 strains (43.2%), partially deleted in 47 strains (35.6%) and 115 strains (87.1%) had at least one region of the gene unaffected by CN variation. Duplications and deletions were primarily observed in the Threonine-rich region or Serine/Threonine-rich region located in the center or end of the *FLO5* gene, respectively. To better resolve intra-genic CN variation of *FLO11,* whose repeat unit is shorter than that of *FLO5*, we recalled CN variants with a smaller window size of 25 bp and re-evaluated CN variation (Figure S12 from Steenwyk and Rokas,

2017). Using this window size, we found extensive duplications in 97 strains (73.5%) between gene coordinates 250-350 bp. Furthermore, duplications were observed in the hydrophobic Serine/Threonine-rich regions (Figure S12 from Steenwyk and Rokas, 2017), which are associated with the flocculation phenotype (Fidalgo et al., 2006; Ramsook et al., 2010).

In contrast to *FLO5* and *FLO11*, other members of the *FLO* gene family did not exhibit intragenic CN variation. For example, CN variation in *FLO1* (*YAR050W*) and *FLO9* (*YAL063C*) typically spanned most or all of the sequence of each gene. Specifically, 125 strains (99.2%) had deletions spanning ≥80% of the gene in *FLO1* and only 2 strains (1.5%) had the entirety of the gene intact. *FLO9* had deletions in 99 strains (75%) that spanned ≥75% of the gene, 11 strains (8.3%) that had a partial deletion spanning <75% of the gene, whereas 1 strain (0.8%) had a CN of 2, and the remaining 21 strains (15.9%) had a CN of 1. In contrast, *FLO10* (*YKR102W*) showed limited CN variation. Specifically, 108 strains (81.8%) had no CN variation while 6 strains (4.5%) had deletions spanning the entirety of the gene. No duplications spanned the entirety of the gene but partial duplications were observed in 17 strains (12.9%) and were located in or just before the Serine/Threonine-rich region.

Functional Implications CN Variable Genes

To determine the functional impact of deleted CN variable genes, we examined the relative growth of deleted CN variable genes (denoted with the Δ symbol) relative to the wild-type (WT) *S. cerevisiae* strain S288c across 418 conditions using the Hillenmeyer et al. 2008 data (Figure S13 from Steenwyk and Rokas, 2017and File S5 from Steenwyk and Rokas, 2017). To determine the impact of duplicated genes, we examined growth fitness of the WT strain with low (~2-3

gene copies) or high plasmid CN (~8-24 gene copies), where each plasmid contained a single gene of interest from previously published data, relative to WT (Figure S14 from Steenwyk and Rokas, 2017 and File S6 from Steenwyk and Rokas, 2017; Payen et al. 2016).

Among deleted genes, 42 / 69 genes for which data exist showed negative and positive fitness effects in at least one tested condition in the S288c genetic background. Furthermore, we found that 12 / 42 genes that are commonly deleted among wine strains typically resulted in a fitness gain in conditions that resembled the fermentation environment. These conditions include growth at 23°C and at 25°C, temperatures within the 15-28°C range that wine is fermented in (Molina et al., 2007) and growth in minimal media, which is commonly used to understand fermentation-related processes (Seki et al., 1985; Govender et al., 2008; Vilela-Moura et al., 2008).

When examining fitness effects when grown at 23°C or at 25°C for 5 or 15 generations for the 12 commonly deleted genes, we observed at least one deletion that resulted in a fitness gain or loss for each condition. However, we observed extensive deletions in the *MAL1* locus (Figure 25) and therefore prioritized reporting the fitness impact of deletions in *MAL11*, *MAL12* and *MAL13*. *ΔMAL11* resulted in a fitness gain for growth at 23°C and 25°C for 5 (0.45x and 0.27x, respectively) and 15 generations (0.20x and 0.52x, respectively). *ΔMAL12* resulted in a gain of fitness at only 25°C after 15 generations (0.46x) and in a loss of fitness ranging from -0.36x to -1.29x in the other temperature conditions. Similarly, *ΔMAL13* resulted in fitness gains and losses dependent on the number of generations. For example, when grown for 15 generations at 25°C a fitness gain of 0.50x was observed while a fitness loss of -0.82x was observed at 23°C.

We next determined the fitness effect of deleted genes in minimal media after 0, 5, and 10 generations. Similar patterns of complex fitness gain and loss were observed as for the other conditions. For example, *ΔTHI12* resulted in a loss of fitness of -4.13x and -1.97x after 0 and 5 generations, but a fitness gain of 0.63x after 10 generations. In contrast, other genes resulted in positive fitness effects. For example, *ΔMAL12* resulted in a fitness gain of 7.25x and 10.41x for 0 generations and 10 generations.

Among duplicated genes, we focused on growth in glucose- and phosphate-limited conditions because glucose becomes scarce toward the end of fermentation prior to the diauxic shift and phosphate limitation is thought to contribute to stuck fermentations (Bisson, 1999; Marsit and Dequin, 2015). Among the 35 of the 69 genes where data were available, 14 genes had duplications among the 132 strains.

When examining fitness effects of duplicated genes in a glucose-limited environment in the S288c background, we found that fitness effects were small in magnitude and dependent on condition and plasmid CN (File S6). For example, *MAL32* low CN increased growth fitness by 0.02x but decreased fitness by -0.01x at a high CN (Figure S14 from Steenwyk and Rokas, 2017). Interestingly, the most prevalent CN for *MAL32* across the 132 strains was 2 (96 strains, 72.7%), with only 3 strains showing a CN of 3 and none a higher CN. Another gene found at low CN in 37 strains (28%) was *HXT17*. Low plasmid CN in a glucose-limited conditions resulted in a fitness gain of 0.06x. In contrast, *MAL13* low or high plasmid CN resulted in a negative growth fitness of -0.02x and -0.01x, respectively. Interestingly, *MAL13* duplication is only observed in 4 strains (3%) and deletions are observed in 61 strains (46.2%).

Similar to the glucose-limited condition, we found fitness was dependent on high or low plasmid CN in the phosphate-limited condition. For example, *MAL31*, a gene present at low CN in 100 strains had a fitness gain of 0.04x at high plasmid CN but low plasmid CN resulted in a fitness loss of -0.02x. In contrast, *MAL32*, which was present at low CN in 99 strains, had a small fitness gain of 0.002x at low plasmid CN and a fitness loss at a high plasmid CN of -0.02x. A total of 6 genes resulted in a disadvantageous growth effect when present at low CN, such as *DDR48*, which resulted in a fitness loss of -0.04x. Altogether, our results suggest that the deleted and duplicated CN variable genes we observe (Figure 26) modulate cellular processes that result in advantageous fitness effects in conditions that resemble the fermentation environment.

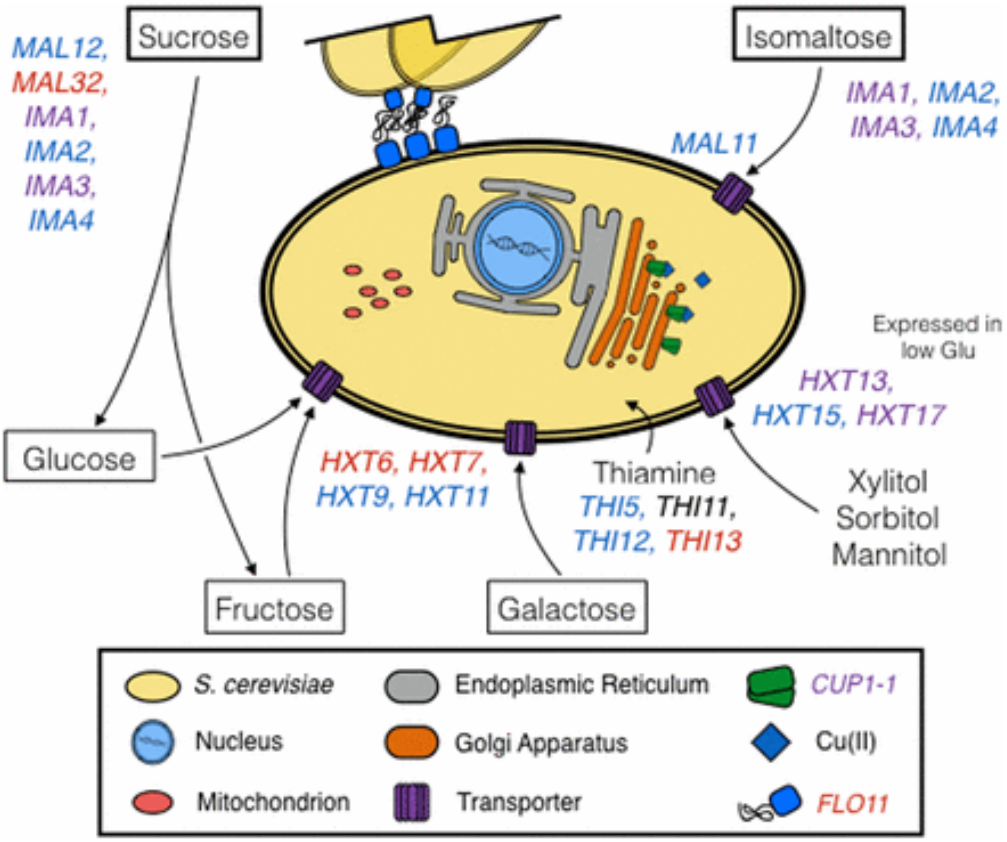

Figure 26. Model summary of CN variable genes in wine yeast strains and their cellular functions.
Genes that are deleted among wine strains are blue, whereas those that are duplicated are in red. Genes that were observed to be both duplicated and deleted (*IMA1*, *IMA3*, *HXT13*, and *CUP1-1*) are purple. Disaccharides are in thick-lined boxes, monosaccharides in thin-lined boxes, and alcohols are unboxed.

Identifying loci absent from CN variation analysis

The present study was able to capture loci represented in the WT/S288c laboratory strain. To identify loci absent from the reference strain, we assembled unmapped reads for 20 strains with the lowest percentage of reads mapped and determined their identity (see methods; Figure S4 from Steenwyk and Rokas, 2017). Across the 20 strains, we identified 429 loci absent from S288c but present in other sequenced *S. cerevisiae* strains. These loci had an average length of 6.9 kb and an average coverage of 107.2X. The 20 loci with the highest bitscore alongside with the number of strains containing the locus are shown in Table S2 from Steenwyk et al. 2021d. All but two of these loci were present only in one of the 20 strains we examined. The two exceptions were: the EC1118_1N26_0012p locus, which we found in 8 / 20 strains, which originates from horizontal gene transfer from *Zygosaccharomyces rouxii* to the commercial EC1118 wine strain of *S. cerevisiae* (Novo et al., 2009); and the EC1118_1O4_6656p locus, which we found in 7 / 20 strains. This locus was also originally found in the EC1118 strain (Novo et al., 2009) and contains a gene similar to a conserved hypothetical protein found in *S. cerevisiae* strain AWRI1631 (Borneman et al., 2008).

Discussion

CN variant loci are known to contribute to the genomic and phenotypic diversity (Perry et al.,

2007; Cutler and Kassner, 2008; Orozco et al., 2009). However, the extent of CN variation in wine strains of *S. cerevisiae* and its impact on phenotypic variation remains less understood. Our examination of structural variation in 132 yeast strains representative of the 'wine clade' showed that CN variants are a significant contributor to the genomic diversity of wine strains of *S. cerevisiae*. Importantly, CN variant loci overlap with diverse genes and gene families functionally related to the fermentation environment such as *CUP, FLO, THI, MAL, IMA* and *HXT* (summarized in Figure 26).

The characteristics of CN variation in wine yeast (Figure 23A; Figure S6 from Steenwyk and Rokas, 2017; Table S1 from Steenwyk and Rokas, 2017) were found to be similar to those of the recently described beer yeast lineage (Gallone et al., 2016). For example, both lineages exhibited a similar size range of CNVRs (Figure 23A; Figure S6 from Steenwyk and Rokas, 2017; Table S1 from Steenwyk and Rokas, 2017) as well as a higher prevalence of CNVRs in the subtelomeric regions (Figure 23D). However, wine strains had a smaller fraction of their genome affected by CN variation (Figure S6 from Steenwyk and Rokas, 2017) than beer strains (Gallone et al., 2016).

Wine yeast strains are thought to be partially domesticated due to the seasonal nature of wine-making, which allows for outcrossing with wild populations (Marsit and Dequin 2015; Gallone et al. 2016; Gonçalves et al. 2016). One human-driven signature of domestication is thought to be the duplication of the *CUP1* locus because multiple copies confer copper resistance and copper sulfates have been used to combat powdery mildews in vineyards since the early 1800s (Warringer et al., 2011; Marsit and Dequin, 2015). Consistent with this 'partial domestication'

view (Marsit and Dequin 2015; Gallone et al. 2016; Gonçalves et al. 2016), many wine strains were not CN variable for *CUP1-1* and *CUP1-2* or had one or both genes deleted (Figure 25).

An alternative, albeit not necessarily conflicting, hypothesis is that wine yeasts underwent domestication for specific but diverse wine flavor profiles (Hyma et al., 2011). Consistent with this view is the deletion (in >90% of the strains) of the *THI5* gene (Figure 25), whose activity is known to produce an undesirable rotten-egg sensory perception via higher SH_2 production and is associated with sluggish fermentations (Bartra et al., 2010). In contrast to wine strains, duplications of *THI5* have been observed across the *Saccharomyces* genus, including in several strains of *S. cerevisiae* (CBS1171, 2 copies; S288c, 4 copies; EM93, 5 copies), *S. paradoxus* (5 copies), and the lager brewing yeast hybrid *Saccharomyces pastorianus* (syn. *S. carlsbergensis*; 2+ copies) (Wightman and Meacock, 2003). In contrast, *THI13*, which is duplicated in 62.1% of strains, shows an increase in its expression 6-100-fold in *S. cerevisiae* when grown on medium containing low concentrations of thiamine allowing for the compensation of low thiamine levels (Li et al., 2010a). Low levels of thiamine in wine fermentation have been associated with stuck or slow fermentations (Ough et al., 1989; Bataillon et al., 1996). Similar to *THI5* deletions, *THI13* duplications may have also been driven by human activity due to the advantageous effect of increased expression within the fermentation environment.

Two other gene families subject to CN variation were the *MAL* and *HXT* gene families. The S288c strain that we used as a reference contained two *MAL* loci (*MAL1* and *MAL3*), each containing three genes – a maltose permease (*MALx1*), a maltase (*MALx2*), and an MAL trans-activator (*MALx3*) – and located near the ends of different chromosomes (Michels et al., 1992).

MAL1 has been observed to be duplicated in beer strains of *S. cerevisiae* (Gallone et al., 2016; Gonçalves et al., 2016) while wine strains primarily lack this locus (Figure 25; Gonçalves et al. 2016). In contrast to the deletion of the *MAL1* locus, *MAL3* duplication in wine yeasts (Figure 25; Gonçalves et al. 2016) is surprising because maltose is absent from the grape must (Gallone et al., 2016). However, knockout studies have demonstrated *MAL32* is necessary for growth on turanose, maltotriose, and sucrose (Brown et al., 2010), which are present in small quantities in wines (Victoria and Carmen 2013). Due to the prominent duplication of *MAL3*, in particular the enzymatic genes *MAL31* and *MAL32*, we speculate that the *MAL3* locus may be utilized to obtain sugars less prevalent in the wine environment or serve other purposes.

The *HXT* gene family in the S288c strain that we used as a reference contains 16 *HXT* paralogs, *GAL2*, *SNF3* and *RGT2*. The expansion of the HXT gene family is positively correlated with aerobic fermentation in *S. paradoxus* and *S. cerevisiae* (Lin and Li, 2011). *HXT6* and *HXT7* are high-affinity glucose transporters expressed at low glucose levels and repressed at high glucose levels (Reifenberger et al., 1995). In contrast to the recently described Asia (Sake), Britain (Beer) and Mosaic lineages (Gallone et al., 2016), we detected duplications in the *HXT6* and *HXT7* genes in wine yeasts (Figure 25). This may confer an advantage toward the end of fermentation and before the diauxic shift when glucose becomes a scarce resource. Evidence potentially supporting this hypothesis is that *HXT6* and *HXT7* are up-regulated by 9.8 and 5.6-fold, respectively, through wine fermentation in the *S. cerevisiae* strain Vin13 (Marks et al., 2008). Furthermore, *HXT6* or *HXT7* is found to be duplicated in experimentally evolved populations in glucose-limited environments (Dunham et al., 2002; Gresham et al., 2008; Dunn et al., 2012).

In summary, these results together with recent studies of CN variation in beer yeast strains (Gallone et al., 2016; Gonçalves et al., 2016), suggest that this type of variation significantly contributes to the genomic diversity of domesticated yeast strains. Furthermore, as most studies of CN variation, including ours, use reference strains, they are likely conservative in estimating the amount of CN variation present in populations. This caveat notwithstanding, examination of publically available data regarding the functional impact of duplicated or deleted genes (again in the context provided by the reference strain's genetic background) suggests that CN variation in several, but not all, of the wine yeast genes confer fitness advantages in conditions that resemble the fermentation environment. Our results raise the questions of the extent to which CN variation contributes to fungal, and more generally microbial, domestication as well as whether the importance of CN variants in natural yeast populations, including those of other *Saccharomyces* yeasts, is on par to their importance in domestication environments.

CHAPTER 7

Extensive loss of cell cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts[6]

Introduction

Genome maintenance is largely attributed to the fidelity of cell cycle checkpoints, DNA repair pathways, and their interaction (Lindahl, 1999). Dysregulation of these processes often leads to the loss of genomic integrity (Hakem, 2008) and hypermutation, or the acceleration of mutation rates (Broustas and Lieberman, 2014). For example, improper control of cell cycle and DNA repair processes can lead to 10- to 100-fold increases in mutation rate (Pal et al., 2007). Furthermore, deletions of single genes can have profound effects on genome stability. For example, the deletion of *MEC3*, which is involved in sensing DNA damage in the G1 and G2/M cell cycle phases, can lead to a 54-fold increase in the gross chromosomal rearrangement rate (Myung et al., 2001). Similarly, nonsense mutations in mismatch repair proteins account for the emergence of hypermutator strains in the yeast pathogens *Cryptococcus deuterogattii* (Billmyre et al., 2017) and *Cryptococcus neoformans* (Boyce et al., 2017; Rhodes et al., 2017a). Due to their importance in ensuring genomic integrity, most genome maintenance-associated processes are thought to be evolutionarily ancient and broadly conserved (Barnum and O'Connell, 2014).

One such ancient and highly conserved process in eukaryotes is the cell cycle

---

[6]This work is published in: Steenwyk, J. L., Opulente, D. A., Kominek, J., Shen, X.-X., Zhou, X., Labella, A. L., et al. (2019). Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLOS Biol. 17, e3000255. doi:10.1371/journal.pbio.3000255.

(Cross et al., 2011; Medina et al., 2016). Landmark features of cell cycle control include cell size control, the mitotic spindle checkpoint, the DNA damage response checkpoint, and DNA replication (Barnum and O'Connell, 2014). Cell size is controlled, in part, through the activity of *WHI5*, which represses the G1/S transition by inhibiting G1/S transcription (Costanzo et al., 2004). Similarly, when kinetochores are improperly attached or are not attached to microtubules, the mitotic spindle checkpoint helps to prevent activation of the anaphase-promoting complex (APC), which controls the G1/S and G2/M transitions (Castro et al., 2005; Barnum and O'Connell, 2014). Additional key regulators in this process are Mad1 and Mad2, which dimerize at unattached kinetochores and delay anaphase. Failure of Mad1:Mad2 recruitment to unattached kinetochores results in failed checkpoint activity (Heinrich et al., 2014). Importantly, many regulators, including but not limited to those mentioned here, are highly similar in structure and function between fungi and animals and are thought to have a shared ancestry (Cross et al., 2011). Interestingly, cell cycle initiation in certain fungi (including *Hanseniaspora*) is achieved through SBF, a transcription factor that is functionally equivalent but evolutionarily unrelated to E2F, the transcription factor that that initiates the cycle in animals, plants, and certain early-diverging fungal lineages (Medina et al., 2016). SBF is postulated to have been acquired via a viral infection, suggesting that evolutionary changes in this otherwise highly conserved process can and do rarely occur (Medina et al., 2016; Hendler et al., 2017).

DNA damage checkpoints can arrest the cell cycle and influence the activation of DNA repair pathways, the recruitment of DNA repair proteins to damaged sites, and the composition and length of telomeres (Zhou and Elledge, 2000). For example, *MEC3* and *RAD9*, function as checkpoint genes required for arrest in the G2 phase after DNA damage has occurred (Weinert et

al., 1994). Additionally, the deletions of DNA damage and checkpoint genes have been known to cause hypermutator phenotypes in the baker's yeast *Saccharomyces cerevisiae* (Serero et al., 2014). Similarly, hypermutator phenotypes are associated with loss-of-function mutations in DNA polymerase genes (Campbell et al., 2017). For example, deletion of the DNA polymerase δ subunit gene, *POL32*, which participates in multiple DNA repair processes, causes an increased mutational load and hypermutation in *S. cerevisiae*, in part, through the increase of genomic deletions and small indels (Huang et al., 2000; Serero et al., 2014). Likewise, the deletion of *MAG1*, a gene encoding a DNA glycosylase that removes damaged bases via the multi-step base excision repair pathway, can cause a 2,500-fold increased sensitivity to the DNA alkylating agent methyl methanesulfonate (Xiao et al., 2001).

In contrast to genes in multi-step DNA repair pathways, other DNA repair genes function individually or are parts of simpler regulatory processes. For example, *PHR1*, a gene that encodes a photolyase, is activated in response to and repairs pyrimidine dimers, one of the most frequent types of lesions caused by damaging UV light (Sebastian et al., 1990; Sebastian and Sancar, 1991). Other DNA repair genes do not interact with DNA but function to prevent the misincorporation of damaged bases. For example, *PCD1* encodes a 8-oxo-dGTP diphosphatase (Nunoshiba, 2004), which suppresses G → T or C → A transversions by removing 8-oxo-dGTP, thereby preventing the incorporation of the base 8-oxo-dG, one of the most abundant endogenous forms of an oxidatively damaged base (Cartwright et al., 2000; De Bont, 2004; Nunoshiba, 2004). Collectively, these studies demonstrate that the loss of DNA repair genes can lead to hypermutation and increased sensitivity to DNA damaging agents.

Hypermutation phenotypes are generally short-lived because most mutations are deleterious and are generally adaptive only in highly stressful or rapidly fluctuating environments (Ram and Hadany, 2012). For example, in *Pseudomonas aeruginosa* infections of cystic fibrosis patients (Oliver, 2000) and mouse gut-colonizing *Escherichia coli* (Giraud et al., 2001), hypermutation is thought to facilitate adaptation to the host environment and the evolution of drug resistance. Similarly, in the fungal pathogens *C. deuterogattii* (Billmyre et al., 2017), *C. neoformans* (Boyce et al., 2017; Rhodes et al., 2017a), and *Candida glabrata* (Healey et al., 2016), hypermutation is thought to contribute to within-host adaptation, which may involve modulating traits such as drug resistance (Healey et al., 2016; Billmyre et al., 2017). However, as adaptation to a new environment increases, hypermutator alleles are expected to decrease in frequency due to the accumulation of deleterious mutations that result as a consequence of the high mutation rate (Sniegowski et al., 1997; Taddei et al., 1997). In agreement with this prediction, half of experimentally evolved hypermutating lines of *S. cerevisiae* had reduced mutation rates after a few thousand generations (McDonald et al., 2012), suggesting hypermutation is a short-lived phenotype and that compensatory mutations can restore or lower the mutation rate. Additionally, this experiment also provided insights to how strains may cope with hypermutation; for example, all *S. cerevisiae* hypermutating lines increased their ploidy to presumably reduce the impact of higher mutation rates (McDonald et al., 2012). Altogether, hypermutation can produce short-term advantages but causes long-term disadvantages, which may explain its repeated but short-term occurrence in clinical environments (Giraud et al., 2001) and its sparseness in natural ones. While these theoretical and experimental studies have provided seminal insights to the evolution of mutation rate and hypermutation, we still lack understanding of the long-term, macroevolutionary effects of increased mutation rates.

Recently, multiple genome-scale phylogenies of species in the budding yeast subphylum Saccharomycotina showed that certain species in the bipolar budding yeast genus *Hanseniaspora* are characterized by very long branches (Riley et al., 2016; Shen et al., 2016b, 2018), which are reminiscent of the very long branches of fungal hypermutator strains (Billmyre et al., 2017; Boyce et al., 2017; Rhodes et al., 2017a). Most of what is known about these cosmopolitan yeasts relates to their high abundance on mature fruits and in fermented beverages (Albertin et al., 2016), especially on grapes and in wine must (Montero et al., 2004; Jordão et al., 2015). As a result, *Hanseniaspora* plays a significant role in the early stages of fermentation and can modify wine color and flavor through the production of enzymes and aroma compounds (Martin et al., 2018). Surprisingly, even with the use of *S. cerevisiae* starter cultures, *Hanseniaspora* species, particularly *Hanseniaspora uvarum,* can achieve very high cell densities, in certain cases comprising greater than 80% of the total yeast population, during early stages of fermentation (Langenberg et al., 2017), suggesting exceptional growth capabilities in this environment.

To gain insight into the long branches and the observed fast growth of *Hanseniaspora,* we sequenced and extensively characterized gene content and patterns of evolution in 25 genomes, including 11 newly sequenced for this study, from 18 / 21 known species in the genus. Our analyses showed that species in the genus *Hanseniaspora* lost many genes involved in diverse processes and delineated two lineages within the genus; a faster-evolving lineage (FEL), which has a strong signature of acceleration in evolutionary rate at its stem branch and has lost many additional genes involved in diverse processes, and a slower-evolving lineage (SEL), which has a weaker signature of evolutionary rate acceleration at its stem branch and underwent fewer gene

losses. Specifically, compared to *S. cerevisiae*, there are 748 genes that were lost from two-thirds of *Hanseniaspora* genomes with FEL yeasts having lost an additional 661 genes and SEL yeasts having lost only an additional 23. Relaxed molecular clock analyses estimate that the FEL and SEL split ~95 million years ago (mya). The degree of evolutionary rate acceleration is commensurate with the preponderance of loss of genes associated with cell cycle and DNA repair processes. Both lineages have lost major cell cycle regulators, including *WHI5* and components of the APC, while FEL species additionally lost numerous genes associated with the spindle checkpoint (e.g., *MAD1* and *MAD2*) and DNA damage checkpoint (e.g., *MEC3* and *RAD9*). Similar patterns are observed among DNA repair-related genes; *Hanseniaspora* species have lost 14 genes, while the FEL yeasts have lost an additional 33 genes. For example, both lineages have lost *MAG1* and *PHR1*, while the FEL has lost additional genes including polymerases (i.e., *POL32* and *POL4*) and multiple telomere-associated genes (e.g., *RIF1*, *RFA3*, *CDC13*, *PBP2*). Compared to the SEL, analyses of substitution patterns in the FEL show higher levels of sequence substitutions, greater instability of homopolymers, and a greater mutational signature associated with the commonly damaged base, 8-oxo-dG (De Bont, 2004). Furthermore, we find that the transition to transversion (or transition / transversion) ratios of the FEL and the SEL are both very close to the ratio expected if transitions and transversions occur neutrally. These results are consistent with the hypothesis that species in the FEL represent a novel example of diversification and long-term evolutionary survival of a hypermutator lineage, which highlights the potential of *Hanseniaspora* for understanding the long-term effects of hypermutation on genome function and evolution.

Materials and Methods

DNA sequencing

For each species, genomic DNA (gDNA) was isolated using a two-step phenol:chloroform extraction previously described to remove additional proteins from the gDNA (Shen et al., 2018). The gDNA was sonicated and ligated to Illumina sequencing adaptors as previously described (Hittinger et al., 2010), and the libraries were submitted for paired-end sequencing (2 x 250) on an Illumina HiSeq 2500 instrument.

Phenotyping

We qualitatively measured growth of species on five carbon sources (maltose, raffinose, sucrose, melezitose, and galactose) as previously described in (Shen et al., 2018). We used a minimal media base with ammonium sulfate and all carbon sources were at a 2% concentration. Yeast were initially grown in YPD and transferred to carbon treatments. Species were visually scored for growth for about a week on each carbon source in three independent replicates over multiple days. A species was considered to utilize a carbon source if it showed growth across ≥ 50% of biological replicates. Growth data for *Hanseniaspora gamundiae* were obtained from Čadež et al., 2019.

Genome assembly and annotation

To generate *de novo* genome assemblies, we used paired-end DNA sequence reads as input to iWGS, version 1.1 (Zhou et al., 2016), a pipeline which uses multiple assemblers and identifies the "best" assembly according to largest genome size and N50 (i.e., the shortest contig length among the set of the longest contigs that account for 50% of the genome assembly's length)

(Yandell and Ence, 2012) as described in (Shen et al., 2018). More specifically, sequenced reads were first quality-trimmed, and adapter sequences were removed used TRIMMOMATIC, version 0.33 (Bolger et al., 2014), and LIGHTER, version 1.1.1 (Song et al., 2014). Subsequently, KMERGENIE, version 1.6982 (Chikhi and Medvedev, 2014), was used to determine the optimal *k*-mer length for each genome individually. Thereafter, six *de novo* assembly tools (i.e., ABYSS, version 1.5.2 (Simpson et al., 2009); DISCOVAR, release 51885 (Weisenfeld et al., 2014); MASURCA, version 2.3.2 (Zimin et al., 2013); SGA, version 0.10.13 (Simpson and Durbin, 2012); SOAPDENOVO2, version 2.04 (Luo et al., 2012); and SPADES, version 3.7.0 (Bankevich et al., 2012)) were used to generate genome assemblies from the processed reads. Using QUAST, version 4.4 (Gurevich et al., 2013), the best assembly was chosen according to the assembly that provided the largest genome size and best N50.

Annotations for eight of the *Hanseniaspora* genomes (i.e., *H. clermontiae, H. osmophila* CBS 313, *H. pseudoguilliermondii, H. singularis, H. uvarum* DSM2768, *H. valbyensis*, *H. vineae* T02 19AF, and *K. hatyaiensis*) and the four outgroup species (i.e., *Cy. jadinii, K. marxianus, S. cerevisiae*, and *W. anomalus*) were generated in a recent comparative genomic study of the budding yeast subphylum (Shen et al., 2018). The other 11 *Hanseniaspora* genomes examined here were annotated by following the same protocol as in (Shen et al., 2018).

In brief, the genomes were annotated using the MAKER pipeline, version 2.31.8 (Holt and Yandell, 2011). The homology evidence used for MAKER consists of fungal protein sequences in the SwissProt database (release 2016_11) and annotated protein sequences of select yeast species from MYCOCOSM (Grigoriev et al., 2014), a web portal developed by the US Department

of Energy Joint Genome Institute for fungal genomic analyses. Three *ab initio* gene predictors were used with the MAKER pipeline, including GENEMARK-ES, version 4.32 (Ter-Hovhannisyan et al., 2008); SNAP, version 2013-11-29 (Korf, 2004); and AUGUSTUS, version 3.2.2 (Stanke and Waack, 2003), each of which was trained for each individual genome. GENEMARK-ES was self-trained on the repeat-masked genome sequence with the fungal-specific option ("–fugus"), while SNAP and AUGUSTUS were trained through three iterative MAKER runs. Once all three *ab initio* predictors were trained, they were used together with homology evidence to conduct a final MAKER analysis in which all gene models were reported ("keep_preds" set to 1), and these comprise the final set of annotations for the genome.

Data acquisition

All publicly available *Hanseniaspora* genomes, including multiple strains from a single species, were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/; S1 File). These species and strains include *H. guilliermondii* UTAD222 (Seixas et al., 2017), *H. opuntiae* AWRI3578, *H. osmophila* AWRI3579, *H. uvarum* AWRI3580 (Sternes et al., 2016), *H. uvarum* 34-9, *H. vineae* T02-19AF (Giorello et al., 2014), *H. valbyensis* NRRL Y-1626 (Riley et al., 2016), and *H. gamundiae* (Čadež et al., 2019). We also included *Saccharomyces cerevisiae* S288C, *Kluyveromyces marxianus* DMKU3-1042, *Wickerhamomyces anomalus* NRRL Y-366-8, and *Cyberlindnera jadinii* NRRL Y-1542, four representative budding yeast species that are all outside the genus *Hanseniaspora* (Shen et al., 2018), which we used as outgroups. Together with publicly available genomes, our sampling of *Hanseniaspora* encompasses all known species in the genus (or its anamorphic counterpart, *Kloeckera*), except *Hanseniaspora lindneri*, which likely belongs to the FEL based on a four-locus phylogenetic study (Cadez, 2006), and

Hanseniaspora taiwanica, which likely belongs to the SEL based on neighbor-joining analyses of the LSU rRNA gene sequence (Chang et al., 2012).

Assembly assessment and identification of orthologs

To determine genome assembly completeness, we calculated contig N50 (Yandell and Ence, 2012) and assessed gene content completeness using multiple databases of curated orthologs from BUSCO, version 3 (Waterhouse et al., 2018a). More specifically, we determined gene content completeness using orthologous sets of genes constructed from sets of genomes representing multiple taxonomic levels, including Eukaryota (superkingdom; 100 species; 303 BUSCOs), Fungi (kingdom; 85 species; 290 BUSCOs), Dikarya (subkingdom; 75 species; 1,312 BUSCOs), Ascomycota (phylum; 75 species; 1,315 BUSCOs), Saccharomyceta (no rank; 70 species; 1,759 BUSCOs), and Saccharomycetales (order; 30 species; 1,711 BUSCOs).

Genomes sequenced in the present project were sequenced at an average depth of 63.49 ± 52.57 (S1 File). Among all *Hanseniaspora*, the average scaffold N50 was 269.03 ± 385.28 kb, the average total number of scaffolds was 980.36 ± 835.20 (398.32 ± 397.97 when imposing a 1kb scaffold filter), and the average genome assembly size was 10.13 ± 1.38 Mb (9.93 ± 1.35 Mb when imposing a 1kb scaffold filter). Notably, the genome assemblies and gene annotations created in the present project were comparable to publicly available ones. For example, the genome size of publicly available *Hanseniaspora vineae* T02 19AF is 11.38 Mb with 4,661 genes, while our assembly of *Hanseniaspora vineae* NRRL Y-1626 was 11.15 Mb with 5,193 genes.

We found that our assemblies were of comparable quality to those from publicly available genomes. For example, *Hanseniaspora uvarum* NRRL Y-1614 (N50 = 267.64 kb; genome size = 8.82 Mb; number of scaffolds = 258; gene number = 4,227), which was sequenced in the present study, and *H. uvarum* AWRI3580 (N50 = 1,289.09 kb; genome size = 8.81 Mb; number of scaffolds = 18; gene number = 4,061), which is publicly available (Sternes et al., 2016) had similar single-copy BUSCO genes present in the highest and lowest ORTHODB (Waterhouse et al., 2013) taxonomic ranks (Eukaryota and Saccharomycetales, respectively). Specifically, *H. uvarum* NRRL Y-1614 and *H. uvarum* AWRI3580 had 80.20% (243 / 303) and 79.87% (242 / 303) of universally single-copy orthologs in Eukaryota present in each genome respectively, and 52.31% (895 / 1,711) and 51.49% (881 / 1,711) of universally single-copy orthologs in Saccharomycetales present in each genome, respectively.

To identify single-copy orthologous genes (OGs) among all protein coding sequences for all 29 taxa, we used ORTHOMCL, version 1.4 (Li et al., 2003). ORTHOMCL clusters genes into OGs using a Markov clustering algorithm [122; https://micans.org/mcl/)] from gene similarity information acquired from a blastp 'all-vs-all' using NCBI's BLAST+, version 2.3.0 (S2 Fig from Steenwyk et al., 2019a; (Camacho et al., 2009)) and the proteomes of species of interest as input. The key parameters used in blastp 'all-vs-all' were: e-value = $1e^{-10}$, percent identity cut-off = 30%, percent match cutoff = 70%, and a maximum weight value = 180. To conservatively identify OGs, we used a strict ORTHOMCL inflation parameter of 4.

To identify additional OGs suitable for use in phylogenomic and molecular sequence analyses, we identified the single best putatively orthologous gene from OGs with full species

representation and a maximum of two species with multiple copies using PHYLOTREEPRUNER, version 1.0 (Kocot et al., 2013). To do so, we first aligned and trimmed sequences in 1,143 OGs out a total of 11,877 that fit the criterion of full representation and a maximum of two species with duplicate sequences. More specifically, we used MAFFT, version 7.294b (Katoh and Standley, 2013), with the BLOSUM62 matrix of substitutions (Mount, 2008), a gap penalty of 1.0, 1,000 maximum iterations, the 'genafpair' parameter, and TRIMAL, version 1.4 (Capella-Gutierrez et al., 2009), with the 'automated1' parameter to align and trim individual sequences, respectively. The resulting OG multiple sequence alignments were then used to infer gene phylogenies using FASTTREE, version 2.1.9 (Price et al., 2010), with 4 and 2 rounds of subtree-prune-regraft and optimization of all 5 branches at nearest-neighbor interchanges, respectively, as well as the 'slownni' parameter to refine the inferred topology. Internal branches with support lower than 0.9 Shimodaira-Hasegawa-like support implemented in FASTTREE (Price et al., 2010) were collapsed using PHYLOTREEPRUNER, version 1.0 (Kocot et al., 2013), and the longest sequence for species with multiple sequences per OG were retained, resulting a robust set of OGs with every taxon being represented by a single sequence. OGs were realigned (MAFFT) and trimmed (TRIMAL) using the same parameters as above.

Phylogenomic analyses

To infer the *Hanseniaspora* phylogeny, we performed phylogenetic inference using maximum likelihood (Felsenstein, 1981) with concatenation (Rokas et al., 2003; Philippe et al., 2005) and coalescence (Edwards, 2009) approaches. To determine the best-fit phylogenetic model for concatenation and generate single-gene trees for coalescence, we constructed trees per single-copy OG using RAXML, version 8.2.8. (Stamatakis, 2014a), where each topology was

determined using 5 starting trees. Single-gene trees that did not recover all outgroup species as the earliest diverging taxa when serially rooted on outgroup taxa were discarded. Individual OG alignments or trees were used for species tree estimation with RAxML (i.e., concatenation) using the LG (Le and Gascuel, 2008) model of substitution, which is the most commonly supported model of substitution (874 / 1,034; 84.53% genes), or ASTRAL-II, version 4.10.12 (i.e., coalescence) (Mirarab and Warnow, 2015). Branch support for the concatenation and coalescence phylogenies was determined using 100 rapid bootstrap replicates (Stamatakis et al., 2008) and local posterior support (Edwards, 2009), respectively.

Several previous phylogenomic studies have shown that the internal branches preceding the *Hanseniaspora* FEL and SEL are long (Riley et al., 2016; Shen et al., 2016b). To examine whether the relationship between the length of the internal branch preceding the FEL and the length of the internal branch preceding the SEL was consistent across genes in our phylogeny, we used NEWICK UTILITIES, version 1.6 (Junier and Zdobnov, 2010) to remove the 88 single-gene trees where either lineage was not recovered as monophyletic and calculated their difference for the remaining 946 genes.

Estimating divergence times

To estimate divergence times among the 25 *Hanseniaspora* genomes, we used the Bayesian method MCMCTree in the PAML, version 4.9 (Yang, 2007), and the concatenated 1,034-gene matrix. The input tree was derived from the concatenation-based ML analysis under a single LG+G4 (Le and Gascuel, 2008) model (Fig 27A). The in-group root (i.e., the split between the

FEL and SEL) age was set between 0.756 and 1.177 time units (1 time unit = 100 million years ago [mya]), which was adopted from a recent study (Shen et al., 2018).

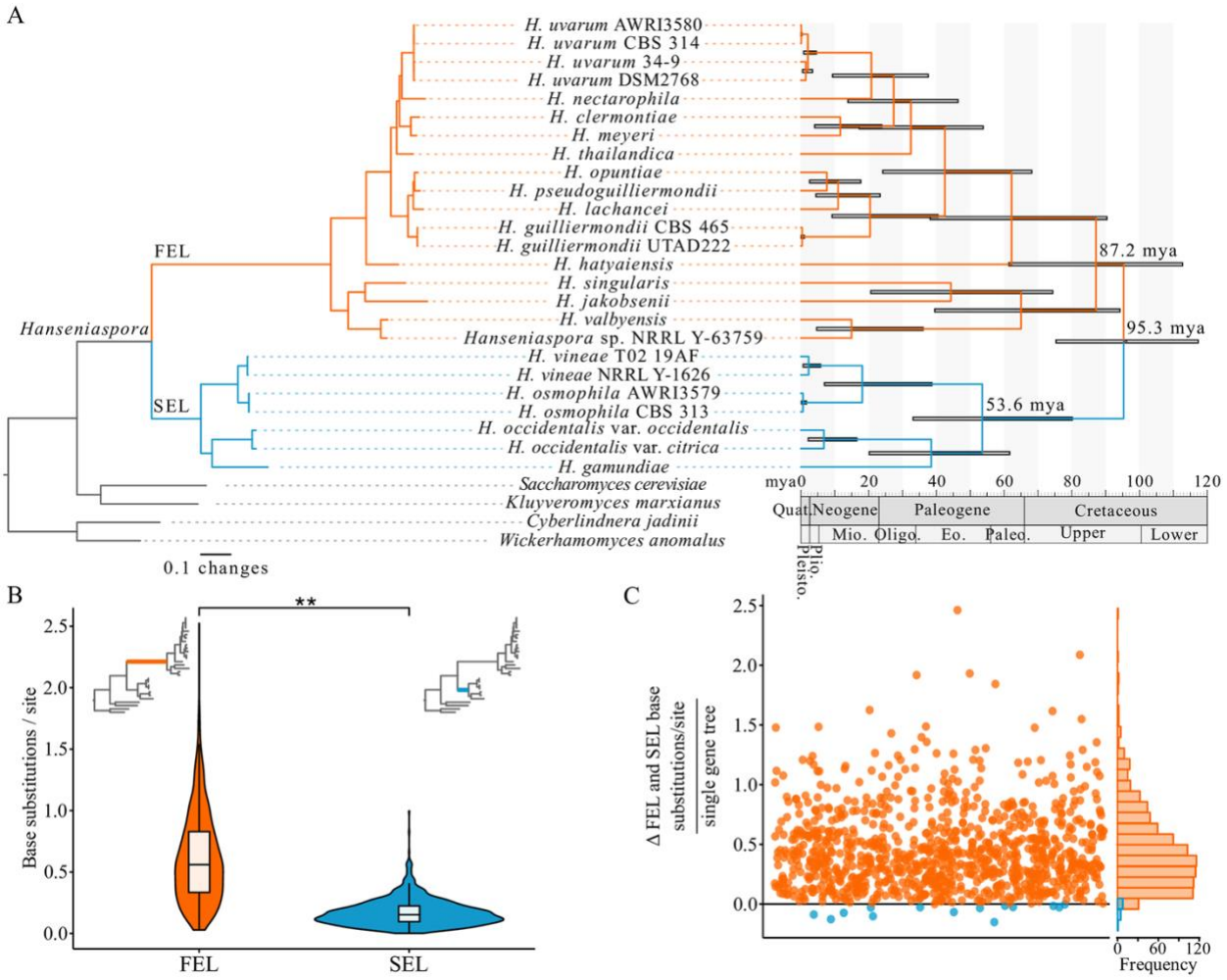

Fig 27. The evolutionary history, rate, and timeline of *Hanseniaspora* diversification.
(A) Phylogenomic and relaxed molecular clock analysis of 1,034 single-copy OGs from a near-complete set of *Hanseniaspora* species revealed two well-supported lineages termed the FEL and SEL, which began diversifying around 87.2 and 53.6 mya after diverging 95.3 mya. (B) Among single-gene phylogenies in which the FEL and SEL were monophyletic (n = 946), the FEL stem branch was consistently and significantly longer (0.62 ± 0.38 base substitutions/site) than the SEL stem branch (0.17 ± 0.11 base substitutions/site) ($p < 0.001$; paired Wilcoxon rank–sum test). (C) Examination of the difference between FEL and SEL: stem branch lengths per single-gene tree revealed that 932 single-gene phylogenies had a longer FEL stem branch (depicted in orange with values greater than 0), while only 14 single-gene phylogenies had a longer SEL stem branch (depicted in blue with values less than 0). Across all single-gene phylogenies, the average

difference in stem branch length between the two lineages was 0.45.
figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. AWRI, Australian Wine Research Institute; CBS, Centraalbureau voor Schimmelcultures; DSM2768, Dutch State Mines 2768; Eo., Eocene; FEL, faster-evolving lineage; Mio., Miocene; mya, million years ago; NRRL, Northern Regional Research Laboratory; OG, orthologous gene; Oligo., Oligocene; Paleo., Paleocene; Pleisto., Pleistocene; Plio., Pliocene; Quat., Quaternary; SEL, slower-evolving lineage; UTAD222, University of Trás-os-Montes and Alto Douro 222.

To infer the *Hanseniaspora* timetree, we first estimated branch lengths under a single LG+G4 (Le and Gascuel, 2008) model with codeml in the PAML, version 4.9 (Yang, 2007), package and obtained a rough mean of the overall mutation rate. Next, we applied the approximate likelihood method (Reis and Yang, 2011; dos Reis et al., 2016) to estimate the gradient vector and Hessian matrix with Taylor expansion (option usedata = 3). Last, we assigned (a) the gamma-Dirichlet prior for the overall substitution rate (option rgene_gamma) as G(1, 1.55), with a mean of 0.64, (b) the gamma-Dirichlet prior for the rate-drift parameter (option sigma2 gamma) as G(1, 10), and (c) the parameters for the birth-death sampling process with birth and death rates $\lambda=\mu=1$ and sampling fraction $\rho=0$. We employed the independent-rate model (option clock=2) to account for the rate variation across different lineages and used soft bounds (left and right tail probabilities equal 0.025) to set minimum and maximum values for the in-group root mentioned above. The MCMC run was first run for 1,000,000 iterations as burn-in and then sampled every 1,000 iterations until a total of 30,000 samples was collected. Two separate MCMC runs were compared for convergence, and similar results were observed.

Gene presence and absence analysis

To determine the presence and absence of genes in *Hanseniaspora* genomes, we built hidden Markov models (HMMs) for each gene present in *Saccharomyces cerevisiae* and used the

resulting HMM profile to search for the corresponding homolog in each *Hanseniaspora* genome, as well as outgroup taxa. More specifically, for each of the 5,917 verified open reading frames from *S. cerevisiae* (Cherry et al., 2012a) (downloaded Oct 2018 from the *Saccharomyces* genome database), we searched for putative homologs in NCBI's Reference Sequence Database for Fungi (downloaded June 2018) using NCBI's BLAST+, version 2.3.0 (Madden, 2013), blastp function, and an e-value cut-off of $1e^{-3}$ as recommended for homology searches (Pearson, 2013). We used the top 100 hits for the gene of interest and aligned them using MAFFT, version 7.294b (Katoh and Standley, 2013), with the same parameters described above. The resulting gene alignment was then used to create an HMM profile for the gene using the hmmbuild function in HMMER, version 3.1b2 (Eddy, 2011). The resulting HMM profile was then used to search for each individual gene in each *Hanseniaspora* genome and outgroup taxa using the hmmsearch function with an expectation value cutoff of 0.01 and a score cutoff of 50. This analysis was done for the 5,735 genes with multiple blast hits allowing for the creation of a HMM profile. To evaluate the validity of constructed HMMs, we examined their ability to recall genes in *S. cerevisiae* and found that we recovered all nuclear genes.

To determine if any functional categories were over- or under-represented among genes present or absent among *Hanseniaspora* species, we conducted gene ontology (GO) (GeneOntologyConsortium, 2004) enrichment analyses using GOATOOLS, version 0.7.9 (Klopfenstein et al., 2018). We used a background of all *S. cerevisiae* genes and a *p*-value cut-off of 0.05 after multiple-test correction using the Holm method (Holm, 1979). Plotting gene presence and absence among pathways was done by examining depicted pathways available

through the KEGG project (Kanehisa et al., 2016) and the *Saccharomyces* Genome Database (Cherry et al., 2012a).

We examined the validity of the gene presence and absence pipeline by examining under-represented terms and the presence or absence of essential genes in *S. cerevisiae* (Winzeler et al., 1999). We hypothesized that under-represented GO terms will be associated with basic molecular processes and that essential genes will be under-represented among the set of absent genes. In agreement with these expectations, GO terms associated with basic biological processes and essential *S. cerevisiae* genes are under-represented among genes that are absent across *Hanseniaspora* genomes. For example, among all genes absent in the FEL and SEL, the molecular functions BASE PAIRING, GO:0000496 ($p < 0.001$); GTP BINDING, GO:0005525 ($p < 0.001$); and ATPASE ACTIVITY, COUPLED TO MOVEMENT OF SUBSTANCES, GO:0043492 ($p < 0.001$), are significantly under-represented (S4 File). Similarly, *S. cerevisiae* essential genes are significantly under-represented ($p < 0.001$; Fischer's exact test for both lineages) among lost genes with 134 and 23 *S. cerevisiae* essential genes having been lost from the FEL and SEL genomes, respectively (lists of essential *S. cerevisiae* genes absent among *Hanseniaspora* genomes are available through figshare 10.6084/m9.figshare.7670756).

Ploidy estimation

To determine ploidy, we leveraged base frequency distributions at variable sites by mapping each genome's reads to its assembly. This approach is widely employed to determine ploidy from next generation sequencing data and has been implemented in several pieces of software (Boeva et al., 2012; Augusto Corrêa dos Santos et al., 2017; Weiß et al., 2018) and studies

(Yoshida et al., 2013; Zhu et al., 2016). In short, examination of base frequency distributions between a frequency of 20 and 80 can provide insight into ploidy status. More specifically, haploid genomes lack biallelic sites so their base frequency distributions will peak at high and low base frequencies and be depleted in positions with base frequencies near 50 (or a 'smiley-pattern'); diploid genomes typically have two alleles for a locus and are expected to exhibit a unimodal distribution centered around a base frequency of 50; finally, triploid genomes typically have one allele on one chromosome and the other allele in the other two chromosomes and are expected to exhibit a bimodal distribution centered around base frequencies of 33 and 66. Note that this approach assumes that there is a sufficient amount of heterozygosity in the genome, and that ploidy changes may be go undetected in genomes lacking heterozygosity. To ensure high-quality read mapping, we first quality-trimmed reads suing TRIMMOMATIC, version 0.36 (Bolger et al., 2014), using the parameters leading:10, trailing:10, slidingwindow:4:20, and minlen:50. Reads were subsequently mapped to their respective genome using BOWTIE2, version 1.1.2 (Langmead and Salzberg, 2012), with the "sensitive" parameter and converted the resulting file to a sorted bam format using SAMTOOLS, version 1.3.1 (Li et al., 2009a). We next used NQUIRE (Weiß et al., 2018), which extracts base frequency information at segregating sites with a minimum frequency of 0.2. Prior to visualization, we removed background noise by utilizing the Gaussian Mixture Model with Uniform noise component (Weiß et al., 2018).

Molecular evolution and mutation analysis

Molecular sequence rate analysis along the phylogeny.

To determine the rate of sequence evolution over the course of *Hanseniaspora* evolution, we examined variation in the rate of nonsynonymous (dN) to the rate of synonymous (dS)

substitutions (dN/dS or ω) across the species phylogeny. We first obtained codon-based alignments of the protein sequences used during phylogenomic inference by threading nucleotides on top of the amino acid sequence using PAL2NAL, version 14 (Suyama et al., 2006), and calculated ω values under the different hypotheses using the CODEML module in PAML, version 4.9 (Yang, 2007). For each gene tested, we set the null hypothesis (H_o) where all internal branches exhibit the same ω (model = 0) and compared it to four different alternative hypotheses. Under the $H_{FEL\text{-}SEL\ branch}$ hypothesis, the branches immediately preceding the FEL and SEL were assumed to exhibit distinct ω values from the background (model = 2) (Fig 28Bi). Under the H_{FEL} hypothesis, the branch immediately preceding the FEL was assumed to have a distinct ω value, all FEL crown branches were assumed to have their own collective ω value, and all background branches were assumed to have their own collective ω value (model = 2) (Fig 28Ci). The H_{SEL} hypothesis assumed the branch preceding the lineage had its own ω value, all SEL crown branches had their own collective ω value, and all background branches were assumed to have their own collective ω value (model = 2) (Fig 28Di). Lastly, the $H_{FEL\text{-}SEL\ crown}$ hypothesis assumed that all FEL crown branches had their own collective ω value, all SEL crown branches had their own collective ω value, and the rest of the branches were assumed to have their own collective ω value (model = 2) (Fig 28Ei). To determine if each of the alternative hypotheses was significantly different from the null hypothesis, we used the likelihood ratio test (LRT) (α = 0.01). A few genes could not be analyzed due to fatal interruptions or errors during use in PAML, version 4.9 (Yang, 2007), which have been reported by other users (Liu et al., 2017); these genes were removed from the analysis. Thus, this analysis was conducted for 989 genes for three tests ($H_{FEL\text{-}SEL\ branch}$, H_{FEL}, and H_{SEL} hypotheses) and 983 genes for one test ($H_{FEL\text{-}SEL\ crown}$ hypothesis).

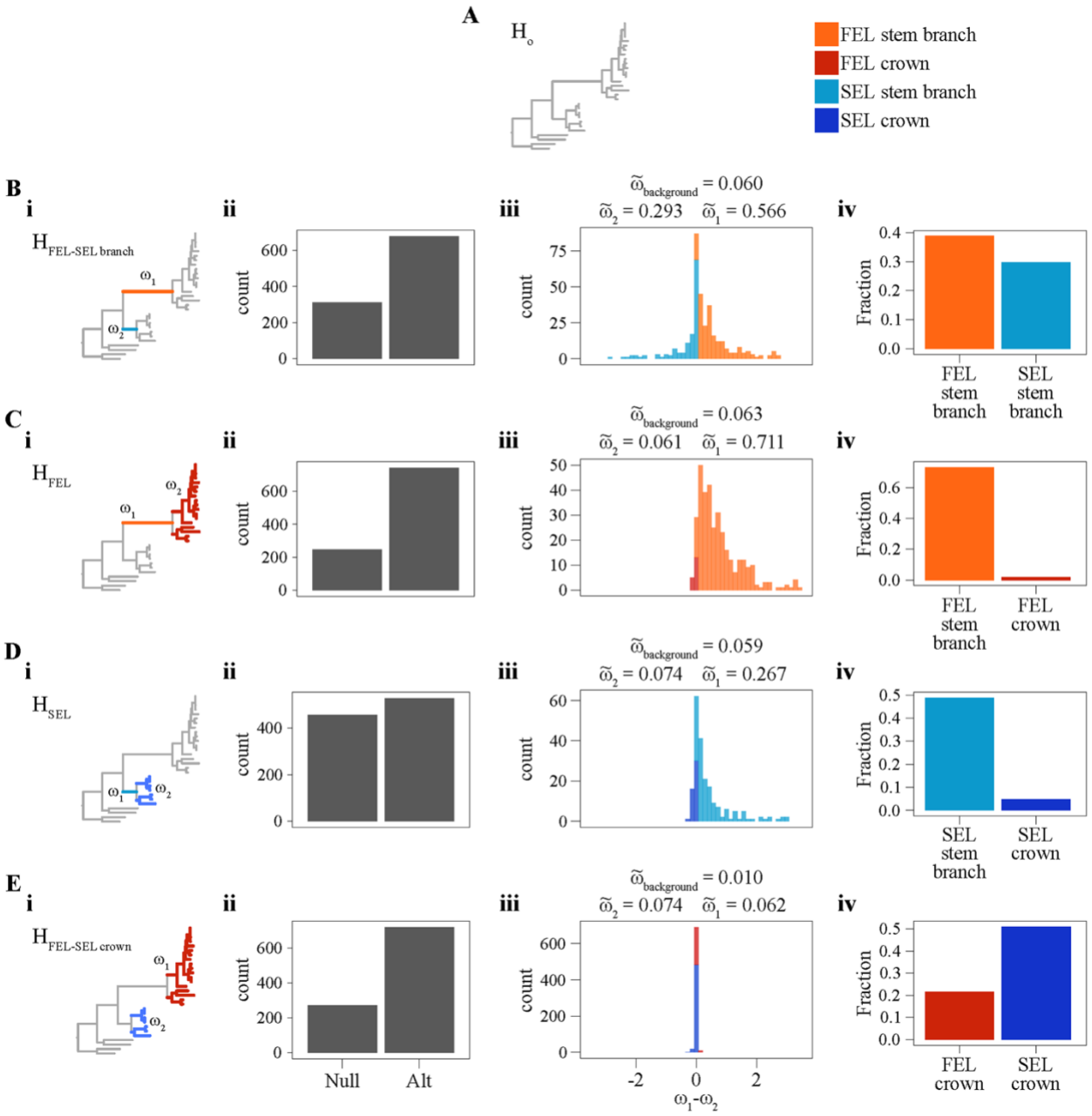

Fig 28. dN/dS (ω) analyses support a historical burst of accelerated evolution in the FEL. (A) The null hypothesis (H_O) that all branches in the phylogeny have the same ω value. Alternative hypotheses (B–E) evaluate ω along three sets of branches. (Bi) The alternative hypothesis ($H_{FEL–SEL\ branch}$) examined ω values along the FEL and SEL stem branches. (Bii) 311 (31.45%) genes supported H_O, and 678 (68.55%) genes supported $H_{FEL–SEL\ branch}$. (Biii) Among the genes that supported $H_{FEL–SEL\ branch}$, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Bi. Here, a range of $\omega_1–\omega_2$ of −3.5 to 3.5 is shown in the histogram. Additionally, we report the median ω_1 and ω_2 values, which are 0.57 and 0.29, respectively. (Biv) 384 (38.83%) genes significantly rejected H_O and were faster in the FEL than the SEL, while 237 (23.96%) significantly rejected H_O and were faster in the SEL than the FEL.

(Ci) The alternative hypothesis (H_{FEL}) examined ω values along the FEL stem branch (ω_1) and crown branches (ω_2). (Cii) 246 (24.87%) genes supported H_O, and 743 (75.13%) genes supported H_{FEL}. (Ciii) Among the genes that supported H_{FEL}, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Ci. The median ω_1 and ω_2 values were 0.71 and 0.06, respectively. (Civ) 725 (73.31%) genes significantly rejected H_O and had higher ω_1 values than ω_2 values, while 18 (1.82%) genes significantly rejected H_O and had higher ω_2 than ω_1 values. (Di) The alternative hypothesis (H_{SEL}) examined ω values along the SEL stem branch (ω_1) and crown branches (ω_2). (Dii) 455 (46.29%) genes supported H_O, and 528 (53.71%) genes supported H_{SEL}. (Diii) Among the genes that supported H_{SEL}, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Di. The median ω_1 and ω_2 values were 0.27 and 0.07, respectively. (Div) 481 (48.93%) genes significantly rejected H_O and had higher ω_1 than ω_2 values, while 47 (4.78%) genes significantly rejected H_O and had higher ω_2 than ω_1 values. (Ei) The alternative hypothesis ($H_{FEL–SEL\ crown}$) examined ω values in the FEL crown branches (ω_1) and SEL crown branches (ω_2). (Eii) 272 (27.50%) genes supported H_O, and 717 (72.50%) genes supported $H_{FEL–SEL\ crown}$. (Eiii) Among the genes that supported $H_{FEL–SEL\ crown}$, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Di. The median ω_1 and ω_2 values were 0.06 and 0.07, respectively. (Eiv) 481 (21.54%) genes significantly rejected H_O and had higher ω_1 than ω_2 values, while 504 (50.96%) genes had higher ω_2 than ω_1 values. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. dN, rate of nonsynonymous substitutions; dS, rate of synonymous subsitutions; FEL, faster-evolving lineage; SEL, slower-evolving lineage.

Examination of mutational signatures

To conservatively identify base substitutions, insertions, and deletions found in taxa in the FEL or SEL, we examined the status of each nucleotide at each position in codon-based and amino acid-based OG alignments. We examined base substitutions, insertions, and deletions at sites that are conserved in the outgroup (i.e., all outgroup taxa have the same character state for a given position in an alignment). For base substitutions, we determined if the nucleotide or amino acid residue in a given *Hanseniaspora* species differed from the conserved outgroup nucleotide or amino acid residue at the same position. To measure if amino acid substitutions in each lineage were conservative or radical (i.e., a substitution to a similar amino acid residue versus a substitution to an amino acid residue with different properties), we used Sneath's index of dissimilarity, which considers 134 categories of biological activity and chemical change to

quantify dissimilarity of amino acid substitutions, and Epstein's coefficient of difference, which considers differences in polarity and size of amino acids to quantify dissimilarity. Notably, Sneath's index is symmetric (i.e., isoleucine to leucine is equivalent to leucine to isoleucine), whereas Epstein's coefficient is not (i.e., isoleucine to leucine is not equivalent to leucine to isoleucine). For indels, we used a sliding window approach with a step size of one nucleotide. We considered positions where a nucleotide was present in all outgroup taxa but a gap was present in *Hanseniaspora* as deletions, and positions where a gap was present in all outgroup taxa and a nucleotide was present in *Hanseniaspora* species as insertions. Analyses were conducted using custom PYTHON, version 3.5.2 (https://www.python.org/), scripts, which use the BIOPYTHON, version 1.70 (Cock et al., 2009a), and NUMPY, version 1.13.1 (Van Der Walt et al., 2011), modules.

We discovered that all *Hanseniaspora* species lack the *PHR1* gene, which is associated with the repair of UV radiation damage but the FEL has lost additional genes that participate in other pathways that can repair UV damage like the base-excision and nucleotide-excision repair pathway (Huang et al., 2000; Budden and Bowden, 2013). UV radiation induces high levels of C → T substitutions at CC sites and more rarely double substitutions of CC → TT (Ikehata and Ono, 2011; Huang et al., 2017). To examine signatures of UV radiation damage across *Hanseniaspora*, we examined the number of C → T substitutions at CC sites (or G → A substitutions at GG sites) as well as the less frequent CC → TT (or GG → AA) double substitutions.

Results

An exceptionally high evolutionary rate in the FEL stem branch

Concatenation and coalescence analyses of a data matrix of 1,034 single-copy orthologous genes (OGs) (522,832 sites; 100% taxon-occupancy) yielded a robust phylogeny of the genus *Hanseniaspora* (Fig 27A, S1 Fig from Steenwyk et al., 2019a, S2 Fig from Steenwyk et al., 2019a). Consistent with previous analyses (Shen et al., 2016b, 2018; Čadež et al., 2019), our phylogeny identified two major lineages, each of which had a long stem branch; we hereafter refer to the lineage with the longer stem branch as the faster-evolving lineage (FEL) and to the other as the slower-evolving lineage (SEL). Relaxed molecular clock analysis suggests that the FEL and SEL split 95.34 (95% credible interval (CI): 117.38 – 75.36) mya, with the origin of their crown groups estimated at 87.16 (95% CI: 112.75 – 61.38) and 53.59 (95% CI: 80.21 – 33.17) mya, respectively (Fig 27A, S3 Fig from Steenwyk et al., 2019a and S2 File from Steenwyk et al., 2019a).

The FEL stem branch is much longer than the SEL stem branch in the *Hanseniaspora* phylogeny (Fig 27) (see also phylogenies in: (Shen et al., 2016b, 2018)). To determine whether this difference in branch length was a property of some or all single-gene phylogenies, we compared the difference in length of the FEL and SEL stem branches among all single-gene trees where each lineage was recovered monophyletic (n = 946). We found that the FEL stem branch was nearly four times longer (0.62 ± 0.38 substitutions / site) than the SEL stem branch (0.17 ± 0.11 substitutions / site) (Fig 27B; $p < 0.001$; Paired Wilcoxon Rank Sum test). Furthermore, of the 946 gene trees examined, 932 had a much longer FEL stem branch (0.46 ± 0.33 Δ substitutions / site), whereas only 14 had a slightly longer SEL stem branch (0.06 ± 0.05 Δ substitutions / site).

The genomes of FEL species have lost substantial numbers of genes

Examination of GC content, genome size, and gene number revealed that the some of the lowest GC content values, as well as the smallest genomes and lowest gene numbers, across the subphylum Saccharomycotina are primarily observed in FEL yeasts (S4 Fig from Steenwyk et al., 2019a). Specifically, the average GC contents for FEL yeasts (33.10 ± 3.53%), SEL yeasts (37.28 ± 2.05%), and all other Saccharomycotina yeasts (40.77 ± 5.58%) are significantly different from one another ($\chi^2(2) = 30.00$, $p < 0.001$; Kruskal-Wallis rank sum test). Pairwise comparisons of GC contents between FEL, SEL, and all other Saccharomycotina were not significant, except in the comparison between the FEL and other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction).

For genome size and gene number, FEL yeast genomes have average sizes of 9.71 ± 1.32 Mb and contain 4,707.89 ± 633.56 genes, respectively, while SEL yeast genomes have average sizes of 10.99 ± 1.66 Mb and contain 4,932.43 ± 289.71 genes. In contrast, all other Saccharomycotina have average genome sizes and gene numbers of 13.01 ± 3.20 Mb and 5,726.10 ± 1,042.60, respectively. Statistically significant differences were observed between the FEL, SEL, and all other Saccharomycotina (genome size: $\chi^2(2) = 33.47$, $p < 0.001$ and gene number: $\chi^2(2) = 31.52$, $p < 0.001$; Kruskal-Wallis rank sum test for both). Pairwise comparisons of genome size and gene number between FEL, SEL, and all other Saccharomycotina revealed that the only significant difference for genome size was between FEL and other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction), while both the FEL and SEL had smaller gene sets compared to other Saccharomycotina yeasts

($p < 0.001$ and $p = 0.008$, respectively; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction). The lower numbers of genes in the FEL (especially) and SEL lineages were also supported by gene content completeness analyses using orthologous sets of genes constructed from sets of genomes representing multiple taxonomic levels across eukaryotes (S5 Fig from Steenwyk et al., 2019a) from the ORTHODB database (Waterhouse et al., 2013).

To further examine which genes have been lost in the genomes of FEL and SEL species relative to other representative Saccharomycotina genomes, we conducted HMM-based sequence similarity searches using annotated *S. cerevisiae* genes as queries in HMM construction (see *Methods*) (S6 Fig from Steenwyk et al., 2019a). Because we were most interested in broad patterns of gene losses in the FEL and SEL, we focused our analyses on genes lost in at least two-thirds of each lineage (i.e., ≥ 11 FEL taxa or ≥ 5 SEL taxa). Using this criterion, we found that 1,409 and 771 genes have been lost in the FEL and SEL, respectively (Fig 2A). Among the genes lost in each lineage, 748 genes were lost across both lineages, 661 genes were uniquely lost in the FEL, and 23 genes were uniquely lost in the SEL (S3 File from Steenwyk et al., 2019a).

To identify the likely functions of genes lost from each lineage, we conducted GO enrichment analyses. Examination of significantly over-represented GO terms for the sets of genes that have been lost in *Hanseniaspora* genomes revealed numerous categories related to metabolism (e.g., MALTOSE METABOLIC PROCESS, GO:0000023, $p = 0.006$; SUCROSE ALPHA-GLUCOSIDASE ACTIVITY, GO:0004575, $p = 0.003$) and genome-maintenance processes (e.g., MEIOTIC CELL

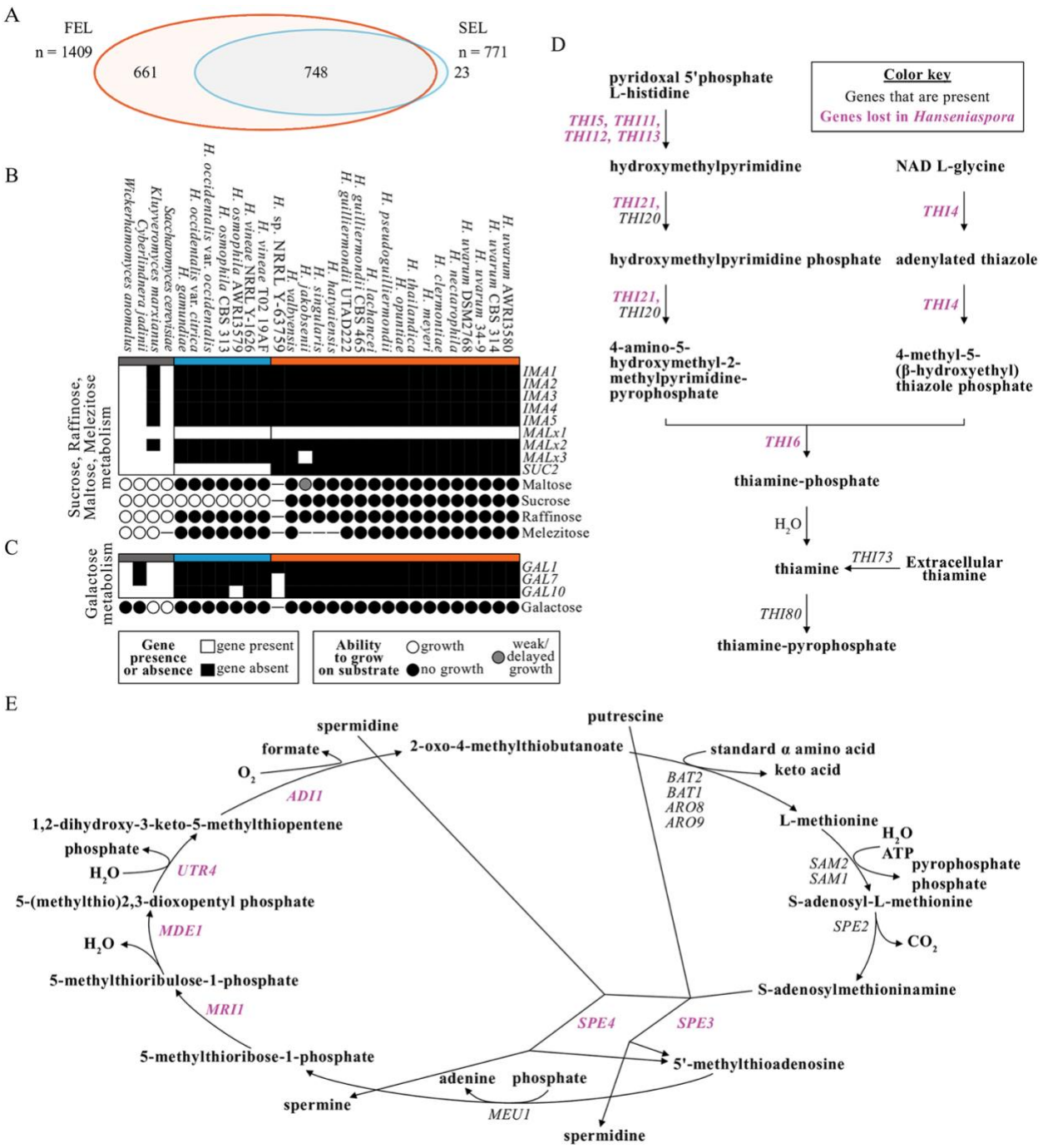

Fig 29. Gene presence and absence analyses reflect phenotype and reveal disrupted pathways. (A) Examination of gene presence and absence (see Methods) revealed numerous genes that were lost across *Hanseniaspora*. Specifically, 1,409 were lost in the FEL, and 771 genes were lost in the SEL. A Euler diagram represents the overlap of these gene sets. Both lineages have lost 748 genes, the FEL has lost an additional 661, and the SEL has lost an additional 23. (B) The *IMA* gene family (*IMA1–5*) encoding α-glucosidases, *MAL* (*MALx1–3*) loci, and *SUC2* are

associated with growth on maltose, sucrose, raffinose, and melezitose. The *IMA* and *MAL* loci are largely absent among *Hanseniaspora* with the exception of homologs *MALx1*, which encode diverse transporters of the major facilitator superfamily whose functions are difficult to predict from sequence; as expected, *Hanseniaspora* spp. cannot grow on maltose, raffinose, and melezitose, with the sole exception of *H. jakobsenii*, which has delayed/weak growth on maltose and is the only *Hanseniaspora* species with *MALx3*, which encodes a homolog of the *MAL*-activator protein. (C) The genes involved with galactose degradation are largely absent among *Hanseniaspora* species, which correlates with their inability to grow on galactose. Genes that are present are depicted in white, and genes that are absent are depicted in black. The ability to grow, the ability to weakly grow/exhibit delayed growth on a given substrate, or the inability to grow is specified using white, gray, and black circles, respectively; dashes indicate no data. (D) Most genes involved in the thiamine biosynthesis pathway are absent among all *Hanseniaspora*. (E) Many genes involved in the methionine salvage pathway are absent among all *Hanseniaspora*. Absent genes are depicted in purple.
figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. *ADI*, Acireductone Dioxygenase; *ARO*, AROmatic amino-acid requiring; AWRI, Australian Wine Research Institute; *BAT*, Branched-chain Amino-acid Transaminase; CBS, Centraalbureau voor Schimmelcultures; DSM2768, Dutch State Mines 2768; FEL, faster-evolving lineage; *GAL*, GALactose metabolism; *IMA*, IsoMAltase; *MAL*, MALtose fermentation; *MDE*, Methylthioribulose-1-phosphate DEhydratase; *MEU*, Multicopy Enhancer of Upstream activation site; *MRI*, MethylthioRibose-1-phosphate Isomerase; NRRL, Northern Regional Research Laboratory; *SAM*, S-AdenosylMethionine requiring; SEL, slower-evolving lineage; *SUC2*, SUCrose; *THI*, THIamine regulon; UTAD222, University of Trás-os-Montes and Alto Douro 222; *UTR*, Unidentified Transcript.

CYCLE, GO:0051321, $p < 0.001$) (S4 File from Steenwyk et al., 2019a). Additional terms, such as CELL CYCLE, GO:0007049 ($p < 0.001$), CHROMOSOME SEGREGATION, GO:0007059 ($p < 0.001$), CHROMOSOME ORGANIZATION, GO:0051276 ($p = 0.009$), and DNA-DIRECTED DNA POLYMERASE ACTIVITY, GO:0003887 ($p < 0.001$), were significantly over-represented among genes absent only in the FEL. Next, we examined in more detail the identities and likely functional consequences of extensive gene losses across *Hanseniaspora* associated with metabolism, cell cycle, and DNA repair.

Metabolism-associated gene losses.

Examination of the genes causing over-representation of metabolism-associated GO terms revealed gene losses in the *IMA* gene family and the *MAL* loci, both of which are associated with growth primarily on maltose but can also facilitate growth on sucrose, raffinose, and melezitose (Kurtzman and Fell, 1998; Opulente et al., 2018). All *IMA* genes have been lost in *Hanseniaspora*, whereas *MALx3*, which encodes the *MAL*-activator protein (Charron et al., 1989) has been lost in all but one species (*Hanseniaspora jakobsenii*; Fig 29B). Consistent with these losses, *Hanseniaspora* species cannot grow on the carbon substrates associated with these genes (i.e., maltose, raffinose, and melezitose) with the exception of *H. jakobsenii*, which has weak/delayed growth on maltose (Fig 29B and S5 File from Steenwyk et al., 2019a). The growth of *H. jakobsenii* on maltose may be due to a cryptic α-glucosidase gene or represent a false positive, as *MALx2* encodes the required enzyme for growth on maltose and is absent in *H. jakobsenii*. Because these genes are also associated with growth on sucrose in some species (Opulente et al., 2018), we also examined their ability to grow on this substrate. In addition to the *MAL* loci conferring growth on sucrose, the invertase Suc2 can also break down sucrose into glucose and fructose (Koschwanez et al., 2011). We found that FEL yeasts have lost *SUC2* and are unable to grow on sucrose, while SEL yeasts have *SUC2* and are able to grow on this substrate (Fig 29B and S5 File from Steenwyk et al., 2019a). Altogether, patterns of gene loss are consistent with known metabolic traits.

Examination of gene sets associated with growth on other carbon substrates revealed that *Hanseniaspora* species also cannot grow on galactose, consistent with the loss of one or more of the three genes involved in galactose assimilation (*GAL1*, *GAL7*, and *GAL10*) from their

genomes (Fig 29C and S5 File from Steenwyk et al., 2019a). Additionally, all *Hanseniaspora* genomes appear to have lost two key genes, *PCK1* and *FBP1*, encoding enzymes in the gluconeogenesis pathway (S7A Fig from Steenwyk et al., 2019a); in contrast, all *Hanseniaspora* have an intact glycolysis pathway (S7B Fig from Steenwyk et al., 2019a).

Altogether these metabolism-associated gene losses may reflect *Hanseniaspora* ecology. More specifically, among wine strains of *S. cerevisiae*, genes associated with maltose and thiamine metabolism are frequently absent in their genomes (Gallone et al., 2016; Steenwyk and Rokas, 2017) and are thought to reflect their ecology in the grape must environment (Steenwyk and Rokas, 2018). Interestingly, similar gene losses are observed among *Hanseniaspora* species but are often more pronounced; for example, *Hanseniaspora* species lack most of the thiamine biosynthesis pathway while wine strains of *S. cerevisiae* typically lack a single member of the *THI* gene family.

Manual examination of other metabolic pathways revealed that *Hanseniaspora* genomes are also lacking some of their key genes. For example, we found that THIAMINE BIOSYNTHETIC PROCESS, GO:0009228 ($p = 0.003$), was an over-represented GO term among genes absent in both the FEL and SEL due to the absence of *THI* and *SNO* family genes. Further examination of genes present in the thiamine biosynthesis pathway revealed extensive gene loss (Fig 29D), which is consistent with their inability to grow on vitamin-free media (Kurtzman and Fell, 1998) (S5 File from Steenwyk et al., 2019a). Notably, *Hanseniaspora* are still predicted to be able to import extracellular thiamine via Thi73 and convert it to its active cofactor via Thi80, which may explain why they can rapidly consume thiamine (Martin et al., 2018). Similarly, examination of

amino acid biosynthesis pathways revealed the methionine salvage pathway was also largely disrupted by gene losses across all *Hanseniaspora* (Fig 29E). Lastly, we found that *GDH1* and *GDH3* from the glutamate biosynthesis pathway from ammonium are absent in FEL yeasts (S3 File from Steenwyk et al., 2019a). However, *Hanseniaspora* have *GLT1*, which enables glutamate biosynthesis from glutamine.

Cell cycle and genome integrity-associated gene losses.

Many genes involved in cell cycle and genome integrity, including cell cycle checkpoint genes, have been lost across *Hanseniaspora* (Fig 30). For example, *WHI5* and *DSE2*, which are responsible for repressing the Start (i.e., an event that determines cells have reached a critical size before beginning division) (Jorgensen, 2002) and help facilitate daughter-mother cell separation through cell wall degradation (Colman-Lerner et al., 2001), have been lost in both lineages. Additionally, the FEL has lost the entirety of the DASH complex (i.e., *ASK1*, *DAD1*, *DAD2*, *DAD3*, *DAD4*, *DUO1*, *DAM1*, *HSK3*, *SPC19*, and *SPC34*), which forms part of the kinetochore and functions in spindle attachment and stability, as well as chromosome segregation, and the MIND complex (i.e., *MTW1*, *NNF1*, *NSL1*, and *DSN1*), which is required for kinetochore bi-orientation and accurate chromosome segregation (S3 and S4 Files from Steenwyk et al., 2019a). Similarly, FEL species have lost *MAD1* and *MAD2*, which are associated with spindle checkpoint processes and have abolished checkpoint activity when their encoded proteins are unable to dimerize (Heinrich et al., 2014). Lastly, components of the anaphase-promoting complex, a major multi-subunit regulator of the cell cycle, are lost in both lineages (i.e., *CDC26* and *MND2*) or just the FEL (i.e., *APC2*, *APC4*, *APC5*, and *SWM1*).

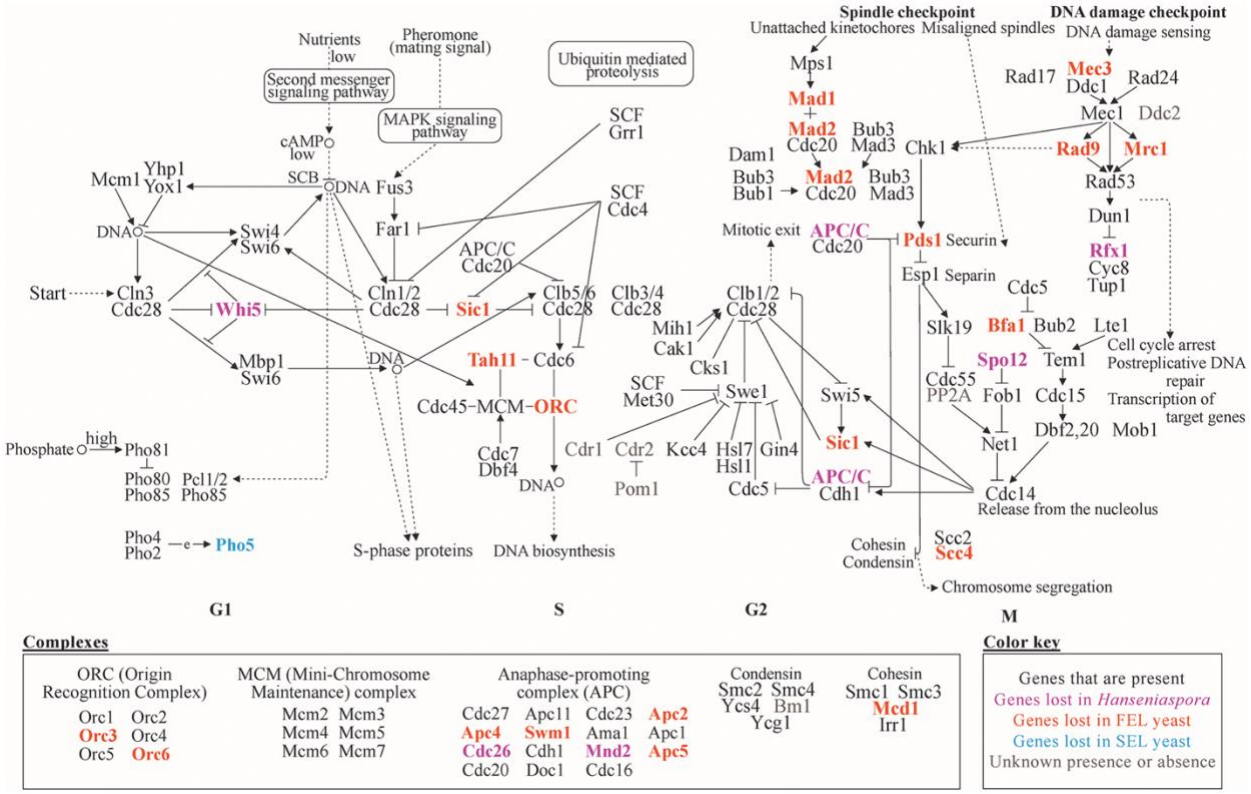

Fig 30. Gene presence and absence in the budding yeast cell cycle.
Examination of cell-cycle genes revealed numerous genes that are absent in *Hanseniaspora* genomes. The genes not present in *Hanseniaspora* participate in diverse functions and include key regulators such as *WHI5*, components of spindle checkpoint processes and segregation such as *MAD1* and *MAD2*, and components of DNA-damage–checkpoint processes such as *MEC3*, *RAD9*, and *RFX1*. Genes absent in both lineages, the FEL, or the SEL are colored purple, orange, or blue, respectively. The "e" in the PHO cascade represents expression of Pho4:Pho2. Dotted lines with arrows indicate indirect links or unknown reactions. Lines with arrows indicate molecular interactions or relations. Circles indicate chemical compounds such as DNA. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. Ama, Activator of meiotic anaphase-promoting complex; APC (or APC/C), Anaphase-Promoting Complex; Bfa, Byr-four-alike; Bm1, Biomimetic moiety glutathionesulfonic acid; Bub, Budding uninhibited by benzimidazole; Cak1, Cyclin-dependent kinase-activating kinase; cAMP, cyclic AdenosineMonoPhosphate; Cdc, Cell division cycle; Cdh, CDC20 homolog; Cdr, *Candida* drug resistance; Chk, Checkpoint kinase; Cks, Cdc28 kinase subunit; Clb, Cyclin B; Cln, Cyclin; Cyc, Cytochrome C; Dam, Duo1 and Mps1 interacting; Dbf, Dumbbell former; Ddc, DNA Damage Checkpoint; Doc, Destruction of Cyclin B; Dun, DNA-damage UNinducible; Esp1, Extra spindle pole bodies 1; Far1, Factor ARrest; FEL, faster-evolving lineage; Fob, Fork Blocking less; Fus3, cell fusion 3; Gin4, Growth inhibitory 4; Grr, Glucose repression-resistant; Hsl, Histone synthetic lethal; Irr, Irregular cell behavior; Kcc, K^+-Cl^- cotransporters; Lte, Low temperature essential; *MAD*, Mitotic Arrest-Deficient; MAPK, Mitogen-Activated Protein Kinase; Mbp, Mlul-box–binding protein; Mcd, Mitotic chromosome determinant; MCM, Mini-Chromosome Maintenance; *MEC3*, Mitosis Entry Checkpoint 3; Met30, Methionine requiring 30; Mih1,

Mitotic inducer homolog; Mnd, Meiotic nuclear divisions; Mob, Mps one binder; Mps, Monopolar spindle; Mrc, Mediator of the Replication Checkpoint; Net, Nucleolar silencing establishing factor and telophase regulator; ORC, Origin Recognition Complex; Pds, Precocious Dissociation of Sisters; PHO, PHOsphate; Pom, Polarity misplaced; PP2A, Protein Phosphatase 2A; *RAD9*, RADiation sensitive; *RFX1*,; SCB, Swi4,6-dependent cell ycle box; Scc, Sister Chromatid Cohesion; SCF, S-phase kinase-associated protein, Cullin, F-box containing complex; SEL, slower-evolving lineage; Sic, Sucrose NonFermenting; Slk, Synthetic lethal karyogamy; Smc, Stability of minichromosomes; Spo, Sporulation; Swe, *Saccharomyces* Wee1; Swi, Switching deficient; Swm, Spore Wall Maturation; Tah11, Topo-A Hypersensitive; Tem, Termination of M phase; Tup, deoxythymidine monophosphate-uptake; *WHI5*, WHIskey 5; Ycg, Yeast cap G; Ycs, Yeast condensing subunit; Yhp1, Yeast Homeo-Protein 1; Yox1, Yeast homeobox 1.

Another group of genes that have been lost in *Hanseniaspora* are genes associated with the DNA damage checkpoint and DNA damage sensing. For example, both lineages have lost *RFX1*, which controls a late point in the DNA damage checkpoint pathway (Lubelsky et al., 2005), whereas the FEL has lost *MEC3* and *RAD9*, which encode checkpoint proteins required for arrest in the G2 phase after DNA damage has occurred (Weinert et al., 1994). Since losses in DNA damage checkpoints and dysregulation of spindle checkpoint processes are associated with genomic instability, we next evaluated the ploidy of *Hanseniaspora* genomes (Galgoczy and Toczyski, 2001). Using base frequency plots, we found that the ploidy of genomes of FEL species ranges between 1 and 3, with evidence suggesting that certain species, such as *H. singularis*, *H. pseudoguilliermondii*, and *H. jakobsenii*, are potentially aneuploid (S8 Fig). In contrast, the genomes of SEL species have ploidies of 1-2 with evidence of potential aneuploidy observed only in *H. occidentalis* var. *citrica.* Greater variance in ploidy and aneuploidy in the FEL compared to the SEL may be due to the FEL's loss of a greater number of components of the anaphase-promoting complex (APC), whose dysregulation is thought to increase instances of aneuploidy (Kim et al., 2017).

Lastly, we examined losses among genes related to meiosis. Although little is known about meiosis and sexual reproduction in *Hanseniaspora*, recent attempts to induce sporulation and sexual reproduction in different *Hanseniaspora* species have been unsuccessful (Chang et al., 2012; Diawara et al., 2015; Albertin et al., 2016; Langenberg et al., 2017). In contrast, other species (i.e., *Hanseniaspora thailandica*, *Hanseniaspora singularis*, and *Hanseniaspora gamundiae*) are able to sporulate (Jindamorakot et al., 2009; Čadež et al., 2019). These inconsistences may be due to the infrequency of sporulation or reduced total number of spores produced which, may be linked to the losses of genes associated with coordinating meiosis such as the major regulator *IME1* (Kassir et al., 1988) and genes associated with spore formation such as *SSP1* (Nag et al., 1997) and *GIP1* (Tachikawa et al., 2001) (S9 Fig from Steenwyk et al., 2019a).

Pronounced losses of DNA repair genes in the FEL.

Examination of other GO-enriched terms revealed numerous genes associated with diverse DNA repair processes that have been lost among *Hanseniaspora* species, and especially the FEL (Fig 31). We noted 14 lost DNA repair genes across all *Hanseniaspora*, including the DNA glycosylase gene *MAG1* (Xiao and Chow, 1998), the photolyase gene *PHR1* that exclusively repairs pyrimidine dimers (Sebastian et al., 1990), and the diphosphatase gene *PCD1*, a key contributor to the purging of mutagenic nucleotides, such as 8-oxo-dGTP, from the cell (Nunoshiba, 2004). An additional 33 genes were lost specifically in the FEL such as *TDP1*, which repairs damage caused by topoisomerase activity (Nitiss et al., 2006); the DNA polymerase gene *POL32* that participates in base-excision and nucleotide-excision repair and

whose null mutants have increased genomic deletions (Huang et al., 2000); and the *CDC13* gene that encodes a telomere-capping protein (Lustig, 2001).

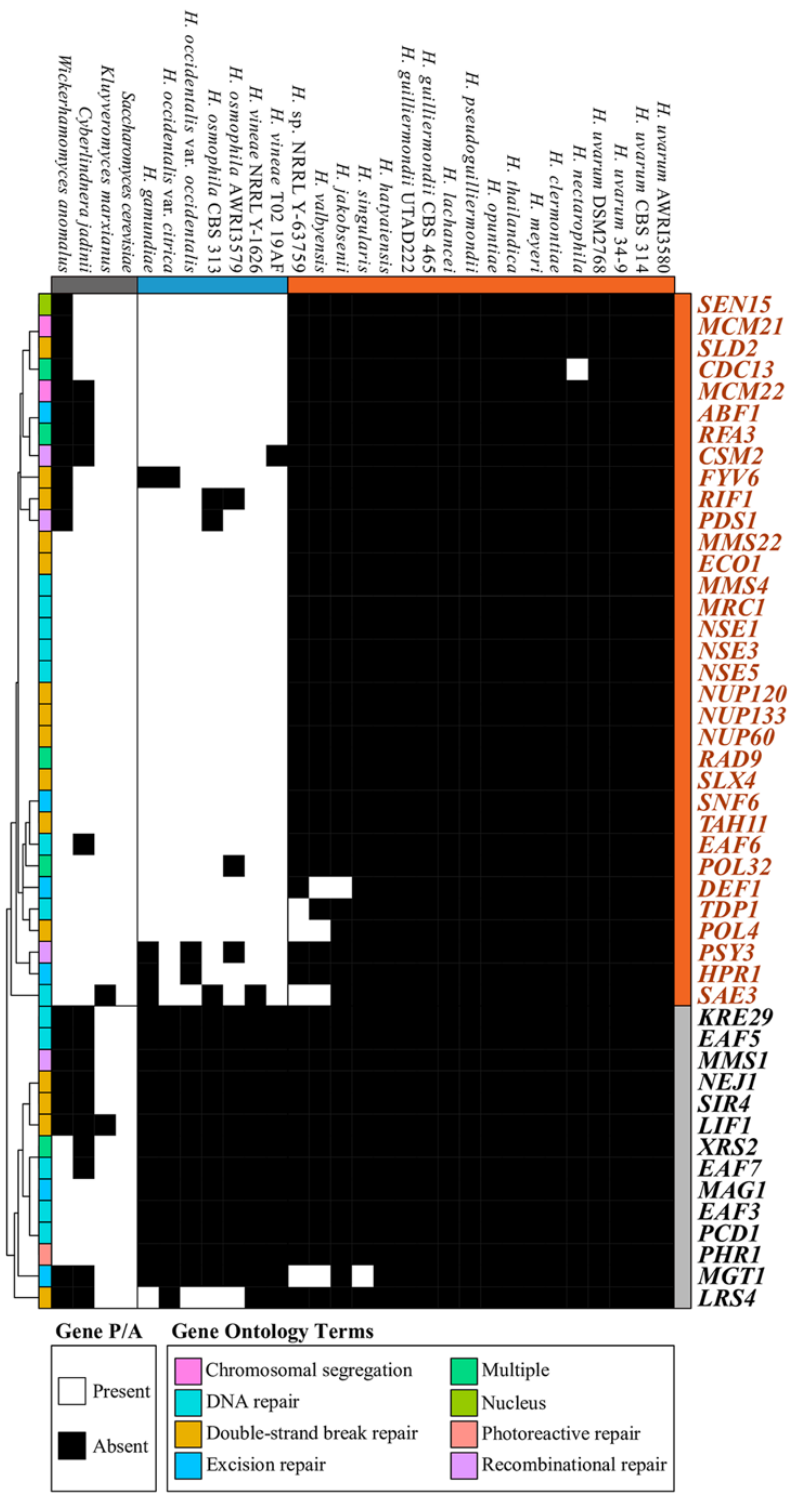

Fig 31. A panoply of genome-maintenance and DNA repair genes are absent among *Hanseniaspora*, especially in the FEL.
Genes annotated as DNA repair genes according to GO (GO:0006281) and child terms were examined for presence and absence in at least two-thirds of each lineage, respectively (268 total genes). 47 genes are absent among the FEL species, and 14 genes are absent among the SEL. Presence and absence of genes was clustered using hierarchical clustering (cladogram on the left) where each gene's ontology is provided as well. Genes with multiple gene annotations are denoted as such using the "multiple" term.
figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. *ABF1*, Autonomously replicating sequence-Binding Factor 1; AWRI, Australian Wine Research Institute; CBS, Centraalbureau voor Schimmelcultures; *CDC13*, Cell Division Cycle 13; *CSM2*, Chromosome Segregation in Meiosis 2; *DEF1*, RNA polymerase II Degradation Factor 1; DSM2768, Dutch State Mines 2768; *EAF6*, Essential something about silencing 2-related acetyltransferase 1-Associated Factor 6; *ECO1*, Establishment of Cohesion 1; FEL, faster-evolving lineage; *FYV6*, Function required for Yeast Viability 6; GO, gene ontology; *HPR1*, HyPerRecombination 1; *KRE29*, Killer toxin Resistant 29; *LIF1*, Ligase Interacting Factor 1; *LRS4*, Loss of RDNA Silencing 4; *MAG1*, 3-MethylAdenine DNA Glycosylase 1; *MCM21*, Mini-Chromosome Maintenance 21; *MGT1*, O-6-MethylGuanine-DNA methylTransferase 1; *MMS22*, Methyl MethaneSulfonate sensitivity 22; *MRC1*, Mediator of the Replication Checkpoint 1; *NEJ1*, Nonhomologous End-Joining defective 1; NRRL, Northern Regional Research Laboratory; *NSE1*, NonStructural maintenance of chromosomes Element 1; *NUP120*, NUclear Pore 120; *PCD1*, Peroxisomal Coenzyme A Diphosphatase 1; *PDS1*, Precocious Dissociation of Sisters 1; *PHR1*, PHotoreactivation Repair deficient 1; *POL32*, POLymerase 32; *PSY3*, Platinum SensitivitY 3; P/A, presence or absence; *RAD9*, RADiation sensitive 9; *RFA3*, Replication Factor A 3; *RIF1*, Repressor/activator site binding protein-Interacting Factor 1; *SAE3*, Sporulation in the Absence of sporulation Eleven; SEL, slower-evolving lineage; *SEN15*, Splicing ENdonuclease 15; *SIR4*, Silent Information Regulator 4; *SLD2*, Synthetically Lethal with DNA polymerase B (II)-1 2; *SLX4*, Synthetical Lethal of unknown (X) function 4; *SNF6*, Sucrose NonFermenting 6; *TAH11*, Topo-A Hypersensitive 11; *TDP1*, Tyrosyl-DNA Phosphodiesterase 1;UTAD222, University of Trás-os-Montes and Alto Douro 222; *XRS2*, X-Ray Sensitive 2.

FEL gene losses are associated with accelerated sequence evolution

Loss of DNA repair genes is associated with a burst of sequence evolution.

To examine the mutational signatures of losing numerous DNA repair genes on *Hanseniaspora* substitution rates, we tested several different hypotheses that postulated changes in the ratio of the rate of nonsynonymous (dN) to the rate of synonymous substitutions (dS) (dN/dS or ω) along the phylogeny (Table 1 from Steenwyk et al., 2019a and Fig 28). For each hypothesis tested, the null was that the ω value remained constant across all branches of the phylogeny. Examination of

the hypothesis that the ω values of both the FEL and SEL stem branches were distinct from the background ω value ($H_{FEL\text{-}SEL\ branch}$; Fig 28B), revealed that 678 genes (68.55% of examined genes) significantly rejected the null hypothesis (Table 1 from Steenwyk et al., 2019a; α = 0.01; LRT; median FEL stem branch ω = 0.57, median SEL stem branch ω = 0.29, and median background ω = 0.060). Examination of the hypothesis that the ω value of the FEL stem branch and the ω value of the FEL crown branches were distinct from the background ω value (H_{FEL}; Fig 28C) revealed 743 individual genes (75.13% of examined genes) that significantly rejected the null hypothesis (Table 1 from Steenwyk et al., 2019a; α = 0.01; LRT; median FEL stem branch ω = 0.71, median FEL crown branches ω = 0.06, median background ω = 0.063). Testing the same hypothesis for the SEL (H_{SEL}; Fig 28D) revealed 528 individual genes (53.7% of examined genes) that significantly rejected the null hypothesis (Table 1 from Steenwyk et al., 2019a; α = 0.01; LRT; median SEL stem branch ω = 0.267, median SEL crown branches ω = 0.074, median background ω = 0.059). Finally, testing of the hypothesis that the FEL and SEL crown branches have ω values distinct from each other and the background ($H_{FEL\text{-}SEL\ crown}$; Fig 28E) revealed 717 genes (72.5% of examined genes) that significantly rejected the null hypothesis (Table 1 from Steenwyk et al., 2019a; α = 0.01; LRT; median FEL crown branches ω = 0.062, median SEL crown branches ω = 0.074, median background ω = 0.010). These results suggest a dramatic, genome-wide increase in evolutionary rate in the FEL stem branch (Fig 28B and 28C), which coincided with the loss of a large number of genes involved in DNA repair.

The FEL has a greater number of base substitutions and indels.

To better understand the mutational landscape in the FEL and SEL, we characterized patterns of base substitutions across the 1,034 OGs. Focusing on first (n = 240,565), second (n = 318,987),

and third (n = 58,151) codon positions that had the same character state in all outgroup taxa, we first examined how many of these sites had experienced base substitutions in FEL and SEL species (Fig 32A). We found significant differences between the proportions of base substitutions in the FEL and SEL (F(1) = 196.88, $p < 0.001$; Multi-factor ANOVA) at each codon position (first: $p < 0.001$; second: $p < 0.001$; and third: $p = 0.02$; Tukey Honest Significance Differences post-hoc test).

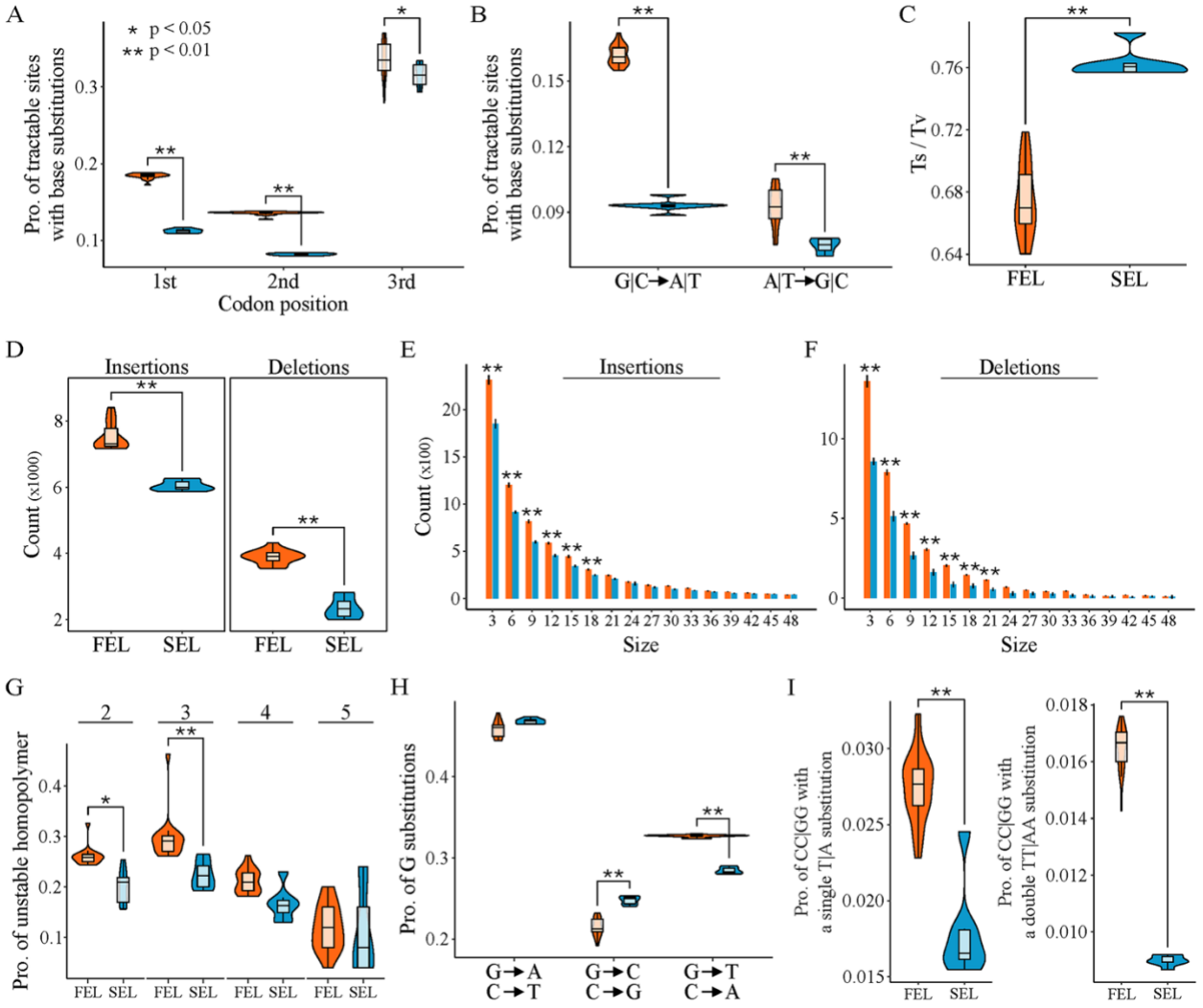

Fig 32. Analyses of base substitutions and indels reveal a higher mutational load in the FEL compared to the SEL.

(A) Analyses of substitution patterns among codon-based alignments of 1,034 OGs revealed a higher number of base substitutions in the FEL compared to the SEL ($F(1) = 196.88$, $p < 0.001$; multifactor ANOVA) and an asymmetric distribution of base substitutions at codon sites ($F(2) = 1{,}691.60$, $p < 0.001$; multifactor ANOVA). A Tukey honest significance differences post hoc test revealed a higher proportion of substitutions in the FEL compared to the SEL at the first ($n = 240{,}565$; $p < 0.001$), second ($n = 318{,}987$; $p < 0.001$), and third ($n = 58{,}151$; $p = 0.02$) codon positions. (B) Analyses of the direction of base substitutions (i.e., G|C → A|T or A|T → G|C) revealed significant differences between the FEL and SEL ($F(1) = 447.1$, $p < 0.001$; multifactor ANOVA) as well as differences in the directionality of base substitutions ($F(1) = 914.5$, $p < 0.001$; multifactor ANOVA). A Tukey honest significance differences post hoc test revealed a significantly higher proportion of substitutions were G|C → A|T compared to A|T → G|C among sites that are G|C ($n = 232{,}546$) and A|T ($n = 385{,}157$) ($p < 0.001$), suggesting a general AT bias of base substitutions. Additionally, there was a significantly higher proportion of sites with base substitutions in the FEL compared to the SEL ($p < 0.001$). Specifically, a higher number of base substitutions was observed in the FEL compared to the SEL for both G|C → A|T ($p < 0.001$) and A|T → G|C mutations ($p < 0.001$), but the bias toward AT was greater in the FEL. (C) Examinations of transition/transversion ratios revealed a lower transition/transversion ratio in the FEL compared to the SEL ($p < 0.001$; Wilcoxon rank–sum test). (D) Comparisons of insertions and deletions revealed a significantly greater number of insertions ($p < 0.001$; Wilcoxon rank–sum test) and deletions ($p < 0.001$; Wilcoxon rank–sum test) in the FEL ($\bar{X}_{\text{insertions}} = 7,521.11 \pm 405.34$; $\bar{X}_{\text{deletions}} = 3,894.11 \pm 208.16$) compared to the SEL ($\bar{X}_{\text{insertions}} = 6,049.571 \pm 155.85$; $\bar{X}_{\text{deletions}} = 2,346.71 \pm 326.22$). (E and F) When adding the factor of size per insertion or deletion, significant differences were still observed between the lineages ($F(1) = 2{,}102.87$, $p < 0.001$; multifactor ANOVA). A Tukey honest significance differences post hoc test revealed that most differences were caused by significantly more small insertions and deletions in the FEL compared to the SEL. More specifically, there were significantly more insertions in the FEL compared to the SEL for sizes 3–18 ($p < 0.001$ for all comparisons between each lineage for each insertion size), and there were significantly more deletions in the FEL compared to the SEL for sizes 3–21 ($p < 0.001$ for all comparisons between each lineage for each deletion size). Black lines at the top of each bar show the 95% confidence interval for the number of insertions or deletions for a given size. (G) Evolutionarily conserved homopolymers of sequence length 2 ($n = 17{,}391$), 3 ($n = 1{,}062$), 4 ($n = 104$), and 5 ($n = 5$) were examined for substitutions and indels. Statistically significant differences of the proportion mutated bases (i.e., [base substitutions + deleted bases + inserted bases]/total homopolymer bases) were observed between the FEL and SEL ($F(1) = 27.68$, $p < 0.001$; multifactor ANOVA). Although the FEL had more mutations than the SEL for all homopolymers, a Tukey honest significance differences post hoc test revealed differences were statistically significant for homopolymers of two ($p = 0.02$) and three ($p = 0.003$). Analyses of homopolymers using additional factors of mutation type (i.e., base substitution, insertion, deletion) and homopolymer sequence type (i.e., A|T and C|G homopolymers) can be seen in S10 Fig. (H) G → T or C → A mutations are associated with the common and abundant oxidatively damaged base, 8-oxo-dG. When examining all substituted G positions for each species and their substitution direction, we found significant differences between different substitution directions

($F(2) = 5{,}682$, $p < 0.001$; multifactor ANOVA). More importantly, a Tukey honest significance differences post hoc test revealed an over-representation of G → T or C → A in the FEL compared to the SEL ($p < 0.001$). (I) Signatures of UV-damage–associated single and double substitutions (i.e., C → T at CC sites and CC → TT) double substitutions are greater in the FEL compared to the SEL ($p < 0.001$ for both tests; Wilcoxon rank–sum test).
figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; OG, orthologous gene; Pro., Proportion; SEL, slower-evolving lineage.

We next investigated differences in the direction of substitutions. Specifically, we examined if substitutions were biased in the AT- direction (i.e., G|C → A|T) or GC- direction (i.e., A|T → G|C) as well as if there are differences among substitutions in these directions between FEL and SEL. We observed significant differences among substitutions in the AT- and GC-directions between the FEL and SEL ($F(1) = 447.1$, $p < 0.001$; Multi-factor ANOVA), as well as between overall AT- and GC-bias across both lineages among G|C ($n = 232{,}546$) and A|T ($n = 385{,}157$) sites ($F(1) = 914.5$, $p < 0.001$; Multi-factor ANOVA) (Fig 32B). There were significantly more base substitutions in the FEL compared to the SEL and a significant bias toward A|T across both lineages ($p < 0.001$ for both tests; Tukey Honest Significance Differences post-hoc test).

We next examined patterns of transition / transversion ratios and observed a lower transition / transversion ratio in the FEL (0.67 ± 0.02) compared to the SEL (0.76 ± 0.01) (Fig 32C; $p < 0.001$; Wilcoxon Rank Sum test); this finding is in contrast to the transition / transversion ratios found in most known organisms, whose values are substantially above 1.00 [56–59]. Altogether, these analyses reveal more base substitutions in the FEL and SEL across all codon positions, a significant AT-bias in base substitutions across all *Hanseniaspora*, and a low transition / transversion ratio across FEL and SEL.

Examination of indels revealed that the total number of insertions or deletions was significantly greater in the FEL ($mean_{insertions}$ = 7521.11 ± 405.34; $mean_{deletions}$ = 3894.11 ± 208.16) compared to the SEL ($mean_{insertions}$ = 6049.571 ± 155.85; $mean_{deletions}$ = 2346.71 ± 326.22) (Fig 32D; $p < 0.001$ for both tests; Wilcoxon Rank Sum test). The difference in number of indels between the FEL and SEL remained significant after taking into account indel size ($F(1) = 2102.87$, $p < 0.001$; Multi-factor ANOVA). Further analyses revealed there are significantly more insertions in the FEL compared to the SEL for insertion sizes 3-18 bp ($p < 0.001$ for all comparisons between each lineage for each insertion size; Tukey Honest Significance Differences post-hoc test), while there were significantly more deletions in the FEL compared to the SEL for deletion sizes 3-21 bp ($p < 0.001$ for all comparisons between each lineage for each deletion size; Tukey Honest Significance Differences post-hoc test). These analyses suggest that there are significantly more indels in the FEL compared to the SEL and that this pattern is primarily driven by short indels.

Greater sequence instability in the FEL and signatures of endogenous and exogenous DNA damage

The FEL has greater instability of homopolymers.

Examination of the total proportion of mutated bases among homopolymers (i.e., stretches of the same base) in codon-based alignments of the 1,034 orthologous genes (i.e., (substituted bases + deleted bases + inserted bases) / total homopolymer bases) revealed significant differences between the FEL and SEL (Fig 32G; $F(1) = 27.68$, $p < 0.001$; Multi-factor ANOVA). Although the FEL had a higher proportion of mutations among homopolymers across all sizes of two (n = 17,391), three (n = 1,062), four (n = 104), and five (n = 5), significant differences were observed

for homopolymers of length two and three ($p = 0.02$ and $p = 0.003$, respectively; Tukey Honest Significance Differences post-hoc). To gain more insight into the stability of different homopolymer runs (i.e., A|T or C|G) and the types of sequence changes that occur among homopolymers, we considered the additional factors of homopolymer sequence type (i.e., A|T or C|G) and mutation type (i.e., base substitution, insertion, or deletion) (S10 Fig from Steenwyk et al., 2019a). In addition to recapitulating differences between the types of mutations that occur at homopolymers ($F(2) = 1686.70$, $p < 0.001$; Multi-factor ANOVA), we observed that base substitutions occurred more frequently than insertions and deletions ($p < 0.001$ for both tests; Tukey Honest Significance Differences post-hoc test). For example, among A|T and C|G homopolymers of length two and C|G homopolymers of length three, base substitutions were higher in the FEL compared to the SEL ($p = 0.009$, $p < 0.001$, and $p < 0.001$, respectively; Tukey Honest Significance Differences post-hoc test). Additionally, there were significantly more base substitutions in A|T homopolymers of length five in the FEL compared to the SEL ($p < 0.001$; Tukey Honest Significance Differences post-hoc test). Altogether, these analyses reveal greater instability of homopolymers in the FEL compared to the SEL due to more base substitutions.

The FEL has a stronger signature of endogenous DNA damage from 8-oxo-dG.

Examination of mutational signatures associated with common endogenous and exogenous mutagens revealed greater signatures of mutational load in the FEL compared to the SEL, as well as in both FEL and SEL compared to the outgroup taxa. The oxidatively damaged guanine base, 8-oxo-dG, is a commonly observed endogenous form of DNA damage that causes the transversion mutation of G → T or C → A (De Bont, 2004). Examination of the direction of base substitutions among all sites with a G base in all outgroup taxa revealed differences in the

direction of base substitutions ($F(2) = 5{,}682$, $p < 0.001$; Multi-factor ANOVA). Moreover, there are significantly more base substitutions at G sites associated with 8-oxo-dG damage in the FEL compared to the SEL (Fig 32H; $p < 0.001$; Tukey Honest Significance Differences post-hoc test). These analyses reveal that FEL genomes have higher proportions of G site substitutions associated with the mutational signature of a common endogenous mutagen.

Hanseniaspora *have a greater genomic signature of UV-damage.*

UV damage can result in C → T substitutions at CC sites and CC → TT double substitutions (Huang et al., 2000; Budden and Bowden, 2013). Although both the FEL and SEL have lost *PHR1*, a gene encoding a DNA photolyase that repairs pyrimidine dimers, FEL has lost additional genes in other pathways that can repair UV damage (e.g. *POL32* in the excision repair pathways). We hypothesized the FEL would have a greater signature of UV damage due to these gene losses. We found significantly greater number of single and double substitutions in CC sites indicative of UV damage in the FEL compared to SEL (Fig 32I; $p < 0.001$ for both tests; Wilcoxon Rank Sum test).

Lastly, we examined if all of these mutations were associated with more radical amino acid changes in the FEL compared to the SEL using two measures of amino acid change: Sneath's index (Sneath, 1966) and Epstein's coefficient of difference (Epstein, 1967). For both measures, we observed significantly more radical amino acid substitutions in the FEL compared to the SEL (S11 Fig from Steenwyk et al., 2019a; $p < 0.001$; Wilcoxon Rank Sum test for both metrics). Altogether, these analyses reveal greater DNA sequence instability in the FEL compared to the SEL, which is also associated with more radical amino acid substitutions.

Discussion

Species in the genus *Hanseniaspora* exhibit the longest branches among budding yeasts and their genomes have some of the lowest numbers of genes, lowest GC contents, and smallest assembly sizes in the subphylum (Fig 27, S4 Fig from Steenwyk et al., 2019a) (Riley et al., 2016; Shen et al., 2016b, 2018). Through the analysis of the genomes of nearly every known *Hanseniaspora* species, this study presents multiple lines of evidence suggesting that one lineage of *Hanseniaspora*, which we have named FEL, is a lineage of long-term, hypermutator species that have undergone extensive gene loss (Figs 27-30 as well as S2, S5, S7 and S8 Figs from Steenwyk et al., 2019a).

Evolution by gene loss is gaining increasing attention as a major mode of genome evolution (Albalat and Cañestro, 2016; Shen et al., 2018) and is mainly possible due to the dispensability of the majority of genes. For example, 90% of *E. coli* (Baba et al., 2006), 80% of *S. cerevisiae* (Giaever et al., 2002), and 73% of *Candida albicans* (Segal et al., 2018) genes are dispensable in laboratory conditions. The loss of dispensable genes can be selected for (Koskiniemi et al., 2012) and is common in lineages of obligate parasites or symbionts, such as in the microsporidia, intracellular fungi which have lost key metabolic pathways such as amino acid biosynthesis pathways (Katinka et al., 2001; Keeling and Slamovits, 2004), and myxozoa, a group of cnidarian obligate parasites that infect vertebrates and invertebrates (Chang et al., 2015). Similar losses are also increasingly appreciated in free-living organisms, such as the budding yeasts [this study; 34,35,76–78] and animals (Albalat and Cañestro, 2016). For example, the loss of *SUC2*, a gene known to enable sucrose utilization (Koschwanez et al., 2011), in the FEL reflects the

inability of species in the FEL to grow on sucrose, while its presence in the SEL reflects its species' ability to grow on sucrose (Fig 28).

However, *Hanseniaspora* species have experienced not just the typically observed losses of metabolic genes (Figs 28A and 28B), but more strikingly, the atypical loss of dozens of cell cycle and DNA damage, response, and repair genes (Figs 29 and 30). Losses of cell cycle genes are extremely rare (Medina et al., 2016), and most such losses are known in the context of cancers (Hartwell, 1992). Losses of individual or a few DNA repair genes have also been observed in individual hypermutator fungal isolates (Billmyre et al., 2017; Boyce et al., 2017; Rhodes et al., 2017a). In contrast, the *Hanseniaspora* losses of cell cycle and DNA repair genes are not only unprecedented in terms of the numbers of genes lost and their striking impact on genome sequence evolution, but also in terms of the evolutionary longevity of the lineage.

Lost checkpoint processes are associated with fast growth and bipolar budding.

Hanseniaspora species lost numerous components of the cell cycle (Fig 29), such as *WHI5*, which causes accelerated G1/S transitions in knock-out *S. cerevisiae* strains (Jorgensen, 2002; Costanzo et al., 2004), as well as components of APC (i.e., *CDC26* and *MND2*), which may accelerate the transition to anaphase (Castro et al., 2005). These and other cell cycle gene losses are suggestive of rapid cell division and growth and consistent with the known ability of *Hanseniaspora* yeast of rapid growth in the wine fermentation environment (Langenberg et al., 2017).

One of the distinguishing characteristics of the *Hanseniaspora* cell cycle is bipolar budding, which is known only in the genera *Wickerhamia* (Debaryomycetaceae) and *Nadsonia* (Dipodascaceae), as well as in *Hanseniaspora* and its sister genus *Saccharomycodes* (both in the family Saccharomycodaceae) (Kurtzman and Fell, 1998; Tavares et al., 2018). These three lineages are distantly related to one another on the budding yeast phylogeny (Shen et al., 2018), so bipolar budding likely evolved three times independently in Saccharomycotina, including in the last common ancestor of *Hanseniaspora* and *Saccharomycodes*. Currently, there is only one genome available for *Saccharomycodes* (Tavares et al., 2018), making robust inferences of ancestral states challenging. Interestingly, examination of cell cycle gene presence and absence in the only representative genome from the genus, *Saccharomycodes ludwigii* (Tavares et al., 2018), reveals that *CDC26*, *PCL1*, *PDS1*, *RFX1*, *SIC1*, *SPO12*, and *WHI5* are absent (S6 File from Steenwyk et al., 2019a), most of which are either absent from all *Hanseniaspora* (i.e., *CDC26*, *RFX1*, *SPO12*, and *WHI5*) or just from the FEL (i.e., *PDS1* and *SIC1*). This evidence raises the hypothesis that bipolar budding is linked to the dysregulation of cell cycle processes due to the absence of cell cycle genes and in particular cell cycle checkpoints (Fig 29).

Some gene losses may be compensatory.

Deletion of many of the genes associated with DNA maintenance that have been lost in *Hanseniaspora* lead to dramatic increases of mutation rates and gross genome instability (Huang et al., 2000; Costanzo et al., 2004; Castro et al., 2005), raising the question of how these gene losses were tolerated in the first place. Examination of the functions of the genes lost in *Hanseniaspora* suggests that at least some of these gene losses may have been compensatory. For example, *POL4* knock-out strains of *S. cerevisiae* can be rescued by the deletion of *YKU70*

(Sterling, 2005), both of which were lost in the FEL. Similarly, the loss of genes responsible for key cell cycle functions (e.g., kinetochore functionality and chromosome segregation) appears to have co-occurred with the loss of checkpoint genes responsible for delaying the cell cycle if its functions fail to complete, which may have allowed *Hanseniaspora* cells to bypass otherwise detrimental cell cycle arrest. Specifically, *MAD1* and *MAD2*, which help delay anaphase when kinetochores are unattached (Heinrich et al., 2014); the 10-gene DASH complex, which participates in spindle attachment, stability, and chromosome segregation (Jenni and Harrison, 2018); and the 4-gene MIND complex, which is required for kinetochore bi-orientation and accurate chromosome segregation (Dimitrova et al., 2016), were all lost in the FEL.

Lastly, the telomere capping protein *CDC13* was lost in FEL but is essential not only in *S. cerevisiae* but also in mammalian cells. However, additional losses in DNA damage response genes (i.e., *SGS1*, *EXO1*, and *RAD9*) can allow yeast cells to survive in the absence of *CDC13* (Ngo and Lydall, 2010). In addition to *CDC13*, FEL has also lost the checkpoint protein *RAD9* and other genes in the DNA damage checkpoint pathway, including *MRC1* and *MEC3*. We hypothesize that the loss of *CDC13* was compensated by losses in the DNA damage response pathway as has been observed in *S. cerevisiae* (Ngo and Lydall, 2010).

Long-term hypermutation and the subsequent slowing of sequence evolution.

Estimates of the substitution rate ratio ω suggest the FEL and SEL, albeit to a much lower degree in the latter, underwent a burst of accelerated sequence evolution in their stem lineages, followed by a reduction in the pace of sequence evolution (Fig 31). This pattern is consistent with theoretical predictions that selection against mutator phenotypes will reduce the overall rate of

sequence evolution (Ram and Hadany, 2012), as well as with evidence from experimental evolution of hypermutator lines of *S. cerevisiae* that showed that their mutation rates were quickly reduced (McDonald et al., 2012). Although we do not know the catalyst for this burst of sequence evolution, hypermutators may be favored in maladapted populations or in conditions where environmental parameters frequently change (McDonald et al., 2012; Ram and Hadany, 2012). While the environment occupied by the *Hanseniaspora* last common ancestor is unknown, it is plausible that environmental instability or other stressors favored hypermutators in *Hanseniaspora*. Extant *Hanseniaspora* species are well known to be associated with the grape environment (Chavan et al., 2009; Seixas et al., 2017; Martin et al., 2018). Interestingly, grapes appear to have originated (Wikstrom et al., 2001) around the same time window that *Hanseniaspora* did (Fig 27B), leading us to speculate that the evolutionary trigger of *Hanseniaspora* hypermutation could have been adaptation to the grape environment.

Losses of DNA repair genes are reflected in patterns of sequence evolution.

Although the relationship between genotype and phenotype is complex, the loss of genes involved in DNA repair can have predictable outcomes on patterns of sequence evolution in genomes. In the case of the observed losses of DNA repair genes in *Hanseniaspora*, the mutational signatures of this loss and the consequent hypermutation can be both general (i.e., the sum total of many gene losses), as well as specific (i.e., can be putatively linked to the losses of specific genes or pathways). Arguably the most notable general mutational signature is that *Hanseniaspora* genome sequence evolution is largely driven by random (i.e., neutral) mutagenic processes with a strong AT-bias. For example, whereas the transition / transversion ratios of eukaryotic genomes are typically within the 1.7 and 4 range (Vignal et al., 2002; Marth et al.,

2011; Zhu et al., 2014a; Wang et al., 2015), *Hanseniaspora* ratios are ~0.66-0.75 (Fig 32C), which are values on par with estimates of transition / transversion caused by neutral mutations alone (e.g., 0.6-0.95 in *S. cerevisiae* (Lynch et al., 2008a; Zhu et al., 2014a), 0.92 in *E. coli* (Lynch, 2007), 0.98 in *Drosophila melanogaster* (Keightley et al., 2009a), and 1.70 in humans (Lynch, 2010)). Similarly, base substitutions across *Hanseniaspora* genomes are strongly AT-biased, especially in the FEL (Fig 32), an observation consistent with the general AT-bias of mutations observed in diverse organisms, including numerous bacteria (Hershberg and Petrov, 2010), *Drosophila* fruit flies (Keightley et al., 2009a), *S. cerevisiae* (Zhu et al., 2014a), and humans (Lynch, 2010).

In addition to these general mutational signatures, examination of *Hanseniaspora* sequence evolution also reveals mutational signatures that can be linked to the loss of specific DNA repair genes. For example, we found a higher proportion of base substitutions associated with the most abundant oxidatively damaged base, 8-oxo-dG, which causes G → T or C → A transversions (De Bont, 2004), in the FEL compared to the SEL, which reflects specific gene losses. Specifically, *Hanseniaspora* yeasts have lost *PCD1*, which encodes a diphosphatase that contributes to the removal of 8-oxo-dGTP (Nunoshiba, 2004) and thereby reduces the chance of misincorporating this damaged base. Once 8-oxo-dG damage has occurred, it is primarily repaired by the base excision repair pathway (De Bont, 2004). Notably, the FEL has lost a key component of the base excision repair pathway, a DNA polymerase δ subunit, encoded by *POL32*, which aids in filling the gap after excision (Seeberg et al., 1995). Accordingly, the proportion of G|C sites with substitutions indicative of 8-oxo-dG damage (i.e., G → T or C → A transversions) is significantly greater in the FEL compared to the SEL (Fig 32H). Similarly, the

numbers of dinucleotide substitutions of CC → TT associated with UV-induced pyrimidine dimers (Huang et al., 2017) are higher across *Hanseniaspora* compared to other yeasts due to the loss of *PHR1*, which encodes a DNA photolyase that repairs pyrimidine dimers (Fig 32I) (Sebastian et al., 1990).

Our analyses provide the first major effort to characterize the genome function and evolution of the enigmatic genus *Hanseniaspora.* Our analyses focus on genomic differences between two lineages and identify major and extensive losses of genes associated with metabolism, cell cycle, and DNA repair processes. These extensive losses and the concomitant acceleration of evolutionary rate mean that levels of amino acid sequence divergence within each of the two *Hanseniaspora* lineages alone, but especially within the FEL, are similar to those observed within plant classes and animal subphyla (S12 Fig from Steenwyk et al., 2019a). These discoveries set the stage for further examination of intra-lineage or intra-species variation in genomic features and content. More interestingly, our analyses lay the foundation for fundamental molecular and evolutionary investigations among *Hanseniaspora*, such as potential novel rewiring of cell cycle and DNA repair processes.

CHAPTER 8

Examination of gene loss in the DNA mismatch repair pathway and its mutational consequences in a fungal phylum[7]

Introduction

An ensemble of DNA repair pathways and cell cycle checkpoints is responsible for detecting and repairing DNA damage, ensuring faithful maintenance of the genome (Friedberg et al., 2005; Giglia-Mari et al., 2010). Among DNA repair pathways, the DNA mismatch repair (MMR) pathway is one of the best characterized (Marti et al., 2002). The MMR pathway is responsible for repairing bases that were incorrectly paired during DNA replication via five steps: recognition, incision, removal, re-synthesis, and ligation (Fig. 33A) (Marti et al., 2002; Fukui, 2010; Hsieh and Zhang, 2017). The MMR pathway is highly conserved in both bacteria and eukaryotes; cells that experience reduction or loss of function in this pathway have increased levels of mutation, as seen in cancer and drug-resistant fungal pathogen strains (Fukui, 2010; Billmyre et al., 2017, 2020; Campbell et al., 2017; Dos Reis et al., 2019).

Although DNA repair genes are generally highly conserved, certain fungal lineages have been reported to have a more limited repertoire, particularly within the phylum Ascomycota. For example, budding yeasts (subphylum Saccharomycotina) and fission yeasts (Taphrinomycotina) have fewer DNA repair genes than filamentous fungi (Pezizomycotina)

---

[7]This work is published in: Phillips, M. A., Steenwyk, J. L., Shen, X.-X., and Rokas, A. (2021). Examination of Gene Loss in the DNA Mismatch Repair Pathway and Its Mutational Consequences in a Fungal Phylum. Genome Biol. Evol. 13. doi:10.1093/gbe/evab219.

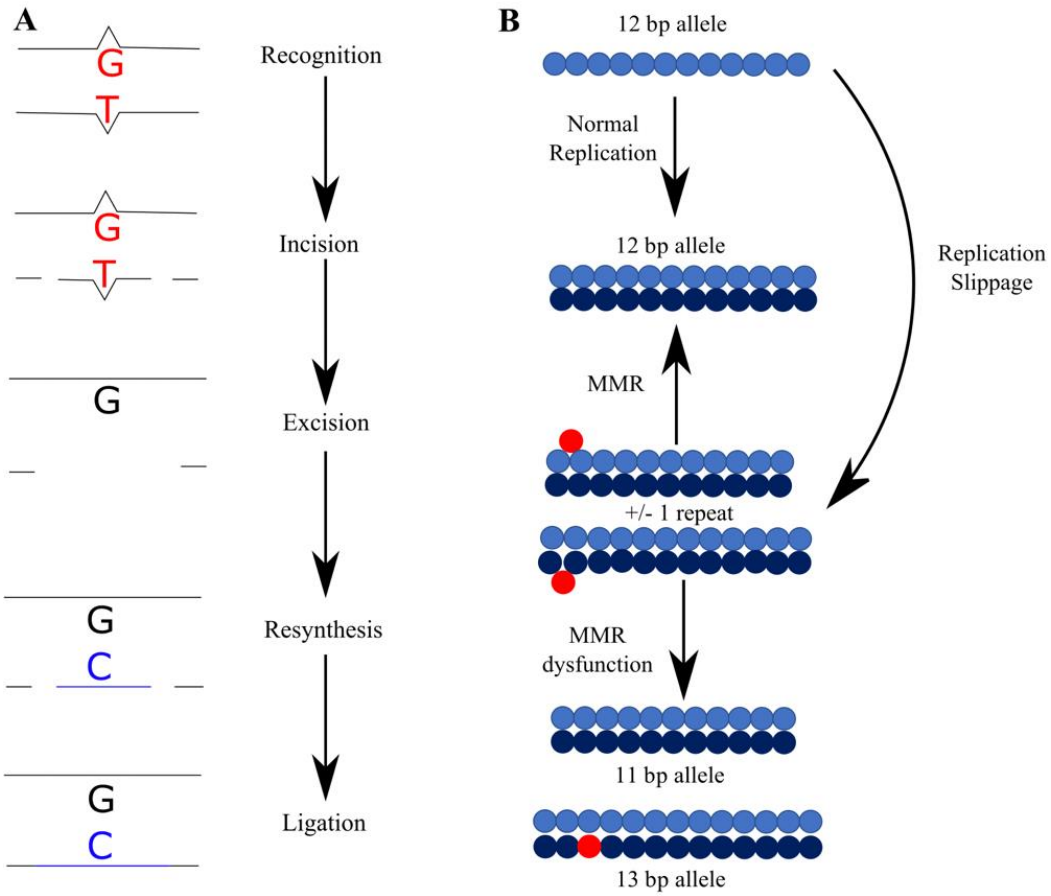

Fig. 33. The DNA Mismatch Repair (MMR) pathway corrects mismatched bases produced during DNA replication and prevents instability in microsatellites.
(A) The pathway is comprised of five conserved steps: recognition of mispaired bases by a sliding clamp, incision of the DNA strand by an endonuclease, excision of the incorrectly paired bases, resynthesis of the DNA strand, and ligation of the newly synthesized segment to the DNA strand. Table S2 from Phillips et al., 2021 includes a full list of MMR genes and their categorization into one of the five steps. (B) The MMR pathway also recognizes base or repeat addition and skipping errors during replication and corrects them. However, dysfunction in this pathway can leave replication slippage unrepaired, leading to the and the addition or deletion of base pairs or repeats, especially in highly repetitive regions such as microsatellites.

(Milo et al., 2019; Shen et al., 2020b). Furthermore, DNA repair genes that were lost from budding yeasts and fission yeasts are more likely to also be lost in filamentous fungi (Milo et al., 2019).

One lineage that has experienced extensive losses in its repertoire of DNA repair genes is the *Hanseniaspora* genus of budding yeasts (Steenwyk et al. 2019). *Hanseniaspora* species have undergone punctuated sequence evolution and have accumulated large numbers of diverse types of substitutions, including those associated with specific gene losses such as UV damage, suggesting that DNA repair is impaired by the high levels of DNA repair gene loss. These findings suggest that DNA repair genes are not universally conserved across fungi and that their loss is compatible with long-term evolutionary survival and diversification of fungal lineages.

One well-established consequence of MMR dysfunction is mutation in microsatellite regions of the genome. Microsatellites are repetitive tracts of DNA, with motifs 1-6 bp long repeated at least five times (Beier et al., 2017). Microsatellites are typically highly polymorphic between individuals and are commonly used as markers in population biology, forensics, paternity testing, and tumor characterization (Richman, 2015). Due to their repetitive nature, microsatellites are prone to experiencing polymerase slippage, which is usually corrected by the MMR pathway (Fig. 33B) (Ellegren, 2004; Richman, 2015). If the MMR pathway does not recognize these errors, as is the case in cancer, microsatellite instability (MSI) can occur (Campbell et al., 2017). MSI is defined by a hypermutable phenotype resulting from a loss of function in the MMR pathway (Boland and Goel, 2010). Instability in microsatellites trends towards elongation in these regions, but contraction can also occur (Ellegren, 2004).

Beyond increased mutation in microsatellite regions, aberrant function of the MMR pathway is associated with genome-wide signatures of genetic instabilities (Boland and Goel, 2010; Billmyre et al., 2017, 2020). MMR mutations have been implicated in the development of

hypermutant and ultrahypermutant human cancers, which constitute approximately 15% of human tumors and less than 1% of tumors, respectively (Campbell et al., 2017). Interestingly, very few tumors with low mutation rates contained mutations in the MMR pathway, whereas more than a third of hypermutant tumors and virtually all the ultrahypermutants contained mutations in MMR genes (Campbell et al., 2017). Hypermutant tumors had high levels of MSI suggesting their hypermutant phenotype is due, at least in part, to MMR dysfunction (Campbell et al., 2017). Hypermutation has also been observed in fungal pathogen strains that have lost MMR pathway genes, potentially driving within-host adaptation and the evolution of drug resistance. For example, Rhodes et al. (2017) found that hypermutation caused by mutations in three MMR pathway genes, including *MSH2*, resulted in a rapid increase in the mutation rate of the human pathogenic fungus *Cryptococcus neoformans*, contributing to infection relapse. Similarly, Billmyre et al. (2017) sequenced multiple strains of the human pathogenic fungus *Cryptococcus deuterogattii* (phylum Basidiomycota) and found that a group of strains with mutations in the *MSH2* gene experienced higher rates of mutation when compared with closely related strains harboring an intact *MSH2* gene. Hypermutation in *C. deuterogattii* mediated rapid evolution of antifungal drug resistance (Billmyre et al., 2017, 2020).

In contrast to MMR gene loss in the microevolutionary context of genetic or infectious disease, the extent of MMR gene loss across lineages spanning multiple species remains understudied. To determine the macroevolutionary impact of MMR gene conservation and loss, we characterized patterns of MMR gene presence and absence in the fungal phylum Ascomycota (Fig. 34). We found that the MMR pathway was highly conserved across Ascomycota, with the median species having 49 / 52 MMR genes present. However, we found that *Blumeria graminis* and species in

the powdery mildew genus *Erysiphe* (subphylum Pezizomycotina, class Leotiomycetes), a group of obligate plant parasites, had many fewer MMR genes and a faster rate of sequence evolution

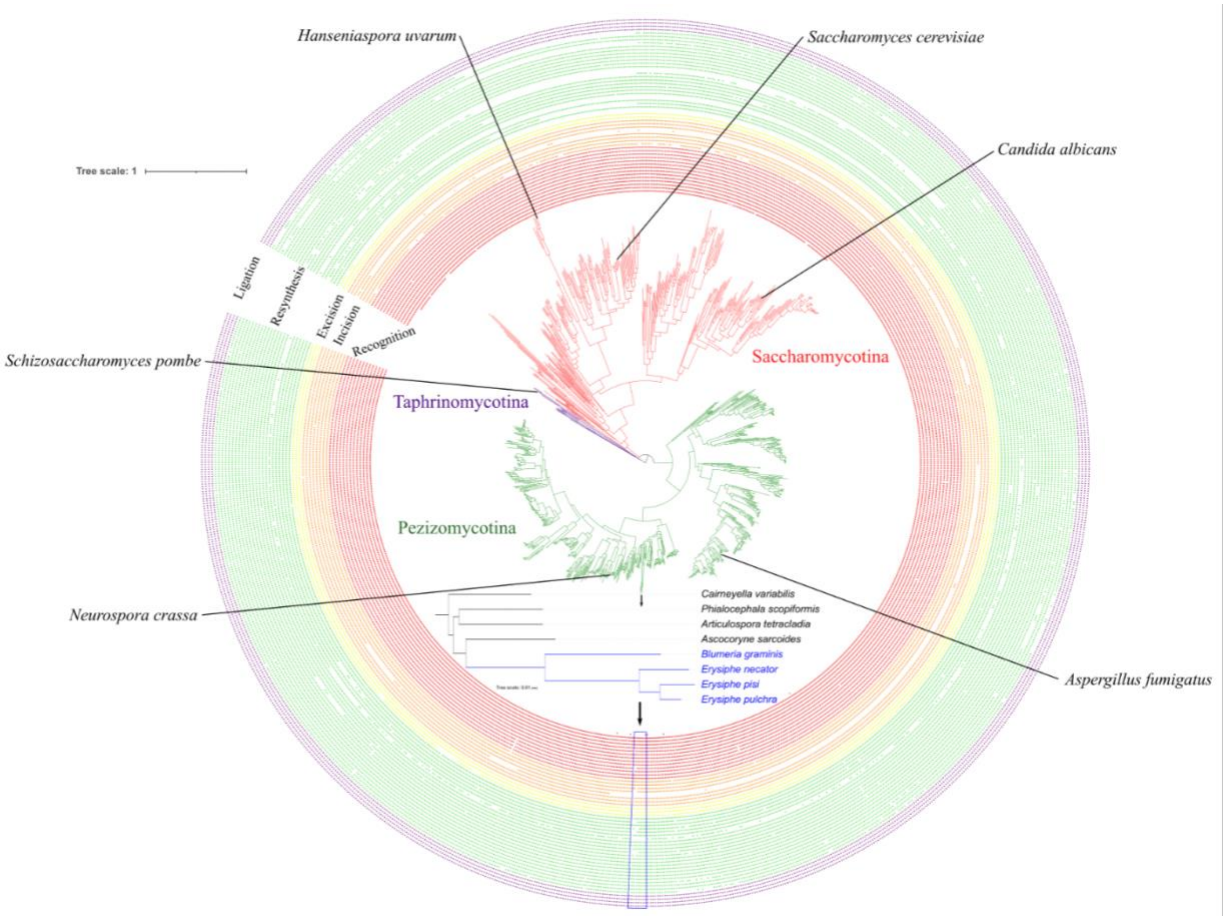

Fig. 34. Conservation of mismatch repair (MMR) pathway genes across the fungal phylum Ascomycota.
MMR genes are generally highly conserved across the phylum. A few model organisms and species of particular interest to medicine and agriculture are labeled as well as a representative species of the faster-evolving *Hanseniaspora* lineage. Gene presences are indicated in the bands surrounding the phylogeny with genes colored according to their function; red is recognition, orange is incision, yellow is excision, green is resynthesis, and purple is ligation. Branches are colored by subphylum; budding yeasts / Saccharomycotina (n = 332 species) are in red, fission yeasts / Taphrinomycotina (n = 14 species) are in purple, and filamentous fungi / Pezizomycotina (n = 761 species) are in green. Taxon names have been omitted from the phylogeny for visualization purposes; the phylogenetic tree with taxon names can be found in Figure S1 from Phillips et al., 2021 and Shen et al. (2020). The inset phylogenetic tree shows the higher loss taxa (HLT; in blue) and lower loss taxa (LLT; in black), with the blue box beneath highlighting them.

than their relatives and most other fungal taxa. Specifically, *Erysiphe necator* has lost 9 MMR genes, *Erysiphe pisi* has lost 21 MMR genes, *Erysiphe pulchra* has lost 7 MMR genes, and *Blumeria graminis* has lost 5 MMR genes (Fig. 35). In contrast, species closely related to *Erysiphe* and *Blumeria* have lost only 1 – 2 MMR genes, consistent with the high degree of MMR gene conservation in the rest of the phylum. Evolutionary genomic analyses revealed that MMR gene losses in *Erysiphe* and *Blumeria* (hereafter referred to as higher loss taxa or HLT) are associated with a proliferation of mononucleotide runs and elongation of microsatellites of all motif lengths, both of which are hallmarks of MMR pathway dysfunction. Reflecting these losses, *Erysiphe* and *Blumeria* genomes also have more pronounced mutational biases and accelerated rates of mutation. These results suggest that obligate plant parasites in the genera *Blumeria* and *Erysiphe* have diversified while lacking otherwise highly conserved MMR genes.

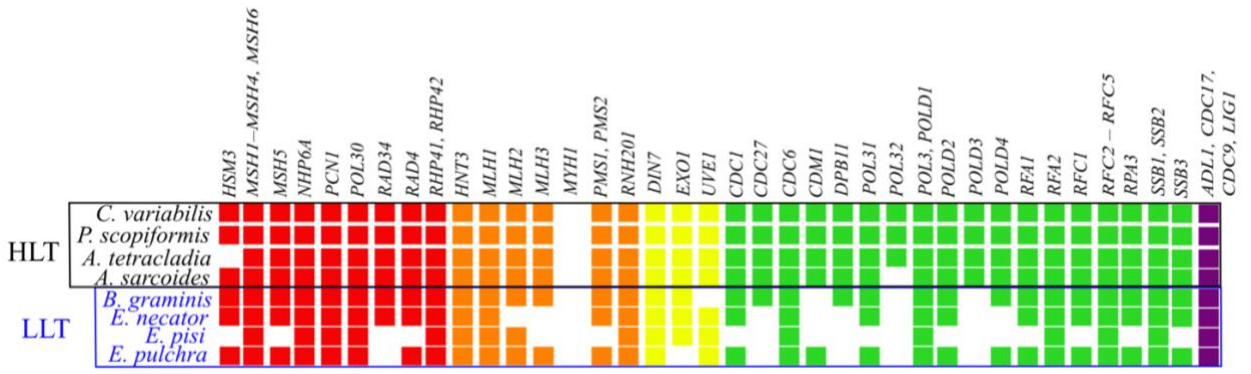

Fig. 35. The powdery mildews *Erysiphe* and *Blumeria* have lost many more mismatch repair (MMR) pathway genes than closely related species.
Higher loss taxa (HLT; shown in blue font) have lost 5 – 21 MMR genes, while lower loss taxa (LLT; shown in black font) have lost 1 – 2 genes and the median ascomycete has lost 3 genes. Note that the losses of *EXO1*, *MLH2*, *MLH3*, *MSH5*, *PMS1*, *PMS2*, and *RFC1* are uniquely observed in HLT. Genes are colored according to their function; red is recognition, orange is incision, yellow is excision, green is resynthesis, and purple is ligation.

Materials and Methods

Curation of the set of DNA mismatch repair pathway genes

To investigate the presence and absence of MMR genes across the fungal phylum Ascomycota, we curated a dataset of MMR genes from the genomes of three fungal model organisms representing the three different subphyla: *Saccharomyces cerevisiae* (subphylum Saccharomycotina), *Neurospora crassa* (Pezizomycotina), and *Schizosaccharomyces pombe* (Taphrinomycotina). We used three sources to curate genes that are part of the MMR pathway: the Kyoto Encyclopedia of Genes and Genomes (KEGG, genome.jp/kegg/; Kanehisa & Goto, 2000), the *Schizosaccharomyces pombe* database (PomBase, pombase.org/; Lock et al., 2019; The Gene Ontology Consortium, 2019), and the *Saccharomyces* Genome Database (SGD, yeastgenome.org/; Cherry et al., 2012). Aiming to be inclusive when selecting genes to be included as MMR, genes in the KEGG diagram of the MMR pathway for each species were included and the gene ontology (GO) term "mismatch repair" was used to search for the genes on SGD and Pombase (Ashburner et al., 2000). We used both computationally and manually curated genes from SGD. We began curating our set of MMR genes in *S. cerevisiae*, with a total of 30 MMR genes identified with KEGG and SGD. Next, we searched KEGG and Pombase for genes in *S. pombe* that had not been annotated as part of the MMR pathway in *S. cerevisiae* (n = 15). We concluded by searching for *N. crassa* MMR genes in KEGG which had not already been categorized as MMR genes in the other two species (n =7). KEGG listed two sequences for the *N. crassa* gene *LIG1*; however, since our sequence similarity search analyses with both sequences yielded identical patterns of loss, we present them as one gene. This approach yielded a total of 52 genes associated with MMR (Table S2 from Phillips et al., 2021).

MMR gene conservation analysis

To examine the conservation of MMR genes across Ascomycota, we implemented a sensitive probabilistic modeling approach using profile Hidden Markov Models (pHMMs) (Johnson et al., 2010) of MMR genes across the genomes of 1,107 species (Shen et al., 2020b). To construct pHMMs, we first searched for putative homologs of MMR genes in the fungal RefSeq protein database using the blastp function of BLAST+, v2.8.1, with a bitscore threshold of 50 and an e-value cutoff of 1 x 10^{-3} (Pearson, 2013). We retrieved the top 100 hits using SAMTOOLS, v1.6 (Li et al., 2009b) with the 'faidx' function. We used MAFFT, v7.402 (Katoh et al., 2002), with the 'genafpair' and 'reordered' parameters, a maximum of 1000 cycles of iterative refinement, the BLOSUM62 matrix, a gap opening penalty of 1.0, and the retree parameter set to 1, to align the sequences following previously established protocol (Steenwyk et al., 2019b).We then used the aligned amino acid sequences as input to the 'hmmbuild' function in HMMER-3.1B2 to construct each pHMM. We ran the pHMMs of the 52 proteins against all 1,107 proteomes using the 'hmmsearch' function. For a gene to be considered present, we set a bitscore threshold of at least 50 and an e-value threshold of less than 1 x 10^{-10}. We used the TBLASTN function of BLAST+, v2.8.1 with a bitscore threshold of 50, e-value cutoff of 1 x 10^{-6}, and 50% minimum query coverage to verify MMR gene absence using the protein sequence of the gene in question and the 1,107 Ascomycota genomes. We used the Interactive Tree of Life (iTOL), v4 (Letunic and Bork, 2019) to visualize the conservation of MMR genes on the Ascomycota phylogeny and to map losses on it. To further verify gene absences in HLT and LLT with *C. variabilis* amino acid sequences, we used the TBLASTN function of BLAST+, v2.8.1 with an e-value threshold of less than 1 x 10^{-5}, a word size of 5 or more, and minimum query coverage of 80% (Milo et al., 2019).

Microsatellite identification and characterization

To identify microsatellites and evaluate their numbers and lengths between genomes with substantial MMR gene loss against those with higher levels of MMR gene conservation, we used the Microsatellite Identification tool (MISA), v2.0 (Beier et al., 2017). Specifically, we compared the microsatellites of two groups of taxa, each of which contained four species. The group of higher loss taxa (HLT) contains the powdery mildews *Blumeria graminis, Erysiphe necator, Erysiphe pisi,* and *Erysiphe pulchra,* which show high levels of MMR gene loss relative to other ascomycetes. The group of lower loss taxa (LLT) contains four closely related species with low levels of MMR gene loss, similar to patterns seen across the rest of the phylum: *Articulospora tetracladia, Ascocoryne sarcoides, Cairneyella variabilis,* and *Phialocephala scopiformis*. The length minimums used for MISA to identify a microsatellite are as follows: 1 base pair (bp) motifs must repeat 12 times, 2 bp motifs must repeat 6 times, 3-6 bp motifs must repeat 5 times. All values used are MISA defaults, except the mononucleotide parameter, which was increased from the default value of 10 repeats to 12 (Temnykh et al., 2001; Beier et al., 2017). A 2-way ANOVA test was performed to test for significance in the number of microsatellites controlled by genome size of each motif length between HLT and LLT. If the 2-way ANOVA rejected the null hypothesis ($\alpha = 0.05$), pairwise comparisons were made with the Tukey Honest Significant Differences (HSD) test. We performed the statistical analysis using R, v3.4.1 (https://www.r-project.org/) and made the figures using ggplot2, v3.1.0 (Wickerham, 2016), and ggpubfigs, v1.0.0 (Steenwyk, 2020).

Estimation of mutational bias and rate of sequence evolution

To characterize the mutational spectra and estimate the rate of sequence evolution between HLT and LLT, we first identified and aligned orthologous sequences across all eight genomes. Orthologous single-copy protein sequences from genes present in all eight genomes (n = 823) were identified using the BUSCO, v4.0.4 (Waterhouse et al., 2018b) pipeline and the OrthoDB, v10, Ascomycota database (Creation date: 2019-11-20) (Kriventseva et al., 2019). We hereafter refer to the 823 single-copy genes as BUSCO genes. BUSCO genes were aligned using MAFFT, v7.402 (Katoh et al., 2002), using the same settings described above. Codon-based alignments were generated by threading the corresponding nucleotide sequences onto the protein alignment using 'thread_dna' function in PhyKIT, v0.1 (Steenwyk et al., 2020a).

To examine patterns of substitutions, we used codon-based alignments to identify nucleotides that differed in a given taxon of interest compared to *C. variabilis*, which was the sister taxon to a clade comprised of the other seven genomes of interest in the Ascomycota phylogeny. More specifically, we compared the character states for a species of interest to *C. variabilis* for each site of each alignment, tracking codon position information (i.e., first, second, or third codon position). We also determined if the substitution was a transition or transversion and examined substitution directionality (e.g., A|T to G|C or G|C to A|T) using *C. variabilis* as the outgroup. These analyses were completed using custom python scripts that utilize functions in Biopython, v1.70 (Cock et al., 2009b).

Finally, we used the codon alignments to compare the rate of sequence evolution between HLT and LLT. Specifically, we measured the ratio of the rate of nonsynonymous substitutions to the

rate of synonymous substitutions (dN/dS or ω) along the species phylogeny for each gene using the CODEML function in PAML, v4.9 (Yang, 2007). For each test, the null hypothesis (H_o) was that all branches had the same ω value (model = 0); the alternative hypothesis (H_A) was that all HLT branches, including the branch of their most recent common ancestor, had one ω value and all other branches had a distinct ω value (model = 2). To determine if the alternative hypothesis was a better fit than the null hypothesis ($\alpha = 0.05$) we used a likelihood ratio test.

Data availability

Supporting statistical analysis, the Ascomycota phylogeny with species names, and 2 supplementary data files (MMR gene presence/absence matrix and ω output) are available via figshare at https://doi.org/10.6084/m9.figshare.14410994. The data supporting the phylogeny of Ascomycota are available at https://doi.org/10.6084/m9.figshare.12751736.

Results

MMR genes are highly conserved across the fungal phylum Ascomycota

By examining the presence of 52 MMR genes using a combination of sequence similarity search algorithms across the genomes of 1,107 fungal species, we found that the MMR pathway is highly conserved across Ascomycota (a median of 49 / 52 MMR genes per species; Fig. 34; File S1 from Phillips et al., 2021). Sixteen genes were present in all species; these included five recognition genes (*MSH1*, *MSH2*, *MSH3*, *MSH4*, and *MSH6*), one incision gene (*MLH1*), one removal gene (*DIN7*), five resynthesis genes (*CDC6*, *RFC2*, *RFC3*, *RFC4*, and *RFC5*), and all four ligation genes (*ADL1*, *CDC17*, *CDC9*, and *LIG1*). Few genes experienced extensive loss. Of

the 11 most commonly lost genes, which were lost in >5% of species, two (*MYH1* and *UVE1*) were lost in the common ancestor of Saccharomycotina, in addition to losses observed in other taxa. The remaining nine genes are unevenly distributed across functions; seven are involved in DNA resynthesis (*CDC27*, *CDM1*, *POL32*, *POLD3*, *POLD4*, *RPA3*, and *SSB3*), one is involved in mismatch recognition (*HSM3*), one is involved in incision (*HNT3*). These findings suggest that genes in the MMR pathway are well conserved across Ascomycota.

A comparison of our results with those reported in Milo et al. (2019) revealed similar patterns of gene presence and absence. For example, Milo et al. (2019) found that *MYH1* was absent from much of Pezizomycotina, which is consistent with our results. However, we did identify a few differences (inferred losses by Milo et al. (2019) vs. inferred presence in our analyses), which suggest that our pipeline is more conservative in classifying gene losses. We surmise that these differences stem from differences in the gene detection pipelines employed and the divergent objectives of the two studies; Milo et al. (2019) aimed to identify orthologs via a reciprocal best BLAST hit approach, whereas we aimed to identify homologs using pHMMs with absences verified using TBLASTN. Importantly, analysis of the human proteome using our pipeline detected all known human orthologs except for *HSM3*, *POL32*, *RPA3*, and *SSB3*, suggesting that our pipeline is well suited to detect MMR genes within Ascomycota, but that a few MMR genes may be more rapidly evolving and therefore more difficult to detect.

Extensive loss of MMR genes in a lineage of powdery mildews

Although MMR genes are highly conserved across Ascomycota, we found that a lineage of obligate plant parasite powdery mildews have among the fewest MMR genes of the 1,107

Ascomycota species examined. We further verified gene losses in HLT and in closely related taxa that experienced fewer losses (hereafter referred to as lower loss taxa or LLT) by performing TBLASTN using the amino acid sequences found by the pHMM for each MMR gene of *Cairneyella variabilis*, an LLT with a highly conserved MMR pathway, as queries and with thresholds described by Milo et al. (2019). This resulted in 9 MMR genes losses in *Erysiphe necator*, 21 in *Erysiphe pisi*, and 7 in *Erysiphe pulchra* (Fig. 35). *E. necator* has been previously documented to have a high rates of genome evolution (Milo et al., 2019) and genomic instability (Jones et al., 2014). *Blumeria graminis*, which is sister to the *Erysiphe* genus, has lost 5 MMR genes; previous studies reported extensive gene loss in diverse pathways in this species, generally in genes thought to be unnecessary for its biotrophic lifestyle (Spanu et al., 2010). In contrast, the closely related species *C. variabilis* and *Phialocephala scopiformis* only lack *MYH1*, an adenine DNA glycosylase that is lost in most filamentous fungi (Chang et al., 2001). In addition to *MYH1*, closely related species *Articulospora tetracladia* lacks *HSM3*, which has been lost in many Pezizomycotina genomes. *Ascocoryne sarcoides* lacks *MYH1* and *POL32*, a DNA polymerase δ subunit, which is part of a larger complex that participates in multiple DNA repair pathways, including nucleotide excision repair and base excision repair (Gerik et al., 1998). Much like the rest of the phylum, genes associated with resynthesis are lost more frequently, but *Erysiphe* and *Blumeria* have lost genes associated with all MMR functions (Table S2 from Phillips et al., 2021) except ligation. In addition, seven of the observed MMR gene losses occur nowhere else in Ascomycota: *EXO1* (excision), *MLH2* (incision), *MLH3* (incision), *MSH5* (recognition), *PMS1* (incision), *PMS2* (incision), and *RFC1* (resynthesis). Taken together, these results raise the hypothesis that HLT may have a partially functional MMR pathway.

While select *Erysiphe* taxa have lost more genes than any other species, there are other species with moderate to high levels of MMR gene loss across the phylum. A total of 318 species have lost 5 or more genes across Ascomycota: 5 species in subphylum Taphrinomycotina, 239 in Saccharomycotina, and 74 in Pezizomycotina. The disproportionate number of Saccharomycotina and Taphrinomycotina species is consistent with our knowledge that organisms in these lineages have, on average, a smaller number of DNA repair genes compared to Pezizomycotina (Milo et al., 2019). MMR gene loss in certain Saccharomycotina lineages, such as in some species from the genera *Hanseniaspora* (Steenwyk et al., 2019a), *Tetrapisispora,* and *Dipodascus*, is comparable to the loss observed in HLT. However, only 9 other species in Pezizomycotina showed MMR gene loss to the same degree as any *Erysiphe* species. In general, species with elevated levels of gene loss primarily lost genes noted as commonly lost earlier in this paper (see "MMR genes are highly conserved across the fungal phylum Ascomycota"), with occasional losses in other genes.

There was a notable discrepancy between the presence and absence of MMR genes inferred by pHMM versus TBLASTN in HLT that was not observed in other species. When measured by pHMM, *E. pulchra* lost 42 MMR genes, as opposed to 9 when using TBLASTN with model organism sequences as queries and 7 when using TBLASTN with *C. variabilis* sequences as queries to verify the absences. *E. necator* and *E. pisi* lost 51 MMR genes according to our pHMMs, as opposed to 10 and 22 by model organism TBLASTN and 9 and 21 with *C. variabilis* TBLASTN, respectively. *B. graminis* lost 9 MMR genes by pHMM, 6 by model organism TBLASTN, and 5 by *C. variabilis* TBLASTN. In the closely related LLT *C. variabilis*, *P. scopiformis, A. tetracladia,* and *A. sarcoides,* genes deemed absent by pHMMs were also

deemed absent in our model organism and *C. variabilis* TBLASTN searches; the sole exception was *HSM3*, which was identified as present in *A. sarcoides* using *C. variabilis* TBLASTN. Even though pHMMs are more sensitive in sequence similarity searches and typically outperform TBLASTN when detecting genes on an evolutionary timescale (Yoon, 2009), this discrepancy is likely explained by the lower annotation quality of some HLT species and the lower genome quality for *E. pisi* (Table S1 from Phillips et al., 2021).

Higher MMR gene loss taxa show increased number and length of microsatellites

Examination of microsatellites revealed microsatellite expansions in HLT in comparison to LLT. Specifically, we found statistically significant increases in the number and length of microsatellites in HLT compared to LLT (Fig. 36A, Tables 1 from Phillips et al., 2021, S3 from Phillips et al., 2021, and S4 from Phillips et al., 2021). Overall, after controlling for genome size, HLT had significantly more microsatellites than LLT ($F = 34.83$; $p < 0.001$; ANOVA; Table S3 from Phillips et al., 2021). This effect was driven by the very large increase in the number of mononucleotide runs in *Erysiphe* and *Blumeria* (Fig. 36C) ($p < 0.001$; Tukey HSD; Table S3 from Phillips et al., 2021). There was no statistically significant difference between the groups in the numbers of microsatellites with a 2-6 bp motif length (Table S3 from Phillips et al., 2021). HLT showed significantly higher average microsatellite lengths at every motif size than LLT (Fig. 36B) ($p < 0.01$ for 1 bp, $p < 0.001$ for all other motif lengths; Wilcoxon rank sum test; Table S4 from Phillips et al., 2021). HLT have an increased number of mononucleotide runs (after controlling for genome size) and an increase in length of microsatellites of all motif lengths, suggesting that the MMR pathway's function is compromised in these species.

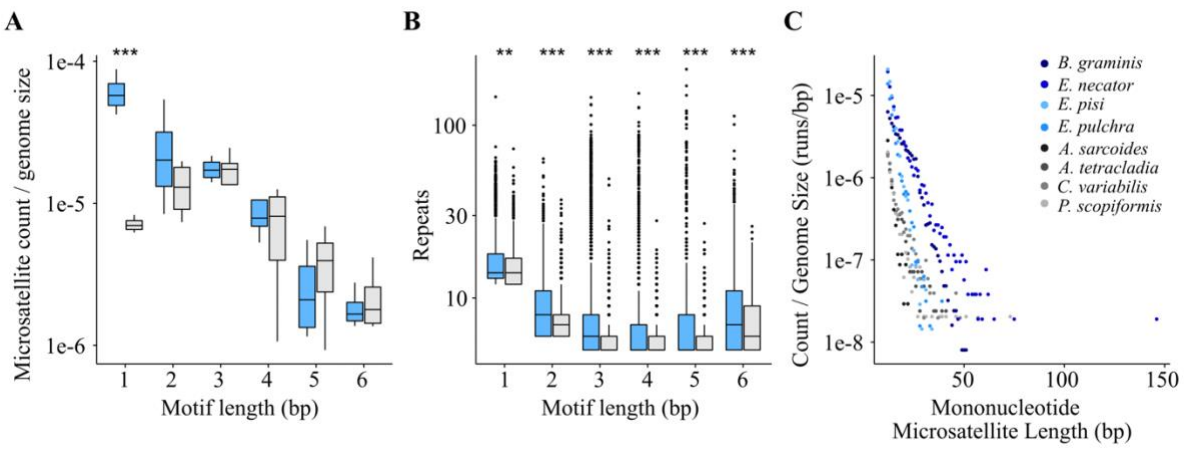

Fig. 36. Genomes of higher loss taxa (HLT; blue bars) show a proliferation of mononucleotide runs and an increase in their microsatellite lengths compared to lower loss taxa (LLT; grey bars). (A) Examination of microsatellites in HLT and LLT (gray bars) showed a significant increase in the number of mononucleotide runs in HLT ($p < 0.001$; ANOVA, Tukey HSD; Table S3 from Phillips et al., 2021). Asterisks in graph indicate significance; **: $p < 0.01$; ***: $p < 0.001$. (B) Microsatellites of each motif length are significantly longer in HLT ($p < 0.01$ for 1 bp, $p < 0.001$ for all other motif lengths; Wilcoxon rank sum test; Table S4). (C) Mononucleotide runs are longer and more numerous in HLT than LLT (Tables S3 from Phillips et al., 2021 and S4 from Phillips et al., 2021).

Higher loss taxa show mutational biases

By examining patterns of substitutions among HLT and LLT we found that HLT displayed stronger mutational biases associated with impaired DNA repair pathway function in comparison to LLT. For example, significantly more substitutions were observed at all codon positions in HLT vs. LLT (Fig. 36A) ($p < 0.01$; Tukey HSD; Table S5 from Phillips et al., 2021) and a significant bias towards substitutions in the A|T direction (Fig. 36B) ($p < 0.001$; Tukey HSD; Table S6 from Phillips et al., 2021). HLT also had a lower ratio of transitions to transversions (0.92 ± 0.04) than LLT (0.99 ± 0.02), though this is not statistically significant (Fig. 36C) ($p = 0.06$; Wilcoxon rank sum exact test; Table S7 from Phillips et al., 2021). Additionally, HLT had lower GC content (HLT: $40.10 \pm 0.02\%$ vs. LLT: $47.49 \pm 0.01\%$). Linear regression revealed a

significant decrease (F = 8.661, p = 0.026; Table S8 from Phillips et al., 2021) in GC content of the genomes as the number of MMR genes lost increases (Fig. 36D).

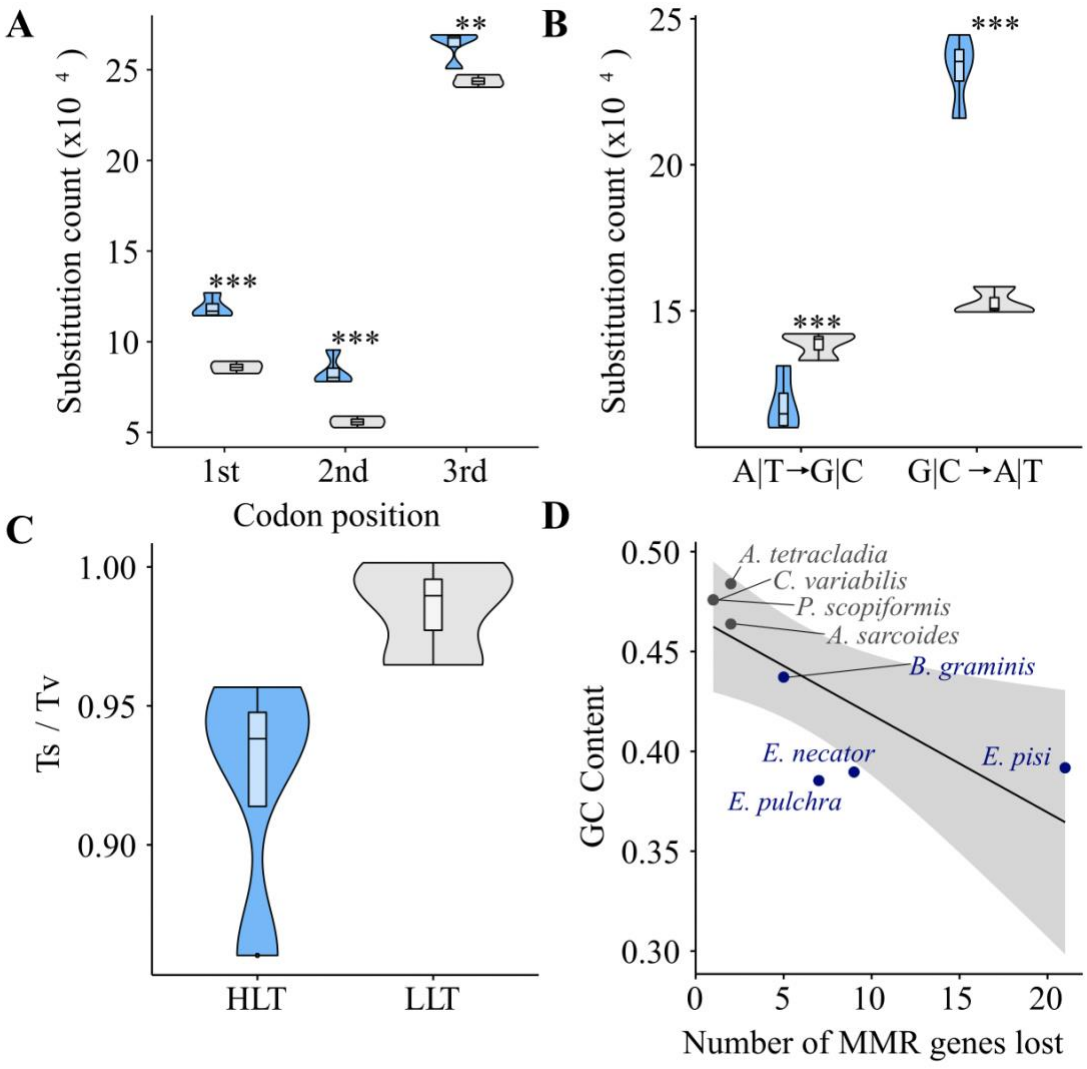

Fig. 37. Higher loss taxa (HLT) show diverse types of mutational bias compared to lower loss taxa (LLT).
(A) HLT (blue bars / fonts) show increased counts in base substitution at every codon position when compared to LLT (gray bars / fonts) ($p < 0.01$; ANOVA, Tukey HSD; Table S5 from Phillips et al., 2021). (B) HLT show significant mutational bias towards mutation in the A|T direction, while this trend is not significant in LLT ($p < 0.001$; $p = 0.27$; ANOVA, Tukey HSD; Table S6 from Phillips et al., 2021). (C) HLT show a decreased ratio of transitions to transversions when compared to LLT, although this difference is not statistically significant ($p = 0.06$; Table S7 from Phillips et al., 2021). (D) Genome GC content decreases with increasing

MMR gene loss (adjusted $R^2 = 0.5225$; $p = 0.026$; linear regression; Table S8 from Phillips et al., 2021). Asterisks in graphs indicate significance; **: $p < 0.01$; ***: $p < 0.001$.

Higher loss taxa have experienced accelerated rates of sequence evolution

To test if the rate of evolution of HLT differed from that of LLT, we performed ω-based branch tests. Our null hypothesis was that all branches of the phylogeny for our selected eight species had the same rate of evolution, while our alternate hypothesis posited that HLT branches, including the branch of their most recent common ancestor, experienced a different rate of sequence evolution than LLT branches. We found that 60.75% of genes rejected the null hypothesis ($\alpha = 0.05$; $n = 500$) and 39.25% failed to reject the null ($n = 323$) (Fig. 38A; File S2 from Phillips et al., 2021). Of the genes which rejected the null hypothesis, 79.80% ($n = 399$) experienced higher rates of substitution in HLT, which constitutes 48.48% of all genes tested (Fig. 6A). Among the genes that rejected the null hypothesis, the difference between the ω values for the HLT (median $\omega = 0.0899$) and the LLT (median $\omega = 0.0567$) showed accelerated rates of substitution for HLT branches (Fig. 38B). These results suggest that MMR gene loss is associated with a genome-wide signature of accelerated mutation rates.

Discussion

Using sequence similarity searches, we examined the conservation of the MMR pathway across 1,107 ascomycete species. The near universal conservation of the vast majority of MMR genes across the phylum confirms this pathway's known critical role for DNA maintenance (Schofield and Hsieh, 2003; Kunkel and Erie, 2005; Fukui, 2010). However, we also discovered that a lineage of *Erysiphe* and *Blumeria* powdery mildews, named HLT, have experienced significant

MMR gene loss (Fig. 35). HLT exhibit increases in mononucleotide run count, microsatellite length, mutational biases, and rate of evolution (Figs. 36, 37, and 38), suggesting that the

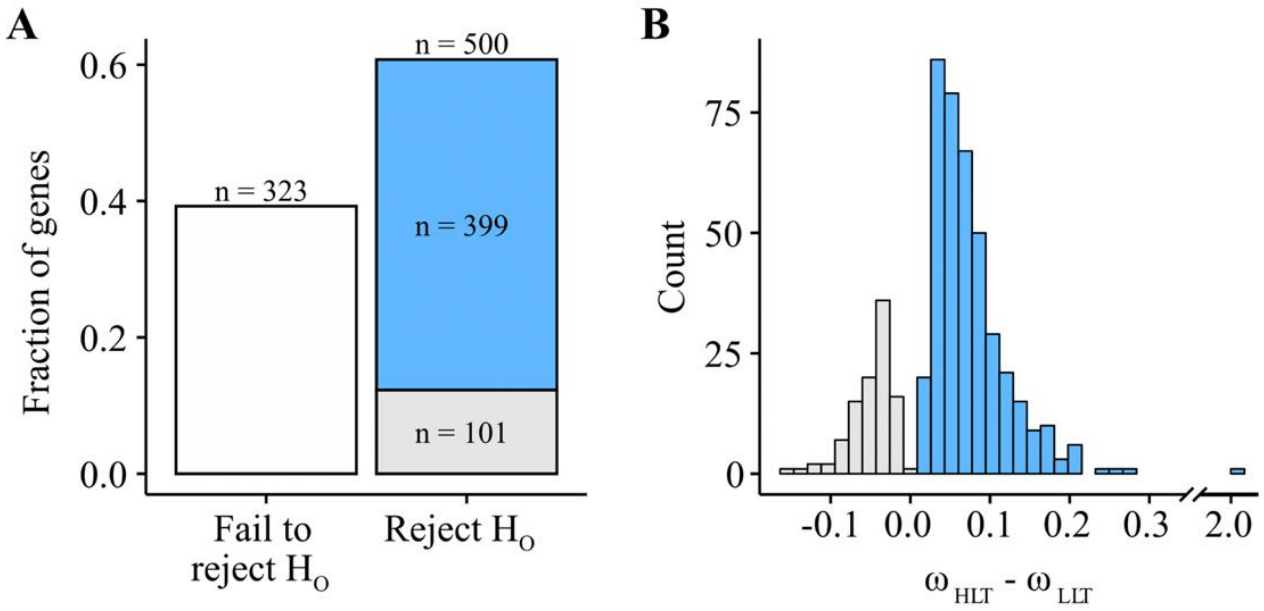

Fig. 38. Powdery mildew higher loss taxa (HLT) show accelerated rates of evolution. (A) Most (60.75%; n = 500) BUSCO genes reject the null hypothesis that HLT branches experience the same rate of substitution as LLT branches. 48.48% of BUSCO genes support a higher rate of evolution for HLT (n = 399; in blue), 12.27% support a higher rate of evolution for LLT (n = 101; in grey), and 39.25% (n = 323) fail to reject the null hypothesis that the rate of substitution is uniform across HLT and LLT branches (in white). Among genes that support the alternative hypothesis, 79.80% (n = 399) support that *Erysiphe* and *Blumeria* evolve more quickly than LLT. (B) Among genes which reject the null hypothesis, the distribution of differences between ω values for HLT and LLT branches show elevated substitution rates in HLT.

function of their MMR pathway may be impaired. While DNA repair mechanisms are present in all eukaryotes and are highly conserved, there is mounting evidence of exceptions to this rule in the fungal kingdom (Steenwyk et al., 2019a; Steenwyk, 2021b). The increased MMR gene absence observed in HLT correlates with changes in the microsatellite compositions of their genomes. The significant difference between HLT and LLT in the number of mononucleotide

runs is consistent with mutational patterns present in human cancers and MMR deficient yeast cells (Arzimanoglou et al., 1998; Lang et al., 2013). Mononucleotide runs are the most prone to replication fork slippage and are used to diagnose MSI in tumors (Richman, 2015). In addition to an increase in the number of mononucleotide runs (Fig. 36A), HLT showed significantly longer microsatellites for each motif length than LLT (Fig. 36B), which suggests impaired MMR function and increased replication fork slippage.

Examination of HLT genomes revealed mutational signatures suggesting that the MMR pathway has been impaired by the observed gene losses. Patterns of substitutions suggest the loss of MMR genes leads to increased substitution rates (Fig. 37A) and lower GC content (Fig. 37D). More specifically, the prominent A|T bias of substitutions in the HLT is likely driven by the known A|T bias of mutations previously observed in bacteria and eukaryotes, including *S. cerevisiae* (Keightley et al., 2009b; Hershberg and Petrov, 2010; Lynch, 2010; Zhu et al., 2014b). Furthermore, GC content decreases in proportion to the number of MMR genes lost in the HLT and LLT, which was also observed among *Hanseniaspora*, a lineage of budding yeasts that have lost diverse DNA repair genes (Steenwyk et al., 2019a). There is no significant difference between the transition to transversion (Ts/Tv) ratios of the HLT (0.92 ± 0.04) and LLT (0.99 ± 0.02), though the trend follows what we would expect for HLT having less efficient DNA repair. Both lineages exhibit near-neutral Ts/Tv ratios (Lynch et al., 2008b; Zhu et al., 2014b). Examination of ω values suggests that faster rates of sequence evolution in HLT compared to the LLT may be associated with MMR gene loss. Long branches, which reflect more substitutions per site, have been previously reported elsewhere for *E. necator* (Milo et al., 2019), providing independent support to our findings.

Species in the LLT lineage show a diversity of ecologies. For example, *A. sarcoides* is saprobic fungus which grows on trees (Gianoulis et al., 2012), *A. tetracladia* is a globally-distributed aquatic hyphomycete (Seena et al., 2012), *C. variabilis* is an ericoid mycorrhizal fungus (Midgley et al., 2016), and *P. scopiformis* is a foliar endophyte (Walker et al., 2016). In contrast, species in the genera *Erysiphe* and *Blumeria* are all powdery mildews and obligate plant parasites. *B. graminis* is the only species in the genus *Blumeria* (Inuma et al., 2007), whereas there are ~450 known species in the genus *Erysiphe* (Takamatsu et al., 2015). The *Erysiphe* species we sampled are distributed across the phylogeny of the genus (Takamatsu et al., 2015; Ellingham et al., 2019); phylogenetic analyses by Takamatsu et al. (2015) placed *E. pisi* and *E. pulchra* in separate phylogenetic groups that diverged 15-20 million years ago, and analyses by Ellingham et al. (2019) showed that *E. pisi* and *E. necator* are distantly related (Ellingham et al., 2019). Considering our taxon sampling and obligate plant parasitic lifestyle of *Erysiphe* species, we hypothesize that our findings likely apply to all species in the genus.

In our approach to investigate the conservation of the MMR pathway in Ascomycota, we chose to be inclusive when selecting genes that function as part of the pathway and conservative when ruling genes absent. Some genes were computationally annotated as belonging to the MMR pathway based on sequence homology and others are implicated in multiple pathways. That said, of the 24 MMR genes lost in at least one HLT species, 16 are included in the KEGG MMR pathway entries for our three model organisms, suggesting that many of the observed losses concern genes directly involved in MMR. In addition, there may be other contributing factors to the observed mutational differences between HLT and LLT, such as dysfunction in other DNA

repair pathways, loss of methyl transferases, and differences in chromatin structure (Steenwyk et al., 2019a; Möller et al., 2021). For example, previous studies have identified the loss of genes involved in the repeat-induced point (RIP) mutation pathway in powdery mildews (Spanu et al., 2010; van Wyk et al., 2021). Loss of genes in the RIP pathway in HLT could contribute to the elevated mutation rates we observed relative to LLT. The MMR pathway is required for maintaining heterochromatin stability in *S. cerevisiae*; dysfunction in this pathway could potentially lead to genome instability within the HLT (Dahal et al., 2017). MMR is more error prone in heterochromatin than euchromatin, likely due at least in part to mechanical accessibility of the MMR machinery (Sun et al., 2016; Dahal et al., 2017). While base-base mismatches are repaired less efficiently in heterochromatin than in euchromatin, single nucleotide insertions and deletions are repaired with similar efficiency in euchromatin and heterochromatin (Dahal et al., 2017); therefore, differences in chromatin structure could have contributed to some of the observed mutational differences but not to the observed differences in mononucleotide runs and microsatellite repeats.

Loss of function in the MMR pathway may be advantageous in certain environments or under certain lifestyles. For example, strains of human pathogens with impaired MMR function are found in environments where antifungal drugs are present. Some of these strains have evolved drug resistance, so the elevated mutation rate generated by MMR gene loss may be adaptive under certain stressful situations (Billmyre et al., 2017; Rhodes et al., 2017a; Billmyre et al., 2020). Species with higher levels of MMR gene loss may be associated with a parasitic lifestyle, though not all parasites have high levels of MMR gene loss; dysfunction in this pathway may be more adaptive, or at least less detrimental, to these organisms, as seen in the HLT. Loss of DNA

repair pathways is also observed in other parasites and may contribute to elevated mutation rates (Gill and Fast, 2007; Derilus et al., 2021). Organisms with parasitic lifestyles tend to evolve more rapidly than free-living organisms; while these mechanisms are unknown, previous work has identified that the loss of the classical nonhomologous end joining (C-NHEJ) pathway is common in this lifestyle and may even be a contributing factor (Nenarokova et al., 2019). Previous studies of genome structure in *E. necator* have found genome expansion largely driven by transposable elements and suggest that genome instability, particularly in copy number variants, can mediate rapid evolution of fungicidal resistance (Jones et al., 2014). The evolution of fungicide resistance in powdery mildews has implications for agriculture; major crops are impacted by these pathogens and some are able to quickly evolve resistance to antifungal chemicals, with resistance evolution accelerated by increased use (Jones et al., 2014; Vielba-Fernández et al., 2020). More broadly, genome instability among HLT taxa reflects their parasitic lifestyle, which is associated with gene loss and plastic genomic architecture (Schmidt and Panstruga, 2011). Gene loss in primary and secondary metabolism, enzymes acting on carbohydrates, and transporters has been documented in *B. graminis*, as well as massive expansion in retrotransposons and genome size, reflecting extreme genomic changes associated with its parasitic lifestyle (Spanu et al., 2010). The lost genes are involved in diverse pathways, including anaerobic fermentation, biosynthesis of glycerol from glycolytic intermediates, and nitrate assimilation, and include multiple subfamilies of transporters (Spanu et al., 2010). Given their extreme genomic changes and importance to agriculture, *Blumeria* and *Erysiphe* may be novel models to study the outcome and evolutionary trajectory of sustained loss of MMR pathways.

CHAPTER 9

Pathogenic allodiploid hybrids of *Aspergillus* fungi[8]

Introduction

Interspecific hybridization can result in the formation of new species that substantially differ in their genomic and phenotypic characteristics from either parental species. One common mechanism by which interspecific hybrids can originate is allopolyploidy, which merges and multiplies the parental species' chromosomes (Baack and Rieseberg, 2007; Abbott et al., 2013). Allopolyploid hybrids may be more similar to one parent in some traits, reflect both parents in others, or may differ from both in the rest. Hybrids' distinct phenotypic profiles means that they can potentially colonize new habitats (Rieseberg, 2003; Rieseberg et al., 2007), whereas their polyploidy means that they can quickly become reproductively isolated from both parental species, forming a new species in the process (Baack and Rieseberg, 2007). Allopolyploids are relatively common in plants, but are also found in several other lineages, including in animals (Mable et al., 2011) and fungi (Dunn and Sherlock, 2008; Depotter et al., 2016).

Among fungi, the most well-known example of allopolyploidy is the whole genome duplication in an ancestor of the baker's yeast *Saccharomyces cerevisiae* (Wolfe and Shields, 1997; Marcet-Houben and Gabaldón, 2015; Wolfe, 2015). Importantly, allopolyploidy is known from both fungal pathogens of plants (Depotter et al., 2016; Stukenbrock, 2016) and of animals

[8]This work is published in: Steenwyk, J. L., Lind, A. L., Ries, L. N. A., dos Reis, T. F., Silva, L. P., Almeida, F., et al. (2020). Pathogenic Allodiploid Hybrids of *Aspergillus* Fungi. Curr. Biol. 30, 2495-2507.e7. doi:10.1016/j.cub.2020.04.071.

(Lin et al., 2009; Mixão and Gabaldón, 2018). For example, the crucifer crop pathogens *Verticillium longisporum* and *Verticillium dahliae* are allopolyploid hybrids, as is the onion pathogen *Botrytis allii* (Nielsen and Yohalem, 2001; Inderbitzin et al., 2011; Depotter et al., 2016). Among fungal pathogens that infect humans, allopolyploidy has been reported in the ascomycete budding yeasts *Candida metapsilosis* (Pryszcz et al., 2015) and *Candida orthopsilosis* (Schröder et al., 2016) and in the basidiomycete yeast *Cryptococcus neoformans* x *Cryptococcus deneoformans* (Rhodes et al., 2017b). To our knowledge, allopolyploidy has never been reported in human pathogenic filamentous fungi.

Aspergillus-related diseases, collectively known as aspergillosis, are caused by various species in the *Aspergillus* genus of filamentous fungi (Barnes and Marr, 2006). Although the saprophytic and ubiquitous airborne species *Aspergillus fumigatus* (section *Fumigati*) is responsible for most infections, several other species are also pathogenic (Brown and Goldman, 2016; Paulussen et al., 2017; van de Veerdonk et al., 2017; Rokas et al., 2020a). One such species is *A. nidulans* (section *Nidulantes*); even though rarely pathogenic, *A. nidulans* is of interest because it is a major cause of invasive aspergillosis infections in chronic granulomatous disease (CGD) patients (Henriet et al., 2012). CGD is a genetic disorder that compromises the ability of phagocytes to produce reactive oxygen species, which act as broad range antimicrobial chemicals (Fang, 2011; Henriet et al., 2012). Strikingly, among CGD patients, *A. nidulans* is harder to treat than the more common pathogen *A. fumigatus* (Dotis and Roilides, 2004; Henriet et al., 2012).

To gain insights into *A. nidulans* pathogenicity, we sequenced 9 clinical isolates that were originally identified as *A. nidulans* from patients with various pulmonary diseases, including 2

isolates from CGD patients. Two of these isolates belong to *A. nidulans* and have been described in detail elsewhere (Bastos et al., 2020a). However, our phenotypic and genomic analyses showed that 6 of the remaining 7 isolates are not *A. nidulans* but rather allodiploid hybrids of *Aspergillus latus*, a species that arose from allodiploid hybridization between *Aspergillus spinulosporus* and an unknown second parental species closely related to *Aspergillus quadrilineatus* (both from section *Nidulantes*). Our analyses also revealed that the seventh isolate belongs to *A. spinulosporus*.

Phenotypic characterization of *A. latus* isolates, their parental species, and *A. nidulans* for diverse infection-relevant traits revealed two key findings. The first finding is that *A. latus* isolates exhibit strain heterogeneity for several infection-relevant traits. For example, we observed wide variation between *A. latus* isolates in their virulence in a disease model as well as in their interactions with human immune cells. The second finding is that *A. latus* isolates are phenotypically distinct from *A. spinulosporus*, *A. quadrilineatus*, and *A. nidulans*. For example, we found that *A. latus* hybrid spores are better at evading macrophage engulfment as well as evading hyphal killing and inhibition of germination by neutrophils than *A. nidulans* or *A. spinulosporus* and are more resistant to antifungals and oxidative stressors than *A. nidulans* and *A. quadrilineatus*. From a clinical perspective, our discovery of allodiploid fungal pathogens of humans suggests that accurate taxonomic identification of *Aspergillus* clinical isolates is a key first step to disease management. From an evolutionary perspective, our discovery suggests that allodiploid hybridization is a general mechanism of genomic and phenotypic diversification among human fungal pathogens.

Materials and Methods

Fluorescence-assisted cell sorting for DNA content determination

Asexual spores (conidia) were collected, centrifuged (13,000 rounds per minute for 3 minutes) and washed with sterile 1 x phosphate-buffered saline (8 grams sodium chloride, 0.2 grams potassium chloride, 1.44 grams disodium phosphate, 0.24 grams monopotassium phosphate per liter of sterilized water). For cell staining, the protocol described by Almeida *et al.* (Almeida et al., 2007) was followed with modifications. Following overnight fixation with 70% ethanol (volume / volume) at 4ºC, spores were harvested, washed and suspended in 850 microliter of sodium citrate buffer (50 millimolar sodium citrate; pH=7.5). Spores were subsequently sonicated using four ultrasound pulses at 40W for 2 seconds with an interval of 1 to 2 seconds between pulses. Sonicated spores were treated for 1 hour at 50ºC with RNase A (0.50 milligrams / milliliter; Invitrogen, Waltham, Massachusetts, USA) and for 2 hours at 50ºC with proteinase K (1 mg/mL; Sigma-Aldrich, St. Louis, Missouri, USA). Spores were stained overnight with SYBR Green 10,000x (Invitrogen™, Carlsbad, CA, USA) diluted 10-fold in Tris-ethylenediaminetetraacetic acid (pH 8.0), at a concentration of 2% (volume / volume) at 4ºC. Finally, Triton® X-100 (Sigma-Aldrich) was added to samples at a final concentration of 0.25% (volume / volume). Stained spores were analyzed in a Fluorescence-assisted cell sorting LSRII flow cytometer (Becton Dickinson, NJ, USA) with a 488 nanometer excitation laser. Signals from a minimum of 30,000 cells per sample were captured in fluorescein isothiocyanate channel (530 nanometers ± 30 nanometers) at low flow rate of ~1,000 cells / second and an acquisition protocol was defined to measure forward scatter and side scatter on a four-decade logarithmic scale and green fluorescence on a linear scale. Fluorescence-assisted cell sorting DIVA was used

as the acquisition software. Results were analyzed with the R package FLOWVIZ, version 1.46.1 (Sarkar et al., 2008).

Asexual spore size measurements

The diameters of 100 spores for each isolate were measured under a Carl Zeiss (Jena, Germany) AxioObserver.Z1 fluorescent microscope equipped with a 100-W HBO mercury lamp using the 100 x magnification oil immersion objective and the AXIOVISION, software v.3.1.

DNA extraction and sequencing

Frozen mycelia of all isolates were ground in liquid nitrogen and genomic DNA was extracted as previously described (Malavazi and Goldman, 2012; Mead et al., 2019b). Standard techniques for manipulation of DNA were used (J. Sambrook, D.W. Russell, 2001). The genomes of all clinical isolates and the type strain of *A. latus* (9 in total) were sequenced at the Genomic Services Lab of Hudson Alpha (Huntsville, Alabama, USA) on an Illumina HiSeq 2500 sequencer; the sole exception was *A. quadrilineatus* NRRL 201^{T}, which was sequenced using on a NovaSeq S4 at the Vanderbilt Technologies for Advanced Genomes facility (Nashville, Tennessee, USA). All isolates were sequenced using paired-end sequencing (150 bp) with the Illumina TruSeq library kit. Additionally, the type strain of *A. latus* and the clinical isolates MM151978 and NIH were also sequenced using mate-pair sequencing (150 bp) using the Illumina Nextera Mate Pair Library kit with an insert size of 4 kilobases. The genome coverage of each isolate was greater than 150X. Both the raw short-read sequence data and the genome assemblies are publicly available (see File S4 for accession numbers).

Genome assembly and annotation

All genomes were assembled with the iWGS pipeline (Zhou et al., 2016) using DIPSPADES, version 1.0 (Safonova et al., 2015) or SPADES, version 3.6.2 (Bankevich et al., 2012). Optimal *k*-mer lengths were selected using KMERGENIE, version 1.6982 (Chikhi and Medvedev, 2014), and assembly quality was evaluated using QUAST, version 3.2 (Gurevich et al., 2013). Protein-coding genes were predicted using AUGUSTUS, version 3.3 (Stanke and Waack, 2003), using *Aspergillus nidulans* gene annotation parameters. Predicted genes in each *A. latus* hybrid genome were annotated by reciprocal-best-BLAST against a database of all *A. nidulans* and *A. spinulosporus* proteins.

Prediction of secondary metabolic gene clusters

Secondary metabolic gene clusters (SMGCs) were predicted in all assembled genomes using ANTISMASH, version 3.0.5.1 (Weber et al., 2015). In addition, we used literature-curated SMGC annotations from the well-characterized *A. nidulans* A4 genome (Galagan et al., 2005) to identify SMGCs not captured by the ANTISMASH software.

Assigning genes in hybrid genomes to parents of origin

To determine the most likely parent-of-origin for each gene in a hybrid genome, we measured the sequence divergence between every gene in a hybrid genome and its ortholog in the *A. spinulosporus* NRRL 2395^{T} parent (Ortiz-Merino et al., 2017). Specifically, for each gene in a hybrid genome, we used BLASTP, version 2.3.0 from NCBI's BLAST+ package (Madden, 2013), to identify the putative *A. spinulosporus* ortholog. The resulting pair was then aligned using MAFFT, version 7.294b (Katoh and Standley, 2013), with the BLOSUM 62 matrix of substitutions

(Mount, 2008), a gap penalty of 1.0, 1,000 maximum iterations, a single guide tree, and the 'genafpair' parameter. The associated DNA sequences were then forced onto the protein alignment using PAL2NAL, version 14 (Suyama et al., 2006), and the synonymous substitution rate *Ks* was calculated according to LWL85m method using YN00 module in PAML, version 4.9 (Yang, 2007). By examining the bimodal distribution of *Ks* values, genes with relatively low *Ks* values ($0 \leq Ks < 0.05$) were inferred to be from the *A. spinulosporus* parent genome and genes with high *Ks* values ($0.05 \leq Ks < 10$) were inferred to stem from the *A. quadrilineatus*-like parent genome. We performed the same analysis on the known, non-hybrid haploid genome of *Aspergillus fumigatus* strain A1163 compared to *A. fumigatus* strain Af293 (Nierman et al., 2005) and on the known hybrid diploid genome of *Zygosaccharomyces parabailii* strain NBRC1047/ATCC56075 compared to their closest homologs in the known parent *Zygosaccharomyces bailii* strain CLIB213 (Ortiz-Merino et al., 2017) as controls.

To determine the completeness of each parental genome and evaluate the efficacy of assigning genes in hybrid genomes to parent-of-origin, we created separate proteome FASTA files for genes from the *A. spinulosporus* and *A. quadrilineatus*-like regions of the genome and evaluated how many near-universally single copy orthologous genes were present in each parental genome using the BUSCO pipeline (Simão et al., 2015) with the ASCOMYCOTA database (creation date: 2016-02-13, number of species: 75, number of BUSCOs: 1315) from ORTHODB, version 9 (Waterhouse et al., 2013).

To determine homeolog pairs between each parental genome, we used a reciprocal best blast hit approach. Specifically, we used the FASTA files of genes created in the previous step and

conducted a reciprocal best blast hit analysis between the two parental genomes using an e-value cutoff of 10^{-3}. To determine if one gene was putatively pseudogenized, we employed a previously established approach, which compares gene lengths between the two homeologs; genes whose length is substantially shorter (we used an upper threshold of 80%) than their homeolog pair are inferred to be pseudogenes (Ortiz-Merino et al., 2017).

Maximum likelihood phylogenetic and phylogenomic analyses

To establish the taxonomic identity of the sequenced isolates, we retrieved their β-tubulin and calmodulin sequences by performing a nucleotide BLAST of the *Aspergillus nidulans* A4 β-tubulin and calmodulin sequences against each assembled genome. In addition, β-tubulin and calmodulin sequences from two strains of each of *Aspergillus foveolatus, Aspergillus latus, Aspergillus nidulans, Aspergillus pachycristatus, Aspergillus rugulosus, Aspergillus spinulosporus,* and *Aspergillus striatus* and from one strain of *Aspergillus corrugatus* were retrieved from GenBank. We also retrieved the same sequences from one strain of *Aspergillus sydowii,* which served as an outgroup for the phylogeny (Chen et al., 2016). Sequences were aligned with MAFFT, version 7.310 (Katoh and Standley, 2013), gaps were removed with TRIMAL, version 1.2.rev59 (Capella-Gutierrez et al., 2009), using the 'gappyout' parameter, and the β-tubulin and calmodulin sequences for each individual strain were concatenated. A phylogeny was constructed from the concatenated sequences with RAxML, version 8.2.11 (Stamatakis, 2014b), with the GTRGAMMAX model and 1,000 rapid bootstrap replicates. Branches with bootstrap support less than 80 were collapsed.

To determine the number of hybridization events that gave rise to *A. latus* isolates, we conducted phylogenomic analyses to reconstruct the evolutionary history of the *A. spinulosporus* parental genome in the *A. latus* hybrids and of the genes of *A. spinulosporus* strains NRRL2395 and 4060. We first identified single-copy orthologous genes across the 9 strains using ORTHOFINDER, version 2.3.8 (Li et al., 2003) with default parameters, which employs a Markov clustering algorithm (van Dongen, 2000) on gene sequence similarity information derived from an 'all-vs-all' approach using NCBI's BLAST+, version 2.3.0. Out of the inferred 12,596 groups of orthologous genes, 5,894 were single-copy (i.e., each of the 9 isolates were represented by a single sequence). All 5,894 sets of corresponding nucleotide sequences were aligned using MAFFT, version 7.402 (Katoh and Standley, 2013), with the 'genafpair' parameter and 1,000 maximum iterative sequence alignment refinements. Alignments were trimmed using TRIMAL, version 1.2rev59 (Capella-Gutierrez et al., 2009), using the 'gappyout' parameter. The resulting sequences were concatenated into a single nucleotide data matrix (8,405,004 sites). The strain phylogeny was inferred using IQTREE, version 1.6.1 (Nguyen et al., 2015), with the nbest parameter set to 10 to increase the number of best trees used during the search. Bipartition support was evaluated using 5,000 ultrafast bootstrap approximation replicates (Hoang et al., 2018). The evolutionary history of the *A. quadrilineatus*-like parental genome among the *A. latus* hybrids and two *A. quadrilineatus* strains (NRRL201^{T} and CBS 853.96) was inferred using the same approach; in this case, the nucleotide data matrix was comprised of 7,385,465 sites (from 5,079 single-copy orthologous genes out of a total of 11,814 groups of orthologous genes).

Lastly, we conducted approximately unbiased topology tests (Shimodaira, 2002) using the data matrices from each parental genome with the 'au' parameter in IQTREE, version 1.6.1 (Nguyen et

al., 2015). Specifically, we examined if the two topologies among the 7 *A. latus* hybrids inferred using the *A. spinulosporus* or *A. quadrilineatus*-like parental genomes were significantly different for either genome-scale data matrix inferred from the parental genomes. During constrained tree search, we used a substitution model of a general time-reversible model with empirical base frequencies, a discrete Gamma model with 4 rate categories, and allowed for a proportion of invariable sites (GTR+I+F+G4) (Tavaré, 1986; Yang, 1994, 1996; Vinet and Zhedanov, 2011).

Examination of loss of heterozygosity using copy number variation analysis

To conservatively and accurately instances of loss of heterozygosity, we measured copy number variation using CONTROL-FREEC, version 9.1 (Boeva et al., 2012) and CNVNATOR, version 0.3.2 (Abyzov et al., 2011). More specifically, we evaluated false discovery rate (FDR) and false positive rate (FPR) using reads from *A. spinulosporus* that were aligned to a "concatenated" genome of the *A. spinulosporus* and *A. nidulans* FGSC-A4 using 10 different window sizes (500, 750, 1000, 1250, 1500, 1750, 2000, 3000, 4000, and 5000 base pairs). Reads were aligned with BWA-MEM, version 0.7.12 (Li, 2013), and duplicates reads were removed with PICARDTOOLS, version 2.1.1 (http://broadinstitute.github.io/picard/). In this way, we inferred the window size parameter that minimizes the FDR and FPR CN variable regions identified in the hybrid genomes. Specific parameters used for CONTROL-FREEC include a minimum and maximum expected GC-content of 0.3 and 0.5, respectively, and a telocentrometric parameter of 10,000. Default parameters were used for CNVNATOR. To identify statistically significant CN variable loci, we implemented a Wilcoxon Rank Sum test (Wallace, 2004) and a Kolmogorov-Smirnov test (Panchenko, 2006), in the case of CONTROL-FREEC and a T-test in the case of CNVNATOR.

Using the resulting set of significant CN variants per window size for each program, FDR and FPR were calculated using the following formulas:

$$FDR = 1 - {}^{TP}/_{(TP + FP)}$$

$$FPR = {}^{FP}/_{(FP + TN)}$$

where *TP* represents true positives, *FP* represents false positives, and *TN* represents true negatives. Across the 10 different window sizes, we found that CONTROL-FREEC often, but not always, slightly outperformed CNVNATOR (Fig. S10 from Steenwyk et al., 2020c). More importantly, we found that CONTROL-FREEC had an FDR and FPR of 0 (i.e., had no false negatives or false positives) when using a window size of 1000 and 1250. Therefore, we used CONTROL-FREEC with a window size of 1000 to identify CN variable loci in all other isolates.

Macrophage isolation

To obtain macrophages for a phagocytosis assay of *Aspergillus* asexual spores (conidia), we used bone marrow-derived macrophages that were isolated as described previously (Weischenfeldt and Porse, 2008). Briefly, macrophages were recovered from femurs of C57BL/6 wild-type mice (6-weeks old) and were incubated in an RPMI medium (Gibco) supplemented with 30% (volume / volume) L929 growth conditioning media, 20% inactivated fetal bovine serum (Gibco), 2 millimolar glutamine and 100 units / milliliter of penicillin-streptomycin (Life Technologies). Fresh media was added after 4 days of cultivation and macrophages were collected after 7 days and used for subsequent experiments.

In vitro phagocytosis by macrophages

Phagocytosis of asexual spores (conidia) by wild-type macrophages were carried out according as previously described with modifications (Bom et al., 2015). 24-well plates containing a 15-mm-diameter coverslip in each well (phagocytosis assay) and $2x10^5$ macrophages per well were incubated with 1 ml of RPMI-FBS [(RPMI medium (Gibco) supplemented with 10% inactivated fetal bovine serum (Gibco), 2 millimolar glutamine and 100 units / milliliter of penicillin-streptomycin (Life Technologies)] at 37ºC, 5% carbon dioxide for 24 hours. Wells were washed with 1 milliliter of phosphate-buffered saline before the same volume of RPMI-FBS medium supplemented with $1x10^6$ spores (1:5 macrophage / spore ratio) was added in the same conditions.

To determine phagocytosis, macrophages were incubated with spores for 1.5 hours before supernatant was removed and 500 µl of phosphate-buffered saline containing 3.7% formaldehyde was added for 15 minutes at room temperature. Sample coverslips were washed with 1 milliliter of ultrapure water and incubated for 20 minutes with 500 microliters of 0.1 milligrams / milliliter calcofluor white to stain the cell wall of non-phagocytosed spores. Samples were washed and coverslips were viewed under a Zeiss Observer Z1 fluorescence microscope. In total, 100 spores were counted per sample and the phagocytosis index was calculated. Experiments were performed in biological triplicates.

Viability of *Aspergillus* hyphae

Human neutrophils from fresh venous blood of healthy adult volunteers were isolated according to a previous study with slight modifications (Drewniak et al., 2013), through centrifugation over

isotonic Percoll, lysed, and resuspended in 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid-buffered saline solution. Since we did not observe any neutrophil-mediated killing of *Aspergillus* asexual spores, as also previously described (Gazendam et al., 2016), we followed the protocol previously reported for neutrophil-mediated inhibition of germination. *Aspergillus* asexual spores were incubated with neutrophils (0.5, 1.0, or $2.5x10^5$ cells / milliliter; effector : target cell ratios of 1:1000, 1:500, or 1:200, respectively) in a 96-well plate overnight at 37ºC in RPMI 1640 medium containing glutamine and 10% fetal calf serum (Life). The neutrophils were lysed in water / sodium hydroxide (pH 11.0) and spore germination was determined using an MTT (thiazolyl blue; Sigma-Aldrich) assay as previously reported (Dos Reis Almeida et al., 2011). Each isolate's viability was calculated relative to incubation without neutrophils, which was set at 100% for each isolate and evaluated separately. To evaluate the viability of *Aspergillus* hyphae in the presence of neutrophils, we used a previously described protocol (Gazendam et al., 2016). *Aspergillus* asexual spores were incubated overnight at 37ºC in RPMI 1640 medium containing glutamine and 10% fetal calf serum (Life) upon formation of a monolayer, as verified by microscope. Freshly isolated neutrophils (0, 1.0, 2.0, or 3.0 x 105 cells / milliliter) were cultured for 1 hour on the *Aspergillus* monolayer at 37ºC. Subsequently, the cells were lysed in water / sodium hydroxide (pH 11.0) and the MTT assay was performed. Each isolate's hyphal viability was calculated as a percentage of its viability after incubation without neutrophils. The experiments were repeated three times, each performed in triplicate.

NETosis assays

Human polymorphonuclear cells (PMN) were isolated from 8 mL of peripheral blood of adult male healthy volunteers by density centrifugation using Mono-Poly™ Resolving Medium (MP

Biomedicals LLC, Irvine, CA, USA) according to the manufacturer's instruction. PMN ($5x10^6$ / mL) were resuspended in Hank's Balanced Salt Solution, without calcium or magnesium, containing 5% FBS (Gibco®, South American, Brazil). For flow cytometry analysis, 100 µL ($5x10^5$ / tube) of PMN were seeded in sterile round-bottom polystyrene tubes, stimulated with 10 nM of phorbol 12-myristate 13-acetate (PMA) (Sigma-Aldrich, St. Louis, MO, USA) or fungi samples ($5x10^5$ / tube) and then incubated for 3 hours at 37ºC and 5% CO2. In the last 30 minutes, 1000x diluted LIVE/DEAD™ (Invitrogen, Eugene, OR, USA) was added. After that, PMN were made to react with SYTOX™ Green Nucleic Acid Stain (1 µM) (Invitrogen) for 10 minutes at room temperature. Data on cells were acquired by flow cytometry using a BD FACSCanto II instrument (BD Bioscience, Franklin Lakes, NJ, USA). One hundred thousand events per sample were collected, doublet discrimination was performed using Forward Scatter Area (FSC-A) versus Forward Scatter Height (FSC-H) parameters, and the PMN were gated according to size (FSC-A) and granularity (Side Scatter Area, SSC-A). LIVE/DEAD™ and SYTOX™ positive cells were analyzed with FlowJo software (TreeStar, Ashland, OR, USA). For fluorescence microscopy, 100 µL ($5x10^5$ / well) of PMN were seeded on 13 mm glass coverslips in 24-wells plate and pre-incubated for 30 minutes at 37ºC and 5% CO^2. After adherence, PMN were stimulated with PMA (10 nM) or fungi samples ($5x10^5$ / tube) and then incubated for 3 hours at 37ºC and 5% CO2. PMN were made to react with SYTOX™ Green (1 µM) (Life Technologies) for 10 minutes at room temperature. The glass coverslips were removed and the slides were fixed with ProLong Gold Antifade Mountant with DAPI (Invitrogen). The images were obtained using a Leica DMI6000 Fluorescence Microscope (Leica Microsystems, Wetzlar, Germany). Details on the flow cytometry gating strategy used and microscopy images are available in the figshare repository (doi: 10.6084/m9.figshare.8114114).

Growth in the presence of different stresses

To study variation in infection-relevant phenotypes between the hybrid isolates, we compared the radial growth of *A. nidulans*, *A. spinulosporus*, *A. latus*, and of all the clinical isolates in different temperatures (30°C, 37°C and 44°C), in the presence of increasing concentrations of oxidative stress-inducing compounds (paraquat and menadione), and on iron starvation. Although the importance of oxidative stress susceptibility is contentious for *Aspergillus* virulence (Lessing et al., 2007; Lambou et al., 2010), we chose to examine these phenotypes because it is well established that hosts produce reactive oxygen species in response to infection (Warris and Ballou, 2019). To test for the effects of iron starvation on fungal growth, MM was prepared without any iron source and supplemented or not with 200 μM of the iron chelators Bathophenanthrolinedisulfonic acid (4,7-diphenyl-1,10-phenanthrolinedisulfonic acid [BPS]) (Sigma) and 300 μM of 3-(2-pyridyl)-5,6-bis(4-phenylsulfonic acid)-1,2,4-triazine (ferrozine) (Sigma). For radial growth, isolates were grown in triplicate from 10^5 spores and incubated at 37°C (except for the temperature test) for 5 days. Growth results in the presence of oxidative stress were expressed as ratio, dividing colony radial diameter (cm) of growth in the stress condition by colony radial diameter in the control (no stress condition).

Hydrogen peroxide tolerance

To test the viability of *Aspergillus* hyphae after exposure to hydrogen peroxide, we performed the XTT (2,3-bis(2-methoxy-4-nitro-5-sulfophenyl)2H-tetrazolium- 5-carboxanilide sodium salt]) (Sigma) assay as described by Henriet et al. (2011), but with modifications. We obtained hyphae from each strain by incubating $1x10^5$ asexual spores/well in 96-well plates containing

MM. After 16 hours at 37°C, the medium was removed and the wells washed twice with PBS. Subsequently, 100 microliters of minimal media supplemented or not (control) with different concentrations of hydrogen peroxide (1, 3 and 5 millimolar) were added to each well. Hyphal viability was assayed after 90 minutes of incubation (37°C) to avoid overgrowth of hyphae. One hundred microliters of PBS and 100 microliters of XTT-menadione solution was added to each well, obtaining a final concentration of 200 micrograms / milliliter XTT and 4.3 micrograms / milliliter menadione and the plates were incubated for 2 hours in the dark. After centrifugation (3,000 x g for 10 min), the supernatants were transferred to another plate and read at 450 nm in a spectrophotometer. Fungal damage was defined as the percent reduction in metabolic activity compared to that of controls without hydrogen peroxide (viability = 100%).

Antifungal susceptibility assays

Antifungal susceptibility testing for voriconazole (Sigma-Aldrich), posaconazole (Sigma-Aldrich), itraconazole (Sigma-Aldrich) and amphotericin B was performed by determining the minimal inhibitory concentration (MIC) according to the protocol established by the Clinical and Laboratory Standards Institute (CLSI, 2017). For caspofungin (Sigma) susceptibility, the radial growth in MM supplemented with different concentration of the drug was carried out similarly as described before (see Growth in the presence of different stresses). The results were presented as a ratio: growth in caspofungin(cm)/growth in the control/without caspofungin (cm).

Lastly, to determine the extent of phenotypic differences across all strains tested and all traits measured, we conducted principal component analysis. To do so, we first scaled (i.e., standardized) the data to account for variables that are measured in different scales (e.g., MIC,

radial growth, virulence). We then conducted principal component analysis and examined each variable's contribution to the variance along principal components using the R, version 3.5.2, packages FACTOEXTRA, version 1.0.5 (Kassambara and Mundt, 2017), and FACTOMINER, version 1.41 (Lê et al., 2008).

Data availability

All data is publicly available through NCBI or a figshare repository. Genome assemblies are available through BioProject IDs PRJNA542678 and PRJNA542141; raw reads are available through BioProject IDs PRJNA542395, PRJNA542181, and PRJNA542141. Strain specific BioProject or BioSample IDs can be found in File S4 from Steenwyk et al., 2020c. The figshare repository (doi: 10.6084/m9.figshare.8114114) contains phenotypic measurements, parent of origin analysis, homeologs per hybrid isolate, fluorescence-assisted cell sorting files, phylogenetic and phylogenomic data matrices and tree files, and biosynthetic gene cluster prediction results.

Results

Six clinical isolates previously characterized as *A. nidulans* are diploid

To gain insights into the genetic diversity of clinical isolates of *A. nidulans*, we analyzed 7 isolates from patients with different pulmonary diseases and compared them to haploid (A4) and the laboratory-induced diploid (R21/R153) reference strains of *A. nidulans* (Table 1 from Steenwyk et al., 2020c). Using microscopy-based and/or molecular biology methods, all 7 isolates had previously been identified as *A. nidulans*, all are similar in appearance when grown in standard laboratory conditions (Fig. S1 from Steenwyk et al., 2020c), and two were analyzed

as *A. nidulans* isolates in previous experimental studies (Lee et al., 2015). Examination of DNA content revealed that 6 / 7 isolates were more similar to the diploid *A. nidulans* R21/ R153 strain than to the haploid *A. nidulans* A4 strain, suggesting that these 6 isolates are diploid (Fig. 39A). The volume of asexual spores (conidia) is frequently proportional to the DNA content of the nucleus (Heagy and Roper, 1952) and examination of their size showed that the same 6 isolates and the diploid *A. nidulans* strain have significantly larger spores than isolates with haploid genomes ($p < 0.001$, respectively; Dunn's test with Benjamini-Hochberg method of multi-test correction for both tests) (Fig. 39B).

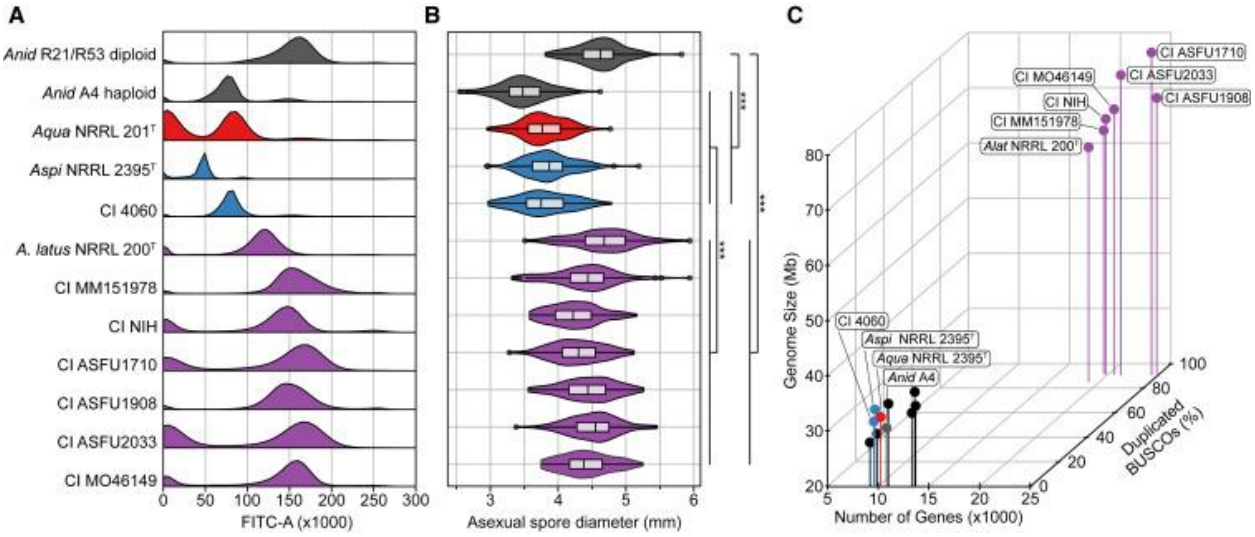

Figure 39. Six Clinical Isolates Previously Characterized as *Aspergillus nidulans* and the Type Strain of *Aspergillus latus* Are Diploids

(A) Fluorescence-assisted cell sorting analysis suggests that the type strain of *Aspergillus latus* NRRL 200^T and 6 clinical isolates (MM151978, NIH, ASFU1710, ASFU1908, ASFU2033, and MO46149) previously identified as *Aspergillus nidulans* have diploid genomes. In contrast, *Aspergillus spinulosporus* NRRL2395 and clinical isolate 4060 have haploid genomes. The haploid *A. nidulans* strain A4 and the laboratory-induced diploid *A. nidulans* strain R21/R23 were used as references of haploid and diploid genomes, respectively.

(B) Asexual spore diameter is significantly different between the 6 diploid clinical isolates, the haploid *A. quadrilineatus*, *A. spinulosporus*, and *A. nidulans*, and the laboratory-induced diploid *A. nidulans* ($\chi^2 = 399.54$; df = 2; $p < 0.001$; Kruskal-Wallis rank sum test). Additional pairwise comparisons are shown by brackets; all comparisons used Dunn's test with Benjamini-Hochberg method of multi-test correction. $^{***}p \leq 0.001$. Boxplot hinges correspond to the first

and third quartiles. Boxplot whiskers extend to values no greater than or less than 1.5 times the interquartile range. Data beyond this range are plotted individually.
(C) The 6 diploid clinical isolates and the *A. latus* NRRL 200^T strain have substantially larger genome sizes, gene numbers, and percent duplicated BUSCO genes compared to haploid genomes of representative *Aspergillus* species (*A. clavatus* NRRL 1, *A. flavus* NRRL 3357, *A. fumigatus* Af293, *A. nidulans* A4, *A. niger* CBS 513.88, *A. sydowii* CBS 593.65, and *A. versicolor* CBS 583.65). Genus and species names are abbreviated using the following scheme: *A. latus* (*Alat*); *A. spinulosporus* (*Aspi*); *A. quadrilineatus* (*Aqua*); and *A. nidulans* (*Anid*). CI represents clinical isolates. Dark gray represents *A. nidulans*, red represents *A. quadrilineatus*, blue represents *A. spinulosporus* and CI 4060, and purple represents *A. latus* and diploid isolates.

To gain further insight into the genomes of the 6 diploid isolates and 1 haploid isolate, we sequenced them and compared their genome size and gene number with those of representative *Aspergillus* species known to be haploid (*A. clavatus* NRRL 1, *A. flavus* NRRL 3357, *A. fumigatus* Af293, *A. nidulans* A4, *A. niger* CBS 513.88, *A. sydowii* CBS 593.65, and *A. versicolor* CBS 583.65) (Galagan et al., 2005; Nierman et al., 2005, 2015; Pel et al., 2007; Fedorova et al., 2008; de Vries et al., 2017). We found that the genomes and gene numbers of the diploids were significantly larger (average genome size = 69.09 ± 5.68 Mb, average gene number = 21,321.57 ± 2,342.13) than those of the haploid representative *Aspergillus* species (average genome size = 32.62 ± 3.05 Mb, average gene number = 11,330.75 ± 1,838.70) (Fig. 39C; Fig. S2; $p < 0.001$; Wilcoxon rank sum test for both tests; File S1). Similarly, examination of gene content completeness revealed a significantly higher number of duplicated near-universally single copy (BUSCO) genes in the diploids relative to representative *Aspergillus* species (Fig. 39C; $p = 0.001$; Wilcoxon rank sum test). Thus, we concluded that 6 / 7 clinical isolates are diploids.

Diploid clinical isolates are Aspergillus latus, a species of hybrid origin

To examine the evolutionary origin of the clinical isolates, we retrieved their calmodulin and β-tubulin sequences and performed molecular phylogenetic analysis in the context of sequences of the two genes from all available taxa in the section *Nidulantes* phylogeny (Chen et al., 2016). We found that the haploid clinical isolate 4060 had nearly identical calmodulin and β-tubulin sequences to other strains of *A. spinulosporus* and formed a monophyletic group with them, suggesting that it belongs to *A. spinulosporus* (Fig. 40A; figshare: 10.6084/m9.figshare.8114114). Notably, we found that all 6 diploid clinical isolates contained two different copies of the calmodulin and β-tubulin genes; one copy was nearly identical to *A. spinulosporus* sequences whereas the other was nearly identical to *A. latus* ones (Fig. 40A; figshare: 10.6084/m9.figshare.8114114), raising the hypothesis that the diploids originated from interspecific hybridization between *A. spinulosporus* and *A. latus*.

To test this hypothesis, we analyzed the genome of *A. spinulosporus* strain NRRL 2395^{T} (Steenwyk et al., 2019c) and sequenced the type strain NRRL 200^{T} of *A. latus*. Examination of the DNA content and asexual spore size of these two species' genomes showed that *A. spinulosporus* NRRL 2395^{T} had similar values as clinical isolate 4060 (Fig. 39) and was also placed in the same phylogenetic clade (Fig. 40A); these findings confirm that clinical isolate 4060 belongs to *A. spinulosporus* and that *A. spinulosporus* is one of the parental species involved in the interspecific hybridization event that gave rise to the 6 clinical isolates.

In contrast, the DNA content and spore size of the genome of the type strain of *A. latus* NRRL 200^{T} were similar to those of the 6 diploid clinical isolates (Fig. 39). Furthermore, like the 6

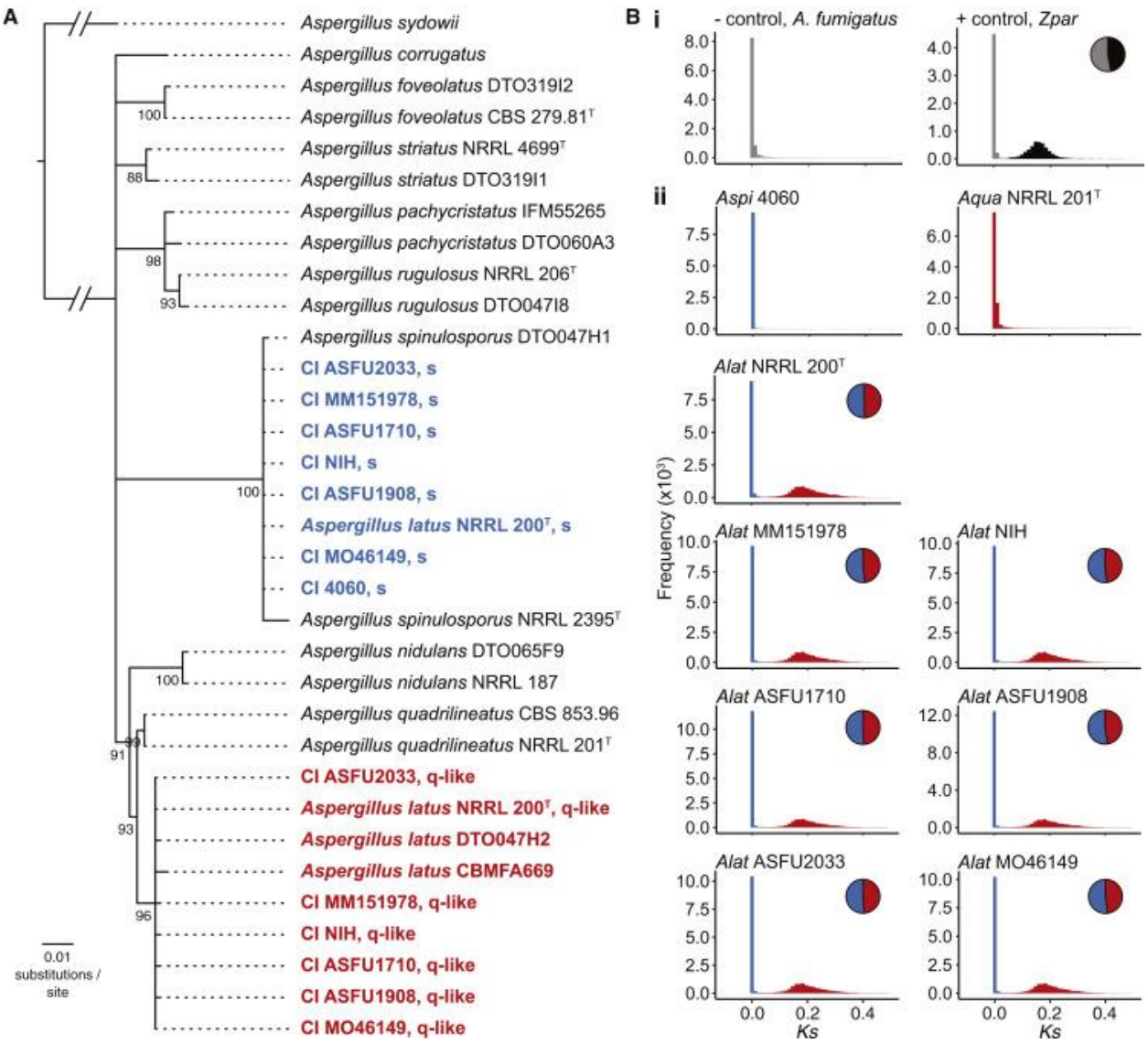

Figure 40. The 6 Clinical Diploids Belong to *A. latus*, an Allodiploid Species Formed via Hybridization of *A. spinulosporus* and a Close Relative of *A. quadrilineatus*

(A) The type strain of *A. latus* NRRL 200^{T} and the 6 diploid clinical isolates have each two copies of the taxonomic markers β-tubulin and calmodulin. Phylogenetic analysis of their β-tubulin and calmodulin sequences together with sequences from representative taxa in section *Nidulantes* suggests that clinical isolate 4060 belongs to *A. spinulosporus*, whereas *A. latus* NRRL 200^{T} and the 6 diploid clinical isolates are derived from two parental genomes. Interestingly, neither of the parental genomes is *A. nidulans*; rather, one is *A. spinulosporus* and the other is a species closely related to *Aspergillus quadrilineatus*. Newly sequenced isolates are shown in red and blue. (B) Examination of sequence divergence (*Ks*; x axis) between each gene in an allodiploid and its best blast hit in *A. spinulosporus* confirms that the 6 diploid clinical isolates and the type strain of *A. latus* are allodiploid hybrids. In contrast, the 7th clinical isolate (4060) is a haploid *A. spinulosporus*. Similarly, we found no evidence of *A. quadrilineatus* NRRL 201^{T} forming via allodiploid hybridization. (Bi) Examination of the

haploid, non-hybrid genome of *A. fumigatus* strain Af293 (negative control) shows a unimodal distribution, whereas examination of the diploid, hybrid genome *Zygosaccharomyces parabailii* strain NBRC1047/ATCC56075 (Zpar) shows a bimodal distribution (positive control; gray represents genes from one parent; black represents genes from the other parent). Red represents genes assigned to the *A.-quadrilineatus*-like parental genome; blue represents genes assigned to the *A. spinulosporus* parental genome.

---

clinical isolates, *A. latus* NRRL 200^T also had two copies of the calmodulin and β-tubulin gene sequences; one copy was nearly identical to *A. spinulosporus* sequences and the other copy was closely related to, but distinct from, *A. quadrilineatus* sequences (Fig. 40A; figshare: 10.6084/m9.figshare.8114114). These results suggest that the 6 diploid clinical isolates belong to *A. latus*, and that *A. latus* is an allodiploid hybrid species that originated via interspecific hybridization between *A. spinulosporus* and a species closely related to *A. quadrilineatus*.

We tested this hypothesis by performing two different sets of analyses. In the first set of analyses, we sequenced, assembled, and annotated the genome of the type strain (NRRL 201^T) of *A. quadrilineatus*. Consistent with our hypothesis that *A. latus* is an allodiploid hybrid, we found that the *A. quadrilineatus* genome contains a single copy of the calmodulin and β-tubulin gene sequences, that these sequences form a monophyletic group with their orthologous sequences retrieved from the genome of a different *A. quadrilineatus* strain (strain CBS 853.96; https://www.ncbi.nlm.nih.gov/sra/SRX5010607), and that the *A. quadrilineatus* sequences form a sister group with one of the two sets of the *A. latus* sequences (Fig. 40A).

In the second set of analyses, we estimated the sequence divergence of each gene in the genomes of the 7 clinical isolates as well as of *A. latus* NRRL 200^T and *A. quadrilineatus* NRRL 201^T from *A. spinulosporus* NRRL 2395^T. Under this analysis, the genomes of non-hybrids are

expected to show a unimodal distribution (e.g., see control non-hybrid, *A. fumigatus*; Fig. 40Bi left), whereas the genomes of hybrids are expected to show a bimodal distribution whose two modes correspond to the distributions of gene sequence divergence values from each parental genome (e.g., see control hybrid, *Zygosaccharomyces parabailii*; Fig. 40Bi right). We found that the haploid *A. spinulosporus* 4060 clinical isolate and *A. quadrilineatus* NRRL 201^T had unimodal distributions reflecting a history devoid of hybridization while the 6 diploid clinical isolates and *A. latus* NRRL 200^T had bimodal distributions consistent with allodiploidy (Fig. 40Bii). Furthermore, all 6 diploid isolates and *A. latus* NRRL 200^T contained nearly equal percentages of *A. spinulosporus* and *A. quadrilineatus*-like genes (51.43 ± 0.74% *A. spinulosporus* : 48.57 ± 0.74% *A. quadrilineatus*-like; Fig. 40B, pie charts), including nearly the full sets of *A. spinulosporus* and *A. quadrilineatus*-like secondary metabolic gene clusters (File S2 from Steenwyk et al., 2020c; figshare: 10.6084/m9.figshare.8114114). Putative homeologs exhibited an average nucleotide sequence divergence of 7.15 ± 0.03%, a value very similar to the average divergence of 7.14% observed between the 8,523 orthologs of *A. spinulosporus* NRRL 2395^T and *A. quadrilineatus* NRRL 201^T (Fig. S3 from Steenwyk et al., 2020c). These two sets of analyses confirm that the 6 diploid clinical isolates belong to *A. latus*, and that *A. latus* is an allodiploid hybrid species that originated via interspecific hybridization between *A. spinulosporus* and a species closely related to *A. quadrilineatus*.

We next assessed whether the allodiploid hybrid species *A. latus* stems from a single hybridization event by comparing the genome-scale phylogenies constructed from the *A. spinulosporus* and the *A. quadrilineatus*-like parental genomes of the *A. latus* isolates (Fig. S4 from Steenwyk et al., 2020c). We found that the relationships of the *A. latus* isolates differed

between the two phylogenies (Fig. S4 from Steenwyk et al., 2020c). This incongruence may stem from biological reasons (e.g., multiple hybridization events or recombination between the two parental genomes). However, the low level of support for relationships among isolates, especially in the phylogeny from the *A. spinulosporus* parental genome (Fig. S4A from Steenwyk et al., 2020c), means that we cannot exclude the possibility that the two phylogenies are not statistically significantly different. To test this, we evaluated whether the two topologies were statistically different using the approximately unbiased topology constraint test (Shimodaira, 2002). Using the *A. spinulosporus* data matrix, we found that we could not reject the topology inferred based on the *A. quadrilineatus*-like data matrix as statistically inferior; similarly, we could not reject the *A. spinulosporus* topology when we using the *A. quadrilineatus*-like data matrix (p-value = 0.50 for both tests). These results are consistent with the hypothesis that the two parental genomes of *A. latus* share the same evolutionary history.

To provide more insight on whether the two parental genomes *A. latus* hybrids undergo recombination, we first examined whether *A. latus* hybrids undergo the sexual cycle to produce sexual spores (ascospores). We found that all *A. latus* hybrids produce sexual spores and that the viability of these spores is similar to that of the sexual spores of their parental species (Fig. S5 from Steenwyk et al., 2020c). We next examined if any contigs in the genomes of *A. latus* isolates had evidence of recombination events. Examination of long (≥ 100 kb) contigs revealed that most genes in most contigs contained genes from one or the other parent and that very few contigs contained substantial percentages of genes from both parents (Fig. S6 from Steenwyk et al., 2020c). For example, only an average of 2.67 ± 0.71% contigs per *A. latus* hybrid genome contained substantial percentages of genes from both parental species (Fig. S6 from Steenwyk et

al., 2020c). However, interpretation of these data is challenging for two reasons. First, the high sequence similarity of the two parental genomes means that identification of parent of origin for highly-conserved genes is difficult and likely explains the sporadic presence of one or a handful of genes from one parent in contigs comprised mostly of genes from the other parent. Second, alignment of several of the contigs that contain large numbers of genes from both parents to the *A. nidulans* A4 reference genome suggests that they are often patchworks of *A. nidulans* contigs; for example, a long stretch of an *A. latus* contig that matches one parent is homologous to *A. nidulans* chromosome 5 and the rest of the contig, which matches the other parent, is homologous to *A. nidulans* chromosome 7. The absence of *A. latus* contigs that contain genes from both parental species and map to a single *A. nidulans* chromosome suggests that *A. latus* contigs that contain genes from both parental species may stem from assembly artifacts. These results suggest that *A. latus* hybrids likely undergo little to no recombination between the two parental genomes.

The genomes of the A. latus allodiploid hybrid isolates are stable

To assess the genome stability of the *A. latus* isolates, we begun by examining the gene content completeness of each parental genome. We found that each parental genome contained nearly all of the 1,315 near-universally single-copy orthologous (BUSCO) genes from the fungal phylum Ascomycota (93.50 ± 1.88% *A. spinulosporus* and 94.30 ± 0.40% *A. quadrilineatus*-like) (Fig. S7 from Steenwyk et al., 2020c). Considering that gene content completeness from each parent is only slightly below that from haploid representative species (average = 96.33 ± 0.78%; min = 95.70%, *A. spinulosporus*; max = 97.3%, *A. nidulans* A4), these results suggest little loss of each parental genome by either aneuploidy or loss of heterozygosity events.

To further test this observation genome-wide, we examined the fraction of orthologous genes shared between the *A. spinulosporus* NRRL 2395^{T} strain and the parental genomes of *A. latus* isolates that stem from *A. spinulosporus*. We found that the *A. spinulosporus* parental genomes from *A. latus* hybrids shared a minimum of 9,227 / 9,611 orthologous genes with *A. spinulosporus* NRRL 2395^{T}; the sole exception was *A. latus* NRRL 200^{T}, which shared 8,749 orthologs (figshare: 10.6084/m9.figshare.8114114). Interestingly, the *A. spinulosporus* parental genome of *A. latus* NRRL 200^{T} shows by far the highest evolutionary rate in our phylogenomic analyses (Fig. S4 from Steenwyk et al., 2020c), suggesting that the *A. spinulosporus* parental genome of this strain may be more genetically unstable than those of the clinical isolates.

Examination of loss of heterozygosity and aneuploidy events in *A. latus* genomes revealed relatively little evidence for either. Two isolates contained loss of heterozygosity regions. The *A. latus* NRRL 200^{T} strain contained a ~1.2 Mb region homologous to the end of *A. nidulans* chromosome VIII that contained two copies of the *A. quadrilineatus*-like parental genome and lacked a copy of the *A. spinulosporus* genome. This region contains several BUSCO genes, which explains why this strain has a higher proportion of missing BUSCO genes from the *A. spinulosporus* parental genome compared to the 6 clinical isolates (Fig. S7 from Steenwyk et al., 2020c). The clinical isolate MO46149 contained a ~1 Mb region homologous to the beginning of *A. nidulans* chromosome V with two copies of the *A. spinulosporus* genome and lacked a copy of the *A. quadrilineatus*-like genome (File S3 from Steenwyk et al., 2020c). We did not find evidence for chromosome-scale aneuploidies (File S3 from Steenwyk et al., 2020c).

Lastly, by comparing the gene lengths of homeolog pairs as a signature of pseudogenization, we found evidence of pseudogenization in at least one gene among an average of 11.67 ± 0.004% homeologs (Fig. S3 from Steenwyk et al., 2020c). These results suggest that the genomes of the *A. latus* allodiploid hybrids are generally stable, that loss of heterozygosity is rare, that major aneuploidies have not occurred, and that both genes in ~88% of homeolog pairs are intact.

Hybrids exhibit wide variation for infection-relevant traits

To examine variation in infection-relevant traits between the hybrid isolates, one of their known parental species (*A. spinulosporus*), the closest known relative of their other parental species (*A. quadrilineatus*), and the species they were originally identified as (*A. nidulans*), we tested the virulence of all isolates in an invertebrate disease model and phenotypically characterized them across a wide variety of infection-relevant conditions, including interactions with host immune cells, drug susceptibility, oxidative stress, iron starvation, and temperature stress (Fig. 41 and S8 from Steenwyk et al., 2020c). Principal component analysis (PCA) and examination of the traits with the greatest contributions to the observed variance among isolates revealed two major findings. First, the 7 *A. latus* hybrids exhibit substantial heterogeneity in their phenotypic profiles (Fig. 41A). Second, the *A. latus* hybrids are phenotypically distinct from *A. nidulans* and their parental species but are more similar to *A. spinulosporus* than to *A. quadrilineatus* (Fig. 41A). Among the traits tested, those with the largest contributions to the observed variation among isolates and species were interactions with host immune cells, antifungal drug susceptibility, and oxidative stress resistance (Fig. 41A and S9 from Steenwyk et al., 2020c). Here, we discuss exemplary phenotypic traits that highlight these two major findings (see Fig. S8 from Steenwyk et al., 2020c for other phenotypes).

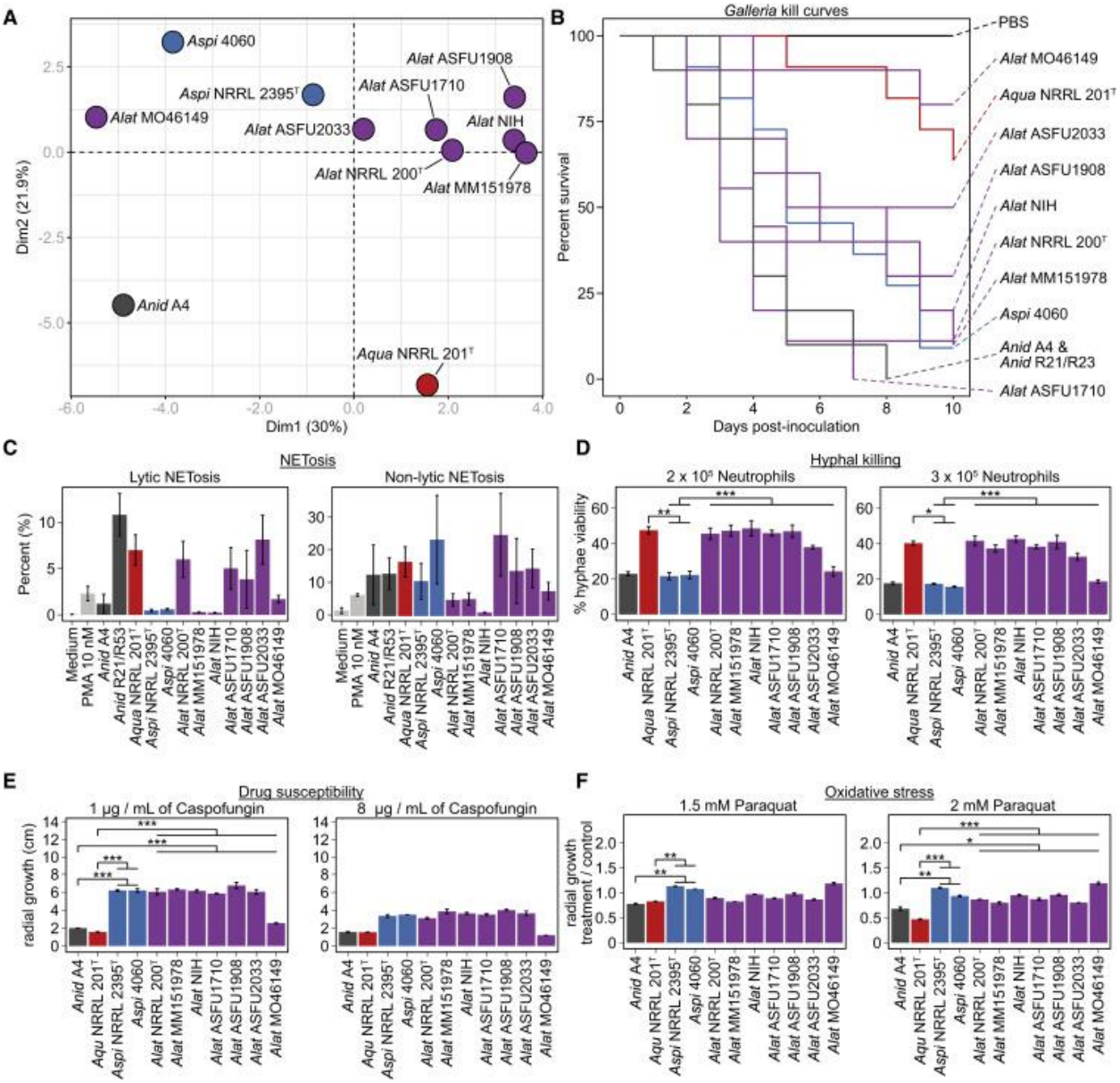

Figure 41. *A. latus* Hybrids Exhibit Strain Heterogeneity and Differ from Parental Species, *A. quadrilineatus*, and *A. nidulans* in Infection-Relevant Phenotypes
(A) Principal-component analysis of diverse infection-relevant phenotypes reveals strain heterogeneity among *A. latus* hybrids and that they differ from the closest relative of the unknown parent, *A. quadrilineatus*, and the *A. nidulans* A4 strain. Data for each phenotype were scaled prior to principal-component analysis. (B and C) Wide phenotypic variation among strains of *A. latus* hybrids as well as among the various species tested was observed for virulence in the *Galleria* moth model of disease (B) and for stimulation of NETosis, a process where neutrophils release neutrophil extracellular traps, or NETs, to kill microbes (C). (D) Examination of the percentage of hyphal viability between the various species revealed significant differences ($F(3) = 24.514$; $p < 0.001$; multi-factor ANOVA). (E) Examination of the caspofungin drug susceptibility profiles among *A. nidulans*, *A. spinulosporus*, and the *A. latus* revealed differences

between the three species ($F(3) = 56.01$; $p < 0.001$; multi-factor ANOVA). At 1 μg/mL caspofungin treatment, *A. spinulosporus* and *A. latus* hybrids grew more than *A. nidulans* and *A. quadrilineatus* ($p < 0.001$ for both comparisons; Tukey honest significant differences test). We found no statistically significant differences between the various species at 8 μg/mL of caspofungin but observed a qualitative difference similar to growth in 1 μg/mL of caspofungin. (F) Examination of growth in the presence of the oxidative stress agent paraquat revealed differences among the various species ($F(3) = 30.25$; $p < 0.001$; multi-factor ANOVA). Dark gray represents *A. nidulans*; red represents *A. quadrilineatus*; blue represents *A. spinulosporus*; and purple represents *A. latus*. All pairwise comparisons shown by brackets were examined using a Tukey honest significant differences test. $^{*}0.01 \leq p \leq 0.05$; $^{**}0.001 \leq p \leq 0.01$; $^{***}p \leq 0.001$. Barplots are displayed with error bars that correspond to one standard deviation from the mean.

Phenotypic variation or strain heterogeneity among *A. latus* hybrids was observed for nearly every trait measured (Fig. 41 and S8 from Steenwyk et al., 2020c). For example, examination of virulence in the invertebrate greater wax moth (*Galleria mellonella*) model revealed substantial variation among isolates ($p < 0.001$; log-rank test; Fig. 41B). Specifically, we observed that *A. latus* isolate ASFU1710 was the most virulent and *A. latus* isolate MO46149 was the least virulent. Similarly, we found substantial strain heterogeneity in how much lytic and non-lytic NETosis (a process where neutrophils release neutrophil extracellular traps, or NETs, to kill microbes; (Branzk et al., 2014)) was stimulated by *A. latus* hybrid isolates (Fig. 41C). For example, *A. latus* NIH did not substantially stimulate NETosis while *A. latus* ASFU1710 did. Strain heterogeneity was less pronounced for other traits, such as hyphal viability, drug susceptibility, and oxidative stress, yet all exhibited variation across isolates (Fig. 41D, E, and F). One *A. latus* isolate that was consistently different from the rest is MO46149; for example, this isolate was twice as susceptible to hyphal killing by neutrophils compared to the other *A. latus* isolates, it was the isolate most sensitive to the antifungal caspofungin, as well as the isolate most tolerant to the oxidative stressor paraquat.

Phenotypic variation was also pronounced when we compared infection-relevant traits between *A. latus*, its two parental species, and *A. nidulans* (Fig. 41A). For example, we found that *A. latus* hybrids (and *A. quadrilineatus*) were less susceptible to killing by neutrophils compared to *A. spinulosporus* (Fig. 41D). In contrast, we found *A. latus* isolates (and *A. spinulosporus*) differed in their susceptibility to low doses of caspofungin (Fig. 41E) or to high doses of oxidative stress (Fig. 41F) from *A. quadrilineatus* and *A. nidulans*.

In summary, we found substantial heterogeneity among *A. latus* hybrid isolates as well as between *A. latus* and closely related or parental species for diverse infection-relevant traits. Generally, *A. latus* hybrids are more similar to their known parent, *A. spinulosporus*, compared to the closest known relative of their other parent, *A. quadrilineatus*. Importantly, *A. latus* hybrids are also phenotypically distinct from *A. nidulans*, the species they were originally misdiagnosed as.

Discussion

Infections by filamentous fungal pathogens affect hundreds of thousands of humans and exhibit very high mortality rates (Brown et al., 2012), so understanding the evolutionary mechanisms underlying their pathogenicity is of great interest. We have discovered several clinical isolates previously identified as *A. nidulans*, an important pathogen of CGD patients, which in reality are allodiploid hybrids that arose via interspecific hybridization between *A. spinulosporus* and a close relative of *A. quadrilineatus* and belong to *A. latus*. In line with clinical misidentification of these species, *A. nidulans*, *A. spinulosporus*, *A. quadrilineatus*, and *A. latus* are known to be

nearly indistinguishable with the exception of aspects of their ascospore micromorphology and their secondary metabolic profiles (Chen et al., 2016). The allodiploid hybrids show strain heterogeneity in their phenotypic profiles and differ from their parental species and *A. nidulans* with respect to several infection-relevant traits. Below, we discuss the implications of these results for disease management and the evolution of fungal pathogenicity.

Application of molecular typing, and more recently genomic, approaches to delineate fungal species and pathogens has revealed the existence of multiple, closely related species that are morphologically indistinguishable but genomically distinct from each other (Taylor et al., 2000). This "hidden" or "cryptic" genomic diversity is found in species from many genera that harbor major fungal pathogens, including *Aspergillus* (Geiser et al., 1998; Balajee et al., 2005; Pringle et al., 2005). Alarmingly, application of molecular methods on fungal clinical isolates has too begun to reveal that a significant portion of fungal infections are caused by these cryptic species. In the case of *Aspergillus*, studies in both the USA and Spain report that 10 to 15% of aspergillosis infections stem from cryptic species (Alastruey-Izquierdo et al., 2014; Perlin et al., 2017). Understanding the biology of these cryptic species is essential for guidance in therapy, as many show high levels of antifungal drug resistance (Alastruey-Izquierdo et al., 2014; Verweij et al., 2015; Perlin et al., 2017). Several *A. fumigatus*-related cryptic species exhibit decreased susceptibility (relative to *A. fumigatus*) to antifungal drugs (Van Der Linden et al., 2011); similarly, we found notable differences in certain phenotypic traits, including drug susceptibility, between the allodiploid hybrids and *A. nidulans* (Figs. 41E and S8 from Steenwyk et al., 2020c).

A growing body of literature suggests that phenotypic heterogeneity among isolates of the same species is an under-appreciated factor in understanding fungal pathogenicity (Keller, 2017). For example, several recent studies have identified phenotypic and genomic differences between *A. fumigatus* strains that are associated with virulence (Kowalski et al., 2016, 2019; Ries et al., 2019); strain heterogeneity is also observed among *A. nidulans* strains (Bastos et al., 2020a). In line with these studies, our work reveals considerable heterogeneity in infection-relevant traits among *A. latus* hybrids (Fig. 3 and S8), further highlighting the importance of strain heterogeneity in understanding *Aspergillus* pathogenicity (Keller, 2017).

Allopolyploid hybrids typically have unstable genomes (Mixão and Gabaldón, 2018). Evolutionary paths to achieve stability after hybridization include whole-genome duplication, total or partial chromosome loss, gene loss, and loss of heterozygosity (Mixão and Gabaldón, 2018). For example, an ancient allodiploid hybridization event in the budding yeast lineage that includes the baker's yeast *Saccharomyces cerevisiae* (Marcet-Houben and Gabaldón, 2015) was quickly followed by rapid gene loss in the parental genomes (Scannell et al., 2007), with estimates suggesting that 10% of genes were lost in the first 10 million years following hybridization (Scannell et al., 2007). Similar rates of gene loss have been reported in plants (Bowers et al., 2003; Paterson et al., 2004) and animals (Brunet et al., 2006), suggesting that rapid gene loss is a common outcome of hybridization. In contrast to these studies, we found that most *A. latus* hybrids (with the possible exception of *A. latus* NRRL 200^{T}) contain both copies of most homeolog gene pairs and have relatively stable genomes (Fig. S3 from Steenwyk et al., 2020c and S6 from Steenwyk et al., 2020c). Consistent with our genomic analyses, *A. latus* isolates exhibit minimal sectoring when grown in culture (Fig. S1 from Steenwyk et al., 2020c)

and have similar ascospore viability compared to their closest relatives, *A. spinulosporus* and *A. quadrilineatus* (Fig. S5 from Steenwyk et al., 2020c). Furthermore, laboratory studies that have created synthetic *Aspergillus* hybrids—including those that created interspecies hybrids from distinct species in section *Nidulantes*—reported that some of these hybrids are relatively stable (Kevei and Peberdy, 1979; Kevei and Perberdy, 1984; Olarte et al., 2015; Macdonald et al., 2018). Taken together, these results suggest that *Aspergillus* hybrids may be more stable than other hybrid allopolyploids.

Although several examples of interspecies hybridization are known in fungi (Wolfe and Shields, 1997; Nielsen and Yohalem, 2001; Inderbitzin et al., 2011; Marcet-Houben and Gabaldón, 2015; Pryszcz et al., 2015; Wolfe, 2015; Depotter et al., 2016; Schröder et al., 2016; Stukenbrock, 2016; Rhodes et al., 2017b; Mixão and Gabaldón, 2018), most of them are ancient. Consequently, the steps that led to hybrid formation and maintenance are harder to elucidate. As the *A. latus* allodiploid hybrids originated much more recently, and the mechanisms that underlie synthetic hybrid formation in *Aspergillus* have been extensively studied (Kevei and Peberdy, 1979; Kevei and Perberdy, 1984; Olarte et al., 2015; Macdonald et al., 2018), we can propose a model to explain the origin and lifecycle of *A. latus* (Fig. 42). Under our model, the first step in the formation of the *A. latus* allodiploid hybrid was the cellular fusion (or plasmogamy) of an *A. spinulosporus* parental isolate and an *A. quadrilineatus*-like parental isolate through a parasexual or a sexual cycle. In the next step, the distinct nuclei contributed by the two parental isolates underwent nuclear fusion (or karyogamy) to create a single nucleus with a diploid genome comprised from the *A. spinulosporus* and *A. quadrilineatus*-like genomes. Once formed, the allodiploid hybrid species *A. latus* has been capable of undergoing both asexual and sexual

reproduction and forms viable asexual spores (conidia; Fig. 39) and sexual spores (ascospores; Fig. S5).

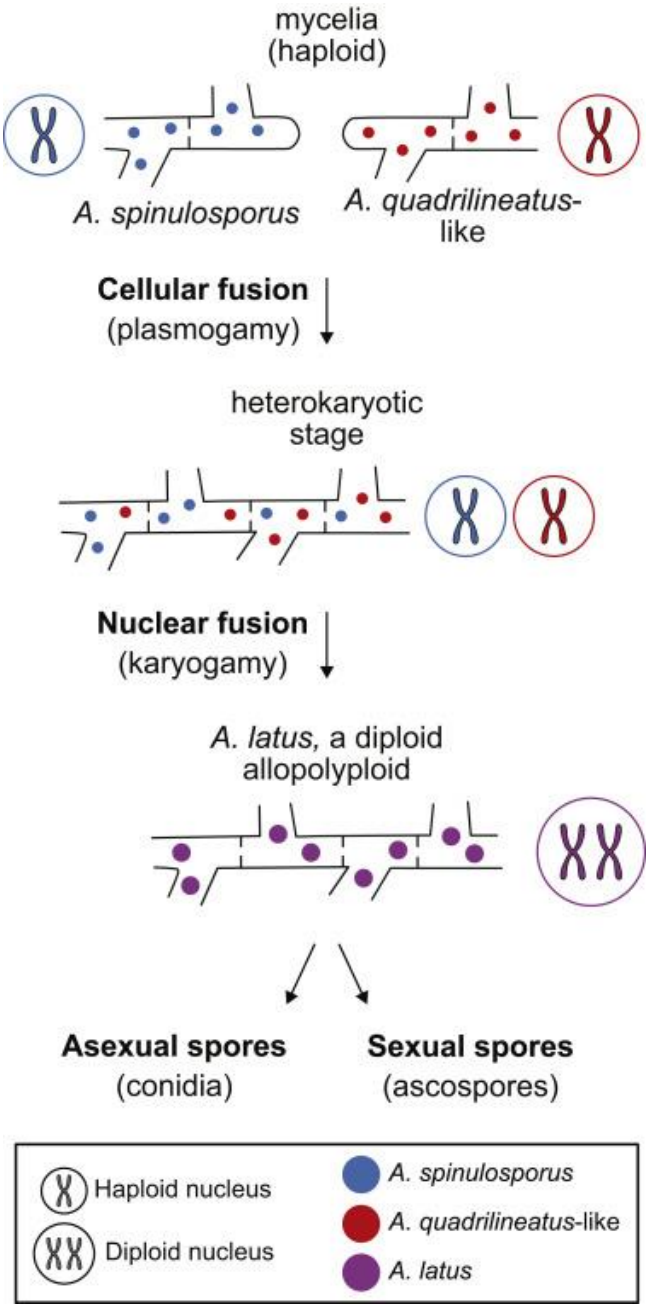

Figure 42. Proposed Model for the Evolution of *A. latus* via Allodiploid Hybridization
Under the model, haploid nuclei of an *A. spinulosporus* isolate and of an isolate from an *A. spinulosporus*-like species underwent cellular fusion (plasmogamy), forming a heterokaryotic mycelium (i.e., a mycelium where cells contain two distinct nuclei). Next, nuclear fusion (karyogamy) resulted in the merging of the two genetically distinct nuclei and their genomes into

a single one, giving rise to the allodiploid species *A. latus*, which is capable of undergoing asexual and sexual reproduction to produce asexual spores (conidia) or sexual spores (ascospores).

Hybrids have been observed in several fungal pathogens of animals and plants (Pryszcz et al., 2015; Depotter et al., 2016; Rhodes et al., 2017b; Mixão and Gabaldón, 2018), suggesting that hybridization of pathogenic fungi poses threats to plant and animal health. Hybridization can result in the acquisition of new traits, such as host expansion or increased virulence. For example, two powdery mildew species that specialize in infecting different species of plants have been shown to hybridize and infect a plant species that neither parent can (Menardo et al., 2016). Similarly, hybridization is thought to contribute to virulence among human yeast pathogens but clear examples are currently lacking (Mixão and Gabaldón, 2018). To our knowledge, our results are the first report of hybrid clinical isolates in a filamentous fungal pathogen of humans (Fig. 40). Like previous studies, we observe that the hybrids exhibit infection-relevant traits that differentiate them from their parental relatives and may provide a fitness advantage inside human hosts (Fig. 41). However, it should be noted that the potential selection pressure for hybrids in patients is unknown; although we isolated the hybrids from patients with diverse pulmonary diseases, we do not know if the hybrids' primary lifestyle is that of a pathogen or whether they originated inside a host. Importantly, the type strain of *A. latus* (NRRL 200^{T}) is not a clinical isolate, suggesting that the species is also found in the environment. More broadly, the existence of hybrids in a diverse set of pathogenic fungi infecting a diverse set of animal and plant hosts raise the hypothesis that allodiploid hybridization contributes to the evolution and diversity of all kinds of fungal pathogens, perhaps to a greater extent than currently realized.

In summary, viewed from a medical mycology perspective, our discovery of allodiploid hybrid clinical isolates reveals the importance of accurate isolate identification and strain heterogeneity. Viewed from an evolutionary perspective, our and previous results suggest that hybridization contributes to the genomic and phenotypic diversification of filamentous fungal pathogens of humans and argue that hybridization represents a general mechanism that can be potentially employed by all fungal pathogens to adapt to all kinds of hosts.

CHAPTER 10

BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data[9]

Introduction

Bioinformatics is the application of computational tools to process and analyze biological data, such as nucleotide or amino acid sequences in the form of genome assemblies, gene annotations, and multiple sequence alignments (Bayat, 2002). Diverse disciplines in the biological sciences rely on bioinformatic methods and software (Wren, 2016). Recently, researchers have acknowledged the need to consider diverse types of biological scientists with different levels of experience when developing software (Kumar and Dudley, 2007). It is also essential to implement high standards of software development that ensure software functionality and archival stability (Mangul et al., 2019b, 2019a). For example, code quality can be improved by utilizing unit and integration tests, which ensure faithful function of code (Darriba et al., 2018). As a result, the development of effective and user-friendly software for diverse biologists often requires an interdisciplinary team of software engineers, biologists, and others.

Even though numerous bioinformatic pieces of software are available, there are still several barriers to creating seamless and reproducible workflows (Kim et al., 2018). This issue in part stems from different pieces of software requiring different input file formats, being unable to account for non-standard biological phenomena such as the use of alternative genetic codes, or

[9]This work is published in: Steenwyk, J. L., Buida, T. J., Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., et al. (2021). BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. bioRxiv, 2021.10.02.462868. doi:10.1101/2021.10.02.462868.

can only be executed using web servers or graphical user interfaces, which cannot be incorporated into high-throughput pipelines. Another factor is that multiple pieces of software or custom scripts are typically needed to execute different steps in a larger bioinformatic pipeline; for example, bioinformatic workflows often rely on one software/script for converting file formats, another software/script for translating sequences using standard and non-standard genetic codes, another software/script to examine the properties of genomes or multiple sequence alignments, and so on. As a result, maintaining efficacious bioinformatic workflows is cumbersome (Kulkarni et al., 2018). Thus, the bioinformatic community would benefit from a multi-purpose toolkit that contains diverse processing and analysis functions.

To address this need, we—an interdisciplinary team of software engineers, evolutionary biologists, molecular biologists, microbiologists, and others—developed BioKIT, a versatile toolkit with 40 functions, several of which were community sourced, that conduct routine and novel processing and analysis of diverse sequence files including genome assemblies, multiple sequence alignments, protein coding sequences, and sequencing data (Table 1 from Steenwyk et al., 2020c). Functions implemented in BioKIT facilitate a wide variety of standard bioinformatic analyses, including genome assembly quality assessment (e.g., N50, L50, assembly size, guanine-cytosine (GC) content, number of scaffolds, and others), the calculation of multiple sequence alignment properties (i.e., number of taxa, alignment length, the number of constant sites, the number of parsimony-informative sites, and the number of variable sites), and processing and analysis of protein coding sequences (e.g., translation using 26 genetic codes including user-specified translation tables, GC content at the first, second, and third codon positions, and relative synonymous codon usage). To demonstrate the utility of BioKIT, we

examined the genome assembly quality of 901 eukaryotic genomes, evaluated the properties of 10 phylogenomic data matrices, calculated relative synonymous codon usage in 171 fungal genomes, and estimated codon optimization in each gene from two *Saccharomyces* budding yeast species using a novel metric, gene-wise relative synonymous codon usage (gw-RSCU). BioKIT comes complete with common and novel functions that will help improve reproducibility and accessibility of diverse bioinformatic analysis and facilitate discovery in the biological sciences.

Materials and Methods

BioKIT is an easy-to-install command-line software that conducts diverse bioinformatic analyses in the UNIX programming environment. BioKIT is written in the Python programming language and has few dependencies, namely Biopython (Cock et al., 2009a) and numPy (Van Der Walt et al., 2011).

BioKIT currently has 40 functions that process and analyze sequence files such as genome assemblies, multiple-sequence alignments, protein coding sequences, and sequencing data (Table 1 from Steenwyk et al., 2020c). Processing functions include those that convert various file formats, subset sequence reads from FASTQ files, rename entries in FASTA files, and others. Analysis functions include those that trim sequence reads in FASTQ files according to quality and length thresholds, calculate relative synonymous codon usage, estimate codon optimization, and others. Similar to other software we have developed (Steenwyk et al., 2020b, 2021b; Steenwyk and Rokas, 2021b), we plan on continuing to develop and incorporate additional functions into BioKIT to meet the needs of the research community.

Details about each function, their usage, tutorials, and other information such as how to request additional functions can be found in the online documentation (https://jlsteenwyk.com/BioKIT). To demonstrate the utility of BioKIT, we highlight four use-cases: (i) genome assembly quality assessment, (ii) summarizing properties of multiple sequence alignments, (iii) determination of relative synonymous codon usage using different genetic codes, and (iv) determination of a novel metric for estimation of gene-wise codon optimization, gene-wise relative synonymous codon usage (gw-RSCU).

Genome assembly quality assessment

Determination of genome assembly properties is essential when evaluating assembly quality (Gurevich et al., 2013; Hunt et al., 2013). To facilitate these analyses, the *genome_assembly_metrics* function in BioKIT calculates 14 diverse properties of genome assemblies that evaluate assembly quality and characteristics including:

• assembly size: sum length of all contigs/scaffolds;

• L50 (and L90): the number of contigs/scaffolds that make up 50% (or, in the case of L90, 90%) of the total length of the genome assembly;

• N50 (and N90): the length of the contig/scaffold which, along with all contigs/scaffolds longer than or equal to that contig/scaffold, contain 50% (or, in the case of N90, 90%) the length of a particular genome assembly;

• GC content: fraction of total bases that are either G or C;

• number of scaffolds: total number of contigs/scaffolds;

• number and sum length of large scaffolds: total number and sum length of contigs/scaffolds above 500 nucleotides in length (length threshold of a "large scaffold" can be modified by the user); and
• frequency of nucleotides: fraction of occurrences for adenine (A), thymine, (T), G, and C nucleotides.

Each metric can also be called using individual functions (e.g., the *n50* function calculates the N50 of an assembly and the *number_of_large_scaffolds* function calculates the number of large scaffolds in an assembly). We anticipate the ability of BioKIT to summarize genome assembly properties will be helpful for assessing genome quality as well as in comparative studies of genome properties, such as the evolution of genome size and GC content (Walker et al., 2015; Shen et al., 2020b). Other pieces of software that conduct similar analyses include QUAST, REAPR, and GenomeQC (Gurevich et al., 2013; Hunt et al., 2013; Manchanda et al., 2020).

Processing and assessing the properties of multiple sequence alignments

Multiple sequence alignments—the alignment of three or more biological sequences—contain a wealth of information. To facilitate easy use and manipulation of multiple sequence alignments, BioKIT implements 16 functions that process or analyze alignments including: generating consensus sequences; generating a position-specific score matrix (which represents the frequency of observing a particular amino acid or nucleotide at a specific position); recoding an alignment using different schemes, such as the RY-nucleotide scheme for nucleotide alignments (Woese et al., 1991; Phillips et al., 2001) or the Dayhoff-6, S&R-6, and KGB-6 schemes for amino acid alignments (Embley et al., 2003; Hrdy et al., 2004; Kosiol et al., 2004; Susko and Roger, 2007);

converting alignments among the following formats: FASTA, Clustal, MAF, Mauve, PHYLIP, PHYLIP-sequential, PHYLIP-relaxed, and Stockholm; extracting entries in FASTA files; removing entries from FASTA file; removing short sequences from a FASTA file; and others.

We highlight the *alignment_summary* function, which calculates numerous summary statistics for a multiple sequence alignment, a common step in many molecular evolutionary analyses (Plomion et al., 2018; Winterton et al., 2018). More specifically, the *alignment_summary* function calculates:

• alignment length: the total number of sites in an alignment;

• number of taxa: the total number of sequences in an alignment;

• number of parsimony-informative sites: a site in an alignment with at least two distinct nucleotides or amino acids that each occur at least twice;

• number of variable sites: a site in an alignment with at least two distinct nucleotides or amino acids;

• number of constant sites: sites with the same nucleotide or amino acid (excluding gaps); and

• the frequency of all character states: the fraction of occurrence for all nucleotides or amino acids (including gap characters represented as '-' or '?' in an alignment.

Like the *genome_assembly_metrics* function, each metric can be calculated individually (e.g., the *constant_sites* function calculates the number of constant sites in an alignment and the *character_frequency* function calculates the frequency of all character states). We anticipate the *alignment_summary* function will assist researchers in statistically evaluating the properties of their alignments. Other pieces of software that perform similar operations include AMAS (Borowiec, 2016) and Mesquite (Mesquite Project Team, 2014).

Examining features of coding sequences including relative synonymous codon usage

BioKIT contains multiple functions that process or analyze protein coding sequences including translating protein coding sequences into amino acids using one of 26 genetic codes or a user-specified translation table as well as determining the GC content at the first, second, and third codon positions.

Here, we highlight the *relative_synonymous_codon_usage* function, which calculates relative synonymous codon usage, the ratio of the observed frequency of synonymous codons to an expected frequency in which all synonymous codons are used equally (Xu et al., 2008). In this analysis, overrepresented codons have relative synonymous codon usage values greater than one whereas underrepresented codons have relative synonymous codon usage values less than one. Relative synonymous codon usage values of one fit the neutral expectation. The *relative_synonymous_codon_usage* function can be used with one of 26 genetic codes including user-specified translation tables. The ability of BioKIT to account for diverse genetic codes makes it uniquely suitable for analyses of lineages that contain multiple genetic codes (Krassowski et al., 2018; LaBella et al., 2019). Other software that conduct similar analyses include DAMBE and GCUA (McInerney, 1998; Xia, 2013).

We also highlight the *gene_wise_relative_synonymous_codon_usage* function, which calculates a novel metric, gw-RSCU, to examine biases in codon usage among individual genes encoded in a genome. More specifically, the gw-RSCU is calculated by determining the mean or median relative synonymous codon usage value for all codons in each gene based on their genome-wide

values. Thus, BioKIT calculates relative synonymous codon usage for each codon based on codon usage in an entire set of protein coding genes, individually reexamines each gene and the relative synonymous codon usage value for each codon therein, and then determines the mean or median relative synonymous codon usage value for the individual gene. The formula for the mean gw-RSCU calculation is as follows:

$$gw\text{-}RSCU^{a} = \frac{\sum_{i=1}^{j} RSCU_i}{n}$$

where gw-RSCU^{a} is the gene that gw-RSCU is being calculated for, RSCU$_i$ is the relative synonmyouse codon usage value (calculated from all protein coding genes in a genome) for the *i*th codon of *j* codons in a gene, and *n* is the number of codons in a gene. To evaluate within-gene variation in relative synonymous codon usage, BioKIT also reports the standard deviation of relative synonymous codon usage values for each gene. Like the *relative_synonymous_codon_usage* function, gw-RSCU can be calculated using alternative genetic codes including user-specified ones. Taken together, these functions can be used individually or in tandem to investigate diverse biological phenomena, including codon usage bias (Brandis and Hughes, 2016; LaBella et al., 2019).

Implementing high standards of software development

Archival instability is a concern for bioinformatic tools and threatens the reproducibility of bioinformatic research. For example, in an analysis that aimed to evaluate the "installability" of bioinformatic software, 28% of over 36,000 bioinformatic tools failed to properly install due to implementation errors (Mangul et al., 2019b). To ensure archival stability of BioKIT, we implemented a previously established protocol (Steenwyk et al., 2020b, 2021b; Steenwyk and Rokas, 2021b) for high standards of software development and design practices. More

specifically, we wrote 327 unit and integration tests that ensure faithful functionality of BioKIT and span 95.46% of the codebase. We also implemented a continuous integration pipeline, which builds, packages, installs, and tests the functionality of BioKIT across Python versions 3.6, 3.7, 3.8, and 3.9. To accommodate diverse installation workflows, we also made BioKIT freely available under the MIT license across popular platforms including GitHub (https://github.com/JLSteenwyk/BioKIT), PyPi (https://pypi.org/project/jlsteenwyk-biokit/), and the Anaconda Cloud (https://anaconda.org/jlsteenwyk/jlsteenwyk-biokit). To make BioKIT more user-friendly, we wrote online documentation, user tutorials, and instructions for requesting new features (https://jlsteenwyk.com/BioKIT). We anticipate our rigorous strategy to implement high standards of software development, coupled to our approach to facilitate easy software installation and extensive documentation, will address instabilities observed among many bioinformatic software and increase the long-term usability of BioKIT.

Results

Genome assembly quality and characteristics among 901 eukaryotic genomes

To demonstrate the utility of BioKIT for the examination of genome assembly quality and characteristics, 14 diverse genome assembly metrics were determined among 901 scaffold-level haploid assemblies of eukaryotic genomes, which were obtained from NCBI, and span three major classes of animals (Mammalia; N = 350), plants (Magnoliopsida; N = 336), and fungi (Eurotiomycetes; N = 215). Genome assembly properties exhibited variation both within and between the three classes (Figure 43). For example, fungi had the smallest average genome size of 32.71 ± 7.04 Megabases (Mbs) whereas mammals had the largest average genome size of 2,645.50 ± 487.48 Mbs. Extensive variation in genome size within each class corroborates

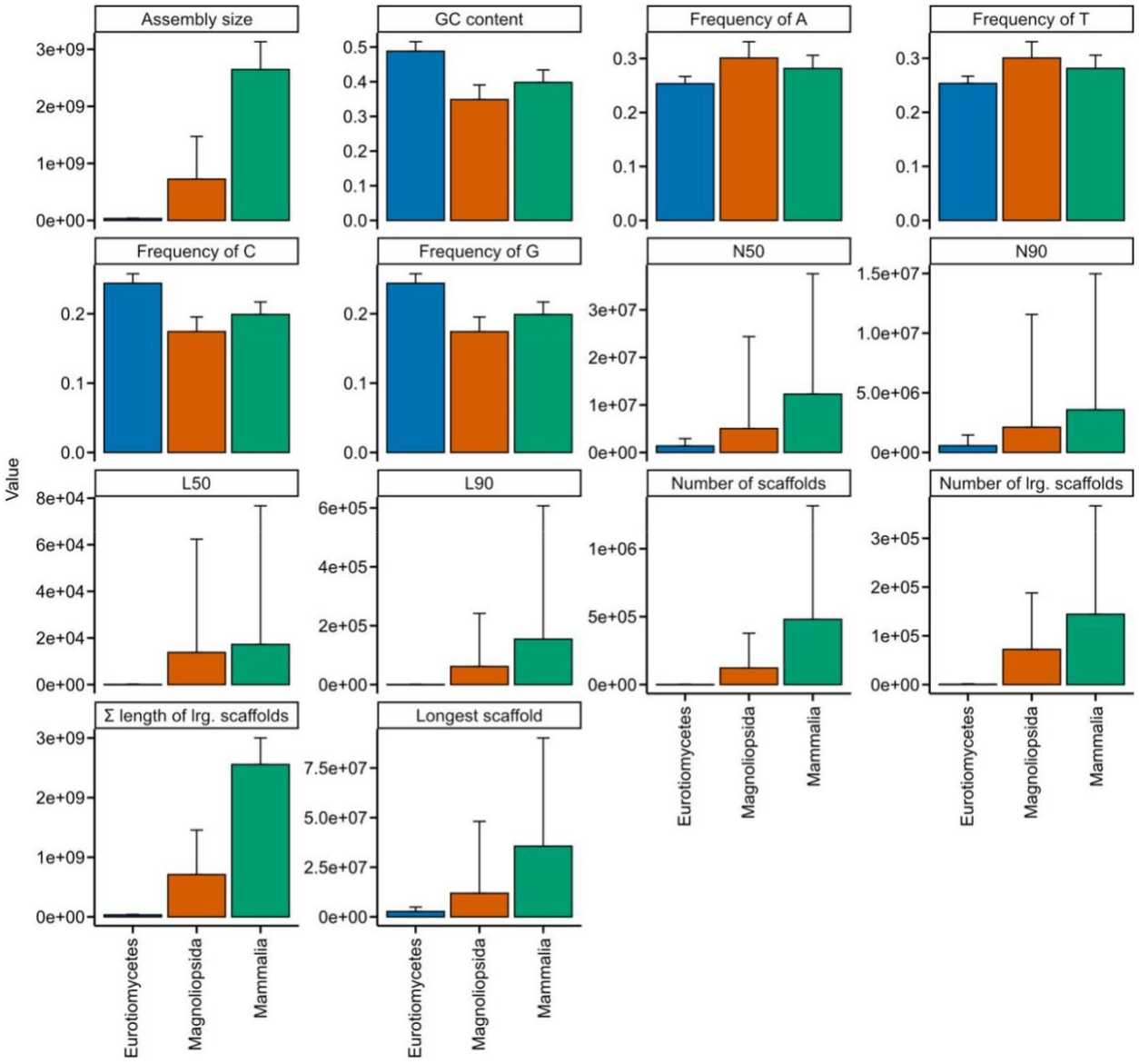

Fig. 43. Summary of genome assembly metrics across 901 genomes from three eukaryotic classes.

Nine hundred and one scaffold-level genome assemblies from three major eukaryotic classes (215 Eurotiomycetes (kingdom: Fungi), 336 Magnoliopsida (kingdom: Plantae), 350 Mammalia (kingdom: Animalia)) were obtained from NCBI and examined for diverse metrics including assembly size, GC content, frequency of A, T, C, and G, N50, N90, L50, L90, number of scaffolds, number of large scaffolds (defined as being greater than 500 nucleotides, which can be modified by the user), sum length of large scaffolds, and longest scaffold in the assembly. Bar plots represent the mean for each taxonomic class. Error bars represent the standard deviation of values.

previous findings of extreme genome size variation among eukaryotes (Elliott and Gregory, 2015). Variation in GC content, a genome property that has been actively investigated for decades (Galtier et al., 2001; Romiguier et al., 2010; Serres-Giardi et al., 2012), was observed among the three eukaryotic classes—animals, plants, and fungi had an average GC content of 0.40 ± 0.04, 0.35 ± 0.04, and 0.49 ± 0.03, respectively. Lastly, there was wide variation in genome assembly metrics associated with continuity of assembly. For example, the average N50 values for animals, plants, and fungi were 12,287.64 ± 25,317.31 Mbs, 5,030.15 ± 19,358.58 Mbs, and 1,370.77 ± 1,552.13 Mbs, respectively. Taken together, these results demonstrate BioKIT can assist researchers in summarizing diverse genome assembly properties, which may be helpful not only for evaluating genome assembly quality, but also for studying genome evolution.

Properties of multiple sequence alignment from 10 phylogenomic studies

To demonstrate the utility of BioKIT in calculating summary statistics for multiple sequence alignments, we calculated six properties across 10 previously published phylogenomic data matrices of amino acid sequences (Misof et al., 2014; Nagy et al., 2014; Borowiec et al., 2015; Chen et al., 2015; Struck et al., 2015; Whelan et al., 2015; Yang et al., 2015; Shen et al., 2016b, 2018; Steenwyk et al., 2019c) (Figure 44). Phylogenomic data matrices varied in the number of taxa (mean = 109.50 ± 87.26; median = 94; max = 343; min = 36). Alignment length is associated with greater phylogenetic accuracy and bipartition support (Shen et al., 2016a); however, recent analyses suggest that in some instances shorter alignments that contain a wealth of informative sites (such as parsimony-informative sites) harbor robust phylogenetic signal (Steenwyk et al., 2020b). Interestingly, the longest observed alignment (1,806,035 sites; *Chen*,

Vertebrates in Figure 2) (Chen et al., 2015) contained the highest number of constant sites (N = 610,994), which are phylogenetically uninformative, as well as the highest number of variable sites (N = 1,195,041), which are phylogenetically informative (Shen et al., 2016a). In contrast to

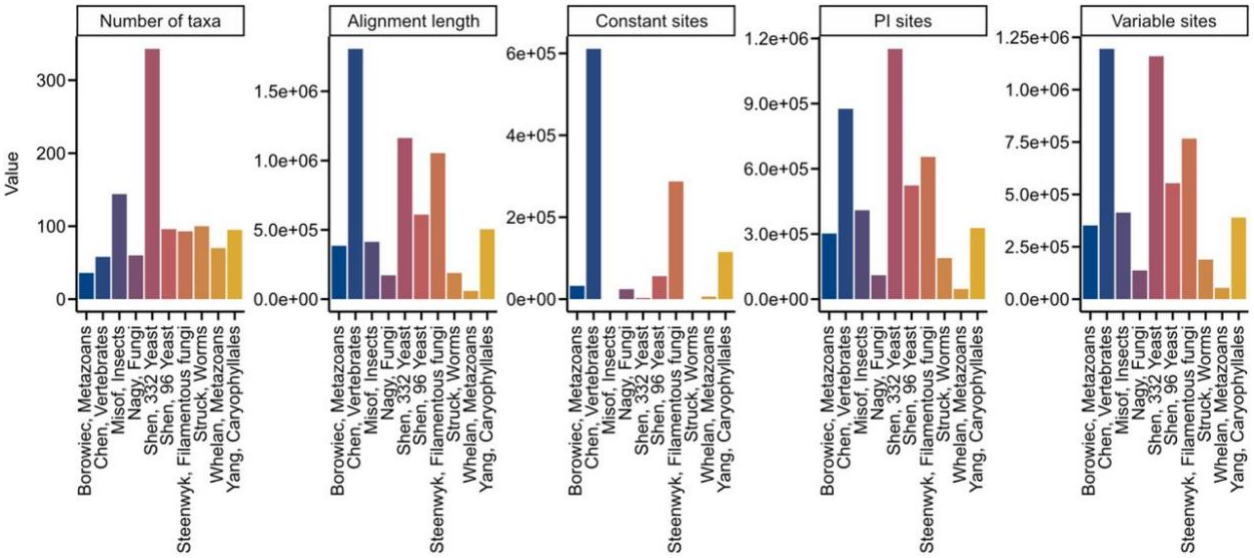

Fig. 44. Summary metrics among multiple sequence alignments from phylogenomic studies. Multiple sequence alignments of amino acid sequences from ten phylogenomic data matrices (Borowiec et al., 2015; Chen et al., 2015; Misof et al., 2014; Nagy et al., 2014; Shen et al., 2018; X.-X. Shen, Zhou, et al., 2016; Steenwyk et al., 2019; Struck et al., 2015; Whelan et al., 2015; Yang et al., 2015) were examined for five metrics: number of taxa, alignment length, number of constant sites, number of parsimony-informative sites, and number of variable sites. The x-axis depicts the last name of the first author of the phylogenomic study followed by a description of the organisms that were under study. The abbreviation PI represents parsimony-informative sites. Although excluded here for simplicity and clarity, BioKIT also determines character state frequency (nucleotide or amino acid) when summarizing alignment metrics.

the multiple sequence alignment of vertebrate sequences, the second longest alignment of budding yeast sequences (1,162,805 sites; *Shen, 332 Yeast* in Figure 44) has few constant sites (N = 2,761) and many parsimony-informative (N = 1,152,145) and variable sites (N = 1,160,044). This observation may be driven in part by the rapid rate of budding yeast evolution compared to animals (Shen et al., 2018). These results demonstrate BioKIT is useful in summarizing multiple sequence alignments.

Relative synonymous codon usage in 107 budding yeast and filamentous fungi

To demonstrate the utility of BioKIT in analyzing protein coding sequences, we calculated the relative synonymous codon usage of all codons in the protein coding sequences of 103 Eurotiomycetes (filamentous fungi) and 68 Saccharomycetes (budding yeasts) genomes obtained from the RefSeq database of NCBI (Figure 45). This example also demonstrates the flexibility of

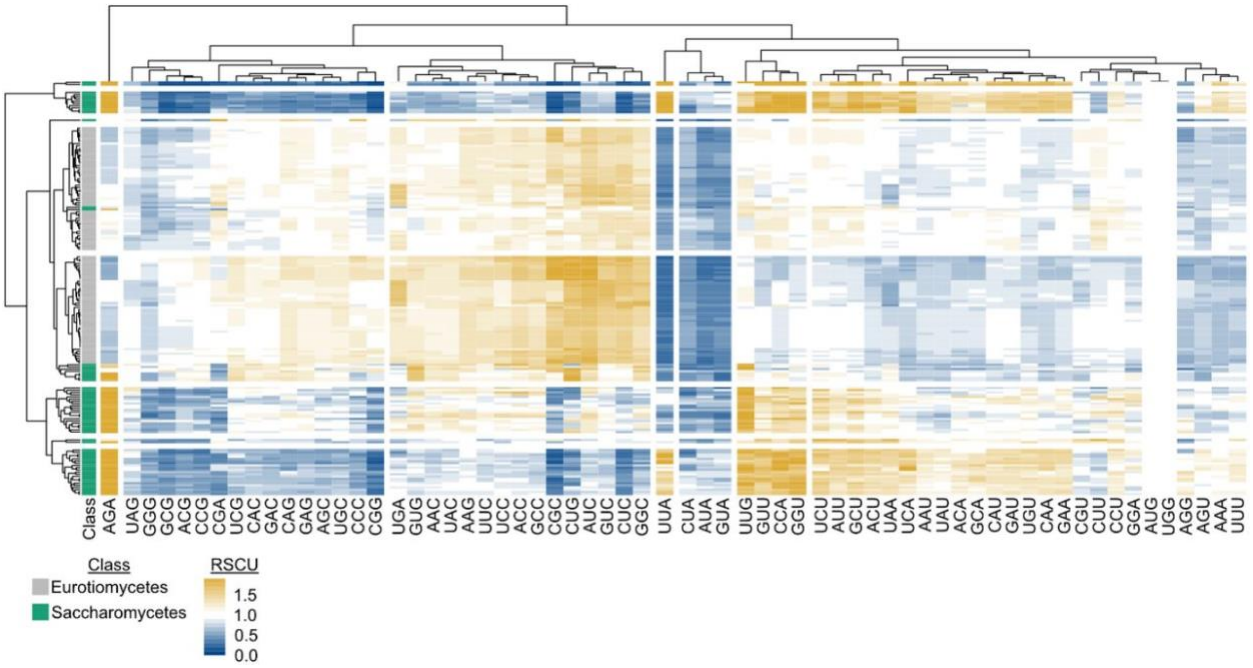

Fig. 45. Relative synonymous codon usage across 171 fungal genomes.
Relative synonymous codon usage (RSCU) was calculated from the coding sequences of 103 Eurotiomycetes (filamentous fungi) and 68 Saccharomycetes (budding yeasts) genomes obtained from NCBI. Hierarchical clustering was conducted across the fungal species (rows) and codons (columns). Eight groups of clustered rows were identified; seven groups of clustered columns were identified. Broad differences were observed in the RSCU values of Eurotiomycetes and Saccharomycetes genomes. For example, Saccharomycetes tended to have higher RSCU values for the AGA codon, whereas Eurotiomycetes tended to have higher RSCU values for the CUG codon. To account for the use of an alternative genetic code in budding yeast genomes from the CUG-Ser1 and CUG-Ser2 lineages, the alternative yeast nuclear code—which is one of 26 alternative genetic codes incorporated into BioKIT—was used during RSCU determination. User's may also provide their own genetic code if it is unavailable in BioKIT. Overrepresented codons (RSCU>1) are depicted in a gold gradient; underrepresented codons (RSCU<1) are

depicted in a blue gradient. RSCU values greater than 2 are depicted with the maximum gold color. Eurotiomycetes are depicted in grey; Saccharomycetes are depicted in green.

BioKIT to account for non-standard genetic codes, which are observed among some budding yeasts that use the CUG codon to encode a serine or alanine rather than a leucine (Krassowski et al., 2018). Hierarchical clustering of relative synonymous codon usage values per codon (columns in Figure 45) revealed similar patterns across groups of codons. For example, CUA, AUA, and GUA—three of the four codons that end in UA—were underrepresented in all fungi. Hierarchical clustering of relative synonymous codon usage values per species (rows in Figure 45) revealed filamentous fungi and budding yeasts often clustered separately. For example, UGA, GUG, AAC, UAC, AAG, UUC, UCC, ACC, GCC, CGC, CUG, AUC, GUC, CUC, and GGC are more often overrepresented among filamentous fungi in comparison to budding yeasts; in contrast, UUG, GUU, CCA, and GGU are more often overrepresented among budding yeasts in comparison to filamentous fungi. Variation within each lineage was also observed; for example, UUA was underrepresented in most, but not all, budding yeasts.

Patterns of gene-wise codon usage bias can be used to assess codon optimization and predict steady-state gene expression levels

To evaluate the utility of BioKIT in examining gene-wise codon usage biases, we calculated the mean and median gw-RSCU value, a novel metric introduced in the present manuscript, for individual protein coding genes in the genome of *S. cerevisiae* (Figure 46A). Mean and median gw-RSCU values were often, but not always, similar—the average absolute difference between mean and median gw-RSCU is 0.05 ± 0.04. In *S. cerevisiae*, as well as other organisms, genes encoding ribosomal components and histones are known to be codon optimized and highly

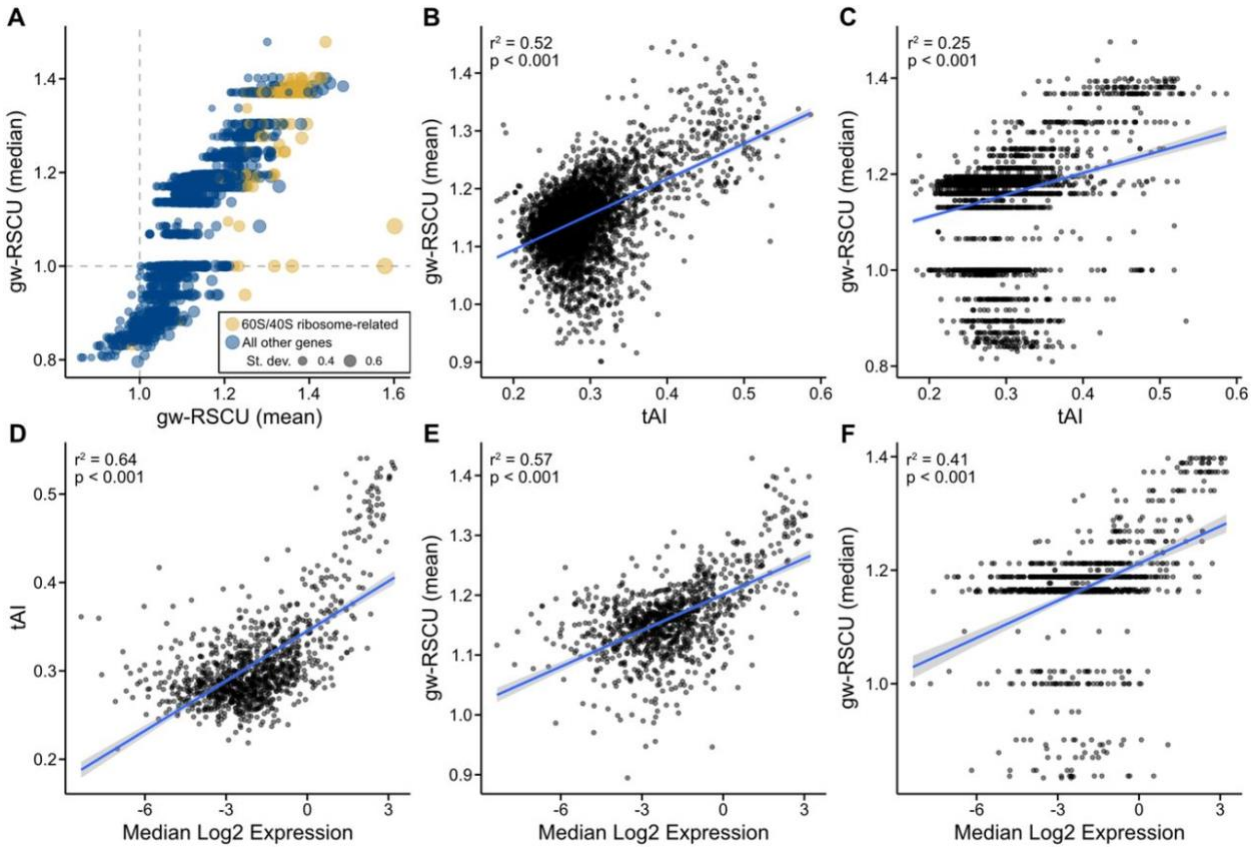

Fig. 46. Mean gene-wise relative synonymous codon usage accurately estimates codon optimization.

(A) Gene-wise relative synonymous codon usage (gw-RSCU), the mean (x-axis) or median (y-axis) relative synonymous codon usage value per gene (based on RSCU values calculated from the entire set of protein coding genes), was calculated from the coding sequences of the model budding yeast *Saccharomyces cerevisiae*. (B, C) In *S. cerevisiae*, a significant correlation was observed between tRNA adaptation index (tAI), a well-known measure of codon optimization, and mean as well as median gw-RSCU ($r^2 = 0.52$, $p < 0.001$ and $r^2 = 0.25$, $p < 0.001$, respectively; Pearson's Correlation Coefficient). (D) Using previously published data, a correlation is observed between median log2 gene expression and tAI in *Saccharomyces mikatae*, which is evidence of tAI values being indicative of codon optimization. Comparison of mean and median gw-RSCU (E and F, respectively) and median log2 gene expression revealed similarly strong correlations ($r^2 = 0.57$, $p < 0.001$ and $r^2 = 0.41$, $p < 0.001$, respectively; Pearson's Correlation Coefficient). Of note, mean gw-RSCU had a strong correlation to gene expression than median gw-RSCU. Each gene is represented by a dot. In panel A, the size of each dot represents the standard deviation of RSCU values observed in the gene and the color of each dot represents if the protein encoded by the gene has functions related to the 60S and 40S ribosomal subunits (gold) or a different function (blue).

expressed (Sharp et al., 1986; Hershberg and Petrov, 2009; LaBella et al., 2021). Therefore, we hypothesized that genes with high gw-RSCU values will have functions related to ribosomes or histones because patterns of gene-wise codon usage bias may be indicative of codon optimization. Supporting this hypothesis, examination of the 10 genes with the highest mean gw-RSCU revealed five genes with ribosome-related functions [RPL41B (YDL133C-A), mean gw-RSCU: 1.60; RPL41A (YDL184C), mean gw-RSCU: 1.58; RPS14A (YCR031C), mean gw-RSCU: 1.44; RPS9B (YBR189W), mean gw-RSCU: 1.43; and RPL18A (YOL120C), mean gw-RSCU: 1.43] and four genes with histone-related functions [HHF1 (YBR009C), mean gw-RSCU: 1.45; HTA2 (YBL003C), mean gw-RSCU: 1.44; HHF2 (YNL030W), mean gw-RSCU: 1.43; and HTA1 (YDR225W), mean gw-RSCU: 1.43]. Examination of the 10 most optimized genes according to median gw-RSCU revealed similar observations wherein nine genes had ribosome-related functions [RPS14A (YCR031C), median gw-RSCU: 1.48; RPS12 (YOR369C), median gw-RSCU: 1.40; RPS30B (YOR182C), median gw-RSCU: 1.40; RPP2A (YOL039W), median gw-RSCU: 1.40; RPL18A (YOL120C), median gw-RSCU; RPS3 (YNL178W), median gw-RSCU: 1.40; RPL13B (YMR142C), median gw-RSCU: 1.40; RPP0 (YLR340W), median gw-RSCU: 1.40; and RPS0B (YLR048W), median gw-RSCU: 1.40]. More broadly, genes associated with the 60S and 40S ribosomal units (gold color in Figure 46A) tended to have high gw-RSCU values. These results suggest gw-RSCU values may be useful for estimating codon optimization.

To further explore the relationship between gw-RSCU and codon optimization, we compared gw-RSCU values to the values of the tRNA adaptation index, a measure of codon optimization (Sabi and Tuller, 2014), in *S. cerevisiae* as well as in steady state gene expression data from

Saccharomyces mikatae (LaBella et al., 2019). In *S. cerevisiae*, strong correlation was observed between mean gw-RSCU and tRNA adaptation index values (Figure 46B) and a less robust, but still significant, correlation was observed between median gw-RSCU and tRNA adaptation index values (Figure 46C). Examination of gw-RSCU and gene expression data from *S. mikatae* revealed a robust correlation (Figure 46E and 46F) suggesting gw-RSCU, and in particular the mean gw-RSCU, can serve as a measure of gene-wise codon optimization.

Discussion

BioKIT is a multi-purpose toolkit that has diverse applications for bioinformatics research. The utilities implemented in BioKIT aim to facilitate the execution of seamless bioinformatic workflows that handle diverse sequence file types. Implementation of state-of-the-art software development and design principles in BioKIT help ensure faithful function and archival stability. BioKIT will be helpful for bioinformaticians with varying levels of expertise and biologists from diverse disciplines including molecular biology.

CHAPTER 11

PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data[10]

Introduction

Multiple sequence alignments (MSAs) and phylogenetic trees are widely used in numerous disciplines, including bioinformatics, evolutionary biology, molecular biology, and structural biology. As a result, the development of user-friendly software that enables biologists to process and analyze MSAs and phylogenetic trees is an active area of research (Kapli et al., 2020).

In recent years, numerous methods have proven useful for diagnosing potential biases and inferring biological events in genome-scale phylogenetic (or phylogenomic) datasets. For example, methods that evaluate sequence composition biases in MSAs (Phillips and Penny, 2003), signatures of clock-like evolution in phylogenetic trees (Liu et al., 2017), phylogenetic treeness (Lanyon, 1988; Phillips and Penny, 2003), taxa whose long branches may cause variation in their placement on phylogenetic trees (Struck, 2014), and others have assisted in summarizing the information content in phylogenomic datasets and improved phylogenetic inference (Felsenstein, 1978; Philippe et al., 2011; Salichos and Rokas, 2013; Doyle et al., 2015; Liu et al., 2017; Smith et al., 2018; Walker et al., 2019).

[10]This work is published in: Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X.-X., and Rokas, A. (2021). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics. doi:10.1093/bioinformatics/btab096.

Other methodological innovations include identifying significant gene-gene covariation of evolutionary rates, which has been shown to accurately and sensitively identify genes that have shared functions, are co-expressed, and/or are part of the same multimeric complexes (Sato et al., 2005; Clark et al., 2012). Furthermore, gene-gene covariation serves as a powerful evolution-based genetic screen for predicting gene function (Brunette et al., 2019). Lastly, a recently developed method has enabled the identification of unresolved internal branches or polytomies in species trees (Sayyari and Mirarab, 2018; One Thousand Plant Transcriptomes Initiative, 2019); such branches can stem from rapid radiation events or from lack of data (Rokas and Carroll, 2006).

Despite the wealth of information in MSAs and phylogenetic trees, there is a dearth of tools that enable researchers to conduct these analyses in a unified framework. For example, to utilize the functions mentioned in the previous paragraphs, a combination of web-server applications, 'hard-coded' scripts available through numerous repositories and supplementary material, standalone software, and/or extensive programming in languages including R, Python, or C is currently required (Cock et al., 2009a; Junier and Zdobnov, 2010; Revell, 2012; Talevich et al., 2012; Kück and Longo, 2014; Struck, 2014; Wolfe and Clark, 2015; Huerta-Cepas et al., 2016; Brown et al., 2017; Hernández et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). As a result, integrating these functions into bioinformatic pipelines can be challenging, reducing their accessibility to the scientific community.

To facilitate the integration of these methods into bioinformatic pipelines, we introduce PhyKIT, a UNIX shell toolkit with 30 functions (Table 1 from Steenwyk et al., 2021b) that have broad

utility for analyzing and processing MSAs and phylogenetic trees. Exemplary functions implemented in PhyKIT include measuring topological similarity of phylogenetic trees, creating codon-based MSAs, concatenating sets of MSAs into phylogenomic datasets, editing and/or viewing alignments and phylogenetic trees, and identifying putatively spurious homologs in MSAs. We highlight three uses of PhyKIT: (1) calculating diverse statistics that summarize the information content and potential biases (e.g., sequence- or phylogeny-based biases) in MSAs and phylogenetic trees; (2) creating a gene-gene covariation network; and (3) inferring the presence of polytomies from phylogenomic data. The diverse functions implemented in PhyKIT will likely be of interest to bioinformaticians, molecular biologists, evolutionary biologists, and others.

Materials and Methods

PhyKIT is a command line tool for the UNIX shell environment written in the Python programming language (https://www.python.org/). PhyKIT requires few dependencies (Biopython (Cock et al., 2009a) and SciPy (Virtanen et al., 2020)) making it user-friendly to install and integrate into existing bioinformatic pipelines. Online documentation of PhyKIT comes complete with tutorials that detail use cases for various functions. Lastly, PhyKIT is modularly designed to allow straightforward integration of additional functions in future versions.

PhyKIT has 30 different functions that help process and analyze MSAs and phylogenetic trees (Table 1 from Steenwyk et al., 2021b). The 30 functions can be grouped into broad categories that assist in conducting analyses of MSAs and phylogenies or in processing/editing them. For

example, “analysis” functions help examine information content biases, gene-gene covariation, and polytomies in phylogenomic datasets; “processing/editing” functions help prune tips from phylogenies, collapse poorly supported bipartitions in phylogenetic trees, concatenate sets of MSAs into a single data matrix, or create codon-based alignments from protein alignments and their corresponding nucleotide sequences.

Detailed information about each one of PhyKIT’s functions and tutorials for using the software can be found in the online documentation (https://jlsteenwyk.com/PhyKIT). Here, we focus on three specific groups of functions implemented in PhyKIT that enable researchers to summarize information content in phylogenomic datasets, create gene-gene evolutionary rate covariation networks, and identifying polytomies in phylogenomic data.

Evaluating information content and biases in phylogenomic datasets

MSAs and phylogenetic trees are frequently examined to evaluate their information content and potential biases in characteristics such as sequence composition or branch lengths (Phillips and Penny, 2003; Philippe et al., 2011; Struck, 2014; Doyle et al., 2015; Shen et al., 2016a; Liu et al., 2017; Smith et al., 2018). PhyKIT implements numerous functions for doing so. We demonstrate the application of 14 functions:

(1) *Alignment length.* The length of a multiple sequence alignment, which is associated with robust bipartition support and tree accuracy (Shen et al., 2016a; Walker et al., 2019);

(2) *Alignment length with no gaps*. The length of a multiple sequence alignment after excluding sites with gaps, which is associated with robust bipartition support and tree accuracy (Shen et al., 2016a);

(3) *Degree of violation of a molecular clock (DVMC)*. A metric used to determine the clock-like evolution of a gene using the standard deviation of branch lengths for a single gene tree (Liu et al., 2017). DVMC is calculated using the following formula:

$$DVMC = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(i_j - \bar{\imath}\right)^2}$$

where *N* represents the number of tips in a phylogenetic tree, i_j being the distance between the root of the tree and species *j*, and $\bar{\imath}$ represents the average root to tip distance. DVMC can be used to identify genes with clock-like evolution for divergence time estimation (Liu et al., 2017);

(4) *Internal branch lengths*. Summary statistics of internal branch lengths in a phylogenetic tree are reported including mean, median, 25th percentile, 75th percentile, minimum, maximum, standard deviation, and variance values. Examination of internal branch lengths is useful in evaluating phylogenetic tree shape;

(5) *Long branch score*. A metric that examines the degree of taxon-specific long branch attraction (Struck, 2014; Weigert et al., 2014). Long branch scores of individual taxa are calculated using the following formula:

$$LB_i = \left(\frac{\overline{PD_i}}{\overline{PD_{all}}} - 1\right) \times 100$$

where $\overline{PD_i}$ represents the average pairwise patristic distance of taxon *i* to all other taxa, $\overline{PD_{all}}$ represents the average patristic distance across all taxa, and LB_i represents the long branch score of taxon *i*. Long branch scores can be used to evaluate heterogeneity in tip-to-root distances and identify taxa that may be susceptible to long branch attraction;

(6) *Pairwise identity*. Pairwise identity is a crude approximation of the evolutionary rate of a gene and is calculated by determining the average number of sites in an MSA that are the same

character state between all pairwise combinations of taxa. This can be used to group genes based on their evolutionary rates (e.g., faster-evolving genes vs. slower-evolving ones) (Chen et al., 2017);

(7) *Patristic distances*. Patristic distances refer to all distances between all pairwise combinations of tips in a phylogenetic tree (Fourment and Gibbs, 2006), which can be used to evaluate the rate of evolution in gene trees or taxon sampling density in species trees;

(8) *Parsimony-informative sites*. Parsimony-informative sites are those sites in an MSA that have a least two character states (excluding gaps) that occur at least twice (Kumar et al., 2016); the number of parsimony-informative sites is associated with robust bipartition support and tree accuracy (Shen et al., 2016a; Steenwyk et al., 2020b);

(9) *Variable sites*. Variable sites are those sites in an MSA that contain at least two different character states (excluding gaps) (Kumar et al., 2016); the number of variable sites is associated with robust bipartition support and tree accuracy (Shen et al., 2016a);

(10) *Relative composition variability*. Relative composition variability is the average variability in the sequence composition among taxa in an MSA. Relative composition variability is calculated using the following formula:

$$Relative\ composition\ variability = \sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left|c_{ij} - \overline{c_i}\right|}{s \times n}$$

where c is the number of different character states per sequence type, n is the number of taxa in an MSA, c_{ij} is the number of occurrences of the ith character state for the jth taxon, $\overline{c_i}$ is the average number of the ith c character state across n taxa, and s refers to the total number of sites (characters) in an MSA. Relative composition variability can be used to evaluate potential sequence composition biases in MSAs, which in turn violate assumptions of site composition homogeneity in standard models of sequence evolution (Phillips and Penny, 2003);

(11) *Saturation.* Saturation refers to when an MSA contains many sites that have experienced multiple substitutions in individual taxa. Saturation is estimated from the slope of the regression line between patristic distances and pairwise identities. Saturated MSAs have reduced phylogenetic information and can result in issues of long branch attraction (Lake, 1991; Philippe et al., 2011);

(12) *Total tree length*. Total tree length refers to the sum of internal and terminal branch lengths and is calculated using the following formula:

$$total\ tree\ length = \sum_{i=1}^{a} l_i + \sum_{j=1}^{b} l_j$$

Where l_i is the branch length of the *i*th branch of *a* internal branches and l_j is the branch length of the *j*th branch of *b* terminal branches. Total tree length measures the inferred total amount or rate of evolutionary change in a phylogenetic tree;

(13) *Treeness*. Treeness (also referred to as stemminess) is a measure of the inferred relative amount or rate of evolutionary change that has taken place on internal branches of a phylogenetic tree (Lanyon, 1988; Phillips and Penny, 2003) and is calculated using the following formula:

$$treeness = \sum_{u=1}^{b} \frac{l_u}{l_t}$$

where l_u is the branch length of the *u*th branch of *b* internal branches, and l_t refers to the total branch length of the phylogenetic tree. Treeness can be used to evaluate how much of the total tree length is observed among internal branches;

(14) *Treeness divided by relative composition variability*. This function combines two metrics to measure both composition bias and other biases that may negatively influence phylogenetic inference. High treeness divided by relative composition variability values have been shown to

be less susceptible to sequence composition biases and are associated with robust bipartition support and tree accuracy (Phillips and Penny, 2003; Shen et al., 2016a).

Calculating gene-gene evolutionary rate covariation or coevolution

Genes that share similar rates of evolution through speciation events (or coevolve) tend to have similar functions, expression levels, or are parts of the same multimeric complexes (Sato et al., 2005; Clark et al., 2012). Thus, identifying significant coevolution between genes (i.e., identifying genes that are significantly correlated in their evolutionary rates across speciation events) can be a powerful evolution-based screen to determine gene function (Brunette et al., 2019).

To measure gene-gene evolutionary rate covariation, PhyKIT implements the mirror tree method (Pazos and Valencia, 2001; Sato et al., 2005), which examines whether two trees have correlated branch lengths. Specifically, PhyKIT calculates the Pearson correlation coefficient between branch lengths in two phylogenetic trees that share the same tips and topology. To account for differences in taxon representation between the two trees, PhyKIT first automatically determines which taxa are shared and prunes one or both such that the same set of taxa is present in both trees. PhyKIT requires that the two input trees have the same topology, which is typically the species tree topology inferred from whole genome or proteome data. Thus, the user will typically first estimate a gene's branch lengths by constraining the topology to match that of the species tree. When running this function, users should be aware that many biological factors, such as horizontal transfer (Doolittle and Bapteste, 2007), incomplete lineage sorting (Degnan and Salter, 2005), and introgression / hybridization (Sang and Zhong, 2000), can lead to gene

histories that deviate from the species tree. In these cases, constraining a gene's history to match that of a species may lead to errors in the covariation analysis.

Due to factors including time since speciation and mutation rate, correlations between uncorrected branch lengths result in a high frequency of false positive correlations (Sato et al., 2005; Clark et al., 2012; Chikina et al., 2016). To ameliorate the influence of these factors, PhyKIT first transforms branch lengths into relative rates. To do so, branch lengths are corrected by dividing the branch length in the gene tree by the corresponding branch length in the species tree. Previous work revealed that one or a few outlier branch length values can be responsible for false positive correlations and should be removed prior to analysis (Clark et al., 2012). Thus, PhyKIT removes outlier data points defined as having corrected branch lengths greater than five (i.e., removing gene tree branch lengths that are five or more times greater than their corresponding species tree branch lengths). Lastly, values are converted into relative rates using a Z-transformation. The resulting relative rates are used when calculating Pearson correlation coefficients.

Identifying polytomies in phylogenomic data

Rapid radiations or diversification events have occurred throughout the tree of life including among mammals, birds, plants, and fungi (Jarvis et al., 2014; Liu et al., 2017; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). Polytomies correspond to internal branches whose length is 0 (or statistically indistinguishable from 0) and can be driven either by biological (e.g., rapid radiations) or analytical (e.g., low amount of data) factors. Thus, polytomies are useful for inferring rapid radiation or diversification events and exploring

incongruence in phylogenies (Sayyari and Mirarab, 2018; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020).

To identify polytomies, a modified approach to a previous strategy was implemented (Sayyari and Mirarab, 2018). More specifically, the support for three alternative topologies is calculated among all gene trees from a phylogenomic dataset. For example, in species tree *((A,B),C), D);*, if examining the presence of a polytomy at the ancestral bipartition of tips *A, B,* and *C*, PhyKIT will determine the number of gene trees that support *((A,B),C);*, *((A,C),B);*, and *((B,C),A);* using the rooted gene trees provided by the user. Equal support for the three topologies (i.e., the presence of a polytomy) among a set of gene trees is assessed using a Chi-squared test. Failing to reject the null hypothesis is indicative of a polytomy (Sayyari and Mirarab, 2018). Note that this approach is distinct from the approach of Sayyari and Mirarab to identify polytomies because PhyKIT uses a gene-based signal rather than a quartet-based signal. The difference between the two methods is that each gene contributes equally to the inference of a polytomy when a gene-based signal is used, whereas genes with greater taxon representation (which contain a greater number of quartets) will contribute a greater signal during polytomy identification when a quartet-based signal is used. From a technical perspective, both approaches are simple to implement and require only a single line of code in the command-line.

Data availability

Data are available on figshare (doi: 10.6084/m9.figshare.13118600).

Results

We outline three example uses of PhyKIT: 1) summarizing information content and identifying potential biases in animal, plant, yeast, and filamentous fungal phylogenomic datasets (Shen et al., 2016b; Laumer et al., 2019; One Thousand Plant Transcriptomes Initiative, 2019; Steenwyk et al., 2019c), 2) constructing a network of significant gene-gene covariation, which reveals genes of shared functions from empirical data spanning ~550 million years of evolution among fungi (Shen et al., 2020b), and 3) illustrating how to identify polytomies using simulated and empirical data (Steenwyk et al., 2019c).

Summarizing information content and biases in phylogenomic data

Examining information content in phylogenomic datasets can help diagnose potential biases that stem from low signal-to-noise ratios, multiple substitutions, non-clocklike evolution, and other biological or analytical factors. To demonstrate the utility of PhyKIT to summarize the information content in phylogenomic datasets, we calculated 14 different metrics known to help diagnose potential biases in phylogenomic datasets or be associated with accurate and well supported phylogenetic inferences (Felsenstein, 1978; Phillips and Penny, 2003; Philippe et al., 2011; Struck, 2014; Doyle et al., 2015; Shen et al., 2016a; Liu et al., 2017; Smith et al., 2018) using four empirical phylogenomic datasets from animals (201 tips; 2,891 genes) (Laumer et al., 2019), budding yeast (332 taxa; 2,408 genes) (Shen et al., 2018), filamentous fungi (93 taxa; 1,668 genes) (Steenwyk et al., 2019c), and plants (1,124 taxa; 403 genes) (One Thousand Plant Transcriptomes Initiative, 2019) (Figure 47, Table 1 from Steenwyk et al., 2021b).

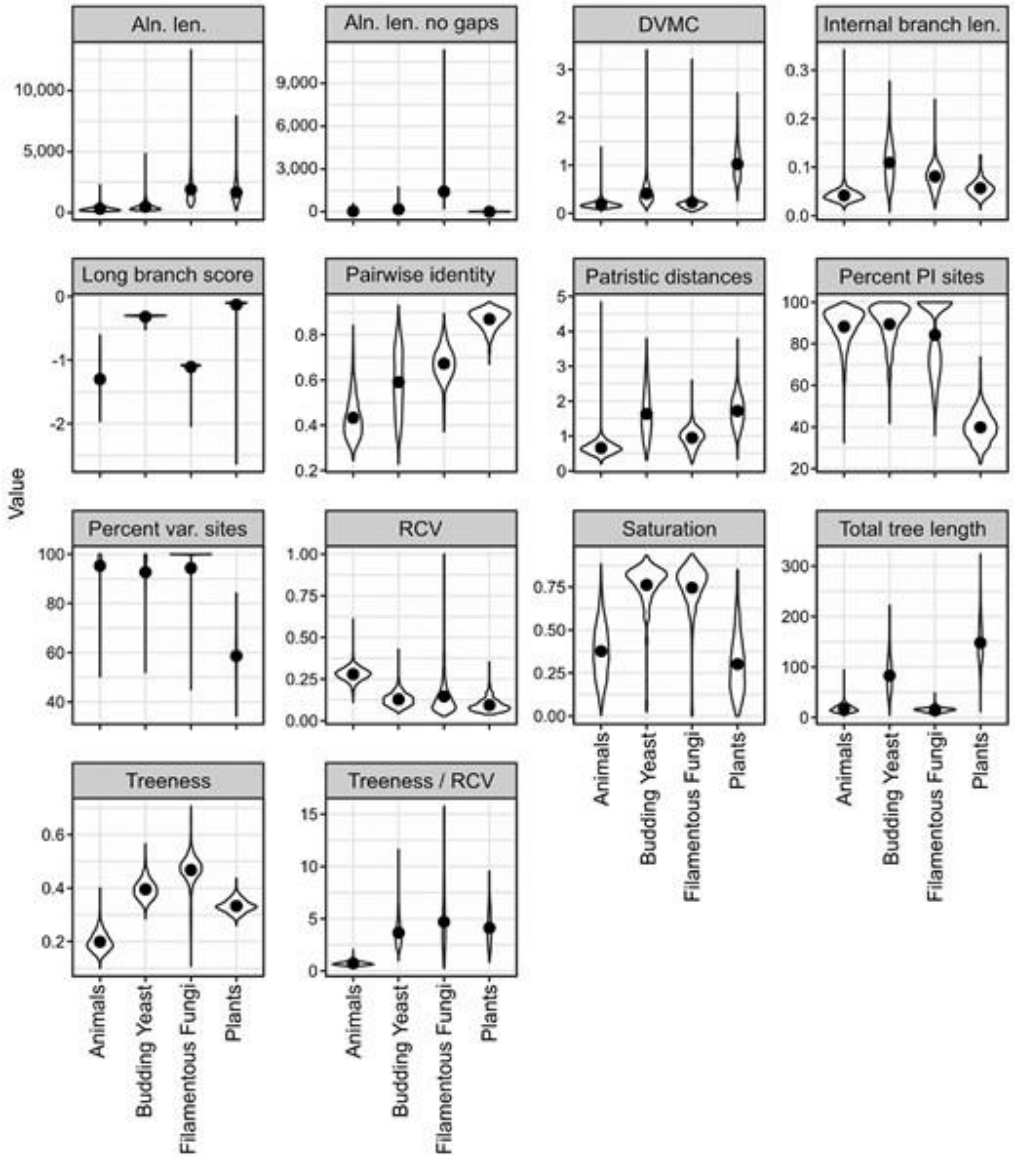

Fig. 47. Summary of information content in four empirical phylogenomic datasets.
Fourteen exemplary metrics implemented in PhyKIT help summarize the information content and identify potential biases in phylogenomic datasets. Each graph displays a violin plot with a black point representing the mean. Error bars indicate one standard error above and below the mean; however, these are difficult to see in nearly all graphs because they were often near the mean. Abbreviations are as follows: Aln. len.: alignment length; Aln. len. no gaps: alignment length excluding sites with gaps; DVMC: degree of violation of a molecular clock; Internal branch len.: average internal branch length; patristic distances: average patristic distance in a gene tree; percent PI sites: percentage of parsimony-informative sites in an MSA; percent var. sites: percentage of variable sites in an MSA; RCV: relative composition variability

Examination of the distributions of the values of the 14 different metrics revealed inter- and intra-dataset heterogeneity (Figure 47). For example, inter-dataset heterogeneity was observed among animal and plant datasets, which had the lowest and highest average pairwise identity across alignments, respectively; intra-dataset heterogeneity was observed in the uniform distribution of pairwise identities in the budding yeast datasets. Similarly, inter-dataset heterogeneity was observed in estimates of saturation where the budding yeast and filamentous fungal MSAs were less saturated by multiple substitutions than the plant and animal datasets; intra-data heterogeneity was also observed in all four datasets. Varying degrees of inter- and intra-dataset heterogeneity was observed for other information content statistics, which may be due biological (e.g., mutation rate) or analytical factors (e.g., taxon sampling, distinct alignment, trimming, and tree inference strategies).

In summary, PhyKIT is useful for examining the information content of phylogenomic datasets. For example, the generation of different phylogenomic data submatrices by selecting subsets of genes or taxa with certain properties (e.g., retention of genes with the highest numbers of parsimony-informative sites or following removal of taxa with high long branch scores) can facilitate the exploration of the robustness of species tree inference or estimating time since divergence (Salichos and Rokas, 2013; Liu et al., 2017; Shen et al., 2018; Steenwyk et al., 2019c; Walker et al., 2019; Li et al., 2020; Shen et al., 2020b).

A network of gene-gene covariation reveals neighborhoods of genes with shared function

Genes with similar evolutionary histories often have shared functions, are co-expressed, or are parts of the same multimeric complexes (Sato et al., 2005; Clark et al., 2012). Using PhyKIT, we

examined gene-gene covariation using 815 genes spanning 1,107 genomes and ~563 million years of evolution among fungi (Shen et al., 2020b). By examining 331,705 pairwise combinations of genes, we found 298 strong signatures of gene-gene covariation (defined as $r > 0.825$). The two genes with the strongest signatures of covariation were *SEC7* and *TAO3* ($r = 0.87$), suggesting that their protein products have similar or shared functions. Supporting this hypothesis, Sec7p contributes to cell-surface growth in the model yeast *Saccharomyces cerevisiae* (Novick and Schekman, 1979) and genes with the Sec7 domain are transcriptionally coregulated with yeast-hyphal switches in the human pathogen *C. albicans* (Song et al., 2008). Similarly, Tao3p in both *S. cerevisiae* and *C. albicans* is part of a RAM signaling network, which controls hyphal morphogenesis, polarized growth, and cell-cycle related processes including cell separation, cell proliferation, and phase transitions (Bogomolnaya et al., 2006; Song et al., 2008).

Complex relationships of gene-gene covariation can be visualized as a network (Figure 48). Examination of network neighborhoods identified groups of genes that have shared functions and are parts of the same multimeric complexes. For example, the proteins encoded by *NDC80* and *NUF2* are part of the same kinetochore-associated complex termed the NDC80 complex—which is required for efficient mitosis (Sundin et al., 2011)—and significantly covary with one another ($r = 0.84$). Similarly, multiple genes that encode proteins involved in DNA replication and repair (i.e., *POL2*, *MSH6*, *RAD26*, *CDC9*, and *EXO1*) were part of the same network neighborhood, consistent with previous work suggesting an intimate interplay between DNA replication and multiple DNA repair pathways (Tsubouchi and Ogawa, 2000; Lujan et al., 2012; Boiteux and Jinks-Robertson, 2013). Other network neighborhoods of genes with shared function such as

ribosome biogenesis, Golgi apparatus-related transport, and control of DNA replication were identified (Figure 48).

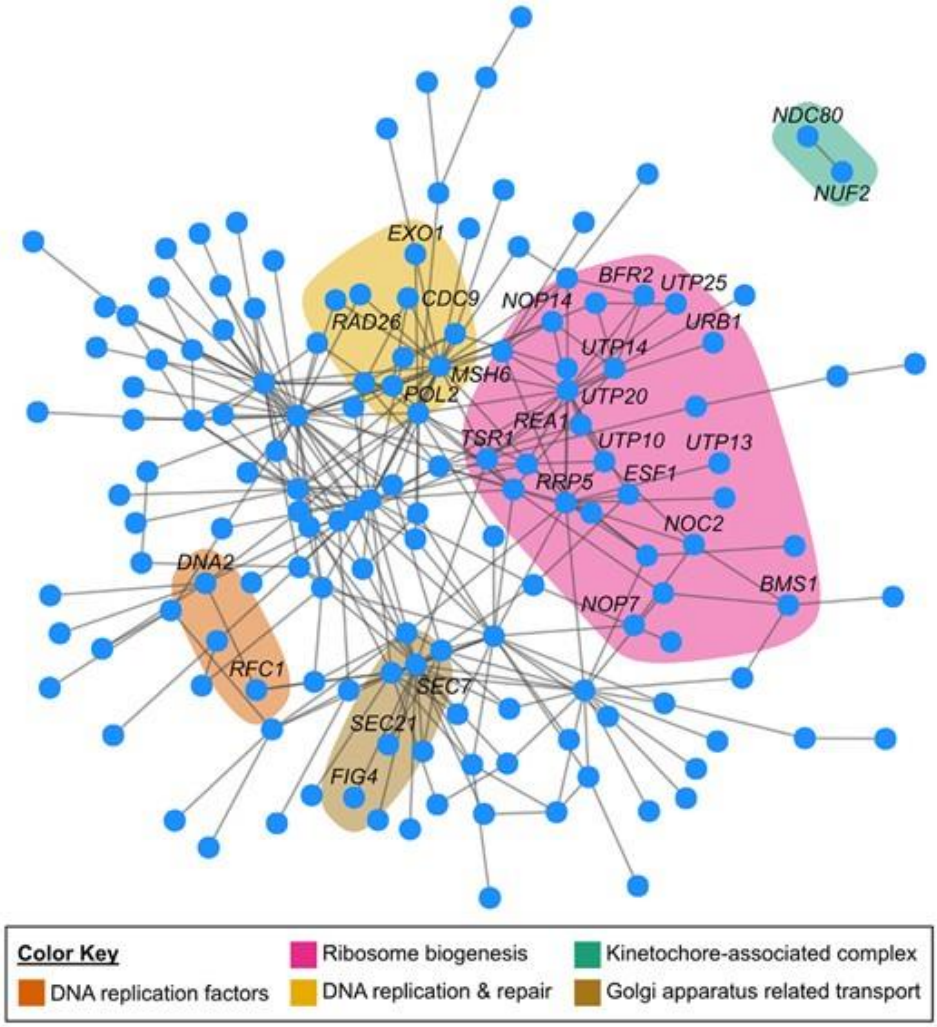

Fig. 48. Gene–gene covariation network inferred from ~550 million years of evolution across 1107 fungi.

A network of significant gene–gene coevolution identifies network neighborhoods representative of associated functional categories. For example, the NDC80 and NUF2 genes (toward the top right of the network) were identified to be significantly coevolving with one another ($r = 0.84$, $P < 0.01$, Pearson's correlation test); they both encode proteins that are part of the same multimeric kinetochore-associated complex (green). Similarly, genes that are DNA replication factors (orange), contribute to DNA replication and repair processes (yellow), participate in

Golgi apparatus-related transport (brown) or ribosome biogenesis (pink) were found to be neighbors in the network

---

Taken together, these results indicate PhyKIT is a useful tool for evaluating gene-gene covariation and predicting genes' functions (Sato et al., 2005; Clark et al., 2012; Brunette et al., 2019). Thus, we anticipate PhyKIT will be helpful for evaluating gene-gene covariation and conducting evolution-based screens for gene functions across the tree of life.

Identifying polytomies in phylogenomic datasets

Rapid radiations or diversification events have occurred throughout the tree of life (Jarvis et al., 2014; Liu et al., 2017; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). One approach to identifying rapid radiations is by testing for the existence of polytomies in species trees (Sayyari and Mirarab, 2018; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). Polytomies can also arise when the amount of data at hand is insufficient for resolution (Walsh et al., 1999). To demonstrate the utility of PhyKIT to identify polytomies, we examined the ability of our approach to identify a simulated polytomy (Figure 49A). PhyKIT was able to conservatively identify the simulated polytomy demonstrating the efficacy of our approach.

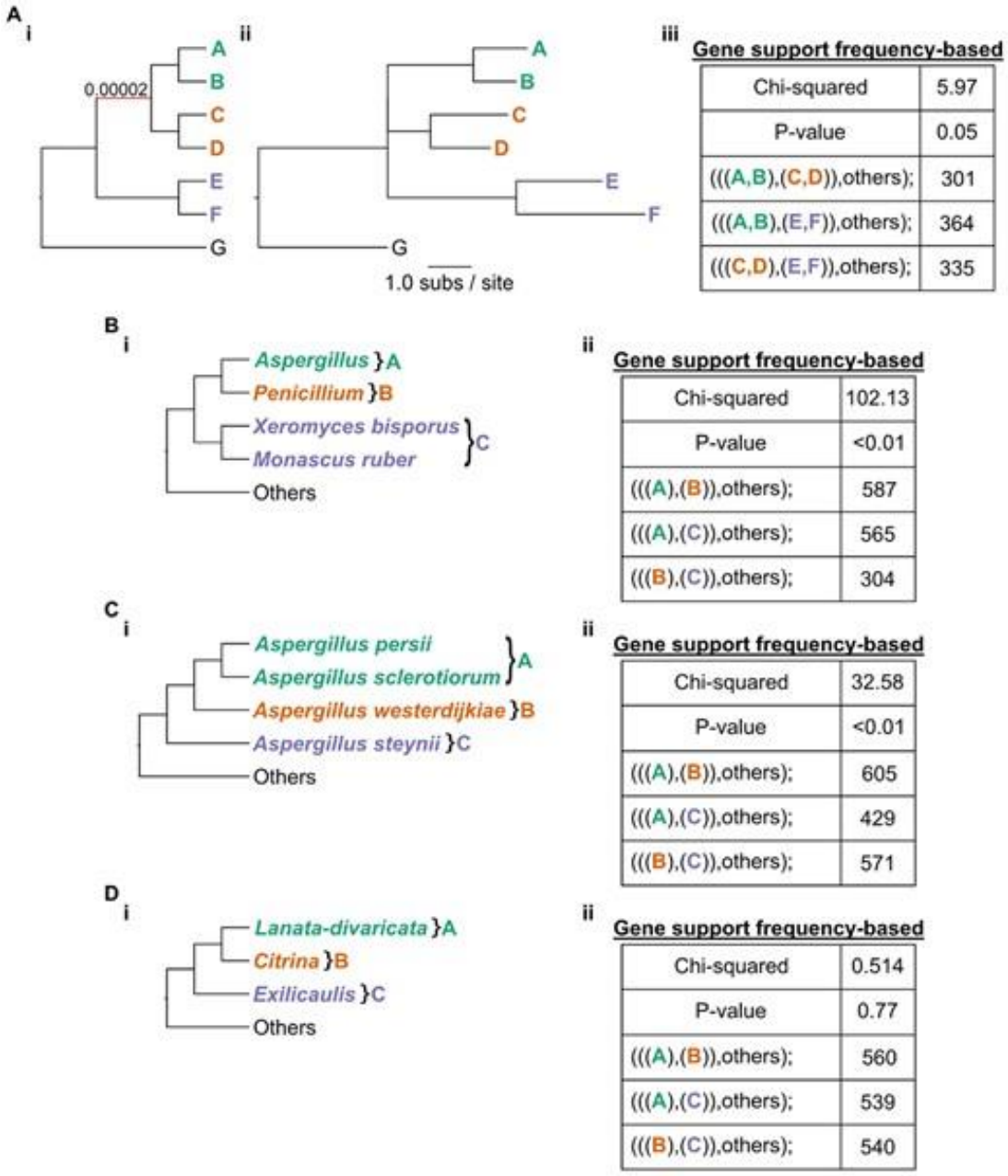

Fig. 49. Identifying polytomies from phylogenomic data.
(Ai) A cladogram of a simulated species phylogeny with tip names A–G. The red branch has a very short branch length of $2 \times 10 - 5$ substitutions per site. (Aii) Phylogram of the same phylogeny shows that all other branches are much longer (≥1.0 substitutions per site). (Aiii) After reconstructing the evolutionary history from 1000 alignments simulated from the phylogeny in Aii, the hypothesis of a polytomy was tested using gene-support frequencies for three alternative rooted topologies defined by the clades of green, orange and purple taxa. Failure to reject the null hypothesis of equal support among genes for each topology is indicative of a polytomy ($\chi 2 = 5.97$, P-value = 0.05, Chi-squared test). (B–D) The same approach was then used to examine if there is evidence for a polytomy at three different branches in a phylogeny of filamentous fungi. (D) Support for a polytomy ($\chi 2 = 0.514$, P-value = 0.77, Chi-squared test) was observed for the relationships between three different sections of Penicillium fungi. These results

demonstrate the utility of gene-support frequencies for evaluating polytomies and examining incongruence in phylogenomic datasets.

We next examined if there is evidence of polytomies in the evolutionary history of filamentous fungi from the genera *Aspergillus* and *Penicillium*. We examined three branches. The first two branches—one dating back ~110 million years ago (Figure 49B), and another dating back ~25 million years ago (Figure 49C)—were not polytomies. In contrast, examination of a ~60 million-year-old branch involving *Lanata-divaricata, Citrina,* and *Exilicaulis* (Figure 49D), which are major lineages (or sections) in the genus *Penicillium*, was consistent with a polytomy. Given the large number of gene trees used in our analysis (n=1,668), these results are consistent with a rapid radiation or diversification event in the history of *Penicillium* species.

In summary, these results suggest that PhyKIT is useful in identifying polytomies in simulated and empirical datasets. More broadly, these results support the notion that polytomies can be used to identify rapid radiation events. Beyond polytomy identification, PhyKIT can be used for exploring incongruence in phylogenies by calculating gene support frequencies. Calculations of gene-based support among different topologies can be used in diverse applications, including identifying putative introgression / hybridization events and conducting phylogenetically-based genome-wide association (PhyloGWAS) studies (Pease et al., 2016; Steenwyk et al., 2019c).

Discussion

We developed PhyKIT, a comprehensive toolkit for processing and analyzing MSAs and trees in phylogenomic datasets. Executing functions implemented in PhyKIT would otherwise require extensive programming, multiple software, and/or web-based applications (Table 1 from

Steenwyk et al., 2021b); thus, PhyKIT offers users a way to streamline approaches and pipelines by relying on only one software. PhyKIT is freely available on GitHub (https://github.com/JLSteenwyk/PhyKIT), PyPi (https://pypi.org/project/phykit/), and the Anaconda Cloud (https://anaconda.org/JLSteenwyk/phykit) under the MIT license with extensive documentation and user tutorials (https://jlsteenwyk.com/PhyKIT). PhyKIT is a fast and flexible toolkit for the UNIX shell environment, which allows it to be easily integrated into bioinformatic pipelines. We anticipate PhyKIT will be of interest to biologists from diverse disciplines and with varying degrees of experience in analyzing MSAs and phylogenies. In particular, PhyKIT will likely be helpful in addressing one of the greatest challenges in biology, building, understanding, and deriving meaning from the tree of life.

CHAPTER 12

ClipKIT: a multiple sequence alignment-trimming software for accurate phylogenomic inference[11]

Introduction

Multiple sequence alignment (MSA) of a set of homologous sequences is an essential step of molecular phylogenetics, the science of inferring evolutionary relationships from molecular sequence data. Errors in phylogenetic analysis can be caused by erroneously inferring site homology or saturation of multiple substitutions (Talavera and Castresana, 2007), which often present as highly divergent sites in MSAs. To remove errors and phylogenetically-uninformative sites, several methods "trim" or filter highly divergent sites using calculations of site/region dissimilarity from MSAs (Talavera and Castresana, 2007; Capella-Gutierrez et al., 2009; Criscuolo and Gribaldo, 2010; Jarvis et al., 2014). A beneficial by-product of MSA trimming, especially for studies that analyse hundreds of MSAs from thousands of taxa (Shen et al., 2020a), is that trimming MSAs reduces the computational time and memory required for phylogenomic inference. Nowadays, MSA trimming is a routine part of molecular phylogenetic inference (Kapli et al., 2020).

Despite the overwhelming popularity of MSA trimming strategies, a recent study revealed that trimming often decreases, rather than increases, the accuracy of phylogenetic inference (Tan et al., 2015). This decrease suggests that current strategies may remove

---

[11]This work is published in: Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biol. 18, e3001007. doi:10.1371/journal.pbio.3001007.

phylogenetically-informative sites (e.g., parsimony-informative and variable sites) that have previously been shown to contribute to phylogenetic accuracy (Shen et al., 2016a). Furthermore, it was shown that phylogenetic inaccuracy is positively associated with the number of removed sites (Tan et al., 2015), revealing a speed-accuracy trade-off wherein trimmed MSAs decrease the computation time of phylogenetic inference but at the cost of reduced accuracy. More broadly, these findings highlight the need for alternative MSA trimming strategies.

To address this need, we developed ClipKIT, an MSA-trimming algorithm based on a conceptually novel framework. Rather than aiming to identify and remove putatively phylogenetically-uninformative sites in MSAs, ClipKIT instead focuses on identifying and retaining parsimony-informative sites, which (alongside other types of sites and features of MSAs, such as variable sites and alignment length) have previously been shown to be phylogenetically informative (Shen et al., 2016a). ClipKIT implements a total of five different trimming strategies. Certain ClipKIT trimming strategies allow users to also retain constant sites, which inform base frequencies in substitution models (Nguyen et al., 2015), and / or trim alignments based on the fraction of taxa represented by gaps per site (or site gappyness). We tested the accuracy and support of phylogenetic inferences using ClipKIT and other alignment trimming software using nearly 140,000 alignments from empirical datasets of mammalian and budding yeast sequences (Shen et al., 2016a) and simulated datasets of metazoans, plants, filamentous fungi, and a larger sampling of budding yeasts sequences (Xi et al., 2014; Whelan et al., 2015; Shen et al., 2016b; Steenwyk et al., 2019c). We found that ClipKIT-trimmed alignments led to accurate and well supported phylogenetic inferences that consistently outperformed other alignment trimming software. Additionally, we note that ClipKIT-trimmed

alignments can save computation time during phylogenetic inference. Taken together, our results demonstrate that alignment trimming based on identifying and retaining parsimony-informative sites is a robust alignment trimming strategy.

Materials and Methods

ClipKIT availability and usage

ClipKIT is a standalone software written in the Python programming language (https://www.python.org/) and is available from GitHub, https://github.com/JLSteenwyk/ClipKIT, and PyPi, https://pypi.org/. Complete documentation is available online (https://jlsteenwyk.com/ClipKIT/). ClipKIT differs from most multiple sequence alignment (MSA) trimming software in that it focuses on identifying and retaining parsimony-informative sites from MSAs rather than on removing highly divergent ones. To do so, ClipKIT conducts site-by-site examination of MSAs and determines whether they should be retained or trimmed based on the strategy of ClipKIT being used and how the site has been classified. During site-by-site examination of MSAs, sites are either classified as parsimony-informative, as constant sites, or neither. Note that other types of sites and features of MSAs have previously been shown to be phylogenetically-informative (e.g., variable sites and MSA length), however, ClipKIT focuses on parsimony-informative sites. Parsimony-informative sites are defined as sites that contain at least two character states that occur in at least two taxa. Constant sites are defined as sites with only one character state that appears in at least two taxa (Kumar et al., 2016). Across the various ClipKIT strategies, parsimony-informative sites are always retained, constant sites are either retained or removed, and sites that are neither parsimony-informative nor constant are always removed.

Previous work (Jarvis et al., 2014) identified two types of "aberrant sites": (1) sites where only one sequence is represented in the alignment, and (2) sites where only two taxa are represented and lack homology (defined by a null model of genome-wide sequence similarity based on species-level divergences) to any other taxa in the alignment. For the first strategy, sites with these features in multiple sequence alignments may stem from a genuine insertion event in one taxon or from assembly, annotation, and/or alignment errors; for the second strategy, homology is defined according to a null model of expected homology based on species-level sequence divergence. ClipKIT removes any sites that are not parsimony-informative or constant; it also removes sites that contain high percentages of gaps. Thus, such "aberrant sites" are typically removed by ClipKIT.

Lastly, ClipKIT can also perform alignment trimming based on site gappyness, which is defined as the percentage of taxa that contain a gap character state (as opposed to a nucleotide or amino acid character state) at a given site. The five ClipKIT trimming strategies are summarized as follows:

1) kpi: a strategy that retains sites that are parsimony-informative, which is specified with the following command:

clipkit <MSA> -m kpi;

This strategy executes the following pseudocode:

FOR site in alignment:

>IF site is parsimony-informative

>>keep the site

>ELSE

>>remove the site

ENDFOR

2) kpic: a strategy that retains sites that are either parsimony-informative or constant, which is specified with the following command:

clipkit <MSA> -m kpic;

This strategy executes the following pseudocode:

FOR site in alignment:

>IF site is parsimony-informative or constant

>>keep the site

>ELSE

>>remove the site

ENDFOR

3) gappy: a strategy that removes sites that are gappy-rich (defined as sites with ≥90% gaps), which is specified with the following command:

clipkit <MSA> -m gappy,

Because gappy-based trimming is the default strategy, it can also be executed with the following command:

clipkit <MSA>;

This strategy executes the following pseudocode:

FOR site in alignment:

>IF site has ≥90% gaps

>>keep the site

>ELSE

>>remove the site

ENDFOR

4) kpi-gappy: a combination of strategies 1 and 3, which is specified with the following command:

clipkit <MSA> -m kpi-gappy;

This strategy executes the following pseudocode:

FOR site in alignment:

>IF site is parsimony-informative AND has ≥90% gaps

>>keep the site

>ELSE

>>remove the site

ENDFOR

and 5) kpic-gappy: a combination of strategies 2 and 3, which is specified with the following command:

clipkit <MSA> -m kpic-gappy.

This strategy executes the following pseudocode:

FOR site in alignment:

>IF site is (parsimony-informative OR constant) AND has ≥90% gaps

>>keep the site

>ELSE

>>remove the site

ENDFOR

All output files have the same name as the input files with the addition of the suffix ".clipkit." Users can specify output files names with the -o/--output option. For example, an alignment may have the output name "ClipKIT_trimmed_aln.fa" with the following command:
clipkit <MSA> -o ClipKIT_trimmed_aln.fa.

To enable users to fine-tune alignment trimming parameters, we provide an additional option for users to specify their own gappyness threshold, which can range between zero and one. For example, to retain sites with ≥95% gaps, the following command would be used:

clipkit <MSA> -g 0.95

This gappyness threshold would execute the following pseudocode:
FOR site in alignment:

>IF site has ≥95% gaps

>>keep the site

>ELSE

>>remove the site

ENDFOR

In practice, we recommend the use of very high gappyness thresholds; the use of lower thresholds may remove too many sites, which may worsen phylogenetic inferences (Shen et al., 2016a).

To enable users to examine the trimmed sites/regions from MSAs, we have also implemented a logging option in ClipKIT. When used, the logging option outputs an additional four-column file with the following information: column 1, position in the alignment (starting at 1); column 2, whether or not the site was trimmed or kept; column 3, reports if the site was parsimony-informative, constant, or neither and; column 4, reports the gappyness of the site. Log files are generated using the -l/--log option:

clipkit <MSA> -l

We anticipate this information will be helpful for alignment diagnostics, fine-tuning of trimming parameters, and other reasons.

To enable seamless integration of ClipKIT into pre-existing pipelines, eight file types can be used as input. More specifically, ClipKIT can input and output *fasta*, *clustal*, *maf*, *mauve*, *phylip*, *phylip-sequential*, *phylip-relaxed*, and *stockholm* formatted MSAs. By default, ClipKIT automatically determines the input file format and creates an output file of the same format; however, users can specify either with the -if/--input_file_format and -of/--output_file_format options. For example, an input file of *fasta* format and a desired output file of *clustal* format can be specified using the following command:

clipkit <MSA> -if fasta -of clustal

Recent analyses indicate that ~28% of available computational tools fail to install due to implementation errors (Mangul et al., 2019b). To overcome this hurdle and ensure archival stability of ClipKIT, we implemented state-of-the-art software development practices and design principles. More specifically, ClipKIT is composed of highly modular, extensible, and reusable code, which allows for easy debugging and seamless integration of new functions and features. We wrote a total of 118 unit and integration tests resulting in 97% code coverage. We also implemented a robust continuous integration (CI) pipeline to automatically build, package, and test ClipKIT whenever code is modified. This CI pipeline runs a testing matrix for Python versions 3.6, 3.7, and 3.8. Given the current configuration, building and testing ClipKIT for future versions of Python will be straightforward. Lastly, central ClipKIT functions rely on few dependencies (i.e., BioPython (Cock et al., 2009a) and NumPy (Van Der Walt et al., 2011)). In summary, we have taken several measures to ensure ClipKIT implements MSA trimming strategies that do not sacrifice the accuracy of phylogenetic inference but also safeguard that ClipKIT will be a long-lasting computational tool for the field of molecular phylogenetics.

Practical considerations when using ClipKIT

Although ClipKIT strategies performed well across empirical genome-scale and simulated datasets, we acknowledge that testing every possible evolutionary scenario is impossible. This is further complicated by the lack of large-scale phylogenomic data matrices in which the true evolutionary relationships among organisms are known. Therefore, we recommend using multiple trimming strategies available in ClipKIT and examining the resulting ABS values for trees. Considering high ABS values often corresponded to lower nRF values (Supplementary

Figure 4 from Steenwyk et al., 2020b and 5 from Steenwyk et al., 2020b), using the resulting phylogeny with the highest ABS value may be representative of the phylogeny that most closely approximates the true evolutionary history. This may require substantially greater computation time. To potentially ameliorate the computation time issue that may arise, we recommend creating subsets of larger datasets that span alignments of various lengths and testing multiple trimming strategies on the reduced dataset.

Although constant sites are thought to be important for informing parameters of substitution models (Nguyen et al., 2015), we observed variation in the performance of ClipKIT strategies that retain only parsimony-informative sites (kpi and kpi-gappy) and the performance strategies that retain parsimony-informative and constant sites (kpic and kpic-gappy). More specifically, at times strategies kpi and kpi-gappy outperformed kpic and kpic-gappy suggesting constant sites may not always be informative to substitution models. However, we note that trimming nucleotide sequences with strategies kpi and kpi-gappy may warrant ascertainment bias correction for nucleotide sequences because constant sites are absent from the trimmed alignments.

Dataset acquisition and generation

To test the efficacy of strategies from ClipKIT and other alignment trimming software (Table 1 from Steenwyk et al., 2020b), we used a total of eight empirical and simulated datasets. For empirical datasets, we obtained publicly available untrimmed amino acid and nucleotide MSAs from 24 mammals ($N_{alignments}$=4,004) and 12 budding yeasts ($N_{alignments}$=5,664) totalling four datasets (Shen et al., 2016a). Publicly available amino acid alignments were generated with

MAFFT, v. 7.164, using the G-INS-I strategy with a gap penalty of 1.53 (Katoh and Standley, 2013). Publicly available nucleotide alignments were generated by mapping nucleotide sequences onto the amino acid alignments. For simulated datasets, we simulated sequence evolution along proposed species phylogenies of 93 filamentous fungi (Steenwyk et al., 2019c), and from simulated amino acid sequence evolution along the species phylogenies of 70 metazoans (Whelan et al., 2015), 46 flowering plants (Xi et al., 2014), and 96 budding yeasts (Shen et al., 2016b) ($N_{alignments}$=50 alignments per dataset or 200 total).

Simulated sequences were generated with INDELible, v1.03 (Fletcher and Yang, 2009) using parameters suggested by the software developers. INDELible was chosen to generate simulated sequences because of its ability to also simulate insertion and deletion events, which are represented by gaps, a common feature in multiple sequence alignments. Nucleotide alignments for filamentous fungi were generated using the general time reversible (GTR) substitution model (Waddell and Steel, 1997). Additional parameters specified were state frequencies values of 0.1, 0.2, 0.3, and 0.4 for T, C, A, and G nucleotides, respectively. We specified the substitution rate matrix using the scheme outlined in Supplementary table 1 from Steenwyk et al., 2020b. Insertion and deletion rates were set to be 5% as frequent as single substitutions. Insertion and deletions occurred according to the power law distribution (a=1.7, M=500). The tree's root length was set to 1,000. For amino acids, all parameters were the same except the insertion and deletion rates were set to be 1% as frequent as single substitutions using the WAG model of substitutions, which was also used to specify state frequencies (Whelan and Goldman, 2001).

The resulting empirical and simulated MSAs were trimmed using 14 popular alignment trimming strategies (Table 1 from Steenwyk et al., 2020b). Altogether, we generated a total of 138,152 MSAs [(4,004 mammalian + 5,664 yeast + 200 simulated MSAs) * (14 trimming strategies, including a "no trimming" strategy) = 138,152 MSAs], which were used to evaluate the performance of ClipKIT and other alignment trimming strategies.

Measuring accuracy and support of phylogenetic inferences

Phylogenetic inferences from MSAs were made using IQ-TREE, v1.6.11 (Nguyen et al., 2015). For nucleotide sequences, we used a GTR substitution model (Tavaré, 1986) with empirical base frequencies and a discrete Gamma model with four rate categories (Yang, 1994) or "GTR+F+G"; for amino acid sequences, we used the general WAG model of substitutions (Whelan and Goldman, 2001) with empirical base frequencies and a discrete Gamma model with four rate categories (Yang, 1994) or "WAG+F+G."

Tree accuracy was measured using normalized Robinson-Foulds (nRF) distances as calculated by ape, v5.3 (Paradis et al., 2004), R package (https://cran.r-project.org/), by comparing the inferred gene phylogenies to their species phylogenies. Tree support was measured using average bipartition support (ABS) from 5,000 ultrafast bootstrap approximations in IQ-TREE (Hoang et al., 2018). To determine if alignment trimming strategies resulted in substantially different alignment lengths, nRF values, and ABS values, we conducted principal component analysis using the R packages FactoMineR, v2.3 (Lê et al., 2008), and factoextra, v.1.0.6 (Kassambara and Mundt, 2017). All plots were made with FactoMineR, factoextra, and ggplot2, v2.3.1 (Wickham, 2009), in the R, 3.6.2 (https://cran.r-project.org/), programming environment.

To summarize nRF and ABS values into a single summary metric, we used desirability functions. Desirability functions rescale a distribution of values to be between zero and one depending on whether or not low values (e.g., nRF) or high values (ABS) are best. More specifically, these transformations were conducted using the following approach:

for nRF values:

$$desirability_{low} = \begin{cases} 0 \\ \dfrac{Y - A}{B - A} \\ 1 \end{cases} \middle| \begin{matrix} Y > B \\ A \leq Y \leq B \\ Y < A \end{matrix}$$

where *Y* is the variable value, *A* is the maximum nRF value, and *B* is the minimum nRF value;

for ABS values

$$desirability_{high} = \begin{cases} 0 \\ \dfrac{Y - A}{B - A} \\ 1 \end{cases} \middle| \begin{matrix} Y < A \\ A \leq Y \leq B \\ Y > B \end{matrix}$$

where *Y* is the variable value, *A* is the minimum ABS value, and *B* is the maximum ABS value. These transformations were conducted for the 14 different trimming strategies on a per gene basis. The resulting values were used to rank the relative performance of the 14 trimming strategies.

To examine the accuracy of branch lengths among single-gene trees, Spearman rank correlations of branch lengths were calculated between untrimmed (control) and trimmed (treatment) simulated multiple sequence alignments. To do so, the topologies of the untrimmed and trimmed phylogenies must be identical. Therefore, branch lengths were inferred along phylogenies that were constrained to match the reference tree topology using IQ-TREE (Nguyen et al., 2015). This analysis was only done for simulated sequences because high confidence in alignment

quality and true tree topology is required. Spearman rank correlations were conducted using the ggpubr, v.0.2.5 (Kassambara, 2020), package in the R, 3.6.2 (https://cran.r-project.org/), programming environment.

For species-level phylogenetic inferences, we used concatenated alignments of trimmed MSAs as input to IQ-TREE (Nguyen et al., 2015). Species-level phylogenetic inferences were also examined when using the quartet-based approach implemented in ASTRAL, v5.7.3 (Mirarab and Warnow, 2015), in which single-gene trees were used as input. Lastly, support among single-gene trees for references topologies was assessed using the information theory-based measure tree certainty (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016), which is implemented in RAxML, v8.2.10 (Stamatakis, 2014a).

Software availability

ClipKIT is available from GitHub, https://github.com/JLSteenwyk/ClipKIT, and PyPi, https://pypi.org/project/clipkit. Complete ClipKIT documentation is available online (https://jlsteenwyk.com/ClipKIT/).

Data availability

All alignments and phylogenies inferred in this study will be available from figshare (doi: 10.6084/m9.figshare.12401618) upon publication. The following link is provided for review purposes only https://figshare.com/s/bd07b70b510bca3155b9.

Results

To test the efficacy of ClipKIT, we examined the accuracy and support of single-gene and species-level phylogenetic trees inferred from untrimmed MSAs and MSAs trimmed using 14 different strategies (Table 1 from Steenwyk et al., 2020b) across four empirical genome-scale datasets and four simulated datasets. The four empirical datasets correspond to the untrimmed amino acid and nucleotide MSAs from 24 mammals ($N_{alignments}$=4,004) and 12 budding yeasts ($N_{alignments}$=5,664) (Shen et al., 2016a). The four simulated datasets ($N_{alignments}$=50 alignments per dataset or 200 total) stem from simulated nucleotide sequence evolution along the species phylogeny of 93 filamentous fungi (Steenwyk et al., 2019c), and from simulated amino acid sequence evolution along the species phylogenies of 70 metazoans (Whelan et al., 2015), 46 flowering plants (Xi et al., 2014), and 96 budding yeasts (Shen et al., 2016b). MSAs were trimmed using popular alignment trimming software (Table 1 from Steenwyk et al., 2020b) generating a total of 138,152 MSAs [(4,004 mammalian + 5,664 yeast + 200 simulated MSAs) * (14 trimming strategies, including a "no trimming" strategy) = 138,152 MSAs]. However, Gblocks and BMGE with an entropy threshold of 0.3 were not used for performance assessment of simulated datasets because they frequently removed entire MSAs.

We found that the 14 strategies examined occupied distinct regions of feature space suggestive of substantial differences between MSAs (Figure 50). Variation in feature space was largely driven by normalized Robinson Foulds (nRF) and average bipartition support (ABS) measures along the first dimension and alignment length along the second dimension for both empirical

and simulated datasets (Figure S1 from Steenwyk et al., 2020b). In empirical datasets, we found that some ClipKIT strategies removed few sites while others removed many and, at times, the most sites (Supplementary figure 2 from Steenwyk et al., 2020b). Among simulated datasets, ClipKIT trimmed substantial portions of MSAs but variation was observed across MSAs and datasets (Supplementary figure 3 from Steenwyk et al., 2020b). Examination of nRF and ABS values revealed ClipKIT performed well, and at times the best, among the MSA-trimming strategies tested, suggesting that phylogenetic inferences made with ClipKIT-trimmed MSAs were both accurate and well supported (Supplementary figure 4 from Steenwyk et al., 2020b and 5 from Steenwyk et al., 2020b). Finally, counter to previous evidence suggestive of a trade-off between trimming and phylogenetic accuracy (Tan et al., 2015), we found that ClipKIT aggressively trimmed MSAs in the empirical datasets without compromising phylogenetic tree accuracy and support (Supplementary figure S2 from Steenwyk et al., 2020b and S4 from Steenwyk et al., 2020b).

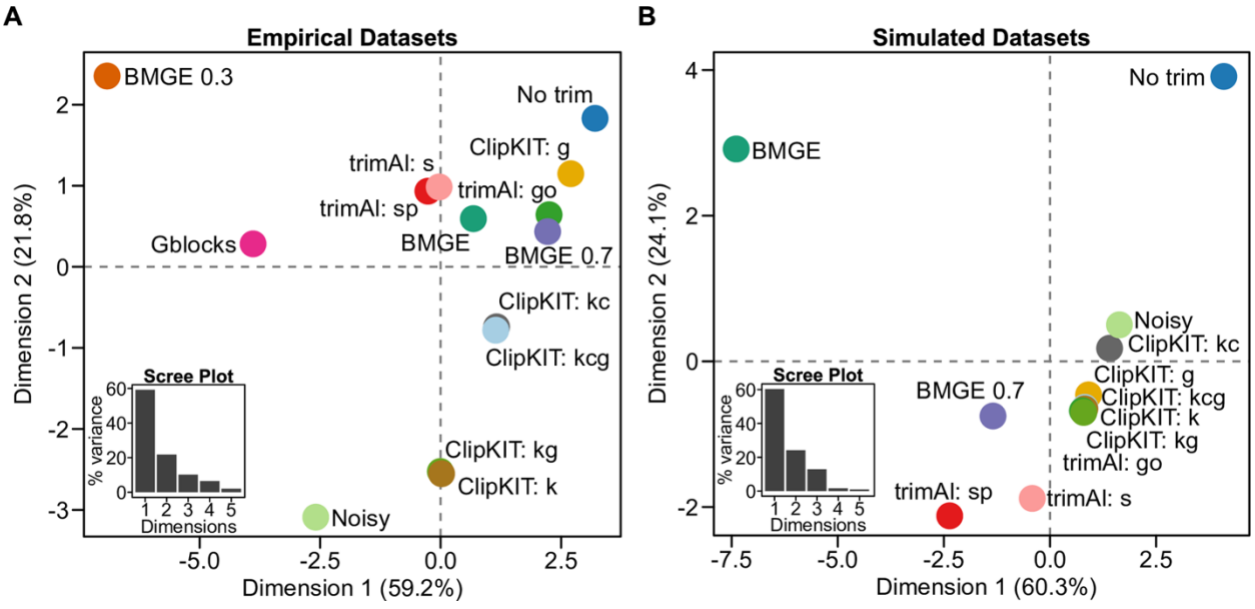

Fig 50. The 14 alignment trimming strategies tested differ in resulting MSAs and metrics of phylogenetic tree accuracy and support.
Principal component analysis of alignment length, nRF, and ABS values across the 14 MSA trimming strategies for 4 empirical datasets (A) and 4 simulated datasets (B). Insets of scree plots depict the percentage of variation explained (y-axis) for the first 5 dimensions (x-axis). Data were scaled prior to conducting principal component analysis. Note that the BMGE 0.3 and Gblocks strategies are not represented in Fig 50B because they frequently removed entire alignments and were therefore removed from the analysis of simulated sequenced. Data used to generate this figure can be found on figshare (doi: 10.6084/m9.figshare.12401618). ABS, average bipartition support; BMGE, Block Mapping and Gathering with Entropy; MSA, multiple sequence alignment; nRF, normalized Robinson–Foulds.

To obtain a summary of overall performance, we ranked the 14 strategies' performance for each dataset using objective desirability-based integration of nRF and ABS values (Eidem et al., 2018) (Figure 51). We found that the five ClipKIT strategies outperformed all others for amino acid sequences in the empirical mammalian dataset (Figure 51A) as well as in the simulated metazoan and flowering plant datasets (Figure 51E and F). Other strategies that performed well included trimAl with the 'gappyout' parameter for empirical datasets and Noisy for simulated datasets (Dress et al., 2008; Capella-Gutierrez et al., 2009). To evaluate MSA trimming strategy performance for empirical and simulated datasets, we examined average ranks across each set of four datasets and found that ClipKIT strategies were among the best performing (Figure 51 I-J). In empirical datasets, ClipKIT's gappy strategy outperformed all others followed by no trimming, trimAl with the 'gappyout' parameter, and then four other ClipKIT strategies (Figure 51I). In simulated datasets, all strategies generally performed better than in empirical datasets; the no trimming strategy ranked best followed by all five ClipKIT strategies (Figure 51J). These results suggest that ClipKIT, which focuses on retaining parsimony-informative sites, was on par with no trimming and frequently outperformed strategies that focus on removing highly divergent sites.

To examine the accuracy of branch lengths across MSA trimming strategies, we conducted correlation analysis between individual branch lengths in gene trees inferred from trimmed

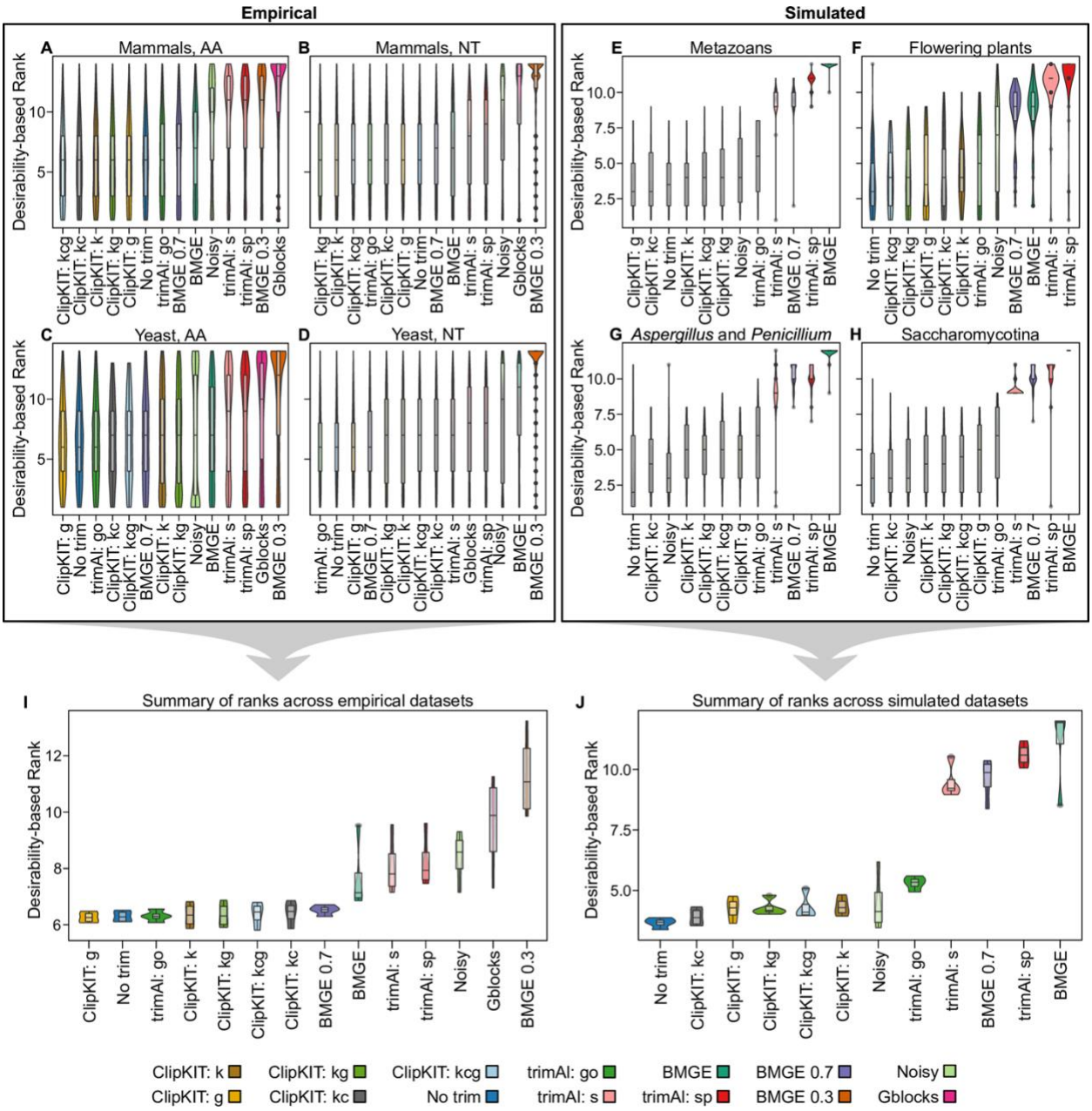

Fig 51. ClipKIT is a top-performing software for trimming MSAs.
Desirability-based integration of accuracy and support metrics per MSA facilitated the comparison of relative performance of the 14 different MSA trimming strategies for empirical (A–D) and simulated (E–H) datasets. Examination of performance for individual datasets and

average performance across empirical (I) and simulated (J) datasets revealed that ClipKIT is a top-performing software. MSA trimming strategies are ordered along the x-axis from the highest-performing strategy to the lowest-performing one according to average desirability–based rank. Boxplots embedded in violin plots have upper, middle, and lower hinges that represent the first, second, and third quartiles. Whiskers extend to 1.5 times the interquartile range. Data used to generate this figure can be found on figshare (doi: 10.6084/m9.figshare.12401618). AA, amino acid; BMGE, Block Mapping and Gathering with Entropy; MSA, multiple sequence alignment; NT, nucleotide.

---

alignments (treatment) and those inferred from untrimmed alignments (control). Because this analysis requires that untrimmed alignments are highly accurate, we conducted it only for individual simulated gene alignments. Notably, in our experimental set up, branch length estimates using trimmed alignments cannot be 'more' accurate than untrimmed alignments. Thus, an alignment trimming algorithm that does not negatively influence branch length estimates will have a Spearman rank correlation coefficient of 1.0. Examination of Spearman rank correlation coefficients revealed that branch lengths of trimmed alignments were typically very highly correlated with the branch lengths of untrimmed alignments (Supplementary figures 6-9 from Steenwyk et al., 2020b); ClipKIT strategies had correlation coefficients of 1.0 suggesting branch lengths inferred using ClipKIT trimmed alignments are accurate.

To evaluate the performance of the 14 strategies for species-level phylogenetic inference, we conducted concatenation- and quartet-based phylogenetic inference using IQ-TREE and ASTRAL, v5.7.3 (Mirarab and Warnow, 2015), respectively. We found that all strategies resulted in nearly identical and well supported phylogenies (Supplementary figures 10-12 from Steenwyk et al., 2020b). We also calculated tree certainty, an information theory-based measure of tree incongruence, which was used to summarize the degree of agreement with the reference topology across gene trees. The output from this analysis is a single value ranging from zero to

one where low values reflect high levels of incongruence among gene trees in the reference topology and high values reflect low levels of incongruence with the reference topology among gene trees (Salichos and Rokas, 2013). Tree certainty values were typically high and similar across all trimming strategies except for a few instances where certain strategies, which do not include ClipKIT strategies, significantly underperformed compared to all the others (Supplementary figure 13 from Steenwyk et al., 2020b). Among simulated datasets, we found that ClipKIT strategies reduced computation time by an average of ~20% compared to no trimming.

Discussion

Current state-of-the-art MSA trimming strategies focus on the removal of highly divergent sites. Highly divergent sites are thought to lack phylogenetic signal either because they represent sites that have become mutationally saturated due to the occurrence of multiple substitutions or because they are the result of inaccurate inference of homology (Lake, 1991). A previous analysis suggested that MSA trimming strategies often decreased the accuracy of phylogenetic inference (Tan et al., 2015), highlighting the need for new strategies.

To address this need, we developed ClipKIT, an alignment trimming software that focuses on identifying and retaining parsimony-informative sites. Examination of the accuracy and support of phylogenetic inferences revealed that ClipKIT strategies consistently and frequently outperformed other MSA trimming strategies and were on par with no trimming. These results suggest that MSA-trimming strategies focused on retaining phylogenetically-informative sites, such as parsimony-informative sites, hold promise for developing more accurate MSA trimming

strategies. We anticipate ClipKIT will be useful for phylogenomic inference and the quest to build the tree of life.

CHAPTER 13

orthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees[12]

Introduction

Molecular evolution studies, such as species tree inference, genome-wide surveys of positive selection, evolutionary rate estimation, measures of gene-gene coevolution, and others typically rely on single-copy orthologs (SC-OGs), a group of homologous genes that originated via speciation and are present in single-copy among species of interest (Rokas et al., 2003; Jeffares et al., 2015b; Li et al., 2017; Wu et al., 2017; Dong et al., 2019; Steenwyk et al., 2021e). In contrast, paralogs, homologous genes that originated via duplication and are often members of large gene families, are typically absent from these studies (Fig 52). Gene families of orthologs and paralogs often encode functionally significant proteins such as transcription factors, transporters, and olfactory receptors (Ozcan and Johnston, 1999; Malnic et al., 2004; Wingender et al., 2013; Niimura et al., 2014). The exclusion of SC-OGs from gene families has not only hindered our understanding of their evolution and phylogenetic informativeness but is also artificially reducing the number of gene markers available for molecular evolution studies. Furthermore, as the number of species and / or their evolutionary divergence increases in a data set, the number of SC-OGs decreases (Emms and Kelly, 2018; Thomas et al., 2020); case in

[12]This work is published in: Steenwyk, J. L., Goltz, D. C., Buida, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2021). orthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. bioRxiv, 2021.10.30.466607. doi:10.1101/2021.10.30.466607.

point, no SC-OGs were identified in a dataset of 42 plants (Emms and Kelly, 2018). As the number of available genomes across the tree of life continues to increase, our ability to identify SC-OGs present in many taxa will become more challenging.

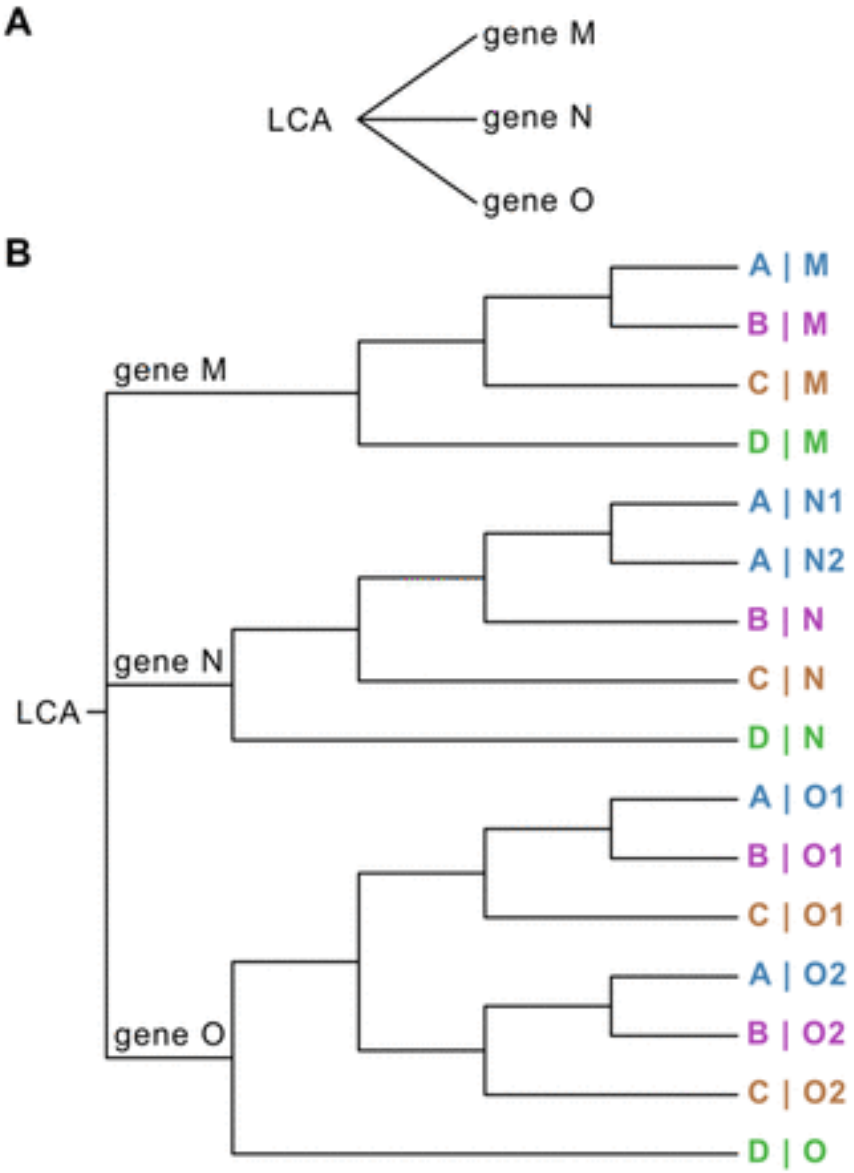

Fig. 52. Cartoon depiction of three classes of paralogs: outparalogs, inparalogs, and coorthologs. (A) Paralogs refer to related genes that have originated via gene duplication, such as genes M, N, and O. (B) Outparalogs and inparalogs refer to paralogs that are related to one another via a duplication event that took place prior to or after a speciation event, respectively. With respect to the speciation event that led to the split of taxa A, B, and C from D, genes M, N, and O are outparalogs because they arose prior to the speciation event; genes O1 and O2 in taxa A, B, and C are inparalogs because they arose after the speciation event. Species-specific inparalogs are paralogous genes observed only in one taxon in a dataset, such as gene N1 and N2 in taxon A.

Species-specific inparalogs N1 and N2 in taxon A are also coorthologs of gene N in taxa B, C, and D; the same is true for inparalogs O1 and O2 from taxon A, which are coorthologs of gene O from taxon D.

In light of these issues, several methods have been developed to account for paralogs in specific types of molecular evolution studies—for example, in species tree reconstruction (Smith and Hahn, 2021). Methods such as SpeciesRax, STAG, ASTRAL-PRO, and DISCO can be used to infer a species tree from a set of SC-OGs and gene families composed of orthologs and paralogs (Emms and Kelly, 2018; Zhang et al., 2020; Morel et al., 2021; Willson et al., 2021). Other methods such as PHYLDOG (Boussau et al., 2013) and guenomu (de Oliveira Martins and Posada, 2017) jointly infer the species and gene trees, but require abundant computational resources, which has hindered their use for large datasets. Other software, such as PhyloTreePruner (Kocot et al., 2013), can conduct species-specific inparalog trimming, whereas Agalma (Dunn et al., 2013), as part of a larger automated phylogenomic workflow, can prune gene trees into maximally inclusive subtrees wherein each taxon is represented by one sequence. Although these methods have expanded the numbers of gene markers used in species tree reconstruction, they were not designed to facilitate the retrieval of as broad a set of SC-OGs as possible for downstream molecular evolution studies such as surveys of positive selection. Furthermore, the phylogenetic information content of these gene families remains unknown.

To address this need, we developed orthoSNAP, a novel tree traversal algorithm that conducts tree splitting and species-specific inparalog pruning to identify SC-OGs nested within larger gene families. We term SC-OGs identified by orthoSNAP as SNAP-OGs because they were retrieved using a splitting and pruning procedure. orthoSNAP takes as input a gene family

phylogeny and associated FASTA file and will output individual FASTA files populated with sequences from SNAP-OGs (Fig 2). During tree traversal, tree uncertainty can be accounted for by orthoSNAP via collapsing poorly supported branches. In a set of four eukaryotic datasets that

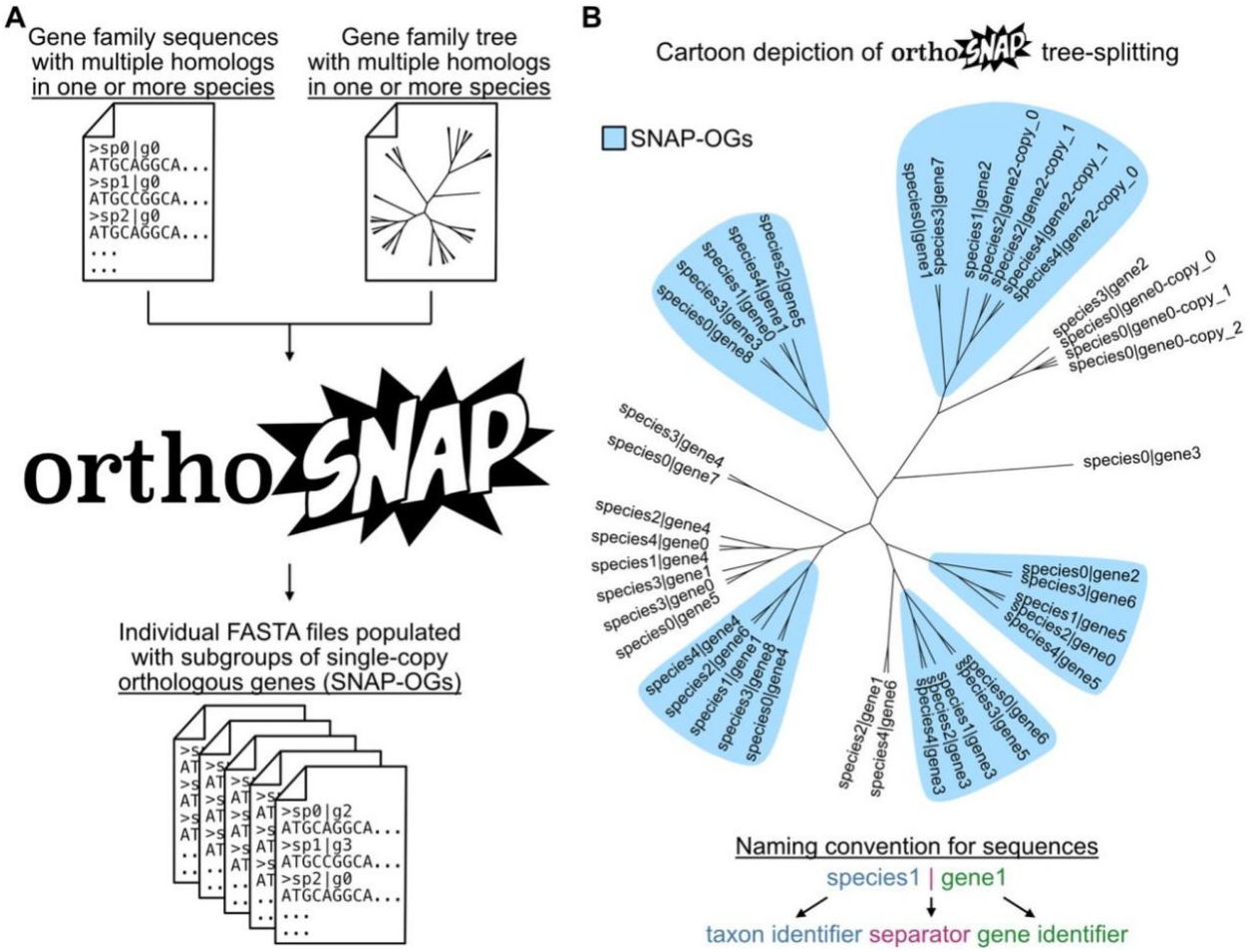

Fig. 53. Cartoon depiction of orthoSNAP workflow.
(A) orthoSNAP takes as input two files: a FASTA file of a gene family with multiple homologs observed in one or more species and the associated gene family tree. The outputted file(s) will be individual FASTA files of SNAP-OGs. (B) A cartoon phylogenetic tree that depicts the evolutionary history of a gene family and five SNAP-OGs therein. While identifying SNAP-OGs, orthoSNAP also identifies and prunes species-specific inparalogs (e.g., species2|gene2-copy_0 and species2|gene2-copy_1), retaining only the inparalog with the longest sequence, a practice common in transcriptomics. Note, orthoSNAP requires that sequence naming schemes must be the same in both sequences and follow the convention in which a taxon identifier and gene identifier are separated by pipe (or vertical bar; "|") character.

contained 6,634 SC-OGs, we used orthoSNAP to identify an additional 6,630 SNAP-OGs. Using a combination of multivariate statistics and phylogenetic measures, we demonstrate that SNAP-

OGs and SC-OGs have similar phylogenetic information content in all four datasets. We also observed similar patterns of support among SNAP-OGs and SC-OGs in a contentious branch in the tree of life. Taken together, these results suggest that orthoSNAP is helpful for expanding the set of gene markers available for molecular evolutionary studies.

Materials and Methods

orthoSNAP availability and documentation

orthoSNAP is a command-line software written in the Python programming language (https://www.python.org/) and requires Biopython (Cock et al., 2009a) and NumPy (Van Der Walt et al., 2011). orthoSNAP is available under the MIT license from GitHub (https://github.com/JLSteenwyk/orthosnap), PyPi (https://pypi.org/project/orthosnap), and the Anaconda cloud (https://anaconda.org/JLSteenwyk/orthosnap). Documentation describes the orthoSNAP algorithm, parameters, and provides user tutorials (https://jlsteenwyk.com/orthosnap).

orthoSNAP algorithm description and usage

We next describe how orthoSNAP identifies SNAP-OGs. orthoSNAP requires two files as input: one is a FASTA file that contains two or more homologous sequences in one or more species and the other the corresponding gene family phylogeny in Newick format. In both the FASTA and Newick file, users must follow a naming scheme—wherein taxon identifiers and gene sequences identifiers are separated by a vertical bar (also known as a pipe character or "|")—which allows orthoSNAP to determine which sequences were encoded in the genome of each taxon. After initiating orthoSNAP, the gene family phylogeny is first mid-point rooted and then SNAP-OGs

are identified using a tree-traversal algorithm. To do so, orthoSNAP will loop through the internal branches in the gene family phylogeny and evaluate the number of distinct taxa identifiers among children terminal branches. If the number of unique taxa identifiers is greater than or equal to the taxon occupancy threshold (default: 50% of total taxa in the inputted phylogeny; users can specify an integer threshold), then all children branches and termini are examined further; otherwise, orthoSNAP will examine the next internal branch. Next, orthoSNAP will collapse branches with low support (default: 80, which is motivated by using ultrafast bootstrap approximations (Hoang et al., 2018) to evaluate bipartition support; users can specify an integer threshold) and conduct species-specific inparalog trimming wherein the longest sequence is maintained, a practice common in transcriptomics. Species-specific inparalogs are defined as sequences from the same taxon that are sister to one another or belong to the same polytomy (Kocot et al., 2013). The resulting set of taxa and sequences are examined to determine if one taxon is represented by one sequence and ensure these sequences have not yet been assigned to a SNAP-OG. If so, they are considered a SNAP-OG; if not, orthoSNAP will examine the next internal branch.

The orthoSNAP algorithm is also described using the following pseudocode:

FOR internal branch in midpoint rooted gene family phylogeny:

> IF taxon occupancy among children termini is greater than or equal to taxon occupancy threshold;

>> Collapse poorly supported bipartitions and trim species-specific inparalogs;

>> IF each taxon among the trimmed set of taxa is represented by only one sequence and no sequences being examined have been assigned to a SNAP-OG yet;

>>> Sequences represent a SNAP-OG and are outputted to a FASTA file

>> ELSE

>>> examine next internal branch

> ELSE

>> examine next internal branch

ENDFOR

To enhance the user experience, arguments or default values are printed to the standard output, a progress bar informs the user of how of the analysis has been completed, and the number of SNAP-OGs identified as well as the names of the outputted FASTA files are printed to the standard output.

Development practices and design principles to ensure long-term software stability

Archival instabilities among software threatens the reproducibility of bioinformatics research (Mangul et al., 2019a). To ensure long-term stability of orthoSNAP, we implemented previously established rigorous development practices and design principles (Steenwyk et al., 2020b, 2021b, 2021a; Steenwyk and Rokas, 2021b). For example, orthoSNAP features a refactored codebase, which facilitates debugging, testing, and future development. We also implemented a continuous integration pipeline to automatically build, package, and install orthoSNAP across Python versions 3.8, 3.8, and 3.9. The continuous integration pipeline also conducts 29 unit and integration tests, which span 95.92% of the codebase and ensure faithful function of orthoSNAP.

Dataset generation

To generate a dataset for identifying SNAP-OGs and comparing them to SC-OGs, we first identified putative groups of orthologous genes across four empirical datasets. To do so, we first downloaded proteomes for each dataset, which were obtained from publicly available repositories on NCBI (S1 and S7 Fig from Steenwyk et al., 2021c; Table S1 from Steenwyk et al., 2021c and S7 from Steenwyk et al., 2021c) or figshare (Shen et al., 2018). Each dataset varied in its sampling of sequence diversity and in the evolutionary divergence of the sampled taxa. The dataset of 24 budding yeasts spans approximately 275 million years of evolution (Shen et al., 2018); the dataset of 36 filamentous fungi spans approximately 94 million years of evolution (Steenwyk et al., 2019c); the dataset of 26 mammals spans approximately 160 million years of evolution (Tarver et al., 2016); and the dataset of 28 placental mammals—which was used to study the contentious deep evolutionary relationships among placental mammals—concerns an ancient divergence that occurred approximately 160 million years ago (Luo et al., 2011). Putatively orthologous groups of genes were identified using OrthoFinder, v2.3.8 (Emms and Kelly, 2019), with default parameters, which resulted in 46,645 orthologous groups of genes with at least 50% taxon occupancy (Table S8 from Steenwyk et al., 2021c).

To infer the evolutionary history of each orthologous group of genes, we first individually aligned and trimmed each group of sequences using MAFFT, v7.402 (Katoh and Standley, 2013), with the "auto" parameter and ClipKIT, v1.1.3 (Steenwyk et al., 2020b), with the "smart-gap" parameter, respectively. Thereafter, we inferred the best-fitting substitution model using Bayesian information criterion and evolutionary history of each orthologous group of genes

using IQ-TREE2, v2.0.6 (Minh et al., 2020). Bipartition support was examined using 1,000 ultrafast bootstrap approximations (Hoang et al., 2018).

To identify SNAP-OGs, the FASTA file and associated phylogenetic tree for each gene family with multiple homologs in one or more species was used as input for orthoSNAP, v0.0.1 (this study). Across 40,011 gene families with multiple homologs in one or more species in all datasets, we identified 6,630 SNAP-OGs with at least 50% taxon occupancy (S2 Fig from Steenwyk et al., 2021c; Table S8 from Steenwyk et al., 2021c). Unaligned sequences of SNAP-OGs were then individually aligned and trimmed using the same strategy as described above. To determine gene families that were SC-OGs, we identified orthologous groups of genes with at least 50% taxon occupancy and each taxon was represented by only one sequence—6,634 orthologous groups of genes were SC-OGs.

Measuring and comparing information content among SC-OGs and SNAP-OGs

To compare the information content of SC-OGs and SNAP-OGs, we calculated nine properties of multiple sequence alignments and phylogenetic trees associated with robust phylogenetic signal in the budding yeasts, filamentous fungi, and mammalian datasets (Table S4 from Steenwyk et al., 2021c). More specifically, we calculated information content from phylogenetic trees such as measures of tree certainty (average bootstrap support), accuracy (Robinson-Foulds distance (Robinson and Foulds, 1981)), signal-to-noise ratios (treeness (Phillips and Penny, 2003)), and violation of clock-like evolution (degree of violation of a molecular clock or DVMC (Liu et al., 2017)). Information content was also measured among multiple sequence alignments by examining alignment length and the number of parsimony-informative sites, which are

associated with robust and accurate inferences of evolutionary histories (Shen et al., 2016) as well as biases in sequence composition (RCV (Phillips and Penny, 2003)). Lastly, information content was also evaluated using metrics that consider characteristics of phylogenetic trees and multiple sequence alignments such as the degree of saturation, which refers to multiple substitutions in multiple sequence alignments that underestimate the distance between two taxa (Philippe et al., 2011), and treeness / RCV, a measure of signal-to-noise ratios in phylogenetic trees and sequence composition biases (Phillips and Penny, 2003). For tree accuracy, phylogenetic trees were compared to species trees reported in previous studies (Tarver et al., 2016; Shen et al., 2018; Steenwyk et al., 2019c). All properties were calculated using functions in PhyKIT, v1.1.2 (Steenwyk et al., 2021b). The function used to calculate each metric and additional information are described in Table S4 from Steenwyk et al., 2021c.

Principal component analysis across the nine properties that summarize phylogenetic information content was used to qualitatively compare SC-OGs and SNAP-OGs in reduced dimensional space. Principal component analysis, visualization, and determination of property contribution to each principal component was conducted using factoextra, v1.0.7 (Kassambara and Mundt, 2017), and FactoMineR, v2.4 (Lê et al., 2008), in the R, v4.0.2 (https://cran.r-project.org/), programming environment. Statistical analysis using a multi-factor ANOVA was used to quantitatively compare SC-OGs and SNAP-OGs using the res.aov() function in R.

Information theory-based approaches were used to evaluate incongruence among SC-OGs and SNAP-OGs phylogenetic trees. More specifically, we calculated tree certainty and tree certainty-all (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016), which are conceptually

similar to entropy values and are derived from examining support among a set of gene trees and the two most supported topologies or all topologies that occur with a frequency of ≥5%, respectively. More simply, tree certainty values range from 0 to 1 in which low values are indicative of low congruence among gene trees and high values are indicative of high congruence among gene trees. Tree certainty and tree certainty-all values were calculated using RAxML, v8.2.10 (Stamatakis, 2014a).

To examine patterns of support in a contentious branch concerning deep evolutionary relationships among placental mammals, we calculated gene support frequencies and ΔGLS. Gene support frequencies were calculated using the "polytomy_test" function in PhyKIT, v1.1.2 (Steenwyk et al., 2021b). To account for uncertainty in gene tree topology, we also examined patterns of gene support frequencies after collapsing bipartitions with less than 75 ultrafast bootstrap approximation support using the "collapse" function in PhyKIT. To calculate ΔGLS, partition log-likelihoods were calculated using the "wpl" parameter in IQ-TREE2 (Minh et al., 2020), which required as input a phylogeny in Newick format that represented either hypothesis one or hypothesis two (Fig 54A) and a concatenated alignment of SC-OGs and SNAP-OGs with partition information. Thereafter, gene-wise log-likelihood scores associated with hypothesis two were subtracted from gene-wise log-likelihood scores associated with hypothesis one. The resulting score is ΔGLS wherein values greater than zero support hypothesis one and values less than zero support hypothesis two.

Data Availability

Results and data are available from figshare (doi: 10.6084/m9.figshare.16875904).

Results

SC-OGs and SNAP-OGs have similar information content

To compare SC-OGs and SNAP-OGs, we first independently inferred orthologous groups of genes in three eukaryotic datasets of 24 budding yeasts, 36 filamentous fungi, and 26 mammals (S1 Fig from Steenwyk et al., 2021c; Table S1 from Steenwyk et al., 2021c). There was variation in the number of SC-OGs and SNAP-OGs in each lineage (S2 Fig from Steenwyk et al., 2021c; Table S2 from Steenwyk et al., 2021c). Interestingly, the ratio of SNAP-OGs : SC-OGs among budding yeasts, filamentous fungi, and mammals was 0.46, 0.83, and 5.53, respectively, indicating SNAP-OGs can substantially increase the number of gene markers in certain lineages. The number of SNAP-OGs identified in a gene family with multiple homologs in one or more species also varied (S3 Fig from Steenwyk et al., 2021c).

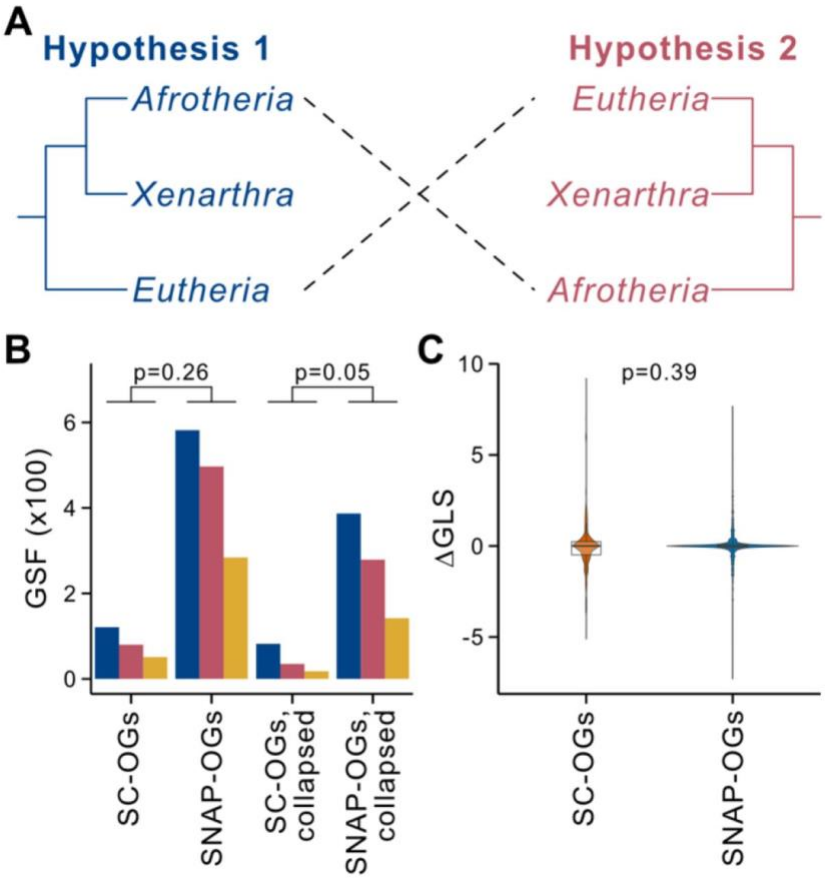

Fig. 54. SC-OGs and SNAP-OGs display similar patterns of support in a contentious branch concerning deep evolutionary relationships among placental mammals.
(A) Two leading hypotheses for the evolutionary relationships among Eutheria, which have implications for the evolution and biogeography of the clade, are that Afrotheria and Xenarthra are sister to all other Eutheria (hypothesis one; blue) and that Afrotheria are sister to all other Eutheria (hypothesis two; pink). (B) Comparison of gene support frequency (GSF) values for hypotheses one, hypothesis two, as well as a third hypothesis (Xenarthra as sister to all other Eutheria represented in yellow) among 252 SC-OGs and 1,428 SNAP-OGs using an α level of 0.01 revealed no differences in support ($p = 0.26$, Fischer's exact test with Benjamini-Hochberg multi-test correction). Comparison after accounting for gene tree uncertainty by collapsing bipartitions with lower than 75 ultrafast bootstrap approximation support (SC-OGs collapsed vs. SNAP-OGs collapsed) also revealed no differences ($p = 0.05$; Fischer's exact test with Benjamini-Hochberg multi-test correction). (C) Examination of the distribution of gene-wise log-likelihood scores (ΔGLS) revealed no difference between SNAP-OGs and SC-OGs ($p = 0.39$; Wilcoxon rank sum test). ΔGLS values greater than zero are indicative of genes with greater support for hypothesis one; values less than zero are indicative of genes with greater support for hypothesis two.

Similar taxon occupancy and best fitting models of substitutions were observed among SC-OGs and SNAP-OGs (S4 Fig from Steenwyk et al., 2021c; Table S3 from Steenwyk et al., 2021c), raising the question of whether SC-OGs and SNAP-OGs have similar information content. To answer this, we calculated nine properties of phylogenetic information content from multiple sequence alignments and phylogenetic trees from SC-OGs and SNAP-OGs (S5 Fig from Steenwyk et al., 2021c; Table S4 from Steenwyk et al., 2021c) and compared them using multivariate analysis and statistics as well as information theory-based phylogenetic measures. Principal component analysis enabled qualitative comparisons between SC-OGs and SNAP-OGs in reduced dimensional space and revealed a striking similarity (Fig 55, S6 Fig from Steenwyk et al., 2021c). Multivariate statistics, namely multi-factor analysis of variance, facilitated a quantitative comparison of SC-OGs and SNAP-OGs and revealed no difference between SC-OGs and SNAP-OGs ($p = 0.63$, $F = 0.23$, $df = 1$; Table S5 from Steenwyk et al., 2021c) and no

interaction between the nine properties and SC-OGs and SNAP-OGs (p = 0.16, F = 1.46, df = 8).

Similarly, multi-factor analysis of variance using an additive model, which assumes each factor

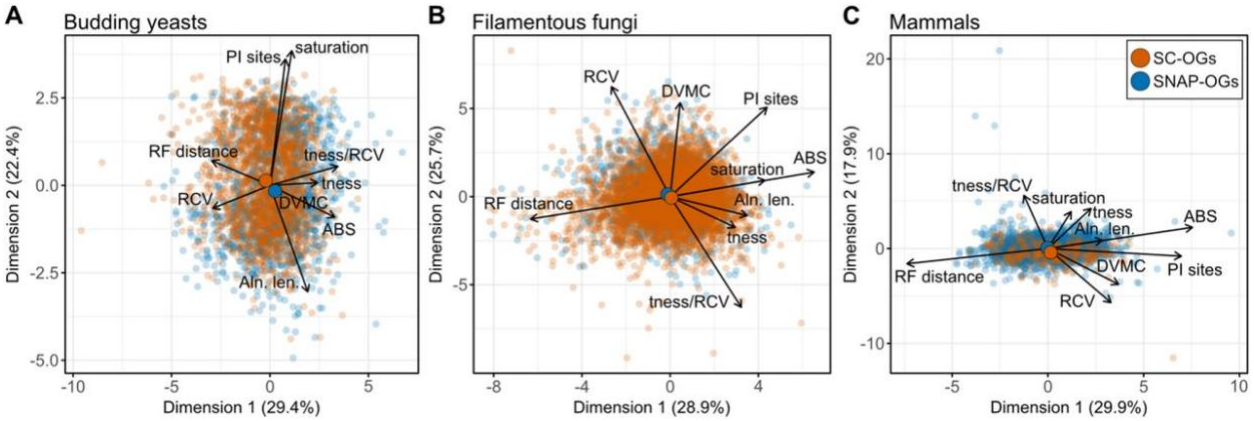

Fig. 55. SC-OGs and SNAP-OGs have similar phylogenetic information content.
To evaluate similarities and differences between SC-OGs (orange dots) and SNAP-OGs (blue dots), we examined each gene's phylogenetic information content by measuring nine properties of multiple-sequence alignments and phylogenetic trees. We performed these analyses on 12,764 gene families from three datasets—24 budding yeasts (1,668 SC-OGs and 1,392 SNAP-OGs) (A), 36 filamentous fungi (4,393 SC-OGs and 2,035 SNAP-OGs) (B), and 26 mammals (321 SC-OGs and 1,775 SNAP-OGs) (C). Principal component analysis revealed striking similarities between SC-OGs and SNAP-OGs in all three datasets. For example, the centroid (i.e., the mean across all metrics and genes) for SC-OGs and SNAP-OGs, which is depicted as an opaque and larger dot, are very close to one another in reduced dimensional space. Supporting this observation, multi-factor analysis of variance with interaction effects of the 6,630 SNAP-OGs and 6,634 SC-OGs revealed no difference between SC-OGs and SNAP-OGs (p = 0.63, F = 0.23, df = 1) and no interaction between the nine properties and SC-OGs and SNAP-OGs (p = 0.16, F = 1.46, df = 8). Multi-factor analysis of variance using an additive model yielded similar results wherein SC-OGs and SNAP-OGs do not differ (p = 0.65, F = 0.21, df = 1). There are also very few outliers of individual SC-OGs and SNAP-OGs, which are represented as translucent dots, in all three panels. For example, SNAP-OGs outliers at the top of panel C are driven by high treeness/RCV values, which is associated with a high signal-to-noise ratio and/or low composition bias (Phillips and Penny, 2003); SNAP-OG outliers at the right of panel C are driven by high average bootstrap support values, which is associated with greater tree certainty (Salichos and Rokas, 2013); and the single SC-OG outlier observed in the bottom right of panel C is driven by a SC-OG with a high degree of violation of a molecular clock (Song et al., 2012), which is associated with lower tree certainty (Doyle et al., 2015). Multiple-sequence alignment and phylogenetic tree properties used in principal component analysis and abbreviations thereof are as follows: average bootstrap support (ABS), degree of violation of the molecular clock (DVMC), relative composition variability, Robinson-Foulds distance (RF distance), alignment length (Aln. len.), the number of parsimony informative sites (PI sites), saturation, treeness (tness), and treeness/RCV (tness/RCV).

is independent and there are no interactions, also revealed no differences between SC-OGs and SNAP-OGs ($p = 0.65$, $F = 0.21$, $df = 1$). Next, we calculated tree certainty, an information theory-based measure of tree congruence from a set of gene trees, and found similar levels of congruence among phylogenetic trees inferred from SC-OGs and SNAP-OGs (Table S6 from Steenwyk et al., 2021c). Taken together, these analyses demonstrate that SC-OGs and SNAP-OGs have similar phylogenetic information content.

SC-OGs and SNAP-OGs have similar patterns of support in a contentious branch in the tree of life

To further compare SC-OGs and SNAP-OGs, we investigated patterns of support in a difficult-to-resolve branch in the tree of life. More specifically, we evaluated the support between two leading hypotheses concerning deep evolutionary relationships among placental mammals: (1) Xenarthra (placental mammals from the Americas) and Afrotheria (placental mammals from Africa) are sister to all other Eutheria (Hallström et al., 2007; Wildman et al., 2007) or (2) Afrotheria are sister to all other Eutheria (Murphy, 2001; Murphy et al., 2001) (Fig 54A). Resolution of this conflict has important implications for understanding the historical biogeography of these organisms. To do so, we first obtained protein-coding gene sequences from six Afrotheria, two Xenarthra, 12 other Eutheria, and eight outgroup taxa from NCBI (S7 Fig from Steenwyk et al., 2021c; Table S7 from Steenwyk et al., 2021c), which represent all annotated and publicly genome assemblies at the time of this study (Table S8 from Steenwyk et al., 2021c). Using the protein translations of these gene sequences as input to OrthoFinder, we identified 252 SC-OGs shared across taxa; application of orthoSNAP identified an additional 1,428 SNAP-OGs, which represents a greater than five-fold increase in the number of gene

markers for this dataset (Table S8 from Steenwyk et al., 2021c). There was variation in the number of SNAP-OGs identified per orthologous group of genes (S8 Fig from Steenwyk et al., 2021c). The highest number of SNAP-OGs identified in an orthologous group of genes was 10, which was a gene family of olfactory receptors and are known to have expanded in the evolutionary history of placental mammals (Niimura et al., 2014). The best fitting substitution models were similar between SC-OGs and SNAP-OGs (S9 Fig from Steenwyk et al., 2021c).

Two independent tests examining support between alternative hypotheses of deep evolutionary relationships among placental mammals revealed similar patterns of support between SC-OGs and SNAP-OGs. More specifically, no differences were observed in gene support frequencies—the number of genes that support one of three possible hypotheses at a given branch in a phylogeny—without or with accounting for single-gene tree uncertainty by collapsing branches with low support values ($p = 0.26$ and $p = 0.05$, respectively; Fischer's exact test with Benjamini-Hochberg multi-test correction; Fig 54B; Table S9 from Steenwyk et al., 2021c). We next conducted a second test of single-gene support for hypothesis one or hypothesis two by measuring gene-wise log-likelihood scores (ΔGLS), which is the difference in the log-likelihood score of a single gene when constrained to the topologies of the two hypotheses. In this case, positive ΔGLS are reflective of greater support for hypothesis one and negative ΔGLS are reflective of greater support for hypothesis two. No difference was observed in the distribution of ΔGLS values ($p = 0.39$; Wilcoxon rank sum test). Examination of patterns of support in a contentious branch in the tree of life using two independent tests revealed SC-OGs and SNAP-OGs are similar and further supports the observation that they contain similar phylogenetic information.

In summary, 46,645 orthologous groups of genes across four datasets contained 6,634 SC-OGs; application of orthoSNAP identified an additional 6,630 SNAP-OGs, doubling the number of gene markers. Comprehensive comparison of the phylogenetic information content among SC-OGs and SNAP-OGs revealed no differences in phylogenetic information content. Strikingly, this observation held true when conducting hypothesis testing in a difficult-to-resolve branch in the tree of life. These findings suggest that SNAP-OGs may be useful for exploring patterns of molecular evolution ranging from genome-wide surveys of positive selection, phylogenomics, gene-gene coevolution analysis, and others.

Discussion

Molecular evolution studies typically rely on SC-OGs. Recently developed methods can integrate gene families of orthologs and paralogs into species tree inference but are not designed to broadly facilitate the retrieval of gene markers for molecular evolution analyses. Furthermore, the phylogenetic information content of gene families of orthologs and paralogs remains unknown. This observation underscores the need for algorithms that can identify SC-OGs nested within larger gene families, which can be in turn be incorporated into diverse molecular evolution analyses, and a comprehensive assessment of their phylogenetic properties.

To address this need, we developed orthoSNAP, a tree splitting and pruning algorithm that identifies SNAP-OGs, which refers to SC-OGs nested within larger gene families wherein species specific inparalogs have also been pruned. Comprehensive examination of the phylogenetic information content of SNAP-OGs and SC-OGs from four empirical datasets of

diverse eukaryotic species revealed striking similarities. In certain datasets, SNAP-OGs were five times more prevalent than SC-OGs indicating SNAP-OGs can substantially increase the size of molecular evolution datasets. We note that our results are qualitatively similar to those reported recently by Smith et al. (Smith et al., 2021), which retrieved SC-OGs nested within larger families from 26 primates and examined their performance in gene tree and species tree inference. Three noteworthy differences are that we also conduct species-specific inparalog trimming, provide a user-friendly command-line software for SNAP-OG identification, and evaluated the phylogenetic information content of SNAP-OGs and SC-OGs across four diverse datasets. We also note that our algorithm can account for diverse types of paralogy—outparalogs, inparalogs, and species-specific inparalogs—whereas other software like PhyloTreePruner, which conducts species-specific inparalog trimming (Kocot et al., 2013), and Agalma, which identifies single-copy outparalogs and inparalogs (Dunn et al., 2013), can account for some, but not all, types of paralogs. Our results, together with other studies, demonstrate the utility of SC-OGs that are nested within larger families (van der Heijden et al., 2007; Dunn et al., 2013; Smith et al., 2021; Willson et al., 2021).

Next, we discuss some practical considerations when using orthoSNAP. In the present study, we inferred orthology information using OrthoFinder (Emms and Kelly, 2019), but several other approaches can be used upstream of orthoSNAP. For example, other graph-based algorithms such as OrthoMCL (Li et al., 2003) or sequence similarity-based algorithms such as orthofisher (Steenwyk and Rokas, 2021b), can be used to infer gene families. Similarly, sequence similarity search algorithms like BLAST+ (Camacho et al., 2009), USEARCH (Edgar, 2010), and

HMMER (Eddy, 2011), can be used to retrieve homologous sets of sequences that are used as input for orthoSNAP.

We suggest employing “best practices” when inferring groups of putatively orthologous genes, including SNAP-OGs. Specifically, orthology information can be further scrutinized using phylogenetic methods. Orthology inference errors may occur upstream of orthoSNAP; for example, SNAP-OGs may be susceptible to erroneous inference of orthology during upstream clustering of putatively orthologous genes. One method to identify putatively spurious orthology inference is by identifying long terminal branches (Shen et al., 2018). Terminal branches of outlier length can be identified using the “spurious_sequence” function in PhyKIT (Steenwyk et al., 2021b). Other tools, such as PhyloFisher, UPhO, and other orthology inference pipelines employ similar strategies to refine orthology inference (Yang and Smith, 2014; Ballesteros and Hormiga, 2016; Tice et al., 2021).

Taken together, we suggest that orthoSNAP is useful for retrieving single-copy orthologous groups of genes from gene family data and that the identified SNAP-OGs have similar phylogenetic information content compared to SC-OGs. In combination with other phylogenomic toolkits, orthoSNAP may be helpful for reconstructing the tree of life and expanding our understanding of the tempo and mode of evolution therein.

CHAPTER 14

orthofisher: a broadly applicable tool for automated gene identification and retrieval[13]

Introduction

Sequence similarity searches of genomic data are commonly employed in diverse fields of biology. Several pieces of software have been designed to infer statistically homologous sequences from databases of sequence data, such as BLAST, DIAMOND, and HMMER (Camacho et al., 2009; Eddy, 2011; Madden, 2013; Buchfink et al., 2015). One frequent use of sequence similarity search methods is for the identification of orthologs, sequences present in the common ancestor of two species, and homologs, sequences that stem from the same common ancestral sequence (Gabaldón and Koonin, 2013). For example, the OrthoFinder software conducts BLAST all-vs-all searches across proteomes to infer groups of putatively orthologous genes (Emms and Kelly, 2019). Similarly, the BUSCO software aims to identify putatively orthologous genes using a predetermined set of profile Hidden Markov Model sequence alignments (pHMMs) derived from single-copy orthologous proteins from the OrthoDB database (Waterhouse et al., 2013, 2018a).

The results of these or similar pieces of software can facilitate diverse downstream analyses (Remm et al., 2001; Li et al., 2003; Train et al., 2017; Waterhouse et al., 2018a; Emms and Kelly, 2019). However, global analyses, such as those conducted by OrthoFinder, are

[13]This work is published in: Steenwyk, J. L., and Rokas, A. (2021). orthofisher: a broadly applicable tool for automated gene identification and retrieval. G3 Genes|Genomes|Genetics 11. doi:10.1093/g3journal/jkab250.

computationally expensive and may be beyond the scope of a research project (e.g., studies focused on a few genes). Similarly, software that rely on databases, such as BUSCO, are constrained to the orthologs therein. As a result, there is a need for bioinformatic software that can conduct automated identification and retrieval of putative homologs and orthologs across sequence databases using user-specified query sequences and output files that facilitate downstream analyses.

We introduce orthofisher, a command-line toolkit for automated identification of highly similar sequences across proteomes using custom pHMMs. orthofisher facilitates downstream analyses by creating multi-FASTA files populated with highly similar sequences identified during pHMM searches. Default parameters are designed to identify sequences with the highest sequence similarity (i.e., putative orthologous genes), but users can customize its use to best fit their research question (e.g., relaxed thresholds can be used to obtain all putatively homologous genes; similarly, searches in databases that contain gene isoforms can be used to retrieve all isoforms of a particular gene). We demonstrate the efficacy of orthofisher by evaluating the precision and recall for identification of sequences with high similarity to query pHMMs in a multiple sequence FASTA (multi-FASTA) files from animals, plants, and fungi. Comparison of orthofisher, BUSCO, and OrthoFinder revealed similar performance in identification of sequences with high sequence similarity. Thus, orthofisher aims to streamline gene identification and retrieval from genomic data, which is the first step of many bioinformatic analyses and projects. We anticipate orthofisher will be of interest to diverse fields of computational biology and to biologists and bioinformaticians.

orthofisher requires two files as input (Figure 56). One file—specified with the -m, --hmm argument—provides the paths to query pHMMs that will be used during sequence similarity search; the other file—specified with the -f, --fasta argument—provides the paths to FASTA files that will be used as the sequence search database. orthofisher then loops through each FASTA file and uses each pHMM to search for similar sequences using HMMER3 (Eddy, 2011) with an expectation-value threshold of 0.001 (which can be modified with the -e, --evalue argument). orthofisher then parses the resulting HMMER3 output using biopython (Cock et al., 2009a) and identifies top hits. Top hits are defined following criteria used in the BUSCO pipeline (Waterhouse et al., 2018a) wherein all sequences with scores that are greater than or equal to 85% of the score of the best hit are maintained. Users can modify this threshold using the -b, --bitscore argument. Top hits are considered homologous genes.

orthofisher outputs three directories and two text files that enable researchers to easily evaluate results from sequence similarity search and facilitate downstream analyses. The three directories are

- *hmmsearch_output*: HMMER3 output files,
- *all_sequences*: one multi-FASTA file per pHMM, which are populated with homologous sequences identified during the sequence similarity search step, and
- *scog*: one multi-FASTA file per pHMM, which are populated with only those homologous sequences that are present at most only once in each genome.

The two text files are

• *short_summary.txt*: the number and percentage of sequences present in single-copy, multi-copy, or absent sequences per pHMM search, and

• *long_summary.txt*: the homologous sequences identified during pHMM search for every query and sequence database.

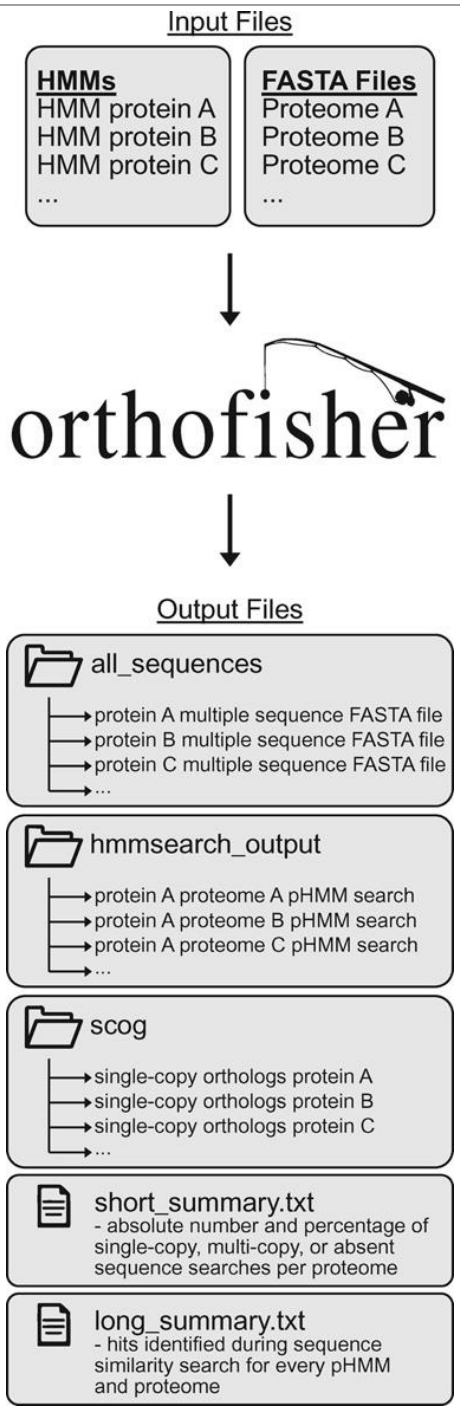

Fig. 56. Workflow overview for orthofisher.
orthofisher takes two files as input, which specify the location of query pHMMs and the FASTA files wherein sequence similarity searches will be performed. orthofisher then outputs three directories and two text files that summarize results and facilitate downstream analyses. Contents of output files will be heavily dependent on user parameters, the pHMMs used, and the input files. For example, transcriptomic data may require additional processing steps such as collapsing isoforms into a single representative sequence per gene. The intent of orthofisher—which is to identify single-copy orthologous genes—is flexible enough to capture paralogous sequences as well. A tutorial for how to use orthofisher is publicly available as part of the online documentation https://jlsteenwyk.com/orthofisher/tutorial.

Nearly 30% of bioinformatic tools fail to install (Mangul et al., 2019b), which poses a nontrivial problem for the reproducibility of computational experiments. To remedy this issue, we implemented state-of-the-art standards of software development practices and design principles (Darriba et al., 2018) following previously established protocol (Steenwyk et al., 2020b, 2021b). For example, whenever changes to code are made, faithful function of orthofisher is tested using a continuous integration pipeline, a process that automatically builds, packages, and tests installation and function using Python versions 3.6, 3.7, and 3.8. We also wrote several unit and integration tests that span 95% of the orthofisher code.

orthofisher comes complete with comprehensive documentation (https://jlsteenwyk.com/orthofisher/), is freely available under the MIT license, and is available for download from GitHub (https://github.com/JLSteenwyk/orthofisher), PyPi (https://pypi.org/project/orthofisher/), and the Anaconda Cloud (https://anaconda.org/jlsteenwyk/orthofisher). The proteomes, pHMMs, and outputs of orthofisher, BUSCO, and OrthoFinder are available through figshare (doi: 10.6084/m9.figshare.14399150).

Results

To determine the similarities and differences between orthofisher and other algorithms that identify putative orthologs, we compared results obtained from orthofisher with that of BUSCO and OrthoFinder. BUSCO and OrthoFinder are both widely adopted methods of identifying orthologous genes across multiple proteomes. As noted in the introduction, each software differs – more specifically, BUSCO conducts homology searches using a predefined set of pHMMs and OrthoFinder conducts proteome-wide analysis to identify groups of orthologous genes. Thus, we expect that if orthofisher can identify putative orthologs across proteomes, it will identify the same genes BUSCO identifies during its sequence similarity search. Given that both algorithms conduct pHMM-based searches, we anticipate that both will exhibit near identical performances. When comparing orthofisher and BUSCO to OrthoFinder, we anticipate the sequences identified during sequence similarity search by orthofisher and BUSCO will be in the same orthologous group of genes inferred by OrthoFinder.

orthofisher and BUSCO obtain similar results

To evaluate the efficacy of orthofisher, we compared results obtained from orthofisher to those obtained from BUSCO, v4.0.4 (Waterhouse et al., 2018a). To do so, both algorithms were used to identify 255 near-universally single-copy orthologous genes obtained from the Eukaryota OrthoDB, v10 (Waterhouse et al., 2013), database across the proteomes of animals (*Homo sapiens*: GCF_000001405.39; *Mus musculus*: GCF_000001635.27), plants (*Arabidopsis thaliana*, NCBI accession: GCA_000001735.2; *Solanum lycopersicum*: GCF_000188115.4), and

fungi (*Saccharomyces cerevisiae*, NCBI accession: GCA_000146045.2; *Candida albicans*: GCA_000182965.3). Measures of precision and recall were calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where *TP* represents true positives, *FP* represents false positives, and *FN* represents false negatives of single-copy orthologous genes. Precision and recall values range from 0 to 1 and higher values reflect better performance.

Near perfect values of precision and recall (0.98 or [231 / [231 + 4]] and 1.0 or [231 / [231 + 0]], respectively) reveal orthofisher is able to automate the identification and retrieval of sequences with high similarity to the query pHMM. A low false positive rate of 0.02 was observed. The difference in the performance of BUSCO and orthofisher stems from an additional set of gene-specific score and length thresholds used by the BUSCO software, which are not implemented in orthofisher. These results demonstrate that orthofisher can accurately identify homologous genes.

To demonstrate the importance of using a score threshold of 85% of the score observed in the best hit following the BUSCO pipeline (Waterhouse et al., 2018a), we highlight an example where absence of a score threshold would have led to identification of additional putatively orthologous genes. A HMMER search using the query BUSCO pHMM 1001705at2759 and a e-value threshold of 1e-10 in the proteome of *A. thaliana* reports the gene as multi-copy whereas both orthofisher and BUSCO report this gene to be single-copy. More specifically, when using only an e-value threshold of 1e-10, the following nine genes are reported: AEE76455.1,

AEE78573.1, AEC10322.1, ANM68500.1, AED93406.1, AEE76521.1, AEE82221.1, AED98328.1, and AEE29324.1; however, AEE76455.1 has a score of 242.5 and the next best hit, AEE78573.1, has a score of 64.5. Thus, a score threshold of 85% of the best hit (in this case 242.5*0.85) is helpful during sequence similarity searches.

orthofisher and BUSCO perform similarly to OrthoFinder

Comparison of the results of BUSCO and orthofisher to OrthoFinder, a global (or whole proteome) ortholog calling algorithm revealed BUSCO, orthofisher, and OrthoFinder produce similar results. To perform these comparisons, we first used OrthoFinder, v2.3.8 (Emms and Kelly, 2019), to identify putative orthologous groups of genes in the same animal, plant, and fungal proteomes described above using an inflation parameter of 1.5 and DIAMOND, v0.9.24.125 (Buchfink et al., 2015). Then, we determined if genes identified as multi-copy are part of the same or different orthologous group(s) of genes and also assessed if genes identified as single-copy in BUSCO or orthofisher were also single-copy in OrthoFinder.

Among multi-copy genes, we found BUSCO and OrthoFinder had nearly identical performance in the proteomes of *A. thaliana*, *S. lycopersicum*, and *C. albicans*. For *S. cerevisiae*, one gene, 1545004at2759, out of 255 differed between BUSCO and OrthoFinder wherein BUSCO identified two homologs and OrthoFinder split these two genes into different orthologous groups of genes. A similar scenario was observed among 12 / 255 and 3 / 255 genes in the human and mouse proteomes, respectively. For orthofisher, a similar scenario was observed for 1 / 255 genes in *S. lycopersicum*; 1 / 255 genes in *A. thaliana*; 8 / 255 genes in *S. cerevisiae*; 4 / 255 genes in *C. albicans*; 13 / 255 genes in the human proteome; and 4 / 255 genes in the mouse

proteome. We note that isoforms of the same gene sequence were present in the analysed proteomes and were accounted for in these analyses.

Among single-copy genes, we observed a few instances where single-copy genes in BUSCO were multi-copy in OrthoFinder. More specifically, this was observed for 8 genes in *S. lycopersicum*; 16 genes in *A. thaliana*; 2 genes in *S. cerevisiae*; 2 genes in *C. albicans*; 36 genes in the human proteome; and 26 genes in the mouse proteome. Similar results were observed for orthofisher. More specifically, 16 / 255 genes in *A. thaliana* were identified as single-copy by orthofisher but were in multi-copy orthologous groups of genes in OrthoFinder. The same observation was made for 7 / 255 genes in *S. lycopersicum*; 1 / 255 gene in *S. cerevisiae*; 2 / 255 genes in *C. albicans*; 35 / 255 genes in the human proteome; and 24 / 255 genes in the mouse proteome.

In summary, sequence similarity searches of 255 genes in 6 proteomes identified differences among 105 genes (6.86%; 105 / 1,530) between BUSCO and OrthoFinder; similarly, we identified differences among 116 genes (7.58%; 116 / 1,530) between orthofisher and OrthoFinder. These differences likely stem from differences in the approach of each algorithm to identify putative orthologs. Specifically, OrthoFinder uses DIAMOND and Markov clustering to identify orthologous groups, BUSCO uses pHMM-based search and gene-specific score and length thresholds using OrthoDB, and orthofisher uses pHMM-based similarity search thresholds. Also, these differences are in part driven by each algorithm reporting different results (i.e., OrthoFinder reports groups of putatively orthologous genes and BUSCO and orthofisher report putative orthologous genes).

orthofisher is helpful for estimating the number of members in a gene family

To demonstrate how to use orthofisher to estimate the number of gene family members, we estimate the number of DNA photolyase (PFam: PF00875) and zinc finger, C2H2 type (PFam: PF00096) homologs in *S. cerevisiae*, *C. albicans*, two species from the *Hanseniaspora* genus (*H. uvarum* NRRL Y1614 and *H. vineae* NRRL Y17529, both of which are known to lack DNA photolyases (Steenwyk et al., 2019a)), and three *Aspergillus* species (*A. niger* CBS 513.88, *A. fumigatus* Af293, and *A. flavus* NRRL 3357). When estimating gene family number, we recommend lowering the score threshold to, for example, 25% of the best hit, which we have done here. In line with previous reports, we found that *Hanseniaspora* species lacked DNA photolyases whereas *S. cerevisiae*, *C. albicans*, and all *Aspergillus* species had one or two DNA photolyases. In contrast, proteins with Zinc finger domains are more abundant across all species with copies ranging from 16 (*H. vineae*) to 39 (*A. flavus*).

Discussion

The intended use of orthofisher is to help identify orthologous genes across species using accurate and sensitive pHMM-based searches. We encourage users to evaluate results produced by orthofisher using additional approaches (e.g., phylogenetic inference) to infer precise relationships of orthology and paralogy among sequences. We note that orthofisher is not explicitly designed to identify a single-representative sequence if multiple isoforms encoded by one gene sequence are present in a proteome. Thus, we also suggest users collapse isoforms prior to or after orthofisher analysis following standard protocol in many transcriptomics studies.

In summary, orthofisher is a command-line tool for automated identification and retrieval of genes of interest from genomic data. We anticipate orthofisher will be useful for evaluating genome completeness, performing phylogenomic inferences, estimating gene family size, and other analyses that rely on identification and retrieval of homologous genes from genomic data.

CHAPTER 15

treehouse: a user-friendly application to obtain subtrees from large phylogenies[14]

Introduction

Evolutionary biology relies on understanding the phylogenetic relationships among sets of genes, traits, and organisms under investigation. However, large phylogenies that contain hundreds of taxa are increasingly becoming inaccessible to researchers interested in the relationships of just a few representatives. For example, some phylogenies are so large that taxon information is often challenging or impossible to visualize and is often excluded (Hug et al., 2016; Peter et al., 2018; Shen et al., 2018; Varga et al., 2019); similarly, the lengths of many internal branches are often very short and the constraints of displaying a large tree in a letter-sized page make the tracing of relationships among a subset of taxa challenging and unnecessarily time-consuming. These issues will increase in frequency as the numbers of taxa included in phylogenies of genes, metagenomes, genomes, etc. continues to rapidly rise.

To address these issues, we introduce *treehouse*, a user-friendly application with minimal dependencies that facilitates the retrieval of subtrees from any user-specified set of taxa in a given phylogeny. Our simple three-step workflow allows users to obtain subtrees from a curated and growing database of large-scale phylogenetic trees from across the tree of life. Additionally, users may obtain subtrees from their own phylogenies which, can facilitate data exploration and

[14]This work is published in: Steenwyk, J. L., and Rokas, A. (2019). Treehouse: a user-friendly application to obtain subtrees from large phylogenies. BMC Res. Notes 12, 541. doi:10.1186/s13104-019-4577-5.

inter-disciplinary collaboration. For easy integration into pre-existing project workflows, subtrees obtained from *treehouse* can be easily be downloaded as a newick file or PDF file that retains branch length information. *Treehouse* enables beginner and expert evolutionary biologists alike to reap the benefits of large-scale phylogenetic projects and use them to test evolutionary-based hypotheses.

Materials and Methods

Data acquisition

The *treehouse* contains a database of 20 representative large phylogenies from across the tree of life (Table 1 from Steenwyk and Rokas, 2019).

Description of the software

Using *treehouse* requires the R packages PHYTOOLS, version 0.6-60 (Revell, 2012), and SHINY, version 1.2.0 (https://shiny.rstudio.com/). Dependencies of PHYTOOLS includes MAPS, version 3.3.0 (https://cran.r-project.org/web/packages/maps/index.html), and APE, version 5.3 (Paradis et al., 2004). To present the phylogeny as depicted by the original authors, phylogenies from *treehouse's* database are rooted. The taxa chosen to root the phylogeny on are inferred from figures presented in the original manuscript or, in the case of phylogenies presented without taxa names, personal communications with the authors. Phylogenies are rooted using PHYTOOLS'S root() function. Using the list of taxa provided by the user, *treehouse* determines the list of taxa to remove from the phylogeny using the setdiff() function. The resulting list is then used to remove taxa in the phylogeny using PHYTOOLS'S drop.tip() function. To write out the resulting phylogeny in a newick-formatted text file or display it in a scalable-vector-graphic-formatted pdf

file, we use the write.tree() and plot.phylo() functions in APE, respectively. To create a user-friendly and intuitive user-interface, we used SHINY.

Results

A three-step workflow to obtain subtrees

treehouse is designed to have a simple user-interface that guides a user through an intuitive three-step workflow (Figure 57A) and user interface (Figure 57B).

(1) Tree Selection

A user can choose between five tabs - userTree, Animals, Fungi, Plants, and Tree of Life - located at the top of the user interface (Figure 57Ba). When using phylogenies from the *treehouse* database, a user selects the desired phylogeny using a dropdown menu (Figure 57Bi; left). In userTree, a user selects a phylogeny in newick format from their local computer (Figure 57Bi; right).

(2) Selection of Taxa

A user next uploads a text file containing the single-column list of taxa that they want a subtree for (Figure 57Bii). Here, each taxon name must be identical to a taxon name in the full phylogeny.

(3) Subtree output

By clicking the 'Update' button, the user launches *treehouse* subtree retrieval. The subtree is plotted to the right of the side panel and buttons that allow the user to download a pdf or text

file of the subtree are below it (Figure 57Biii). Lastly, the full set of taxa in the currently uploaded *treehouse* phylogeny is listed (Figure 57Bc; left).

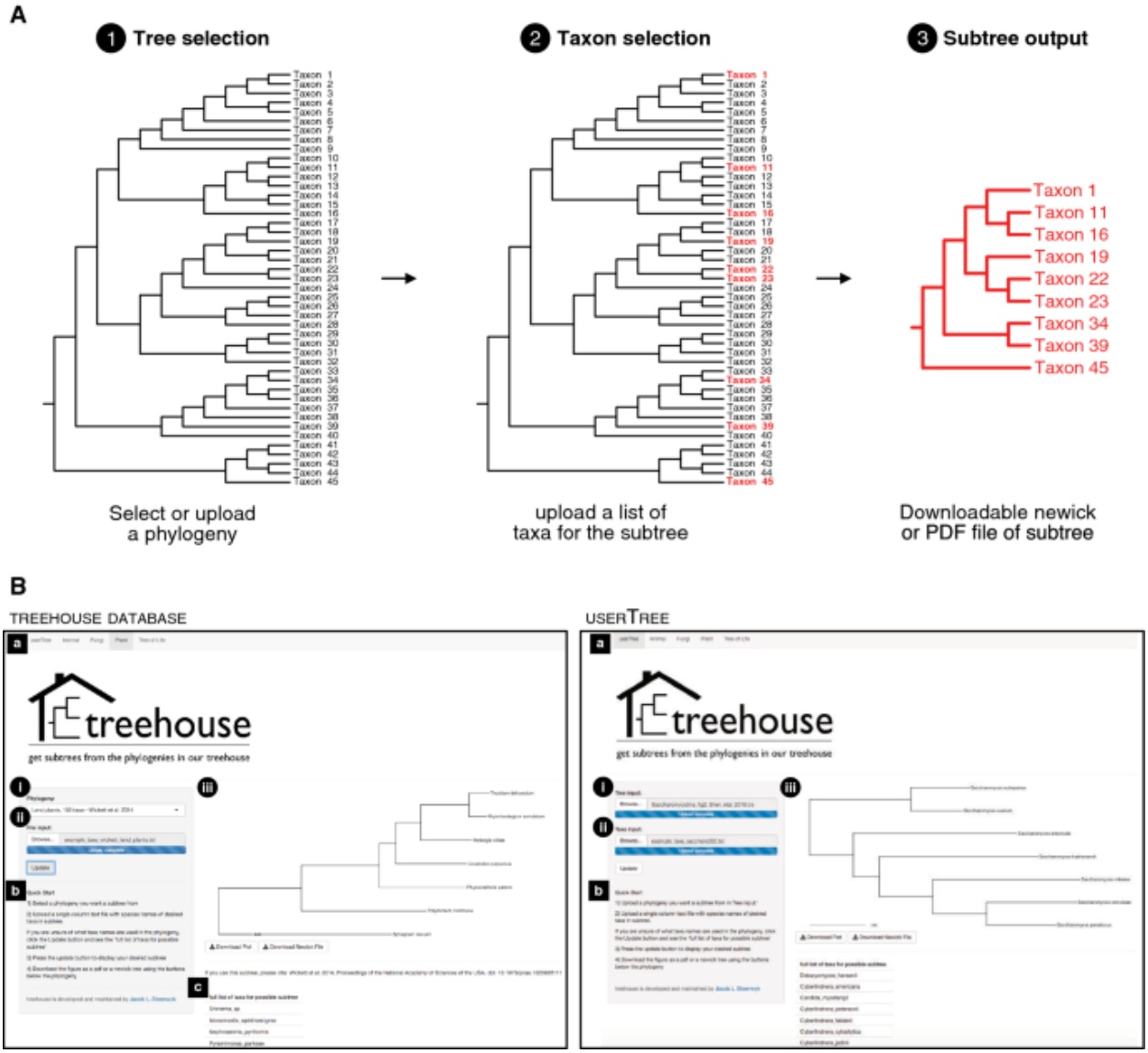

Fig. 57. A simple three-step workflow for using *treehouse*.
A Using *treehouse* requires three simple steps: (1) Tree selection: select a phylogeny from the *treehouse* database or a user-provided phylogeny that you want a subtree for; (2) Taxon selection: upload a list of taxa that a user wants to include in the subtree; and (3) Subtree output: download the newick-formatted text file or scalable-vector-graphic-formatted pdf file of the subtree. **B** *Treehouse's* user interface features a navigation bar (**a**) to toggle between phylogenies available in *treehouse's* databases for animals, fungi, plants, and the tree of life (left) and a user provided phylogeny in userTree (right). **b** To enable easy usage of *treehouse*, quick start directions are displayed. **i** A dropdown menu allows for selection of a larger phylogeny to obtain

a subtree from when using phylogenies in *treehouse's* database. When using userTree, a browser function allows a user to upload their own phylogeny. **ii** A browser function allows the user to upload a list of taxa for the desired subtree. **c** A list of all possible taxa in phylogeny is provided.

Discussion

treehouse is a simple and powerful tool that facilitates subtree retrieval from large phylogenies.

treehouse's functionality rests on the performance of one task, namely removing taxa from a phylogeny. To the experienced phylogenetic or phylogenomic researcher, this might seem to be a trivial task but is not so for most users of phylogenetic trees and no other user-friendly methods are available. Thus, we anticipate the 'typical' *treehouse* users to be researchers that use phylogenies to form hypotheses but do not routinely infer phylogenies themselves. We also anticipate *treehouse* to be a useful teaching tool.

CHAPTER 16

ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures[15]

Introduction

Scientific figures are graphical representations of scientific data (Rougier et al., 2014). Several tools have been developed to generate scientific figures in numerous computer programming languages, including seaborn (Waskom, 2021) and Matplotlib (Matplotlib Org., 2019) in the Python programming language, and lattice (Sarkar, 2017) and ggplot2 (Wickham, 2009) in the R programming language. These and other data visualization pieces of software have empowered researchers with the ability to generate scientific figures from diverse sources of quantitative data. As a result, methods and standards for effective scientific figures—which we define as accurate, clear, and precise representations of scientific data—have been a topic of rigorous debate that is in part influenced by field, audience, and data type (Lau and Vande Moere, 2007; Bertini et al., 2011; Moere and Purchase, 2011).

Although certain rules of effective scientific figures are context-dependent and subject to change, some rules are broadly applicable to several disciplines, including in microbiology and the life sciences. These include two rules from Rougier et al.'s article titled *Ten Simple Rules for Better Figures*: "Do Not Trust the Defaults" and "Use Color Effectively" (Rougier et al., 2014). For the

[15]This work is published in: Steenwyk, J. L., and Rokas, A. (2021). ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures. Microbiol. Resour. Announc. 10. doi:10.1128/MRA.00871-21.

first rule, the authors suggest that default plotting parameters (e.g., font size, ticks, etc.) are sufficient to make a scientific figure but insufficient to make the "best" scientific figure; for the second, the authors suggest that color is an important component of human vision and therefore is equally important when making scientific figures. Effective color use can also make scientific figures more accessible. For example, 8% and 0.4% of European Caucasian men and women, respectively, are red-green color deficient (Birch, 2012). Thus, effective figure making is also a matter of inclusion.

To facilitate generating effective figure making, we present ggpubfigs, an R package with colorblind friendly color palettes and ggplot2 extensions that facilitates the generation of publication-quality scientific figures for quantitative data (https://github.com/JLSteenwyk/ggpubfigs). More specifically, ggpubfigs contains six color palettes that are colorblind friendly and aim to increase accessibility of scientific figures and eight "themes," which modify 21 parameters of a default ggplot2 figure. To demonstrate how ggpubfigs can improve scientific figures in R, we compare default ggplot2 settings (Figure 58A) to those modified using extensions or colorblind friendly color palettes available in ggpubfigs (Figure 58B-G). Users can create additional modifications to a scientific figure according to their specific needs.

Materials and Methods

ggpubfigs is freely available under the MIT license and is available for download on GitHub (https://github.com/JLSteenwyk/ggpubfigs). The GitHub repository comes complete with installation instructions and tutorials. Installing ggpubfigs is simple and only requires executing

one command. Tutorials detail how to use color palettes for qualitative and continuous and discrete quantitative data as well as utilizing ggplot2 theme extensions.

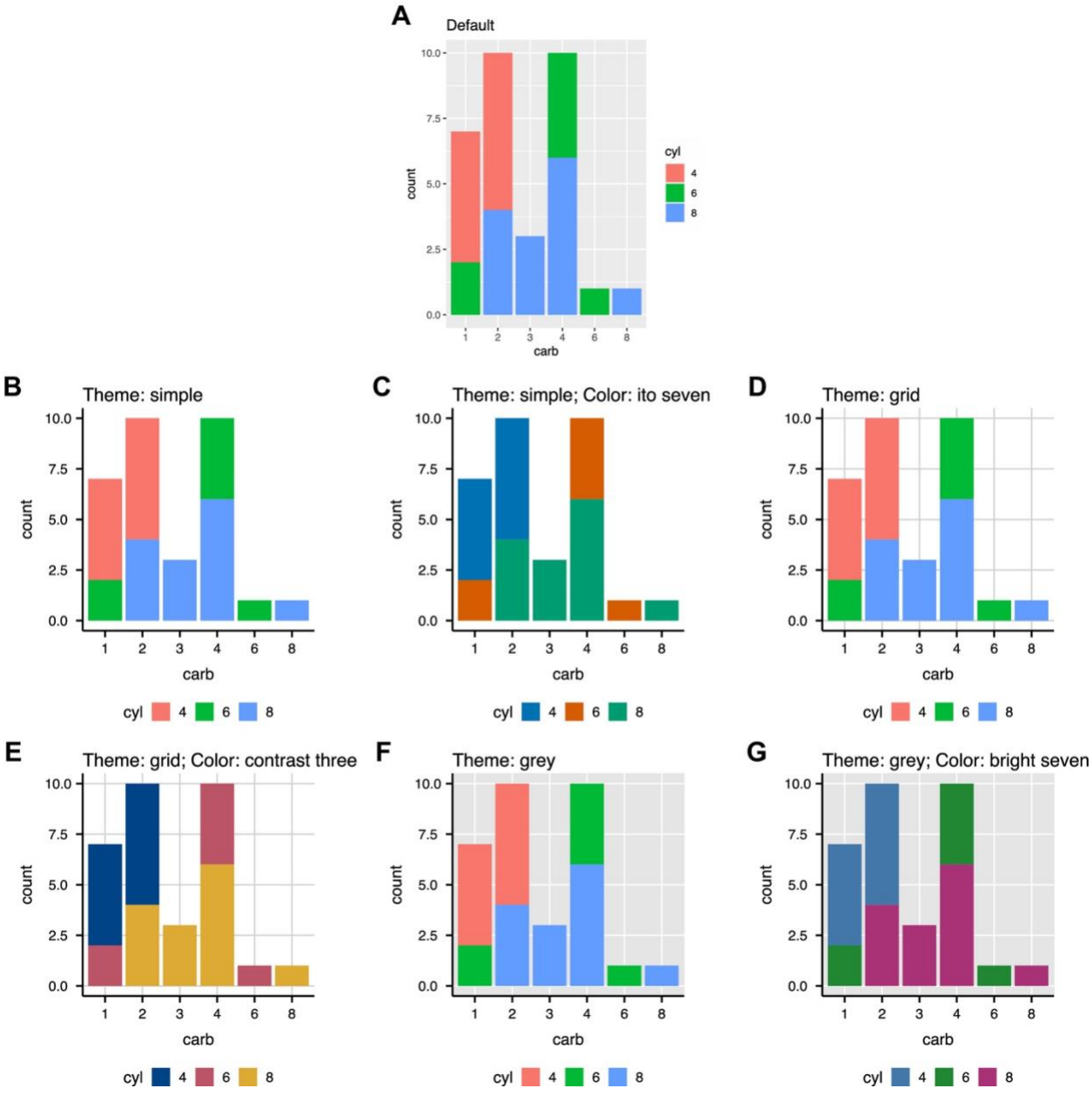

Fig. 58. Examples of ggplot2 extensions and color palettes available in ggpubfigs.
(A) Default scientific visualization made using ggplot2. (B to G) Modified scientific figures made using the simple theme (B), the simple theme with the ito seven color palette (C), the grid

theme (D), the grid theme with the contrast three color palette (E), the gray theme (F), and the gray theme with the bright seven color palette (G). Data are from the mtcars data frame available through the data sets package.

Results

Color palettes can be accessed using the “friendly_pal()” function. For example, friendly_pal("contrast_three") will provide users access to an object of class “palette” that contains the hex codes for the “contrast_three” color palette. Color palettes can be converted into a colormap of *N* colors—which may be useful for plotting data as a heatmap—using the following command: friendly_pal("contrast_three", *N*, type="continuous"). Themes that modify ggplot2 plots can be appended to the end of ggplot2 plotting command. For example, to use the “simple theme” in ggpubfigs, theme_simple(), an object of class gg theme, should be appended to the end of a ggplot2 plotting command.

Discussion

We anticipate ggpubfigs will assist researchers in generating effective scientific figures that are accessible to broad audiences including those that are colorblind.

CHAPTER 17

Concluding discussion and future directions

Here, I presented flagship research from my predoctoral work, which was focused on the evolutionary dynamics of fungi—in particular fungal pathogens and rapid evolutionary processes among fungi—and the development of new software for the life sciences with a focus on their application to evolutionary studies. This work describes novel findings and insights, but also raises exciting new questions. Thus, I hope that this work sets the stage for future research.

Among studies aiming to unravel the evolutionary history of fungal pathogens, our finding that numerous fungal nonpathogens hold many of the same "cards of pathogenicity" as pathogenic fungi (Steenwyk et al., 2020c; Mead et al., 2021) suggests there may be some metaphorical shortcomings in the "pathogenic hand of cards" (Casadevall, 2007). For example, as Dr. Matthew Mead mentioned during a Dr. Antonis Rokas laboratory meeting, when "cards" are played, it is not just the identity of the "card" that is important but also its quantity. In other words, when and how much genetic determinants of virulence are "utilized" is an important consideration when studying fungal pathogenesis. This notion may be further examined using RNA-seq, dual-seq, or proteomics in virulence-related conditions.

Studies evaluating rapid evolutionary processes among fungi have revealed novel research themes ripe for investigation. For example, how often and frequently eukaryotes lose DNA repair genes is currently unknown. Beyond punctuated sequence evolution, it is unclear what other aspects of genome evolution or life history are impacted by DNA repair gene loss, but we can speculate. For example, I hypothesize that chromosome rearrangement may occur at elevated

frequency in organisms that have lost numerous DNA damage response genes compared to close relatives that have not experienced these losses. I also hypothesize that the loss of DNA repair genes may contribute to the diversification of species. This idea is loosely supported by species boundaries typically being defined by genetic similarity (or lack thereof) (Jain et al., 2018); thus, as multiple lineage arise due to the increased frequency of *de novo* mutations, and as these lineages continue to diverge, one may find that the loss of a DNA repair gene can contribute to the emergence of a species-rich lineage.

Our finding that *A. latus* arose via allodiploid hybridization, straddles the previous two research themes—the evolution of fungal pathogenicity and rapid evolutionary processes in fungi. Thus, this system may serve as a focal point for diverse studies and equally diverse insights. Some future research directions include elucidating the precise impact of hybridization on pathogenicity, which has been demonstrated in other fungal species (Stukenbrock, 2016; Mixão and Gabaldón, 2018). Also, evolutionarily, it remains unknown how many hybridization events gave rise to *A. latus*. Filling these knowledge gaps will be greatly enabled by additional population-level sequencing efforts as well as identifying, sequencing, and phenotyping the unknown parent that is closely related to *A. quadrilineatus*. Beyond basic research, this study also has clinical implications that can be used to inform disease management strategies. For example, our finding that *A. latus* isolates were frequently misidentified as *A. nidulans* suggests the frequency of infections caused by *A. latus* is unknown. Incorporating genomics into species-level identification in the clinical setting (and other realms) will elucidate how frequently *A. latus* is causing disease. This information will be critical for shedding light on the epidemiology of *A. latus* and other pathogens.

More broadly, this finding underscores the importance of hybridization as a mechanism of evolution. This notion has been appreciated among other lineages, such as animals including the bears (Pongracz et al., 2017) and butterflies (Pardo-Diaz et al., 2012) and plants (Ellstrand et al., 1996). Except for a few cases, the precise impact of hybridization on organismal fitness is not always clear. As a result, beyond identifying hybrid lineages, a deeper understanding of the biological impact of hybridization is warranted.

The major insight gleaned from developing software—as I see it—is that current ecosystem for software engineers in the biosciences may not be long-term sustainable. Supporting this idea, approximately 30% of bioinformatic tools cannot be installed (Mangul et al., 2019b). This issue may stem from software no longer being maintained due to lack of support or they conduct analyses on file formats or data types that are no longer in use. Nonetheless, there are potential solutions to remedy this issue—for example, coursework at the junction of computer science and biology that enables biologists to learn industry-level standards of software development can alleviate some issues. Moreover, such a practice would represent a healthy cultural shift in placing greater importance on software development, a necessary component of keeping pace with data generation. Furthermore, the offerings of computer science have hardly been tapped by the biological sciences. Integrating more computer science courses into biological curriculum will help biologists realize their full potential.

In my humble attempt to contribute to this cultural shift, I have implemented numerous practices commonplace to most software engineers, but less frequently discussed among biologists. For

example, all software I develop implements rigorous testing protocols (i.e., unit and integration tests) to ensure faithful functionality of software. In combination with continuous integration pipelines, the software I have developed is tested across multiple versions of Python, further strengthening long-term stability, and ensuring faithful function. Although these practices took a lot of effort early in the project, it has proven to be very helpful during later stages. For example, extensive unit and integration tests have led to code quality improvements and faster debugging when handling.

In summary, my thesis—like those that came before mine and those that will come after—represents a small but mighty advance in the sciences. Empowered by a stellar network of collaborators, we have furthered our understanding and identified novel knowledge gaps that raise exciting new questions in the fields of fungal pathogenesis, evolution, and software development that warrant further research and effort. I hope that this thesis serves as a source of inspiration for future discoveries. Further, I hope the journey of completing future research in these disciplines and beyond enables scientists to discover what is most important: oneself and their relationship with others through the lens of teamwork, camaraderie, compassion, courage, hard work, humility, forgiveness and more.

References

Abad, A., Victoria Fernández-Molina, J., Bikandi, J., Ramírez, A., Margareto, J., Sendino, J., et al. (2010). What makes Aspergillus fumigatus a successful pathogen? Genes and molecules involved in invasive aspergillosis. *Rev. Iberoam. Micol.* 27, 155–182. doi:10.1016/j.riam.2010.10.003.

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. (2013). Hybridization and speciation. *J. Evol. Biol.* 26, 229–246. doi:10.1111/j.1420-9101.2012.02599.x.

Abdolrasouli, A., Rhodes, J., Beale, M. A., Hagen, F., Rogers, T. R., Chowdhary, A., et al. (2015a). Genomic Context of Azole Resistance Mutations in Aspergillus fumigatus Determined Using Whole-Genome Sequencing. *MBio* 6, e00536. doi:10.1128/mBio.00536-15.

Abdolrasouli, A., Rhodes, J., Beale, M. A., Hagen, F., Rogers, T. R., Chowdhary, A., et al. (2015b). Genomic Context of Azole Resistance Mutations in Aspergillus fumigatus Determined Using Whole-Genome Sequencing. *MBio* 6. doi:10.1128/mBio.00536-15.

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi:10.1101/gr.114876.110.

Afiyatullov, S. S., Kalinovskii, A. I., Pivkin, M. V., Dmitrenok, P. S., and Kuznetsova, T. A. (2005). Alkaloids from the Marine Isolate of the Fungus Aspergillus fumigatus. *Chem. Nat. Compd.* 41, 236–238. doi:10.1007/s10600-005-0122-y.

Alanio, A., Dellière, S., Fodil, S., Bretagne, S., and Mégarbane, B. (2020). Prevalence of putative invasive pulmonary aspergillosis in critically ill patients with COVID-19. *Lancet*

Respir. Med. 8, e48–e49. doi:10.1016/S2213-2600(20)30237-X.

Alastruey-Izquierdo, A., Alcazar-Fuoli, L., and Cuenca-Estrella, M. (2014). Antifungal Susceptibility Profile of Cryptic Species of Aspergillus. *Mycopathologia* 178, 427–433. doi:10.1007/s11046-014-9775-z.

Alastruey-Izquierdo, A., Mellado, E., Peláez, T., Pemán, J., Zapico, S., Alvarez, M., et al. (2013). Population-Based Survey of Filamentous Fungi and Antifungal Resistance in Spain (FILPOP Study). *Antimicrob. Agents Chemother.* 57, 3380–3387. doi:10.1128/AAC.00383-13.

Albalat, R., and Cañestro, C. (2016). Evolution by gene loss. *Nat. Rev. Genet.* 17, 379–391. doi:10.1038/nrg.2016.39.

Albert, A. W., Chen, J., Kuron, G., Hunt, V., Huff, J., Hoffman, C., et al. (1980). Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme A reductase and a cholesterol-lowering agent. *Proc. Natl. Acad. Sci. U. S. A.* 77, 3957–3961. doi:10.1073/pnas.77.7.3957.

Albertin, W., Setati, M. E., Miot-Sertier, C., Mostert, T. T., Colonna-Ceccaldi, B., Coulon, J., et al. (2016). Hanseniaspora uvarum from Winemaking Environments Show Spatial and Temporal Genetic Clustering. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.01569.

Almeida, A. J., Matute, D. R., Carmona, J. A., Martins, M., Torres, I., McEwen, J. G., et al. (2007). Genome size and ploidy of Paracoccidioides brasiliensis reveals a haploid DNA content: Flow cytometry and GP43 sequence analysis. *Fungal Genet. Biol.* doi:10.1016/j.fgb.2006.06.003.

Alshareef, F., and Robson, G. D. (2014). Genetic and virulence variation in an environmental population of the opportunistic pathogen Aspergillus fumigatus. *Microbiol. (United*

Kingdom). doi:10.1099/mic.0.072520-0.

Ames, R. M., Money, D., Ghatge, V. P., Whelan, S., and Lovell, S. C. (2012). Determining the evolutionary history of gene families. *Bioinformatics* 28, 48–55. doi:10.1093/bioinformatics/btr592.

Aminov, R. I. (2010). A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Front. Microbiol.* 1, 134. doi:10.3389/fmicb.2010.00134.

An integrated map of genetic variation from 1,092 human genomes (2012). *Nature* 491, 56–65. doi:10.1038/nature11632.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., et al. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat. Ecol. Evol.* 1, 0020. doi:10.1038/s41559-016-0020.

Arlt, M. F., Rajendran, S., Birkeland, S. R., Wilson, T. E., and Glover, T. W. (2014). Copy number variants are produced in response to low-dose ionizing radiation in cultured cells. *Environ. Mol. Mutagen.* 55, 103–113. doi:10.1002/em.21840.

Armstrong-James, D., Youngs, J., Bicanic, T., Abdolrasouli, A., Denning, D. W., Johnson, E., et al. (2020). Confronting and mitigating the risk of COVID-19 associated pulmonary aspergillosis. *Eur. Respir. J.* 56, 2002554. doi:10.1183/13993003.02554-2020.

Arzimanoglou, I. I., Gilbert, F., and Barber, H. R. K. (1998). Microsatellite instability in human solid tumors. *Cancer* 82, 1808–1820. doi:https://doi.org/10.1002/(SICI)1097-0142(19980515)82:10<1808::AID-CNCR2>3.0.CO;2-J.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* doi:10.1038/75556.

Ashu, E. E., Hagen, F., Chowdhary, A., Meis, J. F., and Xu, J. (2017). Global Population Genetic

Analysis of Aspergillus fumigatus. *mSphere*. doi:10.1128/msphere.00019-17.

Askew, D. S. (2008). Aspergillus fumigatus: virulence genes in a street-smart mold. *Curr. Opin. Microbiol.* 11, 331–337. doi:10.1016/j.mib.2008.05.009.

Augusto Corrêa dos Santos, R., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* 33, 2575–2576. doi:10.1093/bioinformatics/btx204.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., et al. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495, 360–4. doi:10.1038/nature11837.

Baack, E. J., and Rieseberg, L. H. (2007). A genomic view of introgression and hybrid speciation. *Curr. Opin. Genet. Dev.* 17, 513–518. doi:10.1016/j.gde.2007.09.001.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2. doi:10.1038/msb4100050.

Balajee, S. A., Gribskov, J. L., Hanley, E., Nickle, D., and Marr, K. A. (2005). Aspergillus lentulus sp. nov., a New Sibling Species of A. fumigatus. *Eukaryot. Cell* 4, 625–632. doi:10.1128/EC.4.3.625-632.2005.

Ballester, A.-R., Marcet-Houben, M., Levin, E., Sela, N., Selma-Lázaro, C., Carmona, L., et al. (2015). Genome, Transcriptome, and Functional Analyses of Penicillium expansum Provide New Insights Into Secondary Metabolism and Pathogenicity. *Mol. Plant-Microbe Interact.* 28, 232–248. doi:10.1094/MPMI-09-14-0261-FI.

Ballesteros, J. A., and Hormiga, G. (2016). A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Mol. Biol. Evol.* 33, 2117–2134.

doi:10.1093/molbev/msw069.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.

Barnes, P. D., and Marr, K. A. (2006). Aspergillosis: Spectrum of Disease, Diagnosis, and Treatment. *Infect. Dis. Clin. North Am.* 20, 545–561. doi:10.1016/j.idc.2006.06.001.

Barnum, K. J., and O'Connell, M. J. (2014). "Cell Cycle Regulation by Checkpoints," in, 29–40. doi:10.1007/978-1-4939-0888-2_2.

Barrs, V. R., van Doorn, T. M., Houbraken, J., Kidd, S. E., Martin, P., Pinheiro, M. D., et al. (2013). Aspergillus felis sp. nov., an Emerging Agent of Invasive Aspergillosis in Humans, Cats, and Dogs. *PLoS One* 8, e64871. doi:10.1371/journal.pone.0064871.

Barton, A. B., Su, Y., Lamb, J., Barber, D., and Kaback, D. B. (2003). A Function for Subtelomeric DNA in Saccharomyces cerevisiae. *Genetics* 165, 929–934.

Bartra, E., Casado, M., Carro, D., Campamà, C., and Piña, B. (2010). Differential expression of thiamine biosynthetic genes in yeast strains with high and low production of hydrogen sulfide during wine fermentation. *J. Appl. Microbiol.* 109, 272–281. doi:10.1111/j.1365-2672.2009.04652.x.

Basenko, E. Y., Pulman, J. A., Shanmugasundram, A., Harb, O. S., Crouch, K., Starns, D., et al. (2018). FungiDB: An integrated bioinformatic resource for fungi and oomycetes. *J. Fungi*. doi:10.3390/jof4010039.

Bastos, R. W., Valero, C., Silva, L. P., Schoen, T., Drott, M., Brauer, V., et al. (2020a). Functional characterization of clinical isolates of the opportunistic fungal pathogen Aspergillus nidulans. *bioRxiv*, 2020.01.28.917278. doi:10.1101/2020.01.28.917278.

Bastos, R. W., Valero, C., Silva, L. P., Schoen, T., Drott, M., Brauer, V., et al. (2020b). Functional Characterization of Clinical Isolates of the Opportunistic Fungal Pathogen Aspergillus nidulans. *mSphere* 5. doi:10.1128/mSphere.00153-20.

Bataillon, M., Rico, A., Sablayrolles, J. M., Salmon, J. M., and Barre, P. (1996). Early thiamin assimilation by yeasts under enological conditions: Impact on alcoholic fermentation kinetics. *J. Ferment. Bioeng.* 82, 145–150. doi:10.1016/0922-338X(96)85037-9.

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ* 324, 1018–1022. doi:10.1136/bmj.324.7344.1018.

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi:10.1093/bioinformatics/btx198.

Beimforde, C., Feldberg, K., Nylinder, S., Rikkinen, J., Tuovila, H., Dörfelt, H., et al. (2014). Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data. *Mol. Phylogenet. Evol.* 78, 386–398. doi:10.1016/j.ympev.2014.04.024.

Bell, M. A., and Lloyd, G. T. (2015). strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. *Palaeontology* 58, 379–389. doi:10.1111/pala.12142.

Benedict, K., Jackson, B. R., Chiller, T., and Beer, K. D. (2019). Estimation of Direct Healthcare Costs of Fungal Diseases in the United States. *Clin. Infect. Dis.* 68, 1791–1797. doi:10.1093/cid/ciy776.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B.* doi:10.1111/j.2517-6161.1995.tb02031.x.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2007). GenBank. *Nucleic Acids Res.* 36, D25–D30. doi:10.1093/nar/gkm929.

Berbee, M. L., and Taylor, J. W. (2001). "Fungal molecular evolution: gene trees and geologic time," in *The Mycota*, 229–246.

Bertini, E., Tatu, A., and Keim, D. (2011). Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Trans. Vis. Comput. Graph.* 17, 2203–2212. doi:10.1109/TVCG.2011.229.

Bertuzzi, M., van Rhijn, N., Krappmann, S., Bowyer, P., Bromley, M. J., and Bignell, E. M. (2020). On the lineage of Aspergillus fumigatus isolates in common laboratory use. *Med. Mycol.* doi:10.1093/mmy/myaa075.

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33, W451–W454. doi:10.1093/nar/gki487.

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* 22, 778–790. doi:10.1101/gr.133967.111.

Bignell, E., Cairns, T. C., Throckmorton, K., Nierman, W. C., and Keller, N. P. (2016). Secondary metabolite arsenal of an opportunistic pathogenic fungus. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20160023. doi:10.1098/rstb.2016.0023.

Billmyre, R. B., Applen Clancey, S., Li, L. X., Doering, T. L., and Heitman, J. (2020). 5-fluorocytosine resistance is associated with hypermutation and alterations in capsule biosynthesis in Cryptococcus. *Nat. Commun.* 11, 127. doi:10.1038/s41467-019-13890-z.

Billmyre, R. B., Clancey, S. A., and Heitman, J. (2017). Natural mismatch repair mutations

mediate phenotypic diversity and drug resistance in Cryptococcus deuterogattii. *Elife* 6. doi:10.7554/eLife.28802.

Birch, J. (2012). Worldwide prevalence of red-green color deficiency. *J. Opt. Soc. Am. A* 29, 313. doi:10.1364/JOSAA.29.000313.

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62. doi:10.1111/j.1469-8137.2009.03087.x.

Birchler, J. A., and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14746–53. doi:10.1073/pnas.1207726109.

Bisson, L. F. (1999). Stuck and sluggish fermentations. *Am. J. Enol. Vitic.* 50, 107–119.

Bisson, L. F. (2012). Geographic origin and diversity of wine strains of saccharomyces. *Am. J. Enol. Vitic.* 63, 165–176. doi:10.5344/ajev.2012.11083.

Blachowicz, A., Raffa, N., Bok, J. W., Choera, T., Knox, B., Lim, F. Y., et al. (2020). Contributions of Spore Secondary Metabolites to UV-C Protection and Virulence Vary in Different Aspergillus fumigatus Strains. *MBio* 11. doi:10.1128/mBio.03415-19.

Blackwell, M. (2011). The Fungi: 1, 2, 3 … 5.1 million species? *Am. J. Bot.* 98, 426–438. doi:10.3732/ajb.1000298.

Blanc-Potard, A. B., and Groisman, E. A. (1997). The Salmonella selC locus contains a pathogenicity island mediating intramacrophage survival. *EMBO J.* doi:10.1093/emboj/16.17.5376.

Bodinaku, I., Shaffer, J., Connors, A. B., Steenwyk, J. L., Biango-Daniels, M. N., Kastman, E. K., et al. (2019). Rapid Phenotypic and Metabolomic Domestication of Wild Penicillium

Molds on Cheese. *MBio* 10. doi:10.1128/mBio.02445-19.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., et al. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28, 423–425. doi:10.1093/bioinformatics/btr670.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., et al. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27, 268–269. doi:10.1093/bioinformatics/btq635.

Bogomolnaya, L. M., Pathak, R., Guo, J., and Polymenis, M. (2006). Roles of the RAM signaling network in cell cycle progression in Saccharomyces cerevisiae. *Curr. Genet.* 49, 384–92. doi:10.1007/s00294-006-0069-y.

Boiteux, S., and Jinks-Robertson, S. (2013). DNA Repair Mechanisms and the Bypass of DNA Damage in Saccharomyces cerevisiae. *Genetics* 193, 1025–1064. doi:10.1534/genetics.112.145219.

Bok, J. W., Chung, D. W., Balajee, S. A., Marr, K. A., Andes, D., Nielsen, K. F., et al. (2006). GliZ, a transcriptional regulator of gliotoxin biosynthesis, contributes to Aspergillus fumigatus virulence. *Infect. Immun.* doi:10.1128/IAI.00780-06.

Boland, C. R., and Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. *Gastroenterology* 138, 2073-2087.e3. doi:10.1053/j.gastro.2009.12.064.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bom, V. L. P., de Castro, P. A., Winkelströter, L. K., Marine, M., Hori, J. I., Ramalho, L. N. Z., et al. (2015). The Aspergillus fumigatus sitA Phosphatase Homologue Is Important for Adhesion, Cell Wall Integrity, Biofilm Formation, and Virulence. *Eukaryot. Cell* 14, 728–

744. doi:10.1128/EC.00008-15.

Bongomin, F., Gago, S., Oladele, R., and Denning, D. (2017). Global and Multi-National Prevalence of Fungal Diseases—Estimate Precision. *J. Fungi* 3, 57. doi:10.3390/jof3040057.

Borneman, A. R., Forgan, A. H., Kolouchova, R., Fraser, J. A., and Schmidt, S. A. (2016). Whole Genome Comparison Reveals High Levels of Inbreeding and Strain Redundancy Across the Spectrum of Commercial Wine Strains of Saccharomyces cerevisiae. *G3 Genes|Genomes|Genetics*, g3.115.025692. doi:10.1534/g3.115.025692.

Borneman, A. R., Forgan, A. H., Pretorius, I. S., and Chambers, P. J. (2008). Comparative genome analysis of a Saccharomyces cerevisiae wine strain. in *FEMS Yeast Research*, 1185–1195. doi:10.1111/j.1567-1364.2008.00434.x.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4, e1660. doi:10.7717/peerj.1660.

Borowiec, M. L., Lee, E. K., Chiu, J. C., and Plachetzki, D. C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics* 16, 987. doi:10.1186/s12864-015-2146-4.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res.* 23, 323–330. doi:10.1101/gr.141978.112.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi:10.1038/nature01521.

Boyce, K. J., Wang, Y., Verma, S., Shakya, V. P. S., Xue, C., and Idnurm, A. (2017). Mismatch

Repair of DNA Replication Errors Contributes to Microevolution in the Pathogenic Fungus Cryptococcus neoformans. *MBio* 8. doi:10.1128/mBio.00595-17.

Brandis, G., and Hughes, D. (2016). The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genet.* 12, e1005926. doi:10.1371/journal.pgen.1005926.

Branzk, N., Lubojemska, A., Hardison, S. E., Wang, Q., Gutierrez, M. G., Brown, G. D., et al. (2014). Neutrophils sense microbe size and selectively release neutrophil extracellular traps in response to large pathogens. *Nat. Immunol.* 15, 1017–1025. doi:10.1038/ni.2987.

Brion, C., Ambroset, C., Delobel, P., Sanchez, I., and Blondin, B. (2014). Deciphering regulatory variation of THI genes in alcoholic fermentation indicate an impact of Thi3p on PDC1 expression. *BMC Genomics* 15, 1085. doi:10.1186/1471-2164-15-1085.

Broustas, C. G., and Lieberman, H. B. (2014). DNA Damage Response Genes and the Development of Cancer Metastasis. *Radiat. Res.* 181, 111–130. doi:10.1667/RR13515.1.

Brown, C. A., Murray, A. W., and Verstrepen, K. J. (2010). Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts. *Curr. Biol.* 20, 895–903. doi:10.1016/j.cub.2010.04.027.

Brown, C. J., Todd, K. M., and Rosenzweig, R. F. (1998). Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* 15, 931–42. doi:10.1093/oxfordjournals.molbev.a026009.

Brown, G. D., Denning, D. W., Gow, N. A. R., Levitz, S. M., Netea, M. G., and White, T. C. (2012). Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* 4, 165rv13-165rv13. doi:10.1126/scitranslmed.3004404.

Brown, J. W., Walker, J. F., and Smith, S. A. (2017). Phyx: phylogenetic tools for unix. *Bioinformatics* 33, 1886–1888. doi:10.1093/bioinformatics/btx063.

Brown, N. A., and Goldman, G. H. (2016). The contribution of Aspergillus fumigatus stress responses to virulence and antifungal resistance. *J. Microbiol.* 54, 243–253. doi:10.1007/s12275-016-5510-4.

Brüggemann, R. J., van de Veerdonk, F. L., and Verweij, P. E. (2020). The challenge of managing COVID-19 associated pulmonary aspergillosis. *Clin. Infect. Dis.* doi:10.1093/cid/ciaa1211.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., et al. (2006). Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Mol. Biol. Evol.* 23, 1808–1816. doi:10.1093/molbev/msl049.

Brunette, G. J., Jamalruddin, M. A., Baldock, R. A., Clark, N. L., and Bernstein, K. A. (2019). Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. *Proc. Natl. Acad. Sci.* 116, 19593–19599. doi:10.1073/pnas.1906559116.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Budden, T., and Bowden, N. (2013). The Role of Altered Nucleotide Excision Repair and UVB-Induced DNA Damage in Melanomagenesis. *Int. J. Mol. Sci.* 14, 1132–1151. doi:10.3390/ijms14011132.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* 459, 657–662. doi:10.1038/nature08064.

Cadez, N. (2006). Phylogenetic placement of Hanseniaspora-Kloeckera species using multigene sequence analysis with taxonomic implications: descriptions of Hanseniaspora pseudoguilliermondii sp. nov. and Hanseniaspora occidentalis var. citrica var. nov. *Int. J.*

Syst. Evol. Microbiol. 56, 1157–1165. doi:10.1099/ijs.0.64052-0.

Čadež, N., Bellora, N., Ulloa, R., Hittinger, C. T., and Libkind, D. (2019). Genomic content of a novel yeast species Hanseniaspora gamundiae sp. nov. from fungal stromata (Cyttaria) associated with a unique fermented beverage in Andean Patagonia, Argentina. *PLoS One* 14, e0210792. doi:10.1371/journal.pone.0210792.

Caesar, L. K., Kvalheim, O. M., and Cech, N. B. (2018). Hierarchical cluster analysis of technical replicates to identify interferents in untargeted mass spectrometry metabolomics. *Anal. Chim. Acta* 1021, 69–77. doi:10.1016/j.aca.2018.03.013.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., et al. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* 171, 1042-1056.e10. doi:10.1016/j.cell.2017.09.048.

Campbell, J., Lin, Q., Geske, G. D., and Blackwell, H. E. (2009). New and Unexpected Insights into the Modulation of LuxR-Type Quorum Sensing by Cyclic Dipeptides. *ACS Chem. Biol.* 4, 1051–1059. doi:10.1021/cb900165y.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi:10.1093/bioinformatics/btp348.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics* 25, 288–289. doi:10.1093/bioinformatics/btn615.

Cartwright, J. L., Gasmi, L., Spiller, D. G., and McLennan, A. G. (2000). The Saccharomyces cerevisiae PCD1 gene encodes a peroxisomal nudix hydrolase active toward coenzyme A and its derivatives. *J. Biol. Chem.* doi:10.1074/jbc.M005015200.

Casadevall, A. (2007). Determinants of virulence in the pathogenic fungi. *Fungal Biol. Rev.* 21, 130–132. doi:10.1016/j.fbr.2007.02.007.

Casadevall, A., and Pirofski, L. (2007). Accidental Virulence, Cryptic Pathogenesis, Martians, Lost Hosts, and the Pathogenicity of Environmental Microbes. *Eukaryot. Cell* 6, 2169–2174. doi:10.1128/EC.00308-07.

Castoe, T. A., de Koning, A. P. J., Kim, H.-M., Gu, W., Noonan, B. P., Naylor, G., et al. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci.* 106, 8986–8991. doi:10.1073/pnas.0900233106.

Castro, A., Bernis, C., Vigneron, S., Labbé, J.-C., and Lorca, T. (2005). The anaphase-promoting complex: a key factor in the regulation of cell cycle. *Oncogene* 24, 314–325. doi:10.1038/sj.onc.1207973.

Cavalieri, D., McGovern, P. E., Hartl, D. L., Mortimer, R., and Polsinelli, M. (2003). Evidence for S. cerevisiae Fermentation in Ancient Wine. in *Journal of Molecular Evolution* doi:10.1007/s00239-003-0031-2.

Chain, E., Florey, H. W., Gardner, A. D., Heatley, N. G., Jennings, M. A., Orr-Ewing, J., et al. (1940). Pencillin as a chemotherapeutic agent. *Lancet* 236, 226–228. doi:10.1016/S0140-6736(01)08728-1.

Chamilos, G., and Kontoyiannis, D. P. (2005). Update on antifungal drug resistance mechanisms of Aspergillus fumigatus. *Drug Resist. Updat.* 8, 344–358. doi:10.1016/j.drup.2006.01.001.

Chang, C.-F., Huang, L.-Y., Chen, S.-F., and Lee, C.-F. (2012). Kloeckera taiwanica sp. nov., an

ascomycetous apiculate yeast species isolated from mushroom fruiting bodies. *Int. J. Syst. Evol. Microbiol.* 62, 1434–1437. doi:10.1099/ijs.0.034231-0.

Chang, D.-Y., Gu, Y., and Lu, A.-L. (2001). Fission yeast (Schizosaccharomyces pombe) cells defective in the MutY-homologous glycosylase activity have a mutator phenotype and are sensitive to hydrogen peroxide. *Mol. Genet. Genomics* 266, 336–342. doi:10.1007/s004380100567.

Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., et al. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14912–7. doi:10.1073/pnas.1511468112.

Charron, M. J., Read, E., Haut, S. R., and Michels, C. A. (1989). Molecular evolution of the telomere-associated MAL loci of Saccharomyces. *Genetics* 122, 307–16.

Chavan, P., Mane, S., Kulkarni, G., Shaikh, S., Ghormade, V., Nerkar, D. P., et al. (2009). Natural yeast flora of different varieties of grapes used for wine making in India. *Food Microbiol.* 26, 801–808. doi:10.1016/j.fm.2009.05.005.

Chen, A. J., Frisvad, J. C., Sun, B. D., Varga, J., Kocsubé, S., Dijksterhuis, J., et al. (2016). Aspergillus section Nidulantes (formerly Emericella): Polyphasic taxonomy, chemistry and biology. *Stud. Mycol.* 84, 1–118. doi:10.1016/J.SIMYCO.2016.10.001.

Chen, M.-Y., Liang, D., and Zhang, P. (2015). Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. *Syst. Biol.* 64, 1104–1120. doi:10.1093/sysbio/syv059.

Chen, M.-Y., Liang, D., and Zhang, P. (2017). Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences. *Genome Biol. Evol.* 9, 1998–2012. doi:10.1093/gbe/evx147.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395, 507–513. doi:10.1016/S0140-6736(20)30211-7.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* 65, 997–1008. doi:10.1093/sysbio/syw037.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., et al. (2012a). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705. doi:10.1093/nar/gkr1029.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., et al. (2012b). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705. doi:10.1093/nar/gkr1029.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30, 31–37. doi:10.1093/bioinformatics/btt310.

Chikina, M., Robinson, J. D., and Clark, N. L. (2016). Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Mol. Biol. Evol.* 33, 2182–2192. doi:10.1093/molbev/msw112.

Choin, J., Mendoza-Revilla, J., Arauna, L. R., Cuadros-Espinoza, S., Cassar, O., Larena, M., et al. (2021). Genomic insights into population history and biological adaptation in Oceania. *Nature* 592, 583–589. doi:10.1038/s41586-021-03236-5.

Chowdhary, A., Sharma, C., Hagen, F., and Meis, J. F. (2014). Exploring azole antifungal drug resistance in Aspergillus fumigatus with special reference to resistance mechanisms. *Future Microbiol.* 9, 697–711. doi:10.2217/fmb.14.27.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* 6, 80–92. doi:10.4161/fly.19695.

Clark, N. L., Alani, E., and Aquadro, C. F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.* 22, 714–720. doi:10.1101/gr.132647.111.

CLSI (2008). M38-A2 Reference Method for Broth Dilution Antifungal Susceptibility Testing of Filamentous Fungi; Approved Standard—Second Edition. *Clin. Lab. Stand. Inst.*

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009a). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009b). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163.

Colman-Lerner, A., Chin, T. E., and Brent, R. (2001). Yeast Cbk1 and Mob2 Activate Daughter-Specific Genetic Programs to Induce Asymmetric Cell Fates. *Cell* 107, 739–750. doi:10.1016/S0092-8674(01)00596-7.

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi:10.1093/bioinformatics/btx364.

Costanzo, M., Nishikawa, J. L., Tang, X., Millman, J. S., Schub, O., Breitkreuz, K., et al. (2004). CDK Activity Antagonizes Whi5, an Inhibitor of G1/S Transcription in Yeast. *Cell* 117, 899–913. doi:10.1016/j.cell.2004.05.024.

Cox, M. J., Loman, N., Bogaert, D., and O'Grady, J. (2020). Co-infections: potentially lethal and unexplored in COVID-19. *The Lancet Microbe* 1, e11. doi:10.1016/S2666-5247(20)30009-4.

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10, 210. doi:10.1186/1471-2148-10-210.

Cromie, G. a, Hyma, K. E., Ludlow, C. L., Garmendia-Torres, C., Gilbert, T. L., May, P., et al. (2013). Genomic sequence diversity and population structure of Saccharomyces cerevisiae assessed by RAD-seq. *G3 (Bethesda).* 3, 2163–71. doi:10.1534/g3.113.007492.

Cross, F. R., Buchler, N. E., and Skotheim, J. M. (2011). Evolution of networks and sequences in eukaryotic cell cycle control. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 366, 3532–44. doi:10.1098/rstb.2011.0078.

Cutler, G., and Kassner, P. D. (2008). Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. *Cytogenet Genome Res* 123, 297–306. doi:10.1159/000184721.

Da Lage, J.-L., Binder, M., Hua-Van, A., Janeček, Š., and Casane, D. (2013). Gene make-up: rapid and massive intron gains after horizontal transfer of a bacterial α-amylase gene to Basidiomycetes. *BMC Evol. Biol.* 13, 40. doi:10.1186/1471-2148-13-40.

Da Silva Ferreira, M. E., Luiz Capellaro, J., Dos Reis Marques, E., Malavazi, I., Perlin, D., Park, S., et al. (2004). In vitro evolution of itraconazole resistance in Aspergillus fumigatus involves multiple mechanisms of resistance. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.48.11.4405-4413.2004.

Dagenais, T. R. T., and Keller, N. P. (2009). Pathogenesis of Aspergillus fumigatus in Invasive

Aspergillosis. *Clin. Microbiol. Rev.* 22, 447–465. doi:10.1128/CMR.00055-08.

Dahal, B. K., Kadyrova, L. Y., Delfino, K. R., Rogozin, I. B., Gujar, V., Lobachev, K. S., et al. (2017). Involvement of DNA mismatch repair in the maintenance of heterochromatic DNA stability in Saccharomyces cerevisiae. *PLOS Genet.* 13, e1007074.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The State of Software for Evolutionary Biology. *Mol. Biol. Evol.* 35, 1037–1046. doi:10.1093/molbev/msy014.

Davies, J., and Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433. doi:10.1128/MMBR.00016-10.

De Bont, R. (2004). Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* 19, 169–185. doi:10.1093/mutage/geh025.

de Oliveira Martins, L., and Posada, D. (2017). "Species Tree Estimation from Genome-Wide Data with guenomu," in, 461–478. doi:10.1007/978-1-4939-6622-6_18.

de Vries, R. P., Riley, R., Wiebenga, A., Aguilar-Osorio, G., Amillis, S., Uchima, C. A., et al. (2017). Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. *Genome Biol.* 18, 28. doi:10.1186/s13059-017-1151-0.

Degnan, J. H., and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution (N. Y).* 59, 24–37. doi:10.1111/j.0014-3820.2005.tb00891.x.

Depotter, J. R., Seidl, M. F., Wood, T. A., and Thomma, B. P. (2016). Interspecific hybridization impacts host range and pathogenicity of filamentous microbes. *Curr. Opin. Microbiol.* 32, 7–13. doi:10.1016/j.mib.2016.04.005.

Derilus, D., Rahman, M. Z., Serrano, A. E., and Massey, S. E. (2021). Proteome size reduction in Apicomplexans is linked with loss of DNA repair and host redundant pathways. *Infect.*

Genet. Evol. 87, 104642. doi:https://doi.org/10.1016/j.meegid.2020.104642.

Diawara, B., Kando, C., Anyogu, A., Ouoba, L. I. I., Nielsen, D. S., Sutherland, J. P., et al. (2015). Hanseniaspora jakobsenii sp. nov., a yeast isolated from Bandji, a traditional palm wine of Borassus akeassii. *Int. J. Syst. Evol. Microbiol.* 65, 3576–3579. doi:10.1099/ijsem.0.000461.

Dimitrova, Y. N., Jenni, S., Valverde, R., Khin, Y., and Harrison, S. C. (2016). Structure of the MIND Complex Defines a Regulatory Focus for Yeast Kinetochore Assembly. *Cell* 167, 1014-1027.e12. doi:10.1016/j.cell.2016.10.011.

Dolan, S. K., Owens, R. A., O'Keeffe, G., Hammel, S., Fitzpatrick, D. A., Jones, G. W., et al. (2014). Regulation of Nonribosomal Peptide Synthesis: bis-Thiomethylation Attenuates Gliotoxin Biosynthesis in Aspergillus fumigatus. *Chem. Biol.* 21, 999–1012. doi:10.1016/j.chembiol.2014.07.006.

Dong, Y., Chen, S., Cheng, S., Zhou, W., Ma, Q., Chen, Z., et al. (2019). Natural selection and repeated patterns of molecular evolution following allopatric divergence. *Elife* 8. doi:10.7554/eLife.45199.

Doolittle, W. F., and Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci.* 104, 2043–2049. doi:10.1073/pnas.0610699104.

Dos Reis Almeida, F. B., Carvalho, F. C., Mariano, V. S., Alegre, A. C. P., Silva, R. do N., Hanna, E. S., et al. (2011). Influence of N-Glycosylation on the Morphogenesis and Growth of Paracoccidioides brasiliensis and on the Biological Activities of Yeast Proteins. *PLoS One* 6, e29216. doi:10.1371/journal.pone.0029216.

dos Reis, M., Donoghue, P. C. J., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* 17, 71–80.

doi:10.1038/nrg.2015.8.

Dos Reis, M., and Yang, Z. (2013). The unbearable uncertainty of Bayesian divergence time estimation. *J. Syst. Evol.* 51, 30–43. doi:10.1111/j.1759-6831.2012.00236.x.

Dos Reis, T. F., Silva, L. P., de Castro, P. A., do Carmo, R. A., Marini, M. M., da Silveira, J. F., et al. (2019). The Aspergillus fumigatus Mismatch Repair MSH2 Homolog Is Important for Virulence and Azole Resistance. *mSphere* 4, e00416-19. doi:10.1128/mSphere.00416-19.

dos Santos, R. A. C., Rivero-Menendez, O., Steenwyk, J. L., Mead, M. E., Goldman, G. H., Alastruey-Izquierdo, A., et al. (2020a). Draft Genome Sequences of Four Aspergillus Section Fumigati Clinical Strains. *Microbiol. Resour. Announc.* 9. doi:10.1128/MRA.00856-20.

dos Santos, R. A. C., Steenwyk, J. L., Rivero-Menendez, O., Mead, M. E., Silva, L. P., Bastos, R. W., et al. (2020b). Genomic and Phenotypic Heterogeneity of Clinical Isolates of the Human Pathogens Aspergillus fumigatus, Aspergillus lentulus, and Aspergillus fumigatiaffinis. *Front. Genet.* 11. doi:10.3389/fgene.2020.00459.

Dotis, J., and Roilides, E. (2004). Osteomyelitis due to Aspergillus spp. in patients with chronic granulomatous disease: comparison of Aspergillus nidulans and Aspergillus fumigatus. *Int. J. Infect. Dis.* 8, 103–110. doi:10.1016/j.ijid.2003.06.001.

Doyle, V. P., Young, R. E., Naylor, G. J. P., and Brown, J. M. (2015). Can We Identify Genes with Increased Phylogenetic Reliability? *Syst. Biol.* 64, 824–837. doi:10.1093/sysbio/syv041.

Dress, A. W., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., et al. (2008). Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms Mol. Biol.* 3, 7. doi:10.1186/1748-7188-3-7.

Drewniak, A., Gazendam, R. P., Tool, A. T. J., van Houdt, M., Jansen, M. H., van Hamme, J. L., et al. (2013). Invasive fungal infection and impaired neutrophil killing in human CARD9 deficiency. *Blood* 121, 2385–2392. doi:10.1182/blood-2012-08-450551.

Drgona, L., Khachatryan, A., Stephens, J., Charbonneau, C., Kantecki, M., Haider, S., et al. (2014). Clinical and economic burden of invasive fungal diseases in Europe: focus on pre-emptive and empirical treatment of Aspergillus and Candida species. *Eur. J. Clin. Microbiol. Infect. Dis.* 33, 7–21. doi:10.1007/s10096-013-1944-3.

Drott, M. T., Bastos, R. W., Rokas, A., Ries, L. N. A., Gabaldón, T., Goldman, G. H., et al. (2020). Diversity of Secondary Metabolism in Aspergillus nidulans Clinical Isolates. *mSphere* 5. doi:10.1128/mSphere.00156-20.

Duan, J., Zhang, J.-G., Deng, H.-W., and Wang, Y.-P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 8, e59128. doi:10.1371/journal.pone.0059128.

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., et al. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16144–16149. doi:10.1073/pnas.242624799.

Dunn, B., Richter, C., Kvitek, D. J., Pugh, T., and Sherlock, G. (2012). Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res.* 22, 908–924. doi:10.1101/gr.130310.111.

Dunn, B., and Sherlock, G. (2008). Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. *Genome Res.* 18, 1610–1623. doi:10.1101/gr.076075.108.

Dunn, C. W., Howison, M., and Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14, 330. doi:10.1186/1471-2105-14-330.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7, e1002195. doi:10.1371/journal.pcbi.1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution (N. Y).* 63, 1–19. doi:10.1111/j.1558-5646.2008.00549.x.

Eidem, H. R., Steenwyk, J. L., Wisecaver, J. H., Capra, J. A., Abbot, P., and Rokas, A. (2018). integRATE: a desirability-based data integration framework for the prioritization of candidate genes across heterogeneous omics and its application to preterm birth. *BMC Med. Genomics* 11, 107. doi:10.1186/s12920-018-0426-y.

El-Elimat, T., Figueroa, M., Ehrmann, B. M., Cech, N. B., Pearce, C. J., and Oberlies, N. H. (2013). High-Resolution MS, MS/MS, and UV Database of Fungal Secondary Metabolites as a Dereplication Protocol for Bioactive Natural Products. *J. Nat. Prod.* 76, 1709–1716. doi:10.1021/np4004307.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435–445. doi:10.1038/nrg1348.

Ellingham, O., David, J., and Culham, A. (2019). Enhancing identification accuracy for powdery mildews using previously underexploited DNA loci. *Mycologia* 111, 798–812. doi:10.1080/00275514.2019.1643644.

Elliott, T. A., and Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140331.

doi:10.1098/rstb.2014.0331.

Embley, M., der Giezen, M. van, Horner, D. S., Dyal, P. L., and Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* 358, 191–203. doi:10.1098/rstb.2002.1190.

Emms, D. M., and Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *bioRxiv*, 267914. doi:10.1101/267914.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y.

Endo, A. (2010). A historical perspective on the discovery of statins. *Proc. Japan Acad. Ser. B* 86, 484–493. doi:10.2183/pjab.86.484.

Endo, A., Kuroda, M., and Tsujita, Y. (1976). ML-236A, ML-236B, and ML-236C, new inhibitors of cholesterogensis produced by Penicillium citrinum. *J. Antibiot. (Tokyo).* 29, 1346–1348. doi:10.7164/antibiotics.29.1346.

Epstein, C. J. (1967). Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature* 215, 355–359. doi:10.1038/215355a0.

Fallon, J. P., Reeves, E. P., and Kavanagh, K. (2010). Inhibition of neutrophil function following exposure to the Aspergillus fumigatus toxin fumagillin. *J. Med. Microbiol.* 59, 625–633. doi:10.1099/jmm.0.018192-0.

Fallon, J. P., Reeves, E. P., and Kavanagh, K. (2011). The Aspergillus fumigatus toxin fumagillin suppresses the immune response of Galleria mellonella larvae by inhibiting the action of haemocytes. *Microbiology* 157, 1481–1488. doi:10.1099/mic.0.043786-0.

Fan, H.-W., Noda, H., Xie, H.-Q., Suetsugu, Y., Zhu, Q.-H., and Zhang, C.-X. (2015). Genomic Analysis of an Ascomycete Fungus from the Rice Planthopper Reveals How It Adapts to an

Endosymbiotic Lifestyle. *Genome Biol. Evol.* 7, 2623–2634. doi:10.1093/gbe/evv169.

Fang, F. C. (2011). Antimicrobial Actions of Reactive Oxygen Species. *MBio* 2. doi:10.1128/mBio.00141-11.

Farrer, R. A., Henk, D. A., Garner, T. W., Balloux, F., Woodhams, D. C., and Fisher, M. C. (2013). Chromosomal copy number variation, selection and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity. *PLoS Genet* 9, e1003703. doi:10.1371/journal.pgen.1003703.

Fedorova, N. D., Khaldi, N., Joardar, V. S., Maiti, R., Amedeo, P., Anderson, M. J., et al. (2008). Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. *PLoS Genet.* 4. doi:10.1371/journal.pgen.1000046.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst. Biol.* 27, 401–410. doi:10.1093/sysbio/27.4.401.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi:10.1007/BF01734359.

Felsenstein, J. (1986). The Newick tree format. *English*.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* 266, 418–27.

Feng, Y.-J., Blackburn, D. C., Liang, D., Hillis, D. M., Wake, D. B., Cannatella, D. C., et al. (2017). Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. *Proc. Natl. Acad. Sci.* 114, E5864–E5870. doi:10.1073/pnas.1704632114.

Ferling, I., Dunn, J. D., Ferling, A., Soldati, T., Hillmann, F., and Goldman, G. H. (2020). Conidial melanin of the human-pathogenic fungus Aspergillus fumigatus disrupts cell

autonomous defenses in amoebae. *MBio*. doi:10.1128/mBio.00862-20.

Fidalgo, M., Barrales, R. R., Ibeas, J. I., and Jimenez, J. (2006). Adaptive evolution by mutations in the FLO11 gene. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11228–33. doi:10.1073/pnas.0601713103.

Fleming, A. (1980). On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. influenzae. *Clin. Infect. Dis.* 2, 129–139. doi:10.1093/clinids/2.1.129.

Fletcher, W., and Yang, Z. (2009). INDELible: A Flexible Simulator of Biological Sequence Evolution. *Mol. Biol. Evol.* 26, 1879–1888. doi:10.1093/molbev/msp098.

Fourment, M., and Gibbs, M. J. (2006). PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evol. Biol.* 6, 1. doi:10.1186/1471-2148-6-1.

Fox, E. M., and Howlett, B. J. (2008). Secondary metabolism: regulation and role in fungal biology. *Curr. Opin. Microbiol.* 11, 481–487. doi:10.1016/j.mib.2008.10.007.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: New insights in genome diversity. *Genome Res.* 16, 949–961. doi:10.1101/gr.3677206.

Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A., and Ellenberger, T. (2005). *DNA Repair and Mutagenesis*. doi:10.1128/9781555816704.

Frisvad, J. C., and Larsen, T. O. (2015). Chemodiversity in the genus Aspergillus. *Appl. Microbiol. Biotechnol.* 99, 7859–7877. doi:10.1007/s00253-015-6839-z.

Frisvad, J. C., and Larsen, T. O. (2016). Extrolites of aspergillus fumigatus and other pathogenic species in aspergillus section fumigati. *Front. Microbiol.* doi:10.3389/fmicb.2015.01485.

Fuchs, B. B., O'Brien, E., Khoury, J. B. E., and Mylonakis, E. (2010). Methods for using Galleria mellonella as a model host to study fungal pathogenesis. *Virulence*. doi:10.4161/viru.1.6.12985.

Fukui, K. (2010). DNA Mismatch Repair in Eukaryotes and Bacteria. *J. Nucleic Acids* 2010.

Furukawa, T., van Rhijn, N., Fraczek, M., Gsaller, F., Davies, E., Carr, P., et al. (2020). The negative cofactor 2 complex is a key regulator of drug resistance in Aspergillus fumigatus. *Nat. Commun.* 11, 427. doi:10.1038/s41467-019-14191-1.

Gabaldón, T., and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14, 360–366. doi:10.1038/nrg3456.

Gabaldón, T., Naranjo-Ortíz, M. A., and Marcet-Houben, M. (2016). Evolutionary genomics of yeast pathogens in the Saccharomycotina. *FEMS Yeast Res.* 16, fow064. doi:10.1093/femsyr/fow064.

Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L.-J., Wortman, J. R., Batzoglou, S., et al. (2005). Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 438, 1105–1115. doi:10.1038/nature04341.

Galgoczy, D. J., and Toczyski, D. P. (2001). Checkpoint Adaptation Precedes Spontaneous and Damage-Induced Genomic Instability in Yeast. *Mol. Cell. Biol.* 21, 1710–1718. doi:10.1128/MCB.21.5.1710-1718.2001.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell* 166, 1397-1410.e16. doi:10.1016/j.cell.2016.08.020.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* 159, 907–911.

doi:10.1093/genetics/159.2.907.

Gangneux, J.-P., Reizine, F., Guegan, H., Pinceaux, K., Le Balch, P., Prat, E., et al. (2020). Is the COVID-19 Pandemic a Good Time to Include Aspergillus Molecular Detection to Categorize Aspergillosis in ICU Patients? A Monocentric Experience. *J. Fungi* 6, 105. doi:10.3390/jof6030105.

Garcia-Rubio, R., Monzon, S., Alcazar-Fuoli, L., Cuesta, I., and Mellado, E. (2018). Genome-Wide Comparative Analysis of Aspergillus fumigatus Strains: The Reference Genome as a Matter of Concern. *Genes (Basel).* 9, 363. doi:10.3390/genes9070363.

Gastebois, A., Blanc Potard, A. B., Gribaldo, S., Beau, R., Latgé, J. P., and Mouyna, I. (2011). Phylogenetic and functional analysis of Aspergillus fumigatus MGTC, a fungal protein homologous to a bacterial virulence factor. *Appl. Environ. Microbiol.* doi:10.1128/AEM.00243-11.

Gaudêncio, S. P., and Pereira, F. (2015). Dereplication: racing to speed up the natural products discovery process. *Nat. Prod. Rep.* 32, 779–810. doi:10.1039/C4NP00134F.

Gauthier, T., Wang, X., Sifuentes Dos Santos, J., Fysikopoulos, A., Tadrist, S., Canlet, C., et al. (2012). Trypacidin, a Spore-Borne Toxin from Aspergillus fumigatus, Is Cytotoxic to Lung Cells. *PLoS One* 7, e29906. doi:10.1371/journal.pone.0029906.

Gaya, E., Fernández-Brime, S., Vargas, R., Lachlan, R. F., Gueidan, C., Ramírez-Mejía, M., et al. (2015). The adaptive radiation of lichen-forming Teloschistaceae is associated with sunscreening pigments and a bark-to-rock substrate shift. *Proc. Natl. Acad. Sci.* 112, 11600–11605. doi:10.1073/pnas.1507072112.

Gazendam, R. P., van Hamme, J. L., Tool, A. T. J., Hoogenboezem, M., van den Berg, J. M., Prins, J. M., et al. (2016). Human Neutrophils Use Different Mechanisms To Kill

Aspergillus fumigatus Conidia and Hyphae: Evidence from Phagocyte Defects. *J. Immunol.* 196, 1272–1283. doi:10.4049/jimmunol.1501811.

Geiser, D. M., Pitt, J. I., and Taylor, J. W. (1998). Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus. *Proc. Natl. Acad. Sci.* 95, 388–393. doi:10.1073/pnas.95.1.388.

GeneOntologyConsortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, 258D – 261. doi:10.1093/nar/gkh036.

Gerik, K. J., Li, X., Pautz, A., and Burgers, P. M. J. (1998). Characterization of the Two Small Subunits of Saccharomyces cerevisiae DNA Polymerase δ *. *J. Biol. Chem.* 273, 19747–19755. doi:10.1074/jbc.273.31.19747.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418, 387–391. doi:10.1038/nature00935.

Gianoulis, T. A., Griffin, M. A., Spakowicz, D. J., Dunican, B. F., Alpha, C. J., Sboner, A., et al. (2012). Genomic Analysis of the Hydrocarbon-Producing, Cellulolytic, Endophytic Fungus Ascocoryne sarcoides. *PLOS Genet.* 8, e1002558.

Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S., and Lemos, B. (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl. Acad. Sci. U. S. A.* 112, 2485–2490. doi:10.1073/pnas.1416878112.

Gibbons, J. G., and Rinker, D. C. (2015). The genomics of microbial domestication in the fermented food environment. *Curr. Opin. Genet. Dev.* 35, 1–8. doi:10.1016/j.gde.2015.07.003.

Gibbons, J. G., and Rokas, A. (2013). The function and evolution of the Aspergillus genome.

Trends Microbiol. 21, 14–22. doi:10.1016/j.tim.2012.09.005.

Gibbons, J. G., Salichos, L., Slot, J. C., Rinker, D. C., McGary, K. L., King, J. G., et al. (2012). The Evolutionary Imprint of Domestication on Genome Variation and Function of the Filamentous Fungus Aspergillus oryzae. *Curr. Biol.* 22, 1403–1409. doi:10.1016/j.cub.2012.05.033.

Giglia-Mari, G., Zotter, A., and Vermeulen, W. (2010). DNA damage response. *Cold Spring Harb. Perspect. Biol.* 3, a000745–a000745. doi:10.1101/cshperspect.a000745.

Gill, E. E., and Fast, N. M. (2007). Stripped-down DNA repair in a highly reduced parasite. *BMC Mol. Biol.* 8, 24. doi:10.1186/1471-2199-8-24.

Giorello, F. M., Berná, L., Greif, G., Camesasca, L., Salzman, V., Medina, K., et al. (2014). Genome Sequence of the Native Apiculate Wine Yeast Hanseniaspora vineae T02/19AF. *Genome Announc.* 2, e00530-14. doi:10.1128/genomeA.00530-14.

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., et al. (2001). Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. *Science (80-.).* 291, 2606–2608. doi:10.1126/science.1056421.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 Genes. *Science (80-.).* 274, 546–567. doi:10.1126/science.274.5287.546.

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., et al. (2016). Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. *Curr. Biol.* 26, 2750–2761. doi:http://dx.doi.org/10.1016/j.cub.2016.08.040.

González-Lobato, L., Real, R., Prieto, J. G., Álvarez, A. I., and Merino, G. (2010). Differential inhibition of murine Bcrp1/Abcg2 and human BCRP/ABCG2 by the mycotoxin fumitremorgin C. *Eur. J. Pharmacol.* 644, 41–48. doi:10.1016/j.ejphar.2010.07.016.

Govender, P., Domingo, J. L., Bester, M. C., Pretorius, I. S., and Bauer, F. F. (2008). Controlled expression of the dominant flocculation genes FLO1, FLO5, and FLO11 in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* 74, 6041–6052. doi:10.1128/AEM.00394-08.

Grahl, N., Shepardson, K. M., Chung, D., and Cramer, R. A. (2012). Hypoxia and Fungal Pathogenesis: To Air or Not To Air? *Eukaryot. Cell* 11, 560–570. doi:10.1128/EC.00031-12.

Gregg, K., and Kauffman, C. (2015). Invasive Aspergillosis: Epidemiology, Clinical Aspects, and Treatment. *Semin. Respir. Crit. Care Med.* 36, 662–672. doi:10.1055/s-0035-1562893.

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., et al. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* 4. doi:10.1371/journal.pgen.1000303.

Gribaldo, S., and Philippe, H. (2002). Ancient Phylogenetic Relationships. *Theor. Popul. Biol.* 61, 391–408. doi:10.1006/tpbi.2002.1593.

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42, D699–D704. doi:10.1093/nar/gkt1183.

Gu, X., Fu, Y. X., and Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* 12, 546–57. doi:10.1093/oxfordjournals.molbev.a040235.

Gupta, S., Paul, K., and Kaur, S. (2020). Diverse species in the genus Cryptococcus : Pathogens and their non-pathogenic ancestors. *IUBMB Life* 72, 2303–2312. doi:10.1002/iub.2377.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086.

Guruceaga, X., Ezpeleta, G., Mayayo, E., Sueiro-Olivares, M., Abad-Diaz-De-Cerio, A., Aguirre Urízar, J. M., et al. (2018). A possible role for fumagillin in cellular damage during host infection by Aspergillus fumigatus. *Virulence* 9, 1548–1561. doi:10.1080/21505594.2018.1526528.

Guruceaga, X., Perez-Cuesta, U., Abad-Diaz de Cerio, A., Gonzalez, O., Alonso, R. M., Hernando, F. L., et al. (2019). Fumagillin, a Mycotoxin of Aspergillus fumigatus: Biosynthesis, Biological Activities, Detection, and Applications. *Toxins (Basel).* 12, 7. doi:10.3390/toxins12010007.

Hakem, R. (2008). DNA-damage repair; the good, the bad, and the ugly. *EMBO J.* 27, 589–605. doi:10.1038/emboj.2008.15.

Halász, J., Podányi, B., Vasvári-Debreczy, L., Szabó, A., Hajdú, F., Böcskei, Z., et al. (2000). Structure Elucidation of Fumagillin-Related Natural Products. *Tetrahedron* 56, 10081–10085. doi:10.1016/S0040-4020(00)00979-0.

Hallett, M., Lagergren, J., and Tofigh, A. (2004). Simultaneous identification of duplications and lateral transfers. in *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04* (New York, New York, USA: ACM Press), 347–356. doi:10.1145/974614.974660.

Hallström, B. M., Kullberg, M., Nilsson, M. A., and Janke, A. (2007). Phylogenomic Data Analyses Provide Evidence that Xenarthra and Afrotheria Are Sister Groups. *Mol. Biol. Evol.* 24, 2059–2068. doi:10.1093/molbev/msm136.

Harrell Jr, F. E. (2015). Package "Hmisc" (v4.0-0).

Hartwell, L. (1992). Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. *Cell*. doi:10.1016/0092-8674(92)90586-2.

Hawksworth, D. L., and Lücking, R. (2017). "Fungal Diversity Revisited: 2.2 to 3.8 Million Species," in *The Fungal Kingdom* (Washington, DC, USA: ASM Press), 79–95. doi:10.1128/9781555819583.ch4.

Heagy, F. C., and Roper, J. A. (1952). Deoxyribonucleic Acid Content of Haploid and Diploid Aspergillus Conidia. *Nature* 170, 713–714. doi:10.1038/170713b0.

Healey, K. R., Zhao, Y., Perez, W. B., Lockhart, S. R., Sobel, J. D., Farmakiotis, D., et al. (2016). Prevalent mutator genotype identified in fungal pathogen Candida glabrata promotes multi-drug resistance. *Nat. Commun.* 7, 11128. doi:10.1038/ncomms11128.

Hedayati, M. T., Pasqualotto, A. C., Warn, P. A., Bowyer, P., and Denning, D. W. (2007). Aspergillus flavus: human pathogen, allergen and mycotoxin producer. *Microbiology* 153, 1677–1692. doi:10.1099/mic.0.2007/007641-0.

Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971–2972. doi:10.1093/bioinformatics/btl505.

Heinrich, S., Sewart, K., Windecker, H., Langegger, M., Schmidt, N., Hustedt, N., et al. (2014). Mad1 contribution to spindle assembly checkpoint signalling goes beyond presenting Mad2 at kinetochores. *EMBO Rep.* 15, 291–298. doi:10.1002/embr.201338114.

Hendler, A., Medina, E. M., Kishkevich, A., Abu-Qarn, M., Klier, S., Buchler, N. E., et al. (2017). Gene duplication and co-evolution of G1/S transcription factor specificity in fungi are essential for optimizing cell fitness. *PLOS Genet.* 13, e1006778. doi:10.1371/journal.pgen.1006778.

Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Hum Mol Genet* 18, R1-8. doi:10.1093/hmg/ddp011.

Henriet, S. S. V., Verweij, P. E., and Warris, A. (2012). Aspergillus nidulans and Chronic Granulomatous Disease: A Unique Host–Pathogen Interaction. *J. Infect. Dis.* 206, 1128–1137. doi:10.1093/infdis/jis473.

Hernández, Y., Bernstein, R., Pagan, P., Vargas, L., McCaig, W., Ramrattan, G., et al. (2018). BpWrapper: BioPerl-based sequence and tree utilities for rapid prototyping of bioinformatics pipelines. *BMC Bioinformatics* 19, 76. doi:10.1186/s12859-018-2074-9.

Hershberg, R., and Petrov, D. A. (2009). General Rules for Optimal Codon Choice. *PLoS Genet.* 5, e1000556. doi:10.1371/journal.pgen.1000556.

Hershberg, R., and Petrov, D. A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genet.* 6, e1001115. doi:10.1371/journal.pgen.1001115.

Hess, J., and Goldman, N. (2011). Addressing Inter-Gene Heterogeneity in Maximum Likelihood Phylogenomic Analysis: Yeasts Revisited. *PLoS One* 6, e22783. doi:10.1371/journal.pone.0022783.

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., et al. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (80-.).* 320, 362–365. doi:10.1126/science.1150021.

Hillmann, F., Novohradská, S., Mattern, D. J., Forberger, T., Heinekamp, T., Westermann, M., et al. (2015). Virulence determinants of the human pathogenic fungus Aspergillus fumigatus protect against soil amoeba predation. *Environ. Microbiol.* doi:10.1111/1462-2920.12808.

Hittinger, C. T., Gonçalves, P., Sampaio, J. P., Dover, J., Johnston, M., and Rokas, A. (2010). Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* 464, 54–58. doi:10.1038/nature08791.

Hittinger, C. T., Rokas, A., and Carroll, S. B. (2004). Parallel inactivation of multiple GAL

pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci.* 101, 14144–14149. doi:10.1073/pnas.0404319101.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35, 518–522. doi:10.1093/molbev/msx281.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. (2007). Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genet.* 3, e7. doi:10.1371/journal.pgen.0030007.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491. doi:10.1186/1471-2105-12-491.

Houbraken, J., de Vries, R. P., and Samson, R. A. (2014). "Modern Taxonomy of Biotechnologically Important Aspergillus and Penicillium Species," in *Advances in Applied Microbiology*, 199–249. doi:10.1016/B978-0-12-800262-9.00004-4.

Houbraken, J., and Samson, R. A. (2011). Phylogeny of Penicillium and the segregation of Trichocomaceae into three families. *Stud. Mycol.* 70, 1–51. doi:10.3114/sim.2011.70.01.

Houbraken, J., Weig, M., Groß, U., Meijer, M., and Bader, O. (2016). Aspergillus oerlinghausenensis , a new mould species closely related to A. fumigatus. *FEMS Microbiol. Lett.* 363, fnv236. doi:10.1093/femsle/fnv236.

Houldcroft, C. J., Beale, M. A., and Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* 15, 183–192. doi:10.1038/nrmicro.2016.182.

Howard, S. J., and Arendrup, M. C. (2011). Acquired antifungal drug resistance in Aspergillus fumigatus: epidemiology and detection. *Med. Mycol.* 49, S90–S95. doi:10.3109/13693786.2010.508469.

Hrdy, I., Hirt, R. P., Dolezal, P., Bardonová, L., Foster, P. G., Tachezy, J., et al. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432, 618–622. doi:10.1038/nature03149.

Hsieh, P., and Zhang, Y. (2017). The Devil is in the details for DNA mismatch repair. *Proc. Natl. Acad. Sci.* 114, 3552 LP – 3554. doi:10.1073/pnas.1702747114.

Hu, G., Wang, J., Choi, J., Jung, W. H., Liu, I., Litvintseva, A. P., et al. (2011). Variation in chromosome copy number influences the virulence of Cryptococcus neoformans and occurs in isolates from AIDS patients. *BMC Genomics* 12, 526. doi:10.1186/1471-2164-12-526.

Hu, W., Sillaots, S., Lemieux, S., Davison, J., Kauffman, S., Breton, A., et al. (2007). Essential Gene Identification and Drug Target Prioritization in Aspergillus fumigatus. *PLoS Pathog.* 3, e24. doi:10.1371/journal.ppat.0030024.

Huang, M.-E., de Calignon, A., Nicolas, A., and Galibert, F. (2000). POL32 , a subunit of the Saccharomyces cerevisiae DNA polymerase δ, defines a link between DNA replication and the mutagenic bypass repair pathway. *Curr. Genet.* 38, 178–187. doi:10.1007/s002940000149.

Huang, M. N., Yu, W., Teoh, W. W., Ardin, M., Jusakul, A., Ng, A. W. T., et al. (2017). Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* 27, 1475–1486. doi:10.1101/gr.220038.116.

Hubert, J., Nuzillard, J.-M., and Renault, J.-H. (2017). Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochem. Rev.*

16, 55–95. doi:10.1007/s11101-015-9448-7.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33, 1635–1638. doi:10.1093/molbev/msw046.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi:10.1038/nmicrobiol.2016.48.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* 14, R47. doi:10.1186/gb-2013-14-5-r47.

Hunter, A. J., Jin, B., and Kelly, J. M. (2011). Independent duplications of alpha-amylase in different strains of Aspergillus oryzae. *Fungal Genet Biol* 48, 438–444. doi:10.1016/j.fgb.2011.01.006.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73. doi:btb043 [pii].

Hyma, K. E., Saerens, S. M., Verstrepen, K. J., and Fay, J. C. (2011). Divergence in wine characteristics produced by wild and domesticated strains of Saccharomyces cerevisiae. *FEMS Yeast Res.* 11, 540–551. doi:10.1111/j.1567-1364.2011.00746.x.

Ikehata, H., and Ono, T. (2011). The Mechanisms of UV Mutagenesis. *J. Radiat. Res.* 52, 115–125. doi:10.1269/jrr.10175.

Inderbitzin, P., Davis, R. M., Bostock, R. M., and Subbarao, K. V. (2011). The Ascomycete Verticillium longisporum Is a Hybrid and a Plant Pathogen with an Expanded Host Range. *PLoS One* 6, e18260. doi:10.1371/journal.pone.0018260.

Inuma, T., Khodaparast, S. A., and Takamatsu, S. (2007). Multilocus phylogenetic analyses within Blumeria graminis, a powdery mildew fungus of cereals. *Mol. Phylogenet. Evol.* 44, 741–751. doi:https://doi.org/10.1016/j.ympev.2007.01.007.

Ioannidis, P., Simao, F. A., Waterhouse, R. M., Manni, M., Seppey, M., Robertson, H. M., et al. (2017). Genomic features of the damselfly Calopteryx splendens representing a sister clade to most insect orders. *Genome Biol. Evol.* 9, 415–430. doi:10.1093/gbe/evx006.

Ishikawa, M., Ninomiya, T., Akabane, H., Kushida, N., Tsujiuchi, G., Ohyama, M., et al. (2009). Pseurotin A and its analogues as inhibitors of immunoglobuline E production. *Bioorg. Med. Chem. Lett.* 19, 1457–1460. doi:10.1016/j.bmcl.2009.01.029.

Ito, T., and Masubuchi, M. (2014). Dereplication of microbial extracts and related analytical technologies. *J. Antibiot. (Tokyo).* 67, 353–360. doi:10.1038/ja.2014.12.

J. Sambrook, D.W. Russell (2001). *Molecular cloning : a laboratory manual, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.* doi:10.3724/SP.J.1141.2012.01075.

Jackson, R. W., Johnson, L. J., Clarke, S. R., and Arnold, D. L. (2011). Bacterial pathogen evolution: breaking news. *Trends Genet.* 27, 32–40. doi:10.1016/j.tig.2010.10.001.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 5114. doi:10.1038/s41467-018-07641-9.

James, T. Y., Stajich, J. E., Hittinger, C. T., and Rokas, A. (2020). Toward a Fully Resolved Fungal Tree of Life. *Annu. Rev. Microbiol.* 74, 291–313. doi:10.1146/annurev-micro-022020-051835.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., et al. (2014). Whole-genome

analyses resolve early branches in the tree of life of modern birds. *Science (80-.).* 346, 1320–1331. doi:10.1126/science.1253451.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8, 14061. doi:10.1038/ncomms14061.

Jeffares, D. C., Rallis, C., Rieux, A., Speed, D., Převorovský, M., Mourier, T., et al. (2015a). The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat. Genet.* 47, 235–41. doi:10.1038/ng.3215.

Jeffares, D. C., Tomiczek, B., Sojo, V., and dos Reis, M. (2015b). "A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome," in, 65–90. doi:10.1007/978-1-4939-1438-8_4.

Jenni, S., and Harrison, S. C. (2018). Structure of the DASH/Dam1 complex shows its role at the yeast kinetochore-microtubule interface. *Science (80-.).* 360, 552–558. doi:10.1126/science.aar6436.

Jindamorakot, S., Ninomiya, S., Limtong, S., Yongmanitchai, W., Tuntirungkij, M., Potacharoen, W., et al. (2009). Three new species of bipolar budding yeasts of the genus Hanseniaspora and its anamorph Kloeckera isolated in Thailand. *FEMS Yeast Res.* 9, 1327–1337. doi:10.1111/j.1567-1364.2009.00568.x.

Johns, A., Scharf, D. H., Gsaller, F., Schmidt, H., Heinekamp, T., Straßburger, M., et al. (2017). A Nonredundant Phosphopantetheinyl Transferase, PptA, Is a Novel Antifungal Target That Directs Secondary Metabolite, Siderophore, and Lysine Biosynthesis in Aspergillus fumigatus and Is Critical for Pathogenicity. *MBio* 8. doi:10.1128/mBio.01504-16.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and

iterative HMM search procedure. *BMC Bioinformatics* 11, 431. doi:10.1186/1471-2105-11-431.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8, 275–282. doi:10.1093/bioinformatics/8.3.275.

Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K. C. H., McGuire, B., Gubler, W. D., et al. (2014). Adaptive genomic structural variation in the grape powdery mildew pathogen, Erysiphe necator. *BMC Genomics* 15, 1081. doi:10.1186/1471-2164-15-1081.

Jordão, A., Vilela, A., and Cosme, F. (2015). From Sugar of Grape to Alcohol of Wine: Sensorial Impact of Alcohol in Wine. *Beverages* 1, 292–310. doi:10.3390/beverages1040292.

Jorgensen, P. (2002). Systematic Identification of Pathways That Couple Cell Growth and Division in Yeast. *Science (80-.).* 297, 395–400. doi:10.1126/science.1070850.

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26, 1669–1670. doi:10.1093/bioinformatics/btq243.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10, 19–31. doi:10.1038/nrg2487.

Käfer, E. (1977). Meiotic and Mitotic Recombination in Aspergillus and Its Chromosomal Aberrations. *Adv. Genet.* doi:10.1016/S0065-2660(08)60245-X.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi:10.1038/nmeth.4285.

Kamei, K., and Watanabe, A. (2005). Aspergillus mycotoxins and their effect on the host. *Med. Mycol.* 43, 95–99. doi:10.1080/13693780500051547.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070.

Kao, K. C., and Sherlock, G. (2008). Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. *Nat. Genet.* 40, 1499–504. doi:10.1038/ng.280.

Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* doi:10.1038/s41576-020-0233-0.

Kassambara, A. (2020). 'ggpubr': "ggplot2" Based Publication Ready Plots. *R Packag. version 0.2.5*.

Kassambara, A., and Mundt, F. (2017). factoextra. *R Packag. v. 1.0.5*.

Kassir, Y., Granot, D., and Simchen, G. (1988). IME1, a positive regulator gene of meiosis in S. cerevisiae. *Cell* 52, 853–62.

Katinka, M. D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., et al. (2001). Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* 414, 450–453. doi:10.1038/35106579.

Kato, N., Suzuki, H., Takagi, H., Asami, Y., Kakeya, H., Uramoto, M., et al. (2009). Identification of Cytochrome P450s Required for Fumitremorgin Biosynthesis in Aspergillus fumigatus. *ChemBioChem* 10, 920–928. doi:10.1002/cbic.200800787.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436.

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J. J., et al. (2019). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* doi:10.1093/nar/gkz882.

Keeling, P. J., and Slamovits, C. H. (2004). Simplicity and complexity of microsporidian genomes. *Eukaryot. Cell* 3, 1363–9. doi:10.1128/EC.3.6.1363-1369.2004.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009a). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201. doi:10.1101/gr.091231.109.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009b). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201. doi:10.1101/gr.091231.109.

Keith, T. P., Green, P., Reeders, S. T., Brown, V. A., Phipps, P., Bricker, A., et al. (1990). Genetic linkage map of 46 DNA markers on human chromosome 16. *Proc Natl Acad Sci U S A* 87, 5754–5758. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2377614.

Keller, N. P. (2017). Heterogeneity confounds establishment of "a" model microbial strain. *MBio* 8. doi:10.1128/mBio.00135-17.

Keller, N. P. (2019). Fungal secondary metabolism: regulation, function and drug discovery. *Nat.*

Rev. Microbiol. 17, 167–180. doi:10.1038/s41579-018-0121-1.

Keller, N. P., Turner, G., and Bennett, J. W. (2005). Fungal secondary metabolism — from biochemistry to genomics. *Nat. Rev. Microbiol.* 3, 937–947. doi:10.1038/nrmicro1286.

Kensche, P. R., Oti, M., Dutilh, B. E., and Huynen, M. A. (2008). Conservation of divergent transcription in fungi. *Trends Genet.* 24, 207–211. doi:10.1016/j.tig.2008.02.003.

Kerr, S. C., Fischer, G. J., Sinha, M., McCabe, O., Palmer, J. M., Choera, T., et al. (2016). FleA Expression in Aspergillus fumigatus Is Recognized by Fucosylated Structures on Mucins and Macrophages to Prevent Lung Infection. *PLOS Pathog.* 12, e1005555. doi:10.1371/journal.ppat.1005555.

Kevei, F., and Peberdy, J. F. (1979). Induced segregation in interspecific hybrids of Aspergillus nidulans and Aspergillus rugulosus obtained by protoplast fusion. *MGG Mol. Gen. Genet.* doi:10.1007/BF00337798.

Kevei, F., and Perberdy, J. F. (1984). Further Studies on Protoplast Fusion and Interspecific Hybridization within the Aspergillus nidulans Group. *Microbiology* 130, 2229–2236. doi:10.1099/00221287-130-9-2229.

Khoufache, K., Puel, O., Loiseau, N., Delaforge, M., Rivollet, D., Coste, A., et al. (2007). Verruculogen associated with Aspergillus fumigatus hyphae and conidia modifies the electrophysiological properties of human nasal epithelial cells. *BMC Microbiol.* 7, 5. doi:10.1186/1471-2180-7-5.

Kim, I. Y., Kwon, H. Y., Park, K. H., and Kim, D. S. (2017). Anaphase-Promoting Complex 7 is a Prognostic Factor in Human Colorectal Cancer. *Ann. Coloproctol.* 33, 139–145. doi:10.3393/ac.2017.33.4.139.

Kim, Y.-M., Poline, J.-B., and Dumas, G. (2018). Experimenting with reproducibility: a case

study of robustness in bioinformatics. *Gigascience* 7. doi:10.1093/gigascience/giy077.

King, N., and Rokas, A. (2017). Embracing Uncertainty in Reconstructing Early Animal Evolution. *Curr. Biol.* 27, R1081–R1088. doi:10.1016/j.cub.2017.08.054.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 31, 151–160. doi:10.1007/BF02109483.

Kjærbølling, I., Vesth, T. C., Frisvad, J. C., Nybo, J. L., Theobald, S., Kuo, A., et al. (2018). Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proc. Natl. Acad. Sci.* 115, E753–E761. doi:10.1073/pnas.1715954115.

Kjærbølling, I., Vesth, T., Frisvad, J. C., Nybo, J. L., Theobald, S., Kildgaard, S., et al. (2020). A comparative genomics study of 23 Aspergillus species from section Flavi. *Nat. Commun.* 11, 1106. doi:10.1038/s41467-019-14051-y.

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., et al. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* 8, 10872. doi:10.1038/s41598-018-28948-z.

Knowles, S. L., Mead, M. E., Silva, L. P., Raja, H. A., Steenwyk, J. L., Goldman, G. H., et al. (2020). Gliotoxin, a Known Virulence Factor in the Major Human Pathogen Aspergillus fumigatus , Is Also Biosynthesized by Its Nonpathogenic Relative Aspergillus fischeri. *MBio* 11. doi:10.1128/mBio.03361-19.

Knowles, S. L., Vu, N., Todd, D. A., Raja, H. A., Rokas, A., Zhang, Q., et al. (2019). Orthogonal Method for Double-Bond Placement via Ozone-Induced Dissociation Mass Spectrometry (OzID-MS). *J. Nat. Prod.* 82, 3421–3431. doi:10.1021/acs.jnatprod.9b00787.

Knox, B. P., Blachowicz, A., Palmer, J. M., Romsdahl, J., Huttenlocher, A., Wang, C. C. C., et

al. (2016). Characterization of Aspergillus fumigatus Isolates from Air and Surfaces of the International Space Station. *mSphere* 1. doi:10.1128/mSphere.00227-16.

Kobayashi, T., Abe, K., Asai, K., Gomi, K., Juvvadi, P. R., Kato, M., et al. (2007). Genomics of Aspergillus oryzae. *Biosci. Biotechnol. Biochem.* 71, 646–670. doi:10.1271/bbb.60550.

Kobert, K., Salichos, L., Rokas, A., and Stamatakis, A. (2016). Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol. Biol. Evol.* 33, 1606–1617. doi:10.1093/molbev/msw040.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., et al. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285. doi:10.1093/bioinformatics/btp373.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22, 568–576. doi:10.1101/gr.129684.111.

Kocot, K. M., Citarella, M. R., Moroz, L. L., and Halanych, K. M. (2013). PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinforma.* 2013, 429–435. doi:10.4137/EBO.S12813.

Kocsubé, S., Perrone, G., Magistà, D., Houbraken, J., Varga, J., Szigeti, G., et al. (2016). Aspergillus is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles. *Stud. Mycol.* 85, 199–213. doi:10.1016/j.simyco.2016.11.006.

Koehler, P., Cornely, O. A., Böttiger, B. W., Dusse, F., Eichenauer, D. A., Fuchs, F., et al. (2020). COVID-19 associated pulmonary aspergillosis. *Mycoses* 63, 528–534. doi:10.1111/myc.13096.

Kolde, R. (2012). Package `pheatmap'. *Bioconductor*, 1–6.

Kolora, S. R. R., Owens, G. L., Vazquez, J. M., Stubbs, A., Chatla, K., Jainese, C., et al. (2021). Origins and evolution of extreme life span in Pacific Ocean rockfishes. *Science (80-.).* 374, 842–847. doi:10.1126/science.abg5332.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5, 59. doi:10.1186/1471-2105-5-59.

Koschwanez, J. H., Foster, K. R., and Murray, A. W. (2011). Sucrose Utilization in Budding Yeast as a Model for the Origin of Undifferentiated Multicellularity. *PLoS Biol.* 9, e1001122. doi:10.1371/journal.pbio.1001122.

Kosiol, C., and Goldman, N. (2005). Different Versions of the Dayhoff Rate Matrix. *Mol. Biol. Evol.* 22, 193–199. doi:10.1093/molbev/msi005.

Kosiol, C., Goldman, N., and H. Buttimore, N. (2004). A new criterion and method for amino acid classification. *J. Theor. Biol.* 228, 97–106. doi:10.1016/j.jtbi.2003.12.010.

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-Driven Gene Loss in Bacteria. *PLoS Genet.* 8, e1002787. doi:10.1371/journal.pgen.1002787.

Kowalski, C. H., Beattie, S. R., Fuller, K. K., McGurk, E. A., Tang, Y.-W., Hohl, T. M., et al. (2016). Heterogeneity among Isolates Reveals that Fitness in Low Oxygen Correlates with Aspergillus fumigatus Virulence. *MBio* 7. doi:10.1128/mBio.01515-16.

Kowalski, C. H., Kerkaert, J. D., Liu, K.-W., Bond, M. C., Hartmann, R., Nadell, C. D., et al. (2019). Fungal biofilm morphology impacts hypoxia fitness and disease progression. *Nat. Microbiol.* 4, 2430–2441. doi:10.1038/s41564-019-0558-7.

Kowalski, C. H., Morelli, K. A., Stajich, J. E., Nadell, C. D., and Cramer, R. A. (2021). A heterogeneously expressed gene family modulates the biofilm architecture and hypoxic growth of aspergillus fumigatus. *MBio*. doi:10.1128/mBio.03579-20.

Krassowski, T., Coughlan, A. Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D. A., et al. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nat. Commun.* 9, 1887. doi:10.1038/s41467-018-04374-7.

Kriventseva, E. V, Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., et al. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47, D807–D811. doi:10.1093/nar/gky1053.

Kück, P., and Longo, G. C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.* 11, 81. doi:10.1186/s12983-014-0081-x.

Kulkarni, N., Alessandrì, L., Panero, R., Arigoni, M., Olivero, M., Ferrero, G., et al. (2018). Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics* 19, 349. doi:10.1186/s12859-018-2296-x.

Kumar, S., and Dudley, J. (2007). Bioinformatics software for biologists in the genomics era. *Bioinformatics* 23, 1713–1717. doi:10.1093/bioinformatics/btm239.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2012). Statistics and Truth in Phylogenomics. *Mol. Biol. Evol.* 29, 457–472. doi:10.1093/molbev/msr202.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* doi:10.1093/molbev/msw054.

Kunkel, T. A., and Erie, D. A. (2005). DNA MISMATCH REPAIR. *Annu. Rev. Biochem.* 74, 681–710. doi:10.1146/annurev.biochem.74.082803.133243.

Kurtzman, C. P., and Fell, J. W. (1998). *The Yeasts - A Taxonomic Study*. 4th ed. , eds. C. P.

Kurtzman and J. W. Fell Elsevier Science.

Kvalheim, O. M., Chan, H., Benzie, I. F. F., Szeto, Y., Tzang, A. H., Mok, D. K., et al. (2011). Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. *Chemom. Intell. Lab. Syst.* 107, 98–105. doi:10.1016/j.chemolab.2011.02.002.

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., and Rokas, A. (2019). Variation and selection on codon usage bias across an entire subphylum. *PLOS Genet.* 15, e1008304. doi:10.1371/journal.pgen.1008304.

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., and Rokas, A. (2021). Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. *PLOS Biol.* 19, e3001185. doi:10.1371/journal.pbio.3001185.

Lake, J. A. (1991). The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a040654.

Lambou, K., Lamarre, C., Beau, R., Dufour, N., and Latge, J.-P. (2010). Functional analysis of the superoxide dismutase family in Aspergillus fumigatus. *Mol. Microbiol.* 75, 910–923. doi:10.1111/j.1365-2958.2009.07024.x.

Lang, G. I., Parsons, L., and Gammie, A. E. (2013). Mutation Rates, Spectra, and Genome-Wide Distribution of Spontaneous Mutations in Mismatch Repair Deficient Yeast. *G3 Genes|Genomes|Genetics* 3, 1453 LP – 1465. doi:10.1534/g3.113.006429.

Langenberg, A.-K., Bink, F. J., Wolff, L., Walter, S., von Wallbrunn, C., Grossmann, M., et al. (2017). Glycolytic Functions Are Conserved in the Genome of the Wine Yeast Hanseniaspora uvarum, and Pyruvate Kinase Limits Its Capacity for Alcoholic Fermentation. *Appl. Environ. Microbiol.* 83. doi:10.1128/AEM.01580-17.

Langford, B. J., So, M., Raybardhan, S., Leung, V., Westwood, D., MacFadden, D. R., et al. (2020). Bacterial co-infection and secondary infection in patients with COVID-19: a living rapid review and meta-analysis. *Clin. Microbiol. Infect.* doi:10.1016/j.cmi.2020.07.016.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Lanyon, S. M. (1988). The Stochastic Mode of Molecular Evolution: What Consequences for Systematic Investigations? *Auk* 105, 565–573. doi:10.1093/auk/105.3.565.

Latgé, J.-P., and Chamilos, G. (2019). Aspergillus fumigatus and Aspergillosis in 2019. *Clin. Microbiol. Rev.* 33. doi:10.1128/CMR.00140-18.

Lau, A., and Vande Moere, A. (2007). Towards a Model of Information Aesthetics in Information Visualization. in *2007 11th International Conference Information Visualization (IV '07)* (IEEE), 87–92. doi:10.1109/IV.2007.114.

Laumer, C. E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., et al. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B Biol. Sci.* 286, 20190831. doi:10.1098/rspb.2019.0831.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi:10.18637/jss.v025.i01.

Le, S. Q., and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* 25, 1307–1320. doi:10.1093/molbev/msn067.

Lee, M. J., Liu, H., Barker, B. M., Snarr, B. D., Gravelat, F. N., Al Abdallah, Q., et al. (2015). The Fungal Exopolysaccharide Galactosaminogalactan Mediates Virulence by Enhancing Resistance to Neutrophil Extracellular Traps. *PLOS Pathog.* 11, e1005187. doi:10.1371/journal.ppat.1005187.

Lessing, F., Kniemeyer, O., Wozniok, I., Loeffler, J., Kurzai, O., Haertl, A., et al. (2007). The Aspergillus fumigatus Transcriptional Regulator AfYap1 Represents the Major Regulator for Defense against Reactive Oxygen Intermediates but Is Dispensable for Pathogenicity in an Intranasal Mouse Infection Model. *Eukaryot. Cell* 6, 2290–2302. doi:10.1128/EC.00267-07.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi:10.1093/nar/gkz239.

Li, B., Zong, Y., Du, Z., Chen, Y., Zhang, Z., Qin, G., et al. (2015). Genomic Characterization Reveals Insights Into Patulin Biosynthesis and Pathogenicity in Penicillium Species. *Mol. Plant-Microbe Interact.* 28, 635–647. doi:10.1094/MPMI-12-14-0398-FI.

Li, F., Wang, B., Wang, L., and Cao, B. (2014). Phylogenetic Analyses on the Diversity of Aspergillus fumigatus Sensu Lato Based on Five Orthologous Loci. *Mycopathologia* 178, 163–176. doi:10.1007/s11046-014-9790-0.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, L., Hsiang, T., Li, Q., Wang, L., and Yu, Z. (2018). Draft Genome Sequence of NRRL 5109, an Ex-Type Isolate of Aspergillus neoellipticus. *Microbiol. Resour. Announc.* 7. doi:10.1128/MRA.01262-18.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503.

Li, M., Petteys, B. J., McClure, J. M., Valsakumar, V., Bekiranov, S., Frank, E. L., et al. (2010a). Thiamine biosynthesis in Saccharomyces cerevisiae is regulated by the NAD+-dependent histone deacetylase Hst1. *Mol. Cell. Biol.* 30, 3329–41. doi:10.1128/MCB.01590-09.

Li, X.-J., Zhang, Q., Zhang, A.-L., and Gao, J.-M. (2012). Metabolites from Aspergillus fumigatus, an endophytic fungus associated with Melia azedarach, and their antifungal, antifeedant, and toxic activities. *J. Agric. Food Chem.* 60, 3424–31. doi:10.1021/jf300146n.

Li, Y., Liu, Z., Shi, P., and Zhang, J. (2010b). The hearing gene Prestin unites echolocating bats and whales. *Curr. Biol.* 20, R55–R56. doi:10.1016/j.cub.2009.11.042.

Li, Y., Steenwyk, J. L., Chang, Y., Wang, Y., James, T. Y., Stajich, J. E., et al. (2020). A genome-scale phylogeny of Fungi; insights into early evolution, radiations, and the relationship between taxonomy and phylogeny. *bioRxiv*, 2020.08.23.262857. doi:10.1101/2020.08.23.262857.

Li, Y., Steenwyk, J. L., Chang, Y., Wang, Y., James, T. Y., Stajich, J. E., et al. (2021). A genome-scale phylogeny of the kingdom Fungi. *Curr. Biol.* 31, 1653-1665.e5. doi:10.1016/j.cub.2021.01.074.

Li, Z., De La Torre, A. R., Sterck, L., Cánovas, F. M., Avila, C., Merino, I., et al. (2017). Single-Copy Genes as Molecular Markers for Phylogenomic Studies in Seed Plants. *Genome Biol. Evol.* 9, 1130–1147. doi:10.1093/gbe/evx070.

Li, Z., Peng, C., Shen, Y., Miao, X., Zhang, H., and Lin, H. (2008). l,l-Diketopiperazines from Alcaligenes faecalis A72 associated with South China Sea sponge Stelletta tenuis. *Biochem. Syst. Ecol.* 36, 230–234. doi:10.1016/j.bse.2007.08.007.

Lin, X., Patel, S., Litvintseva, A. P., Floyd, A., Mitchell, T. G., and Heitman, J. (2009). Diploids in the Cryptococcus neoformans Serotype A Population Homozygous for the α Mating Type Originate via Unisexual Mating. *PLoS Pathog.* 5, e1000283. doi:10.1371/journal.ppat.1000283.

Lin, Z., and Li, W. H. (2011). Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts. *Mol. Biol. Evol.* 28, 131–142. doi:10.1093/molbev/msq184.

Lind, A. L., Smith, T. D., Saterlee, T., Calvo, A. M., and Rokas, A. (2016). Regulation of secondary metabolism by the velvet complex is temperature-responsive in aspergillus. *G3 Genes, Genomes, Genet.* doi:10.1534/g3.116.033084.

Lind, A. L., Wisecaver, J. H., Lameiras, C., Wiemann, P., Palmer, J. M., Keller, N. P., et al. (2017). Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol.* 15. doi:10.1371/journal.pbio.2003583.

Lind, A. L., Wisecaver, J. H., Smith, T. D., Feng, X., Calvo, A. M., and Rokas, A. (2015). Examining the evolution of the regulatory circuit controlling secondary metabolism and development in the fungal genus Aspergillus. *PLoS Genet.* 11, e1005096. doi:10.1371/journal.pgen.1005096.

Lindahl, T. (1999). Quality Control by DNA Repair. *Science (80-.).* 286, 1897–1905. doi:10.1126/science.286.5446.1897.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., et al. (2009). Population genomics of domestic and wild yeasts. *Nature* 458, 337–41. doi:10.1038/nature07743.

Liu, D., Zhang, R., Yang, X., Wu, H., Xu, D., Tang, Z., et al. (2011). Thermostable cellulase

production of Aspergillus fumigatus Z5 under solid-state fermentation and its application in degradation of agricultural wastes. *Int. Biodeterior. Biodegradation* 65, 717–725. doi:10.1016/j.ibiod.2011.04.005.

Liu, L., Zhang, J., Rheindt, F. E., Lei, F., Qu, Y., Wang, Y., et al. (2017). Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci.* 114, E7282–E7290. doi:10.1073/pnas.1616744114.

Lo, W. S., and Dranginis, A. M. (1996). FLO11, a yeast gene related to the STA genes, encodes a novel cell surface flocculin. *J. Bacteriol.* 178, 7144–7151.

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bähler, J., et al. (2019). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res.* 47, D821–D827. doi:10.1093/nar/gky961.

Losada, L., Ajayi, O., Frisvad, J. C., Yu, J., and Nierman, W. C. (2009). Effect of competition on the production and activity of secondary metabolites in Aspergillus species. *Med. Mycol.* 47, S88–S96. doi:10.1080/13693780802409542.

Losada, L., Sugui, J. A., Eckhaus, M. A., Chang, Y. C., Mounaud, S., Figat, A., et al. (2015). Genetic Analysis Using an Isogenic Mating Pair of Aspergillus fumigatus Identifies Azole Resistance Genes and Lack of MAT Locus's Role in Virulence. *PLoS Pathog.* doi:10.1371/journal.ppat.1004834.

Lubelsky, Y., Reuven, N., and Shaul, Y. (2005). Autorepression of Rfx1 Gene Expression: Functional Conservation from Yeast to Humans in Response to DNA Replication Arrest. *Mol. Cell. Biol.* 25, 10665–10673. doi:10.1128/MCB.25.23.10665-10673.2005.

Lujan, S. A., Williams, J. S., Pursell, Z. F., Abdulovic-Cui, A. A., Clark, A. B., Nick McElhinny,

S. A., et al. (2012). Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity. *PLoS Genet.* 8, e1003016. doi:10.1371/journal.pgen.1003016.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1, 18. doi:10.1186/2047-217X-1-18.

Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476, 442–445. doi:10.1038/nature10291.

Lupetti, A., Danesi, R., Campa, M., Tacca, M. Del, and Kelly, S. (2002). Molecular basis of resistance to azole antifungals. *Trends Mol. Med.* 8, 76–81. doi:10.1016/S1471-4914(02)02280-3.

Lupski, J. R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* 1, e49. doi:10.1371/journal.pgen.0010049.

Lustig, A. J. (2001). Cdc13 subcomplexes regulate multiple telomere functions. *Nat. Struct. Biol.* 8, 297–9. doi:10.1038/86157.

Lynch, M. (2007). *The Origins of Genome Architecture*. doi:10.1093/jhered/esm073.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci.* 107, 961–968. doi:10.1073/pnas.0912629107.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., et al. (2008a). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci.* doi:10.1073/pnas.0803466105.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., et al. (2008b). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad.*

Sci. 105, 9272 LP – 9277. doi:10.1073/pnas.0803466105.

M.Victoria, M.-A., and M. Carmen, P. (2013). Wine Chemistry and Biochemistry. *J. Chem. Inf. Model.* 53, 1689–1699. doi:10.1017/CBO9781107415324.004.

Ma, L.-J., Geiser, D. M., Proctor, R. H., Rooney, A. P., O'Donnell, K., Trail, F., et al. (2013). Fusarium Pathogenomics. *Annu. Rev. Microbiol.* 67, 399–416. doi:10.1146/annurev-micro-092412-155650.

Ma, Y. ., Li, Y., Liu, J. ., Song, Y. ., and Tan, R. . (2004). Anti-Helicobacter pylori metabolites from Rhizoctonia sp. Cy064, an endophytic fungus in Cynodon dactylon. *Fitoterapia* 75, 451–456. doi:10.1016/j.fitote.2004.03.007.

Mable, B. K., Alexandrou, M. A., and Taylor, M. I. (2011). Genome duplication in amphibians and fish: an extended synthesis. *J. Zool.* 284, 151–182. doi:10.1111/j.1469-7998.2011.00829.x.

Macdonald, D., Thomson, D. D., Johns, A., Contreras Valenzuela, A., Gilsenan, J. M., Lord, K. M., et al. (2018). Inducible Cell Fusion Permits Use of Competitive Fitness Profiling in the Human Pathogenic Fungus Aspergillus fumigatus. *Antimicrob. Agents Chemother.* 63. doi:10.1128/AAC.01615-18.

Macheleidt, J., Mattern, D. J., Fischer, J., Netzker, T., Weber, J., Schroeckh, V., et al. (2016). Regulation and Role of Fungal Secondary Metabolites. *Annu. Rev. Genet.* 50, 371–392. doi:10.1146/annurev-genet-120215-035203.

MacIntyre, C. R., Chughtai, A. A., Barnes, M., Ridda, I., Seale, H., Toms, R., et al. (2018). The role of pneumonia and secondary bacterial infection in fatal and serious outcomes of pandemic influenza a(H1N1)pdm09. *BMC Infect. Dis.* 18, 637. doi:10.1186/s12879-018-3548-0.

Madden, T. (2013). The BLAST sequence analysis tool. *BLAST Seq. Anal. Tool*, 1–17.

Magan, N., and Lacey, J. (1984). Effects of gas composition and water activity on growth of field and storage fungi and their interactions. *Trans. Br. Mycol. Soc.* 82, 305–314. doi:10.1016/S0007-1536(84)80074-1.

Magditch, D. A., Liu, T. B., Xue, C., and Idnurm, A. (2012). DNA Mutations Mediate Microevolution between Host-Adapted Forms of the Pathogenic Fungus Cryptococcus neoformans. *PLoS Pathog.* doi:10.1371/journal.ppat.1002936.

Malavazi, I., and Goldman, G. H. (2012). "Gene Disruption in Aspergillus fumigatus Using a PCR-Based Strategy and In Vivo Recombination in Yeast," in, 99–118. doi:10.1007/978-1-61779-539-8_7.

Malnic, B., Godfrey, P. A., and Buck, L. B. (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci.* 101, 2584–2589. doi:10.1073/pnas.0307882100.

Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M., et al. (2020). GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* 21, 193. doi:10.1186/s12864-020-6568-2.

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019a). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20, 47. doi:10.1186/s13059-019-1649-8.

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019b). Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLOS Biol.* 17, e3000333. doi:10.1371/journal.pbio.3000333.

Manolio, T. A., Bult, C. J., Chisholm, R. L., Deverka, P. A., Ginsburg, G. S., Goldrich, M., et al. (2021). Genomic medicine year in review: 2021. *Am. J. Hum. Genet.* 108, 2210–2214.

doi:10.1016/j.ajhg.2021.11.006.

Marcet-Houben, M., Ballester, A.-R., de la Fuente, B., Harries, E., Marcos, J. F., González-Candelas, L., et al. (2012). Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest pathogen of citrus. *BMC Genomics* 13, 646. doi:10.1186/1471-2164-13-646.

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biol.* 13, e1002220. doi:10.1371/journal.pbio.1002220.

Marín, Sanchis, Sáenz, Ramos, Vinas, and Magan (1998). Ecological determinants for germination and growth of some Aspergillus and Penicillium spp. from maize grain. *J. Appl. Microbiol.* 84, 25–36. doi:10.1046/j.1365-2672.1997.00297.x.

Marks, V. D., Ho Sui, S. J., Erasmus, D., Van Der Merwe, G. K., Brumm, J., Wasserman, W. W., et al. (2008). Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response. *FEMS Yeast Res.* 8, 35–52. doi:10.1111/j.1567-1364.2007.00338.x.

Marsit, S., and Dequin, S. (2015). Diversity and adaptive evolution of Saccharomyces wine yeast: a review. *FEMS Yeast Res.*, 1–12. doi:10.1093/femsyr/fov067.

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W., et al. (2011). The functional spectrum of low-frequency coding variation. *Genome Biol.* 12, R84. doi:10.1186/gb-2011-12-9-r84.

Marti, T. M., Kunz, C., and Fleck, O. (2002). DNA mismatch repair and mutation avoidance pathways. *J. Cell. Physiol.* 191, 28–41. doi:https://doi.org/10.1002/jcp.10077.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., et al.

(2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* 23, 1817–1828. doi:10.1101/gr.159426.113.

Martin, V., Valera, M., Medina, K., Boido, E., and Carrau, F. (2018). Oenological Impact of the Hanseniaspora/Kloeckera Yeast Genus on Wines—A Review. *Fermentation* 4, 76. doi:10.3390/fermentation4030076.

Matplotlib Org. (2019). Matplotlib: Python plotting. *Matplotlib*.

Mattern, D. J., Schoeler, H., Weber, J., Novohradská, S., Kraibooj, K., Dahse, H.-M., et al. (2015). Identification of the antiphagocytic trypacidin gene cluster in the human-pathogenic fungus Aspergillus fumigatus. *Appl. Microbiol. Biotechnol.* 99, 10151–10161. doi:10.1007/s00253-015-6898-1.

McDonagh, A., Fedorova, N. D., Crabtree, J., Yu, Y., Kim, S., Chen, D., et al. (2008). Sub-Telomere Directed Gene Expression during Initiation of Invasive Aspergillosis. *PLoS Pathog.* 4, e1000154. doi:10.1371/journal.ppat.1000154.

McDonald, M. J., Hsieh, Y.-Y., Yu, Y.-H., Chang, S.-L., and Leu, J.-Y. (2012). The Evolution of Low Mutation Rates in Experimental Mutator Populations of Saccharomyces cerevisiae. *Curr. Biol.* 22, 1235–1240. doi:10.1016/j.cub.2012.04.056.

McInerney, J. O. (1998). GCUA: general codon usage analysis. *Bioinformatics* 14, 372–373. doi:10.1093/bioinformatics/14.4.372.

Mead, M. E., Borowsky, A. T., Joehnk, B., Steenwyk, J. L., Shen, X.-X., Sil, A., et al. (2020). Recurrent Loss of abaA, a Master Regulator of Asexual Development in Filamentous Fungi, Correlates with Changes in Genomic and Morphological Traits. *Genome Biol. Evol.* 12, 1119–1130. doi:10.1093/gbe/evaa107.

Mead, M. E., Knowles, S. L., Raja, H. A., Beattie, S. R., Kowalski, C. H., Steenwyk, J. L., et al.

(2018). Characterizing the pathogenic, genomic, and chemical traits of Aspergillus fischeri, the closest sequenced relative of the major human fungal pathogen Aspergillus fumigatus. *bioRxiv*. doi:https://doi.org/10.1101/430728.

Mead, M. E., Knowles, S. L., Raja, H. A., Beattie, S. R., Kowalski, C. H., Steenwyk, J. L., et al. (2019a). Characterizing the Pathogenic, Genomic, and Chemical Traits of Aspergillus fischeri , a Close Relative of the Major Human Fungal Pathogen Aspergillus fumigatus. *mSphere* 4. doi:10.1128/mSphere.00018-19.

Mead, M. E., Raja, H. A., Steenwyk, J. L., Knowles, S. L., Oberlies, N. H., and Rokas, A. (2019b). Draft Genome Sequence of the Griseofulvin-Producing Fungus Xylaria flabelliformis Strain G536. *Microbiol. Resour. Announc.* 8. doi:10.1128/MRA.00890-19.

Mead, M. E., Steenwyk, J. L., Silva, L. P., de Castro, P. A., Saeed, N., Hillmann, F., et al. (2021). An evolutionary genomic approach reveals both conserved and species-specific genetic elements related to human disease in closely related Aspergillus fungi. *Genetics* 218. doi:10.1093/genetics/iyab066.

Medina, E. M., Turner, J. J., Gordân, R., Skotheim, J. M., and Buchler, N. E. (2016). Punctuated evolution and transitional hybrid network in an ancestral cell cycle of fungi. *Elife* 5. doi:10.7554/eLife.09492.

Menardo, F., Praz, C. R., Wyder, S., Ben-David, R., Bourras, S., Matsumae, H., et al. (2016). Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. *Nat. Genet.* 48, 201–205. doi:10.1038/ng.3485.

Mesquite Project Team (2014). Mesquite: A modular system for evolutionary analysis. *Available from http// mesquiteproject.wikispaces.com/home*. doi:10.1017/CBO9781107415324.004.

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene

function analysis with the PANTHER classification system. *Nat. Protoc.* 8, 1551–66. doi:10.1038/nprot.2013.092.

Miao, Y., Liu, D., Li, G., Li, P., Xu, Y., Shen, Q., et al. (2015). Genome-wide transcriptomic analysis of a superior biomass-degrading strain of A. fumigatus revealed active lignocellulose-degrading genes. *BMC Genomics* 16, 459. doi:10.1186/s12864-015-1658-2.

Michels, C. A., Read, E., Nat, K., and Charron, M. J. (1992). The telomere-associated MAL3 locus of Saccharomyces is a tandem array of repeated genes. *Yeast* 8, 655–665. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1441745.

Midgley, D. J., Rosewarne, C. P., Greenfield, P., Li, D., Vockler, C. J., Hitchcock, C. J., et al. (2016). Genomic insights into the carbohydrate catabolism of Cairneyella variabilis gen. nov. sp. nov., the first reports from a genome of an ericoid mycorrhizal fungus from the southern hemisphere. *Mycorrhiza* 26, 345–352. doi:10.1007/s00572-016-0683-6.

Milo, S., Misgav, R. H., Hazkani-Covo, E., and Covo, S. (2019). Limited DNA repair gene repertoire in Ascomycete yeast revealed by comparative genomics. *Genome Biol. Evol.* doi:10.1093/gbe/evz242.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534. doi:10.1093/molbev/msaa015.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52.

doi:10.1093/bioinformatics/btv234.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* 346, 763–767. doi:10.1126/science.1257570.

Mitsuguchi, H., Seshime, Y., Fujii, I., Shibuya, M., Ebizuka, Y., and Kushiro, T. (2009). Biosynthesis of Steroidal Antibiotic Fusidanes: Functional Analysis of Oxidosqualene Cyclase and Subsequent Tailoring Enzymes from Aspergillus fumigatus. *J. Am. Chem. Soc.* 131, 6402–6411. doi:10.1021/ja8095976.

Mixão, V., and Gabaldón, T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast* 35, 5–20. doi:10.1002/yea.3242.

Mixão, V., and Gabaldón, T. (2020). Genomic evidence for a hybrid origin of the yeast opportunistic pathogen Candida albicans. *BMC Biol.* 18, 48. doi:10.1186/s12915-020-00776-6.

Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sánchez-García, M., et al. (2020). Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nat. Commun.* 11, 5125. doi:10.1038/s41467-020-18795-w.

Moere, A. Vande, and Purchase, H. (2011). On the role of design in information visualization. *Inf. Vis.* 10, 356–371. doi:10.1177/1473871611415996.

Molina, A. M., Swiegers, J. H., Varela, C., Pretorius, I. S., and Agosin, E. (2007). Influence of wine fermentation temperature on the synthesis of yeast-derived volatile aroma compounds. *Appl. Microbiol. Biotechnol.* 77, 675–687. doi:10.1007/s00253-007-1194-3.

Möller, M., Habig, M., Lorrain, C., Feurtey, A., Haueisen, J., Fagundes, W. C., et al. (2021). Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and

changes evolutionary trajectory in a fungal pathogen. *PLOS Genet.* 17, e1009448.

Mondon, P., De Champs, C., Donadille, A., Ambroise-Thomas, P., and Grillot, R. (1996). Variation its virulence of Aspergillus fumigatus strains in a murine model of invasive pulmonary aspergillosis. *J. Med. Microbiol.* doi:10.1099/00222615-45-3-186.

Montero, C. M., Dodero, M. C. R., Sanchez, D. A. G., and Barroso, C. G. (2004). Analysis of low molecular weight carbohydrates in food and beverages: A review. *Chromatographia*. doi:10.1365/s10337-003-0134-3.

Moran, G. P., Coleman, D. C., and Sullivan, D. J. (2011). Comparative Genomics and the Evolution of Pathogenicity in Human Pathogenic Fungi. *Eukaryot. Cell* 10, 34–42. doi:10.1128/EC.00242-10.

Morel, B., Schade, P., Lutteropp, S., Williams, T. A., Szöllősi, G. J., and Stamatakis, A. (2021). SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *bioRxiv*, 2021.03.29.437460. doi:10.1101/2021.03.29.437460.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., and Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics* 24, 1757–1764. doi:10.1093/bioinformatics/btn322.

Mortimer, R. K. (2000). Evolution and variation of the yeast (Saccharomyces) genome. *Genome Res.* 10, 403–409. doi:10.1101/gr.10.4.403.

Mount, D. W. (2008). Using BLOSUM in Sequence Alignments. *Cold Spring Harb. Protoc.* 2008, pdb.top39-pdb.top39. doi:10.1101/pdb.top39.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 17, 53.

doi:10.1186/s13059-016-0917-0.

Murphy, W. J. (2001). Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science (80-.).* 294, 2348–2351. doi:10.1126/science.1067179.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618. doi:10.1038/35054550.

Myung, K., Datta, A., and Kolodner, R. D. (2001). Suppression of Spontaneous Chromosomal Rearrangements by S Phase Checkpoint Functions in Saccharomyces cerevisiae. *Cell* 104, 397–408. doi:10.1016/S0092-8674(01)00227-6.

Nabhan, A. R., and Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* 13, 122–134. doi:10.1093/bib/bbr014.

Nag, D. K., Koonce, M. P., and Axelrod, J. (1997). SSP1, a gene necessary for proper completion of meiotic divisions and spore formation in Saccharomyces cerevisiae. *Mol. Cell. Biol.* 17, 7029–39.

Nagy, L. G., Ohm, R. A., Kovács, G. M., Floudas, D., Riley, R., Gácser, A., et al. (2014). Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat. Commun.* 5, 4471. doi:10.1038/ncomms5471.

Nagy, L. G., Riley, R., Tritt, A., Adam, C., Daum, C., Floudas, D., et al. (2016). Comparative Genomics of Early-Diverging Mushroom-Forming Fungi Provides Insights into the Origins of Lignocellulose Decay Capabilities. *Mol. Biol. Evol.* 33, 959–970. doi:10.1093/molbev/msv337.

Nascimento, A. M., Goldman, G. H., Park, S., Marras, S. A. E., Delmas, G., Oza, U., et al.

(2003). Multiple Resistance Mechanisms among Aspergillus fumigatus Mutants with High-Level Resistance to Itraconazole. *Antimicrob. Agents Chemother.* 47, 1719–1726. doi:10.1128/AAC.47.5.1719-1726.2003.

Nasir, N., Farooqi, J., Mahmood, S. F., and Jabeen, K. (2020). COVID-19-associated pulmonary aspergillosis (CAPA) in patients admitted with severe COVID-19 pneumonia: An observational study from Pakistan. *Mycoses* 63, 766–770. doi:10.1111/myc.13135.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* 16, 60–68. doi:10.1038/s41589-019-0400-9.

Neafsey, D. E., Barker, B. M., Sharpton, T. J., Stajich, J. E., Park, D. J., Whiston, E., et al. (2010). Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Res.* 20, 938–946. doi:10.1101/gr.103911.109.

Nenarokova, A., Záhonová, K., Krasilnikova, M., Gahura, O., McCulloch, R., Zíková, A., et al. (2019). Causes and Effects of Loss of Classical Nonhomologous End Joining Pathway in Parasitic Eukaryotes. *MBio* 10, e01541-19. doi:10.1128/mBio.01541-19.

Ngo, H.-P., and Lydall, D. (2010). Survival and Growth of Yeast without Telomere Capping by Cdc13 in the Absence of Sgs1, Exo1, and Rad9. *PLoS Genet.* 6, e1001072. doi:10.1371/journal.pgen.1001072.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Nielsen, J. C., Grijseels, S., Prigent, S., Ji, B., Dainat, J., Nielsen, K. F., et al. (2017). Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite

production in Penicillium species. *Nat. Microbiol.* 2, 17044. doi:10.1038/nmicrobiol.2017.44.

Nielsen, K., De Obaldia, A. L., and Heitman, J. (2007). Cryptococcus neoformans Mates on Pigeon Guano: Implications for the Realized Ecological Niche and Globalization. *Eukaryot. Cell* 6, 949–959. doi:10.1128/EC.00097-07.

Nielsen, K., and Yohalem, D. S. (2001). Origin of a Polyploid Botrytis Pathogen through Interspecific Hybridization between Botrytis aclada and B. byssoidea. *Mycologia* 93, 1064. doi:10.2307/3761668.

Nierman, W. C., Pain, A., Anderson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., et al. (2005). Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* 438, 1151–1156. doi:10.1038/nature04332.

Nierman, W. C., Yu, J., Fedorova-Abrams, N. D., Losada, L., Cleveland, T. E., Bhatnagar, D., et al. (2015). Genome Sequence of Aspergillus flavus NRRL 3357, a Strain That Causes Aflatoxin Contamination of Food and Feed. *Genome Announc.* 3. doi:10.1128/genomeA.00168-15.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res.* 24, 1485–1496. doi:10.1101/gr.169532.113.

Nitiss, K. C., Malik, M., He, X., White, S. W., and Nitiss, J. L. (2006). Tyrosyl-DNA phosphodiesterase (Tdp1) participates in the repair of Top2-mediated DNA damage. *Proc. Natl. Acad. Sci.* 103, 8953–8958. doi:10.1073/pnas.0603455103.

Novick, P., and Schekman, R. (1979). Secretion and cell-surface growth are blocked in a temperature-sensitive mutant of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.*

76, 1858–62. doi:10.1073/pnas.76.4.1858.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., et al. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proc. Natl. Acad. Sci. U. S. A.* 106, 16333–16338. doi:10.1073/pnas.0904673106.

Nunoshiba, T. (2004). A novel Nudix hydrolase for oxidized purine nucleoside triphosphates encoded by ORFYLR151c (PCD1 gene) in Saccharomyces cerevisiae. *Nucleic Acids Res.* 32, 5339–5348. doi:10.1093/nar/gkh868.

O'Donnell, K., Rooney, A. P., Proctor, R. H., Brown, D. W., McCormick, S. P., Ward, T. J., et al. (2013). Phylogenetic analyses of RPB1 and RPB2 support a middle Cretaceous origin for a clade comprising all agriculturally and medically important fusaria. *Fungal Genet. Biol.* 52, 20–31. doi:10.1016/j.fgb.2012.12.004.

O'Hanlon, K. A., Cairns, T., Stack, D., Schrettl, M., Bignell, E. M., Kavanagh, K., et al. (2011). Targeted Disruption of Nonribosomal Peptide Synthetase pes3 Augments the Virulence of Aspergillus fumigatus. *Infect. Immun.* 79, 3978–3992. doi:10.1128/IAI.00192-11.

Ogundero, V. W. (1983). Factors affecting growth and cellulose hydrolysis by the thermotolerant aspergillus nidulans from composts. *Acta Biotechnol.* 3, 65–72. doi:10.1002/abio.370030116.

Olarte, R. A., Worthington, C. J., Horn, B. W., Moore, G. G., Singh, R., Monacell, J. T., et al. (2015). Enhanced diversity and aflatoxigenicity in interspecific hybrids of Aspergillus flavus and Aspergillus parasiticus. *Mol. Ecol.* 24, 1889–1909. doi:10.1111/mec.13153.

Oliver, A. (2000). High Frequency of Hypermutable Pseudomonas aeruginosa in Cystic Fibrosis Lung Infection. *Science (80-.).* 288, 1251–1253. doi:10.1126/science.288.5469.1251.

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi:10.1038/s41586-019-1693-2.

Opulente, D. A., Rollinson, E. J., Bernick-Roehr, C., Hulfachor, A. B., Rokas, A., Kurtzman, C. P., et al. (2018). Factors driving metabolic diversity in the budding yeast subphylum. *BMC Biol.* 16, 26. doi:10.1186/s12915-018-0498-3.

Orozco, L. D., Cokus, S. J., Ghazalpour, A., Ingram-Drake, L., Wang, S., van Nas, A., et al. (2009). Copy number variation influences gene expression and metabolic traits in mice. *Hum. Mol. Genet.* 18, 4118–4129. doi:10.1093/hmg/ddp360.

Ortiz-Merino, R. A., Kuanyshev, N., Braun-Galleani, S., Byrne, K. P., Porro, D., Branduardi, P., et al. (2017). Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a failed mating-type switch. *PLoS Biol.* 15. doi:10.1371/journal.pbio.2002128.

Osherov, N., Kontoyiannis, D. P., Romans, A., and May, G. S. (2001). Resistance to itraconazole in Aspergillus nidulans and Aspergillus fumigatus is conferred by extra copies of the A. nidulans P-450 14alpha-demethylase gene, pdmA. *J. Antimicrob. Chemother*. 48, 75–81. doi:10.1093/jac/48.1.75.

Ough, C. ., Davenport, M., and Joseph, K. (1989). Effects of Certain Vitamins on Growth and Fermentation Rate of Several Commercial Active Dry Wine Yeasts. *Am. J. Enol. Vitic.* 40, 208–213.

Ozcan, S., and Johnston, M. (1999). Function and regulation of yeast hexose transporters. *Microbiol. Mol. Biol. Rev.* 63, 554–69. doi:10.1128/MMBR.63.3.554-569.1999.

Padilla, P. A., Fuge, E. K., Crawford, M. E., Errett, A., and Werner-Washburne, M. (1998). The highly conserved, coregulated SNO and SNZ gene families in Saccharomyces cerevisiae

respond to nutrient limitation. *J. Bacteriol.* 180, 5718–5726.

Pal, C., Maciá, M. D., Oliver, A., Schachar, I., and Buckling, A. (2007). Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature* 450, 1079–1081. doi:10.1038/nature06350.

Panchenko, P. (2006). "Kolmogorov-Smirnov Test," in *SpringerReference* (Berlin/Heidelberg: Springer-Verlag), 83–90. doi:10.1007/SpringerReference_221375.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412.

Paris, S., Wysong, D., Debeaupuis, J.-P., Shibuya, K., Philippe, B., Diamond, R. D., et al. (2003). Catalases of Aspergillus fumigatus. *Infect. Immun.* 71, 3551–3562. doi:10.1128/IAI.71.6.3551-3562.2003.

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.* 101, 9903–9908. doi:10.1073/pnas.0307901101.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441, 1103–1108. doi:10.1038/nature04789.

Paul, S., Million-Weaver, S., Chattopadhyay, S., Sokurenko, E., and Merrikh, H. (2013). Accelerated gene evolution through replication–transcription conflicts. *Nature* 495, 512–515. doi:10.1038/nature11989.

Paul, S., Zhang, A., Ludeña, Y., Villena, G. K., Yu, F., Sherman, D. H., et al. (2017). Insights from the genome of a high alkaline cellulase producing Aspergillus fumigatus strain obtained from Peruvian Amazon rainforest. *J. Biotechnol.* 251, 53–58.

doi:10.1016/j.jbiotec.2017.04.010.

Paulussen, C., Hallsworth, J. E., Álvarez-Pérez, S., Nierman, W. C., Hamill, P. G., Blain, D., et al. (2017). Ecology of aspergillosis: insights into the pathogenic potency of Aspergillus fumigatus and some other Aspergillus species. *Microb. Biotechnol.* 10, 296–322. doi:10.1111/1751-7915.12367.

Payen, C., Sunshine, A. B., Ong, G. T., Pogachar, J. L., Zhao, W., and Dunham, M. J. (2016). High-Throughput Identification of Adaptive Mutations in Experimentally Evolved Yeast Populations. *PLoS Genet* 12, e1006339. Available at: http://dx.doi.org/10.1371%2Fjournal.pgen.1006339.

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Des. Sel.* 14, 609–614. doi:10.1093/protein/14.9.609.

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Curr. Protoc. Bioinforma.* 42, 3.1.1-3.1.8. doi:10.1002/0471250953.bi0301s42.

Pease, J. B., Haak, D. C., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biol.* 14, e1002379. doi:10.1371/journal.pbio.1002379.

Pel, H. J., de Winde, J. H., Archer, D. B., Dyer, P. S., Hofmann, G., Schaap, P. J., et al. (2007). Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nat. Biotechnol.* 25, 221–231. doi:10.1038/nbt1282.

Pennerman, K. K., Yin, G., Glenn, A. E., and Bennett, J. W. (2020). Identifying candidate Aspergillus pathogenicity factors by annotation frequency. *BMC Microbiol.* 20, 342. doi:10.1186/s12866-020-02031-y.

Perlin, D. S., Rautemaa-Richardson, R., and Alastruey-Izquierdo, A. (2017). The global problem

of antifungal resistance: prevalence, mechanisms, and management. *Lancet Infect. Dis.* 17, e383–e392. doi:10.1016/S1473-3099(17)30316-X.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39, 1256–1260. doi:10.1038/ng2123.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556, 339–344. doi:10.1038/s41586-018-0030-5.

Pezer, Z., Harr, B., Teschke, M., Babiker, H., and Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. *Genome Res* 25, 1114–1124. doi:10.1101/gr.187187.114.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., et al. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol.* 9, e1000602. doi:10.1371/journal.pbio.1000602.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* 36, 541–562. doi:10.1146/annurev.ecolsys.35.112202.130205.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., et al. (2009). Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr. Biol.* 19, 706–712. doi:10.1016/j.cub.2009.02.052.

Phillips, M. A., Steenwyk, J. L., Shen, X.-X., and Rokas, A. (2021). Examination of Gene Loss in the DNA Mismatch Repair Pathway and Its Mutational Consequences in a Fungal Phylum. *Genome Biol. Evol.* 13. doi:10.1093/gbe/evab219.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-Scale Phylogeny and the Detection of Systematic Biases. *Mol. Biol. Evol.* 21, 1455–1458. doi:10.1093/molbev/msh137.

Phillips, M. J., Lin, Y.-H., Harrison, G., and Penny, D. (2001). Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. *Proc. R. Soc. London. Ser. B Biol. Sci.* 268, 1533–1538. doi:10.1098/rspb.2001.1677.

Phillips, M. J., and Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* 28, 171–185. doi:10.1016/S1055-7903(03)00057-5.

Pitt, J. I. (1994). The current role of Aspergillus and Penicillium in human and animal health. *Med. Mycol.* 32, 17–32. doi:10.1080/02681219480000701.

Pitt, J. I. (2002). "Biology and Ecology of Toxigenic Penicillium Species," in *Advances in experimental medicine and biology*, 29–41. doi:10.1007/978-1-4615-0629-4_4.

Pitt, J. I., and Hocking, A. D. (2009). *Fungi and Food Spoilage*. Boston: Springer.

Pitt, J. I., and Taylor, J. W. (2014). Aspergillus , its sexual states and the new International Code of Nomenclature. *Mycologia* 106, 1051–1062. doi:10.3852/14-060.

Plomion, C., Aury, J.-M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., et al. (2018). Oak genome reveals facets of long lifespan. *Nat. Plants* 4, 440–452. doi:10.1038/s41477-018-0172-3.

Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11, 395. doi:10.1186/1471-2105-11-395.

Poukka, H., Aarnisalo, P., Santti, H., Jänne, O. A., and Palvimo, J. J. (2000). Coregulator Small Nuclear RING Finger Protein (SNURF) Enhances Sp1- and Steroid Receptor-mediated

Transcription by Different Mechanisms. *J. Biol. Chem.* 275, 571–579. doi:10.1074/jbc.275.1.571.

Pretorius, I. S. (2000). Tailoring wine yeast for the new millennium: Novel approaches to the ancient art of winemaking. *Yeast* 16, 675–729. doi:10.1002/1097-0061(20000615)16:8<675::AID-YEA585>3.0.CO;2-B.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* 5. doi:10.1371/journal.pone.0009490.

Pringle, A., Baker, D. M., Platt, J. L., Wares, J. P., Latgé, J. P., and Taylor, J. W. (2005). Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus Aspergillus fumigatus. *Evolution* 59, 1886–99.

Pryszcz, L. P., Németh, T., Saus, E., Ksiezopolska, E., Hegedűsová, E., Nosek, J., et al. (2015). The Genomic Aftermath of Hybridization in the Opportunistic Pathogen Candida metapsilosis. *PLOS Genet.* 11, e1005626. doi:10.1371/journal.pgen.1005626.

Puértolas-Balint, F., Rossen, J. W. A., Oliveira dos Santos, C., Chlebowicz, M. M. A., Raangs, E. C., van Putten, M. L., et al. (2019). Revealing the Virulence Potential of Clinical and Environmental Aspergillus fumigatus Isolates Using Whole-Genome Sequencing. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.01970.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

R Development Core Team, R. (2008). *Computational Many-Particle Physics*. , eds. H. Fehske, R. Schneider, and A. Weiße Berlin, Heidelberg: Springer Berlin Heidelberg doi:10.1007/978-3-540-74686-7.

Radosa, S., Ferling, I., Sprague, J. L., Westermann, M., and Hillmann, F. (2019). The different

morphologies of yeast and filamentous fungi trigger distinct killing and feeding mechanisms in a fungivorous amoeba. *Environ. Microbiol.* doi:10.1111/1462-2920.14588.

Raffa, N., and Keller, N. P. (2019). A call to arms: Mustering secondary metabolites for success and survival of an opportunistic pathogen. *PLOS Pathog.* 15, e1007606. doi:10.1371/journal.ppat.1007606.

Raftery, A. E., and Lewis, S. M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. *Pract. Markov Chain Monte Carlo* 7, 763–773. doi:10.1.1.41.6352.

Ram, Y., and Hadany, L. (2012). The evolution of stress-induced hypermutation in asexual populations. *Evolution (N. Y).* 66, 2315–2328. doi:10.1111/j.1558-5646.2012.01576.x.

Rambaut, A. (2009). FigTree, a graphical viewer of phylogenetic trees. *Inst. Evol. Biol. Univ. Edinburgh*.

Ramsook, C. B., Tan, C., Garcia, M. C., Fung, R., Soybelman, G., Henry, R., et al. (2010). Yeast cell adhesion molecules have functional amyloid-forming sequences. *Eukaryot. Cell* 9, 393–404. doi:10.1128/EC.00068-09.

Reifenberger, E., Freidel, K., and Ciriacy, M. (1995). Identification of novel HXT genes in Saccharomyces cerevisiae reveals the impact of individual hexose transporters on glycolytic flux. *Mol. Microbiol.* 16, 157–167. doi:10.1111/j.1365-2958.1995.tb02400.x.

Reis, M. d., and Yang, Z. (2011). Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times. *Mol. Biol. Evol.* 28, 2161–2172. doi:10.1093/molbev/msr045.

Remm, M., Storm, C. E. V., and Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.

doi:10.1006/jmbi.2000.5197.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

Rhodes, J., Beale, M. A., Vanhove, M., Jarvis, J. N., Kannambath, S., Simpson, J. A., et al. (2017a). A Population Genomics Approach to Assessing the Genetic Basis of Within-Host Microevolution Underlying Recurrent Cryptococcal Meningitis Infection. *G3 Genes|Genomes|Genetics* 7, 1165–1176. doi:10.1534/g3.116.037499.

Rhodes, J., Desjardins, C. A., Sykes, S. M., Beale, M. A., Vanhove, M., Sakthikumar, S., et al. (2017b). Tracing Genetic Exchange and Biogeography of Cryptococcus neoformans var. grubii at the Global Population Level. *Genetics* 207, 327–346. doi:10.1534/genetics.117.203836.

Richman, S. (2015). Deficient mismatRichman, S. (2015). Deficient mismatch repair: Read all about it (Review). Int J Oncol, 47(4), 1189–1202. https://doi.org/10.3892/ijo.2015.3119ch repair: Read all about it (Review). *Int J Oncol* 47, 1189–1202. doi:10.3892/ijo.2015.3119.

Ries, L. N. A., Steenwyk, J. L., de Castro, P. A., de Lima, P. B. A., Almeida, F., de Assis, L. J., et al. (2019). Nutritional Heterogeneity Among Aspergillus fumigatus Strains Has Consequences for Virulence in a Strain- and Host-Dependent Manner. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.00854.

Rieseberg, L. H. (2003). Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science (80-.).* 301, 1211–1216. doi:10.1126/science.1086949.

Rieseberg, L. H., Kim, S.-C., Randell, R. A., Whitney, K. D., Gross, B. L., Lexer, C., et al. (2007). Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica* 129, 149–165. doi:10.1007/s10709-006-9011-y.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., et al. (2016). Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci.* 113, 9882–9887. doi:10.1073/pnas.1603941113.

Risch, N. (1990). Linkage Strategies for Genetically Complex Traits. II. The Power of Affected Relative Pairs. *Am. J. Hum. Genet* 46, 229–241. doi:10.2460/javma.239.10.1288.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A. J., Gray, M. W., Philippe, H., et al. (2007). Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. *Curr. Biol.* 17, 1420–1425. doi:10.1016/j.cub.2007.07.036.

Rodríguez-Navarro, S., Llorente, B., Rodríguez-Manzaneque, M. T., Ramne, A., Uber, G., Marchesan, D., et al. (2002). Functional analysis of yeast gene families involved in metabolism of vitamins B1 and B6. *Yeast* 19, 1261–1276. doi:10.1002/yea.916.

Rohlfs, M., Albert, M., Keller, N. P., and Kempken, F. (2007). Secondary chemicals protect mould from fungivory. *Biol. Lett.* 3, 523–525. doi:10.1098/rsbl.2007.0338.

Rohlfs, M., and Churchill, A. C. L. (2011). Fungal secondary metabolites as modulators of interactions with insects and other arthropods. *Fungal Genet. Biol.* 48, 23–34. doi:10.1016/j.fgb.2010.08.008.

Rokas, A., and Abbot, P. (2009). Harnessing genomics for evolutionary insights. *Trends Ecol. Evol.* 24, 192–200. doi:10.1016/j.tree.2008.11.004.

Rokas, A., and Carroll, S. B. (2005). More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. *Mol. Biol. Evol.* 22, 1337–1344. doi:10.1093/molbev/msi121.

Rokas, A., and Carroll, S. B. (2006). Bushes in the Tree of Life. *PLoS Biol.* 4, e352. doi:10.1371/journal.pbio.0040352.

Rokas, A., Mead, M. E., Steenwyk, J. L., Oberlies, N. H., and Goldman, G. H. (2020a). Evolving moldy murderers: Aspergillus section Fumigati as a model for studying the repeated evolution of fungal pathogenicity. *PLOS Pathog.* 16, e1008315. doi:10.1371/journal.ppat.1008315.

Rokas, A., Mead, M. E., Steenwyk, J. L., Raja, H. A., and Oberlies, N. H. (2020b). Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat. Prod. Rep.* doi:10.1039/C9NP00045C.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804. doi:10.1038/nature02053.

Rokas, A., Wisecaver, J. H., and Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.* doi:10.1038/s41579-018-0075-3.

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* 20, 1001–1009. doi:10.1101/gr.104372.109.

Ropars, J., Cruaud, C., Lacoste, S., and Dupont, J. (2012). A taxonomic and ecological overview of cheese fungi. *Int. J. Food Microbiol.* 155, 199–210. doi:10.1016/j.ijfoodmicro.2012.02.005.

Rougier, N. P., Droettboom, M., and Bourne, P. E. (2014). Ten Simple Rules for Better Figures. *PLoS Comput. Biol.* 10, e1003833. doi:10.1371/journal.pcbi.1003833.

Rutsaert, L., Steinfort, N., Van Hunsel, T., Bomans, P., Naesens, R., Mertes, H., et al. (2020). COVID-19-associated invasive pulmonary aspergillosis. *Ann. Intensive Care* 10, 71. doi:10.1186/s13613-020-00686-4.

Rydholm, C., Dyer, P. S., and Lutzoni, F. (2007). DNA sequence characterization and molecular evolution of MAT1 and MAT2 mating-type loci of the self-compatible ascomycete mold Neosartorya fischeri. *Eukaryot. Cell*. doi:10.1128/EC.00319-06.

Sabi, R., and Tuller, T. (2014). Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. *DNA Res.* 21, 511–526. doi:10.1093/dnares/dsu017.

Safonova, Y., Bankevich, A., and Pevzner, P. A. (2015). dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *J. Comput. Biol.* 22, 528–545. doi:10.1089/cmb.2014.0153.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi:10.1038/nature12130.

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* 31, 1261–1271. doi:10.1093/molbev/msu061.

Samson, R. A., Hong, S., Peterson, S. W., Frisvad, J. C., and Varga, J. (2007). Polyphasic taxonomy of Aspergillus section Fumigati and its teleomorph Neosartorya. *Stud. Mycol.* 59, 147–203. doi:10.3114/sim.2007.59.14.

Samson, R. A., Visagie, C. M., Houbraken, J., Hong, S.-B., Hubka, V., Klaassen, C. H. W., et al. (2014). Phylogeny, identification and nomenclature of the genus Aspergillus. *Stud. Mycol.* 78, 141–173. doi:10.1016/j.simyco.2014.07.004.

Sang, T., and Zhong, Y. (2000). Testing Hybridization Hypotheses Based on Incongruent Gene Trees. *Syst. Biol.* 49, 422–434. doi:10.1080/10635159950127321.

Sanguinetti, M., Posteraro, B., La Sorda, M., Torelli, R., Fiori, B., Santangelo, R., et al. (2006). Role of AFR1, an ABC transporter-encoding gene, in the in vivo response to fluconazole and virulence of Cryptococcus neoformans. *Infect. Immun.* 74, 1352–1359.

doi:10.1128/IAI.74.2.1352-1359.2006.

Sarkar, D. (2017). Package “lattice”: Trellis Graphics for R. *R Doc.*

Sarkar, D., Le meur, N., and Gentleman, R. (2008). Using flowViz to visualize flow cytometry data. *Bioinformatics*. doi:10.1093/bioinformatics/btn021.

Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* 21, 3482–3489. doi:10.1093/bioinformatics/bti564.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 37, D5–D15. doi:10.1093/nar/gkn741.

Sayyari, E., and Mirarab, S. (2018). Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes (Basel).* 9. doi:10.3390/genes9030132.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M., and Wolfe, K. H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci.* 104, 8397–8402. doi:10.1073/pnas.0608218104.

Schacherer, J., Shapiro, J. a, Ruderfer, D. M., and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* 458, 342–5. doi:10.1038/nature07670.

Schmidt, S. M., and Panstruga, R. (2011). Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis? *Curr. Opin. Plant Biol.* 14, 392–399. doi:https://doi.org/10.1016/j.pbi.2011.03.006.

Schofield, M. J., and Hsieh, P. (2003). DNA Mismatch Repair: Molecular Mechanisms and Biological Function. *Annu. Rev. Microbiol.* 57, 579–608. doi:10.1146/annurev.micro.57.030502.090847.

Schröder, M. S., Martinez de San Vicente, K., Prandini, T. H. R., Hammel, S., Higgins, D. G., Bagagli, E., et al. (2016). Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species. *PLOS Genet.* 12, e1006404. doi:10.1371/journal.pgen.1006404.

Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference* doi:10.25080/majora-92bf1922-011.

Sebastian, J., Kraus, B., and Sancar, G. B. (1990). Expression of the yeast PHR1 gene is induced by DNA-damaging agents. *Mol. Cell. Biol.* 10, 4630–7.

Sebastian, J., and Sancar, G. B. (1991). A damage-responsive DNA binding protein regulates transcription of the yeast DNA repair gene PHR1. *Proc. Natl. Acad. Sci.* 88, 11251–11255. doi:10.1073/pnas.88.24.11251.

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *BMJ* 349, g7327–g7327. doi:10.1136/bmj.g7327.

Seeberg, E., Eide, L., and Bjørås, M. (1995). The base excision repair pathway. *Trends Biochem. Sci.* 20, 391–397. doi:10.1016/S0968-0004(00)89086-6.

Seena, S., Duarte, S., Pascoal, C., and Cássio, F. (2012). Intraspecific Variation of the Aquatic Fungus Articulospora tetracladia: An Ubiquitous Perspective. *PLoS One* 7, e35884.

Segal, E. S., Gritsenko, V., Levitan, A., Yadav, B., Dror, N., Steenwyk, J. L., et al. (2018). Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a

Stable Haploid Isolate of Candida albicans. *MBio* 9. doi:10.1128/mBio.02048-18.

Seidel, D., Durán Graeff, L. A., Vehreschild, M. J. G. T., Wisplinghoff, H., Ziegler, M., Vehreschild, J. J., et al. (2017). FungiScope ™ -Global Emerging Fungal Infection Registry. *Mycoses* 60, 508–516. doi:10.1111/myc.12631.

Seixas, I., Barbosa, C., Salazar, S. B., Mendes-Faia, A., Wang, Y., Güldener, U., et al. (2017). Genome Sequence of the Nonconventional Wine Yeast Hanseniaspora guilliermondii UTAD222. *Genome Announc.* 5, e01515-16. doi:10.1128/genomeA.01515-16.

Seki, T., Choi, E. H., and Ryu, D. (1985). Construction of killer wine yeast strain. *Appl. Environ. Microbiol.* 49, 1211–1215.

Sener, E. F. (2014). Association of Copy Number Variations in Autism Spectrum Disorders: A Systematic Review. *Chinese J. Biol.* 2014, 1–9. doi:10.1155/2014/713109.

Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P., and Nicolas, A. G. (2014). Mutational landscape of yeast mutator strains. *Proc. Natl. Acad. Sci.* 111, 1897–1902. doi:10.1073/pnas.1314423111.

Serres-Giardi, L., Belkhir, K., David, J., and Glémin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *Plant Cell* 24, 1379–1397. doi:10.1105/tpc.111.093674.

Sewell, T. R., Zhu, J., Rhodes, J., Hagen, F., Meis, J. F., Fisher, M. C., et al. (2019). Nonrandom Distribution of Azole Resistance across the Global Population of Aspergillus fumigatus. *MBio* 10. doi:10.1128/mBio.00392-19.

Seyedmousavi, S., Guillot, J., Arné, P., de Hoog, G. S., Mouton, J. W., Melchers, W. J. G., et al. (2015). Aspergillus and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. *Med. Mycol.* 53, 765–797. doi:10.1093/mmy/myv067.

Shang, Y., Xiao, G., Zheng, P., Cen, K., Zhan, S., and Wang, C. (2016). Divergent and Convergent Evolution of Fungal Pathogenicity. *Genome Biol. Evol.* 8, 1374–1387. doi:10.1093/gbe/evw082.

Sharp, P. M., Tuohy, T. M. F., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14, 5125–5143. doi:10.1093/nar/14.13.5125.

Sharpton, T. J., Stajich, J. E., Rounsley, S. D., Gardner, M. J., Wortman, J. R., Jordar, V. S., et al. (2009). Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. *Genome Res.* 19, 1722–1731. doi:10.1101/gr.087551.108.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* 1, 0126. doi:10.1038/s41559-017-0126.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 175, 1533-1545.e20. doi:10.1016/j.cell.2018.10.023.

Shen, X.-X., Salichos, L., and Rokas, A. (2016a). A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. *Genome Biol. Evol.* 8, 2565–2580. doi:10.1093/gbe/evw179.

Shen, X.-X., Steenwyk, J. L., Labella, A. L., Opulente, D. A., Zhou, X., Kominek, J., et al. (2020a). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *bioRxiv*. doi:10.1101/2020.05.11.088658.

Shen, X.-X., Steenwyk, J. L., LaBella, A. L., Opulente, D. A., Zhou, X., Kominek, J., et al. (2020b). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal

phylum Ascomycota. *Sci. Adv.* 6, eabd0079. doi:10.1126/sciadv.abd0079.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2016b). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3 Genes|Genomes|Genetics* 6, 3927–3939. doi:10.1534/g3.116.034744.

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst. Biol.* 51, 492–508. doi:10.1080/10635150290069913.

Shimodaira, H., and Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* 16, 1114. doi:10.1093/oxfordjournals.molbev.a026201.

Shor, E., Garcia-Rubio, R., DeGregorio, L., and Perlin, D. S. (2020). A Noncanonical DNA Damage Checkpoint Response in a Major Fungal Pathogen. *MBio* 11, e03044-20. doi:10.1128/mBio.03044-20.

Shwab, E. K., Bok, J. W., Tribus, M., Galehr, J., Graessle, S., and Keller, N. P. (2007). Histone Deacetylase Activity Regulates Chemical Diversity in Aspergillus. *Eukaryot. Cell* 6, 1656–1664. doi:10.1128/EC.00186-07.

Sicard, D., and Legras, J. L. (2011). Bread, beer and wine: Yeast domestication in the Saccharomyces sensu stricto complex. *Comptes Rendus - Biol.* 334, 229–236. doi:10.1016/j.crvi.2010.12.016.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.

Simpson, J. T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 22, 549–556. doi:10.1101/gr.126953.111.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123. doi:10.1101/gr.089532.108.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–32. doi:10.1038/nrg3642.

Singh-Babak, S. D., Babak, T., Fraser, H. B., and Johnson, A. D. (2021). Lineage-specific selection and the evolution of virulence in the Candida clade. *Proc. Natl. Acad. Sci.* 118, e2016818118. doi:10.1073/pnas.2016818118.

Sionov, E., Lee, H., Chang, Y. C., and Kwon-Chung, K. J. (2010). Cryptococcus neoformans overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. *PLoS Pathog.* 6, e1000848. doi:10.1371/journal.ppat.1000848.

Skelly, D. A., Merrihew, G. E., Riffle, M., Connelly, C. F., Kerr, E. O., Johansson, M., et al. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* 23, 1496–1504. doi:10.1101/gr.155762.113.

Slot, J. C., and Rokas, A. (2010). Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl. Acad. Sci.* 107, 10136–10141. doi:10.1073/pnas.0914418107.

Smith, J. S. C., Chin, E. C. L., Shu, H., Smith, O. S., Wall, S. J., Senior, M. L., et al. (1997). An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L.): Comparisons with data from RFLPS and pedigree. *Theor. Appl. Genet.* 95, 163–173. doi:10.1007/s001220050544.

Smith, M. L., and Hahn, M. W. (2021). New Approaches for Inferring Phylogenies in the

Presence of Paralogs. *Trends Genet.* 37, 174–187. doi:10.1016/j.tig.2020.08.012.

Smith, M. L., Vanderpool, D., and Hahn, M. W. (2021). Using all gene families vastly expands data available for phylogenomic inference in primates. *bioRxiv*, 2021.09.22.461252. doi:10.1101/2021.09.22.461252.

Smith, S. A., Brown, J. W., and Walker, J. F. (2018). So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLoS One* 13, e0197433. doi:10.1371/journal.pone.0197433.

Sneath, P. H. (1966). Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* 12, 157–95.

Sniegowski, P. D., Gerrish, P. J., and Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature* 387, 703–705. doi:10.1038/42701.

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., et al. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* 76, 71–76. doi:10.1016/j.ijsu.2020.02.034.

Song, L., Florea, L., and Langmead, B. (2014). Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.* 15, 509. doi:10.1186/s13059-014-0509-9.

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci.* 109, 14942–14947. doi:10.1073/pnas.1211733109.

Song, Y., Cheon, S. A., Lee, K. E., Lee, S.-Y., Lee, B.-K., Oh, D.-B., et al. (2008). Role of the RAM network in cell polarity and hyphal morphogenesis in Candida albicans. *Mol. Biol. Cell* 19, 5456–77. doi:10.1091/mbc.e08-03-0272.

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010).

Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism. *Science (80-.).* 330, 1543 LP – 1546. doi:10.1126/science.1194573.

Spikes, S., Xu, R., Nguyen, C. K., Chamilos, G., Kontoyiannis, D. P., Jacobson, R. H., et al. (2008). Gliotoxin Production in Aspergillus fumigatus Contributes to Host-Specific Differences in Virulence. *J. Infect. Dis.* 197, 479–486. doi:10.1086/525044.

Squire, R. (1981). Ranking animal carcinogens: a proposed regulatory approach. *Science (80-.).* 214, 877–880. doi:10.1126/science.7302565.

Stamatakis, A. (2014a). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Stamatakis, A. (2014b). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst. Biol.* 57, 758–771. doi:10.1080/10635150802429642.

Stambuk, B. U., Batista, A. S., and De Araujo, P. S. (2000). Kinetics of active sucrose transport in Saccharomyces cerevisiae. *J. Biosci. Bioeng.* 89, 212–214. doi:10.1016/S1389-1723(00)88742-3.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, ii215–ii225. doi:10.1093/bioinformatics/btg1080.

Steenwyk, J. L. (2020). JLSteenwyk/ggpubfigs: first release of ggpubfigs. doi:10.5281/ZENODO.4126988.

Steenwyk, J. L. (2021a). Evolutionary Divergence in DNA Damage Responses among Fungi. *MBio* 12. doi:10.1128/mBio.03348-20.

Steenwyk, J. L. (2021b). Evolutionary Divergence in DNA Damage Responses among Fungi.

MBio 12, e03348-20. doi:10.1128/mBio.03348-20.

Steenwyk, J. L., Buida, T. J., Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., et al. (2021a). BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. *bioRxiv*, 2021.10.02.462868. doi:10.1101/2021.10.02.462868.

Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X.-X., and Rokas, A. (2021b). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*. doi:10.1093/bioinformatics/btab096.

Steenwyk, J. L., Buida, T. J., LaBella, A. L., Li, Y., Shen, X.-X., and Rokas, A. (2020a). PhyKIT: a UNIX shell toolkit for processing and analyzing phylogenomic data. *bioRxiv*, 2020.10.27.358143. doi:10.1101/2020.10.27.358143.

Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2020b). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biol.* 18, e3001007. doi:10.1371/journal.pbio.3001007.

Steenwyk, J. L., Goltz, D. C., Buida, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2021c). orthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. *bioRxiv*, 2021.10.30.466607. doi:10.1101/2021.10.30.466607.

Steenwyk, J. L., Lind, A. L., Ries, L. N. A., dos Reis, T. F., Silva, L. P., Almeida, F., et al. (2020c). Pathogenic Allodiploid Hybrids of Aspergillus Fungi. *Curr. Biol.* 30, 2495-2507.e7. doi:10.1016/j.cub.2020.04.071.

Steenwyk, J. L., Mead, M. E., de Castro, P. A., Valero, C., Damasio, A., dos Santos, R. A. C., et al. (2021d). Genomic and Phenotypic Analysis of COVID-19-Associated Pulmonary Aspergillosis Isolates of Aspergillus fumigatus. *Microbiol. Spectr.* 9. doi:10.1128/Spectrum.00010-21.

Steenwyk, J. L., Mead, M. E., Knowles, S. L., Raja, H. A., Roberts, C. D., Bader, O., et al. (2020d). Variation Among Biosynthetic Gene Clusters, Secondary Metabolite Profiles, and Cards of Virulence Across Aspergillus Species. *Genetics*, genetics.303549.2020. doi:10.1534/genetics.120.303549.

Steenwyk, J. L., Opulente, D. A., Kominek, J., Shen, X.-X., Zhou, X., Labella, A. L., et al. (2019a). Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. *PLOS Biol.* 17, e3000255. doi:10.1371/journal.pbio.3000255.

Steenwyk, J. L., Phillips, M. A., Yang, F., Date, S. S., Graham, T. R., Berman, J., et al. (2021e). A gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. *bioRxiv*, 2021.07.09.451830. doi:10.1101/2021.07.09.451830.

Steenwyk, J. L., and Rokas, A. (2018). Copy Number Variation in Fungi and Its Implications for Wine Yeast Genetic Diversity and Adaptation. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.00288.

Steenwyk, J. L., and Rokas, A. (2019). Treehouse: a user-friendly application to obtain subtrees from large phylogenies. *BMC Res. Notes* 12, 541. doi:10.1186/s13104-019-4577-5.

Steenwyk, J. L., and Rokas, A. (2021a). ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures. *Microbiol. Resour. Announc.* 10. doi:10.1128/MRA.00871-21.

Steenwyk, J. L., and Rokas, A. (2021b). orthofisher: a broadly applicable tool for automated gene identification and retrieval. *G3 Genes|Genomes|Genetics* 11. doi:10.1093/g3journal/jkab250.

Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H., and Rokas, A. (2019b). A Robust Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the

Genera Aspergillus and Penicillium *MBio* 10, e00925-19. doi:10.1128/mBio.00925-19.

Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H., and Rokas, A. (2019c). A Robust Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the Genera Aspergillus and Penicillium. *MBio* 10. doi:10.1128/mBio.00925-19.

Steenwyk, J. L., Soghigian, J. S., Perfect, J. R., and Gibbons, J. G. (2016). Copy number variation contributes to cryptic genetic variation in outbreak lineages of Cryptococcus gattii from the North American Pacific Northwest. *BMC Genomics* 17, 700. doi:10.1186/s12864-016-3044-0.

Steenwyk, J., and Rokas, A. (2017). Extensive Copy Number Variation in Fermentation-Related Genes Among Saccharomyces cerevisiae Wine Strains. *G3 Genes, Genomes, Genet.* 7. Available at: http://www.g3journal.org/content/7/5/1475#ref-25 [Accessed July 3, 2017].

Sterling, C. H. (2005). DNA Polymerase 4 of Saccharomyces cerevisiae Is Important for Accurate Repair of Methyl-Methanesulfonate-Induced DNA Damage. *Genetics* 172, 89–98. doi:10.1534/genetics.105.049254.

Sternes, P. R., Lee, D., Kutyna, D. R., and Borneman, A. R. (2016). Genome Sequences of Three Species of Hanseniaspora Isolated from Spontaneous Wine Fermentations. *Genome Announc.* 4, e01287-16. doi:10.1128/genomeA.01287-16.

Stierle, A. A., and Stierle, D. B. (2015). Bioactive Secondary Metabolites Produced by the Fungal Endophytes of Conifers. *Nat. Prod. Commun.* 10, 1671–82.

Struck, T. H. (2014). TreSpEx–-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evol. Bioinforma.* 10, EBO.S14239. doi:10.4137/EBO.S14239.

Struck, T. H., Golombek, A., Weigert, A., Franke, F. A., Westheide, W., Purschke, G., et al.

(2015). The Evolution of Annelids Reveals Two Adaptive Routes to the Interstitial Realm. *Curr. Biol.* 25, 1993–1999. doi:10.1016/j.cub.2015.06.007.

Stukenbrock, E. H. (2016). The Role of Hybridization in the Evolution and Emergence of New Fungal Plant Pathogens. *Phytopathology* 106, 104–112. doi:10.1094/PHYTO-08-15-0184-RVW.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., et al. (2010). Diversity of human copy number variation and multicopy genes. *Science (80-.).* 330, 641–646. doi:10.1126/science.1197005.

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science (80-.).* 349, aab3761. doi:10.1126/science.aab3761.

Sugui, J. A., Kwon-Chung, K. J., Juvvadi, P. R., Latge, J.-P., and Steinbach, W. J. (2015). Aspergillus fumigatus and Related Species. *Cold Spring Harb. Perspect. Med.* 5, a019786–a019786. doi:10.1101/cshperspect.a019786.

Sugui, J. A., Pardo, J., Chang, Y. C., Zarember, K. A., Nardone, G., Galvez, E. M., et al. (2007). Gliotoxin Is a Virulence Factor of Aspergillus fumigatus : gliP Deletion Attenuates Virulence in Mice Immunosuppressed with Hydrocortisone. *Eukaryot. Cell* 6, 1562–1569. doi:10.1128/EC.00141-07.

Sugui, J. A., Peterson, S. W., Figat, A., Hansen, B., Samson, R. A., Mellado, E., et al. (2014). Genetic Relatedness versus Biological Compatibility between Aspergillus fumigatus and Related Species. *J. Clin. Microbiol.* 52, 3707–3721. doi:10.1128/JCM.01704-14.

Suh, A. (2016). The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool. Scr.* 45, 50–62. doi:10.1111/zsc.12213.

Sun, L., Zhang, Y., Zhang, Z., Zheng, Y., Du, L., and Zhu, B. (2016). Preferential Protection of Genetic Fidelity within Open Chromatin by the Mismatch Repair Machinery*. *J. Biol. Chem.* 291, 17692–17705. doi:https://doi.org/10.1074/jbc.M116.719971.

Sun, S., Coelho, M. A., David-Palma, M., Priest, S. J., and Heitman, J. (2019). The Evolution of Sexual Reproduction and the Mating-Type Locus: Links to Pathogenesis of Cryptococcus Human Pathogenic Fungi. *Annu. Rev. Genet.* doi:10.1146/annurev-genet-120116-024755.

Sundin, L. J. R., Guimaraes, G. J., and Deluca, J. G. (2011). The NDC80 complex proteins Nuf2 and Hec1 make distinct contributions to kinetochore-microtubule attachment in mitosis. *Mol. Biol. Cell* 22, 759–68. doi:10.1091/mbc.E10-08-0671.

Susko, E., and Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Mol. Biol. Evol.* 24, 2139–2150. doi:10.1093/molbev/msm144.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34. doi:10.1093/nar/gkl315.

Tachikawa, H., Bloecher, A., Tatchell, K., and Neiman, A. M. (2001). A Gip1p-Glc7p phosphatase complex regulates septin organization and spore wall formation. *J. Cell Biol.* 155, 797–808. doi:10.1083/jcb.200107008.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H., and Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature* 387, 700–702. doi:10.1038/42696.

Takahashi-Nakaguchi, A., Muraosa, Y., Hagiwara, D., Sakai, K., Toyotome, T., Watanabe, A., et al. (2015). Genome sequence comparison of Aspergillus fumigatus strains isolated from patients with pulmonary aspergilloma and chronic necrotizing pulmonary aspergillosis.

Med. Mycol. 53, 353–360. doi:10.1093/mmy/myv003.

Takamatsu, S., Ito (Arakawa), H., Shiroya, Y., Kiss, L., and Heluta, V. (2015). First comprehensive phylogenetic analysis of the genus Erysiphe (Erysiphales, Erysiphaceae) I. The Microsphaera lineage. *Mycologia* 107, 475–489. doi:10.3852/15-007.

Talavera, G., and Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Syst. Biol.* 56, 564–577. doi:10.1080/10635150701472164.

Talevich, E., Invergo, B. M., Cock, P. J., and Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* 13, 209. doi:10.1186/1471-2105-13-209.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a040023.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., et al. (2015). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst. Biol.* 64, 778–791. doi:10.1093/sysbio/syv033.

Tarver, J. E., dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., et al. (2016). The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biol. Evol.* 8, 330–344. doi:10.1093/gbe/evv261.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. life Sci.* 17, 57–86.

Tavares, M. J., Güldener, U., Esteves, M., Mendes-Faia, A., Mendes-Ferreira, A., and Mira, N. P. (2018). Genome Sequence of the Wine Yeast Saccharomycodes ludwigii UTAD17.

Microbiol. Resour. Announc. 7. doi:10.1128/MRA.01195-18.

Taylor, J. W. (2015). Evolutionary Perspectives on Human Fungal Pathogens. *Cold Spring Harb. Perspect. Med.* 5, a019588. doi:10.1101/cshperspect.a019588.

Taylor, J. W., Göker, M., and Pitt, J. I. (2016). Choosing one name for pleomorphic fungi: The example of Aspergillus versus Eurotium , Neosartorya and Emericella. *Taxon* 65, 593–601. doi:10.12705/653.10.

Taylor, J. W., Jacobson, D. J., Kroken, S., Kasuga, T., Geiser, D. M., Hibbett, D. S., et al. (2000). Phylogenetic Species Recognition and Species Concepts in Fungi. *Fungal Genet. Biol.* 31, 21–32. doi:10.1006/fgbi.2000.1228.

Tekaia, F., and Latgé, J.-P. (2005). Aspergillus fumigatus: saprophyte or pathogen? *Curr. Opin. Microbiol.* 8, 385–392. doi:10.1016/j.mib.2005.06.017.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. (2001). Computational and Experimental Analysis of Microsatellites in Rice (Oryza sativa L.): Frequency, Length Variation, Transposon Associations, and Genetic Marker Potential. *Genome Res.* 11, 1441–1452.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 18, 1979–1990. doi:10.1101/gr.081612.108.

Teste, M. A., Marie François, J., and Parrou, J. L. (2010). Characterization of a new multigene family encoding isomaltases in the yeast Saccharomyces cerevisiae, the IMA family. *J. Biol. Chem.* 285, 26815–26824. doi:10.1074/jbc.M110.145946.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi:10.1093/nar/gky1055.

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., et al. (2020). Gene content evolution in the arthropods. *Genome Biol.* 21, 15. doi:10.1186/s13059-019-1925-7.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi:10.1101/gr.772403.

Tice, A. K., Žihala, D., Pánek, T., Jones, R. E., Salomaki, E. D., Nenarokov, S., et al. (2021). PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLOS Biol.* 19, e3001365. doi:10.1371/journal.pbio.3001365.

Tomee, and Kauffman (2000). Putative virulence factors of Aspergillus fumigatus. *Clin. Exp. Allergy* 30, 476–484. doi:10.1046/j.1365-2222.2000.00796.x.

Train, C.-M., Glover, N. M., Gonnet, G. H., Altenhoff, A. M., and Dessimoz, C. (2017). Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* 33, i75–i82. doi:10.1093/bioinformatics/btx229.

Tsang, C.-C., Tang, J. Y. M., Lau, S. K. P., and Woo, P. C. Y. (2018). Taxonomy and evolution of Aspergillus, Penicillium and Talaromyces in the omics era – Past, present and future. *Comput. Struct. Biotechnol. J.* 16, 197–210. doi:10.1016/j.csbj.2018.05.003.

Tsubouchi, H., and Ogawa, H. (2000). Exo1 Roles for Repair of DNA Double-Strand Breaks and Meiotic Crossing Over in Saccharomyces cerevisiae. *Mol. Biol. Cell* 11, 2221–2233. doi:10.1091/mbc.11.7.2221.

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., et al. (2019). PHI-base: the pathogen–host interactions database. *Nucleic Acids Res.*

doi:10.1093/nar/gkz904.

Urquhart, A. S., Hu, J., Chooi, Y.-H., and Idnurm, A. (2019). The fungal gene cluster for biosynthesis of the antibacterial agent viriditoxin. *Fungal Biol. Biotechnol.* 6, 9. doi:10.1186/s40694-019-0072-y.

Vallabhaneni, S., Mody, R. K., Walker, T., and Chiller, T. (2016). The Global Burden of Fungal Diseases. *Infect. Dis. Clin. North Am.* 30, 1–11. doi:10.1016/j.idc.2015.10.004.

Valsecchi, I., Sarikaya-Bayram, Ö., Wong Sak Hoi, J., Muszkieta, L., Gibbons, J., Prevost, M.-C., et al. (2017). MybA, a transcription factor involved in conidiation and conidial viability of the human pathogen Aspergillus fumigatus. *Mol. Microbiol.* 105, 880–900. doi:10.1111/mmi.13744.

van Arkel, A. L. E., Rijpstra, T. A., Belderbos, H. N. A., van Wijngaarden, P., Verweij, P. E., and Bentvelsen, R. G. (2020). COVID-19–associated Pulmonary Aspergillosis. *Am. J. Respir. Crit. Care Med.* 202, 132–135. doi:10.1164/rccm.202004-1038LE.

van de Veerdonk, F. L., Gresnigt, M. S., Romani, L., Netea, M. G., and Latgé, J.-P. (2017). Aspergillus fumigatus morphology and dynamic host interactions. *Nat. Rev. Microbiol.* 15, 661–674. doi:10.1038/nrmicro.2017.90.

van der Heijden, R. T., Snel, B., van Noort, V., and Huynen, M. A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8, 83. doi:10.1186/1471-2105-8-83.

Van Der Linden, J. W. M., Warris, A., and Verweij, P. E. (2011). Aspergillus species intrinsically resistant to antifungal agents. *Med. Mycol.* 49, S82–S89. doi:10.3109/13693786.2010.499916.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for

efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi:10.1109/MCSE.2011.37.

van Dongen, S. (2000). Graph clustering by flow simulation. *Graph Stimul. by flow Clust.* PhD thesis, University of Utrecht. doi:10.1016/j.cosrev.2007.05.001.

van Wyk, S., Wingfield, B. D., De Vos, L., van der Merwe, N. A., and Steenkamp, E. T. (2021). Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota . *Front. Microbiol.* 11, 3625.

Varga, T., Krizsán, K., Földi, C., Dima, B., Sánchez-García, M., Sánchez-Ramírez, S., et al. (2019). Megaphylogeny resolves global patterns of mushroom evolution. *Nat. Ecol. Evol.* doi:10.1038/s41559-019-0834-1.

Vargas-Muñiz, J. M., Renshaw, H., Richards, A. D., Lamoth, F., Soderblom, E. J., Moseley, M. A., et al. (2015). The Aspergillus fumigatus septins play pleiotropic roles in septation, conidiation, and cell wall stress, but are dispensable for virulence. *Fungal Genet. Biol.* 81, 41–51. doi:10.1016/j.fgb.2015.05.014.

Verweij, P. E., Ananda-Rajah, M., Andes, D., Arendrup, M. C., Brüggemann, R. J., Chowdhary, A., et al. (2015). International expert opinion on the management of infection caused by azole-resistant Aspergillus fumigatus. *Drug Resist. Updat.* 21–22, 30–40. doi:10.1016/j.drup.2015.08.001.

Vesth, T. C., Nybo, J. L., Theobald, S., Frisvad, J. C., Larsen, T. O., Nielsen, K. F., et al. (2018). Investigation of inter- and intraspecies variation through genome sequencing of Aspergillus section Nigri. *Nat. Genet.* doi:10.1038/s41588-018-0246-1.

Vielba-Fernández, A., Polonio, Á., Ruiz-Jiménez, L., de Vicente, A., Pérez-García, A., and Fernández-Ortuño, D. (2020). Fungicide Resistance in Powdery Mildew Fungi. *Microorg.* 8. doi:10.3390/microorganisms8091431.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34, 275–305. doi:10.1051/gse:2002009.

Vijaykrishna, D., Jeewon, R., and Hyde, K. (2006). Molecular taxonomy, origins and evolution of freshwater ascomycetes. *Fungal Divers.* 23, 351–390.

Vilela-Moura, A., Schuller, D., Mendes-Faia, A., and Côrte-Real, M. (2008). Reduction of volatile acidity of wines by selected yeast strains. *Appl. Microbiol. Biotechnol.* 80, 881–890. doi:10.1007/s00253-008-1616-x.

Vinet, L., and Zhedanov, A. (2011). A 'missing' family of classical orthogonal polynomials. *J. Phys. A Math. Theor.* 44, 085201. doi:10.1088/1751-8113/44/8/085201.

Vinnere Pettersson, O., and Leong, S. L. (2011). "Fungal Xerophiles (Osmophiles)," in *eLS* (Chichester, UK: John Wiley & Sons, Ltd). doi:10.1002/9780470015902.a0000376.pub2.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. doi:10.1038/s41592-019-0686-2.

Visagie, C. M., Houbraken, J., Frisvad, J. C., Hong, S.-B., Klaassen, C. H. W., Perrone, G., et al. (2014). Identification and nomenclature of the genus Penicillium. *Stud. Mycol.* 78, 343–371. doi:10.1016/j.simyco.2014.09.001.

Waddell, P. J., and Steel, M. . (1997). General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites. *Mol. Phylogenet. Evol.* 8, 398–414. doi:10.1006/mpev.1997.0452.

Walker, A. K., Frasz, S. L., Seifert, K. A., Miller, J. D., Mondo, S. J., LaButti, K., et al. (2016). Full Genome of Phialocephala scopiformis DAOMC 229536, a Fungal Endophyte of

Spruce Producing the Potent Anti-Insectan Compound Rugulosin. *Genome Announc.* 4, e01768-15. doi:10.1128/genomeA.01768-15.

Walker, J. F., Walker-Hale, N., Vargas, O. M., Larson, D. A., and Stull, G. W. (2019). Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7, e7747. doi:10.7717/peerj.7747.

Walker, P. J., Firth, C., Widen, S. G., Blasdell, K. R., Guzman, H., Wood, T. G., et al. (2015). Evolution of Genome Size and Complexity in the Rhabdoviridae. *PLOS Pathog.* 11, e1004664. doi:10.1371/journal.ppat.1004664.

Wallace, D. (2004). The Mann-Whitney Test. *J. Am. Soc. Inf. ...*, 1–5.

Walsh, H. E., Kidd, M. G., Moum, T., and Friesen, V. L. (1999). Polytomies and the Power of Phylogenetic Inference. *Evolution (N. Y).* 53, 932. doi:10.2307/2640732.

Wang, F.-Z., Li, D.-H., Zhu, T.-J., Zhang, M., and Gu, Q.-Q. (2011). Pseurotin A 1 and A 2 , two new 1-oxa-7-azaspiro[4.4]non-2-ene-4,6-diones from the holothurian-derived fungus Aspergillus fumigatus WFZ-25. *Can. J. Chem.* 89, 72–76. doi:10.1139/V10-157.

Wang, F., Fang, Y., Zhu, T., Zhang, M., Lin, A., Gu, Q., et al. (2008). Seven new prenylated indole diketopiperazine alkaloids from holothurian-derived fungus Aspergillus fumigatus. *Tetrahedron* 64, 7986–7991. doi:10.1016/j.tet.2008.06.013.

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., and Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* 31, 318–23. doi:10.1093/bioinformatics/btu668.

Warringer, J., Zörgö, E., Cubillos, F. A., Zia, A., Gjuvsland, A., Simpson, J. T., et al. (2011). Trait Variation in Yeast Is Defined by Population History. *PLoS Genet* 7, e1002111+. doi:10.1371/journal.pgen.1002111.

Warris, A., and Ballou, E. R. (2019). Oxidative responses and fungal infection biology. *Semin. Cell Dev. Biol.* 89, 34–46. doi:10.1016/j.semcdb.2018.03.004.

Waskom, M. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6, 3021. doi:10.21105/joss.03021.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018a). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi:10.1093/molbev/msx319.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018b). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi:10.1093/molbev/msx319.

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., and Kriventseva, E. V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* 41, D358–D365. doi:10.1093/nar/gks1116.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., et al. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243. doi:10.1093/nar/gkv437.

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., et al. (2014). Illuminating the Base of the Annelid Tree Using Transcriptomics. *Mol. Biol. Evol.* 31, 1391–1401. doi:10.1093/molbev/msu080.

Weinert, T. A., Kiser, G. L., and Hartwell, L. H. (1994). Mitotic checkpoint genes in budding yeast and the dependence of mitosis on DNA replication and repair. *Genes Dev.* 8, 652–665. doi:10.1101/gad.8.6.652.

Weischenfeldt, J., and Porse, B. (2008). Bone Marrow-Derived Macrophages (BMM): Isolation

and Applications. *Cold Spring Harb. Protoc.* 2008, pdb.prot5080-pdb.prot5080. doi:10.1101/pdb.prot5080.

Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., et al. (2014). Comprehensive variation discovery in single human genomes. *Nat. Genet.* 46, 1350–1355. doi:10.1038/ng.3121.

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., and Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19, 122. doi:10.1186/s12859-018-2128-z.

Whelan, S., and Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* 18, 691–699. doi:10.1093/oxfordjournals.molbev.a003851.

Whelan, N. V., Kocot, K. M., Moroz, L. L., and Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl. Acad. Sci.* 112, 5773–5778. doi:10.1073/pnas.1503453112.

Wickerham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.

Wickham, H. (2009). *ggplot2*. New York, NY: Springer New York doi:10.1007/978-0-387-98141-3.

Wiedner, S. D., Ansong, C., Webb-Robertson, B. J., Pederson, L. A. M., Fortuin, S., Hofstad, B. A., et al. (2013). Disparate proteome responses of pathogenic and nonpathogenic aspergilli to human serum measured by Activity-Based Protein Profiling (ABPP). *Mol. Cell. Proteomics*. doi:10.1074/mcp.M112.026534.

Wiemann, P., Guo, C.-J., Palmer, J. M., Sekonyela, R., Wang, C. C. C., and Keller, N. P. (2013). Prototype of an intertwined secondary-metabolite supercluster. *Proc. Natl. Acad. Sci.* 110,

17065–17070. doi:10.1073/pnas.1313258110.

Wiemann, P., Lechner, B. E., Baccile, J. A., Velk, T. A., Yin, W.-B., Bok, J. W., et al. (2014). Perturbations in small molecule synthesis uncovers an iron-responsive secondary metabolite network in Aspergillus fumigatus. *Front. Microbiol.* 5. doi:10.3389/fmicb.2014.00530.

Wightman, R., and Meacock, P. A. (2003). The THI5 gene family of Saccharomyces cerevisiae: distribution of homologues among the hemiascomycetes and functional redundancy in the aerobic biosynthesis of thiamin from pyridoxine. *Microbiology* 149, 1447–1460. doi:10.1099/mic.0.26194-0.

Wikstrom, N., Savolainen, V., and Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B Biol. Sci.* 268, 2211–2220. doi:10.1098/rspb.2001.1782.

Wildman, D. E., Uddin, M., Opazo, J. C., Liu, G., Lefort, V., Guindon, S., et al. (2007). Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl. Acad. Sci.* 104, 14395–14400. doi:10.1073/pnas.0704342104.

Willson, J., Roddur, M. S., Liu, B., Zaharias, P., and Warnow, T. (2021). DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Syst. Biol.* doi:10.1093/sysbio/syab070.

Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 41, D165–D170. doi:10.1093/nar/gks1123.

Winterton, S. L., Lemmon, A. R., Gillung, J. P., Garzon, I. J., Badano, D., Bakkes, D. K., et al. (2018). Evolution of lacewings and allied orders using anchored phylogenomics (Neuroptera, Megaloptera, Raphidioptera). *Syst. Entomol.* 43, 330–354.

doi:10.1111/syen.12278.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., et al. (1999). Functional characterization of the S-cerevisiae genome by gene deletion and parallel analysis. *Science (80-.).* 285, 901–906. doi:10.1126/science.285.5429.901.

Woese, C. R., Achenbach, L., Rouviere, P., and Mandelco, L. (1991). Archaeal Phylogeny: Reexamination of the Phylogenetic Position of Archaeoglohus fulgidus in Light of Certain Composition-induced Artifacts. *Syst. Appl. Microbiol.* 14, 364–371. doi:10.1016/S0723-2020(11)80311-5.

Wolfe, K. H. (2015). Origin of the Yeast Whole-Genome Duplication. *PLOS Biol.* 13, e1002221. doi:10.1371/journal.pbio.1002221.

Wolfe, K. H., Armisén, D., Proux-Wera, E., ÓhÉigeartaigh, S. S., Azam, H., Gordon, J. L., et al. (2015). Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res.* 15, fov035. doi:10.1093/femsyr/fov035.

Wolfe, K. H., and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713. doi:10.1038/42711.

Wolfe, N. W., and Clark, N. L. (2015). ERC analysis: web-based inference of gene function via evolutionary rate covariation: Fig. 1. *Bioinformatics*, btv454. doi:10.1093/bioinformatics/btv454.

Wren, J. D. (2016). Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades. *Bioinformatics* 32, 2686–2691. doi:10.1093/bioinformatics/btw284.

Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. *Curr.*

Biol. 27, 3025-3033.e5. doi:10.1016/j.cub.2017.08.043.

Xi, Z., Liu, L., Rest, J. S., and Davis, C. C. (2014). Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies. *Syst. Biol.* 63, 919–932. doi:10.1093/sysbio/syu055.

Xia, X. (2013). DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. *Mol. Biol. Evol.* 30, 1720–1728. doi:10.1093/molbev/mst064.

Xiao, W., and Chow, B. L. (1998). Synergism between yeast nucleotide and base excision repair pathways in the protection against DNA methylation damage. *Curr. Genet.* doi:10.1007/s002940050313.

Xiao, W., Chow, B. L., Hanna, M., and Doetsch, P. W. (2001). Deletion of the MAG1 DNA glycosylase gene suppresses alkylation-induced killing and mutagenesis in yeast cells lacking AP endonucleases. *Mutat. Res. - DNA Repair*. doi:10.1016/S0921-8777(01)00113-6.

Xu, X., Liu, Q., Fan, L., Cui, X., and Zhou, X. (2008). Analysis of synonymous codon usage and evolution of begomoviruses. *J. Zhejiang Univ. Sci. B* 9, 667–674. doi:10.1631/jzus.B0820005.

Yamada, A., Kataoka, T., and Nagai, K. (2000). The fungal metabolite gliotoxin: immunosuppressive activity on CTL-mediated cytotoxicity. *Immunol. Lett.* 71, 27–32. doi:10.1016/s0165-2478(99)00155-8.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–42. doi:10.1038/nrg3174.

Yang, Y., Chen, M., Li, Z., Al-Hatmi, A. M. S., de Hoog, S., Pan, W., et al. (2016). Genome Sequencing and Comparative Genomics Analysis Revealed Pathogenic Potential in

Penicillium capsulatum as a Novel Fungal Pathogen Belonging to Eurotiales. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01541.

Yang, Y., Moore, M. J., Brockington, S. F., Soltis, D. E., Wong, G. K.-S., Carpenter, E. J., et al. (2015). Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. *Mol. Biol. Evol.* 32, 2001–2014. doi:10.1093/molbev/msv081.

Yang, Y., and Smith, S. A. (2014). Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. doi:10.1093/molbev/msu245.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39, 306–314. doi:10.1007/BF00160154.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372. doi:10.1016/0169-5347(96)10041-0.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.

Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503. doi:10.1016/S0169-5347(00)01994-7.

Yin, W.-B., Baccile, J. A., Bok, J. W., Chen, Y., Keller, N. P., and Schroeder, F. C. (2013). A Nonribosomal Peptide Synthetase-Derived Iron(III) Complex from the Pathogenic Fungus Aspergillus fumigatus. *J. Am. Chem. Soc.* 135, 2064–2067. doi:10.1021/ja311145n.

Yin, W.-B., Ruan, H.-L., Westrich, L., Grundmann, A., and Li, S.-M. (2007). CdpNPT, an N-Prenyltransferase fromAspergillus fumigatus: Overproduction, Purification and

Biochemical Characterisation. *ChemBioChem* 8, 1154–1161. doi:10.1002/cbic.200700079.

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* 10, 402–415. doi:http://dx.doi.org/10.2174/138920209789177575.

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., et al. (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife* 2. doi:10.7554/eLife.00731.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19, 153. doi:10.1186/s12859-018-2129-y.

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020). ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol. Biol. Evol.* 37, 3292–3307. doi:10.1093/molbev/msaa139.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 10, 451–481. doi:10.1146/annurev.genom.9.081307.164217.

Zhao, C., Fraczek, M. G., Dineen, L., Lebedinec, R., Macheleidt, J., Heinekamp, T., et al. (2019a). High-Throughput Gene Replacement in Aspergillus fumigatus. *Curr. Protoc. Microbiol.* doi:10.1002/cpmc.88.

Zhao, J., Mou, Y., Shan, T., Li, Y., Zhou, L., Wang, M., et al. (2010). Antimicrobial Metabolites from the Endophytic Fungus Pichia guilliermondii Isolated from Paris polyphylla var. yunnanensis. *Molecules* 15, 7961–7970. doi:10.3390/molecules15117961.

Zhao, S., Latgé, J.-P., and Gibbons, J. G. (2019b). Genome Sequences of Two Strains of the

Food Spoilage Mold Aspergillus fischeri. *Microbiol. Resour. Announc.* 8. doi:10.1128/MRA.01328-19.

Zhong, B., Liu, L., Yan, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends Plant Sci.* 18, 492–495. doi:10.1016/j.tplants.2013.04.009.

Zhou, B.-B. S., and Elledge, S. J. (2000). The DNA damage response: putting checkpoints in perspective. *Nature* 408, 433–439. doi:10.1038/35044005.

Zhou, P., Liu, Z., Chen, Y., Xiao, Y., Huang, X., and Fan, X.-G. (2020). Bacterial and fungal infections in COVID-19 patients: A matter of concern. *Infect. Control Hosp. Epidemiol.* 41, 1124–1125. doi:10.1017/ice.2020.156.

Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., Looz, M. von, and Rokas, A. (2017). Quartet-based computations of internode certainty provide accurate and robust measures of phylogenetic incongruence. *bioRxiv*, 168526. doi:10.1101/168526.

Zhou, X., Peris, D., Kominek, J., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2016). in silico Whole Genome Sequencer & Analyzer (iWGS): A Computational Pipeline to Guide the Design and Analysis of de novo Genome Sequencing Studies. *G3 Genes|Genomes|Genetics*. doi:10.1534/g3.116.034249.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2018). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol. Biol. Evol.* 35, 486–503. doi:10.1093/molbev/msx302.

Zhu, Y. O., Sherlock, G., and Petrov, D. A. (2016). Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. *G3 Genes|Genomes|Genetics* 6, 2421–2434. doi:10.1534/g3.116.029397.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. A. (2014a). Precise estimates of mutation

rate and spectrum in yeast. *Proc. Natl. Acad. Sci.* 111, E2310–E2318. doi:10.1073/pnas.1323011111.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. A. (2014b). Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2310–E2318. doi:10.1073/pnas.1323011111.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi:10.1093/bioinformatics/btt476.