THE FUNCTION AND EVOLUTION OF THE *ASPERGILLUS* GENOME

By

John Gregory Gibbons

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

August 2012

Nashville, Tennessee

Approved:

Professor Antonis Rokas

Professor David E. McCauley

Professor David M. Geiser

Professor Katherine L. Friedman

Professor Seth R. Bordenstein

To my loving and supportive family, Vicenta, John and Elena

and

To my wife Ali, whose love and thoughtfulness I am forever grateful for

ACKNOWLEDGEMENTS

My dissertation took many directions over the last five years. I am especially grateful for my committee members who have willingly let this process evolve and have continually given me support and encouragement. Seth Bordenstein provided me with many key insights along the way that positively impacted my research. Kathy Friedman's genetic and molecular biology expertise was invaluable. In particular, I struggled with a molecular protocol for several months and Kathy helped me solve the problem by troubleshooting each step. David Geiser has been my "*Aspergillus* encyclopedia", has helped me to think about the biological context of my data and has shared many samples with me along the way. David McCauley has been a fantastic committee chair. For virtually every chapter in my dissertation, I have spent time in his office talking about statistics and population genetics.

Although it is difficult to thank someone who has impacted you so profoundly in a few sentences, I would like to especially acknowledge my advisor, Antonis Rokas. He has been an amazing mentor, role model and friend. His patience, modesty, drive and enthusiasm for science have been contagious. I will always appreciate his confidence in my work and his treatment of me as a colleague. I will never forget the hours spent in his office brainstorming ideas on his dry-erase board. I only hope to someday guide students with such quality and compassion. Thank you for all that you have done.

I have also had the opportunity to work with some amazing people, both at Vanderbilt and elsewhere. My current and former labmates have been a constant source of insight and entertainment. Special thanks to Jason Slot, who has helped me immensely and has shown me "where there's a will, there's a way", Kris McGary for always

dropping everything to talk science (and anything else), Leonidas Salichos for his continuous collaboration and persistence to “reward ourselves” after a long week and David Rinker for this scientific realism to balance out my over optimism. Thanks also to current member Patricia Soria and former members Ioannis Stergiopoulos, Pad Mahadevan and John Tossberg. I would also like to recognize undergraduate Holly Elmore, who worked unbelievably hard with me over the last two years. Many other collaborators have helped me grow as a scientist in some way, including Patrick Abbot, Chris Hittinger, Eric Jansen, Maren Klich, Corne Klaassen, Natalie Fedorova, Arun Balajee, Sarah Lawson, Jonas King, David Tabb, Hayes McDonald and Jean-Marc Lassance. I have to also express my sincere gratitude toward Jean-Paul Latge and Anne Beauvais for their warmth and hospitably. I also need to acknowledge, Steven Baskauf who has been a great teaching mentor, Roz Johnson and Leslie Maxwell for helping me smoothly jump through the hoops, Carol Wiley for her support during grant writing, Travis Clark for “all-things-Illumina” and to Kristen Porter-Utley and Renate Gebauer for sparking my interest in biology as an undergraduate and keeping it ignited.

Of course, I could not have done any of this without my family. They have been there with comfort when I needed it most, constant pride and unconditional love. Thank you for everything Mom, Dad and Elena. Most importantly, I can never thank my wife Ali enough. She has been the glue that has kept me together, my greatest inspiration, and my best friend.

My work has been generously supported by Vanderbilt University, The Gisela Mosig Travel Fund, The Vanderbilt Graduate School and the National Institutes of Health and National Institute of Allergy and Infectious Diseases (F31AI091343-01).

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

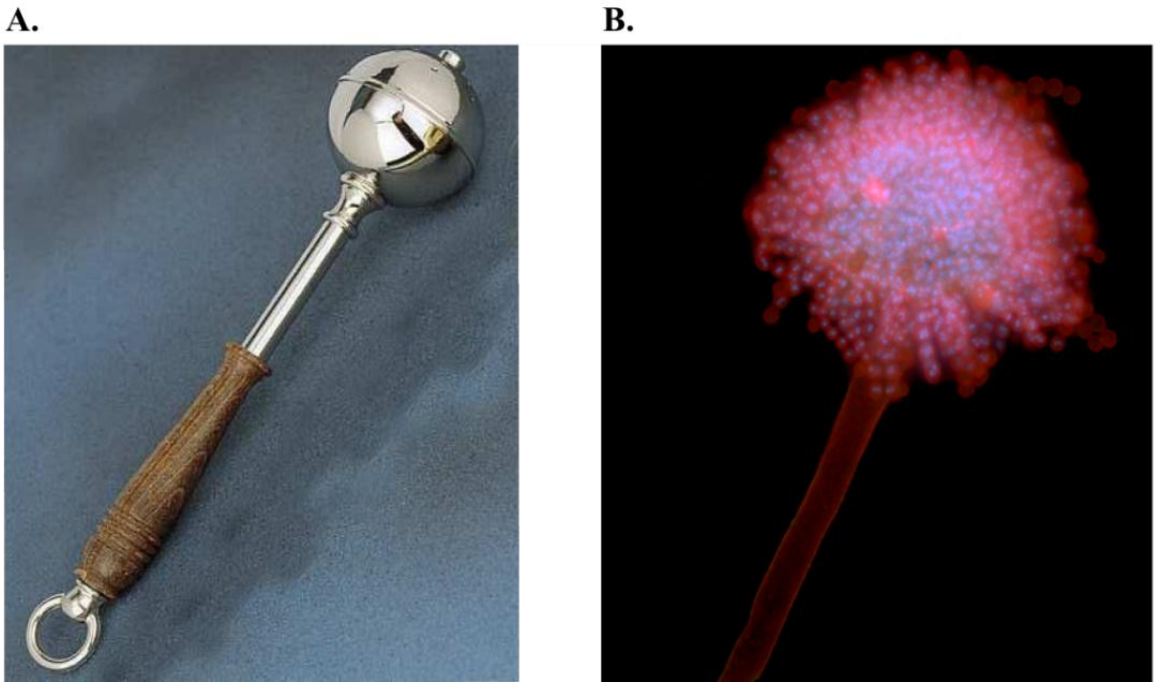
CHAPTER I

INTRODUCTION

The Genus *Aspergillus*

Nearly 300 years ago (1729), priest and botanist Pietro Antonio Micheli first described the asexual reproductive structures of several molds, including *Aspergillus*, which he likened in appearance to the *Aspergillum*, the instrument used to sprinkle holy water in the Roman Catholic Church (Osmani and Goldman 2008) (Figure 1.1). This asexual spore producing structure is the defining microscopic marker distinguishing species of the genus (Machida and Gomi 2010).

Figure 1.1. (A) The aspergillum, a holy water sprinkling instrument, and (B) the asexual reproductive structure of *Aspergillus flavus* (image courtesy of Jonas G King).

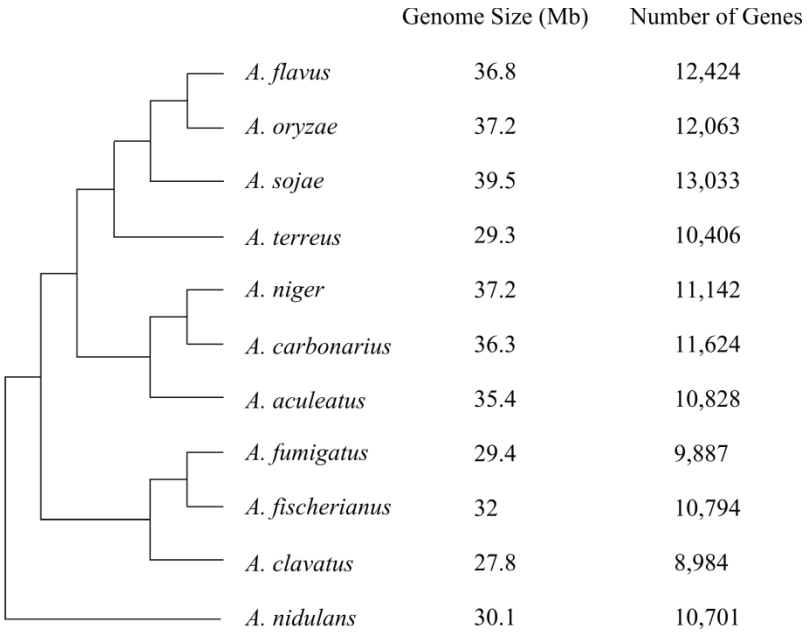


During vegetative growth, mycelium can differentiate becoming enlarged, forming a “T” or “L” shape called the foot cell. The conidiophore, a stalk-like structure, develops from the foot cell and culminates as the spherical vesicle. Primary and secondary sterigmata extend from the vesicle, the latter of which produce the asexual conidiospores (Carlile, Watkinson et al. 2001). In addition to morphological similarities, phylogenetic analysis provides further support that the species of *Aspergillus* belong to a monophyletic group (Peterson 2008). Despite these likenesses, the genus as a whole is extremely diverse and spans an evolutionary distance comparable to humans and fish (Fedorova, Khaldi et al. 2008).

Of the ~250 *Aspergillus* species that have been identified thus far (Samson and Varga 2007); roughly one third can also produce meiotic ascopores via sexual reproduction (Geiser 2009). In the cleistothecium, nuclei divide and form a mass of cells called the ascogenous hyphae. The tips of these hyphae differentiate to form the ascus where two haploid nuclei fuse. The newly formed diploid nucleus divides meiotically then mitotically to produce eight haploid ascospores (Carlile, Watkinson et al. 2001). Genomic features of many *Aspergillus* species known only to reproduce asexually suggest the cryptic ability to reproduce by sexual means (Fisher and Henk 2012). These characteristics include the presence of intact meiosis-related genes (Dyer and O'Gorman 2012), the existence of mating-type loci at near equal ratios (Paoletti, Rydholm et al. 2005; Ramirez-Prado, Moore et al. 2008), and mutational patterns consistent with recombination events (Fisher and Henk 2012; Gibbons, Salichos et al. 2012; Klaassen, Gibbons et al. 2012).

Few genera of fungi are as impactful, both beneficially and damagingly, to human society as the *Aspergillus*. On the positive side, several species are used to make traditional Asian foods and beverages, others are utilized as "cell factories" in the production of industrially important compounds such as citric acid, while further, *A. terreus* is the original source of the cholesterol-lowering drug lovastatin (Baker 2008). Conversely, some *Aspergillus* species produce toxic chemicals that can contaminate crop stocks (Amaike and Keller 2011) and other species can cause infections in humans and other animals (Bennett and Klich 1992; Geiser, Taylor et al. 1998; Latge 1999; Gugnani 2003). As a testament to their importance, representative genomes of several species have been sequenced (Figure 1.2), providing a rich resource for the study of pressing questions concerning the pathogenicity, specialization, and evolutionary history of the *Aspergillus* genome.

Figure 1.2. The phylogenetic relationships and genome characteristics of the sequenced *Aspergillus* species.



Aspergillus Ecology

Aspergillus species are saprophytic and are most frequently found in soil and decaying plant material where they secrete digestive enzymes to externally break down organic matter into more simple nutrients which they then utilize. These fungi play a vital role in the ecosystem, where they recycle carbon and nitrogen (Latge 1999). As generalists, *Aspergillus* can thrive on a variety of different carbon sources and have been isolated from a diversity of substrates including coil, alligator nesting material, aviation fuel, Egyptian mummies, electrical fuses, dead termites and dried bacon (Baker 2008; Gibbons, Salichos et al. 2012). *Aspergillus* spores are very small, typically ranging between 2 – 3μ and are thus easily dispersed through wind (Latge 1999). As such, most *Aspergillus* species are ubiquitously distributed and their conidia have been isolated at extreme environments such high altitude Tibetan glaciers (Zhang, Yao et al. 2002), Antarctica (Wicklow 1968) and the Saharan Desert (Cardwell and Cotty 2002). Although most fungal species are mesophilic (grow at moderate temperatures), particular species, such as *A. fumigatus* are thermotolerant and can grow at temperatures over 50˚C (Bhabhra and Askew 2005).

Approximately 20 *Aspergillus* species are also capable of causing opportunistic infections in humans (Denning 1998; Latge 1999; Brakhage and Langfelder 2002). Individuals with immunodeficiencies due to chronic granulomatous disease, AIDS, chemotherapy treatment or organ transplantation surgery for example, are at much higher risk of acquiring an infection. In particular, *A. fumigatus* makes up the majority of human

disease, followed far less frequently by *A. flavus*, *A. niger*, *A. terreus* and *A. nidulans* (Osmani and Goldman 2008). Although the vast majority of *Aspergillus* infections are caused by *A. fumigatus*, surveys of hospital air have not identified an overabundance of *A. fumigatus* conidia (Schmitt, Blevins et al. 1990; Hospenthal, Kwon-Chung et al. 1998) suggesting that the species is particularly competent at colonizing immunocompromised individuals.

Estimates suggest that most individuals inhale several hundred *A. fumigatus* conidia per day where they can reach the small cavities of the lungs (Latge 1999; Latge 2001). In healthy individuals macrophages phagocytose detected conidia while neutrophils kill the developing hyphae if conidia germinate (Osmani and Goldman 2008). The group of diseases caused by *Aspergillus* is collectively termed aspergillosis and consists of both invasive and localized (aspergilloma) infections. Invasive infections are characterized by their entrance to the blood stream and are nearly always fatal when left untreated, with death normally occurring less than two weeks after diagnosis (Denning 1998). Aspergilloma infections are usually localized to preexisting lung cavities caused by tuberculosis and other pulmonary diseases (Latge 1999). The defining characteristic of aspergilloma is the presence of a "fungus ball" in which the hyphae form a dense network and are embedded in an extracellular matrix made up of proteins, monosaccharides, polysaccharides and secondary metabolites produced by the colony (Beauvais, Schmidt et al. 2007; Loussert, Schmitt et al. 2010). The clinical relevance of this "biofilm" morphology is noteworthy, as the fungi are more virulent and less receptive to antifungal

drug treatments (Mowat, Lang et al. 2008; Seidler, Salvenmoser et al. 2008; Mowat, Williams et al. 2009; Rajendran, Mowat et al. 2011).

Although amphotericin B and caspofungin are used intravenously to treat aspergillosis, their success rate is relatively low, and therefore azole-derived drugs such as itraconazole, voriconazole and posaconazole are more commonly administered (Denning 1998; Howard, Cerar et al. 2009). Azoles inhibit lanosterol demethylase in turn disrupting the production of ergosterol, an essential component of the fungal cell membrane (Sheehan, Hitchcock et al. 1999). The increased frequency of chemotherapy has also resulted in a rise of *Aspergillus* infections. Unfortunately, azole-resistance in *A. fumigatus* has emerged and rapidly spread over the last several decades (Verweij, Dorsthorst et al. 2002; Howard, Webster et al. 2006; Snelders, van der Lee et al. 2008; Howard, Cerar et al. 2009). Functional mutations in the targeted lanosterol demethylase gene have been documented in the majority of resistant isolates, however other isolates with unknown resistance mechanisms are present as well (Mellado, Garcia-Effron et al. 2007).

Though invasive infection of live plants is rare, particular *Aspergillus* species are considered weak pathogens and, perhaps more importantly, agricultural pests. The direct ability of the fungi to penetrate healthy plant material has been demonstrated in stressful conditions; however the entry point of infection is typically induced by insects or other physical means (Gugnani 2003). More frequently however, species such as *A. flavus* and *A. parasiticus* are especially adept for growth on nuts and oilseeds, including cotton,

maize and peanut (Yu, Cleveland et al. 2005). This is particularly troublesome because *A. flavus* and *A. parasiticus* can produce aflatoxin, a toxic and carcinogenic secondary metabolite that was first discovered after thousands of livestock were killed by contaminated groundnut feed (Nesbitt, O'Kelly et al. 1962). Although infrequent, death resulting from acute toxicity of aflatoxin in humans does occur (Krishnamachari, Bhat et al. 1975; Williams, Phillips et al. 2004). In developing countries where aflatoxin levels in food are not tightly regulated, studies suggest an association with elevated frequencies of liver cancer (Pitt 2000).

Industrial Applications of *Aspergillus*

In their natural environment *Aspergillus* species produce and secrete a variety of enzymes to degrade organic matter. These intrinsic characteristics have been exploited by humans for thousands of years in various applications of food and beverage production. In the late 1890s the first commercially available enzyme (an amylase called "takadiastase") produced by *A. oryzae* became available (Liese, Seelbach et al. 2000). Today, many enzymes produced by *Aspergillus* are used in the food and brewing industries, the processing of animal feed and the paper and pulping industries (Machida, Asai et al. 2005; Pel, de Winde et al. 2007). Additionally, the biotechnology industry uses *A. niger* to produce heterologous proteins at very high yields (Punt, van Biezen et al. 2002). *Aspergillus* species are also a rich source of industrially useful metabolites. For example, *A. niger* is the source for virtually the entire global supply citric acid and has been used for this purpose since the 1920s by Pfizer (Papagianni 2007). Furthermore, in 1980

Merck patented the cholesterol lowering drug lovastatin, originally derived from *A. terreus*, which generated over $1 billion of sales (Tobert 2003).

However, perhaps no other *Aspergillus* species shares as long a history with humans as *A. oryzae*. For thousands of years, *A. oryzae* has been used to make sake (rice wine), miso (soybean paste) and shoyu (soy sauce) (Machida, Yamada et al. 2008). The main role of *A. oryzae* is as a metabolic "workhorse"; breaking down its substrate into more palatable forms for humans and simpler components for further processing by yeast and lactic acid bacteria (Baker 2008). Traditionally, *A. oryzae* conidia were mixed with ash and tightly packed into paper bags; a process which preserved spores yet prevented contamination due to alkaline pH and limited oxygen (Machida, Yamada et al. 2008). Interestingly, historical documents chronicling the sake making process described the isolation of mold growing on the ear of rice (Machida, Yamada et al. 2008). Numerous studies have revealed the close similarity of *A. oryzae* to the ubiquitously distributed species *A. flavus*, indicating *A. oryzae* is a domesticated form (Kurtzman, Smiley et al. 1986; Geiser, Pitt et al. 1998; Payne, Nierman et al. 2006; Gibbons, Salichos et al. 2012). However, the two species differ with respect to phenotype, as *A. oryzae* is a safe species used in the food industry while *A. flavus* is a harmful toxin producer.

Aspergillus Secondary Metabolites

Aspergillus species, along with much of the fungal kingdom, produce a diversity of organic compounds called secondary metabolites. As their name indicates, these chemicals are not essential for primary cellular functions and their production is typically

limited to precise developmental stages or environmental circumstances (Keller, Turner et al. 2005). The principal function of secondary metabolites is as defense. *Aspergillus* invest an enormous amount of energy into the external breakdown of food sources and many of their secondary metabolites inhibit the growth of bacteria, fungi, protozoans and other ecological competitors. These compounds include the pigment melanin, the commonly used antibiotic penicillin, the cholesterol reducing drug lovastatin and the toxic carcinogen aflatoxin (Osbourn 2010).

Despite their chemical assortment, most secondary metabolites derive from a limited number of chemical building blocks and fall into four major groups: (i) polyketides, (ii) non-ribosomal peptides, (iii) terpenes and (iv) alkaloids (Keller, Turner et al. 2005; Hoffmeister and Keller 2007). Remarkably, in fungi the genes encoding for the synthesis, regulation and transport of these compounds are most often present in the genome as a physically clustered group of genes residing in close proximity to one another on the chromosome (Osbourn 2010). These gene clusters can be small, such as the 3-gene cluster encoding penicillin in *A. nidulans*, or large, as in the ~70 Kb, 26-gene cluster encoding aflatoxin in *A. flavus* (Osbourn 2010). Individual gene clusters can be highly variable within species with respect to their allelic states, gene content and entire presence (Kusumoto, Nogata et al. 2000; Tominaga, Lee et al. 2006; Gibbons, Salichos et al. 2012).

Insights from the *Aspergillus* Genomes

Since 2005, the genomes of 15 *Aspergillus* isolates from 13 species have been sequenced with an additional number of other species currently underway (Figure 1.2). The diversity in the *Aspergillus* genomes reflects their distant evolutionary relationships and makes them an ideal model for the study of comparative and functional genomics. The genomes are made up of 8 chromosomes, range between ~28 Mb and ~37 Mb, are roughly equally composed of non-coding and coding regions, and contain between ~9,000 and ~13,000 genes (Galagan, Calvo et al. 2005; Machida, Asai et al. 2005; Nierman, Pain et al. 2006; Payne, Nierman et al. 2006; Pel, de Winde et al. 2007; Fedorova, Khaldi et al. 2008; Andersen, Salazar et al. 2011; Sato, Oshima et al. 2011).

Analysis of the *Aspergillus* genomes has yielded several biological insights. For example, comparison of gene content between the pathogen *A. fumigatus* and the closely related and rarely pathogenic *A. fischerianus*, revealed hundreds of genes unique to *A. fumigatus* some of which appear likely involved in virulence (Nierman, Pain et al. 2006). Moreover, genome-wide expression analysis during temperature conditions mimicking the human body identified several genes contributing to thermotolerance in *A. fumigatus* (Nierman, Pain et al. 2006). Study of *A. oryzae* in comparison to *A. fumigatus* and *A. nidulans* brought to light a large-scale genome expansion. Interestingly genes unique to the expanded regions were enriched for metabolic and hydrolytic enzymes; the very functions which make the species an ideal industrial organism (Machida, Asai et al. 2005).

Additionally, multiple strains of *A. fumigatus* and *A. niger* have been sequenced and have uncovered genomic features pertinent to their specializations. Comparative analysis of *A. fumigatus* identified 2% of genes which were unique to each isolate (Fedorova, Khaldi et al. 2008). Of particular interest were two genomic islands unique to one isolate which may play a role in stress response. One of the regions contained genes functioning in arsenic detoxification while the other was associated in the response to osmotic and heavy metal stress (Fedorova, Khaldi et al. 2008). Furthermore, substantial changes between industrial isolates of *A. niger*, including the unique presence of alpha-amylase in the "enzyme factory" strain, highlight the genetic underpinnings of their different commercial applications (Andersen, Salazar et al. 2011).

Chapter Previews

In this dissertation, I analyze the *Aspergillus* genomes to investigate the pathogenicity of *A. fumigatus* and the domestication of *A. oryzae* while also broadly examining the function and evolution of repetitive DNA. In Chapter II, I characterize the transcriptional profile of *A. fumigatus* during its biofilm-like state, where it exhibits heightened pathogenicity and reduced susceptibility to drug treatment. To gain insight into the regulatory response underlying this phenotypic transition, I compared the gene expression differences of a single clinical strain of *A. fumigatus* during *in vitro* biofilm and non-biofilm growth. My analysis was guided by a core set of questions: (1) which genes contribute to enhanced virulence, (2) which genes contribute to drug resistance, and (3) which genes regulate the transition to the biofilm morphology? These results will

add valuable insight into the genetics and transcriptomics of biofilm formation while also generating a resource of candidate genes for use in future functional experiments.

In Chapter III, I evaluate the population biology of Dutch *A. fumigatus* isolates in the context of understanding the potential spread of drug resistance, in the country where it likely originated. I sought to determine if population structure and reproductive lifestyle shape the evolutionary patterns of drug resistance. Specifically I asked, (1) does *A. fumigatus* have population structure in the Netherlands, if so, (2) are resistant isolates distinct, and (3) is sexual recombination contributing to the spread of resistance? The outcomes of this study are critical in the detection, control and treatment of *A. fumigatus* infections and will add to the overall limited knowledge of *A. fumigatus* population biology.

Next, in Chapter IV, I take a holistic approach to understanding how domestication has shaped the genome of *A. oryzae*. Although the domestication of *A. oryzae* from its progenitor *A. flavus* is well supported, there are only few hints into the genetic and functional changes resulting from this event. Here, I examine genome-wide sequence variation, gene expression profiles and protein abundance differences between the two species to assess the outcome of thousands of years of human induced selective pressures. The given results are discussed in comparison to plant and animal domestication models, which have been more comprehensively studied.

The availability of *Aspergillus* genomes also lends itself to addressing more broad evolutionary questions. I will focus on one specific genomic characteristic common to eukaryotes: the presence of tandemly repeated DNA (Li, Korol et al. 2002). These repeated motifs vary in size but are highly polymorphic due to their underlying mutation mechanisms. In the case of smaller tandem repeats, during replication the complementary and template strands can become temporarily disassociated, incorrectly realigned to a different repeat motif and incorporate the mutation into the newly synthesized strand (Levinson and Gutman 1987). Alternatively, recombination in the form of unequal crossing over mediated by homologous repetitive sequences, better models mutational trends of larger tandem repeats (Hancock 1999). Mutation rates of tandem repeats can be several orders of magnitude greater than single nucleotide polymorphisms (Drake, Charlesworth et al. 1998) and importantly, are abundant, subtle, rarely deleterious, and potentially reversible (Kashi and King 2006). Notably, tandem repeats are present at considerably higher levels than predicted by chance (Dieringer and Schlotterer 2003) and have been associated with phenotypic variation (Sawyer, Hennessy et al. 1997; Fondon and Garner 2004; Verstrepen, Jansen et al. 2005; Mirkin 2007; Vinces, Legendre et al. 2009). The *Aspergillus* genomes present a model system on which to advance our understanding about the evolution and function of tandemly repeated DNA.

In Chapter V, I first survey the distribution and variation of tandem repeats across the coding regions of 10 *Aspergillus* genomes. I examine their evolutionary stability, composition and functional associations and relate these results to their implications in *Aspergillus* biology and overall patterns found in other eukaryotes. In Chapter VI, I

directly test the longstanding but sparsely studied hypothesis that subtle variation in tandem repeat length acts as a phenotypic tuning knobs (King 1997, Kashi 1997). Specifically, I examine the relationship between promoter region tandem repeat length and gene expression of over 140 loci in 16 isolates of *A. flavus* and *A. oryzae*.

Collectively, my dissertation addresses looming yet central questions concerning pathogenicity and evolution by utilizing the *Aspergillus* genomes in combination with functional analysis.

REFERENCES

Amaike, S. and N. P. Keller (2011). "Aspergillus flavus." Annual Review of Phytopathology, Vol 49 **49**: 107-133.

Andersen, M. R., M. P. Salazar, et al. (2011). "Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88." Genome Research **21**(6): 885-897.

Baker, S. E. a. B. J. W. (2008). An overview of the genus Aspergillus. The Aspergilli: genomics, medical aspects, biotechnology, and research methods. G. H. Goldman, Osmani S. A. New York, CRC Press**:** 3-14.

Beauvais, A., C. Schmidt, et al. (2007). "An extracellular matrix glues together the aerial-grown hyphae of Aspergillus fumigatus." Cellular Microbiology **9**(6): 1588-1600.

Bennett, J. W. and M. A. Klich (1992). Aspergillus : biology and industrial applications. Boston, Butterworth-Heinemann.

Bhabhra, R. and D. S. Askew (2005). "Thermotolerance and virulence of Aspergillus fumigatus: role of the fungal nucleolus." Med Mycol **43 Suppl 1**: S87-93.

Brakhage, A. A. and K. Langfelder (2002). "Menacing mold: The molecular biology of Aspergillus fumigatus." Annual Review of Microbiology **56**: 433-455.

Cardwell, K. F. and P. J. Cotty (2002). "Distribution of Aspergillus section flavi among field soils from the four agroecological zones of the Republic of Benin, West Africa." Plant Disease **86**(4): 434-439.

Carlile, M. J., S. C. Watkinson, et al. (2001). The fungi. San Diego, Calif. ; London, Academic Press.

Denning, D. W. (1998). "Invasive aspergillosis." Clinical Infectious Diseases **26**(4): 781-803.

Dieringer, D. and C. Schlotterer (2003). "Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species." Genome Research **13**(10): 2242-2251.

Drake, J. W., B. Charlesworth, et al. (1998). "Rates of spontaneous mutation." Genetics **148**(4): 1667-1686.

Dyer, P. S. and C. M. O'Gorman (2012). "Sexual development and cryptic sexuality in fungi: insights from Aspergillus species." Fems Microbiology Reviews **36**(1): 165-192.

Fedorova, N. D., N. Khaldi, et al. (2008). "Genomic Islands in the Pathogenic Filamentous Fungus Aspergillus fumigatus." PLoS Genet **4**(4): e1000046.

Fisher, M. C. and D. A. Henk (2012). "Sex, drugs and recombination: the wild life of Aspergillus." Molecular Ecology **21**(6): 1305-1306.

Fondon, J. W. and H. R. Garner (2004). "Molecular origins of rapid and continuous morphological evolution." Proceedings of the National Academy of Sciences of the United States of America **101**(52): 18058-18063.

Galagan, J. E., S. E. Calvo, et al. (2005). "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." Nature **438**(7071): 1105-15.

Geiser, D. M. (2009). "Sexual structures in Aspergillus: morphology, importance and genomics." Med Mycol **47 Suppl 1**: S21-6.

Geiser, D. M., J. I. Pitt, et al. (1998). "Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus." Proceedings of the National Academy of Sciences of the United States of America **95**(1): 388-393.
Geiser, D. M., J. W. Taylor, et al. (1998). "Cause of sea fan death in the West Indies." Nature **394**(6689): 137-138.
Gibbons, J. G., L. Salichos, et al. (2012). "The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*." Current Biology ***in press***.
Gugnani, H. C. (2003). "Ecology and taxonomy of pathogenic Aspergilli." Frontiers in Bioscience-Landmark **8**: S346-S357.
Hancock, J. M. (1999). Microsatellites and other simple sequences: genomic context and mutational mechanisms. Microsatellites : evolution and applications. Oxford ; New York, Oxford University Press**:** 1-9.
Hoffmeister, D. and N. P. Keller (2007). "Natural products of filamentous fungi: enzymes, genes, and their regulation." Natural Product Reports **24**(2): 393-416.
Hospenthal, D. R., K. J. Kwon-Chung, et al. (1998). "Concentrations of airborne Aspergillus compared to the incidence of invasive aspergillosis: lack of correlation." Medical Mycology **36**(3): 165-168.
Howard, S. J., D. Cerar, et al. (2009). "Frequency and Evolution of Azole Resistance in Aspergillus fumigatus Associated with Treatment Failure." Emerging Infectious Diseases **15**(7): 1068-1076.
Howard, S. J., I. Webster, et al. (2006). "Multi-azole resistance in Aspergillus fumigatus." Int J Antimicrob Agents **28**(5): 450-3.
Kashi, Y. and D. G. King (2006). "Simple sequence repeats as advantageous mutators in evolution." Trends Genet **22**(5): 253-9.
Keller, N. P., G. Turner, et al. (2005). "Fungal secondary metabolism - From biochemistry to genomics." Nature Reviews Microbiology **3**(12): 937-947.
Klaassen, C. H. W., J. G. Gibbons, et al. (2012). "Evidence for genetic differentiation and variable recombination rates among Dutch populations of the opportunistic human pathogen Aspergillus fumigatus." Molecular Ecology **21**(1): 57-70.
Krishnamachari, K. A. V. R., R. V. Bhat, et al. (1975). "Investigations into an Outbreak of Hepatitis in Parts of Western India." Indian Journal of Medical Research **63**(7): 1036-&.
Kurtzman, C. P., M. J. Smiley, et al. (1986). "DNA Relatedness among Wild and Domesticated Species in the Aspergillus-Flavus Group." Mycologia **78**(6): 955-959.
Kusumoto, K., Y. Nogata, et al. (2000). "Directed deletions in the aflatoxin biosynthesis gene homolog cluster of Aspergillus oryzae." Current Genetics **37**(2): 104-111.
Latge, J. P. (1999). "Aspergillus fumigatus and aspergillosis." Clinical Microbiology Reviews **12**(2): 310-+.
Latge, J. P. (2001). "The pathobiology of Aspergillus fumigatus." Trends in Microbiology **9**(8): 382-389.
Levinson, G. and G. A. Gutman (1987). "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." Mol Biol Evol **4**(3): 203-21.

Li, Y. C., A. B. Korol, et al. (2002). "Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review." Molecular Ecology **11**(12): 2453-2465.
Liese, A., K. Seelbach, et al. (2000). Industrial biotransformations. Weinheim ; New York, Wiley-VCH.
Loussert, C., C. Schmitt, et al. (2010). "In vivo biofilm composition of Aspergillus fumigatus." Cellular Microbiology **12**(3): 405-410.
Machida, M., K. Asai, et al. (2005). "Genome sequencing and analysis of Aspergillus oryzae." Nature **438**(7071): 1157-61.
Machida, M. and K. Gomi (2010). Aspergillus : molecular biology and genomics. Wymondham, Caister Academic.
Machida, M., O. Yamada, et al. (2008). "Genomics of Aspergillus oryzae: Learning from the History of Koji Mold and Exploration of Its Future." DNA Research **15**(4): 173-183.
Mellado, E., G. Garcia-Effron, et al. (2007). "A new Aspergillus fumigatus resistance mechanism conferring in vitro cross-resistance to azole antifungals involves a combination of cyp51A alterations." Antimicrobial Agents and Chemotherapy **51**(6): 1897-1904.
Mirkin, S. M. (2007). "Expandable DNA repeats and human disease." Nature **447**(7147): 932-40.
Mowat, E., S. Lang, et al. (2008). "Phase-dependent antifungal activity against Aspergillus fumigatus developing multicellular filamentous biofilms." Journal of Antimicrobial Chemotherapy **62**(6): 1281-1284.
Mowat, E., C. Williams, et al. (2009). "The characteristics of Aspergillus fumigatus mycetoma development: is this a biofilm?" Medical Mycology **47**: S120-S126.
Nesbitt, B. F., J. O'Kelly, et al. (1962). "Aspergillus flavus and turkey X disease. Toxic metabolites of Aspergillus flavus." Nature **195**: 1062-3.
Nierman, W. C., A. Pain, et al. (2006). "Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus (vol 438, pg 1151, 2005)." Nature **439**(7075): 1151-1156.
Osbourn, A. (2010). "Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation." Trends in Genetics **26**(10): 449-457.
Osmani, S. A. and G. H. Goldman (2008). The Aspergilli : genomics, medical aspects, biotechnology, and research methods. Boca Raton, Taylor & Francis.
Paoletti, M., C. Rydholm, et al. (2005). "Evidence for sexuality in the opportunistic fungal pathogen Aspergillus fumigatus." Current Biology **15**(13): 1242-1248.
Papagianni, M. (2007). "Advances in citric acid fermentation by Aspergillus niger: Biochemical aspects, membrane transport and modeling." Biotechnology Advances **25**(3): 244-263.
Payne, G. A., W. C. Nierman, et al. (2006). "Whole genome comparison of Aspergillus flavus and A-oryzae." Medical Mycology **44**: S9-S11.
Pel, H. J., J. H. de Winde, et al. (2007). "Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88." Nat Biotechnol **25**(2): 221-31.
Peterson, S. W. (2008). "Phylogenetic analysis of Aspergillus species using DNA sequences from four loci." Mycologia **100**(2): 205-226.

Pitt, J. I. (2000). "Toxigenic fungi and mycotoxins." British Medical Bulletin **56**(1): 184-192.
Punt, P. J., N. van Biezen, et al. (2002). "Filamentous fungi as cell factories for heterologous protein production." Trends in Biotechnology **20**(5): 200-206.
Rajendran, R., E. Mowat, et al. (2011). "Azole Resistance of Aspergillus fumigatus Biofilms Is Partly Associated with Efflux Pump Activity." Antimicrobial Agents and Chemotherapy **55**(5): 2092-2097.
Ramirez-Prado, J. H., G. G. Moore, et al. (2008). "Characterization and population analysis of the mating-type genes in Aspergillus flavus and Aspergillus parasiticus." Fungal Genetics and Biology **45**(9): 1292-1299.
Samson, R. A. and J. n. Varga (2007). Aspergillus systematics in the genomic era. Utrecht, The Netherlands, CBS Fungal Biodiversity Centre.
Sato, A., K. Oshima, et al. (2011). "Draft Genome Sequencing and Comparative Analysis of Aspergillus sojae NBRC4239." DNA Research **18**(3): 165-176.
Sawyer, L. A., J. M. Hennessy, et al. (1997). "Natural variation in a Drosophila clock gene and temperature compensation." Science **278**(5346): 2117-2120.
Schmitt, H. J., A. Blevins, et al. (1990). "Aspergillus species from hospital air and from patients." Mycoses **33**(11-12): 539-41.
Seidler, M. J., S. Salvenmoser, et al. (2008). "Aspergillus fumigatus Forms Biofilms with Reduced Antifungal Drug Susceptibility on Bronchial Epithelial Cells." Antimicrobial Agents and Chemotherapy **52**(11): 4130-4136.
Sheehan, D. J., C. A. Hitchcock, et al. (1999). "Current and emerging azole antifungal agents." Clinical Microbiology Reviews **12**(1): 40-+.
Snelders, E., H. A. L. van der Lee, et al. (2008). "Emergence of Azole Resistance in Aspergillus fumigatus and Spread of a Single Resistance Mechanism." Plos Medicine **5**(11): 1629-1637.
Tobert, J. A. (2003). "Lovastatin and beyond: The history of the HMG-CoA reductase inhibitors." Nature Reviews Drug Discovery **2**(7): 517-526.
Tominaga, M., Y. H. Lee, et al. (2006). "Molecular analysis of an inactive aflatoxin biosynthesis gene cluster in Aspergillus oryzae RIB strains." Applied and Environmental Microbiology **72**(1): 484-490.
Verstrepen, K. J., A. Jansen, et al. (2005). "Intragenic tandem repeats generate functional variability." Nature Genetics **37**(9): 986-990.
Verweij, P. E., D. T. A. T. Dorsthorst, et al. (2002). "Nationwide survey of in vitro activities of itraconazole and voriconazole against clinical Aspergillus fumigatus isolates cultured between 1945 and 1998." Journal of Clinical Microbiology **40**(7): 2648-2650.
Vinces, M. D., M. Legendre, et al. (2009). "Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability." Science **324**(5931): 1213-1216.
Wicklow, D. T. (1968). "Aspergillus Fumigatus Fresenius Isolated from Ornithogenic Soil Collected at Hallett Station Antarctica." Canadian Journal of Microbiology **14**(6): 717-&.
Williams, J. H., T. D. Phillips, et al. (2004). "Human aflatoxicosis in developing countries: a review of toxicology, exposure, potential health consequences, and interventions." American Journal of Clinical Nutrition **80**(5): 1106-1122.

Yu, J., T. E. Cleveland, et al. (2005). "Aspergillus flavus genomics: gateway to human and animal health, food safety, and crop resistance to diseases." Rev Iberoam Micol **22**(4): 194-202.
Zhang, X. J., T. D. Yao, et al. (2002). "Microorganisms in a high altitude glacier ice in Tibet." Folia Microbiologica **47**(3): 241-245.

CHAPTER II

GLOBAL TRANSCRIPTOME CHANGES UNDERLYING COLONY GROWTH IN THE OPPORTUNISTIC HUMAN PATHOGEN *ASPERGILLUS FUMIGATUS*

John G. Gibbons[1], Anne Beauvais[2], Remi Beau[2], Kriston L. McGary[1], Jean-Paul Latge[2] and Antonis Rokas[1]

[1]*Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN, 37235, United States of America*

[2]*Unité des Aspergillus, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*

This chapter is published in *Eukaryotic Cell*, 2012, 11: 68-78.

ABSTRACT

Aspergillus fumigatus is the most common and deadly pulmonary fungal infection worldwide. In the lung, the fungus usually forms a dense colony of filaments embedded in a polymeric extracellular matrix. To identify candidate genes involved in this biofilm growth (BF), we used RNA-Seq to compare the transcriptomes of BF and liquid plankton growth (PL). Sequencing and mapping of tens of millions sequence reads against the *A. fumigatus* transcriptome identified 3,728 differentially regulated genes in the two conditions. Although many of these genes, including the ones encoding for transcription factors, stress response, the ribosome and the translation machinery, likely reflect the different growth demands in the two conditions, our experiment also identified hundreds of candidate genes for the observed differences in morphology and pathobiology between BF and PL. We found an overrepresentation of up-regulated genes in transport, secondary metabolism, and cell wall and surface functions. Furthermore, up-regulated genes showed significant spatial structure across the *A. fumigatus* genome; they were more likely to occur in subtelomeric regions and co-localized in 27 genomic neighborhoods, many of which overlapped with known or candidate secondary metabolism gene clusters. We also identified 1,164 genes that were down-regulated. This gene set was not spatially structured across the genome and was overrepresented in genes participating in primary metabolic functions, including carbon and amino acid metabolism. These results add valuable insight into the genetics of biofilm formation in *A. fumigatus* and other filamentous fungi and identify many relevant, in the context of biofilm biology, candidate genes for downstream functional experiments.

INTRODUCTION

The filamentous fungal genus *Aspergillus* consists of over 250 saprophytic species (Geiser, Klich et al. 2007). Although some species, such as *A. niger*, *A. terreus* and *A. oryzae*, are exploited commercially for the production of enzymes, pharmaceuticals and traditional Asian foods and beverages, others are capable of colonizing and infecting immunocompromised individuals (Baker and Bennett 2008). The primary opportunistic human pathogen is *A. fumigatus*, the fungus responsible for the highest number of deaths and for the second highest number of infections, behind only *Candida albicans* (Latge and Steinbach 2009).

A. fumigatus produces an abundance of very small (2-3μ) asexual spores (also known as conidia) that easily disperse throughout the air (Latge 1999). Their continuous inhalation can lead to the production of a wide spectrum of diseases, which are collectively known as aspergillosis (Latge and Steinbach 2009). In addition to allergic diseases resulting from conidial inhalation, the growth of *A. fumigatus* hyphae can produce either focal infections in preexisting lung cavities or invasive infections in patients undergoing heavy chemo- and radiotherapies for cancer treatments or organ transplantation.

Historically, *in vitro* studies to understand *A. fumigatus* pathobiology used fungal colonies grown in liquid shake conditions. Although these colonies are not single-celled, they deserve the definition of planktonic (PL) colonies since they live in a fluid environment as opposed to the ones attached to a surface. Importantly, these PL colonies have vastly different pathological and morphological characteristics than those observed

in vivo (Loussert, Schmitt et al. 2010). For example, during mycelial development in the lung, the hyphae form and embed themselves in a dense extracellular matrix (ECM) (Loussert, Schmitt et al. 2010). This ECM, whose function, in part, is to tightly "glue" hyphae together and to protect the fungus from an outside hostile environment, is absent when the fungus is grown under liquid shake conditions (Beauvais, Schmidt et al. 2007; Loussert, Schmitt et al. 2010; Muller, Seidler et al. 2011). In contrast, an ECM is produced when the fungus is grown in aerial static conditions (Beauvais, Schmidt et al. 2007). Importantly, in this aerially grown biofilm-like state (BF), the fungus exhibits reduced susceptibility to antifungal drugs (Mowat, Butcher et al. 2007; Mowat, Lang et al. 2008; Seidler, Salvenmoser et al. 2008; Rajendran, Mowat et al. 2011) and undergoes major metabolic changes that are thought to be involved in virulence (Bruns, Seidler et al. 2010; Loussert, Schmitt et al. 2010).

The differences in pathological and morphological characteristics between PL and *in vivo* grown *A. fumigatus* suggest that PL is a poor *in vitro* disease model. In contrast, the BF model is phenotypically close to the *A. fumigatus in vivo* growth and thus more appropriate for pathobiology studies (Bruns, Seidler et al. 2010; Loussert, Schmitt et al. 2010). To provide a global and accurate profile of *A. fumigatus in vitro* biofilm growth, we utilized RNA-Seq (Mortazavi, Williams et al. 2008) to compare the global gene expression profiles of *A. fumigatus* grown in BF and PL conditions. We identified thousands of differentially regulated genes, whose protein products participate in functions such as transcription, translation and stress response, that likely account for the different growth demands associated with the two conditions tested. However, we also

identified hundreds of differentially regulated genes that constitute candidates for the observed pathobiological and morphological differences between the two conditions. For example, we observed extensive up-regulation of genes whose proteins participate in transport, secondary metabolism, and cell wall and surface functions in BF relative to PL. Interestingly, whereas genes that were up-regulated in BF were significantly overrepresented in subtelomeric regions and localized in genomic neighborhoods with similar regulation, BF down-regulated genes exhibited neither of these two trends. Together, our results provide a fine grain transcriptional examination of *A. fumigatus* grown in biofilm conditions and offer numerous candidates for downstream functional experiments.

MATERIALS AND METHODS

Fungal Strains and Culture Conditions

We chose the *A. fumigatus* ATCC 46645 wild type strain, which has been used previously for biofilm studies (Beauvais, Schmidt et al. 2007). For shake cultures (PL), 500 ml of Brian's medium were inoculated with 10^7 conidia and the flask was incubated at 37˚C in darkness for 16 hours at 150 rpm. For the production of a colony on a 2% agar Brian medium (BF), porous cellophane was deposited on the surface of the agar and 50 µl of a conidial suspension (10^7/ml) per 9 cm Petri dish was spread onto the cellophane using an inoculation spreader. The Petri dishes were also incubated in the dark at 37°C for 16 hours.

RNA Isolation, mRNA Library Construction and Illumina Sequencing

Prior to harvesting each culture was visually inspected to ensure the absence of conidia. We collected fungal tissue from the PL culture by filtration, and from the BF culture using a sterile spatula. Upon harvesting, the fungal colony from the BF culture was immediately frozen in liquid nitrogen and ground it into a fine powder with a mortar and pestle. We extracted total RNA using a previously described phenol/chloroform protocol (Lamarre, Sokol et al. 2008). RNA samples were then treated with DNase and further purified using the Qiagen RNeasy mini kit, following manufacturer's instructions (Gibbons, Janson et al. 2009; Hittinger, Johnston et al. 2010). mRNA libraries were constructed and sequenced at the Vanderbilt Genome Technology Core following Illumina specifications, generating over 20 million reads for each sample. To establish that the RNA-Seq experiments reported in our study do not show significant variation

when replicated, we also generated data from two additional *A. fumigatus* strains (Af293 and CEA10) during PL growth and from one additional technical replicate during BF growth.

Read Mapping and Gene Regulation Quantification

For each dataset, we converted fastq sequence read files to fasta formatted ones using the fq_all2std.pl script in the Maq software package, version 0.7.1 (http://maq.sourceforge.net/index.shtml). We then mapped each dataset to the *A. fumigatus* af293 reference transcriptome (Nierman, Pain et al. 2005) using the SeqMap software, version 1.08 (Jiang and Wong 2008, http://www.stanford.edu/group/wonglab/jiangh/seqmap/), allowing two mismatches per read. We mapped 11,842,153 and 15,385,865 reads to the *A. fumigatus* af293 reference transcriptome (Nierman, Pain et al. 2005) from BF and PL growth, respectively. To quantify global gene expression levels in BF and PL conditions, we calculated the RPKM value, a self-normalized value of absolute transcript abundance, of each reference transcript in the two datasets using the rSeq software, version 0.0.5 (Mortazavi, Williams et al. 2008, http://www.stanford.edu/group/wonglab/jiangh/rseq/).

Identification of Differentially Regulated Genes

To limit the number of false positives, we employed a conservative approach to identify differentially expressed genes by implementing both biological and statistical cutoffs for significance (Pitts, Rinker et al. 2011). For our biological cutoff, we compared the fold difference of RPKM values between BF and PL conditions by calculating the relative

RPKM (rRPKM = $RPKM_{BF}$ / $RPKM_{PL}$) of each gene. We required a 2-fold difference in relative gene expression between conditions (i.e., rRPKM $\geq$ 2 or rRPKM $\leq$ 0.5). For our statistical cutoff, we compared the proportion of reads that mapped to each gene for the BF and PL conditions via Fisher's exact tests, applying a Bonferroni multiple test corrected *p-value* cutoff of 5.5e-06.

The Genome Architecture of Differentially Regulated Genes

To test whether differentially regulated genes were represented disproportionately in subtelomeric regions, we compared the proportion of up-regulated and down-regulated genes in subtelomeric regions to the proportion of up-regulated and down-regulated genes in the rest of the genome (background), independently for each chromosome and for the entire genome. We defined subtelomeric regions as the 300kb regions preceding the telomere ends (Fedorova, Khaldi et al. 2008; McDonagh, Fedorova et al. 2008). We implemented a Bonferroni multiple test corrected *p-value* cutoff of 0.0028.

To test whether differential gene regulation along the *A. fumigatus* chromosomes was structured spatially into genomic neighborhoods, we analyzed the gene order pattern of up-regulated and down-regulated genes across each chromosome as well as across the entire genome. To assess statistical significance we used the Wald-Wolfowitz runs test (Sokal and Rohlf 1995), which compares the expected number of runs, assuming randomness, to the observed number of runs. In our analyses, we coded each gene as "up-regulated" or "non-up-regulated" and defined "runs" as strings of up-regulated genes that

were preceded and followed by non-up-regulated genes or vice versa. We implemented a Bonferroni multiple test corrected *p-value* cutoff of 0.0028.

To characterize further the gene content of genomic neighborhoods of differential gene regulation, we performed sliding window analysis to identify significant clusters of up-regulated or down-regulated genes. Each gene was encoded as up-regulated or down-regulated using the expression cutoffs previously described. For each chromosome, a probability was calculated for each window of 6, 12, or 24 consecutive genes, with a step size of one gene. The probability of each window being up-regulated (or down-regulated) was calculated using a cumulative binomial probability, where the probability of success is the fraction of genes that are up-regulated (or down-regulated). Thus, the probability of a window with X out of N genes being up-regulated (or down-regulated) is given by the following equation:

$$p(X >= k) = \sum_{i=k}^{n} \binom{n}{k} p^i (1-p)^{(n-i)}$$

To account for the testing of multiple hypotheses, we used a false discovery rate (FDR) cutoff of 0.15, where the FDR was estimated empirically. We randomly shuffled expression levels across the genome and calculated the probability for each window in permuted datasets. A false positive (FP) is a window from a random set and a true positive (TP) is a window from the real data. To increase precision of the estimated FDR, we performed 1000 permutations; therefore, the false positives were divided by 1000 to

account for the number of random permutations. For each window size, the cutoff probability for significant windows was determined by calculating a cumulative FDR, where $FDR = [FP / (TP + FP)]$, and choosing the probability with an FDR of 0.1. Regions of chromosomes with overlapping significant windows were manually concatenated and trimmed for clarity and discussion.

Functional Associations of Differentially Regulated Genes

To examine whether differentially regulated genes were preferentially associated with certain functions, we compared the proportion of up-regulated and down-regulated genes belonging to the 2nd and 3rd order FunCat categories (Ruepp, Zollner et al. 2004) and to the 109 *A. fumigatus* annotated KEGG pathways (Kanehisa and Goto 2000) to their corresponding values in the rest of the genome (defined as the number of genes in all other categories). We performed all comparisons using Fisher's exact tests (Sokal and Rohlf 1995). Because the total number of genes in certain FunCat categories and KEGG pathways was small, we did not apply a multiple test corrected *p-value* cutoff and instead used a cutoff of 0.05.

We also examined the functional associations of differentially expressed genes for a number of gene sets with functions relevant to BF or PL growth. These included 409 genes encoding antigens and cell-surface proteins (Latge, Mouyna et al. 2005; The GPI Lipid Anchor Project, http://mendel.imp.ac.at/gpi/gpi_genomes.html; Mari and Scala 2006, The Allergome Project, http://www.allergome.org), 81 genes encoding allergens (Nierman, Pain et al. 2005; Mari and Scala 2006, The Allergome Project,

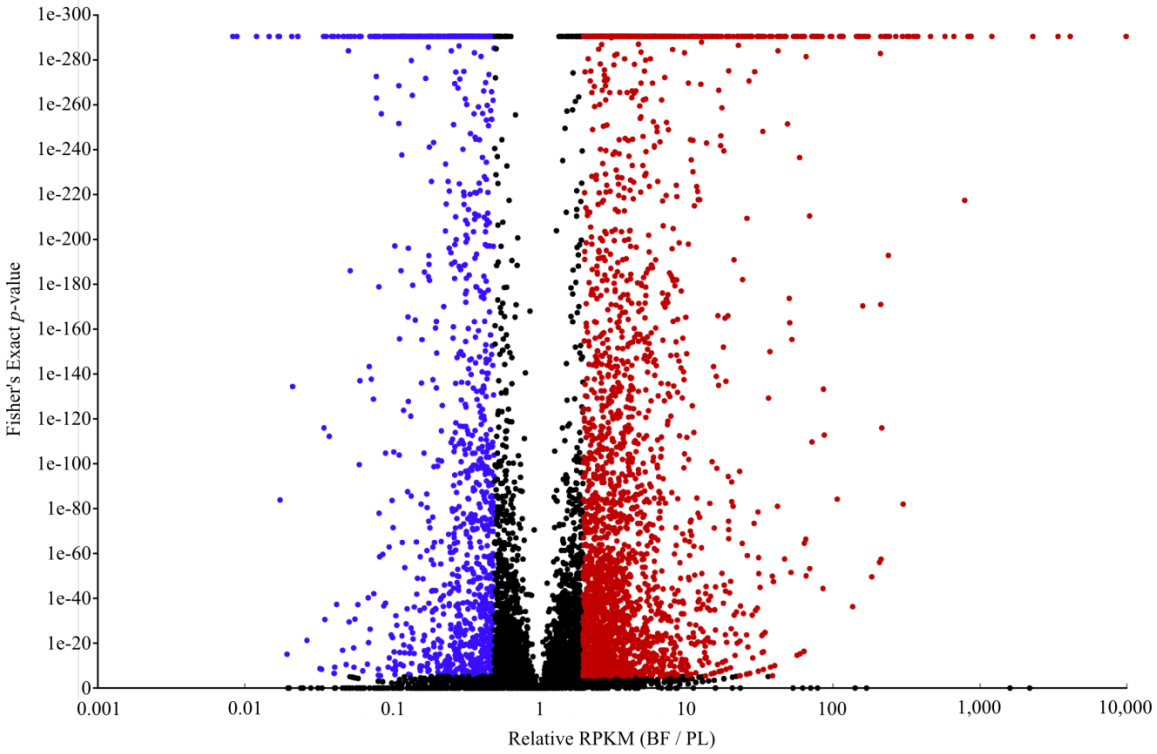
http://www.allergome.org), 319 genes encoding the major facilitator superfamily (MFS) and ATP-Binding Cassette (ABC) membrane transporter proteins (Nierman, Pain et al. 2005; Ren, Chen et al. 2007, http://www.membranetransport.org/), 392 transcription factors (Beri, Whittington et al. 1987; Andrianopoulos and Hynes 1988; Lints, Davis et al. 1995; Vallim, Miller et al. 2000; Strittmatter, Irniger et al. 2001; Mulder, Saloheimo et al. 2004; Nierman, Pain et al. 2005; Vienken, Scherer et al. 2005; Grosse and Krappmann 2008; Soriani, Malavazi et al. 2008; Gravelat, Ejzykowicz et al. 2010) and the 383 genes present in 22 secondary metabolism *A. fumigatus* gene clusters (Perrin, Fedorova et al. 2007).

RESULTS

Thousands of Genes are Differentially Regulated in BF Relative to PL

Of the 9,887 *A. fumigatus* transcripts (Nierman, Pain et al. 2005; Fedorova, Khaldi et al. 2008), 9,525 were expressed in one or both conditions and 362 genes were not expressed in either condition. Of the 9,525 expressed genes, 251 and 175 genes were uniquely expressed in BF and PL growth, respectively. Of the remaining 9,099 genes that were expressed in both conditions, 5,380 showed uniform expression, 2,565 were up-regulated and 1,164 were down-regulated in BF relative to PL (Figure 2.1)

Figure 2.1. Differentially regulated genes between aerial (BF) and submerged (PL) growth. For each gene, the rRPKM value ($RPKM_{BF}$ / $RPKM_{PL}$) was plotted against its respective Fisher's exact p-value. P-values smaller than 1e-290 were reported as 1e-290. The red and blue points correspond to the biologically (2-fold difference) and statistically significant (p < 5.5e-06) up-regulated and down-regulated genes between BF and PL, respectively.

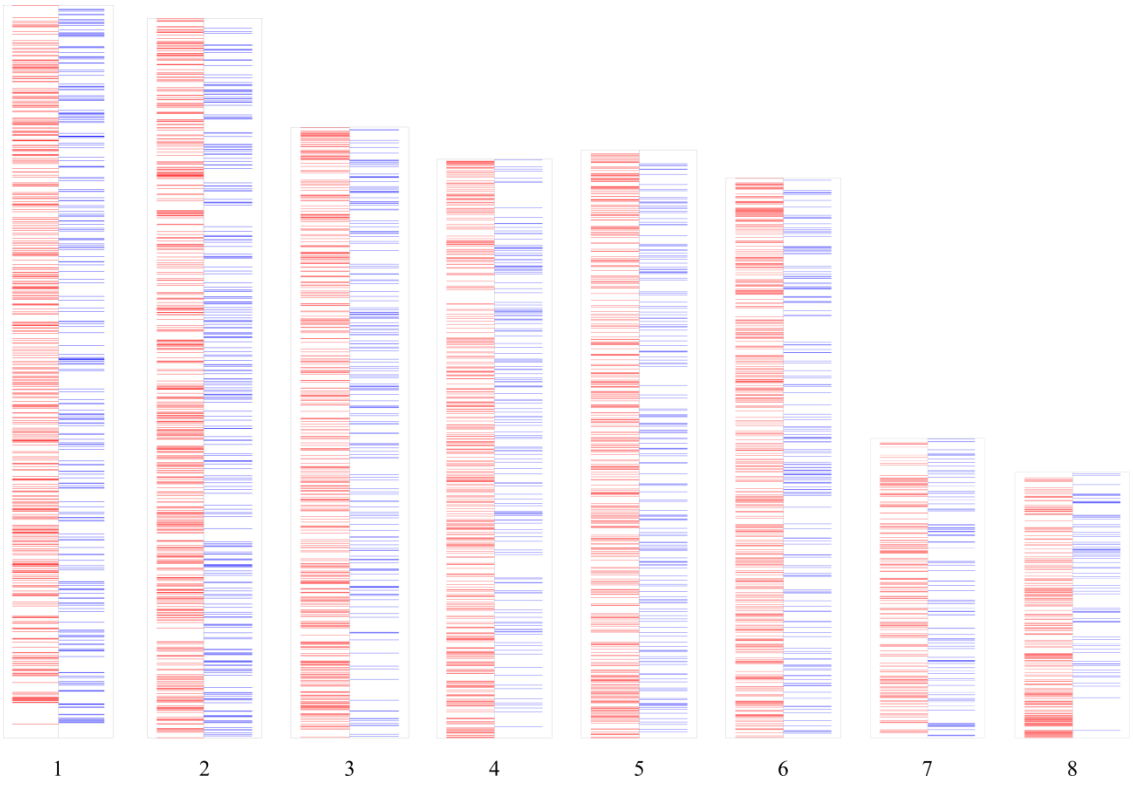


Remarkably, the range of expression values in both samples ranged 7 orders of magnitude. Our experiments showed very high levels of correlation for both the biological and technical replication experiments (in all cases examined, Pearson's $r >$ 0.91, Figure S2.1), on par with similar studies in the literature (Bruno, Wang et al. 2010), indicating that there is very little technical or biological variation for the conditions tested.

Non-random Distribution of Differentially Regulated Genes in the *A. fumigatus* Genome

The subtelomeric regions of the *A. fumigatus* genome harbor clusters of highly variable and recently evolved genes (Fedorova, Khaldi et al. 2008). Considering this, we mapped the chromosomal locations of up- and down-regulated genes to examine whether differentially expressed genes were overrepresented in subtelomeric regions and for the presence of physically linked clusters of co-regulated genes (Figure 2.2). We found significant over-representation of up-regulated genes (p = 2.45e-06) and under-representation of down-regulated genes (p = 4.5e-21) in subtelomeric regions of the *A. fumigatus* genome. Individual chromosomes showed similar trends. For example, chromosome 5 was significantly enriched in up-regulated genes (p = 0.0003) and significantly under-represented in down-regulated genes (p = 0.0005). The only exception was chromosome 1, in which up-regulated genes were significantly under-represented in its subtelomeric regions (p = 1.46e-08).

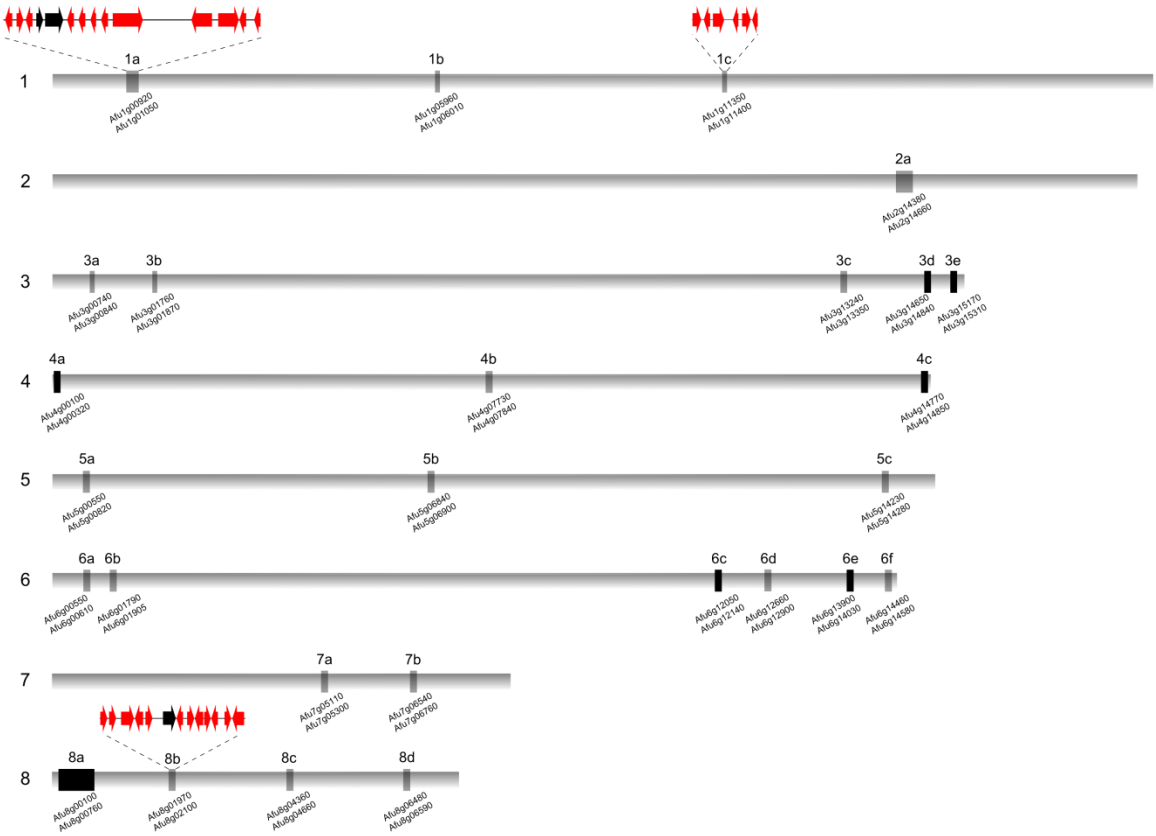
Figure 2.2. The genome-wide distribution of differentially regulated genes. All differentially regulated genes are plotted against the eight *A. fumigatus* chromosomes. Genes are plotted by order (and not physical distance) on the chromosome. For each chromosome, red bars indicate the positions of up-regulated genes, whereas blue bars indicate the positions of down-regulated genes. Genes that are not differentially regulated are not shown.



Up-regulated and down-regulated genes were not randomly distributed across the genome but instead tended to reside within certain genomic neighborhoods ($p_{\text{up-regulated}}$ = 1.09e-22, $p_{\text{down-regulated}}$ = 6.33e-28). Generally, individual chromosome analyses exhibited the same pattern with the exception of chromosome 4 for the up-regulated gene set (p = 0.1166) and chromosomes 2, 4 and 5 for the down-regulated gene set (p = 0.0030, p = 1 and p = 0.1000, respectively). To better characterize the gene content of these genomic neighborhoods, we further scanned the genome using sliding window analyses and

identified 27 up-regulated gene clusters, but no down-regulated ones (Figure 2.3). The percentage of genes that were up-regulated on a typical gene cluster was 72%.

Figure 2.3. *De novo* identification of gene clusters that are up-regulated in BF in the *A. fumigatus* genome. Each bar corresponds to one of the eight *A. fumigatus* chromosomes. Boxes indicate the up-regulated gene clusters identified by our sliding window analysis. Boxes colored in black correspond to annotated secondary metabolism gene clusters (Perrin, Fedorova et al. 2007), while gray boxes correspond to novel up-regulated clusters of unknown function. The first and last gene names of each cluster are given below each box. Three examples of novel up-regulated gene clusters are shown in greater detail. In these clusters, we have indicated order of gene transcription, as well as color-coded up-regulated (in red) and non-differentially regulated (in black) genes (we did not identify any down-regulated genes).



Seven of these 27 up-regulated gene clusters overlapped with clusters predicted or known to be involved in secondary metabolism (Nierman, Pain et al. 2005; Perrin, Fedorova et

al. 2007) (Figure 2.3). Interestingly, 18 of the other 20 clusters contain genes that are hallmarks for secondary metabolism, including polyketide synthases, membrane transporters, transcription factors, cytochrome p450s, reductases and transferases. For example, a 14-gene cluster (cluster 1A, Figure 2.3) on chromosome 1 contains a putative polyketide synthase, a MFS transporter, a FAD-dependent oxidase, a short chain dehydrogenase/reductase, an oxidoreductase, a NACHT domain protein, a transposase and a reverse transcriptase, whereas a 6 gene cluster (cluster 1C, Figure 2.3) also on chromosome 1 contains a cytochrome p450, a MFS transporter, an integral membrane protein, an aldehyde reductase, an oxidoreductase and a hydrolase. However, our search also identified clusters unlikely to be involved in secondary metabolism, such as a 13-gene cluster (cluster 8B, Figure 2.3) on chromosome 8 that contains 3 glycosyl transferases, a β-glucosidase, a glycan biosynthesis protein, two putative sugar transporters, a xylitol dehydrogenase, an extracellular endo-polygalacturonase, and a C6 transcription factor.

Functional Classification of Differentially Regulated Genes

To gauge the functions of differentially expressed genes, we compared the proportion of up-regulated and down-regulated genes in the 2^{nd} and 3^{rd} level FunCat categories (Ruepp, Zollner et al. 2004) and the annotated KEGG pathways (Kanehisa and Goto 2000) to the background. We found that 9 and 17 2^{nd} and 3^{rd} level FunCat categories were overrepresented in the up-regulated gene set, respectively, compared to 6 and 13 in the down-regulated gene set. Additionally, 10 and 10 KEGG pathways were overrepresented in the up-regulated and down-regulated gene sets, respectively. Collectively, these results

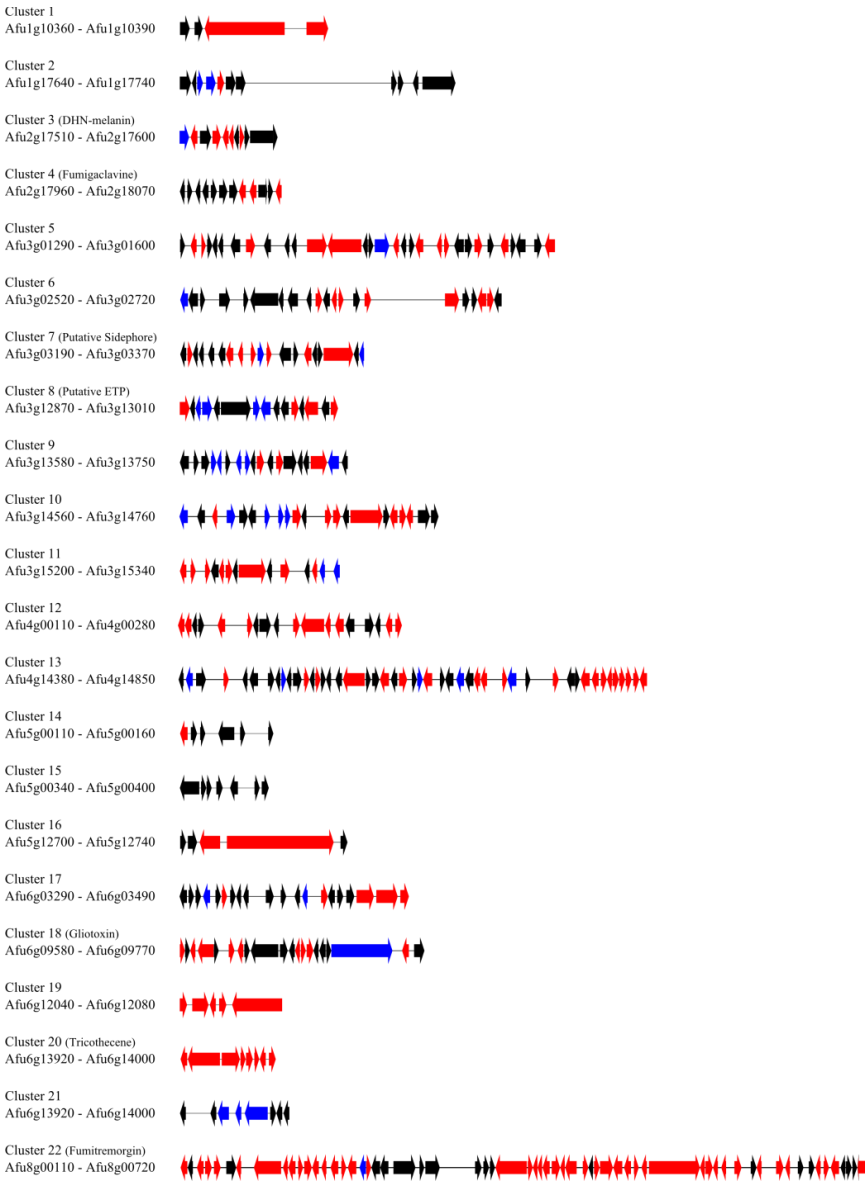
show an overrepresentation of up-regulated genes involved in translation and transport in BF relative to PL, and an overrepresentation of down-regulated genes involved in primary metabolism.

Up-regulation of secondary metabolic gene clusters

The identification of putative and known secondary metabolic gene clusters in up-regulated genomic neighborhoods and the up-regulation of the transcription factor *LaeA*, a global regulator of secondary metabolism in *A. fumigatus* (Perrin, Fedorova et al. 2007) (see the Results section *Widespread Differential Regulation of Transcription Factors*), prompted us to also examine differential expression in all known *A. fumigatus* secondary metabolism gene clusters. We used the 383 genes in 22 gene clusters described by Perrin and co-workers as the set of known *A. fumigatus* secondary metabolism gene clusters (Perrin, Fedorova et al. 2007). We observed widespread up-regulation of secondary metabolism genes and entire clusters. Specifically, 165 of the 383 genes were up-regulated (122) or uniquely expressed in BF (43), whereas only 35 of the 383 were down-regulated in BF (27) or uniquely expressed in PL (8) (Figure 2.4), suggesting that the up-regulated gene set is significantly enriched in genes involved in secondary metabolism relative to background (p = 2.7e-18). For example, all 8 trichothecene cluster genes (Ward, Bielawski et al. 2002; Nierman, Pain et al. 2005) were up-regulated in BF relative to PL (Figure 2.4). Most noticeably, 44 of the 62 genes in the fumitremorgen supercluster (Perrin, Fedorova et al. 2007), which encodes enzymes that produce fumitremorgin, pseurotin A, as well as an unknown compound (Maiya, Grundmann et al. 2006; Maiya, Grundmann et al. 2007; Perrin, Fedorova et al. 2007) were up-regulated (32) or uniquely

expressed (12). Within this supercluster, the putative o-methyltransferase CalO6 (Afu8g00200) was the most up-regulated gene in the genome showing a 9,885-fold up-regulation in BF relative to PL.

Figure 2.4. Expression patterns of known secondary metabolism gene clusters. Gene clusters 1 – 22 correspond to the 22 previously characterized secondary metabolism gene clusters in the *A. fumigatus* genome (Perrin, Fedorova et al. 2007). If previously reported, the name of the product produced by the gene cluster is shown in parentheses. The first and last gene names of each cluster are given. For each cluster, we have indicated order of gene transcription, as well as color-coded up-regulated (in red), down-regulated (in blue) and non-differentially regulated (in black) genes.

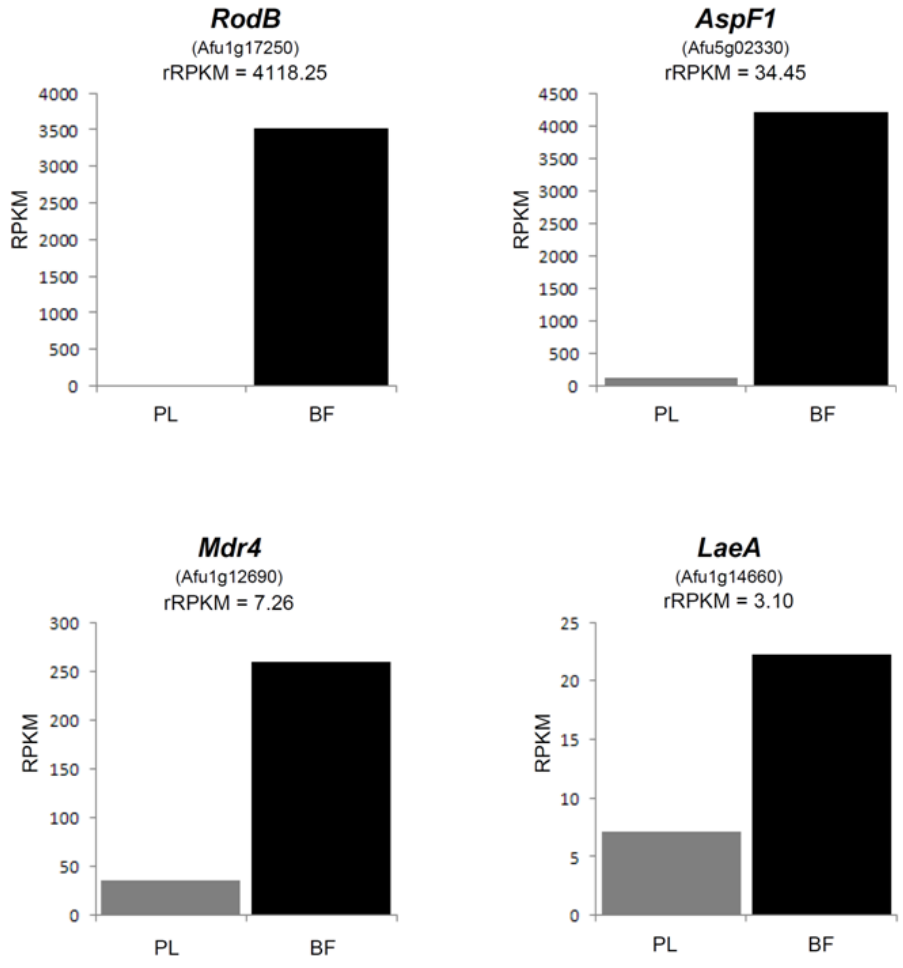


Up-regulation of cell wall genes

Because *A. fumigatus* cells are embedded in an ECM during BF formation (Beauvais, Schmidt et al. 2007; Loussert, Schmitt et al. 2010), we investigated the expression patterns of 409 genes coding for enzymes that degrade, modify, or create glycosidic bonds, including those involved in cell wall and ECM biosynthesis. We found that 169 of the 409 genes were up-regulated (146) or uniquely expressed (23) in BF, whereas 41 were down-regulated in BF (30) or uniquely expressed (11) in PL, suggesting a higher activity for carbohydrate-active enzymes in aerial conditions. However, the expression of the synthases responsible for the constitutive synthesis of α and β 1,3 glucans and chitin did not noticeably change. Among the glycosylhydrolases, the regulation of chitinases was not modified, whereas glucanases and glucosidases were highly expressed when the fungus was grown in an aerial colony; 15 glucanases and 14 glucosidases were up-regulated in BF, and only 2 glucanases and none of the glucosidases were down-regulated. The great majority of transglycosidases was also up-regulated in BF (35 transglycosidases were up-regulated and 2 were down-regulated). We also observed several significant differences between the two conditions in genes encoding for ECM proteins. For example, *RodA* (Afu5g09580), *RodB* (Afu1g17250) and *RodD* (Afu5g01490) were up-regulated in BF relative to PL. Interestingly, *RodB* was up-regulated by over 4,000-fold and was, overall, the second highest up-regulated gene (Figure 2.5).

Of the 81 allergens examined, 39 genes were up-regulated in BF relative to PL and another two were uniquely expressed in BF. In contrast, only 12 genes were down-regulated in BF relative to PL and only 1 gene was uniquely expressed in PL. Some allergens, such as the galactomanoprotein MP1 (Afu4g03240), the mannosidase MsdS (Afu1g14560), and the ribonuclease AspF1 (Afu5g02330) were expressed >30-fold in BF.

Figure 2.5. Examples of up-regulated genes during biofilm growth. RPKM values (Y axis) of the cell surface hydrophobin *RodB* (Afu1g17250), the allergen *AspF1* (Afu5g02330), the multidrug transporter *Mdr4* (Afu7g02690), and the global secondary metabolite regulator *LaeA* (Afu1g14660), are show during PL (gray) and BF (black) growth. For each gene, the rRPKM ($RPKM_{BF}$/ $RPKM_{PL}$) value is reported.



GPIlation and glycosylation of the proteins were the two major protein modifications that were up-regulated in aerially grown mycelia. The biosynthesis of the glycosyphosphatidyl inositol anchor was up-regulated, and so were all 14 genes controlling the steps occurring inside the ER from the palmitoylCoA-dependent inositol acyltransfer, phosphoethanolamine addition, mannosylation to transfer to proteins. In addition, 38 of the 81 genes coding for putative GPI-proteins were up-regulated in BF whereas the genes coding for only 4 of these GPI-proteins were down-regulated. The other protein post-translational modification that showed significant up-regulation in BF was glycosylation. This is true for both O and N glycosylation. The genes encoding for the four proteins known to be involved in O mannosylation – PMT1 (Afu1g07690), PMT2 (Afu3g06450), PMT4 (Afu8g04500), and the ortholog to the *Candida albicans* MNT1 (Afu5g10760) – were up-regulated. Finally, 21 of the 25 genes coding for mannosyltransferases located in the ER lumen and in the Golgi apparatus, which are responsible for N-glycan biosynthesis, were up-regulated in BF relative to PL.

Glycolysis is down-regulated in BF

The glycolysis pathway responsible for the anaerobic degradation of glucose to produce ethanol or lactic acid was down-regulated in BF. Specifically, the first step in the glycolysis pathway that results in the production of glyceraldehyde-3-phosphate, and the second step that leads to the production of the final product, which collectively account for 21 of the 35 pathway genes were down-regulated. Accordingly, lactate production was reduced in aerial conditions (the lactate dehydrogenase, encoded by Afu5g14800, was down-regulated) and so was ethanol production (the pyruvate decarboxylase leading

to the production of acetaldehyde was down-regulated) in BF. In contrast to glycolysis, respiration bridging the respiratory chain to oxidative phosphorylation was up-regulated, including the genes coding for protein complex II, complex III, and complex V controlling oxidative phosphorylation.

Physiological changes in BF growth can be responsible for drug resistance

Membranes are mainly composed of phospholipids and sterols. Although genes involved in glycerol phospholipid metabolism were not differentially regulated, those involved in the sterol metabolism were up-regulated in BF relative to PL. Specifically, 10 out of 20 genes leading to ergosterol from farnesyl di phosphate were up-regulated, including the *cyp51A* gene (Afu4g04820), which is targeted by the azole class of drugs and which was 2.3-fold up-regulated (p < 1e-300).

Reduced susceptibility to drugs in BF conditions (Seidler, Salvenmoser et al. 2008) could also result from an increased activity of efflux pumps and transporter proteins (Rajendran, Mowat et al. 2011). To test this hypothesis, we examined the expression patterns of the 274 genes encoding the major facilitator superfamily (MFS) proteins and the 45 genes encoding the ATP-Binding Cassette (ABC) proteins (Ren, Chen et al. 2007). We found that 146 of these 319 genes were up-regulated (140) or uniquely expressed in BF (6) and 16 were down-regulated in BF (11) or uniquely expressed in PL (5). Among the ABC transporters identified were all members of the MDR1 family (MDR1: Afu5g06070, Afu4g14130, Afu6g03470, MDR4: Afu7g02690, Afu6g03080, Afu3g02760, Afu3g03670, Afu1g12690, Afu3g03430, Afu7g00480, MDR2:

Afu4g10000) suspected to be associated to drug resistance in *A. fumigatus* (Tobin, Peery et al. 1997; Nascimento, Goldman et al. 2003), and the genes belonging to the *ScYOR1* family (Afu4g09150, Afu5g07970, Afu5g08150, Afu5g10510, Afu1g16880, Afu1g10390, Afu2g01500) associated to drug resistance in yeast (Decottignies, Grant et al. 1998).

Widespread differential regulation of transcription factors

The *A. fumigatus* genome contains 392 predicted or experimentally validated transcription factors (Beri, Whittington et al. 1987; Andrianopoulos and Hynes 1988; Lints, Davis et al. 1995; Vallim, Miller et al. 2000; Strittmatter, Irniger et al. 2001; Mulder, Saloheimo et al. 2004; Nierman, Pain et al. 2005; Vienken, Scherer et al. 2005; Grosse and Krappmann 2008; Soriani, Malavazi et al. 2008; Gravelat, Ejzykowicz et al. 2010). We found that nearly half of these transcription factors were differentially expressed (124 up-regulated (119) or uniquely expressed in BF (5) and 71 down-regulated in BF (69) or uniquely expressed in PL (2)), including several that function in asexual and sexual development.

Ribosome and translation

Translation (FunCat ID 12.04) and ribosome biogenesis (FunCat ID 12.01) were up-regulated in BF relative to PL. Given that much of a typical cell's energy is invested in producing ribosomes, it is not surprising that ribosome biogenesis and translation are up-regulated during aerial growth, which is the fastest of the two conditions tested (Beauvais, Schmidt et al. 2007; Cook and Tyers 2007).

DISCUSSION

A New Approach to Understand Essential Changes in *A. fumigatus* Lifestyle

In the lung parenchyma, like in the aerial colony on an agar plate (BF), the *A. fumigatus* mycelium is covered by an ECM (Beauvais, Schmidt et al. 2007; Mowat, Williams et al. 2009; Bruns, Seidler et al. 2010; Loussert, Schmitt et al. 2010; Muller, Seidler et al. 2011; Rajendran, Mowat et al. 2011). We reasoned that better understanding the genome-wide transcriptional configuration during BF would provide insight into the genes involved in this process and potentially a better understanding of the establishment of a colony during *in vivo* growth. Our comparative RNA-Seq analysis of growth in these two conditions (BF and PL) shows that this single change in non-nutritional environmental conditions led to the differential expression of thousands of genes (Figure 2.1). Considering that a recent comparison of BF and PL growth in *A. fumigatus* using microarray and 2-D gel electrophoresis technologies identified only ~700 genes and ~40 proteins that were differentially abundant during development (Bruns, Seidler et al. 2010), RNA-Seq appears to be the most powerful tool for genome-wide functional comparisons of fungal growth to date (Nagalakshmi, Wang et al. 2008; Wilhelm, Marguerat et al. 2008; Wang, Gerstein et al. 2009; Bruno, Wang et al. 2010).

Composition and Organization of the Extracellular Matrix

During biofilm growth, fungi produce an ECM that functions in the cohesive linkage of fungal cells with themselves or to the substratum (Beauvais, Schmidt et al. 2007; Mowat, Williams et al. 2009; Loussert, Schmitt et al. 2010). Although little is known about the chemical composition of fungal biofilms, it is well established that ECM composition

differs greatly between fungal species. In *C. albicans* the ECM is mainly composed of β 1,3 glucan (Nett, Sanchez et al. 2010), whereas the *Cryptococcus* ECM is based on glucuronoxylomannan, the major capsule component (Martinez and Casadevall 2007). In contrast, the *A. fumigatus* ECM is composed of polysaccharides, pigments, hexoses and proteins (Beauvais, Schmidt et al. 2007). One of the three polysaccharides that have been identified so far in the *A. fumigatus* ECM is α 1,3 glucan, a major component of the ECM in an *in vivo* aspergilloma model, where the hyphal network is tightly packed and has binding properties (Fontaine, Beauvais et al. 2010; Loussert, Schmitt et al. 2010). Accordingly, the observed up-regulation of the three α 1,3 glucan synthase genes in BF provides indirect support for their function in hyphal adhesion.

Two types of proteins can help organizing the structure of the colony, hydrophobins and adhesins. Hydrophobins are hydrophobic proteins that contain highly conserved cystein bridges, and which are most often responsible for conferring a hydrophobic character to fungal morphotypes (Sunde, Kwan et al. 2008). For example, the *A. fumigatus* RodAp hydrophobin is a beta-amyloid protein that self organizes into rodlets, confers a hydrophobic character to the conidia, and promotes the adhesion of cells to host proteins (Thau, Monod et al. 1994). Although the RodA and RodB proteins have been purified from conidia (Paris, Debeaupuis et al. 2003), our results show that both genes are transcribed prior to the formation of the conidiophore, as it was verified that no conidiophores were present when the RNA was isolated. Furthermore, *RodB* is the second most highly expressed gene in BF, and *RodA, RodB* and *RodD* are all up-regulated in BF cells, consistent with recent findings suggesting that several

hydrophobins (*RodB*, *RodD*, *RodE* and *RodF*) are expressed in the *A. fumigatus* ECM (Beauvais, Schmidt et al. 2007). The function of hydrophobins during hyphal growth is currently investigated by multiple sequential deletions of all hydrophobin genes.

The other protein family that participates in structuring the colony is the secreted glycosylated adhesins. Adhesins have been characterized in bacteria and yeasts and their role in cell-to-cell adhesion is well established (Soto and Hultgren 1999; Dunne 2002; Verstrepen and Klis 2006; de Groot, Kraneveld et al. 2008). One of the major changes in BF is an increase in the transcription of genes favoring glycosylation. Glycosylation controls many cell interactions in eukaryotes. For example, in *Candida* two families of glycosylated adhesins, the Als proteins from *C. albicans* and the Epa proteins from *C. glabrata,* have been shown to mediate adhesion to epithelia and to play a role in biofilm formation (Dranginis, Rauceo et al. 2007), whereas in humans glycan heterogeneity leads to different ligand-receptor complex formations (Dennis, Nabi et al. 2009).

The up-regulation of glycosylation in BF cells was also directly associated to the up regulation of GPI lation. Although this might have been an expected result, since the vast majority of fungal cell wall proteins is both GPI lated and O- and N glycosylated, this is the first time this correlation is observed. Adhesins have not been investigated biochemically in filamentous fungi, however, a number of putative adhesins have been bioinformatically predicted to exist in the *A. fumigatus* genome (Upadhyay, Mahajan et al. 2009). An up-regulation of 11 of the 25 genes with the highest adhesion probability score (Afu1g09510, Afu1g14430, Afu3g00420, Afu3g01150, Afu3g09690, Afu4g03240,

Afu6g13720, Afu7g00580, Afu7g02460, Afu7g05340 and Afu8g01970) was identified. Most of these proteins are also predicted to be GPI anchored. The expression of glycosylated proteins exposed on hyphal cell wall and present in the extracellular matrix likely results in strong mycelial adhesion between the hyphae of a colony. In addition, increased N glycosylation could protect the fungus against stress, in agreement with other data suggesting that the fungus is under higher stress in BF than in PL growth (Pattison and Amtmann 2009).

Fighting Antifungal Drugs

Whereas most of antifungal drugs are very active *in vitro* against PL cells, one of the major problems in treating aspergillosis is that the same drugs often have very poor *in vivo* impact (Denning and Hope 2010). This is consistent with several reports suggesting that BF cells exhibit higher resistance to drugs than PL cells (Mowat, Butcher et al. 2007; Mowat, Lang et al. 2008; Seidler, Salvenmoser et al. 2008; Rajendran, Mowat et al. 2011). The RNA-Seq data suggests at least three physiological events that could be responsible for the poor *in vivo* drug efficacy.

First, the presence of an ECM rich in sticky polysaccharides and proteins may reduce the penetration of the drugs. Although this has not been investigated in *A. fumigatus*, work in *Candida* has shown that the extracellular β 1,3 glucan matrix sequesters antifungals (Nett, Sanchez et al. 2010; Vediyappan, Rossignol et al. 2010). Second, sterol metabolism regulation in BF versus PL growth is in the direction expected to counteract the efficacy of azoles. Azoles inhibit the ergosterol biosynthesis pathway by targeting *cyp51A*

(Afu4g04820), a gene encoding for the 14-α-sterol demethylase enzyme (Mellado, Garcia-Effron et al. 2007). In *A. fumigatus*, it is currently thought that mutations within the *cyp51A* promoter and coding region alter the protein's structure conferring resistance due to target site alterations (Mellado, Garcia-Effron et al. 2007; Snelders, van der Lee et al. 2008). *cyp51A* is more than 2-fold up-regulated in *A. fumigatus* BF versus PL cells. This higher expression of *cyp51A*, as well as that of other genes responsible for increasing sterol concentration in the membranes of the aerial mycelium, which is expected to lead to an increase in the amount of target sterol, is likely to be associated to azole resistance, as has already been shown in *C. albicans* (Dunkel, Liu et al. 2008). Finally, azole resistance might be facilitated by multi-drug resistance membrane transporters, which play an active role in exporting antimicrobials out of the cytoplasm (Slaven, Anderson et al. 2002; Nascimento, Goldman et al. 2003; Meneau and Sanglard 2005). For example, a recent study showed that *Mdr4* is up-regulated in an *in vivo A. fumigatus* biofilm mouse model during voriconazole treatment (Rajendran, Mowat et al. 2011). Consistent with this hypothesis, we identified an overrepresentation of up-regulated genes coding for MFS and ABC transporters, including *Mdr1*, *Mdr2* and *Mdr4*, which are known to be involved in azole resistance.

Tolerance to Toxic and Aggressive Environments

The ecological niche for *A. fumigatus* is the soil, where it has to survive in a highly toxic environment and competitive microflora made of bacteria, fungi and protozoa (Tekaia and Latge 2005). The wide variety of potent mycotoxins produced by *A. fumigatus* help the fungus survive in this environment, whereas in mammalian hosts, these secondary

metabolites have broad virulent and pathogenic effects including cytotoxicity, mutagenicity, carcinogenicity, and immunosuppression (Keller, Turner et al. 2005). Our data show that many genes within secondary metabolite encoding clusters are up-regulated in BF (Figure 2.4). For example, similar to the findings of Bruns and co-workers (Bruns, Seidler et al. 2010), we observed substantial up-regulation of the fumitremorgen C supercluster (Figures 2.3, 2.4). This cluster encodes for an unknown compound, as well as the known toxins fumitremorgin (Maiya, Grundmann et al. 2006), a tremorgenic mycotoxin capable of inhibiting cell cycle in mammals, and pseurotin A, a neuritogenic compound and chitinase inhibitor (Maiya, Grundmann et al. 2007). Additionally, we observed an up-regulation of an entire gene cluster encoding for a trichothecene mycotoxin, a compound that can inhibit protein synthesis in eukaryotes (Bennett and Klich 2003). Lastly, we observed an up-regulation of *LaeA* (Figure 2.5), a known global regulator of secondary metabolism gene clusters in *A. fumigatus*, which may partially explain the observed widespread up-regulation of genes within secondary metabolite encoding clusters (Figure 2.4). Importantly, our comparison also identified several other up-regulated gene clusters that are candidates for secondary metabolism involvement, but whose function is currently unknown (Figure 2.3).

In addition to the protection against other soil microbes conferred by secondary metabolites, *A. fumigatus* is extremely rich in transporters involved in the efflux of toxins that are present in the soil (Iovdijova and Bencko 2010; Udeigwe, Eze et al. 2011). Our results show that the ion transporter genes involved in the efflux of arsenic and copper toxic ions (6 / 8 genes in the arsenic cluster and 14 / 15 genes in the copper MIP cluster),

two of the most prevalent toxic ions in the soil, are up-regulated in BF. This suggests that the fungus is well armed to be resistant to toxic ions in the soil. However, and in contrast to plants (Bienert and Jahn 2010), arsenite efflux is not relying on major intrinsic proteins or aquaporin that conduct water molecules and selected solutes in and out the cell, as two of the three annotated aquaglyceroporins (Afu4g03390, Afu4g00680 and Afu6g08480) are down-regulated in BF relative to PL.

MFS and ABC transporters are significantly up-regulated in BF relative to PL. Interestingly, saprophytic fungi, such as *Aspergillus*, typically harbor the highest numbers of transporters (Ren, Chen et al. 2007). For example, the TrasportDB database lists 934 transporters in *A. oryzae* (Ren, Chen et al. 2007, http://www.membranetransport.org/). In contrast, yeasts, which reside in relatively benign environments, have much smaller numbers of transporters; for example, *Schizosaccharomyces pombe* and *S. cerevisiae* contain only 185 and 316 genes, respectively (with 9 and 24 ABC transporters). A similar trend exists in bacteria where saprotrophs, like *Pseudomonas*, have at least 3x more transporters than pathogens. Thus, the presence of large numbers of transporters may be a marker for saprotrophism. Given that orthologs of the transcription factors controlling MDR proteins in yeasts (such as *YAP1, PDR1, PDR3,* and *PDR8*) have been identified by BLAST in *A. fumigatus*, one of the outstanding questions is which are the transcription factors controlling their expression.

Saprophytic microbes do only fight against each other but also compete for food. In contrast to MDR pumps, nutrition-associated transporters usually have specific

physiological substrates. We observed that ATP-dependent transporters of the P-type ATP (P-ATPase) superfamily are highly expressed in BF. Specifically, 14 genes, including the gene encoding for the major Pma1 protein ATPase, are up-regulated whereas only 5 are down-regulated in BF from a set of 23 genes. Similarly, pumps transporting putatively Zn, iron, potassium, sodium, and to a lesser extent, calcium and phosphate are up-regulated in BF and so are also amino acid transporters (poorly studied in *A. fumigatus*), peptide transporters, and gaba permeases (20.01.07). Among the nine up-regulated peptide transporters, four are members of the 8 oligopeptide transporter family (Hartmann 2010). Monosaccharide / hexose transporter genes are also up-regulated. This is in agreement with the release of glucose on the surface of the colony since transporters can also function as efflux pumps (Jansen, De Winde et al. 2002) and may be involved in the active secretion of glucose that was found in high amount during BF growth (Beauvais, Schmidt et al. 2007). Interestingly, the transporters for monocarboxylate (such as lactate and pyruvate – the end products of glycolysis) are highly up-regulated in BF. The active transport of nutrient is in agreement with a higher metabolic activity during BF growth.

ACKNOWLEDGMENTS AND CONTRIBUTIONS

We thank members of the Rokas lab for valuable comments on this work, Travis Clark and Chelsea Baker for help with Illumina sequencing and David McCauley for statistical advice. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. J.G.G. is funded by the Graduate Program in Biological Sciences at Vanderbilt University and the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH, NIAID: F31AI091343-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the NIH. Work in J.-P.L's Aspergillus lab is partly funded by the ESF grant Fuminomics and the ALLFUN FP7 project. Research in A.R.'s lab is supported by the Searle Scholars Program and the National Science Foundation (DEB-0844968).

J.G.G, A.R., A.B. and J.P.L. designed the study. R.B. and A.B. set up experiments and extracted RNA. J.G.G. analyzed the data. K.L.M. performed the sliding window FDR analysis.

REFERENCES

Andrianopoulos, A. and M. J. Hynes (1988). "Cloning and analysis of the positively acting regulatory gene amdR from *Aspergillus nidulans*." Mol Cell Biol **8**(8): 3532-41.

Baker, S. E. and J. W. Bennett (2008). An overview of the genus *Aspergillus*. The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods. G. H. Goldman and S. A. Osmani**:** 3-13.

Beauvais, A., C. Schmidt, et al. (2007). "An extracellular matrix glues together the aerial-grown hyphae of *Aspergillus fumigatus*." Cellular Microbiology **9**(6): 1588-600.

Bennett, J. W. and M. Klich (2003). "Mycotoxins." Clin Microbiol Rev **16**(3): 497-516.

Beri, R. K., H. Whittington, et al. (1987). "Isolation and characterization of the positively acting regulatory gene QUTA from *Aspergillus nidulans*." Nucleic Acids Research **15**(19): 7991-8001.

Bienert, G. P. and T. P. Jahn (2010). "Major intrinsic proteins and arsenic transport in plants: new players and their potential role." Advances in experimental medicine and biology **679**: 111-25.

Bruno, V. M., Z. Wang, et al. (2010). "Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq." Genome Res **20**(10): 1451-8.

Bruns, S., M. Seidler, et al. (2010). "Functional genomic profiling of *Aspergillus fumigatus* biofilm reveals enhanced production of the mycotoxin gliotoxin." Proteomics **10**(17): 3097-107.

Cook, M. and M. Tyers (2007). "Size control goes global." Current opinion in Biotechnology **18**(4): 341-50.

de Groot, P. W., E. A. Kraneveld, et al. (2008). "The cell wall of the human pathogen *Candida glabrata*: differential incorporation of novel adhesin-like wall proteins." Eukaryotic Cell **7**(11): 1951-64.

Decottignies, A., A. M. Grant, et al. (1998). "ATPase and multidrug transport activities of the overexpressed yeast ABC protein Yor1p." The Journal of biological chemistry **273**(20): 12612-22.

Denning, D. W. and W. W. Hope (2010). "Therapy for fungal diseases: opportunities and priorities." Trends in microbiology **18**(5): 195-204.

Dennis, J. W., I. R. Nabi, et al. (2009). "Metabolism, cell surface organization, and disease." Cell **139**(7): 1229-41.

Dranginis, A. M., J. M. Rauceo, et al. (2007). "A biochemical guide to yeast adhesins: glycoproteins for social and antisocial occasions." Microbiology and molecular biology reviews : MMBR **71**(2): 282-94.

Dunkel, N., T. T. Liu, et al. (2008). "A gain-of-function mutation in the transcription factor Upc2p causes upregulation of ergosterol biosynthesis genes and increased fluconazole resistance in a clinical *Candida albicans* isolate." Eukaryotic Cell **7**(7): 1180-90.

Dunne, W. M., Jr. (2002). "Bacterial adhesion: seen any good biofilms lately?" Clinical Microbiology Reviews **15**(2): 155-66.

Fedorova, N. D., N. Khaldi, et al. (2008). "Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*." PLoS Genetics **4**(4): e1000046.

Fontaine, T., A. Beauvais, et al. (2010). "Cell wall alpha1-3glucans induce the aggregation of germinating conidia of *Aspergillus fumigatus*." Fungal Genet Biol **47**(8): 707-12.

Geiser, D. M., M. A. Klich, et al. (2007). "The current status of species recognition and identification in *Aspergillus*." Studies in Mycology **59**: 1-10.

Gibbons, J. G., E. M. Janson, et al. (2009). "Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics." Mol Biol Evol **26**(12): 2731-44.

Gravelat, F. N., D. E. Ejzykowicz, et al. (2010). "*Aspergillus fumigatus* MedA governs adherence, host cell interactions and virulence." Cellular Microbiology **12**(4): 473-88.

Grosse, V. and S. Krappmann (2008). "The asexual pathogen *Aspergillus fumigatus* expresses functional determinants of *Aspergillus nidulans* sexual development." Eukaryotic Cell **7**(10): 1724-32.

Hartmann, T. (2010). Nitrogen metabolism in *Aspergillus fumigatus* with emphasis on the oligopeptide transporter (OPT) gene family, Julius - Maximilians - Universität Würzburg.

Hittinger, C. T., M. Johnston, et al. (2010). "Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life." Proc Natl Acad Sci U S A **107**(4): 1476-81.

Iovdijova, A. and V. Bencko (2010). "Potential risk of exposure to selected xenobiotic residues and their fate in the food chain--part I: classification of xenobiotics." Annals of agricultural and environmental medicine : AAEM **17**(2): 183-92.

Jansen, M. L., J. H. De Winde, et al. (2002). "Hxt-carrier-mediated glucose efflux upon exposure of Saccharomyces cerevisiae to excess maltose." Applied and Environmental Microbiology **68**(9): 4259-65.

Jiang, H. and W. H. Wong (2008). "SeqMap: mapping massive amount of oligonucleotides to the genome." Bioinformatics **24**(20): 2395-6.

Kanehisa, M. and S. Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Research **28**(1): 27-30.

Keller, N. P., G. Turner, et al. (2005). "Fungal secondary metabolism - from biochemistry to genomics." Nat Rev Microbiol **3**(12): 937-47.

Lamarre, C., S. Sokol, et al. (2008). "Transcriptomic analysis of the exit from dormancy of *Aspergillus fumigatus* conidia." BMC Genomics **9**: 417.

Latge, J. P. (1999). "*Aspergillus fumigatus* and aspergillosis." Clinical Microbiology Reviews **12**(2): 310-50.

Latge, J. P., I. Mouyna, et al. (2005). "Specific molecular features in the organization and biosynthesis of the cell wall of *Aspergillus fumigatus*." Med Mycol **43 Suppl 1**: S15-22.

Latge, J. P. and W. J. Steinbach, Eds. (2009). *Aspergillus fumigatus* and Aspergillosis. Washington, DC, ASM Press.

Lints, R., M. A. Davis, et al. (1995). "The positively acting amdA gene of *Aspergillus nidulans* encodes a protein with two C2H2 zinc-finger motifs." Mol Microbiol **15**(5): 965-75.

Loussert, C., C. Schmitt, et al. (2010). "*In vivo* biofilm composition of *Aspergillus fumigatus*." Cellular Microbiology **12**(3): 405-10.

Maiya, S., A. Grundmann, et al. (2006). "The fumitremorgin gene cluster of *Aspergillus fumigatus*: identification of a gene encoding brevianamide F synthetase." ChemBioChem **7**(7): 1062-9.
Maiya, S., A. Grundmann, et al. (2007). "Identification of a hybrid PKS/NRPS required for pseurotin A biosynthesis in the human pathogen *Aspergillus fumigatus*." ChemBioChem **8**(14): 1736-43.
Mari, A. and E. Scala (2006). "Allergome: a unifying platform." Arbeiten aus dem Paul-Ehrlich-Institut(95): 29-39; discussion 39-40.
Martinez, L. R. and A. Casadevall (2007). "*Cryptococcus neoformans* biofilm formation depends on surface support and carbon source and reduces fungal cell susceptibility to heat, cold, and UV light." Applied and Environmental Microbiology **73**(14): 4592-601.
McDonagh, A., N. D. Fedorova, et al. (2008). "Sub-telomere directed gene expression during initiation of invasive aspergillosis." PLoS Pathog **4**(9): e1000154.
Mellado, E., G. Garcia-Effron, et al. (2007). "A new *Aspergillus fumigatus* resistance mechanism conferring in vitro cross-resistance to azole antifungals involves a combination of cyp51A alterations." Antimicrobial Agents and Chemotherapy **51**(6): 1897-904.
Meneau, I. and D. Sanglard (2005). "Azole and fungicide resistance in clinical and environmental *Aspergillus fumigatus* isolates." Medical Mycology **43 Suppl 1**: S307-11.
Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-8.
Mowat, E., J. Butcher, et al. (2007). "*Aspergillus fumigatus* biofilms are refractory to antifungal challenge." International Journal of Antimicrobial Agents **29**: S147-S148.
Mowat, E., S. Lang, et al. (2008). "Phase-dependent antifungal activity against *Aspergillus fumigatus* developing multicellular filamentous biofilms." Journal of Antimicrobial Chemotherapy **62**(6): 1281-1284.
Mowat, E., C. Williams, et al. (2009). "The characteristics of *Aspergillus fumigatus* mycetoma development: is this a biofilm?" Medical Mycology **47**: S120-S126.
Mulder, H. J., M. Saloheimo, et al. (2004). "The transcription factor HACA mediates the unfolded protein response in *Aspergillus niger*, and up-regulates its own transcription." Mol Genet Genomics **271**(2): 130-40.
Muller, F. M. C., M. Seidler, et al. (2011). "*Aspergillus fumigatus* biofilms in the clinical setting." Medical Mycology **49**: S96-S100.
Nagalakshmi, U., Z. Wang, et al. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-9.
Nascimento, A. M., G. H. Goldman, et al. (2003). "Multiple resistance mechanisms among *Aspergillus fumigatus* mutants with high-level resistance to itraconazole." Antimicrobial Agents and Chemotherapy **47**(5): 1719-26.
Nett, J. E., H. Sanchez, et al. (2010). "Genetic basis of *Candida* biofilm resistance due to drug-sequestering matrix glucan." J Infect Dis **202**(1): 171-5.
Nierman, W. C., A. Pain, et al. (2005). "Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*." Nature **438**(7071): 1151-6.

Paris, S., J. P. Debeaupuis, et al. (2003). "Conidial hydrophobins of *Aspergillus fumigatus*." Applied and Environmental Microbiology **69**(3): 1581-8.
Pattison, R. J. and A. Amtmann (2009). "N-glycan production in the endoplasmic reticulum of plants." Trends in plant science **14**(2): 92-9.
Perrin, R. M., N. D. Fedorova, et al. (2007). "Transcriptional regulation of chemical diversity in *Aspergillus fumigatus* by LaeA." PLoS Pathog **3**(4): e50.
Pitts, R. J., D. C. Rinker, et al. (2011). "Transcriptome profiling of chemosensory appendages in the malaria vector *Anopheles gambiae* reveals tissue- and sex-specific signatures of odor coding." BMC Genomics **12**(1): 271.
Rajendran, R., E. Mowat, et al. (2011). "Azole resistance of *Aspergillus fumigatus* biofilms is partly associated with efflux pump activity." Antimicrobial Agents and Chemotherapy.
Ren, Q., K. Chen, et al. (2007). "TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels." Nucleic Acids Research **35**(Database issue): D274-9.
Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Research **32**(18): 5539-5545.
Seidler, M. J., S. Salvenmoser, et al. (2008). "*Aspergillus fumigatus* forms biofilms with reduced antifungal drug susceptibility on bronchial epithelial cells." Antimicrobial Agents and Chemotherapy **52**(11): 4130-4136.
Slaven, J. W., M. J. Anderson, et al. (2002). "Increased expression of a novel *Aspergillus fumigatus* ABC transporter gene, atrF, in the presence of itraconazole in an itraconazole resistant clinical isolate." Fungal Genet Biol **36**(3): 199-206.
Snelders, E., H. A. van der Lee, et al. (2008). "Emergence of azole resistance in *Aspergillus fumigatus* and spread of a single resistance mechanism." PLoS Medicine **5**(11): e219.
Sokal, R. R. and F. J. Rohlf (1995). Biometry : the principles and practice of statistics in biological research. New York, Freeman.
Soriani, F. M., I. Malavazi, et al. (2008). "Functional characterization of the *Aspergillus fumigatus* CRZ1 homologue, CrzA." Mol Microbiol **67**(6): 1274-91.
Soto, G. E. and S. J. Hultgren (1999). "Bacterial adhesins: common themes and variations in architecture and assembly." Journal of bacteriology **181**(4): 1059-71.
Strittmatter, A. W., S. Irniger, et al. (2001). "Induction of jlbA mRNA synthesis for a putative bZIP protein of *Aspergillus nidulans* by amino acid starvation." Current genetics **39**(5-6): 327-34.
Sunde, M., A. H. Kwan, et al. (2008). "Structural analysis of hydrophobins." Micron **39**(7): 773-84.
Tekaia, F. and J. P. Latge (2005). "*Aspergillus fumigatus*: saprophyte or pathogen?" Current Opinion in Microbiology **8**(4): 385-92.
Thau, N., M. Monod, et al. (1994). "rodletless mutants of *Aspergillus fumigatus*." Infection and Immunity **62**(10): 4380-8.
Tobin, M. B., R. B. Peery, et al. (1997). "Genes encoding multiple drug resistance-like proteins in *Aspergillus fumigatus* and *Aspergillus flavus*." Gene **200**(1-2): 11-23.

Udeigwe, T. K., P. N. Eze, et al. (2011). "Application, chemistry, and environmental implications of contaminant-immobilization amendments on agricultural soil and water quality." Environment international **37**(1): 258-67.

Upadhyay, S. K., L. Mahajan, et al. (2009). "Identification and characterization of a laminin-binding protein of *Aspergillus fumigatus*: extracellular thaumatin domain protein (AfCalAp)." Journal of Medical Microbiology **58**(Pt 6): 714-22.

Vallim, M. A., K. Y. Miller, et al. (2000). "*Aspergillus* SteA (sterile12-like) is a homeodomain-C2/H2-Zn+2 finger transcription factor required for sexual reproduction." Mol Microbiol **36**(2): 290-301.

Vediyappan, G., T. Rossignol, et al. (2010). "Interaction of *Candida albicans* biofilms with antifungals: transcriptional response and binding of antifungals to beta-glucans." Antimicrobial Agents and Chemotherapy **54**(5): 2096-111.

Verstrepen, K. J. and F. M. Klis (2006). "Flocculation, adhesion and biofilm formation in yeasts." Molecular microbiology **60**(1): 5-15.

Vienken, K., M. Scherer, et al. (2005). "The Zn(II)2Cys6 putative *Aspergillus nidulans* transcription factor repressor of sexual development inhibits sexual development under low-carbon conditions and in submersed culture." Genetics **169**(2): 619-30.

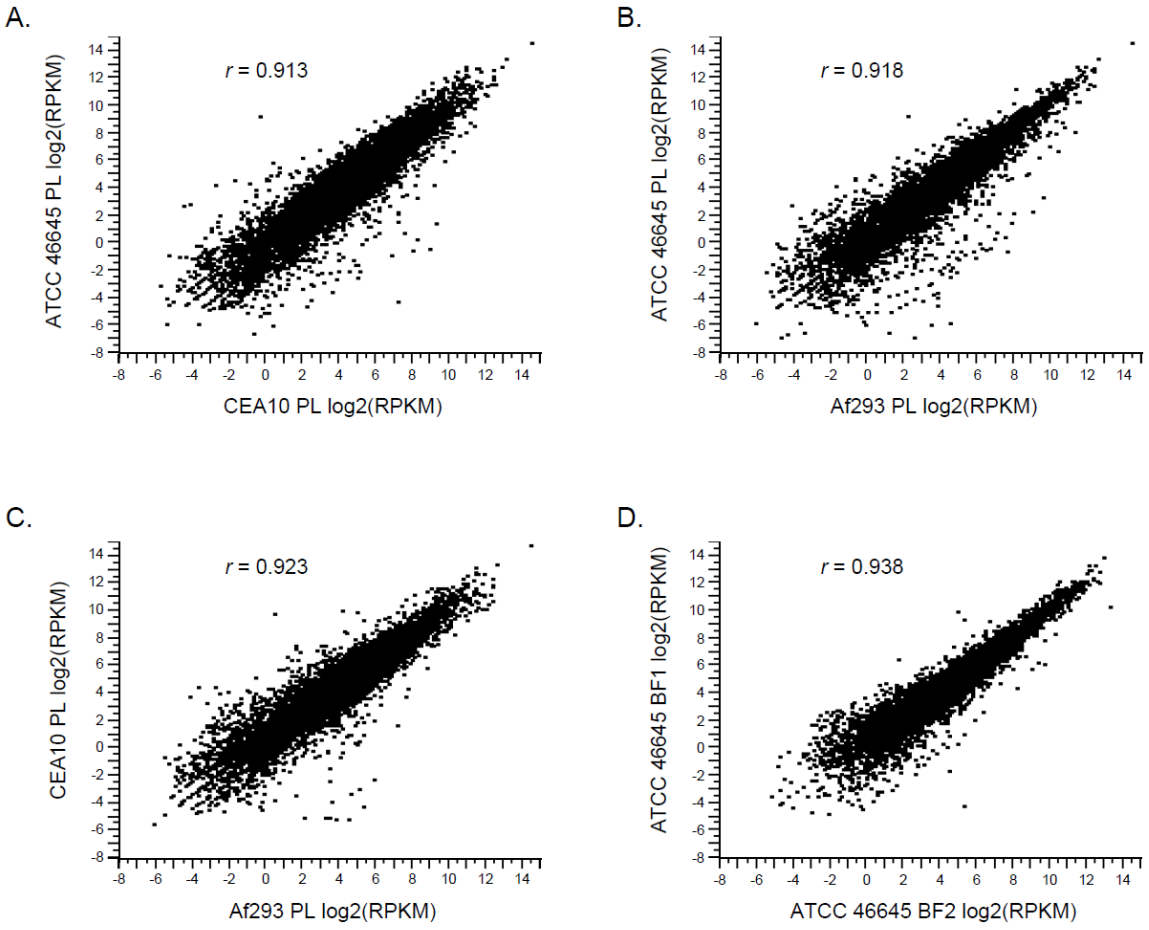
Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Ward, T. J., J. P. Bielawski, et al. (2002). "Ancestral polymorphism and adaptive evolution in the trichothecene mycotoxin gene cluster of phytopathogenic *Fusarium*." Proceedings of the National Academy of Sciences of the United States of America **99**(14): 9278-83.

Wilhelm, B. T., S. Marguerat, et al. (2008). "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution." Nature **453**(7199): 1239-43.

SUPPLEMENTARY FIGURES

Figure S2.1. Correlation analysis between three biological replicates on PL growth performed on three different *A. fumigatus* strains (panels A–C), and one technical replicate performed on BF growth (panel D). Note that Pearson's *r* is > 0.91 in all panels, indicating that there is very little technical or biological variation across the samples tested.



CHAPTER III

EVIDENCE FOR GENETIC DIFFERENTIATION AND VARIABLE RECOMBINATION RATES AMONG DUTCH POPULATIONS OF THE OPPORTUNISTIC HUMAN PATHOGEN *ASPERGILLUS FUMIGATUS*

Corné H.W. Klaassen,[1*] John G. Gibbons,[2*] Natalie D. Fedorova,[3] Jacques F. Meis[1] and Antonis Rokas[2]

[1]*Department of Medical Microbiology and Infectious Diseases, Canisius Wilhelmina Hospital, Nijmegen, the Netherlands*

[2]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA*

[3]*J Craig Venter Institute, Rockville, MD, USA*

[*]These authors contributed equally to this work.

This chapter is published in *Molecular Ecology*, 2012, 21: 57-70.

ABSTRACT

As the frequency of antifungal drug resistance continues to increase, understanding the genetic structure of fungal populations, where resistant isolates have emerged and spread, is of major importance. *Aspergillus fumigatus* is a ubiquitously distributed fungus and the primary causative agent of invasive aspergillosis (IA), a potentially lethal infection in immunocompromised individuals. In the last few years, an increasing number of *A. fumigatus* isolates has evolved resistance to triazoles, the primary drugs for treating IA infections. In most isolates, this multiple-triazole-resistance (MTR) phenotype is caused by mutations in the *cyp51A* gene, which encodes the protein targeted by the triazoles. We investigated the genetic differentiation and reproductive mode of *A. fumigatus* in the Netherlands, the country where the MTR phenotype likely originated, to determine their role in facilitating the emergence and distribution of resistance genotypes. Using 20 genome-wide neutral markers, we genotyped 255 Dutch isolates including 25 isolates with the MTR phenotype. In contrast to previous reports, our results show that Dutch *A. fumigatus* genotypes are genetically differentiated into five distinct populations. Four of the five populations show significant linkage disequilibrium, indicative of an asexual reproductive mode, whereas the fifth population is in linkage equilibrium, indicative of a sexual reproductive mode. Notably, the observed genetic differentiation among Dutch isolates does not correlate with geography, although all isolates with the MTR phenotype nest within a single, predominantly asexual, population. These results suggest that both reproductive mode and genetic differentiation contribute to the structure of Dutch *A. fumigatus* populations, and are likely shaping the evolutionary dynamics of drug resistance in this potentially deadly pathogen.

INTRODUCTION

Invasive fungal infections, which have become a major health issue over the past three decades, are difficult to treat and diagnose. Identifying the genetic structure (i.e., the genetic differentiation and mode of reproduction) of fungal pathogens is essential for the development of better therapeutics against fungal infections. The majority of medically important fungi are haploid organisms that reproduce mostly asexually while occasionally engaging in sexual or parasexual reproduction (Taylor, Geiser et al. 1999; Heitman 2006; Sun and Heitman 2011), resulting in populations that are predominantly non-recombining. This prevalence of "asexual" reproduction among human pathogenic fungi is surprising, not only because recombination can be very advantageous in stressful environments (Goddard, Godfray et al. 2005; Zeyl, Curtin et al. 2005), such as inside a human host, but also because several of these pathogens have closely related non-pathogenic sexual relatives (Nielsen and Heitman 2007; Butler 2010).

One of the most important fungal pathogens of humans is *Aspergillus fumigatus*, the leading cause of a rapidly progressing, and frequently deadly, systemic infection called invasive aspergillosis (IA) (Denning 1998). Although many fungal species, including several human pathogens, show genetic differentiation (Milgroom 1996; Taylor, Turner et al. 2006; Hittinger, Goncalves et al. 2010), several studies have reported that *A. fumigatus* lacks genetic differentiation (Debeaupuis, Sarfati et al. 1997; Pringle, Baker et al. 2005; Rydholm, Szakacs et al. 2006). For example, an early restriction fragment length polymorphism analysis of hundreds of clinical and environmental isolates (Debeaupuis, Sarfati et al. 1997) and a multilocus sequence-based phylogenetic analysis

of clinical and environmental isolates from five continents (Rydholm, Szakacs et al. 2006) both found no evidence of genetic differentiation. This absence of genetic differentiation in *A. fumigatus* is consistent with the species' cosmopolitan distribution, its high abundance, and ease of aerial dispersal. Alternatively, the observed absence of differentiation could be due to the use of markers that are less informative or to sampling design. For example, a different multilocus sequence-based phylogenetic analysis of isolates from around the world identified two globally distributed and genetically differentiated lineages within *A. fumigatus* (Pringle, Baker et al. 2005), arguing that additional studies are required prior to concluding that *A. fumigatus* is not genetically differentiated.

Early evolutionary analyses also suggested that *A. fumigatus* had lost the ability to reproduce sexually (Geiser, Timberlake et al. 1996). However, the presence of intact meiosis-related genes in the *A. fumigatus* genome (Galagan, Calvo et al. 2005; Nierman, Pain et al. 2005; Rokas and Galagan 2008), the presence of mating type loci at near-equal frequencies (Paoletti, Rydholm et al. 2005), and the demonstration that certain isolates can undergo sexual reproduction in the laboratory (O'Gorman, Fuller et al. 2009), suggest that natural *A. fumigatus* populations are likely to reproduce both asexually and sexually. This inference is consistent with the detection of historical recombination in several population genetic studies of *A. fumigatus* populations (Varga and Toth 2003; Paoletti, Rydholm et al. 2005; Pringle, Baker et al. 2005).

Understanding the genetic structure of *A. fumigatus* is important for human affairs because recent reports indicate an increase in the frequency of multi-triazole-resistant (MTR) *A. fumigatus* isolates worldwide. Triazole drugs are the primary and most effective therapy against IA infections (Meis and Verweij 2001; Herbrecht, Denning et al. 2002). In the majority of isolates, MTR resistance is due to two mutations in the *cyp51A* gene that encodes 14α-sterol demethylase, the triazole target (Mellado, Garcia-Effron et al. 2007; Verweij, Mellado et al. 2007; Snelders, van der Lee et al. 2008). The two mutations, known as the TR/L98H allele, involve (i) a duplication in the *cyp51A* promoter, and (ii) a non-synonymous point mutation in the *cyp51A* coding region that results in a L98H amino acid change in the protein product. *A. fumigatus* is not transmitted host-to-host, so the rapid spread of the TR/L98H allele raises the possibility that it might have originated outside the clinical environment (Snelders, van der Lee et al. 2008; Verweij, Snelders et al. 2009). Notably, genetic analysis of a collection of MTR isolates shows that all isolates with the TR/L98H allele are confined within a single clade and are less variable than non-resistant isolates (Snelders, van der Lee et al. 2008), consistent with a single and recent origin.

The recent spread of the TR/L98H allele in *A. fumigatus* presents a unique opportunity to investigate the role of genetic differentiation and reproductive mode in shaping the evolution of drug resistance in this potentially deadly pathogen. Our hypothesis was that sexual reproduction is facilitating the emergence and spread of the TR/L98H allele across the entire *A. fumigatus* population. To test this hypothesis, we studied the genetic structure of *A. fumigatus* in the Netherlands, the country where MTR first emerged.

Using 20 neutral markers dispersed across the *A. fumigatus* genome, we genotyped and analyzed 255 clinical and environmental Dutch isolates, including 25 MTR isolates with the TR/L98H allele. Contrary to our expectation, we found that Dutch *A. fumigatus* genotypes group into five distinct populations that differed in levels of genetic diversity and recombination patterns, which allowed us to infer that four of the five populations were predominantly reproducing asexually. Importantly, the analysis showed that all isolates with the TR/L98H allele nested within one of the four asexually reproducing populations. These results suggest that genetic differentiation and reproductive mode are influencing the dynamics of drug resistance patterns in natural *A. fumigatus* populations.

MATERIALS AND METHODS

Isolate Collection

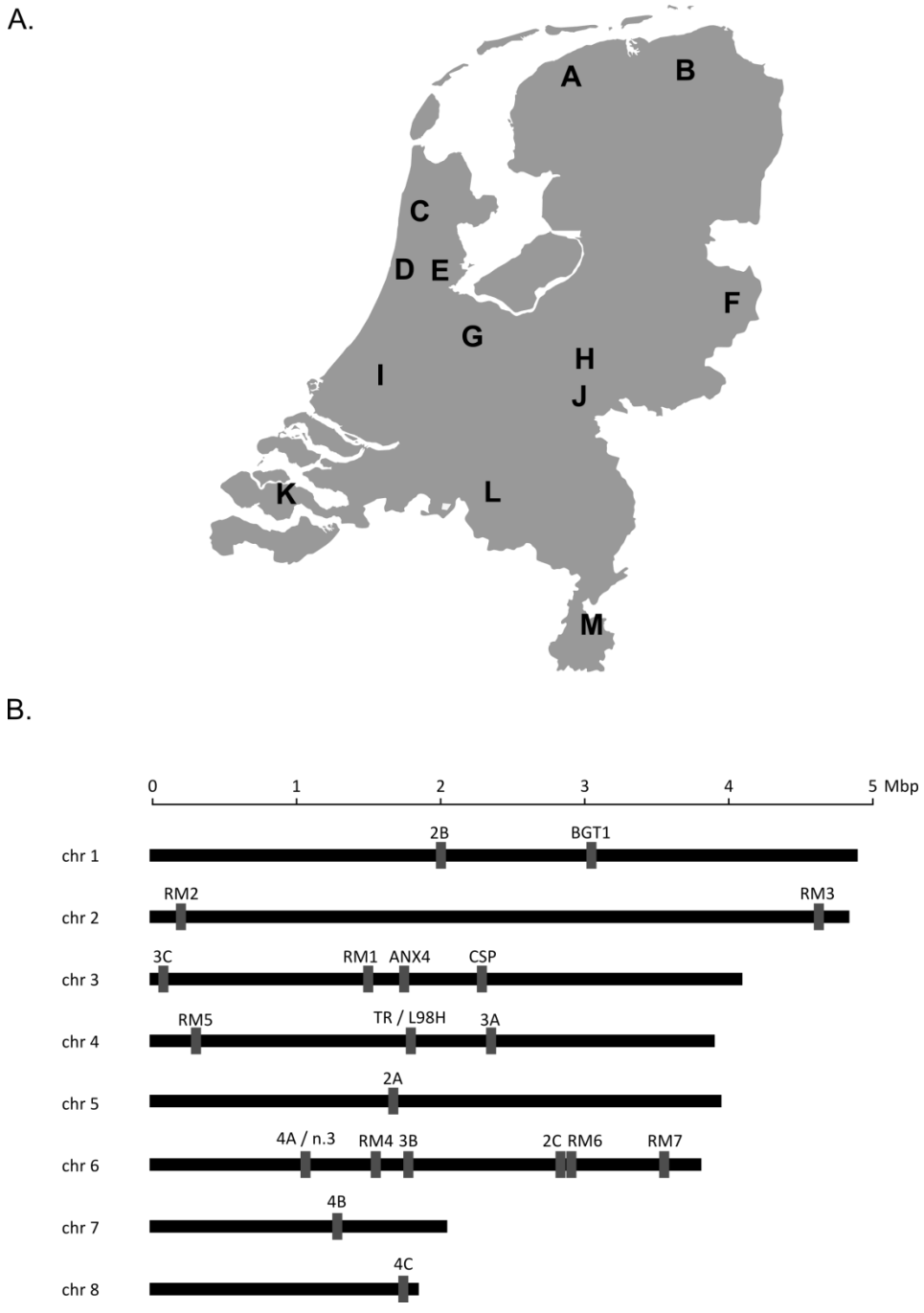
We analysed 255 clinical and environmental isolates from the Netherlands (Figure 3.1A). We included 201 isolates from a nation-wide survey conducted in 2005 (Klaassen, de Valk et al. 2009; Klaassen, de Valk et al. 2010), 12 isolates from the clinical samples collection of the Canisius Wilhelmina Hospital in the city of Nijmegen, and 42 isolates from samples of cultivated garden soils from the greater Nijmegen area from a survey conducted in 2008. In all isolates, we determined the presence / absence of the TR/L98H allele using a recently developed real-time PCR screening method (Klaassen, de Valk et al. 2010). Sixteen of the 213 clinical isolates and nine of the 42 environmental isolates contained the TR/L98H allele.

Molecular Typing

We genotyped all 255 isolates using 20 markers (9 microsatellite, 1 indel, and 10 sequence / PCR-typing markers), distributed across the eight *A. fumigatus* chromosomes (Figure 3.1B, Table S3.1). Typing of the nine microsatellite markers, which are 2A, 2B, 2C, 3A, 3B, 3C, 4A, 4B and 4C, was performed as described previously (de Valk, Meis et al. 2005). We also scored the presence / absence of a one base-pair deletion in the flanking region of microsatellite marker 4A as a separate indel marker; this is the n.3 marker. We sequenced the BGT1 and ANXC4 loci and identified marker alleles using the multilocus sequence-typing scheme described by Bain and co-workers (Bain, Tavanti et al. 2007). We sequence-typed the CSP (cell surface protein) locus as previously described (Balajee, Tay et al. 2007; Klaassen, de Valk et al. 2009). We PCR-typed the MAT

(mating type) locus using the "one common and one specific primer" strategy, which yields amplicons of different length for each of the two MAT alleles (Paoletti, Rydholm et al. 2005). Using a similar approach, we PCR-typed the alleles of six putative HET (for *het*erokaryon or vegetative incompatibility) loci, which are thought to regulate self / non-self-recognition during filamentous growth (Fedorova, Khaldi et al. 2008; Fedorova, Harris et al. 2009). For the MAT and HET loci, which together will be further referred to as recombination markers RM1 – RM7 (Figure 3.1B, Table S3.1), the A-allele corresponds to the allele present in strain Af293 (Nierman, Pain et al. 2005), the B-allele corresponds to the allele present in strain A1163 (Fedorova, Khaldi et al. 2008), whereas the C-allele corresponds to a recently identified third allelic variant (Klaassen, submitted). We scored alleles that failed to yield a PCR product as nulls and confirmed negative results using a different set of PCR amplification primers.

Figure 3.1. (A) Sampling location of the 156 genotypes from the Netherlands and (B) the chromosomal location of the TR/L98H (MTR) locus and the 20 markers used in the present study. In panel A, city and number of isolates and genotypes collected per city, respectively (in parentheses), are as follows: A: Leeuwarden (8, 3), B: Groningen (1, 1), C: Alkmaar (10, 9), D: Haarlem (7, 5), E: Amsterdam (19, 10), F: Enschede (15, 9), G: Utrecht (21, 7), H: Arnhem (6, 3), I: Rotterdam (45, 29), J: Nijmegen (102, 68), K: Goes (7, 5), L: Veldhoven (9, 4), and M: Heerlen (5, 3). The chromosome length scale (in million base pairs or Mbp) is shown on top of panel B.



Neutral Evolution Analysis

Markers undergoing positive selection are poorly suited for the study of population structure (Avise 2000). Therefore, we tested whether the coding sequences containing our markers are likely undergoing positive selection by estimating the ω ratio of the non-synonymous substitution rate (d_N) to the synonymous substitution rate (d_S) for each gene using the CODEML module from the PAML software, version 4.4 (Yang 2007). We first identified and aligned orthologs between the transcriptomes of *A. fumigatus*, *Neosartorya fischeri*, and *A. clavatus* as described previously (Rokas, Payne et al. 2007; Fedorova, Khaldi et al. 2008; Rokas 2009). To test for positive selection in each coding gene in our marker set, we first evaluated the log likelihood of the null M7 model. Under M7, ω values at different codon positions in a gene follow a beta distribution, where ω is constrained to fall between zero and one. We then measured the difference (ΔL) between the log likelihood of the M7 model and that of the alternative M8 model, which, in addition to the zero to one beta distribution for ω values, also allows for a subset of codon sites to have ω values above one (Yang 2006; Scannell, Zill et al. 2011). We excluded all genes showing rates of synonymous substitutions larger than 2, because in these cases substitution saturation is likely to reduce the power and reliability of the performed comparisons. Finally, for the BGT1 and ANXC4 markers, for which typing was done by sequencing, we also evaluated whether the *A. fumigatus* population departed from neutrality by calculating Tajima's D (Tajima 1989), as implemented in the DNASP software, version 5.10.01 (Librado and Rozas 2009). All tests were performed at $p = 0.01$ significance.

Clonal Correction

Inclusion of clonally related genotypes can blur analyses of genetic differentiation, haploid diversity and linkage disequilibrium, because it violates the assumptions of the evolutionary models used in these analyses (Pritchard, Stephens et al. 2000; Jombart, Devillard et al. 2010). Because microsatellite markers have very high mutation rates (Lynch, Sung et al. 2008), when a large number of microsatellite markers is used, clonally related genotypes might not have identical alleles at all microsatellite loci. Furthermore, for organisms, such as *A. fumigatus*, that are capable of reproducing both sexually and asexually, it is difficult to determine *a priori* what the best clonal correction threshold might be. To avoid these problems, we generated a series of clonally corrected data sets by eliminating the genotypes of all but one (randomly chosen) isolates with identical alleles for the n.3, ANXC4, BGT1, RM1-7, and CSP markers, and with 0 – 9 identical microsatellite markers.

Genetic Differentiation Analysis

We examined the genetic differentiation of Dutch *A. fumigatus* isolates using both model-based and non-model based approaches for the microsatellite (9 markers), non-microsatellite (11 markers) and full (20 markers) data sets, as well as for the series of clonally corrected full (20 marker) data sets.

We examined genetic differentiation using the Bayesian model-based approach implemented in the software STRUCTURE, version 2.3.3 (Pritchard, Stephens et al. 2000).

We used “admixture” and “allele frequencies are correlated among populations” as our ancestry and frequency models, respectively. We ran 100 replicates of 200,000 Markov Chain Monte Carlo (MCMC) generations for K = 1–10, where K = number of populations. In each run, we discarded the first 100,000 generations as burn-in. To identify the optimal K value, we used two different approaches. The first approach makes use of calculating the average log probability (LnP(D)) of each K value (Pritchard, Stephens et al. 2000). Under this approach, the optimal K value is the one showing the highest LnP(D) score. The second approach is based on the *ad hoc* statistic ΔK, which calculates the rate of change in the log probability of data between successive runs with different K values (Evanno, Regnaut et al. 2005). Under this approach, the optimal K value is the one that maximizes ΔK. To evaluate the robustness of population assignment across the 100 STRUCTURE replicates we compared population assignments for each replicate to the replicate used in this study and calculated average individual membership coefficients using the CLUMPP software, version 1.1.2 (Jakobsson and Rosenberg 2007).

Because natural populations often violate Hardy-Weinberg equilibrium and linkage equilibrium assumptions, inferences drawn solely from model-based methods can be problematic. Therefore, we also analyzed our data set using the non-model-based multivariate approach DAPC, as implemented in the ADEGENET software, version 1.3-0 (Jombart 2008; Jombart, Devillard et al. 2010). We predicted the optimal number of clusters (populations) using the *k*-means clustering algorithm, “find.clusters”, retaining all principal components. We calculated the Bayesian Information Criterion (BIC) for K = 1–10, where K = number of populations. The optimal number of populations was

identified as the one for which BIC showed the lowest value and after which BIC increased or decreased by the least amount. We then used DAPC to assign individuals into populations, retaining the number of principal components encompassing 80% of the cumulative variance.

Genetic Differentiation by Geography Analysis

We performed two analyses to test the hypothesis that genotype geography was associated with genetic differentiation. In the first analysis, we grouped genotypes into populations based on their city of origin and based on their STRUCTURE population assignment. We then estimated global and pairwise population differentiation (φ_{PT}) values, a suitable measure of population differentiation analogous to F_{ST} for haploids, and performed AMOVA analysis in each data set using the GENALEX software, version 6.41 (Peakall and Smouse 2006). If genotype geography is significantly associated with genetic differentiation, we expect to see greater among–population variation, smaller within–population variation, and significant population differentiation when genotypes are grouped by city of origin compared to when genotypes are grouped to the populations they are assigned to by the STRUCTURE software.

In the second analysis, we performed a χ^2goodness-of-fit test between the *observed* number of genotypes from each location and the *expected* number of genotypes from each location due to chance. For each population, we calculated the expected number of genotypes from each location using the equation $N_{LOC} / (N_{TOT} \otimes N_{POP})$, where N_{LOC} is the

number of genotypes from location X, N_{POP} is the number of genotypes from population X, and N_{TOT} is the total number of genotypes.

Haploid Diversity Analysis

We calculated Nei's unbiased haploid diversity (*uh*), a measure that calculates haploid diversity corrected for sample size, independently for the total marker set as well as the microsatellite marker set, for each one of the populations delineated using the STRUCTURE software, using the GENALEX software, version 6.4.1 (Peakall and Smouse 2006). For comparison, we also calculated haploid diversity from two populations of *A. nidulans* using a set of seven microsatellite markers (Hosid, Grishkan et al. 2008), as well as from a population of *A. flavus* and a population of *A. parasiticus* using a different set of seven microsatellite markers (Tran-Dinh and Carter 2000). Note that the microsatellite markers used in these studies are different from the markers used in this study.

Linkage Disequilibrium Analysis

To calculate the degree of association between alleles in our set of 20 markers and to examine whether patterns of recombination are similar across the five populations delineated using the STRUCTURE software, we calculated linkage disequilibrium (LD) using the MULTILOCUS software, version 1.3b (Agapow and Burt 2001). Specifically, we evaluated the index of association (*Ia*) (Maynard Smith, Smith et al. 1993), which calculates the distance between all possible locus pairs and compares the variance of distances against results expected if there is no association between loci. *Ia* was calculated both globally for each population as well as between all locus pairs within

each population. We assessed statistical significance by comparing the observed *Ia* value against the *Ia* values obtained from 1,000 randomized multilocus genotype data matrices, using $p = 0.05$ as the significance value cutoff. The randomization step shuffles the alleles among isolates independently for each locus and assumes that the population is in linkage equilibrium. Because non-recombining populations are expected to contain higher numbers of clonally related genotypes, we also calculated the number of identical genotypes in a population, or clonal richness, for each population as another indicator of reproductive mode.

Estimation of Divergence Times of *A. fumigatus* Populations

We estimated pairwise divergence times between *A. fumigatus* populations by adapting the methods developed by Zhivotovsky (2001), which have been previously applied in *A. flavus* (Grubisha and Cotty 2010), for our set of microsatellite markers. We excluded markers 3A and 3C from this analysis because they exhibited unusually high levels of variation that may deviate from the generalized stepwise mutation model. We calculated divergence time, in generation units, using the equation $T_D = (D_1 / 2w) – (V_0 / w)$. D_1 is the average over loci of the average squared difference of repeat unit copy number between pairs of alleles sampled (one each from the populations) over 5 replicates (Goldstein, Linares et al. 1995), w is the effective mutation rate (Thuillet, Bataillon et al. 2005), and V_0 is the average over all loci of the within-population variance in the repeat unit in the ancestral population. We estimated early and late boundaries of divergence by setting $V_0 = 0$ and V_0 = variance in the extant populations, respectively (Zhivotovsky 2001; Munkacsi, Stoxen et al. 2008). w was estimated by averaging the average mutation

rate for each loci where mutation rate (μ) = 0.00003R – 0.0001, where R is the repeat unit copy number (Thuillet, Bataillon et al. 2005).

RESULTS

A. fumigatus Markers are Evolving Neutrally

Six of our 20 markers reside in non-coding regions and are unlikely to be under selective pressure (Table S1). For the remaining 14 markers that reside within coding genes, we were able to reliably identify orthologs for 10 genes and reliably estimate ω values for 7 of them. The remaining three genes had d_S values larger than 2, making neutrality testing unreliable. The genes evaluated were associated with markers 2A, 2B, 3A, ANXC4, BGT1, CSP, and RM7. None of the 7 genes could reject the null M7 model in favour of the M8 model (2A: $\Delta L = 0$; 2B: $\Delta L = 0.098$; 3A: $\Delta L = 0$; ANXC4: $\Delta L = 0$; BGT1: $\Delta L = 0$; CSP: $\Delta L = 3.512$; and RM7: $\Delta L = 0$; *p* values for all tests are > 0.01), suggesting that positive selection is unlikely to be acting on them. Consistent with these results, examination of Tajima's D statistic for the ANXC4 and BGT1 loci within *A. fumigatus* showed no evidence of positive or balancing selection (ANXC4: D = -0.194; and BGT1: D = -0.522; *p* values for all tests are > 0.01).

A. fumigatus Genotypes and Clonal Correction

Prior to any clonal correction, the entire collection of 255 isolates yielded 225 different genotypes. Microsatellite markers made the biggest contribution to the observed genotypic diversity. Specifically, 224/225 genotypes were recognized by the microsatellite markers alone, 106/225 by the RM markers, 20/225 by the CSP marker, and 10/225 by the ANXC4 and BGT markers combined. In both ANXC4 and BGT1, we identified two new alleles. Since up to now in both genes only four different alleles were

recognized, we have provisionally numbered the new alleles as the fifth and sixth allele of each marker.

To remove any additional clonally related genotypes, we generated series of clonally corrected data sets by eliminating the genotypes of all but one (randomly chosen) isolates with identical alleles for the n.3, ANXC4, BGT1, RM1-7, and CSP markers, as well as for 9 – 0 microsatellite markers. This filter resulted in the elimination of 29 – 106 genotypes (Figure S3.1), with the remaining number of non-clonal genotypes plateauing to ~150 genotypes when clonal correction of 5 – 0 identical (or 4 – 9 different) microsatellites was applied. Given these results, we decided to use the "5 identical microsatellites" clonal correction threshold, because it coincides with the reaching of the clonal correction plateau. Using this threshold, we removed 99 genotypes (94 with triazole susceptible alleles and 5 with the MTR allele), resulting in a final data set of 156 non-clonally related genotypes, including 20 with the MTR allele. Unless otherwise indicated, all subsequent analyses were performed on this data set.

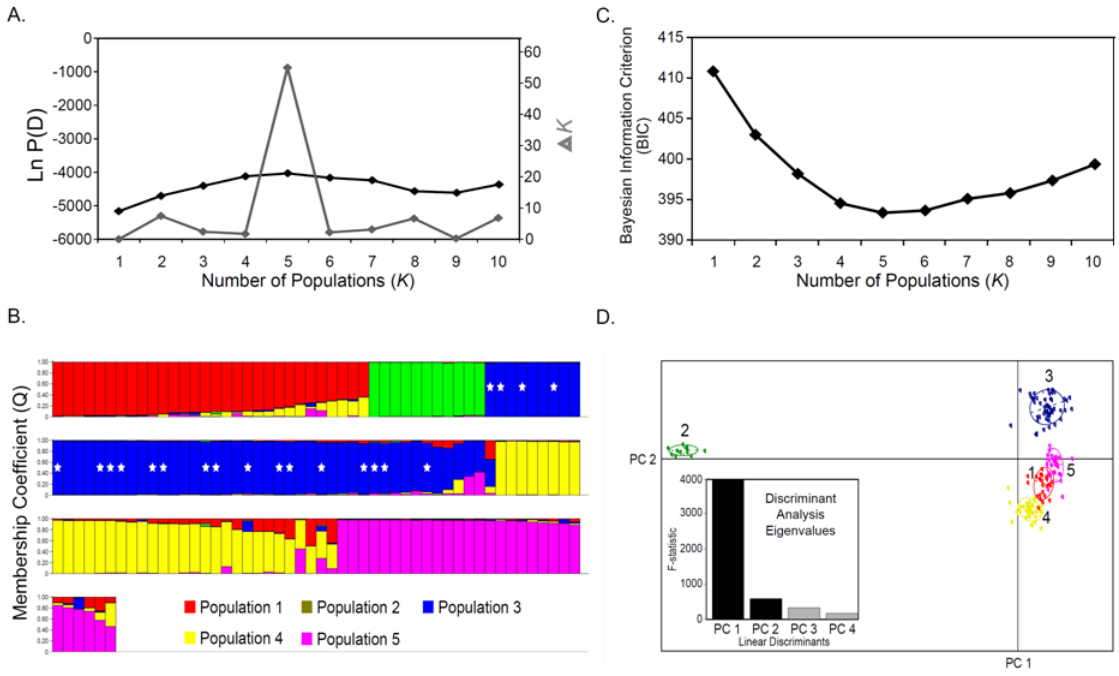
A. fumigatus Genotypes Belong to Five Distinct Populations

We inferred genetic differentiation for the 156 non-clonally related genotypes using the STRUCTURE model-based approach (Pritchard, Stephens et al. 2000) and the DAPC non model-based approach (Jombart, Devillard et al. 2010), independently for the microsatellite, the non-microsatellite, and the full marker data sets. Although analyses with both approaches on the microsatellite and non-microsatellite data sets give different answers as to the optimal number of populations, DAPC analysis of the microsatellite data

set and STRUCTURE analysis of the non-microsatellite data set both estimate that the optimal number of populations is five (Figure S3.2). Analyses of the full marker data set with both the STRUCTURE and the DAPC approach support this inference. Specifically, both the LnP(D) (-4035.01) and ΔK (54.91) approaches in STRUCTURE analysis indicate that $K = 5$ (Figure 3.2A), and so does the BIC in the DAPC analysis, which reaches its minimum value (393.37) for $K = 5$ as well as displays its smallest increase from $K = 5$ (393.37) to $K = 6$ (393.66) (Figure 3.2C).

We also examined whether our genetic differentiation structure inferences differed as we imposed different clonal correction thresholds on our marker data set (Figure S3.3). In the 225 genotype data set (generated by elimination of all but one genotype with identical alleles in all 9 microsatellite markers) the optimal number of populations predicted using the LnP(D) approach was 2. In contrast, the optimal number of populations predicted for all other clonally-corrected data sets was either 4 (for the genotype data sets generated by elimination of all but one genotype with identical alleles in 8 and 7 microsatellite markers, respectively), or 5 (for the genotype data sets generated by elimination of all but one genotype with identical alleles in 6 or fewer microsatellite markers). The very similar numbers of populations inferred by requiring elimination of all but one genotype with identical alleles in 8 or fewer microsatellite markers justifies our use of the "5 identical microsatellites" clonal correction threshold. Unless otherwise indicated, all subsequent analyses use the 156 non-clonally related genotype data set generated by the "5 identical microsatellites" clonal correction threshold, and its 5 inferred populations (as assigned by STRUCTURE).

Figure 3.2. Both STRUCTURE (panels A and B) and DAPC (panels C and D) analyses of 156 non-clonally related clinical and environmental genotypes identify the existence of five *A. fumigatus* populations in the Netherlands. (A) STRUCTURE analysis estimates that the optimal predicted number of populations *K* for our set of genotypes is five. This inference is supported by both the average log probability (LnP(D)) of each *K* value (black line) and by the *ad hoc* statistic Δ*K* (grey line). (B) The STRUCTURE based assignment of 156 genotypes into the five *A. fumigatus* populations. Each column on the X-axis corresponds to a different genotype. The Y-axis represents an individual's membership coefficient to each population. White stars indicate multi-triazole resistant (MTR) individuals. STRUCTURE populations 1 – 5 are indicated by red, green, blue, yellow and pink color, respectively. (C) DAPC analysis estimates that the optimal predicted number of populations *K* for our set of genotypes is five. The Y-axis corresponds to the Bayesian Information Criterion (BIC), a goodness of fit measurement calculated for each *K*. The lowest BIC value (*K* = 5) indicates the optimal number of populations. (D) DAPC clustering of the five populations using the first two principal components (Y-axis and X-axis, respectively). The first four eigenvalue components are show in the lower left panel. DAPC populations 1 – 5 are indicated by red, green, blue, yellow and pink color, respectively and are highly similar to STRUCTURE delineated populations.



Seventeen of the 20 markers used in this study contribute significantly to the observed genetic differentiation (with the CSP marker showing the strongest association with the different populations; Cramér's V statistic = 82.5%), whereas the remaining three markers (RM-1, RM-3 and RM-7) show a random distribution over the five populations (Figures S3.4 – S3.6).

Our results also suggest that the ancestry of several genotypes in the five populations traces to more than one population. For example, the results from the STRUCTURE approach indicate that several genotypes from population 1 share contributions from population 4 and vice versa (Figure 3.2B), indicating either that these two populations share a more recent common ancestor and/or provide evidence of recent admixture between them. Furthermore, the results from the DAPC approach point to a clear separation of population 2, and to a lesser extent of population 3, from all other populations; in contrast, population 1 genotypes show some overlap with genotypes from population 4 and population 5 genotypes (Figure 3.2D). Interestingly, we found that all 20 genotypes containing the MTR allele nested within population 3 (Figure 3.2B), a result that was supported by both approaches.

Finally, we note that the assignment of individual genotypes into the five populations is highly concordant across STRUCTURE replicates as well as between the DAPC and STRUCTURE analyses. Using CLUMPP (Jakobsson and Rosenberg 2007), we calculated average individual membership coefficients for the 100 STRUCTURE replicates and found that population assignments are identical across runs and membership coefficients nearly

identical to the replicate used in this study (Figure S3.7). Between the DAPC and STRUCTURE analyses, only 8/156 genotypes, all of which have considerable STRUCTURE membership coefficients to more than one population, are assigned to different populations when analyzed using the two different approaches. For all subsequent analyses, we used the assignment of individual genotypes into the five populations from the STRUCTURE approach.

A. fumigatus Genotype Geography is not Associated with Genetic Differentiation

Two different analyses failed to provide any evidence in support of the hypothesis that the geographical origin of genotypes was associated with genetic differentiation (as inferred by STRUCTURE). Specifically, we do not identify any global population differentiation when genotypes are grouped into populations by their city of origin (φ_{PT} = 0.004, p = 0.39). Similarly, we find only 4/66 cases of significant population differentiation when we calculate φ_{PT} values pairwise between cities (these were: Alkmaar vs. Haarlem: φ_{PT} = 0.056, p = 0.036; Alkmaar vs. Rotterdam: φ_{PT} = 0.030, p = 0.038; Arnhem vs. Goes: φ_{PT} = 0.153, p = 0.020 and Arnhem vs. Veldhoven: φ_{PT} = 0.126, p = 0.026) (Figure S3.3). Conversely, when genotypes are grouped by STRUCTURE, we find significant global (φ_{PT} = 0.223, p = 0.001) and pairwise differentiation for each population (p = 0.001 for all comparisons) (Figure S3.8). Finally, the AMOVA analysis shows that almost none of the genetic variation among populations is explained when genotypes are grouped by city of origin; in contrast, 23% of the genetic variation among populations is explained when genotypes are grouped by the STRUCTURE-assigned populations (Figure S3.8).

Using a second analysis, we also tested whether genotypes from particular cities are represented disproportionately in specific populations. For each population, our statistical analyses reject an association between geography and population assignment ($p_{population1}$ = 0.96, $p_{population2}$ = 0.83, $p_{population3}$ = 0.78, $p_{population4}$ = 0.45 and $p_{population5}$ = 0.90).

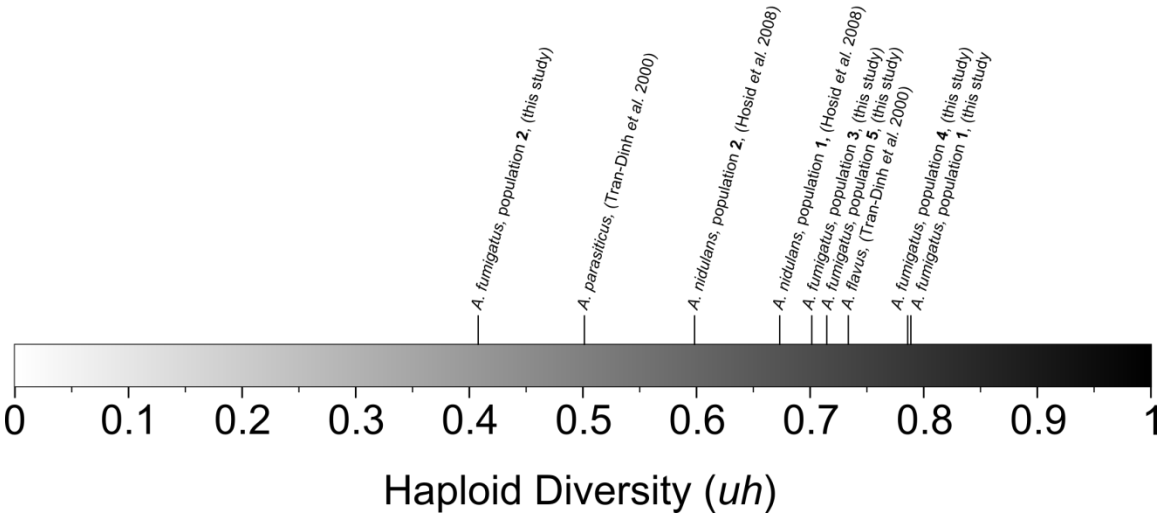
Haploid Diversity in the Five *A. fumigatus* Populations

Levels of haploid diversity are variable across the five *A. fumigatus* populations. Populations 1, 3, 4, and 5 show relatively high levels of haploid diversity ($uh_{population1}$ = 0.599, $uh_{population3}$ = 0.540, $uh_{population4}$ = 0.586 and $uh_{population5}$ = 0.531), whereas population 2 exhibits relatively lower levels of diversity ($uh_{population2}$ = 0.388).

To make estimates of haploid diversity comparable to previously published microsatellite analyses from other *Aspergillus* species (Tran-Dinh and Carter 2000; Hosid, Grishkan et al. 2008), we also calculated *uh* using only our microsatellite markers (Figure 3.3). Again, we found that populations 1, 3, 4, and 5 show higher relative levels of haploid diversity ($uh_{population1}$ = 0.788, $uh_{population3}$ = 0.702, $uh_{population4}$ = 0.787 and $uh_{population5}$ = 0.711), compared to population 2 ($uh_{population2}$ = 0.408). Interestingly, *uh* values obtained for populations 1, 3, 4 and 5 are comparable to those found in two populations of *A. nidulans* ($uh_{population1}$ = 0.675 and $uh_{population2}$ = 0.598) (Hosid, Grishkan et al. 2008), a population of *A. flavus* (*uh* = 0.733) (Tran-Dinh and Carter 2000) and, to a lesser extent, a population of *A. parasiticus* (*uh* = 0.505) (Tran-Dinh and Carter 2000), even though different sets of markers were used in different studies. We observed few fixed loci in all

populations (population 1 = 0, population 3 = 1, population 4 = 0 and population 5 = 0) with the exception of population 2, which contained seven fixed loci.

Figure 3.3. Unbiased haploid diversity (*uh*) measures of the five *A. fumigatus* populations and other representative *Aspergillus* species. Microsatellite-based *uh* values from populations of other representative *Aspergillus* species are from the following studies: *A. flavus* (Tran-Dinh and Carter 2000), *A. parasiticus* (Tran-Dinh and Carter 2000), and two *A. nidulans* populations (Hosid, Grishkan et al. 2008).

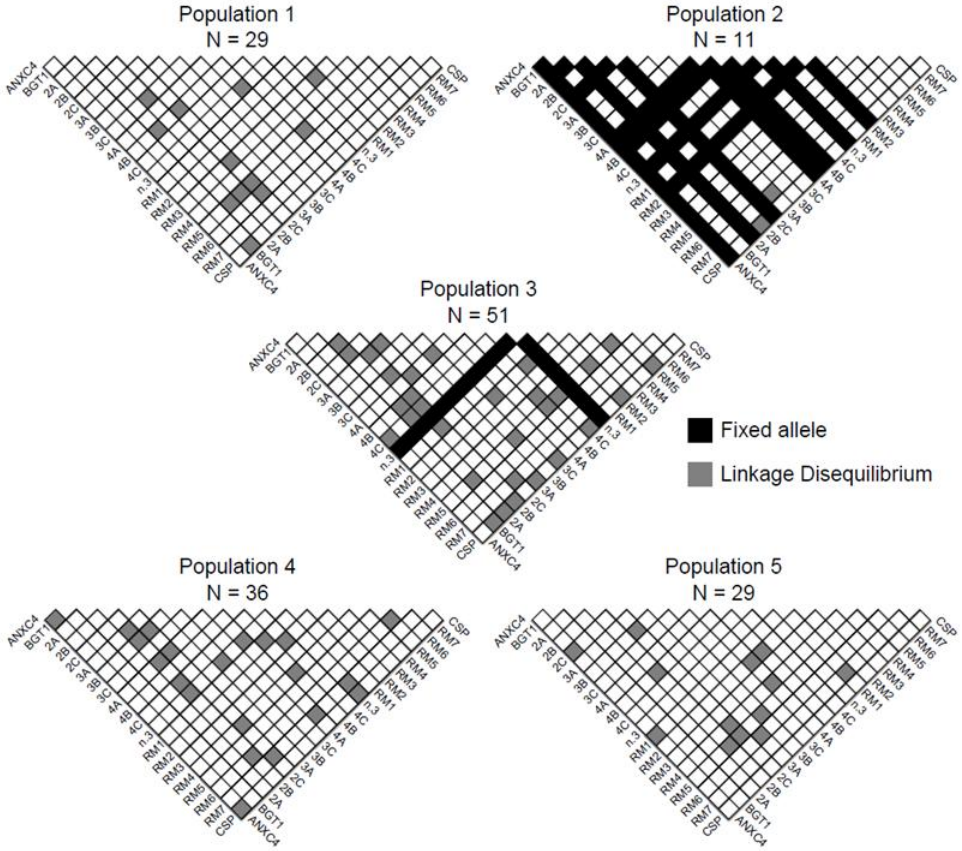


Recombination Levels Vary between the Five *A. fumigatus* Populations

Examination of LD suggests that *A. fumigatus* populations show varying levels of recombination (Figure 3.4). Globally, populations 2 – 5 showed significant *Ia* values indicative of population-level LD (population 2: $Ia = 0.588$, $p = 0.003$; population 3: $Ia = 0.429$, $p < 0.001$; population 4: $Ia = 0.374$, $p < 0.001$; population 5: $Ia = 0.471$, $p < 0.001$), whereas values for population 1 ($Ia = 0.106$, $p = 0.114$) are indicative of a recombining population. Furthermore, of the 190 possible locus pairs tested, 12, 135, 47, 21 and 13 are fixed or in LD in populations 1–5, respectively (Figure 3.4). Of the 76 observed locus pairs that were in LD, only 16 pairs involved markers located on the same

chromosome, of which only nine involved neighboring markers not separated by another marker.

Figure 3.4. Linkage disequilibrium (LD) patterns of the five *A. fumigatus* populations. LD was determined by calculating the Index of Association (*Ia*) for all locus pairs independently for all populations. White, grey and black boxes represent loci in equilibrium, loci in significant LD, and fixed loci, respectively.



Non-recombining populations are expected to contain higher numbers of clonally related genotypes. Thus, clonal richness, i.e., the number of identical genotypes in a population, might be used as another indicator of reproductive mode. Levels of clonal richness across the five *A. fumigatus* populations were 50% (29 of 58 isolates belonging to population 1

were inferred to have identical genotypes) for population 1, 35% (6 of 17) for population 2, 30% (22 of 73) for population 3, 45% (30 of 66) for population 4, and 29% (12 of 41) for population 5.

Divergence Times between *A. fumigatus* Populations

We estimated the average mutation rate, *w*, at 2.97 x 10^{-4}, a value comparable to that obtained for *A. flavus* (Grubisha and Cotty 2010). The pairwise upper and lower divergence time estimates between population 3, which contains the MTR alleles, and all others were (in 1,000 generation multiples): Population 3 vs. 1: 102 – 19; vs. 2: 80 – 23; vs. 4: 111 – 33; and vs. 5: 109 – 51 (Table 3.1). We also estimated of upper and lower boundaries of divergence times (again in 1,000 generation multiples) between the MTR allele containing genotypes within population 3 and populations 1 (113 – 20), 2 (49 – 6), 4 (94 – 16), and 5 (58 – 7) (Table 3.1).

Table 3.1. Estimated times of divergence between *A. fumigatus* populations with upper and lower divergence time estimates given in units of 1,000 generations.

Population	#1	#2	#3	#4	#5
#2	167–67				
#3	102–19	80–23			
#4	76–0	102–32	111–33		
#5	54–0	28–0	109–51	68–6	
Multiple-triazole-resistance-containing isolates within population #3	113–20	49–6	N/A	94–16	58–7

DISCUSSION

To investigate the role of reproductive mode and population structure in evolution of triazole drug resistance, we studied the *A. fumigatus* populations in the Netherlands, where the multi-triazole-resistant TR/L98H allele likely originated and was first reported (Verweij, Mellado et al. 2007). Given the evidence that recombination, commonly associated with sexual reproduction, can be very advantageous in stressful environments (Goddard, Godfray et al. 2005; Zeyl, Curtin et al. 2005), such as inside a human host or upon exposure to a fungicide, we hypothesized that sexual reproduction was facilitating the emergence and/or spread of the TR/L98H allele. However, analysis of the data obtained in this study allowed us to reject our original hypothesis, suggesting instead that all TR/L98H alleles identified in our study nest within a single, predominantly asexual, population and have not spread across populations. These results emphasize the role of asexual reproduction and genetic differentiation in shaping the evolution of azole resistance in *A. fumigatus*.

High-resolution Markers Show Genetic Differentiation in Dutch *A. fumigatus*

To our knowledge, this is the first study that reports the existence of genetic differentiation in any part of the distribution of *A. fumigatus*, one of the most important opportunistic fungal pathogens of humans, and suggests that Dutch *A. fumigatus* genotypes group into five distinct populations. Consistent with our findings, another recent study reported the existence of two genetically differentiated lineages within *A. fumigatus* (Pringle, Baker et al. 2005), although in that case the authors argued, based on phylogenetic analysis, that these lineages were separate species. The detection of genetic

differentiation in *A. fumigatus* argues against the "everything is everywhere" hypothesis, which states that highly abundant microbial eukaryote species with cosmopolitan distributions lack genetic differentiation (Finlay 2002).

One potential explanation for this discrepancy between past studies and ours might be the difference in the density and scale of sampling between studies. For example, in the two most comprehensive multilocus studies to date, 63 and 70 isolates from five different continents were analysed, respectively (Pringle, Baker et al. 2005; Rydholm, Szakacs et al. 2006), whereas our study examined a much larger number of isolates from a relatively small geographical area.

Another potential explanation for the discordance of these results with those from earlier studies might be our use of a much larger and more highly informative panel of markers (Figures S3.4 and S3.5). All previous studies have employed either sequence-based or RFLP-based typing techniques (Debeaupuis, Sarfati et al. 1997; Pringle, Baker et al. 2005; Rydholm, Szakacs et al. 2006). Although these techniques are very reliable (Taylor, Geiser et al. 1999), they are typically less informative when compared to microsatellite markers (Bain, Tavanti et al. 2007). If the absence of differentiation in past studies of *A. fumigatus* population biology is to be explained by the use of less informative markers, then we expect that future studies on isolates from other geographic regions of comparable size using similar or superior markers (e.g., Harris, Feil et al. 2010) to ours are highly likely to identify genetically differentiated *A. fumigatus* populations. Interestingly, a previous study of *A. fumigatus* isolates performed using the

most informative non-microsatellite marker from our panel did not find any evidence of genetic differentiation in North America (Balajee, Tay et al. 2007), suggesting that the pattern of differentiation of this important human pathogen might vary across its range of distribution.

Our results also show that the five identified *A. fumigatus* populations do not correlate with geography. In many eukaryotes, the presence of genetically differentiated populations is often the result of geographical isolation or ecological niche preference. Surprisingly, the populations in our study do not show any correlation with geography or environment of origin (clinical versus soil). A similar lack of correlation with geography was observed in *A. flavus* populations (Grubisha and Cotty 2010). In this study, the genetic diversity of 243 *A. flavus* isolates was analysed using 24 microsatellite loci and the mating type locus. Notably, all *A. flavus* populations were clonal, with no evidence of gene flow between populations.

Varying Levels of Recombination among Dutch *A. fumigatus* Populations

Four of the five Dutch populations show significant levels of LD, suggesting that recombination in these populations has been rare or absent. Significant LD can be due to several different reasons (Maynard Smith, Smith et al. 1993). For example, significant LD is expected in populations of organisms that do not possess any molecular mechanism for recombination such that clonal propagation is their sole means of reproduction. This explanation is unlikely to hold true for *A. fumigatus*. The sexual reproduction machinery in the *A. fumigatus* genome appears intact (Galagan, Calvo et al. 2005; Rokas and

Galagan 2008), the MAT loci are typically at near-equal frequencies in *A. fumigatus* populations (Paoletti, Rydholm et al. 2005) – this is also the case in our populations, and certain isolates can reproduce sexually in the laboratory (O'Gorman, Fuller et al. 2009).

Another reason for significant LD values involves failure to account for the genetic differentiation of the species studied. In such cases, the presence of LD will likely reflect the lack of recombination between populations of the species, potentially masking recombination within populations. For example, analysis of the entire 156-genotype data set reveals significant levels of LD ($Ia = 0.718$, $p < 0.001$), masking the finding that population 1 is not in LD. Given that we first identified the pattern of genetic differentiation in *A. fumigatus* and then calculated LD separately for each genetically differentiated population, it is highly unlikely that our results are affected by this reason.

Furthermore, the observed varying levels of LD suggest that different *A. fumigatus* populations may exhibit different reproductive modes. Nearly two decades ago, Maynard Smith and colleagues distinguished microbial reproductive modes into three models: clonal (predominantly non-recombining), panmictic (predominantly recombining), and epidemic (predominantly recombining but shows significant associations between loci due to recent, explosive increases in particular genotypes) (Maynard Smith, Smith et al. 1993). On a first level of analysis, it appears that the observed pattern of reproductive modes across *A. fumigatus* populations, where four of the five populations are predominantly non-recombining, might be more similar to the epidemic model than to either the clonal or the panmictic one. One expectation of the epidemic model is that

clonal populations are very recent and have undergone explosive growth, so they are expected to harbour little genetic diversity. However, neither the estimated times of divergence nor the levels of genetic diversity for most of the non-recombining *A. fumigatus* populations support the very recent origins. Similarly, levels of genotypic diversity are comparable across most populations, irrespective of the presence or absence of recombination. Rather, it appears that different *A. fumigatus* populations fit into different reproductive mode models.

How this variation in reproductive mode across *A. fumigatus* populations is controlled and whether these distinct populations use either sexual or asexual reproductive modes to regulate gene flow regardless of environmental signals remain open questions. It is known that both asexual and sexual reproductive modes can confer significant advantages as well as disadvantages to fungal populations (Sun and Heitman 2011), so the pattern of reproductive modes across populations might be determined by their balance. It is also theoretically possible that the only population of *A. fumigatus* identified in this study as recombining may in fact be reproducing asexually. This is so because, in fungi, genetic recombination can occur both during meiosis as part of the sexual cycle or during mitosis, as part of the parasexual cycle (Clutterbuck 1996; Taylor, Jacobson et al. 1999). The effect of the parasexual cycle on long-term genetic exchange within fungal populations is thought to be limited because it typically takes place between genetically similar individuals (Clutterbuck 1996), but its true extent is unknown on recombination within *A. fumigatus* populations is unknown.

Population Structure, Recombination and their Implications for the Spread of the MTR Allele

Our finding that all genotypes containing the TR/L98H allele were confined to a single, predominantly asexually reproducing, population and has not yet spread across populations rejects our hypothesis that sexual reproduction facilitated the emergence and/or spread of the TR/L98H allele in *A. fumigatus*. Notably, very recent global surveillance studies have found isolates with the TR/L98H allele in both China and India (Lockhart, Frade et al. 2011; Chowdhary et al. unpublished data), but it is not yet known whether these isolates stem from the same population.

This lack of recombination and lack of gene flow is puzzling; in environments lacking triazole drugs, retention of MTR-like alleles is likely to be costly (Cowen, Kohn et al. 2001; Stergiopoulos, van Nistelrooy et al. 2003; Cowen 2008), whereas in environments containing triazole drugs MTR-like alleles are likely to be strongly advantageous. In both cases, recombination (and associated gene flow) should be favored – to eliminate the costly resistant allele in the first case, or to fix the advantageous resistant allele in the second, and yet our evidence suggests that the population is predominantly asexual. One possible explanation is that the sexual reproduction mode in *A. fumigatus* populations cannot be "switched on" instantaneously once a population has become asexual because of the accumulation of deleterious mutations in mating and meiosis genes (Sun and Heitman 2011), despite the high costs associated with clonal reproduction during episodic selection, e.g., upon sporadic exposure to a fungicide.

When did the TR/L98H allele originate? Recently, Verweij and colleagues raised the hypothesis that the evolution of MTR resistance in *A. fumigatus* might have been a by-product of the use of azole compounds in agriculture (Verweij, Snelders et al. 2009). Dating of the divergence of population 3 genotypes as well as of all TR/L98H allele containing genotypes within population 3 from the other populations (Table 3.1), implies that both groups diverged from the other populations at least 6,000 generations ago. Thus, the genetic background of MTR allele-containing genotypes is likely more ancient than the first use of azole drugs in agriculture or medicine. Nevertheless, it is entirely plausible that the origin of the TR/L98H allele in a population 3 genetic background occurred much more recently (Verweij, Snelders et al. 2009), a hypothesis consistent with the pattern of resistance spreading from the Netherlands to the rest of Europe. Future studies that examine the global genetic structure of *A. fumigatus* using a similarly informative set of markers are likely to be highly instructive on how the genetic differentiation and reproductive mode of *A. fumigatus* populations shape the evolutionary dynamics of drug resistance patterns of this deadly human pathogen.

From a practical standpoint, elucidating the role of genetic differentiation and reproductive mode in influencing the genetic structure of *A. fumigatus* can also facilitate the development of diagnostic tools to detect resistant infections and preventive measures aimed to curb the spread of azole resistance in fungal populations. In several European countries, the frequency of azole resistant *A. fumigatus* isolates has reached 12-55%; as this frequency continues to increase, a better understanding of the origin and spread of MTR alleles across *A. fumigatus* is becoming critical.

ACKNOWLEDGEMENTS

We thank Zeev Frenkel, Abraham Korol, and Nai Tran-Dinh for access to raw *Aspergillus* genotype data, Dr. A. Chowdhary for access to unpublished data, and Thibaut Jombart for assistance with the DAPC approach. We are also grateful to David McCauley and the reviewers for their helpful suggestions and advice. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract numbers N01-AI30071 and / or HHSN272200900007C, the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH, NIAID: F31AI091343-01 to JGG), the Searle Scholars Program (AR), and the National Science Foundation (DEB-0844968 to AR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the NIH.

J.G.G., C.H.W.K., N.D.F., J.F.M. and A.R. designed the study. J.G.G. performed the population genetics analysis. C.H.W.K. genotyped the isolates. A.R. calculated ω for markers.

REFERENCES

Agapow, P. M. and A. Burt (2001). "Indices of multilocus linkage disequilibrium." Molecular Ecology Notes **1**(1-2): 101-102.

Avise, J. C. (2000). Phylogeography: The History and Formation of Species. Cambridge, Harvard Univ. Press.

Bain, J. M., A. Tavanti, et al. (2007). "Multilocus sequence typing of the pathogenic fungus *Aspergillus fumigatus*." Journal of Clinical Microbiology **45**(5): 1469-77.

Balajee, S. A., S. T. Tay, et al. (2007). "Characterization of a novel gene for strain typing reveals substructuring of *Aspergillus fumigatus* across North America." Eukaryotic Cell **6**(8): 1392-9.

Butler, G. (2010). "Fungal sex and pathogenesis." Clinical Microbiology Reviews **23**(1): 140-59.

Clutterbuck, A. J. (1996). "Parasexual recombination in fungi." Journal of Genetics **75**(3): 281-286.

Cowen, L. E. (2008). "The evolution of fungal drug resistance: modulating the trajectory from genotype to phenotype." Nature Reviews Microbiology **6**(3): 187-98.

Cowen, L. E., L. M. Kohn, et al. (2001). "Divergence in fitness and evolution of drug resistance in experimental populations of *Candida albicans*." Journal of Bacteriology **183**(10): 2971-2978.

Cramér, H. (1999). Mathematical Methods of Statistics. Uppsala, Princeton University Press.

de Valk, H. A., J. F. Meis, et al. (2005). "Use of a novel panel of nine short tandem repeats for exact and high-resolution fingerprinting of *Aspergillus fumigatus* isolates." Journal of Clinical Microbiology **43**(8): 4112-20.

Debeaupuis, J. P., J. Sarfati, et al. (1997). "Genetic diversity among clinical and environmental isolates of *Aspergillus fumigatus*." Infection and Immunity **65**(8): 3080-5.

Denning, D. W. (1998). "Invasive aspergillosis." Clinical Infectious Diseases **26**(4): 781-803.

Evanno, G., S. Regnaut, et al. (2005). "Detecting the number of clusters of individuals using the software structure: a simulation study." Molecular Ecology **14**(8): 2611-2620.

Fedorova, N. D., S. Harris, et al. (2009). "Using aCGH to study intraspecific genetic variability in two pathogenic molds, *Aspergillus fumigatus* and *Aspergillus flavus*." Medical Mycology **47 Suppl 1**: S34-41.

Fedorova, N. D., N. Khaldi, et al. (2008). "Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*." PLoS Genetics **4**(4): e1000046.

Finlay, B. J. (2002). "Global dispersal of free-living microbial eukaryote species." Science **296**(5570): 1061-3.

Galagan, J. E., S. E. Calvo, et al. (2005). "Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*." Nature **438**(7071): 1105-15.

Geiser, D. M., W. E. Timberlake, et al. (1996). "Loss of meiosis in *Aspergillus*." Molecular Biology and Evolution **13**(6): 809-17.

Goddard, M. R., H. C. J. Godfray, et al. (2005). "Sex increases the efficacy of natural selection in experimental yeast populations." Nature **434**(7033): 636-640.
Goldstein, D. B., A. R. Linares, et al. (1995). "An evaluation of genetic distances for use with microsatellite loci." Genetics **139**(1): 463-471.
Grubisha, L. C. and P. J. Cotty (2010). "Genetic isolation among sympatric vegetative compatibility groups of the aflatoxin-producing fungus *Aspergillus flavus*." Molecular Ecology **19**(2): 269-80.
Harris, S. R., E. J. Feil, et al. (2010). "Evolution of MRSA during hospital transmission and intercontinental spread." Science **327**(5964): 469-474.
Heitman, J. (2006). "Sexual reproduction and the evolution of microbial pathogens." Current Biology **16**(17): R711-R725.
Herbrecht, R., D. W. Denning, et al. (2002). "Voriconazole versus amphotericin B for primary therapy of invasive aspergillosis." New England Journal of Medicine **347**(6): 408-15.
Hittinger, C. T., P. Goncalves, et al. (2010). "Remarkably ancient balanced polymorphisms in a multi-locus gene network." Nature **464**(7285): 54-8.
Hosid, E., I. Grishkan, et al. (2008). "Diversity of microsatellites in natural populations of ascomycetous fungus, *Emericella nidulans*, in Israel on local and regional scales." Mycological Progress **7**(2): 99-109.
Jakobsson, M. and N. A. Rosenberg (2007). "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure." Bioinformatics **23**(14): 1801-1806.
Jombart, T. (2008). "adegenet: a R package for the multivariate analysis of genetic markers." Bioinformatics **24**(11): 1403-5.
Jombart, T., S. Devillard, et al. (2010). "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations." BMC Genetics **11**: 94.
Klaassen, C. H., H. A. de Valk, et al. (2009). "Utility of CSP typing to sub-type clinical *Aspergillus fumigatus* isolates and proposal for a new CSP type nomenclature." Journal of Microbiological Methods **77**(3): 292-6.
Klaassen, C. H., H. A. de Valk, et al. (2010). "Novel mixed-format real-time PCR assay to detect mutations conferring resistance to triazoles in *Aspergillus fumigatus* and prevalence of multi-triazole resistance among clinical isolates in the Netherlands." Journal of Antimicrobial Chemotherapy **65**(5): 901-5.
Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-1452.
Lockhart, S. R., J. P. Frade, et al. (2011). "Azole resistance in *Aspergillus fumigatus* isolates from the ARTEMIS global surveillance is primarily due to the TR/L98H mutation in the *cyp51A* gene." Antimicrobial Agents and Chemotherapy: in press.
Lynch, M., W. Sung, et al. (2008). "A genome-wide view of the spectrum of spontaneous mutations in yeast." Proceedings of the National Academy of Sciences, USA **105**(27): 9272-7.
Maynard Smith, J., N. H. Smith, et al. (1993). "How clonal are bacteria?" Proceedings of the National Academy of Sciences, USA **90**(10): 4384-4388.
Meis, J. F. and P. E. Verweij (2001). "Current management of fungal infections." Drugs **61 Suppl 1**: 13-25.

Mellado, E., G. Garcia-Effron, et al. (2007). "A new *Aspergillus fumigatus* resistance mechanism conferring in vitro cross-resistance to azole antifungals involves a combination of cyp51A alterations." Antimicrobial Agents and Chemotherapy **51**(6): 1897-904.

Milgroom, M. G. (1996). "Recombination and the multilocus structure of fungal populations." Annual Review of Phytopathology **34**: 457-477.

Munkacsi, A. B., S. Stoxen, et al. (2008). "*Ustilago maydis* populations tracked maize through domestication and cultivation in the Americas." Proceedings of the Royal Society of London. Series B, Biological Sciences **275**(1638): 1037-1046.

Nielsen, K. and J. Heitman (2007). "Sex and virulence of human pathogenic fungi." Advances in Genetics **57**: 143-73.

Nierman, W. C., A. Pain, et al. (2005). "Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*." Nature **438**(7071): 1151-6.

O'Gorman, C. M., H. T. Fuller, et al. (2009). "Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*." Nature **457**(7228): 471-4.

Paoletti, M., C. Rydholm, et al. (2005). "Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*." Current Biology **15**(13): 1242-8.

Peakall, R. and P. E. Smouse (2006). "GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research." Molecular Ecology Notes **6**(1): 288-295.

Pringle, A., D. M. Baker, et al. (2005). "Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus *Aspergillus fumigatus*." Evolution **59**(9): 1886-99.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**: 945-959.

Rokas, A. (2009). "The effect of domestication on the fungal proteome." Trends Genet **25**(2): 60-3.

Rokas, A. and J. E. Galagan (2008). The *Aspergillus nidulans* genome and a comparative analysis of genome evolution in *Aspergillus*. The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods. G. H. Goldman and S. A. Osmani. Boca Raton, FL, CRC Press**:** 43-55.

Rokas, A., G. Payne, et al. (2007). "What can comparative genomics tell us about species concepts in the genus *Aspergillus*?" Studies in Mycology **59**: 11-7.

Rydholm, C., G. Szakacs, et al. (2006). "Low genetic variation and no detectable population structure in *Aspergillus fumigatus* compared to closely related *Neosartorya* species." Eukaryotic Cell **5**(4): 650-7.

Scannell, D. R., O. A. Zill, et al. (2011). "The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus." G3 **1**: 11-25.

Snelders, E., H. A. van der Lee, et al. (2008). "Emergence of azole resistance in *Aspergillus fumigatus* and spread of a single resistance mechanism." PLoS Medicine **5**(11): e219.

Stergiopoulos, I., J. G. M. van Nistelrooy, et al. (2003). "Multiple mechanisms account for variation in base-line sensitivity to azole fungicides in field isolates of *Mycosphaerella graminicola*." Pest Management Science **59**(12): 1333-1343.

Sun, S. and J. Heitman (2011). "Is sex necessary?" BMC biology **9**: 56.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**: 585-595.
Taylor, J., D. Jacobson, et al. (1999). "The evolution of asexual fungi: reproduction, speciation and classification." Annual Review of Phytopathology **37**: 197-246.
Taylor, J. W., D. M. Geiser, et al. (1999). "The evolutionary biology and population genetics underlying fungal strain typing." Clinical Microbiology Reviews **12**(1): 126-46.
Taylor, J. W., E. Turner, et al. (2006). "Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi." Philos Trans R Soc Lond B Biol Sci **361**(1475): 1947-63.
Thuillet, A. C., T. Bataillon, et al. (2005). "Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data." Genetics **169**(3): 1589-1599.
Tran-Dinh, N. and D. Carter (2000). "Characterization of microsatellite loci in the aflatoxigenic fungi *Aspergillus flavus* and *Aspergillus parasiticus*." Molecular Ecology **9**(12): 2170-2172.
Varga, J. and B. Toth (2003). "Genetic variability and reproductive mode of *Aspergillus fumigatus*." Infection, Genetics and Evolution **3**(1): 3-17.
Verweij, P. E., E. Mellado, et al. (2007). "Multiple-triazole-resistant aspergillosis." New England Journal of Medicine **356**(14): 1481-3.
Verweij, P. E., E. Snelders, et al. (2009). "Azole resistance in *Aspergillus fumigatus*: a side-effect of environmental fungicide use?" Lancet Infectious Diseases **9**(12): 789-95.
Yang, Z. (2006). Computational Molecular Evolution. Oxford, Oxford University Press.
Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular Biology and Evolution **24**(8): 1586-91.
Zeyl, C., C. Curtin, et al. (2005). "Antagonism between sexual and natural selection in experimental populations of *Saccharomyces cerevisiae*." Evolution **59**(10): 2109-15.
Zhivotovsky, L. A. (2001). "Estimating divergence time with the use of microsatellite genetic distances: Impacts of population growth and gene flow." Molecular Biology and Evolution **18**(5): 700-709.

SUPPLEMENTAL TABLES

Table S3.1. The nomenclature, description, marker type, genomic location, and origin of the 20 markers used in this study.

Marker	Locus	Description	Marker type	Coordinate (Chr: Mbp)	Reference
2A	AFUA_5G06790	Conserved hypothetical protein	Microsatellite	05:01.7	(de Valk et al. 2005)
2B	AFUA_1G11310	4-Coumarate-coa ligase (intron)	Microsatellite	01:02.0	(de Valk et al. 2005)
2C	AFUA_6G11450	C6 transcription factor, putative	Microsatellite	06:02.8	(de Valk et al. 2005)
3A	AFUA_4G09070	ATP binding L-PSP endoribonuclease family protein	Microsatellite	04:02.4	(de Valk et al. 2005)
3B	N.A.	non-coding	Microsatellite	06:01.8	(de Valk et al. 2005)
3C	N.A.	non-coding	Microsatellite	03:00.1	(de Valk et al. 2005)
4A	N.A.	non-coding	Microsatellite	06:01.1	(de Valk et al. 2005)
4B	N.A.	non-coding	Microsatellite	07:01.3	(de Valk et al. 2005)
4C	N.A.	non-coding	Microsatellite	08:01.7	(de Valk et al. 2005)
n.3	N.A.	non-coding	Indel	06:01.1	(de Valk et al. 2005)
BGT1	AFUA_1G11460	1,3-beta-glucanosyltransferase	Sequence	01:03.0	(Bain et al. 2007)
ANXC4	AFUA_3G07020	Annexin	Sequence	03:01.8	(Bain et al. 2007)
CSP	AFUA_3G08990	Cell-surface protein, putative	PCR	03:02.3	(Balajee et al. 2007)
RM1	AFUA_3G06160	Mating type protein	PCR	03:01.5	(Paoletti et al. 2005)
RM2	AFUA_2G00910	Pfs-NB-ARC-TPR domain protein	PCR	02:00.2	(Fedorova et al. 2008)
RM3	AFUA_2G17420	Pfs-NB-ARC domain protein	PCR	02:04.7	(Fedorova et al. 2008)
RM4	AFUA_6G07030	NACHT-Ankyrin domain protein	PCR	06:01.6	(Fedorova et al. 2008)
RM5	AFUA_4G01110	Beta-1,3-glucanase	PCR	04:00.3	(Fedorova et al. 2008)
RM6	AFUA_6G11710	Conserved hypothetical protein	PCR	06:02.9	(Fedorova et al. 2008)
RM7	AFUA_6G13820	Conserved hypothetical protein	PCR	06:03.5	(Fedorova et al. 2008)

SUPPLEMENTAL FIGURES

Figure S3.1. The effect of different clonal correction thresholds on the number of non-clonal genotypes identified from 255 Dutch *A. fumigatus* isolates. The X-axis corresponds to a wide range of clonal correction thresholds, from requiring that genotypes are identical in all markers (microsatellites and non-microsatellites) before considered clonal to requiring that genotypes are only identical across the non-microsatellite markers. The Y-axis shows the number of unique genotypes identified for the different clonal correction thresholds.

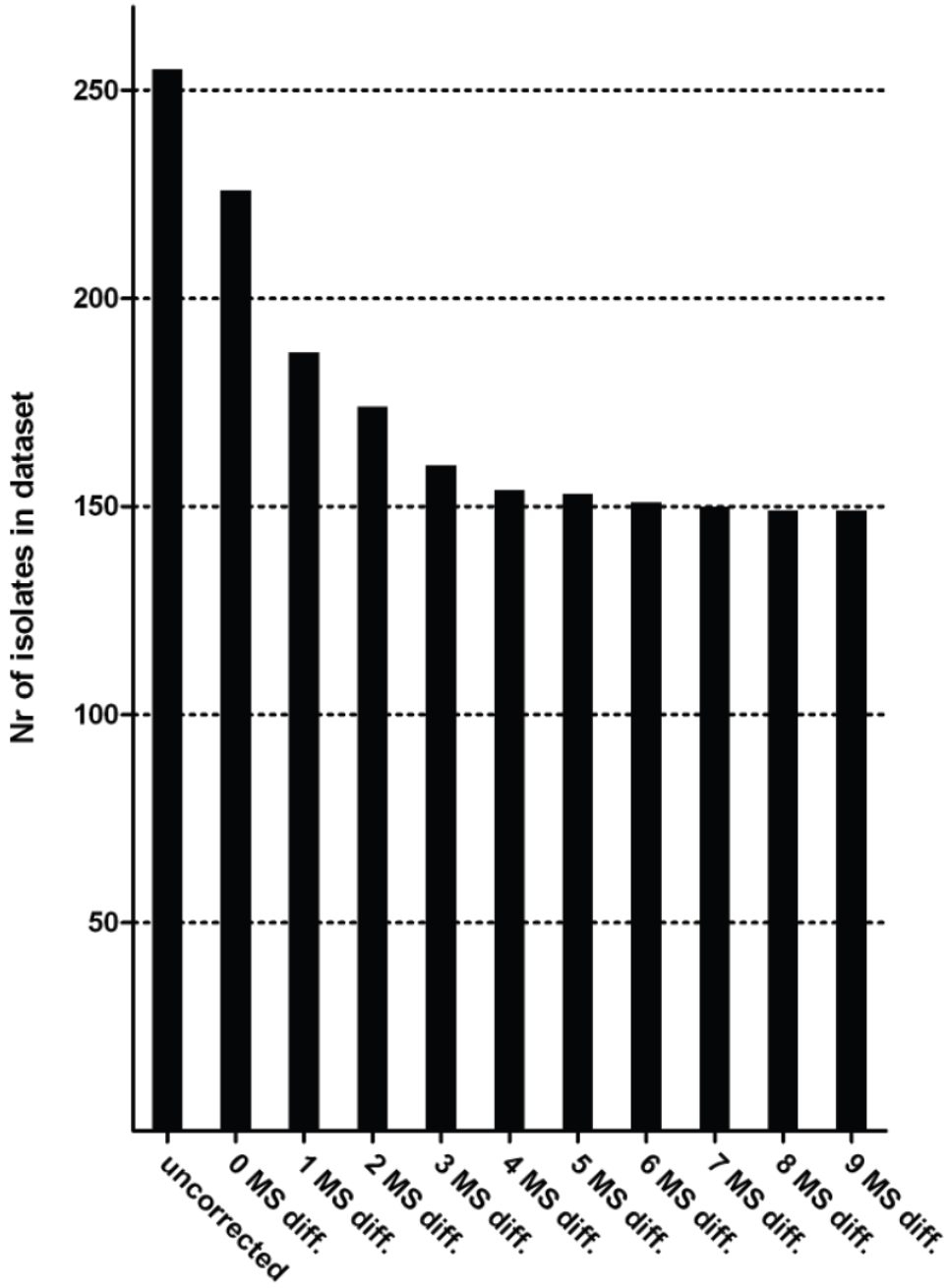


Figure S3.2. STRUCTURE and DAPC analysis of the microsatellite marker (panels A and B), non-microsatellite marker (panels C and D), and full marker (panels E and F) data sets. Both approaches predict $K = 5$ as the optimal number of populations for the full marker data set.

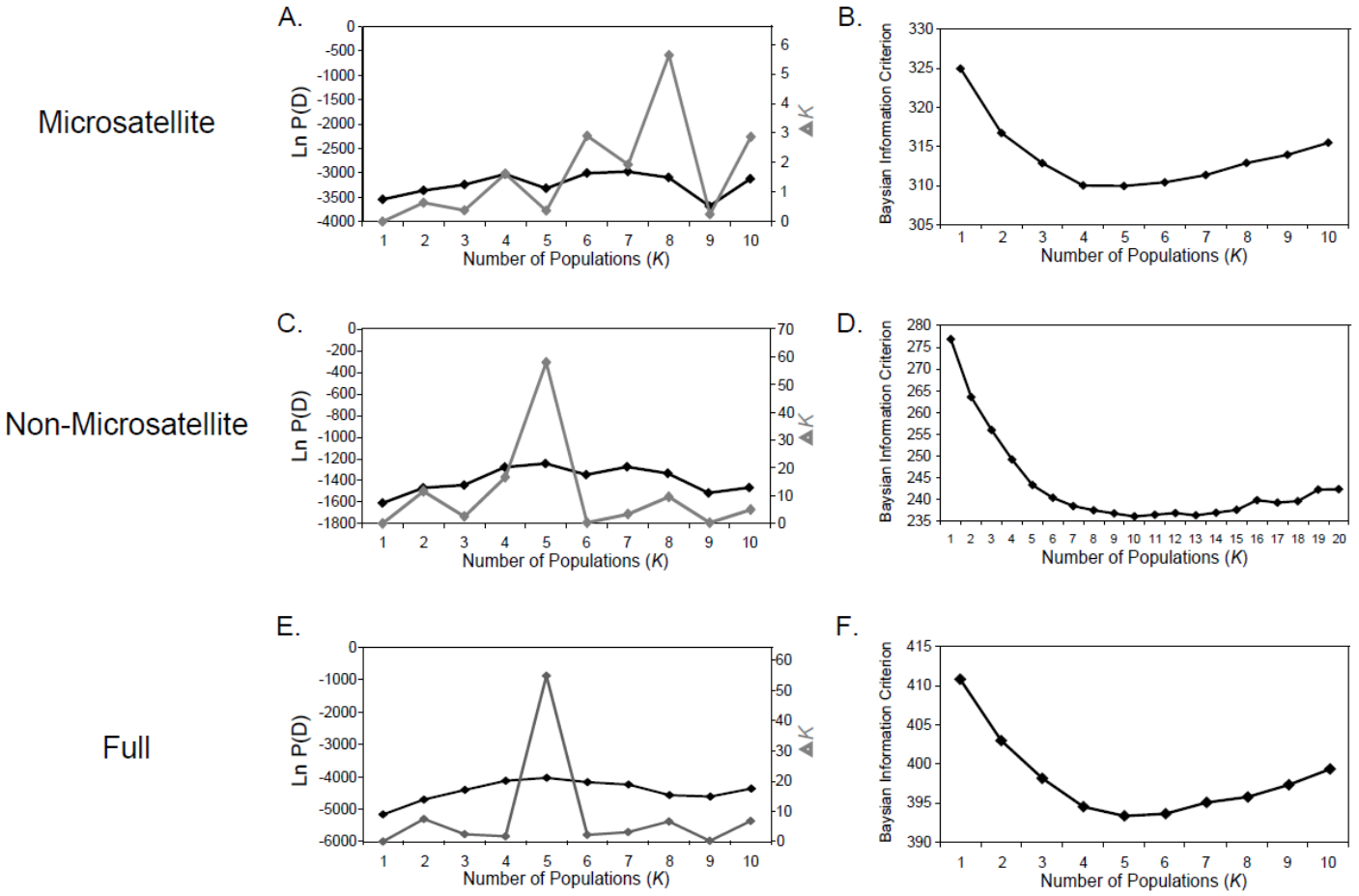


Figure S3.3. The optimal number of populations *K* predicted by STRUCTURE analysis on data sets with varying levels of clonal correction. For each data set, the average log probability (LnP(D)) of each *K* value (black line) and the *ad hoc* statistic Δ*K* (grey line) were calculated. The data sets are as follows: exclusion of all but one randomly chosen genotype with identical alleles for the n.3, ANXC4, BGT1, RM1-7, and CSP markers and with identical alleles in 9 (panel A), 8 (panel B), 7 (panel C), 6 (panel D), 5 (panel E), or 4 – 0 (panel F) identical microsatellite markers.

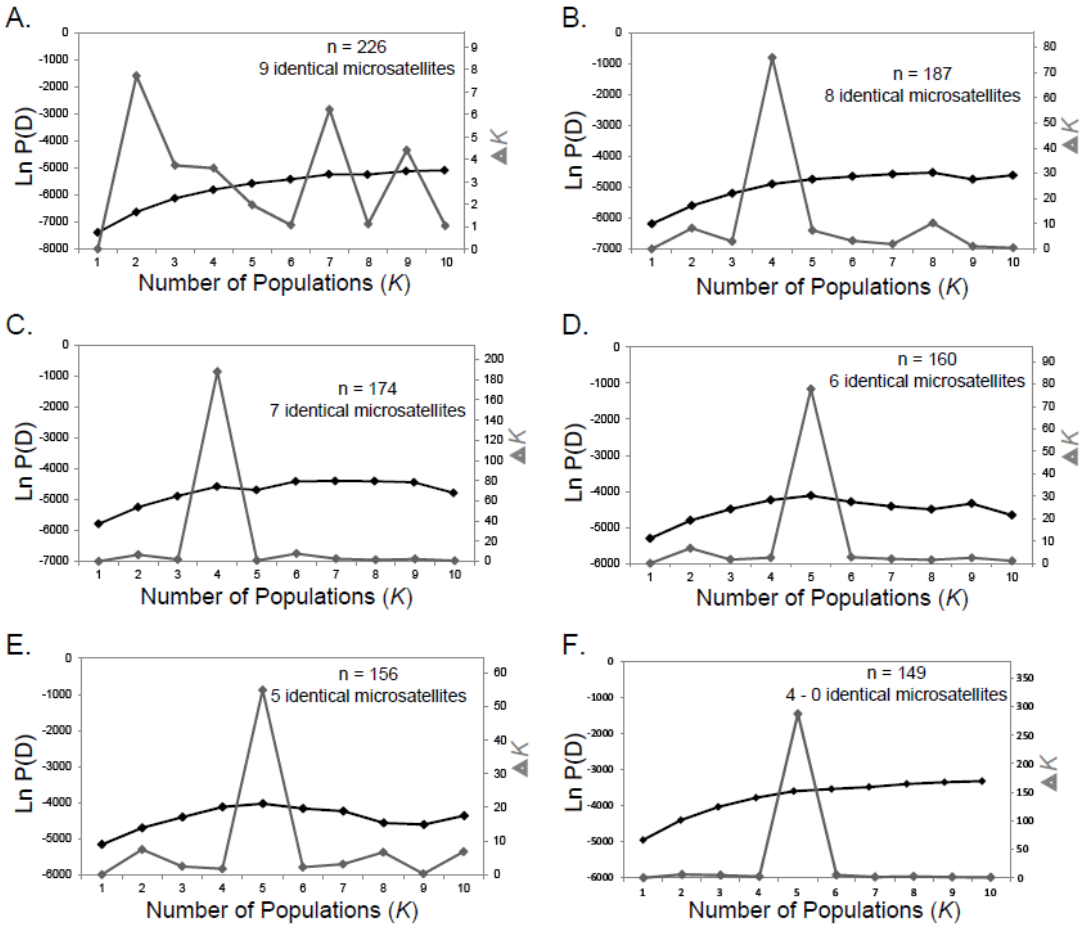


Figure S3.4. The distribution of alleles for each of the 9 microsatellite markers and the one indel marker used in this study across the five *A. fumigatus* populations, as delineated by the STRUCTURE analysis. Populations 1 – 5 are indicated by red, green, blue, yellow and pink color, respectively. The X-axis displays the different alleles for each marker, and the Y-axis the number of genotypes with that allele.

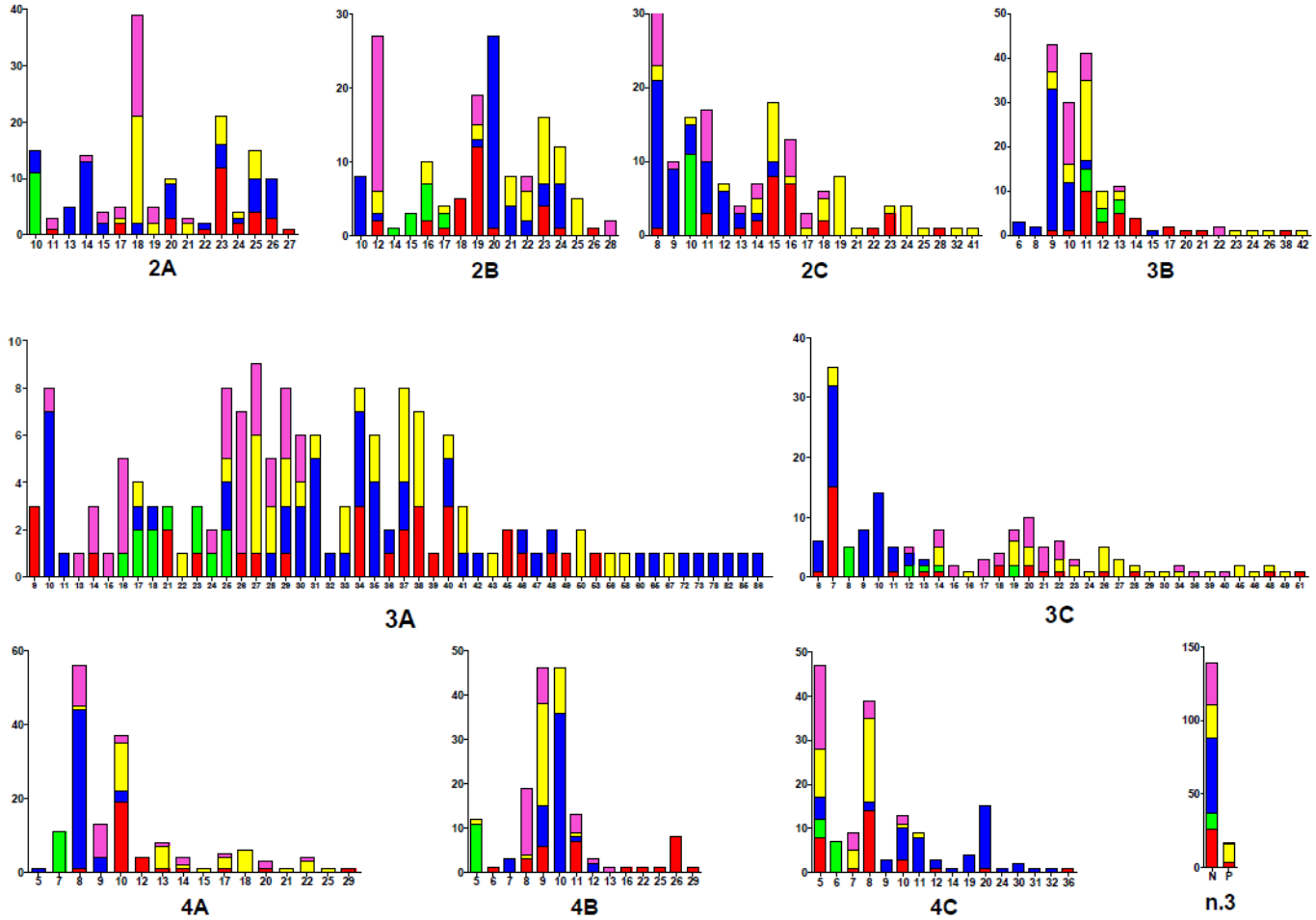


Figure S3.5. The distribution of alleles for each of the 10 sequence / PCR-typing markers used in this study across the five *A. fumigatus* populations, as delineated by the STRUCTURE analysis. Populations 1 – 5 are indicated by red, green, blue, yellow and pink color, respectively. The X-axis displays the different alleles for each marker, and the Y-axis the number of genotypes with that allele.

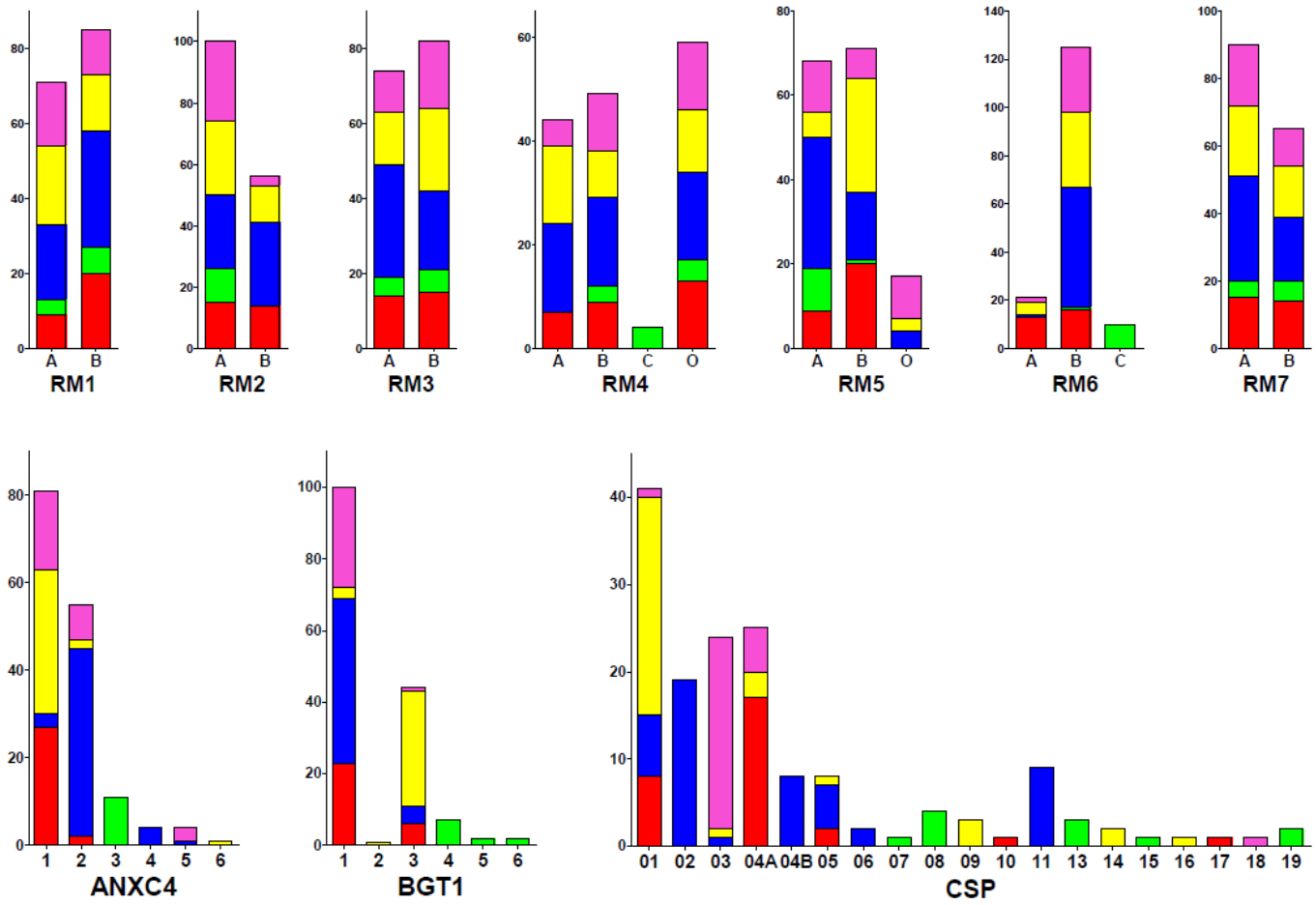


Figure S3.6. Strength of association between markers and populations according to Cramér's V statistic (Cramér 1999). Asterisks (*) denote markers that do not show a statistically significant association with population structure.

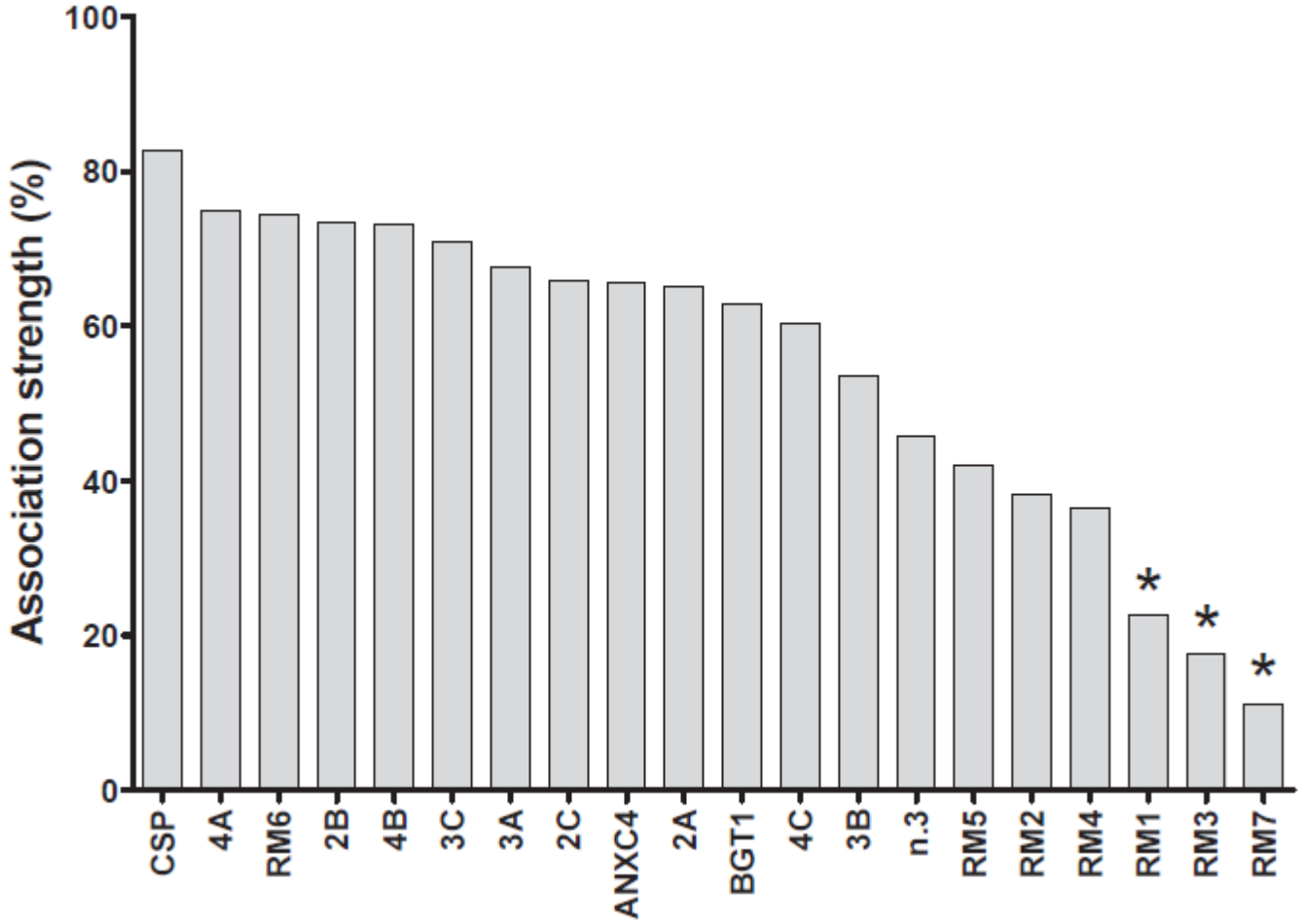


Figure S3.7. STRUCTURE analysis is highly consistent across replicates. Shown are the plots of the 100 STRUCTURE replicates after output data were processed using the CLUMPP software (Jakobsson and Rosenberg 2007) to correct for label switching across replicates. Samples were sorted first by population and then by numeric order (X axis). The Y axis represents an individual's membership coefficient to each population. STRUCTURE populations 1 – 5 are represented by red, green, blue, yellow and pink color, respectively.

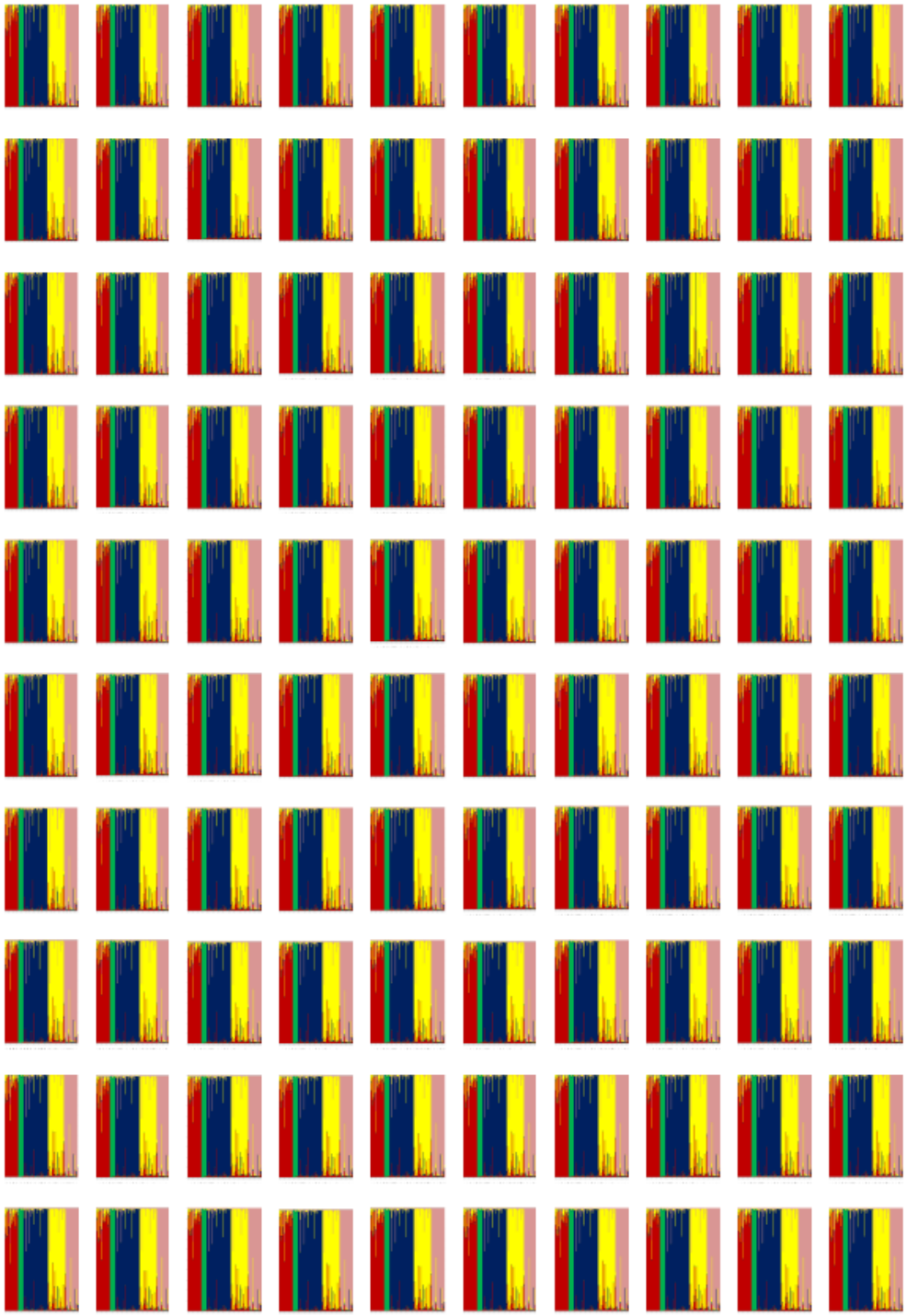
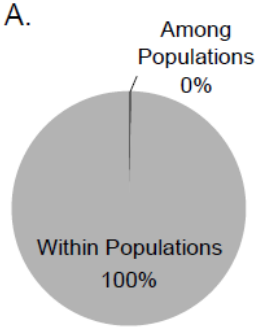


Figure S3.8. Geographic origin is not associated with genetic differentiation. (A) Individuals were grouped into populations based on their city of origin or (B) STRUCTURE population assignment. Pie charts represent AMOVA results explaining the variance found within and among populations. Tables represent pairwise φ_{PT} values (lower diagonal)and probability values of population differentiation based on 999 permutations (upper diagonal). Grey boxes represent significant population differentiation at a *p*-value cutoff of 0.05.

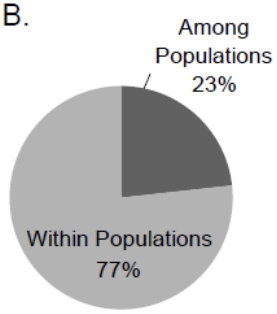
A.



C.

	Leeuwarden	Alkmaar	Haarlem	Amsterdam	Enschede	Utrecht	Arnhem	Rotterdam	Nijmegen	Goes	Veldhoven	Heerlen	
Leeuwarden		0.123	0.423	0.487	0.331	0.484	0.422	0.437	0.320	0.501	0.522	0.480	**Leeuwarden**
Alkmaar	0.061		0.036	0.121	0.482	0.068	0.250	0.038	0.053	0.116	0.162	0.300	**Alkmaar**
Haarlem	0.008	0.056		0.451	0.432	0.459	0.233	0.455	0.453	0.391	0.350	0.467	**Haarlem**
Amsterdam	0.000	0.032	0.000		0.448	0.442	0.224	0.436	0.406	0.406	0.426	0.470	**Amsterdam**
Enschede	0.025	0.000	0.000	0.000		0.451	0.188	0.478	0.460	0.434	0.442	0.455	**Enschede**
Utrecht	0.003	0.042	0.000	0.000	0.000		0.301	0.459	0.217	0.189	0.487	0.494	**Utrecht**
Arnhem	0.009	0.036	0.021	0.039	0.050	0.017		0.239	0.058	0.020	0.026	0.412	**Arnhem**
Rotterdam	0.000	0.030	0.000	0.000	0.000	0.000	0.021		0.119	0.185	0.274	0.470	**Rotterdam**
Nijmegen	0.012	0.026	0.000	0.001	0.000	0.012	0.069	0.007		0.458	0.442	0.333	**Nijmegen**
Goes	0.004	0.047	0.005	0.004	0.000	0.028	0.153	0.018	0.000		0.411	0.378	**Goes**
Veldhoven	0.007	0.050	0.000	0.000	0.000	0.001	0.126	0.012	0.000	0.000		0.261	**Veldhoven**
Heerlen	0.000	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.011	0.035		**Heerlen**
	Leeuwarden	**Alkmaar**	**Haarlem**	**Amsterdam**	**Enschede**	**Utrecht**	**Arnhem**	**Rotterdam**	**Nijmegen**	**Goes**	**Veldhoven**	**Heerlen**	

B.



D.

	Pop1	Pop2	Pop3	Pop4	Pop5	
Pop1		0.001	0.001	0.001	0.001	**Pop1**
Pop2	0.356		0.001	0.001	0.001	**Pop2**
Pop3	0.207	0.387		0.001	0.001	**Pop3**
Pop4	0.113	0.369	0.240		0.001	**Pop4**
Pop5	0.172	0.387	0.196	0.189		**Pop5**
	Pop1	**Pop2**	**Pop3**	**Pop4**	**Pop5**	

CHAPTER IV

THE EVOLUTIONARY IMPRINT OF DOMESTICATION ON GENOME VARIATION AND FUNCTION OF THE FILAMENTOUS FUNGUS *ASPERGILLUS ORYZAE*

John G. Gibbons[1], Leonidas Salichos[1], Jason C. Slot[1], David C. Rinker[1,2], Kriston L. McGary[1], Jonas G. King[1], Maren A. Klich[3], David L. Tabb[4], W. Hayes McDonald[5] and Antonis Rokas[1,2]

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA*

[2]*Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA*

[3]*USDA, ARS, Southern Regional Research Center, New Orleans, LA, USA*

[4]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA*

[5]*Department of Biochemistry, Vanderbilt University Medical Center, Nashville, TN, USA*

This chapter is published in *Current Biology*, August 7, 2012, 22: 1-7.

ABSTRACT

The domestication of animals, plants and microbes fundamentally transformed the lifestyle and demography of the human species (Diamond 2002). Although the genetic and functional underpinnings of animal and plant domestication are well understood, little is known about microbe domestication (Legras, Merdinoglu et al. 2007; Liti, Carter et al. 2009; Rokas 2009; Hyma, Saerens et al. 2011; Libkind, Hittinger et al. 2011). We systematically examined genome-wide sequence and functional variation between the domesticated fungus *Aspergillus oryzae*, whose saccharification abilities humans have harnessed for thousands of years to produce sake, soy sauce and miso from starch-rich grains, and its wild relative *A. flavus*, a potentially toxigenic plant and animal pathogen (Machida, Asai et al. 2005). We discovered dramatic changes in the sequence variation and abundance profiles of genes and wholesale primary and secondary metabolic pathways between domesticated and wild relative isolates during growth on rice. Through selection by humans, our data suggest that an atoxigenic lineage of *A. flavus* gradually evolved into a "cell factory" for enzymes and metabolites involved in the saccharification process. These results suggest that whereas animal and plant domestication was largely driven by Neolithic "genetic tinkering" of developmental pathways, microbe domestication was driven by extensive remodeling of metabolism.

INTRODUCTION

Examination of several plants and animals suggests that domestication was driven by genetic changes in diverse developmental pathways that ultimately led to large fruits, naked grains, small brains and big bodies (Diamond 2002; Doebley, Gaut et al. 2006; Purugganan and Fuller 2009). Although the molecular genetics and phenotypic outcomes of crop and livestock domestication have been extensively studied (Andersson and Georges 2004; Doebley, Gaut et al. 2006; Purugganan and Fuller 2009), the evolutionary paths traversed by domesticated microbes remain poorly understood (Legras, Merdinoglu et al. 2007; Liti, Carter et al. 2009; Rokas 2009; Hyma, Saerens et al. 2011; Libkind, Hittinger et al. 2011). In China, evidence for a fermented beverage based on rice mixed with honey and fruit dates back to 7,000 B.C. (McGovern, Zhang et al. 2004). Over the millennia that followed, the gradual development of the saccharification process, in which filamentous fungi break down the starch-rich rice to sugars that yeast ferments, morphed the beverage into the high-alcohol rice wine known as sake (Teramoto, Hano et al. 2000; McGovern, Zhang et al. 2004; Abe, Gomi et al. 2006; Kobayashi, Abe et al. 2007; Machida, Yamada et al. 2008). The filamentous fungus used in saccharification for making sake, as well as other traditional Japanese products such as soy sauce and miso, is *Aspergillus oryzae* (class Eurotiomycetes, phylum Ascomycota). For sake making, *A. oryzae* spores (koji-kin) are first spread onto steamed rice. After a ~ 2-day growth period, the resulting *A. oryzae*-rice (koji) is mixed with additional steamed rice and water and fermented by *Saccharomyces cerevisiae*, such that the breakdown of the rice starch by *A. oryzae* occurs in parallel with the conversion of sugars to alcohol by *S. cerevisiae* (Yoshizawa 1999). However, the saccharific and more generally proteolytic and

metabolic, activities of *A. oryzae* do not only fuel the yeast, but they also contribute metabolites that influence the flavor and aroma of sake (Yoshizawa 1999).

A. oryzae is closely related to the wild species *A. flavus* (Kurtzman, Smiley et al. 1986; Geiser, Pitt et al. 1998), the two species sharing 99.5% genome-wide nucleotide similarity (Rokas, Payne et al. 2007). However, *A. oryzae* is an atoxigenic domesticate recognized by the U.S. Department of Agriculture as a Generally Regarded As Safe (GRAS) organism (Machida, Asai et al. 2005), whereas *A. flavus* is a destructive agricultural pest of several seed crops and producer of the potent natural carcinogen aflatoxin (Murakami, Takase et al. 1967). This striking contrast between genomic and phenotypic variation makes the *A. oryzae – A. flavus* lineage an excellent microbe domestication model for the study of the functional changes associated with microbe domestication and the impact of the process on genome variation (Machida, Asai et al. 2005; Rokas 2009; Hunter, Jin et al. 2011; Kato, Tokuoka et al. 2011).

MATERIALS AND METHODS

Isolate Selection

Our collection consisted of eight *A. oryzae* and eight *A. flavus* isolates chosen to represent a diversity of industrial uses, ecologies and geographies (Table 4.1). Isolates were genotyped across 16 microsatellite markers to verify strains were not clonally related.

Table 4.1. List of *A. oryzae and A. flavus* isolates analyzed

Species	Isolate	Source	Country of Origin
Aspergillus oryzae	SRRC 302	Sake	Japan
	RIB 331	Miso	Japan
	RIB 333	Miso	Japan
	RIB 537	Sake	Japan
	RIB 632	Sake	Japan
	RIB 642	Sake	Japan
	RIB 949	Shoyu	Japan
	RIB 40 (reference)	Sake	Japan
Aspergillus flavus	SRRC 1273	Soil	USA (Arizona)
	SRRC 1357	Dried bacon	Croatia
	SRRC 2112	Hazelnut	Turkey
	SRRC 2114	Wheat	USA
	SRRC 2524	Dead termites	China
	SRRC 2632	Blood	USA (Chicago, Illinois)
	SRRC 2653	Corneal ulcer	USA (Miami, Florida)
	NRRL 3357 (reference)	Peanut	USA

Illumina Library Sample Preparation

For genomic DNA (gDNA) library preparation, isolates were grown in potato dextrose broth in a tissue rotator for 3-4 days at room temperature. Mycelium was then dried under a vacuum and the tissue was ground with a mortar and pestle in liquid nitrogen. gDNA was extracted using the DNeasy Plant Maxi Kit (Qiagen) implementing an RNase step according to the manufacturer's instructions. gDNA libraries were prepared as described previously (Hittinger, Goncalves et al. 2010).

For mRNA library preparation, we used three sake-derived isolates of *A. oryzae* (RIB 632, RIB 642 and RIB 40) belonging to different clades (Figure 4.1A) and three *A. flavus* isolates (SRRC 1357, SRRC 2524 and NRRL 3357) from different clades. Isolates were grown on rice to model the industrial setting of sake production. Akitakomachi short grain rice (Lundberg Family Farms) was rinsed, soaked in water for 30 minutes then cooked in an autoclave. The cooked rice was plated in petri dishes (100 mm x 15 mm polystyrene), flattened with a sterilized spatula and covered with a layer of sterile porous cellophane. 500 µl of a water conidial suspension (10^7/ml) was then spread onto the plate and incubated in the dark at 30˚C for 24 hours. Mycelium was harvested with a metal spatula, flash frozen in liquid nitrogen and stored at -80˚C. Mycelium was ground with a mortar and pestle in liquid nitrogen. Total RNA was extracted using TRIzol (Life Technologies), DNased then cleaned with an RNeasy column (Qiagen) according to the manufacturer's instructions. Total RNA integrity was quality controlled via Bioanalyzer (Agilent Technologies). mRNA libraries were prepared as previously described (Gibbons, Janson et al. 2009) with the exceptions that (i) mRNA was isolated using

Seradyn Sera-Mag magnetic Oligo(dT) beads (Thermo Scientific) and (ii) 12 cycles of PCR were used for template enrichment.

MudPIT Proteomics Sample Preparation and Peptide Identification

Two biological replicates of *A. oryzae* RIB 40 and of *A. flavus* NRRL 3357 were grown on rice as described above. Mycelium was harvested with a metal spatula and ground with a mortar and pestle in liquid nitrogen. Each of the four samples was then denatured with 8M urea 100 mM tris pH 8.5, reduced with TCEP, and alkylated with iodoacetamide prior to dilution to 2 M urea and overnight digestion with porcine trypsin. The resulting peptides from 0.4 mg of cellular material were subjected to MudPIT essentially as described previously (MacCoss, McDonald et al. 2002) except that only two ammonium acetate steps were utilized, 200 mM and 1000 mM. Data dependent nano LC-MS/MS data acquisition was conducted on a Thermo LTQ instrument equipped with an Eksigent NanoLC-AS1 Autosampler 2.08, an Eksigent NanoLC-1D plus HPLC pump, and Nanospray source. Collections averaged 15,613 tandem mass spectra per reversed phase gradient step.

LC-MS/MS data were translated to mz5 format (Wilhelm, Kirchner et al. 2012) by the ProteoWizard msConvert utility (Kessner, Chambers et al. 2008). Spectra were identified against protein sequence databases translated from open reading frames from the Broad Institute genomes of the two species. Data for *A. oryzae* were identified against the 12,134 protein sequences of that species, and data for *A. flavus* were identified against the 12,658 protein sequences from that species. Each protein was considered in both forward

and reversed orientation, with the latter serving as a decoy sequences for establishing FDR values. MyriMatch 1.6 (Tabb, Fernando et al. 2007) conducted the semi-tryptic peptide comparison, assuming that all Cysteines were carbamidomethylated and allowing for the oxidation of Methionine and the loss of ammonia from N-terminal Glutamines. Precursor ions were allowed to differ from expected average masses by 1.5 m/z, while fragments were required to fall within 0.5 m/z of expected monoisotopic values. Protein assembly took place in IDPicker 2.6 (Ma, Dasari et al. 2009), which applied a 0.05 FDR to filter the peptide-spectrum matches. Subsequently, proteins were required to match two distinct peptides to be included, and parsimony rules removed subset and subsumable proteins. Empirically, the protein FDR for *A. oryzae* was estimated at 1.3%, while the FDR for *A. flavus* was 4.5%.

Illumina Data Processing

Sequence quality of the raw read sets were assessed using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Reads were trimmed at both the 5' and 3' ends using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) so that the first and last position had an average quality score ≥ 20.

Read Mapping, Consensus Sequence Generation and SNP Calling

Genomic DNA reads in FASTQ format were mapped to the *A. oryzae* RIB 40 reference genome using the Mapping and Assembly with Qualities (Maq) software package v0.5.0 (Li, Ruan et al. 2008), allowing 3 mismatches per read. For each sample, we generated

and extracted consensus sequences based on the *A. oryzae* RIB 40 reference genome using the "assemble" and "cns2fq" commands. The resulting FASTQ consensus sequence was converted to FASTA format and all ambiguously called bases (non A, T, C or G) were converted to 'N'. For each sample, we extracted a list of SNPs using the "cns2snp" command and further filtered the list by retaining only SNP-containing sites that had at least 5 reads covering the position and an average quality score ≥ 20, using the "maq.pl SNPfilter". We also performed the same analysis using *A. flavus* NRRL 3357 as the reference.

Identification of *A. oryzae* and *A. flavus* Unique Genomic Regions

We identified regions of the reference genomes that were unique to *A. oryzae* and *A. flavus* by analyzing the consensus genome sequences. Bases which were unmapped were coded ambiguously as 'N'. We then extracted the positions in the genome which had Ns in all isolates of one species and all "non-Ns" in all isolates of the other. Regions separated by ≤ 25 bp were merged and only regions ≥ 250 bp were further analyzed for gene content. We performed 4 comparisons: (i) all *A. oryzae* isolates vs. all *A. flavus* isolates and (ii) all *A. flavus* isolates vs. all *A. oryzae* isolates.

Sequence Analysis of Aflatoxin Locus

To investigate the sequence variation at the aflatoxin locus, we independently BLASTed (REF) the gDNA sequence reads from the subset of isolates for which we had expression data (*A. oryzae* RIB 632, *A. oryzae* RIB 642, *A. flavus* SRRC 1357 and *A. flavus* SRRC 2524) against the both the *A. oryzae* RIB 40 and *A. flavus* NRRL 3357 reference aflatoxin

loci, implementing an E-value cutoff of 1E-6. Reads with significant hits were extracted and assembled with Velvet v1.2.03 (Zerbino and Birney 2008) using the Multiple-*k* method (Surget-Groba and Montoya-Burgos 2010). Assembled contigs were then BLASTed against the loci of interest and inspected for mutations.

Phylogenetic Analysis

We assessed the evolutionary relationship of our isolates using the alignment of our 100,084 "high quality SNPs" which we defined as SNP sites where there were no ambiguous base pairs in any of the 16 strains. We constructed our phylogeny using the maximum parsimony optimality criterion with 1,000 bootstraps replicates in MEGA v5.05 (Tamura, Peterson et al. 2011).

In order to investigate the evolution of 6-gene cluster and 9-gene cluster genes, we retrieved amino acid sequences from a local database of 103 complete and draft fungal proteomes (see (Campbell, Rokas et al. 2012)) using BlastP (Altschul, Gish et al. 1990), retaining hits with >45% similarity that were between 50 and 150% the length of the query. We aligned the sequences with mafft-6.847 under default settings (Katoh and Toh 2008), manually removed divergent taxa and realigned with mafft. We removed alignment columns containing greater than 40% gaps using TrimAL v3 (Capella-Gutierrez, Silla-Martinez et al. 2009), and performed Maximum Likelihood (ML) analysis and 100 ML bootstrap replicates under the PROTGAMMAJTT model of amino acid substitution using RAxML v7.2.6 (Stamatakis 2006).

Population Structure Analysis

We examined the genetic differentiation of our isolates using the STRUCTURE software (Pritchard, Stephens et al. 2000) as previously described (Klaassen, Gibbons et al. 2012). Briefly, we first identified the optimal number of populations (*K*) using a subset of 1,000 randomly selected SNP markers (Evanno, Regnaut et al. 2005), then inferred the population assignments of each isolate using our entire "high quality SNP" set.

Detection of Recent Positive Selection

We identified regions of the *A. oryzae* genome that exhibit a reduction in genetic variation, indicative of recent positive selection. Using the aligned consensus genome sequences, we calculated nucleotide diversity (Θ) for each species in 5 Kb windows with a 500 bp step size using the VariScan software (Vilella, Blanco-Garcia et al. 2005; Hutter, Vilella et al. 2006). For windows having no segregating sites, we normalized Θ as the average of Θ values of all windows having $\geq$ 4,000 sites with one segregating site. Windows which had $\geq$ 2,000 sites where more than two nucleotides were ambiguous (Ns) in either species were removed. This quality filter resulted in the inclusion of 65,894 of 75,629 windows (87%). For each window, we then calculated relative nucleotide diversity as follows:

$$\Theta_{OF} = \log_2(\Theta_{A.\ oryzae} / \Theta_{A.\ flavus})$$

We defined candidates as windows which fell in the 0.25% quantile of Θ_{OF} distribution. Candidate windows which overlapped were collapsed into larger PSSRs.

Gene Expression Quantification, Differential Expression and Differential Protein Abundance

Gene expression levels were quantified as previously described (Gibbons, Beauvais et al. 2012) by mapping reads against the *A. oryzae* RIB 40 reference transcriptome. Differentially expressed genes were identified by comparing the proportion of reads that mapped to each gene for *A. oryzae* and *A. flavus* via Fisher's exact tests with a multiple test-corrected *P* value cutoff of 4.14e-6. We also identified a species-level up-regulated gene set in *A. oryzae* as genes where all isolates were expressed ≥ 10 RPKM and up-regulated by at least 1.5-fold vs. all *A. flavus* isolates. We performed the identical analysis between the atoxigenic (*A. oryzae* RIB 40, RIB 632, RIB 642 and *A. flavus* SRRC 1357) and toxigenic (*A. flavus* SRRC 2524 and NRRL 3357) clades. To identify differential protein abundance between *A. oryzae* RIB 40 and *A. flavus* NRRL 3357 we pooled the spectral mapping results of two biological replicates per species and compared the proportion of spectra that mapped to each protein for *A. oryzae* and *A. flavus* using Fisher's exact tests implementing a *P* value cutoff of 0.01.

Functional Associations of Gene Sets

To examine whether particular sets of genes were preferentially associated with certain functions, we compared the proportion of genes in a given gene set belonging to the 2nd-order FunCat categories (Ruepp, Zollner et al. 2004) and SMURF predicted secondary metabolism associated genes (Khaldi, Seifuddin et al. 2010) to their corresponding values in the remainder of the genome using Fisher’s exact tests applying a multiple test-

corrected *P* value cutoff of 0.0006. For functional associations of differentially expressed genes, we used the set of genes that were statistically differentially expressed between *A. oryzae* RIB 40 and *A. flavus* NRRL 3357 (1,397 up-regulated genes and 1,284 down-regulated genes). Additional functional information for genes was gathered from the DBD transcription factor prediction database (Wilson, Charoensawan et al. 2008), the MEROPS peptidase database (Rawlings, Barrett et al. 2012), the PFAM protein families database (Punta, Coggill et al. 2012), and the TansportDB membrane transporter database (Ren, Chen et al. 2007).

Data Availability

Raw Illumina sequence reads were submitted to the NCBI Sequence Read Archive (SRA) (Accession Numbers: *A. oryzae* gDNA: SRA0502658, *A. flavus* gDNA: SRA052664, *A. oryzae* RNAseq, SRA0502666 and *A. flavus* RNAseq: SRA052667). Raw proteomics data was submitted to Tranche and can be downloaded from the Vanderbilt MSRC Bioinformatics Data page: www.mc.vanderbilt.edu/msrc/bioinformatics/data.php .

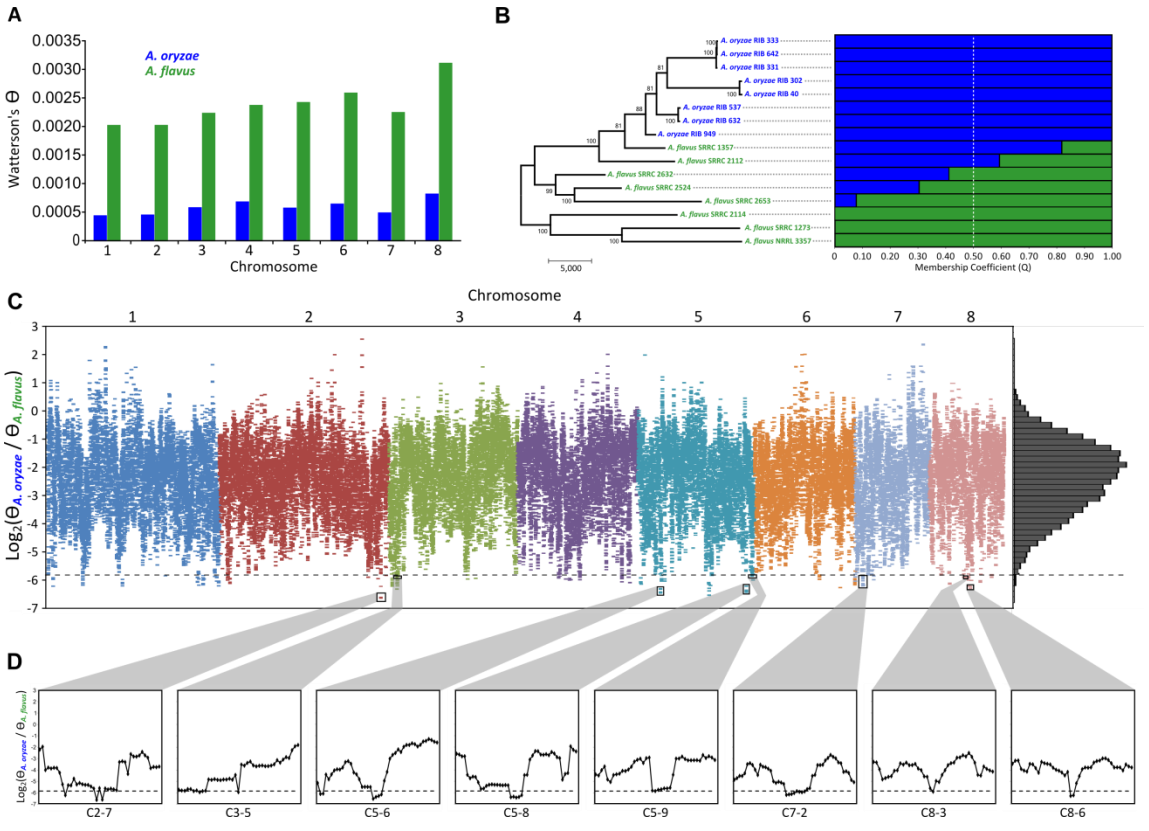
RESULTS AND DISCUSSION

Domesticated organisms have typically been selected for beneficial traits conferred by certain genetic loci and have undergone several rounds of population bottlenecks. Although we previously did not find evidence that the *A. oryzae* genome exhibited a relaxation of selective constraints, a common characteristic accompanying plant an animal domestication (Rokas 2009), whether the *A. oryzae* genome has experienced positive selection during the domestication process remains an open question. To address this question, we Illumina sequenced 14 geographically and industrially diverse isolates from *A. oryzae* and *A. flavus* and jointly analyzed them with the two species' reference genomes (Machida, Asai et al. 2005; Payne, Nierman et al. 2006) (*A. oryzae* RIB 40 and *A. flavus* NRRL 3357; Tables S4.1). Analysis of the genome-wide nucleotide diversity across the 16 isolates showed that the genetic diversity of the *A. oryzae* isolates is ~25% of that found in the *A. flavus* isolates (chromosome average nucleotide variation $\Theta_{A.\ oryzae}$ = 0.0006 versus $\Theta_{A.\ flavus}$ = 0.0024; T-test, P = 4.1e-7), consistent with previous gene-level estimates (Geiser, Pitt et al. 1998; Geiser, Dorner et al. 2000; Chang and Ehrlich 2010) (Figure 4.1A). Evolutionary analysis of 100,084 high quality SNPs suggested that the *A. oryzae* isolates are monophyletic, in agreement with the previous hypotheses that *A. oryzae* originated via a single domestication event (Geiser, Pitt et al. 1998; Geiser, Dorner et al. 2000), and do not group by geography or ecology (Figure 4.1B). Interestingly, two *A. flavus* isolates (SRRC 1357 and SRRC 2112) show closer affinity to *A. oryzae* than to other *A. flavus* isolates (Figure 4.1B), suggesting that *A. oryzae* originated from within *A. flavus*.

One of the footprints of recent selection on the genome is the reduction in variation of regions that are close to the variants under selection (Sabeti, Schaffner et al. 2006). When a beneficial allele is rapidly driven toward fixation, nearby neutral variants are likely to also become fixed as a result of the low rate of recombination between closely linked sites (Smith and Haigh 1974). By estimating the relative genome-wide nucleotide diversity $\Theta_{OF} = \log_2 (\Theta_{A.\ oryzae} / \Theta_{A.\ flavus})$ we identified 61 putative selective sweep regions (PSSRs) (Figure 4.1C, D). Examination of PSSR gene content indicates that the main targets of selection were genes and pathways involved in primary metabolism (PM) and secondary metabolism (SM). For example, the 148 PSSR genes were significantly overrepresented for SM (Fisher's Exact Test (FET), $P = 0.0004$), whereas five PSSRs contained SM gene clusters, including one for the biosynthesis of the tremorgenic mycotoxin aflatrem (PSSR C5-9; Figure 4.1C,D) (Nicholson, Koulman et al. 2009). These results were particularly noteworthy as SM gene families are thought to have expanded and be located in unique genomic regions of the *A. oryzae - A. flavus* lineage compared to the far more distantly related species *A. fumigatus* and *A. nidulans* (Machida, Asai et al. 2005). Furthermore, several PSSR genes are involved in protein and peptide degradation (genes in PSSRs C2-7 and C5-8) and carbohydrate metabolism (C3-5, C5-6) (Figure 4.1 C, D). One of the strongest supported PSSRs (C8-6) contained a glutaminase (Figure 4.1 C, D), which catalyzes the hydrolysis of carbon-nitrogen bonds of L-glutamine to glutamic acid, a widely used food flavor enhancer found at considerable levels in sake (Fujisawa and Yoshino 1998). Strikingly, whereas there are six polymorphic sites within the *A. oryzae* isolates (2 promoter, 4 intron), *A. flavus*

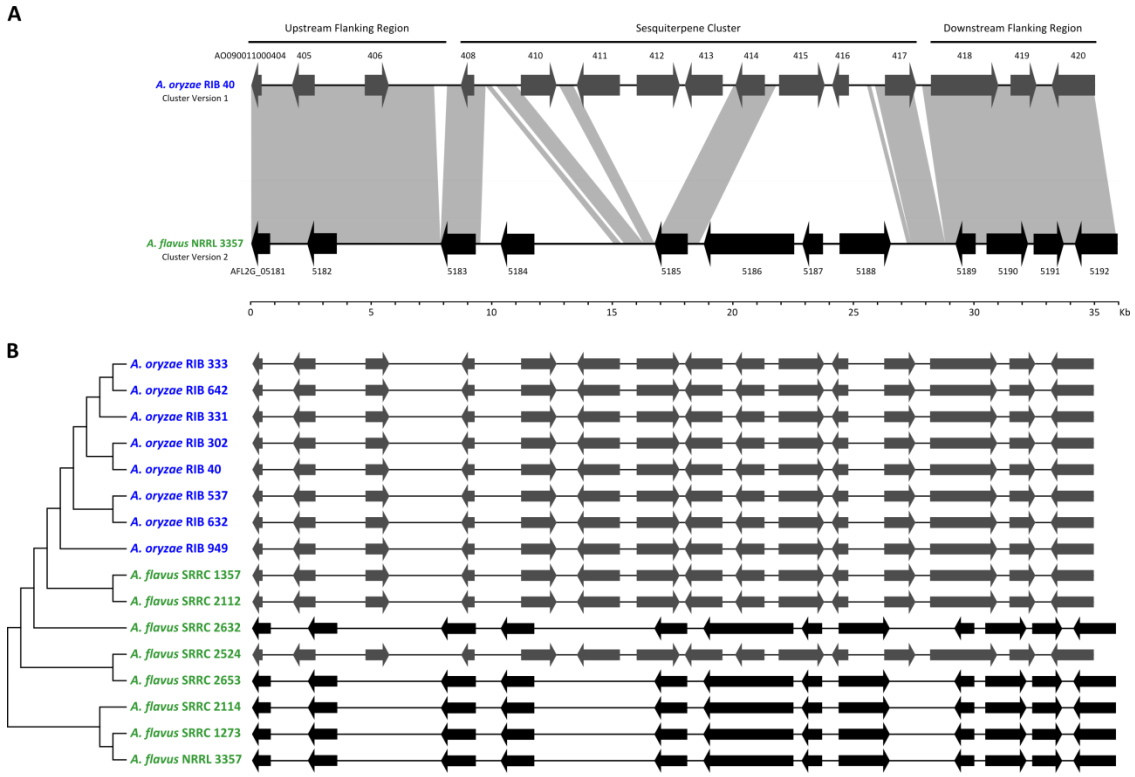
isolates are polymorphic at 86 sites (14 synonymous, 2 nonsynonymous, 18 promoter region and 52 intron) (Figure S4.1).

Figure 4.1. **Phylogenetic relationship and genomic patterns of variation in *A. oryzae* and *A. flavus*.** (A) Chromosomal levels of nucleotide variation (Θ) in A. oryzae (blue) and A. flavus (green). (B) Left panel: Parsimony-inferred phylogeny of the 16 A. oryzae (blue) and A. flavus (green) isolates from the 100,084 high quality genome-wide variant sites. Values near internodes indicate bootstrap support, generated by 1,000 replicates. The scale bar represents the number of changes. Right panel: STRUCTURE-based membership coefficient for each isolate (population number K = 2). The A. oryzae and A. flavus genetic backgrounds are shown in blue and green, respectively. (C) Relative nucleotide diversity scores (ΘOF) for 5-kb windows (65,894 windows) with a 500 bp step size scanning the eight chromosomes. Points below the dotted line represent genomic regions below the empirical 0.25% quantile (164 windows) and comprise the candidate Putative Selective Sweep Regions (PSSRs). The right panel shows the distribution of ΘOF scores. (D) Close-ups of representative PSSRs and flanking regions.



We also examined the isolate genome data to identify differences in genome architecture between the two species. Although our search identified only five genes shared uniquely by all *A. oryzae* isolates and none by *A. flavus* isolates, it did also identify a locus that contains a 9-gene cluster in the *A. oryzae* genome, but contains a 6-gene cluster in the *A. flavus* NRRL 3357 genome (Figure 4.2A). Interestingly, the 9-gene cluster is very similar to the sesquiterpene gene cluster in *Trichoderma virens* (Mukherjee, Horwitz et al. 2006; Mukherjee, Horwitz et al. 2011), whose product belongs to a class of food flavoring aromatic compounds (Janssens, Depooter et al. 1992), whereas the 6-gene cluster comprises of a terpene cyclase and GAPDH from the 9-gene cluster together with four other unrelated genes (Figure S4.2). Remarkably, although *A. oryzae* is fixed for the 9-gene cluster, *A. flavus* is polymorphic; three isolates contain the 9-gene cluster, while the other five contain the alternative 6-gene cluster (Figure 4.2B). Furthermore, the genes contained in the two alternative cluster "alleles" at this locus have different evolutionary histories (Figure S4.2). Most unique genes of the 9-gene cluster group with sequences from *A. clavatus* and very divergent fungi related to *T. virens*, consistent with horizontal transfer, whereas most *A. flavus* unique genes of the alternative cluster group with sequences from *A. aculeatus*, suggesting a very different history.

Figure 4.2. The variable genome architecture of the sesquiterpene cluster locus. (A) Microsynteny of the locus harboring the sesquiterpene encoding gene cluster and its flanking regions in *A. oryzae* RIB 40 and *A. flavus* NRRL 3357 isolates. Gray blocks represent genomic regions exhibiting significant sequence similarity between species. Genes, and the direction of transcription, are symbolized by arrows and labeled. The *A. oryzae* RIB 40 genme contains a 9-gene cluster "allele", whereas the *A. flavus* NRRL 3357 genome contains a 6-gene cluster "allele". Only the terpene cyclase (AO090011000408) and the GAPDH (AO090011000414), as well as a few non-coding regions are homologous between the two "alleles". (B) A graph showing the allele present in each of the 16 isolates. Note that all eight *A. oryzae* contain the 9-gene cluster "allele", whereas *A. flavus* is polymorphic.

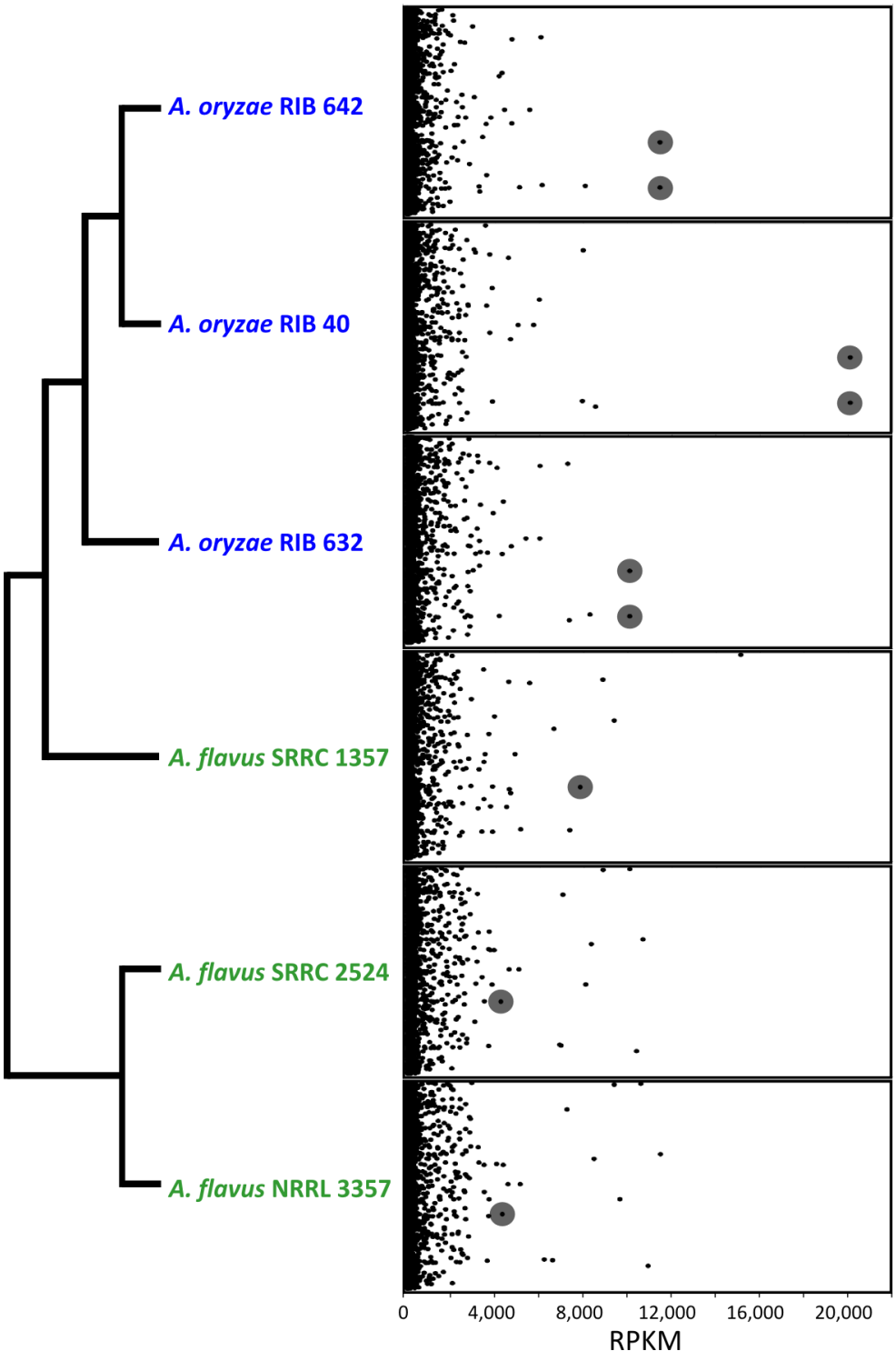


A. oryzae has been grown continually on starch-rich grains, such as rice and soy, for thousands of years (Machida, Asai et al. 2005; Machida, Yamada et al. 2008). To identify functional differences and putative adaptations to this starch-rich diet, we examined the transcriptome profiles of three phylogenetically distinct isolates of sake-derived *A. oryzae*, as well as the proteome profiles of the reference isolate of each species, during growth on rice. Similar to the analyses of the PSSR gene content, comparison of the

transcriptome and proteome profiles between *A. oryzae* and *A. flavus* identified several differentially abundant transcripts, proteins and pathways involved in PM and SM.

All *A. oryzae* isolates possess two or three copies of α-amylase (Machida, Asai et al. 2005; Hunter, Jin et al. 2011), the enzyme that hydrolyzes the α-D-glycosidic bonds of starch to produce dextrin, compared to a single copy in *A. flavus*. We found that the transcript and protein abundance of α-amylase was the highest of any *A. oryzae* gene or protein and was significantly up-regulated compared to *A. flavus* (gene expression: FET; P < 1e-300 and protein abundance: >30-fold, FET, P = 2.15e-51) (Figure 4.3). Several other *A. oryzae* up-regulated genes are involved in carbohydrate PM, including the genome neighbors amylolytic transcriptional activator *amyR* (Gomi, Akeno et al. 2000) (FET; P = 1.68e-97) and saccharide metabolizing enzyme maltase glucoamylase (FET; P = 1.79e-17), as well as the glucose metabolizing enzyme sorbitol dehydrogenase (FET; P = 8.22e-252) (Figures S4.3 and S4.4). Importantly, comparison of the transcriptional profile of the two species showed that both the up-regulated and down-regulated gene sets in *A. oryzae* were overrepresented for carbohydrate PM (FET; P = 6.24e-5 and P = 4.22e-12, respectively), suggesting that differential regulation of PM is a key functional difference between the two species.

Figure 4.3. α-amylase is the most highly expressed transcript in *A. oryzae*. Expression levels (RPKM) (X-axis) of all genes (Y-axis) for each of the six isolates organized by their phylogenetic relatedness. The two α-amylase paralogs are highlighted in gray. Expression levels for the two paralogs are depicted as equal because they have identical coding sequences and differentiation of their expression levels is not possible.

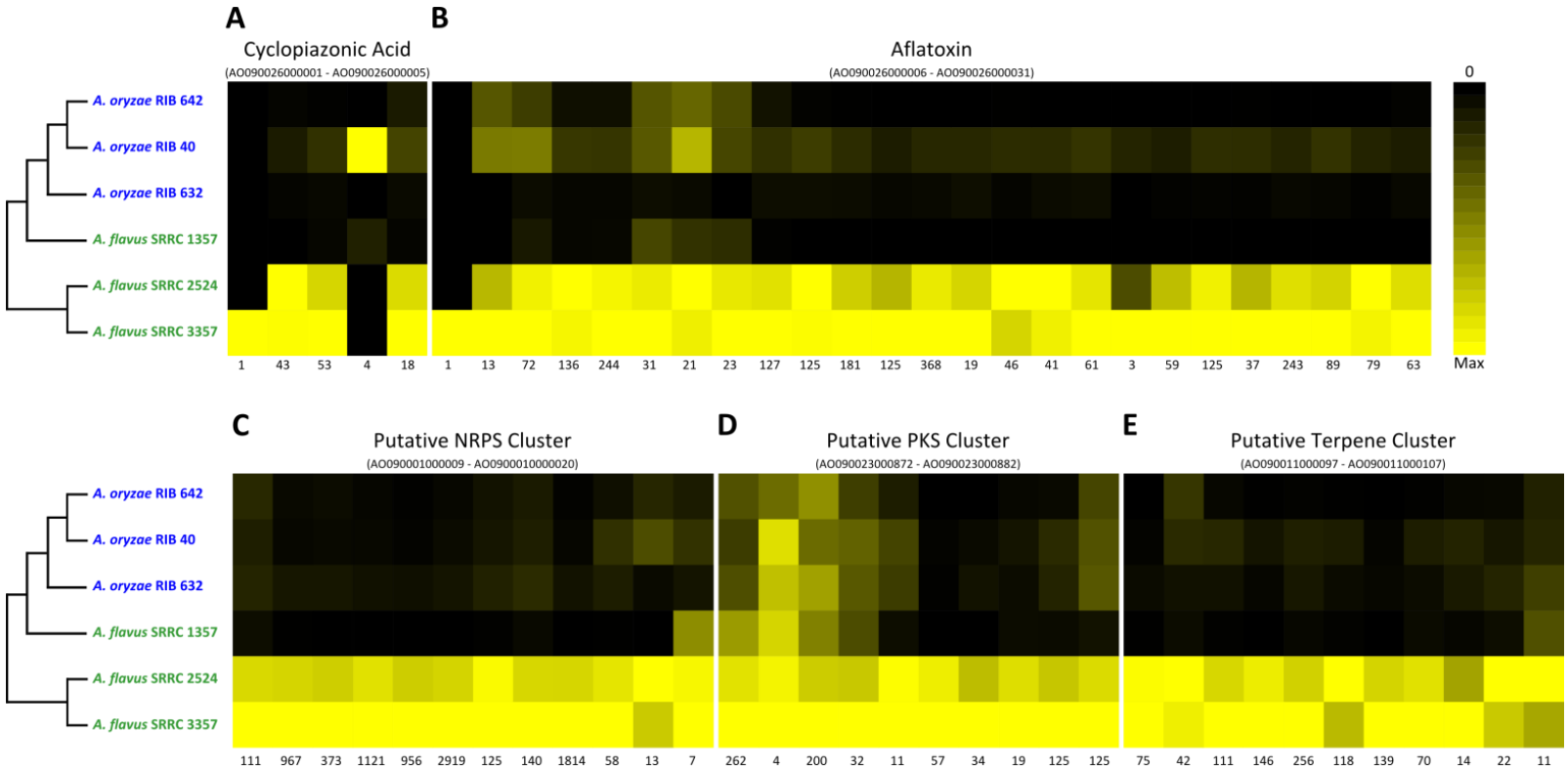


A. oryzae is also equipped with an arsenal of secreted enzymes that break down the proteins and complex polysaccharides of the grain outer layers, providing access to the starch-rich interior layers (Yoshizawa 1999; Kitamoto 2002; Machida, Asai et al. 2005; Kobayashi, Abe et al. 2007). Several protease-encoding genes are located in PSSRs (e.g., the methionine aminopeptidase located in the PSSR C5-8), or are up-regulated (e.g., extracellular cellulase *celA*), or both (e.g., the up-regulated proteinase located in PSSR C2-7) (Figure 4.1 C, D). In contrast, 16 of the 27 plant polysaccharide degrading genes were down-regulated (a few of them are also located in PSSRs, e.g., endoglucanase and feruloyl esterase in PSS C3-5 and endo-1,4-β-xylanase in PSS C5-6). The broad down-regulation of this subset of genes likely reflects differences between *A. oryzae* and *A. flavus*.

Comparison of the gene expression profiles of 610 genes in all 55 predicted SM gene clusters (Khaldi, Seifuddin et al. 2010) against background genes in the two species showed that another general characteristic of the *A. oryzae* transcriptome during growth on rice is SM down-regulation (FET, P = 7.3e-10). This is consistent with the wholesale down-regulation of five SM gene clusters in *A. oryzae* (Figure 4.4). Importantly, both the cyclopiazonic acid and the aflatoxin SM pathway in *A. oryzae* were down-regulated (Figure 4.4A, B), explaining a key phenotypic difference between *A. oryzae* and *A. flavus*, which is the inability of the first to produce either of the two toxins (Machida, Asai et al. 2005; Kato, Tokuoka et al. 2011; Rank, Klejnstrup et al. 2012). We further investigated sequence variation in the isolates with expression data with respect to five

previously characterized types of mutations observed at the aflatoxin gene cluster locus: (*i*) transcription binding site mutations in the *aflR* promoter (Tominaga, Lee et al. 2006), (*ii*) a ~250 bp 3' deletion in the *aflT* coding region (Tominaga, Lee et al. 2006), (iii) a frameshift mutation in the *norA* coding region (Tominaga, Lee et al. 2006), (*iv*) multiple nonsynonymous mutations in the *verA* coding region (Tominaga, Lee et al. 2006), and (*v*) ~40 Kb deletion from *norB* to *norA* genes (Chang, Horn et al. 2005). This analysis revealed mutation *v* in *A. oryzae* RIB 632 and mutations *i – iv* in *A. oryzae* RIB 632 and RIB 40 (Tominaga, Lee et al. 2006) when compared to *A. flavus* NRRL 3357. Furthermore, the *A. oryzae*-like isolate *A. flavus* SRRC 1357, contained 5 and 13 nonsynonymous mutations in the *aflT* (*ii*) and *verA* (*iv*) genes respectively, while *A. flavus* SRRC 2524 was nearly identical to *A. flavus* NRRL 3357 (3 and 1 synonymous mutations in the *norA* (*iii*) and *verA* (*iv*) genes). Interestingly, aflatoxin is genotoxic to *S. cerevisiae* (Keller-Seitz, Certa et al. 2004), suggesting that the atoxicity of *A. oryzae* might have been driven by its impact on yeast survival and, as a consequence, fermentation for making sake.

Figure 4.4. The *A. oryzae* secondary metabolism transcriptome is widely down-regulated during growth on rice. Expression levels of five down-regulated secondary metabolism biosynthesis gene clusters for the six isolates for: (A) cyclopiazonic acid, (B) aflatoxin, (C) putative nonribosomal peptide metabolite, (D) putative polyketide synthase metabolite, and (E) putative terpene. The range of genes included in each gene clusters is given under each cluster's name. For each gene, the color of the heat map cell corresponds to its expression level (in RPKM units), where black is zero expression and yellow is the maximum RPKM for that gene (listed below each gene).



A. flavus natural isolates show substantial variation in SM production and several are known to be atoxigenic (Murakami, Takase et al. 1967; Kusumoto, Nogata et al. 2000; Chang, Horn et al. 2005; Georgianna, Fedorova et al. 2010; Amaike and Keller 2011; Kato, Tokuoka et al. 2011; Rank, Klejnstrup et al. 2012). Interestingly, the SM expression profile of the atoxigenic *A. flavus* SRRC 1357, the isolate most closely related to *A. oryzae* (Figure 4.1A), was more similar to *A. oryzae* than to those of the other *A. flavus* isolates (Figure 4.4A-D), consistent with the hypothesis that *A. oryzae* was domesticated from an atoxigenic clade of *A. flavus*.

During malt rice (koji) production *A. oryzae* also produces a variety of aromatic, flavor-producing volatile compounds and associated enzymes (Ito, Yoshida et al. 1990; Yoshizawa 1999; Yoshizaki, Yamato et al. 2010). In addition to the sequence and genome architecture differences observed in the glutaminase and sesquiterpene loci, we also detected functional differences in other industrially-associated genes. Two particularly interesting examples of up-regulated genes include a glycosyl transferase (FET; P = 1.75e-237), a member of a broad sugar modifier family involved in the making of many sweeteners (Scott 1989), and an asparaginase (gene expression: FET; P = 1.29e-15 and protein abundance: FET, P = 0.006), an enzyme used commercially to reduce acrylamide levels in starch rich foods, such as rice (Friedman 2003). Surprisingly however, of the more than 500 genes annotated as MFS or ABC transporters, only 6 were up-regulated in all *A. oryzae* isolates when compared to all *A. flavus* isolates and an additional 6 were up-regulated in the *A. oryzae* isolates and the closely related *A. flavus* isolate when compared against all other *A. flavus* isolates (Figure S4.3).

In summary, our systematic comparison of sequence, gene expression, and protein abundance variation in the *A. oryzae – A. flavus* lineage indicates that *A. oryzae* domestication was accompanied by dramatic changes in primary and secondary metabolism. In a span of a few millennia, unintentional human breeding of predominantly segregating variation present in *A. flavus* resulted, through the gradual accumulation of small (e.g., Figure 4.2 C, D and Figure S4.1) and large scale (e.g., Figures 4.2 and 4.3) genetic and functional changes (e.g., Figures 4.3 and 4.4), to the evolution of the saccharific and proteolytic *A. oryzae* "cell factory". Although alterations in metabolic pathways were also likely targets of selection during both plant and animal domestication (Aharoni, Giri et al. 2004), the majority of changes was primarily driven by modifications in developmental pathways that affect growth and form. In stark contrast, these and previous (Makarova, Slesarev et al. 2006; Makarova and Koonin 2007; Douglas and Klaenhammer 2010; Kelly, Ward et al. 2010; Borneman, Desany et al. 2011; Libkind, Hittinger et al. 2011; Bachmann, Starrenburg et al. 2012) findings argue that the molecular foundations of microbe domestication largely rested in the restructuring of metabolism.

ACKNOWLEDGMENTS AND CONTRIBUTIONS

We thank members of Rokas lab, Chris Hittinger, Shannon Beltz, David Geiser, Kathy Friedman, David Friedman, Jim Patton, Julian Hillyer, Kamya Rajaram, Abby Olena, Scott Egan, Jonathan Flowers, David McCauley, Travis Clark, Chelsea Baker, and Dr. Osamu Yamada and the National Research Institute of Brewing of Japan. JGG is funded by the Graduate Program in Biological Sciences at Vanderbilt University and the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH, NIAID: F31AI091343-01). Research in A.R.'s lab is supported by the Searle Scholars Program and the National Science Foundation (DEB-0844968).

J.G.G and A.R. designed the study; J.G.G. prepared samples and analyzed the data; L.S. developed variant site extracting software; J.C.S. carried out the evolutionary analysis of the sesquiterpene cluster; D.C.R and J.G.K contributed to the experimental design; K.L.M contributed bioinformatic support; M.A.L. provided biological samples; D.L.T. and W.H.M. generated the proteomics data. J.G.G. and A.R. lead the paper writing together with contributions from all co-authors.

REFERENCES

Abe, K., K. Gomi, et al. (2006). "Impact of Aspergillus oryzae genomics on industrial production of metabolites." Mycopathologia **162**(3): 143-153.

Aharoni, A., A. P. Giri, et al. (2004). "Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species." Plant Cell **16**(11): 3110-3131.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Amaike, S. and N. P. Keller (2011). "Aspergillus flavus." Annual Review of Phytopathology, Vol 49 **49**: 107-133.

Andersson, L. and M. Georges (2004). "Domestic-animal genomics: deciphering the genetics of complex traits." Nature Reviews Genetics **5**(3): 202-212.

Bachmann, H., M. J. C. Starrenburg, et al. (2012). "Microbial domestication signatures of Lactococcus lactis can be reproduced by experimental evolution." Genome Research **22**(1): 115-124.

Borneman, A. R., B. A. Desany, et al. (2011). "Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of Saccharomyces cerevisiae." Plos Genetics **7**(2).

Campbell, M., A. Rokas, et al. (2012). "Horizontal transfer and death of a fungal secondary metabolic gene cluster." Genome Biol Evol **in press**.

Capella-Gutierrez, S., J. M. Silla-Martinez, et al. (2009). "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses." Bioinformatics **25**(15): 1972-1973.

Chang, P. K. and K. C. Ehrlich (2010). "What does genetic diversity of Aspergillus flavus tell us about Aspergillus oryzae?" International Journal of Food Microbiology **138**(3): 189-199.

Chang, P. K., B. W. Horn, et al. (2005). "Sequence breakpoints in the aflatoxin biosynthesis gene cluster and flanking regions in nonaflatoxigenic Aspergillus flavus isolates." Fungal Genetics and Biology **42**(11): 914-923.

Diamond, J. (2002). "Evolution, consequences and future of plant and animal domestication." Nature **418**(6898): 700-707.

Doebley, J. F., B. S. Gaut, et al. (2006). "The molecular genetics of crop domestication." Cell **127**(7): 1309-21.

Douglas, G. L. and T. R. Klaenhammer (2010). "Genomic evolution of domesticated microorganisms." Annu Rev Food Sci Technol **1**: 397-414.

Evanno, G., S. Regnaut, et al. (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." Molecular Ecology **14**(8): 2611-2620.

Friedman, M. (2003). "Chemistry, biochemistry, and safety of acrylamide. A review." Journal of Agricultural and Food Chemistry **51**(16): 4504-4526.

Fujisawa, K. and M. Yoshino (1998). Formation of inosinic acid as the taste compound in the fermentation of Japanese sake. In Developments in Food Science Elsevier.

Geiser, D. M., J. W. Dorner, et al. (2000). "The phylogenetics of mycotoxin and sclerotium production in Aspergillus flavus and Aspergillus oryzae." Fungal Genetics and Biology **31**(3): 169-179.

Geiser, D. M., J. I. Pitt, et al. (1998). "Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus." Proceedings of the National Academy of Sciences of the United States of America **95**(1): 388-393.
Georgianna, D. R., N. D. Fedorova, et al. (2010). "Beyond aflatoxin: four distinct expression patterns and functional roles associated with Aspergillus flavus secondary metabolism gene clusters." Molecular Plant Pathology **11**(2): 213-226.
Gibbons, J. G., A. Beauvais, et al. (2012). "Global Transcriptome Changes Underlying Colony Growth in the Opportunistic Human Pathogen Aspergillus fumigatus." Eukaryotic Cell **11**(1): 68-78.
Gibbons, J. G., E. M. Janson, et al. (2009). "Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics." Mol Biol Evol **26**(12): 2731-44.
Gomi, K., T. Akeno, et al. (2000). "Molecular cloning and characterization of a transcriptional activator gene, amyR, involved in the amylolytic gene expression in Aspergillus oryzae." Bioscience Biotechnology and Biochemistry **64**(4): 816-827.
Hittinger, C. T., P. Goncalves, et al. (2010). "Remarkably ancient balanced polymorphisms in a multi-locus gene network." Nature **464**(7285): 54-U61.
Hunter, A. J., B. Jin, et al. (2011). "Independent duplications of alpha-amylase in different strains of Aspergillus oryzae." Fungal Genetics and Biology **48**(4): 438-444.
Hutter, S., A. J. Vilella, et al. (2006). "Genome-wide DNA polymorphism analyses using VariScan." Bmc Bioinformatics **7**.
Hyma, K. E., S. M. Saerens, et al. (2011). "Divergence in wine characteristics produced by wild and domesticated strains of Saccharomyces cerevisiae." Fems Yeast Research **11**(7): 540-551.
Ito, K., K. Yoshida, et al. (1990). "Volatile Compounds Produced by the Fungus Aspergillus-Oryzae in Rice Koji and Their Changes during Cultivation." Journal of Fermentation and Bioengineering **70**(3): 169-172.
Janssens, L., H. L. Depooter, et al. (1992). "Production of Flavors by Microorganisms." Process Biochemistry **27**(4): 195-215.
Jiang, H. and W. H. Wong (2009). "Statistical inferences for isoform expression in RNA-Seq." Bioinformatics **25**(8): 1026-1032.
Kato, N., M. Tokuoka, et al. (2011). "Genetic Safeguard against Mycotoxin Cyclopiazonic Acid Production in Aspergillus oryzae." Chembiochem **12**(9): 1376-1382.
Katoh, K. and H. Toh (2008). "Recent developments in the MAFFT multiple sequence alignment program." Briefings in Bioinformatics **9**(4): 286-298.
Keller-Seitz, M. U., U. Certa, et al. (2004). "Transcriptional response of yeast to aflatoxin B-1: Recombinational repair involving RAD51 and RAD1." Molecular Biology of the Cell **15**(9): 4321-4336.
Kelly, W. J., L. J. H. Ward, et al. (2010). "Chromosomal Diversity in Lactococcus lactis and the Origin of Dairy Starter Cultures." Genome Biology and Evolution **2**: 729-744.
Kessner, D., M. Chambers, et al. (2008). "ProteoWizard: open source software for rapid proteomics tools development." Bioinformatics **24**(21): 2534-2536.

Khaldi, N., F. T. Seifuddin, et al. (2010). "SMURF: Genomic mapping of fungal secondary metabolite clusters." Fungal Genetics and Biology **47**(9): 736-741.
Kitamoto, K. (2002). "Molecular biology of the Koji molds." Adv Appl Microbiol **51**: 129-53.
Klaassen, C. H. W., J. G. Gibbons, et al. (2012). "Evidence for genetic differentiation and variable recombination rates among Dutch populations of the opportunistic human pathogen Aspergillus fumigatus." Molecular Ecology **21**(1): 57-70.
Kobayashi, T., K. Abe, et al. (2007). "Genomics of Aspergillus oryzae." Bioscience Biotechnology and Biochemistry **71**(3): 646-670.
Kurtzman, C. P., M. J. Smiley, et al. (1986). "DNA Relatedness among Wild and Domesticated Species in the Aspergillus-Flavus Group." Mycologia **78**(6): 955-959.
Kusumoto, K., Y. Nogata, et al. (2000). "Directed deletions in the aflatoxin biosynthesis gene homolog cluster of Aspergillus oryzae." Current Genetics **37**(2): 104-111.
Legras, J. L., D. Merdinoglu, et al. (2007). "Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history." Molecular Ecology **16**(10): 2091-2102.
Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Research **18**(11): 1851-1858.
Libkind, D., C. T. Hittinger, et al. (2011). "Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast." Proceedings of the National Academy of Sciences of the United States of America **108**(35): 14539-14544.
Liti, G., D. M. Carter, et al. (2009). "Population genomics of domestic and wild yeasts." Nature **458**(7236): 337-341.
Ma, Z. Q., S. Dasari, et al. (2009). "IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering." Journal of Proteome Research **8**(8): 3872-3881.
MacCoss, M. J., W. H. McDonald, et al. (2002). "Shotgun identification of protein modifications from protein complexes and lens tissue." Proceedings of the National Academy of Sciences of the United States of America **99**(12): 7900-7905.
Machida, M., K. Asai, et al. (2005). "Genome sequencing and analysis of *Aspergillus oryzae*." Nature **438**(7071): 1157-61.
Machida, M., O. Yamada, et al. (2008). "Genomics of Aspergillus oryzae: Learning from the History of Koji Mold and Exploration of Its Future." DNA Research **15**(4): 173-183.
Makarova, K., A. Slesarev, et al. (2006). "Comparative genomics of the lactic acid bacteria." Proceedings of the National Academy of Sciences of the United States of America **103**(42): 15611-15616.
Makarova, K. S. and E. V. Koonin (2007). "Evolutionary genomics of lactic acid bacteria." Journal of Bacteriology **189**(4): 1199-1208.
McGovern, P. E., J. H. Zhang, et al. (2004). "Fermented beverages of pre- and proto-historic China." Proceedings of the National Academy of Sciences of the United States of America **101**(51): 17593-17598.
Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-8.

Mukherjee, M., B. A. Horwitz, et al. (2006). "A secondary metabolite biosynthesis cluster in Trichoderma virens: evidence from analysis of genes underexpressed in a mutant defective in morphogenesis and antibiotic production." Curr Genet **50**(3): 193-202.
Mukherjee, P. K., B. A. Horwitz, et al. (2011). "Secondary metabolism in Trichoderma - a genomic perspective." Microbiology **158**(Pt 1): 35-45.
Murakami, H., S. Takase, et al. (1967). "Non-Productivity of Aflatoxin by Japanese Industrial Strains of Aspergillus." Journal of General and Applied Microbiology **13**(4): 323-&.
Nicholson, M. J., A. Koulman, et al. (2009). "Identification of Two Aflatrem Biosynthesis Gene Loci in Aspergillus flavus and Metabolic Engineering of Penicillium paxilli To Elucidate Their Function." Applied and Environmental Microbiology **75**(23): 7469-7481.
Payne, G. A., W. C. Nierman, et al. (2006). "Whole genome comparison of Aspergillus flavus and A-oryzae." Medical Mycology **44**: S9-S11.
Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.
Punta, M., P. C. Coggill, et al. (2012). "The Pfam protein families database." Nucleic Acids Research **40**(D1): D290-D301.
Purugganan, M. D. and D. Q. Fuller (2009). "The nature of selection during plant domestication." Nature **457**(7231): 843-848.
Rank, C., M. Klejnstrup, et al. (2012). "Comparative Chemistry of Aspergillus oryzae (RIB40) and A. flavus (NRRL 3357)." Metabolites **2**(1): 36-56.
Rawlings, N. D., A. J. Barrett, et al. (2012). "MEROPS: the database of proteolytic enzymes, their substrates and inhibitors." Nucleic Acids Research **40**(D1): D343-D350.
Ren, Q., K. Chen, et al. (2007). "TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels." Nucleic Acids Res **35**(Database issue): D274-9.
Rokas, A. (2009). "The effect of domestication on the fungal proteome." Trends Genet **25**(2): 60-3.
Rokas, A., G. Payne, et al. (2007). "What can comparative genomics tell us about species concepts in the genus Aspergillus?" Stud Mycol **59**: 11-7.
Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.
Sabeti, P. C., S. F. Schaffner, et al. (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-1620.
Scott, D. (1989). "Specialty Enzymes and Products for the Food-Industry." Acs Symposium Series **389**: 176-192.
Smith, J. M. and J. Haigh (1974). "Hitch-Hiking Effect of a Favorable Gene." Genetical Research **23**(1): 23-35.
Stamatakis, A. (2006). "RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690.

Surget-Groba, Y. and J. I. Montoya-Burgos (2010). "Optimization of de novo transcriptome assembly from next-generation sequencing data." Genome Research **20**(10): 1432-1440.
Tabb, D. L., C. G. Fernando, et al. (2007). "MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis." Journal of Proteome Research **6**(2): 654-661.
Tamura, K., D. Peterson, et al. (2011). "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." Molecular Biology and Evolution **28**(10): 2731-2739.
Teramoto, Y., T. Hano, et al. (2000). "Production and characteristics of an ancient form of sake made with shitogi." Journal of the Institute of Brewing **106**(2): 95-99.
Tominaga, M., Y. H. Lee, et al. (2006). "Molecular analysis of an inactive aflatoxin biosynthesis gene cluster in Aspergillus oryzae RIB strains." Applied and Environmental Microbiology **72**(1): 484-490.
Vilella, A. J., A. Blanco-Garcia, et al. (2005). "VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data." Bioinformatics **21**(11): 2791-2793.
Wilhelm, M., M. Kirchner, et al. (2012). "mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets." Molecular & Cellular Proteomics **11**(1).
Wilson, D., V. Charoensawan, et al. (2008). "DBD--taxonomically broad transcription factor predictions: new content and functionality." Nucleic Acids Res **36**(Database issue): D88-92.
Yoshizaki, Y., H. Yamato, et al. (2010). "Analysis of Volatile Compounds in Shochu Koji, Sake Koji, and Steamed Rice by Gas Chromatography-Mass Spectrometry." Journal of the Institute of Brewing **116**(1): 49-55.
Yoshizawa, K. (1999). "Sake: Production and flavor." Food Reviews International **15**(1): 83-107.
Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome Research **18**(5): 821-829.

SUPPLEMENTARY FIGURES

Figure S4.1. Sequence variation in the glutaminase locus in *A. oryzae* is dramatically reduced. Schematic of the glutaminase locus. Gray and black bars represent non-coding coding regions, respectively. The arrow indicates the direction of transcription. Blue and green stars indicate polymorphic sites in *A. oryzae* and *A. flavus*, respectively. The highlighted region shows the translation of the single nonsynonymous variant identified.

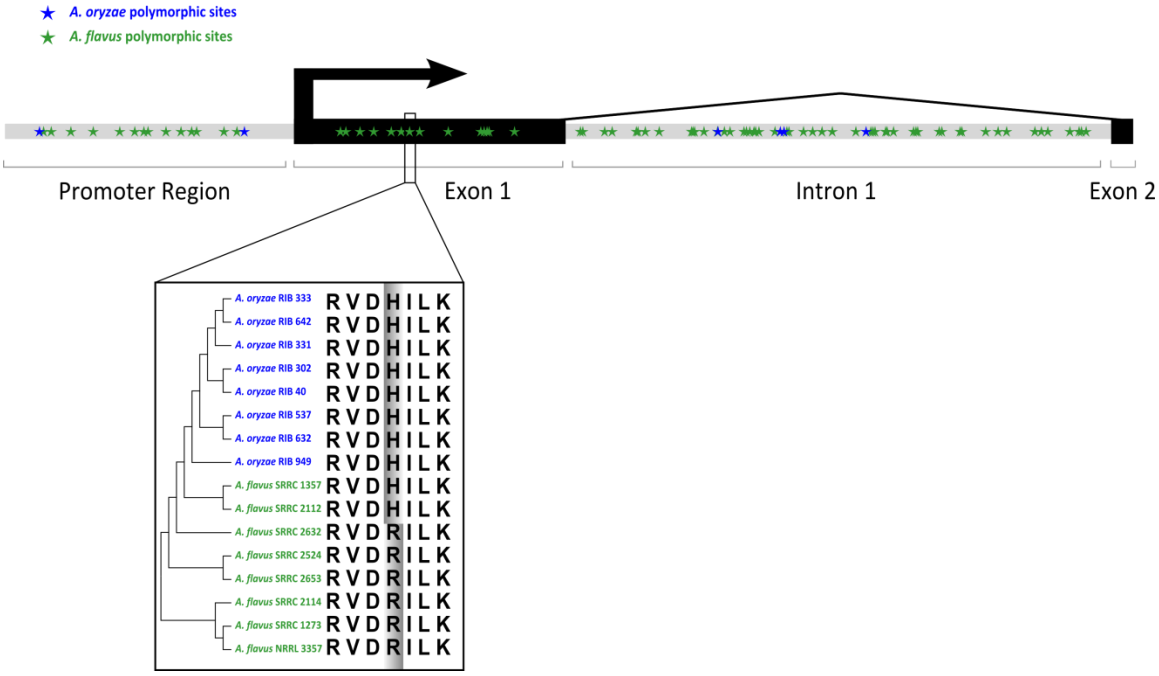
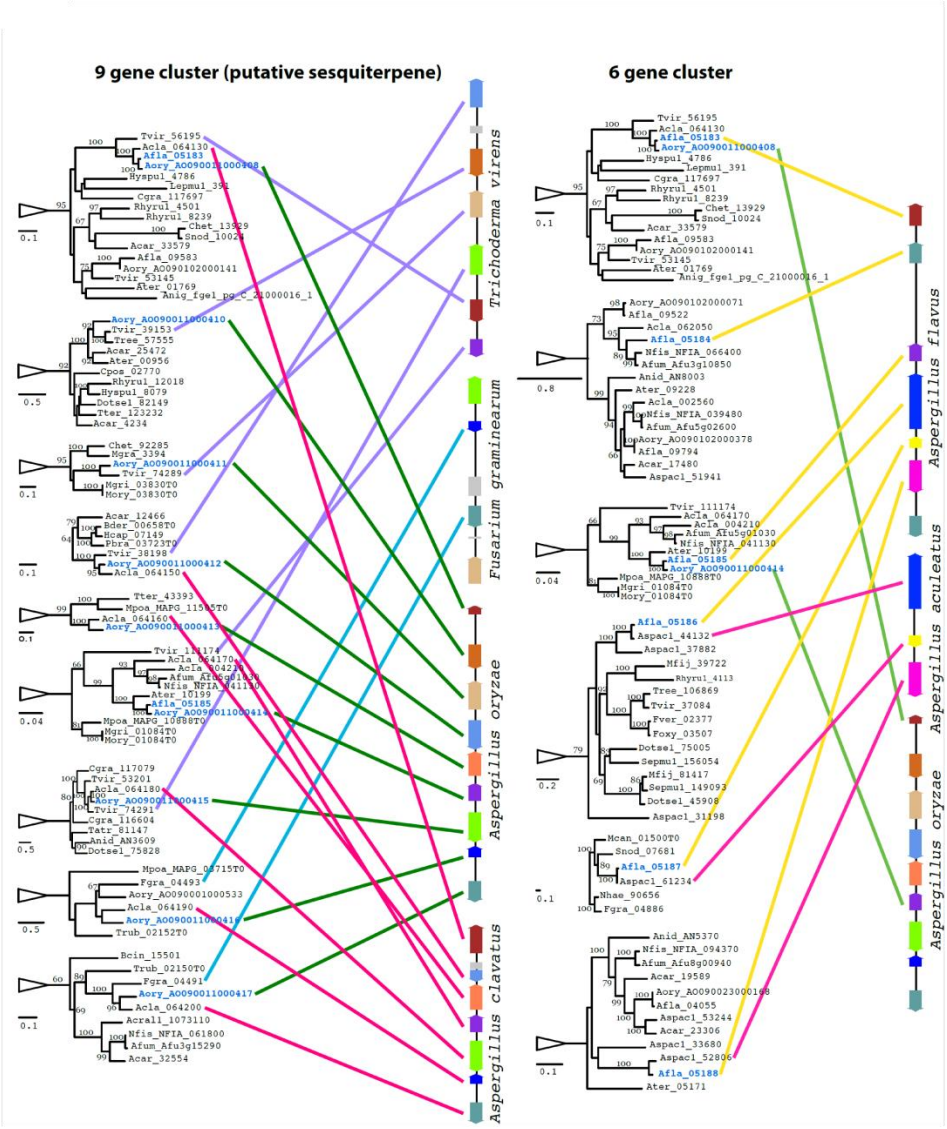


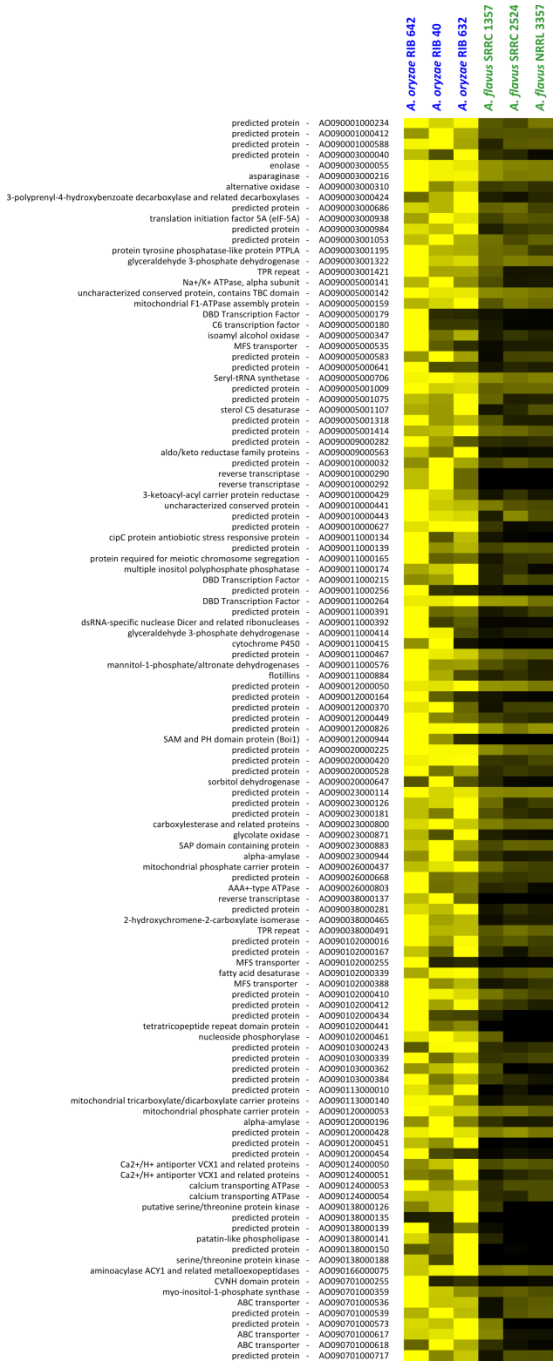
Figure S4.2. Evolutionary history of genes from the two alternative "allele gene clusters" at the sesquiterpene gene cluster locus. Maximum Likelihood (ML) phylogenies of amino acid sequences from genes of the 9- and 6-gene clusters. Phylogenetic trees are arbitrarily rooted and truncated for simplicity (collapsed branch at root). Scale bars denote substitutions per site. Support values are from 100 ML bootstrap replicates when greater or equal to 60. Chromosomal clustering of homologous genes from *A. flavus*, *A. oryzae*, *A. clavatus*, *A. aculeatus*, *F. graminearum*, and *T. virens* is shown to right of trees, connected by lines to corresponding sequence names.



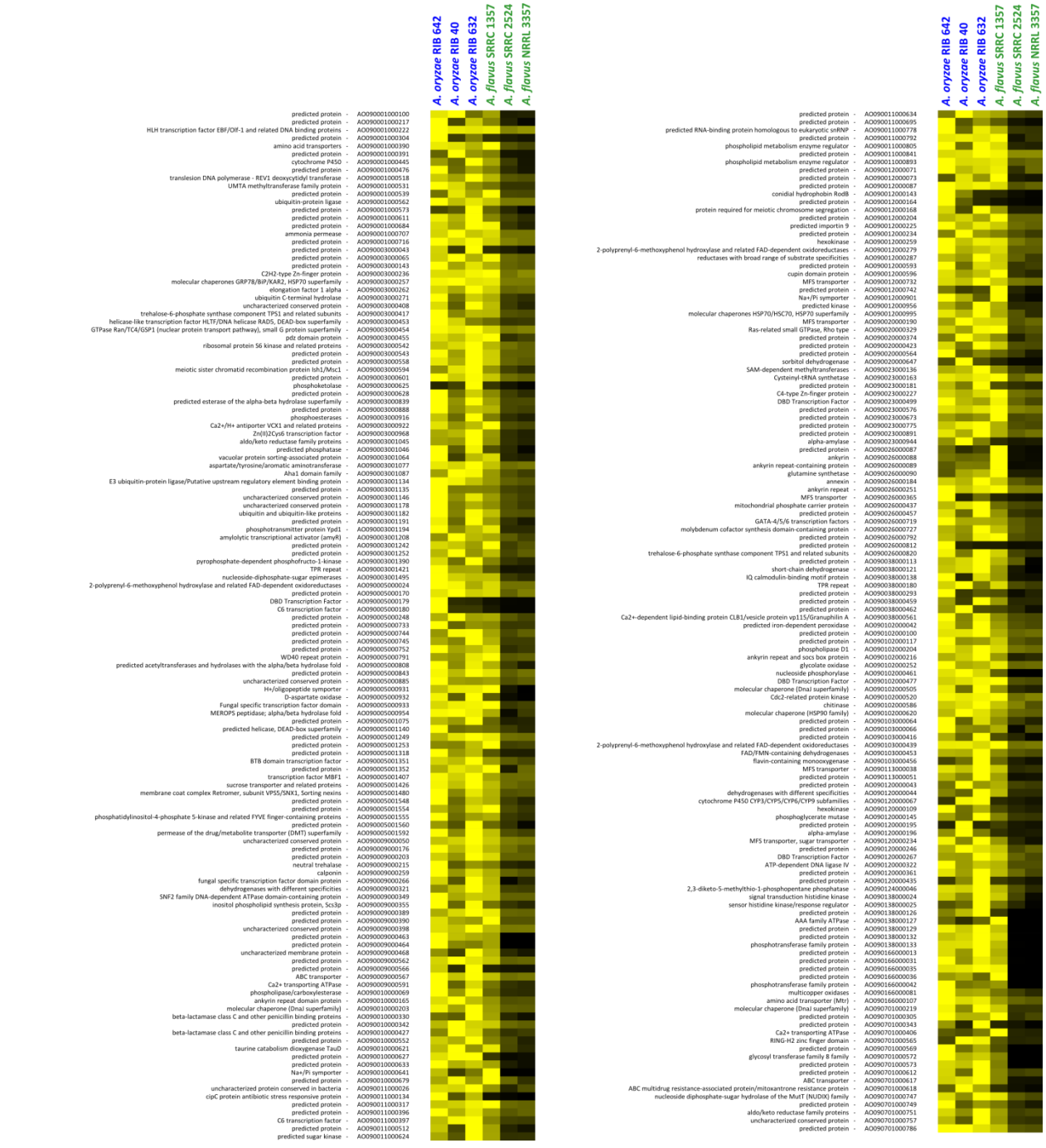
Acar=*Aspergillus carbonarius*
Acla=*A. clavatus*
Acral1=*Acremonium alcalophilum*
Afla=*A. flavus*
Afum=*A. fumigatus*
Anid=*A. nidulans*
Anig=*A. niger*
Aory=*A. oryzae*
Aspac1=*A. aculeatus*
Ater=*A. terreus*
Acral1=*Acremonium alcalophilum*
Bcin=*Botrytis cinerea*
Bder=*Batrachochytrium dendrobatidis*
Cgra=*Cladonia grayi*
Chet=*Cochliobolus heterostrophus*
Cpos=*Coccidioides posadasii*
Dotse1=*Dothiostroma septosporum*
Fgra=*Fusarium graminearum*
Foxy=*F. oxysporum*
Fver=*F. verticillioides*
Hcap=*Histoplasma capsulatum*
Hyspu1=*Hysterium pulicare*
Lepmu1=*Leptosphaeria maculans*
Mcan=*Microsporum canis*
Mfij=*Mycosphaerella fijiensis*
Mgra=*M. graminicola*
Mgri=*Magnaporthe grisea*
Mory=*M. oryzae*
Mpoa=*M. poae*
Nfis=*Neosartorya fischeri*
Nhae=*Nectria haematococca*
Pbra=*Paracoccidiodes brasiliensis*
Rhyru1=*Rhytidhysteron rufulum*
Sepmu1=*Septoria musiva*
Snod=*Stagonospora nodorum*
Tatr=*Trichoderma atroviride*
Tree=*T. reesei*
Tvir=*T. virens*
Trub=*Trichophyton rubrum*
Tter=*Theilavia terrestris*

Figure S4.3. The set of up-regulated *A. oryzae* genes. (A) Expression levels of the 117 genes that are up-regulated by at least 1.5 fold in the three *A. oryzae* isolates relative to the three *A. flavus* isolates. (B) Expression levels of the 257 genes that are up-regulated by at least 1.5 fold in the four atoxigenic isolates (the three *A. oryzae* isolates and the *A. flavus* SRRC 1357 isolate) against the two *A. flavus* toxigenic isolates (SRRC 2524 and NRRL 3357). For each gene, the color of the heat map cell corresponds to its expression level (in RPKM units), where black is zero expression and yellow is the maximum RPKM for that gene. The predicted function of each gene is given next to each gene name.

A



B


A. oryzae RIB 642
A. oryzae RIB 40
A. oryzae RIB 632
A. flavus SRRC 1357
A. flavus SRRC 2524
A. flavus NRRL 3357
predicted protein - AO090001000100
predicted protein - AO090001000217
HLH transcription factor EBF/Olf-1 and related DNA binding proteins - AO090001000222
predicted protein - AO090001000304
amino acid transporters - AO090001000390
predicted protein - AO090001000391
cytochrome P450 - AO090001000445
predicted protein - AO090001000476
translesion DNA polymerase - REV1 deoxycytidyl transferase - AO090001000518
UMTA methyltransferase family protein - AO090001000531
predicted protein - AO090001000539
ubiquitin-protein ligase - AO090001000562
predicted protein - AO090001000573
predicted protein - AO090001000611
predicted protein - AO090001000684
ammonia permease - AO090001000707
predicted protein - AO090001000716
predicted protein - AO090003000043
predicted protein - AO090003000065
predicted protein - AO090003000143
C2H2-type Zn-finger protein - AO090003000236
molecular chaperones GRP78/BiP/KAR2, HSP70 superfamily - AO090003000257
elongation factor 1 alpha - AO090003000262
ubiquitin C-terminal hydrolase - AO090003000271
uncharacterized conserved protein - AO090003000408
trehalose-6-phosphate synthase component TPS1 and related subunits - AO090003000417
helicase-like transcription factor HLTF/DNA helicase RAD5, DEAD-box superfamily - AO090003000453
GTPase Ran/TC4/GSP1 (nuclear protein transport pathway), small G protein superfamily - AO090003000454
pdz domain protein - AO090003000455
ribosomal protein S6 kinase and related proteins - AO090003000542
predicted protein - AO090003000543
predicted protein - AO090003000558
meiotic sister chromatid recombination protein Ish1/Mxc1 - AO090003000594
predicted protein - AO090003000601
phosphoketolase - AO090003000625
predicted protein - AO090003000628
predicted esterase of the alpha-beta hydrolase superfamily - AO090003000839
predicted protein - AO090003000888
phosphoesterases - AO090003000916
Ca2+/H+ antiporter VCX1 and related proteins - AO090003000922
Zn(II)2Cys6 transcription factor - AO090003000968
aldo/keto reductase family proteins - AO090003001045
predicted phosphatase - AO090003001046
vacuolar protein sorting-associated protein - AO090003001064
aspartate/tyrosine/aromatic aminotransferase - AO090003001077
Aha1 domain family - AO090003001087
E3 ubiquitin-protein ligase/Putative upstream regulatory element binding protein - AO090003001134
predicted protein - AO090003001135
uncharacterized conserved protein - AO090003001146
uncharacterized conserved protein - AO090003001178
ubiquitin and ubiquitin-like proteins - AO090003001182
predicted protein - AO090003001191
phosphotransmitter protein Ypd1 - AO090003001194
amylolytic transcriptional activator (amyR) - AO090003001208
predicted protein - AO090003001242
predicted protein - AO090003001252
pyrophosphate-dependent phosphofructo-1-kinase - AO090003001390
TPR repeat - AO090003001421
nucleoside-diphosphate-sugar epimerases - AO090003001495
2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases - AO090005000024
predicted protein - AO090005000170
DBD Transcription Factor - AO090005000179
C6 transcription factor - AO090005000180
predicted protein - AO090005000248
predicted protein - AO090005000733
predicted protein - AO090005000744
predicted protein - AO090005000745
predicted protein - AO090005000752
WD40 repeat protein - AO090005000791
predicted acetyltransferases and hydrolases with the alpha/beta hydrolase fold - AO090005000808
predicted protein - AO090005000843
uncharacterized conserved protein - AO090005000885
H+/oligopeptide symporter - AO090005000931
D-aspartate oxidase - AO090005000932
Fungal specific transcription factor domain - AO090005000933
MEROPS peptidase; alpha/beta hydrolase fold - AO090005000954
predicted protein - AO090005001075
predicted helicase, DEAD-box superfamily - AO090005001140
predicted protein - AO090005001249
predicted protein - AO090005001253
predicted protein - AO090005001318
BTB domain transcription factor - AO090005001351
predicted protein - AO090005001352
transcription factor Mbf1 - AO090005001407
sucrose transporter and related proteins - AO090005001426
membrane coat complex Retromer, subunit VPS5/SNX1, Sorting nexins - AO090005001480
predicted protein - AO090005001548
predicted protein - AO090005001554
phosphatidylinositol-4-phosphate 5-kinase and related FYVE finger-containing proteins - AO090005001555
predicted protein - AO090005001560
permease of the drug/metabolite transporter (DMT) superfamily - AO090005001592
uncharacterized conserved protein - AO090009000050
predicted protein - AO090009000176
predicted protein - AO090009000203
neutral trehalase - AO090009000215
calponin - AO090009000259
fungal specific transcription factor domain protein - AO090009000266
dehydrogenases with different specificities - AO090009000321
SNF2 family DNA-dependent ATPase domain-containing protein - AO090009000349
inositol phospholipid synthesis protein, Scs3p - AO090009000355
predicted protein - AO090009000389
predicted protein - AO090009000390
uncharacterized conserved protein - AO090009000398
predicted protein - AO090009000463
predicted protein - AO090009000464
uncharacterized membrane protein - AO090009000468
predicted protein - AO090009000562
predicted protein - AO090009000566
ABC transporter - AO090009000567
Ca2+ transporting ATPase - AO090009000591
phospholipase/carboxylesterase - AO090010000069
ankyrin repeat domain protein - AO090010000165
molecular chaperone (Dnaj superfamily) - AO090010000203
beta-lactamase class C and other penicillin binding proteins - AO090010000330
predicted protein - AO090010000342
beta-lactamase class C and other penicillin binding proteins - AO090010000427
predicted protein - AO090010000552
taurine catabolism dioxygenase TauD - AO090010000621
predicted protein - AO090010000627
predicted protein - AO090010000633
Na+/Pi symporter - AO090010000641
predicted protein - AO090010000679
uncharacterized protein conserved in bacteria - AO090011000026
cipC protein antibiotic stress responsive protein - AO090011000134
predicted protein - AO090011000317
predicted protein - AO090011000396
C6 transcription factor - AO090011000397
predicted protein - AO090011000512
predicted sugar kinase - AO090011000624
predicted protein - AO090011000634
predicted protein - AO090011000695
predicted RNA-binding protein homologous to eukaryotic snRNP - AO090011000778
predicted protein - AO090011000792
phospholipid metabolism enzyme regulator - AO090011000805
predicted protein - AO090011000841
phospholipid metabolism enzyme regulator - AO090011000893
predicted protein - AO090012000071
predicted protein - AO090012000073
predicted protein - AO090012000087
conidial hydrophobin RodB - AO090012000143
predicted protein - AO090012000164
protein required for meiotic chromosome segregation - AO090012000168
predicted protein - AO090012000204
predicted importin 9 - AO090012000225
predicted protein - AO090012000234
hexokinase - AO090012000259
2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases - AO090012000279
reductases with broad range of substrate specificities - AO090012000287
predicted protein - AO090012000593
cupin domain protein - AO090012000596
MFS transporter - AO090012000732
predicted protein - AO090012000742
Na+/Pi symporter - AO090012000901
predicted kinase - AO090012000956
molecular chaperones HSP70/HSC70, HSP70 superfamily - AO090012000995
MFS transporter - AO090020000190
Ras-related small GTPase, Rho type - AO090020000329
predicted protein - AO090020000374
predicted protein - AO090020000423
predicted protein - AO090020000564
sorbitol dehydrogenase - AO090020000647
SAM-dependent methyltransferases - AO090023000136
Cysteinyl-tRNA synthetase - AO090023000163
predicted protein - AO090023000181
C4-type Zn-finger protein - AO090023000227
DBD Transcription Factor - AO090023000499
predicted protein - AO090023000576
predicted protein - AO090023000673
predicted protein - AO090023000775
predicted protein - AO090023000891
alpha-amylase - AO090023000944
predicted protein - AO090026000087
ankyrin - AO090026000088
ankyrin repeat-containing protein - AO090026000089
glutamine synthetase - AO090026000090
annexin - AO090026000184
ankyrin repeat - AO090026000251
MFS transporter - AO090026000365
mitochondrial phosphate carrier protein - AO090026000437
predicted protein - AO090026000457
GATA-4/5/6 transcription factors - AO090026000719
molybdenum cofactor synthesis domain-containing protein - AO090026000727
predicted protein - AO090026000792
predicted protein - AO090026000812
trehalose-6-phosphate synthase component TPS1 and related subunits - AO090026000820
predicted protein - AO090038000113
short-chain dehydrogenase - AO090038000121
IQ calmodulin-binding motif protein - AO090038000138
TPR repeat - AO090038000180
predicted protein - AO090038000293
predicted protein - AO090038000459
predicted protein - AO090038000462
Ca2+-dependent lipid-binding protein CLB1/vesicle protein vp115/Granuphilin A - AO090038000561
predicted iron-dependent peroxidase - AO090102000042
predicted protein - AO090102000100
predicted protein - AO090102000117
phospholipase D1 - AO090102000204
ankyrin repeat and socs box protein - AO090102000216
glycolate oxidase - AO090102000252
nucleoside phosphorylase - AO090102000461
DBD Transcription Factor - AO090102000477
molecular chaperone (Dnaj superfamily) - AO090102000505
Cdc2-related protein kinase - AO090102000520
chitinase - AO090102000586
molecular chaperone (HSP90 family) - AO090102000620
predicted protein - AO090103000064
predicted protein - AO090103000066
predicted protein - AO090103000416
2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases - AO090103000439
FAD/FMN-containing dehydrogenases - AO090103000453
flavin-containing monooxygenase - AO090103000456
MFS transporter - AO090113000038
predicted protein - AO090113000051
predicted protein - AO090120000043
dehydrogenases with different specificities - AO090120000044
cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies - AO090120000067
hexokinase - AO090120000109
phosphoglycerate mutase - AO090120000145
predicted protein - AO090120000195
alpha-amylase - AO090120000196
MFS transporter, sugar transporter - AO090120000234
predicted protein - AO090120000246
DBD Transcription Factor - AO090120000267
ATP-dependent DNA ligase IV - AO090120000322
predicted protein - AO090120000361
predicted protein - AO090120000435
2,3-diketo-5-methylthio-1-phosphopentane phosphatase - AO090124000046
signal transduction histidine kinase - AO090138000024
sensor histidine kinase/response regulator - AO090138000025
predicted protein - AO090138000126
AAA family ATPase - AO090138000127
predicted protein - AO090138000129
predicted protein - AO090138000132
phosphotransferase family protein - AO090138000133
predicted protein - AO090166000013
predicted protein - AO090166000031
predicted protein - AO090166000035
predicted protein - AO090166000036
phosphotransferase family protein - AO090166000042
multicopper oxidases - AO090166000081
amino acid transporter (Mtr) - AO090166000107
molecular chaperone (Dnaj superfamily) - AO090701000219
predicted protein - AO090701000305
predicted protein - AO090701000343
Ca2+ transporting ATPase - AO090701000406
RING-H2 zinc finger domain - AO090701000565
predicted protein - AO090701000569
glycosyl transferase family 8 family - AO090701000572
predicted protein - AO090701000573
predicted protein - AO090701000612
ABC transporter - AO090701000617
ABC multidrug resistance-associated protein/mitoxantrone resistance protein - AO090701000618
nucleoside diphosphate-sugar hydrolase of the MutT (NUDIX) family - AO090701000747
predicted protein - AO090701000749
aldo/keto reductase family proteins - AO090701000751
uncharacterized conserved protein - AO090701000757
predicted protein - AO090701000786

CHAPTER V

COMPARATIVE AND FUNCTIONAL CHARACTERIZATION OF INTRAGENIC TANDEM REPEATS IN TEN *ASPERGILLUS* GENOMES

John G. Gibbons and Antonis Rokas

Department of Biological Sciences, Vanderbilt University, Box 351634 Station B, Nashville, TN 37235-1634, USA

This chapter is published in *Molecular Biology and Evolution*, 2009, 26: 591-602.

ABSTRACT

Intragenic tandem repeats (ITRs) are consecutive repeats of three or more nucleotides found in coding regions. ITRs are the underlying cause of several human genetic diseases, and have been associated with phenotypic variation, including pathogenesis, in several kingdoms. We have examined the evolution and functional role of ITRs in ten genomes spanning the fungal genus *Aspergillus*, a clade of relevance to medicine, agriculture, and industry. We identified several hundred ITRs in each of the species examined. ITR content varied extensively between species, with an average 79% of ITRs unique to a given species. For the fraction of conserved ITR regions, sequence comparisons within species and between close relatives revealed that they were highly variable. ITR-containing proteins were evolutionarily less conserved, compositionally distinct than other proteins in the genome and overrepresented for domains associated with cell-surface localization and function. Furthermore, ITR-containing proteins were preferentially associated with the functional processes of transcription, cellular communication and cell-type differentiation and disassociated from metabolism and energy. Despite the evolutionary lability of ITR regions, their functional associations appear to be remarkably conserved across eukaryotes. Fungal ITRs are likely involved in a variety of developmental processes involving transcriptional regulation, such as control of self / non-self recognition, as well as to cell-surface associated functions such as cell adhesion. Thus, the contribution of ITRs to the fungal lifestyle may be more general than previously assumed.

INTRODUCTION

A recurrent theme in eukaryotic genomes is the presence of repetitive DNA (Toth, Gaspari et al. 2000; Katti, Ranjekar et al. 2001; Li, Korol et al. 2004; Hancock and Simon 2005; Karaoglu, Lee et al. 2005; Thomas 2005; Kashi and King 2006), notably highlighted by the human genome which is composed of approximately 35% repetitive elements (Venter, Adams et al. 2001). One such class of repetitive DNA is intragenic tandem repeats (ITRs) which comprise of 3 or more nucleotides repeated in tandem, found in protein coding regions. ITRs include short sequence repeats, such as microsatellites (Ellegren 2004; Li, Korol et al. 2004), as well as longer repeats that span tens to hundreds of nucleotides, such as the ~100nt repeat identified in the *FLO*1, *FLO5*, and *FLO9* genes in *Saccharomyces cerevisiae* (Verstrepen, Jansen et al. 2005).

ITR sequences are typically mutationally unstable and prone to local expansion and contraction either via unequal recombination events or slip-strand mispairing (Levinson and Gutman 1987; Schlotterer and Tautz 1994; Bichara, Wagner et al. 2006). This inherent mutational instability frequently results in elevated mutation rates relative to the rest of the genome, especially for short ITRs (Dieringer and Schlotterer 2003; Kashi and King 2006; Moxon, Bayliss et al. 2006). For example, in humans, microsatellite mutation rates are as frequent as 10^{-3} to 10^{-4} per locus per generation (Weber and Wong 1993) compared to a rate of 10^{-8} per generation for single-nucleotide substitutions (Drake, Charlesworth et al. 1998). Not surprisingly, the high mutation rate of ITRs is the cause of several human hereditary disorders (Sherman, Jacobs et al. 1985; Sutherland and Richards 1995; Pearson, Edamura et al. 2005; Mirkin 2007). For example, Huntington's

disease, an inherited autosomal dominant neurodegenerative disorder, is caused by an expansion of a CAG repeat in exon 1 of the *IT-15* gene which results in a glutamine expansion in the protein product (Schilling, Sharp et al. 1995). However, variation in ITRs has also been linked to variation in circadian clock adjustment in the fruit fly (Sawyer, Hennessy et al. 1997) as well as skeletal morphology in domestic dog breeds (Fondon and Garner 2004).

Genotypic (Balajee, Tay et al. 2007; Levdansky, Romano et al. 2007) and phenotypic (Verstrepen, Reynolds et al. 2004; Verstrepen, Jansen et al. 2005; Fidalgo, Barrales et al. 2006; Michael, Park et al. 2007) variation associated with ITRs has also been observed in fungi. For example, ITR variation in the FLO1 protein of *S. cerevisiae* is positively associated with an increase in cell-cell adhesion (Verstrepen, Jansen et al. 2005), whereas ITR variation in the FLO11 protein of the same species contributes to the formation of self-supporting biofilm (Fidalgo, Barrales et al. 2006). ITRs have also been identified in members of the ALS protein family which are thought to play a similar role in mediating adhesion to other cells and substrates in *Candida albicans* and *C. glabrata* (Verstrepen, Reynolds et al. 2004; Oh, Cheng et al. 2005). These findings have lead to the hypothesis that fungal ITRs may be implicated in the generation of variation in cell-surface proteins, molecules with active roles in the colonization of host tissue and evasion of its immune system (Jordan, Snyder et al. 2003; Verstrepen, Reynolds et al. 2004; Levdansky, Romano et al. 2007).

We were particularly interested in studying the evolution and function of ITRs in the filamentous ascomycete *Aspergillus*, a genus with a large societal impact, both beneficial and detrimental. For example, the species *A. oryzae* and *A. niger* are commercially exploited for a variety of industrial purposes (Machida, Asai et al. 2005; Pel, de Winde et al. 2007). In contrast, *A. flavus* is a producer of the carcinogenic compound aflatoxin as well as an agricultural pathogen (corn, cotton, peanuts) that causes annual losses totaling hundreds of millions of dollars (Yu, Whitelaw et al. 2004; Yu, Cleveland et al. 2005). *A. fumigatus* and *A. terreus* are potentially lethal opportunistic pathogens and the leading causes of invasive pulmonary aspergillosis (Patterson, Kirkpatrick et al. 2000; Yu, Whitelaw et al. 2004). It is perhaps a testament to the relevance of this genus to human affairs that ten draft genomes from eight species are already available (Galagan, Calvo et al. 2005; Machida, Asai et al. 2005; Nierman, Pain et al. 2006; Pel, de Winde et al. 2007; Fedorova, Khaldi et al. 2008).

To understand the comparative biology of ITRs in this important fungal genus, we first identified and calculated the frequency and distribution of ITRs across the genomes of eight *Aspergillus* species, and evaluated their relative placement in proteins. We next assessed the evolutionary conservation of ITRs and ITR-containing proteins by comparing the relative proportion of orthologous ITRs and ITR-containing proteins in the entire, background and ITR proteomes. We then examined ITR variation levels by analyzing the observed variation within and between closely-related species. To gain insight into the functional biology of *Aspergillus* ITR-containing proteins we first determined whether ITR-containing proteins were compositionally distinct from the

background proteome. We next bioinformatically examined ITR-containing proteins for the presence of a variety of cell-surface associated protein domains. Finally, we evaluated whether ITR-containing proteins were associated with specific functional categories.

MATERIALS AND METHODS

Genome Sequences

The coding sequences analyzed in this study were downloaded from the *Aspergillus* comparative site at the Broad Institute (http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiHome.html). They are also available in public databases under the accession numbers *A. flavus* NRRL 3357 [Genbank: AAIH01000000]; *A. oryzae* RIB 40 [DDJB: AP007150-AP007177]; *A. terreus* NIH2624 [Genbank: AAJN01000000]; *A. niger* CBS 513.88 [EMBL: AM270980-AM270998]; *A. niger* ATCC 1015 [http://genome.jgi-psf.org/Aspni1/Aspni1.home.html:]; *N. fischeri* NRRL 181 [Genbank: AAKE03000000]; *A. fumigatus* CEA10 [Genbank: ABDB01000000]; *A. fumigatus* Af293 [Genbank: AAHF01000000]; *A. clavatus* NRRL 1 [Genbank: AAKD00000000]; and *A. nidulans* FGSC A4 [Genbank: AACD00000000].

Genomic Identification of ITRs

The EMBOSS ETANDEM software was used to independently identify short (3-39nt) and long (40-500nt) ITRs in each of the 10 analyzed transcriptomes (Rice, Longden et al. 2000). ITRs with a consensus sequence conservation ≥ 85% and an absolute sequence length of at least 24nt (i.e., eight copies of a trinucleotide repeat; six copies of a tetranucleotide repeat; two copies of any large repeat unit) were considered significant, as 24nt is the minimum criteria for trinucleotide repeat detection in the software.

Conservation of ITR-containing Genes

Orthologs were identified using the reciprocal-best-BLAST-hit approach for each pairwise species comparison with a cut-off e-value of 1e-06 (Rokas, Payne et al. 2007; Moreno-Hagelsieb and Latimer 2008). Average conservation was calculated by dividing the total number of orthologs shared by a species pair by the total number of proteins of the species with the smaller proteome. We used this method to calculate the conservation of the entire proteome, background proteome and ITR-containing proteome. ITR conservation was calculated by identifying the number of shared ITRs between orthologs of each species pair and dividing it by the total number of ITRs in the orthologs.

ITR Variation Within and Between Species

Investigation of ITR variation was examined in two intraspecific and two interspecific comparisons: *A. fumigatus* strain Af293 vs. CEA10, *A. niger* strain CBS 518.33 vs. ATCC1015, *A. flavus* vs. *A. oryzae* and *A. fumigatus* Af293 vs. *N. fischeri* (the closely related sexual relative of *A. fumigatus*). In each case, the orthologs of all ITR-containing genes were compared. ITR-orthologs were categorized as a) MONOMORPHIC, b) VARIABLE, c) ORTHOLOG NO ITR or d) NO ORTHOLOG. Because the current annotation of *A. niger* strain CBS 513.88 contains many instances of multiple stop codons per gene, only genes with a single stop codon were used.

The effect of repeat unit copy number and longest pure tract on ITR variation rates were examined via Student's t-test (Sokal and Rohlf 1995). Data from the two interspecific and intraspecific comparisons, respectively, were pooled as the distributions did not significantly deviate from each other. The average copy number and average pure tract

length of monomorphic and variable ITRs was independently assessed in trinucleotide and hexanucleotide repeats.

Amino Acid Composition

For each species, the absolute and relative frequency of each amino acid was calculated in the background and ITR-containing proteomes. The relative proportions of each amino acid were analyzed via Fisher's Exact Test (Sokal and Rohlf 1995). A Bonferroni corrected *p* value = 0.0003125 was implemented in order to limit overall experiment-wise error rates due to multiple comparisons.

Hydropathy Index

For each species, an average hydropathy score was calculated for each protein by assigning each amino acid with a numerical value based on the Kyte and Doolittle Hydropathy Index, and dividing the sum by the total number of amino acids in the protein (Kyte and Doolittle 1982). For each species, the mean hydropathy scores of background and ITR proteomes were compared via Student's t-test (Sokal and Rohlf 1995). A Bonferroni corrected *p* value = 0.00625 was implemented in order to limit experiment-wise error rates due to multiple comparisons.

The Relative Position of ITRs Within Proteins

To test the hypothesis that ITRs were randomly distributed throughout the protein, we adapted the methods of Huntley *et al* (Huntley and Clark 2007). Briefly, each protein was separated into three equal sized segments, the N-terminal, Mid-segment and C-terminal

(Huntley and Clark 2007). For each species, we calculated the midpoints of all ITRs and identified the protein segments the midpoints were located in. We generated the expected frequencies of ITR position by using the following equations: Mid-segment = (L / 3) / (L - l) and N-terminal and C-terminal = ((L / 3) – (l / 2)) / (L – l), where L = protein length and l = total ITR length (Huntley and Clark 2007). For each species, we then averaged the probabilities of each ITR and multiplied this by the total number of ITRs. G-tests were used to assess whether observed frequencies deviated from expected (Sokal and Rohlf 1995).

Functional Annotation and Classification

The SignalP version 3.0 software was used to predict signal peptides using the hidden Markov model constructed with eukaryotic proteins (Bendtsen, Nielsen et al. 2004). GPI anchors were predicted using the big-Pi predictor software (Eisenhaber, Schneider et al. 2004). Transmembrane helices were predicted using the TMHMM version 2.0 software, with hits considered significant when more than 18 amino acids in the transmembrane helix were predicted (Krogh, Larsson et al. 2001). The proportions of each predicted motif in each proteome were assessed via Fisher's Exact Test (Sokal and Rohlf 1995). A Bonferroni corrected p value = 0.00625 was implemented for each motif in order to limit experiment-wise error rates due to multiple comparisons. All statistical analysis was performed using the JMP software, version 5.0.1a (Frenkel and Blumenthal 2002). Putative functional domains were identified using the Pfam annotation of the eight *Aspergillus* proteomes (Finn, Mistry et al. 2006), which were downloaded from the *Aspergillus* comparative site at the Broad Institute .

To test the hypothesis that ITR-containing proteins differed in specific functions, the major FunCat (Ruepp, Zollner et al. 2004) categories for *A. oryzae*, *A. terreus*, *A. fumigatus*, and *A. nidulans* were retrieved from the MIPS PEDANT database (http://pedant.gsf.de/). For each category, the proportion of ITR-containing proteins and background proteins was assessed via Fisher's Exact Test (Sokal and Rohlf 1995).

In a whole genome expression profiling microarray analysis, Nierman *et al.* identified 458 genes in *A. fumigatus* that were differentially expressed at 30˚C, 37˚C and 48˚C (Nierman, Pain et al. 2006). The latter two temperatures represent ones that the species experiences in the human body and compost, respectively, and both are presumed to be more stressful than the 30˚C one. To test whether the ITR-containing genes of *A. fumigatus* were over- or under-represented in the gene set which is differentially-expressed under temperature stress, we examined the proportion of ITR-containing genes relative to the background ones in the gene set via Fisher's Exact Test (Sokal and Rohlf 1995).

RESULTS

Identification and Distribution of ITRs Across *Aspergillus*

We identified short (3-39nt) and long (>39nt) ITRs in each of the ten transcriptomes using the ETANDEM software (Rice, Longden et al. 2000). The total number of ITRs ranged from 172 to 345 per species (Table 5.1, Figure 5.1). The number of ITR-containing genes ranged from 154 to 317 (Table 5.1). Several ITR-containing genes are well characterized and are known to play key roles in fungal life style and pathogenicity (Figure 5.2). In many instances, individual genes harbored multiple ITRs (average = 1.11 ITRs per gene). In all species analyzed, short ITRs were far more abundant than long ITRs (Table 5.1, Figure 5.1). ITR abundance was not associated with genome size (r^2 = 0.15; F = 1.02; n = 8; p = 0.35).

Table 5.1 General Characteristics and ITR summary of the *Aspergilli*

	A. flavus	*A. oryzae*	*A. terreus*	*A. niger*	*N. fischeri*	*A. fumigatus*	*A. clavatus*	*A. nidulans*
Strain	NRRL 3357	RIB 40	NIH 2624	CBS 513.88	NRRL 181	Af 293	NRRL 1	FGSC A4
Number of genes	12587	12063	10406	13912	10403	9887	9120	10665
Total ITRs	235	204	172	345	222	222	313	215
ITR-containing genes	210	182	154	317	194	207	278	200
Genes containing multiple ITRs	20	17	14	22	24	15	28	12
Short ITRs	161	136	123	277	155	187	262	160
Long ITRs	74	68	49	68	67	35	51	55

Figure 5.1. **Distribution of ITRs across the *Aspergilli*.** (A) Phylogenetic relationships (Rokas 2007) and distribution of ITRs across the analyzed *Aspergillus* species. For each species, black bars represent total ITRs identified, white bars represent short ITRs (repeat unit 3 - 39 nt) and gray bars represent long ITRs (repeat unit >= 40nt). (B) Distribution of short and long ITRs across the *Aspergilli*. ITRs were binned according to repeat unit size (X axis). The Y axis represents the total number of occurrences. Each species is indicated by a different color and label on the Z axis.

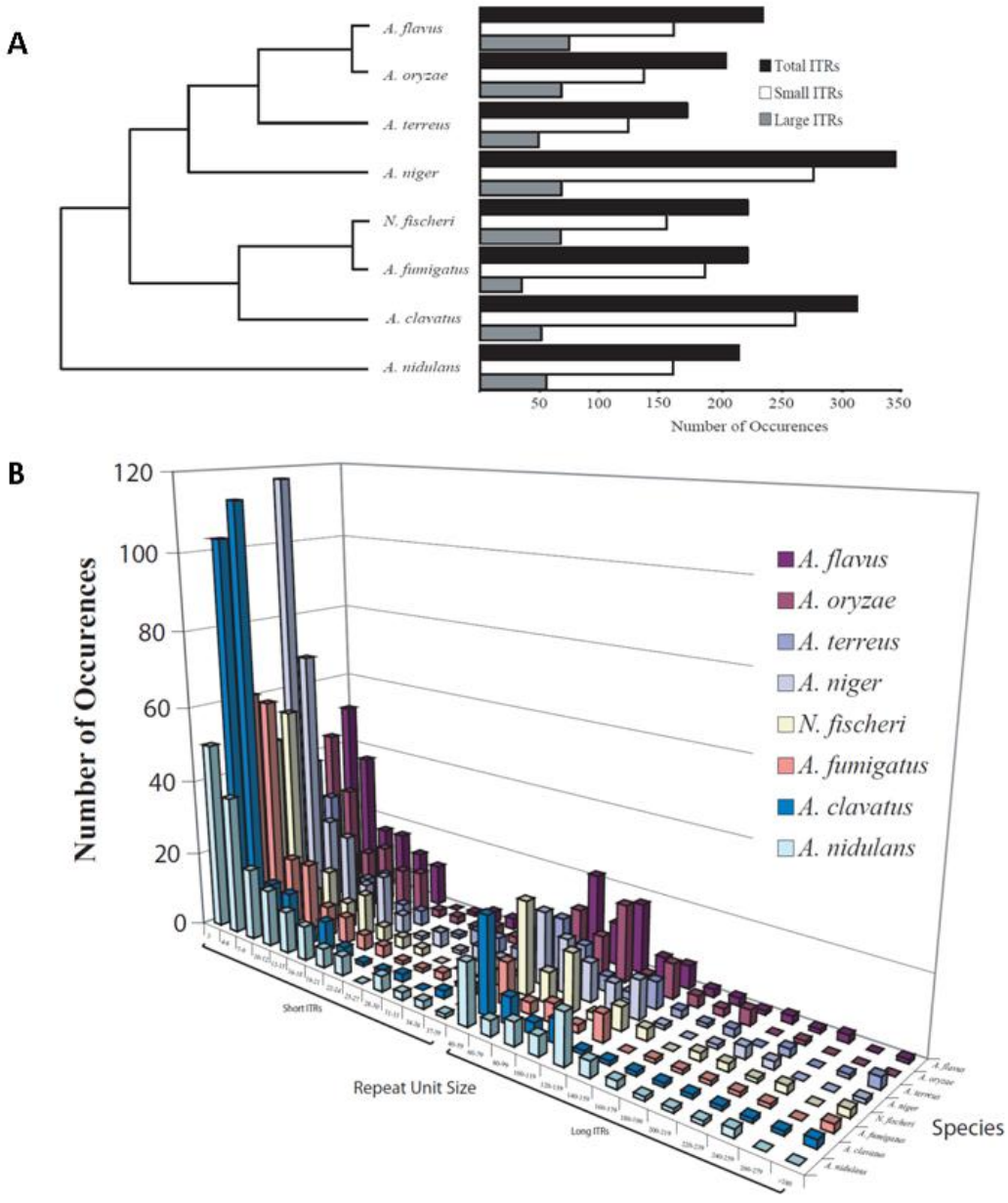
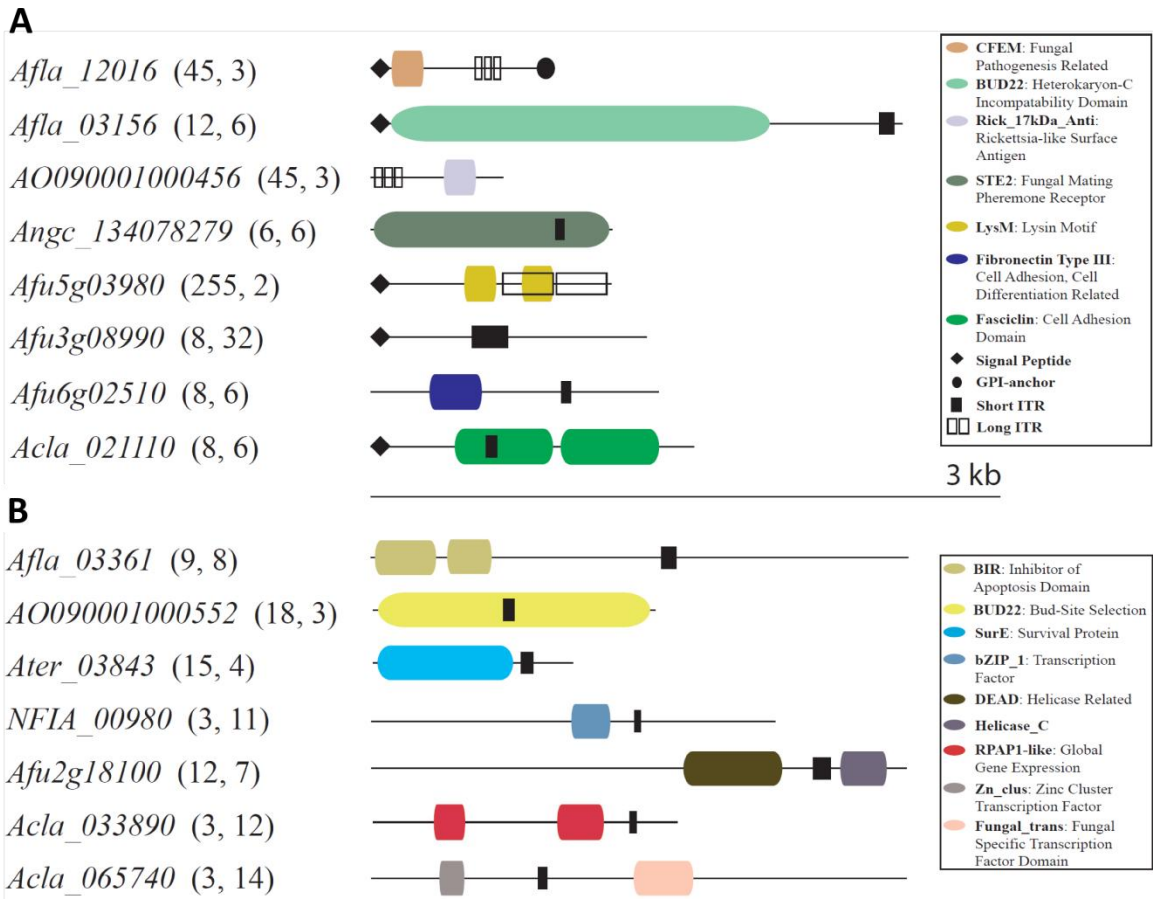


Figure 5.2. Representative *Aspergillus* ITR-containing proteins. (A) Cell-surface associated ITR-containing proteins. (B) ITR-containing proteins associated with transcriptional regulation. The numbers in parentheses after each gene identifier are the repeat unit size and copy number, respectively. Short repeats (repeat unit 3 to 39nt) are represented by a single solid black rectangle. Long repeats (repeat unit 40 to 500nt) are represented by a hollow rectangle for each repeat unit. Pfam functional domains are represented by colored ellipses.



We found that approximately 95% (1835 of 1928) of ITR repeat units had lengths divisible by three, a result likely reflecting selection toward repeat units which do not alter the reading frame (Metzgar, Bytof et al. 2000). The remaining 5% (93 of 1928) of ITRs consisted of repeat units which were not divisible by three. Thirty five such ITRs were identified as tetranucleotides, although 16 of them could be further collapsed to

dinucleotides. The remaining 58 ITRs exhibited repeat unit lengths between 5 and 104 nucleotides, and occurred less frequently.

Approximately 40% (35 of 93) of the ITRs, whose repeat units were not divisible by three, had an absolute sequence length divisible by three. Interestingly, only one out of these 35 ITRs was variable (found in the *A. fumigatus* gene *afu5g06790*), with both repeat unit copy number variants remaining in frame. The remaining 60% (58 of 93) of the ITRs did not have an absolute sequence length divisible by three and was not variable. Because our study was restricted to coding regions without internal stop codons, we currently do not know whether any of these proteins exhibits phase variation (van der Woude and Baumler 2004; Moxon, Bayliss et al. 2006).

ITRs with short repeat unit sizes were more abundant than ITRs with long repeat unit sizes (Figure 5.1). The mean short ITR repeat unit copy number (average = 9.08) was significantly larger than the mean long ITR repeat unit copy number (average = 3.20) (t = -17.214; $d.f.$ = 1926; p = 7.0e-70). The largest repeat unit copy numbers identified were two 97-repeat trinucleotides, found in the non-orthologous ITR-containing genes of *A. niger* (*angc_134081920*) and *A. oryzae* (*ao090010000583*).

Identifying the Relative Position of ITRs within Proteins

The tendency of ITRs to occur toward the end of a protein has been frequently observed in eukaryotes (Alba and Guigo 2004; Siwach, Pophaly et al. 2006; Huntley and Clark 2007). To test whether this was the case in the *Aspergilli* we first identified the relative

midpoint location of each ITR in its respective protein and divided each protein into three equally sized regions (N-terminal, Mid-segment, and C-terminal), following a previously developed protocol (Huntley and Clark 2007). We then compared the observed frequencies to those expected by chance for each species.

We found that the distribution of ITRs across proteins' lengths was random for seven of the eight species. *A. niger* was the only species to significantly deviate from the null distribution ($p = 0.0015$). Similar results were obtained when short and long ITRs were independently tested for each species. When data from all species was pooled, however, we do find a significant deviation from the null distribution (*d.f.* = 2; $g = 25.657$; $p = 2.683e06$). We also observed a clear, but non-significant, trend of ITR overrepresentation in the C-terminal portion of ITR-containing proteins across all species other than *A. oryzae* and *A. clavatus*.

Conservation of ITRs and ITR-containing Proteins

To evaluate the evolutionary conservation of ITR-containing proteins relative to the entire and background proteomes, we identified all orthologs for each pairwise species comparison (Table 5.2A). We found that the ITR proteome was less conserved than the entire and background proteomes, with the latter two being essentially identical. Even more noticeably were the low levels of ITR conservation between species pairs, with an average 21% of ITRs found in a given species present in another one. For example, comparison of the sister species *A. flavus* and *A. oryzae* revealed that 84% of background proteins shared an ortholog, compared to only 75% of ITR-containing proteins (Table

5.2A). Within these ITR-containing proteins, only 56% of ITRs was conserved (monomorphic or variable) (Table 5.2B).

Table 5.2 Evolutionary Conservation of ITRs (A) and ITR-Containing Proteins (B)

	A. flavus	*A. oryzae*	*A. terreus*	*A. niger*	*N. fischeri*	*A. fumigatus*	*A. clavatus*	*A. nidulans*	**Proteome**
									Entire
A. flavus	X								Background
									ITR-containing
	84%								E
A. oryzae	84%	X							B
	75%								I
	73%	69%							E
A. terreus	73%	70%	X						B
	67%	60%							I
	63%	62%	69%						E
A. niger	63%	63%	70%	X					B
	58%	55%	60%						I
	75%	71%	71%	72%					E
N. fischeri	75%	61%	71%	72%	X				B
	72%	61%	62%	64%					I
	74%	72%	71%	72%	86%				E
A. fumigatus	74%	72%	72%	72%	86%	X			B
	65%	58%	59%	65%	83%				I
	79%	76%	76%	78%	86%	82%			E
A. clavatus	80%	77%	77%	78%	87%	83%	X		B
	57%	54%	62%	76%	73%	70%			I
	70%	67%	68%	68%	71%	72%	77%		E
A. nidulans	70%	67%	69%	68%	71%	72%	78%	X	B
	60%	57%	63%	58%	64%	62%	61%		I

Proteome "Entire" or "E" includes all proteins
Proteome "Background" or "B" includes proteins which do not contain an ITR
Proteome "ITR-containing" or "ITR" includes proteins which contain at least one ITR

B. Average ITR conservation

	A. flavus	*A. oryzae*	*A. terreus*	*A. niger*	*N. fischeri*	*A. fumigatus*	*A. clavatus*	*A. nidulans*
A. flavus	x							
A. oryzae	56%	x						
A. terreus	14%	20%	x					
A. niger	16%	22%	23%	x				
N. fischeri	19%	18%	20%	16%	x			
A. fumigatus	19%	20%	25%	16%	47%	x		
A. clavatus	15%	17%	19%	17%	20%	22%	x	
A. nidulans	12%	14%	20%	14%	13%	14%	15%	x

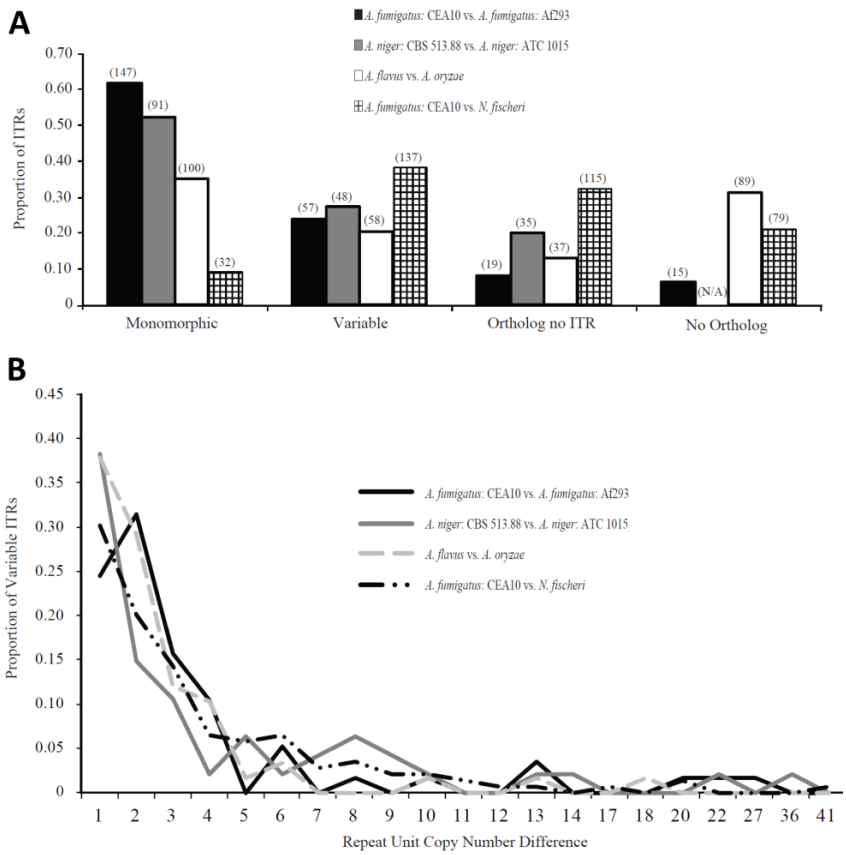
ITR Variation Within and Between Species

ITRs are typically highly variable, within and among species, and it is for this reason that they are frequently used in biomedical and population-based applications (Selkoe and Toonen 2006; Balajee, Tay et al. 2007). To assess ITR variation within and between *Aspergillus* species, we performed two intraspecific (*A. fumigatus*: strain Af293 vs. strain CEA10 and *A. niger*: strain CBS 513.88 vs. strain ATCC 1015) and two interspecific (*A. flavus* vs. *A. oryzae* and *A. fumigatus* vs. *N. fischeri*) comparisons. ITRs were categorized as either MONOMORPHIC (no difference in repeat unit copy number), VARIABLE (difference of at least one repeat unit), ORTHOLOG NO ITR (the ITR was absent in the ortholog) or NO ORTHOLOG.

In the two intraspecific comparisons, as well as the interspecific *A. flavus* vs. *A. oryzae* comparison, approximately 25% of ITRs were in the VARIABLE category, while approximately 40% of ITRs in the *A. fumigatus* vs. *N. fischeri* comparison were categorized as VARIABLE (Figure 5.3A). The most abundant repeat unit variations were single repeat unit differences (~32%), however, repeat unit differences as large as 36 and 41 were identified in the *A. fumigatus* strain comparison and the between-species *A. fumigatus* vs. *N. fischeri* comparison, respectively (Figure 5.3B). The main difference between the intraspecific and interspecific comparisons was that ~32% of ITRs in the *A. fumigatus* vs. *N. fischeri* case belonged to the ORTHOLOG NO ITR category, compared to 13% in the other between species comparison of *A. flavus* and *A. oryzae*, and 8% and 20% in the two within species comparisons (*A. fumigatus* and *A. niger*, respectively) (Figure 5.3A). Additionally, only 9% of ITRs in the *A. fumigatus* vs. *N. fischeri* case

were in the MONOMORPHIC category, compared to 35% in the other between species comparison, and 52% and 62% in the within species (*A. niger* and *A. fumigatus*, respectively) comparisons.

Figure 5.3. ITR variation within and between species. (A) Categorical proportions of ITR-containing genes in two within species and two between species comparisons. Black bars represent *A. fumigatus* strain CEA10 vs. *A. fumigatus* strain Af293, dark gray bars represent *A. niger* strain 513.88 vs. *A. niger* strain ATCC1015, white bars represent *A. flavus* vs. *A. oryzae* and the white mesh bars represent *A. fumigatus* strain CEA10 vs. *N. fischeri*. Numbers in parentheses above bars correspond to the total number of occurrences. The Y axis is the proportion of total ITRs. ITRs were grouped into four categories indicated on the X axis. MONOMORPHIC ITRs showed identical repeat unit copy number whereas VARIABLE ITRs showed a difference of at least one repeat unit copy number. ITRs in the ORTHOLOG NO ITR category consisted of orthologs in which no ITR was present only in one of the two taxa compared. Due to the poor annotation quality the *A. niger* strain ATCC1015, only orthologs with identified ITRs were used in this analysis. ITR-containing genes in one taxon which did not have an identifiable ortholog in the other taxon were placed in the NO ORTHOLOG category (B) ITR repeat unit copy number variation. The X axis is the difference in repeat unit copy number in variable ITRs. The Y axis represents the proportion of total variable ITRs. The solid black line represents the *A. fumigatus* strain CEA10 vs. *A. fumigatus* strain Af293 comparison, the dashed dark gray line represents the *A. niger* strain 513.88 vs. *A. niger* strain ATCC1015 comparison, the dashed light gray line represents the *A. flavus* vs. *A. oryzae* comparison, and the dashed black line represents the *A. fumigatus* strain CEA10 vs. *N. fischeri* comparison. It is important to note that values in the *A. niger* intraspecific comparison are relatively inflated due to exclusion of the NO ORTHOLOG category.



Interestingly, the distribution of ITRs in the MONOMORPHIC and ORTHOLOG NO ITR categories from the interspecific comparison between *A. flavus* and *A. oryzae* were much more similar to the within species cases than to the interspecific comparison between *A. fumigatus* and *N. fischeri* (Figure 5.3A, B). In agreement with several lines of genetic, molecular and genomic data, these observed patterns also suggest that, despite their distinct species status, *A. oryzae* is a domesticated ecotype of *A. flavus* (Kurtzman, Smiley et al. 1986; Geiser, Pitt et al. 1998; Kumeda and Asao 2001; Montiel, Dickinson et al. 2003; Rokas, Payne et al. 2007).

Previous research has indicated a positive association between repeat unit copy number and sequence instability, thereby increasing polymorphism rates (Brohede and Ellegren 1999; Lai, Shinde et al. 2003; Shinde, Lai et al. 2003). Therefore, we examined the relationship between of repeat unit copy number and levels of trinucleotide and hexanucleotide ITR variation, using data from the two interspecific cases. We found no significant difference in average monomorphic and variable repeat unit copy number in both trinucleotide and hexanucleotide repeats (t = -0.899; N = 142; $d.f.$ = 141; p = 0.37 and t = 0.241; N = 101; $d.f.$ = 100; p = 0.81, respectively). However, previous research, focused on intraspecific size variation in microsatellite loci, has indicated that the longest 'pure tract', or uninterrupted repeat tract, may be a more accurate predictor of polymorphism (Lai and Sun 2003; Butland, Devon et al. 2007; Anmarkrud, Kleven et al. 2008). We therefore compared the average longest pure tract, between monomorphic and variable ITRs in the two intraspecific cases. We found that size variable ITRs, had

significantly longer average pure tracts in trinucleotide repeat units (monomorphic = 4.59; variable = 10.43) but not in hexanucleotide repeat units (monomorphic = 2.80; variable = 3.71) compared to monomorphic ITRs (t = 4.22; N = 94; $d.f.$ = 93; p = 5.62e-05 and t = 1.67; N = 56; $d.f.$ = 55; p = 0.10, respectively). However, a larger sample size, drawing from more individuals within-species, may be needed to more confidently assess this relationship in the *Aspergilli*.

Amino Acid Composition of ITR-containing Proteins

To test whether ITR-containing proteins were compositionally distinct from background proteins, we compared amino acid frequencies and mean hydropathy, a measure of a protein's interaction with water, across proteomes. In both analyses, an underlying difference between ITR-containing proteins and background proteins was apparent. Of the twenty amino acids analyzed, only histidine showed no significant difference between the ITR and background proteomes in all eight species (Table 5.3). Twelve of the twenty amino acids (cysteine, glutamine, glutamic acid, isoleucine, leucine, methionine, phenylalanine, proline, serine, threonine, tryptophan, and tyrosine) were significantly differentially distributed across proteomes in all eight species (Table 5.3). Similarly, in each of the eight species, average hydropathy of the ITR and background proteomes differed significantly, with the ITR proteome always being less hydrophobic than the background proteome (Table 5.4).

Table 5.3. Amino Acid Composition of ITR-Containing Proteins

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	Y	W	V
Aspergillus flavus	×	×	×	+	—	+	+	×	×	—	—	+	—	—	+	+	+	—	—	—
Aspergillus oryzae	×	×	×	×	—	+	+	+	×	—	—	+	—	—	+	+	+	—	—	—
Aspergillus terreus	+	×	×	+	—	+	+	×	×	—	—	—	—	—	+	+	+	—	—	—
Aspergillus niger	×	×	×	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Neosartorya fischeri	+	×	—	+	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Aspergillus fumigatus	×	×	×	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—
Aspergillus clavatus	+	—	—	×	—	+	+	+	×	—	—	—	—	—	+	+	+	—	—	×
Aspergillus nidulans	+	×	—	×	—	+	+	×	×	—	—	×	—	—	+	+	+	—	—	—

+ Amino acids that are overrepresented in the ITR-containing proteomes (Bonferroni corrected $P = 0.0003125$).
— Amino acids that are underrepresented in the ITR-containing proteomes (Bonferroni corrected $P = 0.0003125$).
× Amino acids that do not differ in composition between ITR-containing and background proteomes.

Table 5.4. Hydropathy of ITR-Containing Proteome

	Aspergillus flavus		*Aspergillus oryzae*		*Aspergillus terreus*		*Aspergillus niger*		*Neosartorya fischeri*		*Aspergillus fumigatus*		*Aspergillus clavatus*		*Aspergillus nidulans*	
Proteome	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR	Back	ITR
Number of proteins	12,377	210	11,881	182	10,252	154	13,595	317	10,209	194	9,680	207	8,842	278	10,465	200
Average hydropathy	−0.3	−0.6	−0.3	−0.6	−0.3	−0.6	−0.6	−0.8	−0.3	−0.6	−0.3	−0.5	−0.3	−0.8	−0.3	−0.8
P value	*4.62E−20		*8.18E−20		*5.72E−15		*1.81E−10		*4.62E−15		*2.61E−05		*3.15E−05		*7.21E−36	

Proteome "Back" = proteins which do not possess an ITR.
Proteome "ITR" = proteins containing one or more ITRs.
*Significant at Bonferroni corrected *P* value = 0.00625.

Functional Characterization of ITR-containing Proteins

Previous studies have suggested that ITRs may play an important role in fungal pathogenesis by generating structural diversity in cell-surface associated proteins (Verstrepen, Reynolds et al. 2004; Verstrepen, Jansen et al. 2005; Levdansky, Romano et al. 2007). To test this hypothesis we bioinformatically identified signal peptides, indicators of secreted proteins, transmembrane helices, hallmarks of transmembrane proteins, and GPI anchors, molecules attached to some cell surface proteins, in the background and ITR proteomes for each of the 8 species. Signal peptides and GPI anchors were significantly overrepresented in the ITR proteomes while transmembrane helices showed no significant difference across proteomes (Table 5.5). The overrepresentation of signal peptides and GPI anchors in ITR proteomes provides further

support that ITRs may play an active role in cell-surface associated proteins (Hamada, Terashima et al. 1999; Levdansky, Romano et al. 2007).

Table 5.5. Protein Motif Comparison of ITR-Containing Genes and Background Genes

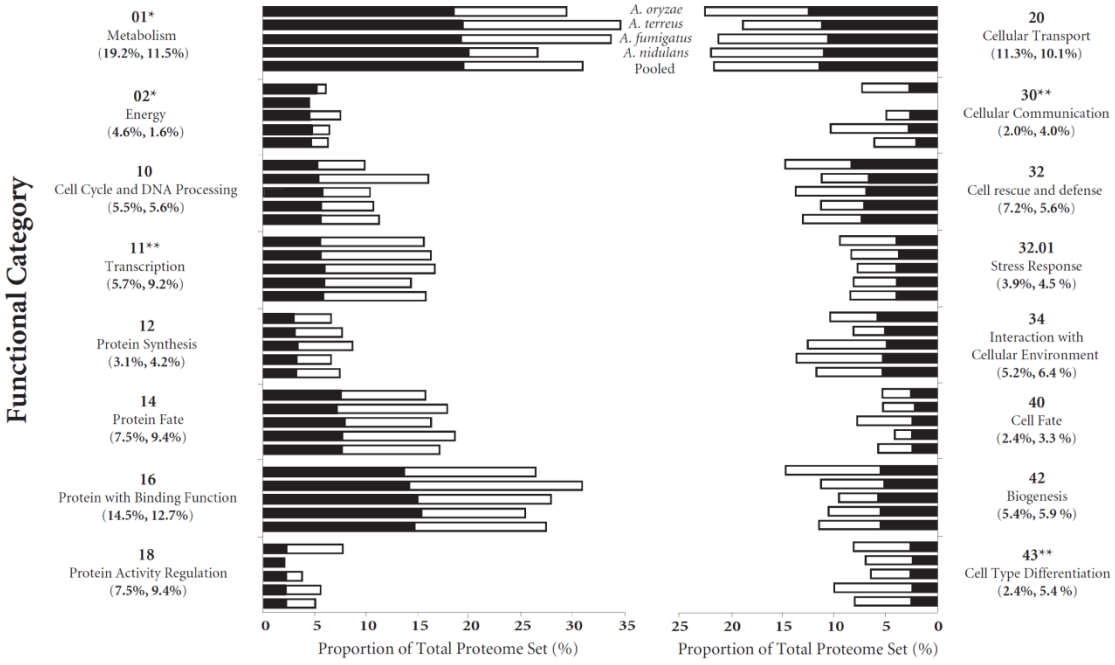
		Signal Peptide				Transmembrane Helix				GPI Anchor			
Species	Proteome	−	+	Prop	*P* Value	−	+	Prop	*P* Value	−	+	Prop	*P* Value
Aspergillus flavus	Back	10,413	1,964	0.189		11,162	1,214	0.109	1.000	12,360	16	0.001	
	ITR	155	55	0.355	**9E−05	191	20	0.105		206	5	0.024	**2E−05
Aspergillus oryzae	Back	10,817	1,064	0.098		10,693	1,188	0.111	0.170	11,864	17	0.001	
	ITR	149	33	0.221	**1E−03	158	24	0.152		179	3	0.017	**1E−v03
Aspergillus terreus	Back	8,629	1,623	0.188		9,205	1,047	0.114	0.504	10,230	22	0.002	
	ITR	121	33	0.273	0.0747	136	18	0.132		150	4	0.027	**5E−04
Aspergillus niger	Back	11,635	1,959	0.168		12,351	1,243	0.101	0.375	13,564	30	0.002	
	ITR	257	61	0.237	*0.0195	294	24	0.082		315	3	0.010	*0.0389
Neosartorya fischeri	Back	8,650	1,560	0.180		9,234	976	0.106	1.000	10,185	25	0.002	
	ITR	147	46	0.313	**2E−03	175	18	0.103		189	4	0.021	**2E−03
Aspergillus fumigatus	Back	8,303	1,377	0.166		8,729	951	0.109	0.195	9,672	8	0.001	
	ITR	160	47	0.294	**1E−03	181	26	0.144		197	10	0.051	**5E−13
Aspergillus clavatus	Back	7,569	1,273	0.168		7,983	859	0.108	0.258	8,819	23	0.003	
	ITR	207	71	0.343	**2E−06	245	33	0.135		270	8	0.030	**3E−06
Aspergillus nidulans	Back	8,894	1,571	0.177		9,450	1,015	0.107	0.717	10,437	28	0.003	
	ITR	159	41	0.258	*0.0362	179	21	0.117		199	1	0.005	0.432

Proteome Back = genes that do not contain the presence of an ITR.
Proteome ITR = genes containing one or more ITR.
"−" = absence of a predicted protein motif.
"+" = presence of predicted protein motif.
Prop = proportion of total proteome set.
*Statistically significant at *P* value = 0.05.
**Statistically significant at Bonferroni *P* value = 0.00625.

We further investigated the functional role of ITRs by examining the occurrences of background and ITR-containing proteins in four of the ten genomes (*A. oryzae*, *A. terreus*, *A. fumigatus*, and *A. nidulans*), according to the FunCat annotation scheme (Ruepp, Zollner et al. 2004). Given the distributional similarities within species (Figure 5.4) and the small percentage of ITR-containing genes, data from the four species was pooled in order to have sufficient data points for reliable statistical analysis. ITR-containing genes were significantly underrepresented in the METABOLISM (p = 3E-5) and ENERGY (p = 0.0015) categories (Figure 5.4). In contrast, ITRs were significantly overrepresented in the TRANSCRIPTION (p = 0.0007), CELLULAR COMMUNICATION/SIGNAL

TRANSDUCTION MECHANISM (p = 0.0073) and CELL TYPE DIFFERENTIATION (p = 0.0007) categories (Figure 5.4).

Figure 5.4. Functional classification of ITR and background proteomes according to the FunCat scheme. The ITR-containing proteome and background proteome proportions of sixteen FunCat categories for *A. oryzae*, *A. terreus*, *A. fumigatus*, *A. nidulans* and the pooled data of all species are reported. The FunCat category number is shown above each FunCat Category. A single star (*) next to the FunCat number represents a statistically significant underrepresentation of ITR-containing proteins whereas a double star (**) represents a statistically significant overrepresentation of ITR-containing proteins. The percentages displayed under each FunCat category are the pooled percentages of the background and ITR proteomes. For each FunCat category and species set, the first black bar is proportion of background proteins belonging to the category, whereas the white bar is the proportion of ITR-containing proteins belonging to the same category.



Interestingly, studies in *Escherichia coli* and *S. cerevisiae* have identified an overrepresentation of microsatellites in stress response genes (Rocha, Matic et al. 2002; Bowen, Roberts et al. 2005). However, there was no association between ITR-containing proteins and the BIOGENESIS OF CELL WALL COMPONENTS (p = 0.66) and STRESS

RESPONSE ($p = 0.53$) categories. Furthermore, examination of the presence of ITR-containing genes in a previously identified *A. fumigatus* gene set that showed differential expression under temperature induced stress (Nierman, Pain et al. 2006) also did not reveal any association ($p = 0.35$).

DISCUSSION

ITRs are frequently associated with genetic disease and pathogenesis (Sherman, Jacobs et al. 1985; Sutherland and Richards 1995; Fondon and Garner 2004; Pearson, Edamura et al. 2005; Verstrepen, Jansen et al. 2005; Mirkin 2007), but also with adaptation to changing environments and phenotypic evolution (Verstrepen, Reynolds et al. 2004; Oh, Cheng et al. 2005; Verstrepen, Jansen et al. 2005; Fidalgo, Barrales et al. 2006; Michael, Park et al. 2007). Nearly two thousand ITRs are distributed throughout the genomes of the eight *Aspergillus* examined in this study. These ITR regions are highly variable, and the proteins containing them are less conserved and compositionally distinct relative to the rest of the proteome. ITR-containing proteins also appear to be functionally distinct. They are more likely to contain signal peptides and GPI anchors, motifs strongly suggestive of functional involvement on or around the cell surface. Furthermore, ITR-containing proteins are preferentially associated with certain cellular processes (transcription, cellular communication and cell-type differentiation) and dissociated from others (metabolism and energy). These results bear on our understanding of the comparative and functional biology of eukaryotic ITRs, as well as the realization of the fungal lifestyle.

The availability of several comparative studies allows us to identify several general features of ITR-containing proteins. ITR abundance in eukaryotic proteomes is not correlated with genome size (Karaoglu, Lee et al. 2005; Huntley and Clark 2007), and ITR content varies extensively between species (Figures 5.1-3) (Huntley and Clark 2007). ITR regions and proteins are, on average, more hydrophilic than the rest (Table 5.4) (Katti, Ranjekar et al. 2001; Kim, Booth et al. 2008). While it has been hypothesized that hydrophilic tandem repeat peptides in regions linking protein domains may produce more tolerated structural formations (Katti, Sami-Subbu et al. 2000), a general explanation explaining the hydrophilic nature of ITR regions is still lacking. Finally, ITR regions are consistently highly variable, both within species as well as between close relatives (Figure 5.3) (Jordan, Snyder et al. 2003; Bowen, Roberts et al. 2005; O'Dushlaine, Edwards et al. 2005; Levdansky, Romano et al. 2007; Kim, Booth et al. 2008). This variation, coupled with the increasing abundance of multiple genomes from a variety of clades, raise the possibility to efficiently identify and develop a suite of clade-wide microsatellite and minisatellite markers that assess variation in taxa separated by hundreds of million years of evolution.

Perhaps more surprisingly, ITR-containing proteins across eukaryotes also share a number of functional features. Most strikingly, ITRs are consistently overrepresented in proteins associated with transcriptional, developmental and signaling processes (Figure 5.4) (Katti, Sami-Subbu et al. 2000; Young, Sloan et al. 2000; Alba and Guigo 2004; O'Dushlaine, Edwards et al. 2005; Huntley and Clark 2007), whereas they are underrepresented in proteins participating in metabolic and housekeeping processes

(Figure 5.4) (Young, Sloan et al. 2000; Huntley and Clark 2007). This conservation of functional association is observed across organisms separated by large evolutionary distances, and persists despite differences across studies in the identification and functional classification of tandem repeats. This enrichment of tandem repeats has been attributed to their general involvement in modulating protein-protein interactions (Hancock and Simon 2005), where slight variations in tandem repeat regions can potentially generate slight variations in the structure of the protein-protein interaction network (King, Soller et al. 1997). Furthermore, the presence and functional role of ITRs in proteins participating in key processes, such as transcription, may also be the explanation as to why human repeat-based disorders are so common and devastating (Gatchel and Zoghbi 2005).

Studies in *S. cerevisiae* and *C. albicans* have shown that ITR variation can modulate the adhesiveness of several cell-surface proteins (Bowen, Roberts et al. 2005; Oh, Cheng et al. 2005; Verstrepen, Jansen et al. 2005), a key trait for understanding fungal pathogenesis and virulence (Oh, Cheng et al. 2005; Verstrepen, Jansen et al. 2005). Several *Aspergillus* species are also capable of colonizing human tissue and causing potentially fatal infections (Patterson, Kirkpatrick et al. 2000; Iversen, Burton et al. 2007). We too found that *Aspergillus* ITR-containing proteins were significantly enriched for cell-surface associated motifs (Figure 5.2, Table 5.5) (Levdansky, Romano et al. 2007), although no association between ITRs and cell-surface proteins could be established on the basis of the FunCat analysis (Figure 5.4). Importantly, the two main adhesion families in *Saccharomyces* and *Candida* (FLO and ALS, respectively) are not

found in *Aspergillus* species (data not shown) (Levdansky, Romano et al. 2007). Instead, the surface of *Aspergillus* is not as clearly understood, but includes a small number of as yet uncharacterized ITR-containing proteins (Levdansky, Romano et al. 2007).

The relationship between ITRs and variation at the fungal cell surface notwithstanding, the enrichment of several functional processes with ITR-containing proteins suggests that their role in fungi might be more diverse than previously thought (Verstrepen, Jansen et al. 2005; Levdansky, Romano et al. 2007). Experimental (Michael, Park et al. 2007; Paoletti, Saupe et al. 2007) and bioinformatic evidence (this study) both suggest that ITR variability may be key in a wide variety of physiological and developmentally important processes. For example, ITR variation in the *Neurospora crassa* protein WC-1 likely plays an important role in regulating the organism's circadian clock behavior in response to environmental cues (Michael, Park et al. 2007). In *Aspergillus* (and other filamentous fungi), an important role for ITR variation may be in the control of self / nonself recognition during somatic cell fusion (Paoletti, Saupe et al. 2007). We noted that several ITR-containing proteins in the *Aspergillus* possessed PFAM domains characteristic of HET proteins. The HET protein family is thought to control self / nonself recognition across filamentous fungi (Espagne, Balhadere et al. 2002; Paoletti, Saupe et al. 2007), with repeat variation in these proteins playing a key role in establishing recognition specificity (Paoletti, Saupe et al. 2007).

ACKNOWLEDGEMENTS AND CONTRIBUTIONS

We thank Melanie Huntley for providing details of her published experimental protocol on the placement of tandem repeats. We also thank Stefanie Butland for providing insight into her methods of calculating longest pure tract repeat. JGG is funded by the Graduate Program in Biological Sciences at Vanderbilt University. Research in AR's lab is supported by the Searle Scholars Program and Vanderbilt University.

J.G.G. and A.R. designed the study. J.G.G. analyzed the data.

REFERENCES

. "The Broad Insitute's *Aspergillus* Comparative Database." from http://www.broad.mit.edu/annotation/genome/aspergillus_terreus/MultiHome.html.

"Munich information center for protein sequences PEDANT 3 database (MIPS)."

Alba, M. M. and R. Guigo (2004). "Comparative analysis of amino acid repeats in rodents and humans." Genome Res **14**(4): 549-54.

Anmarkrud, J. A., O. Kleven, et al. (2008). "Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus HrU10." Bmc Evolutionary Biology **8**: -.

Balajee, S. A., S. T. Tay, et al. (2007). "Characterization of a novel gene for strain typing reveals substructuring of Aspergillus fumigatus across north America." Eukaryotic Cell **6**(8): 1392-1399.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." Journal of Molecular Biology **340**(4): 783-795.

Bichara, M., J. Wagner, et al. (2006). "Mechanisms of tandem repeat instability in bacteria." Mutat Res **598**(1-2): 144-63.

Bowen, S., C. Roberts, et al. (2005). "Patterns of polymorphism and divergence in stress-related yeast proteins." Yeast **22**(8): 659-668.

Brohede, J. and H. Ellegren (1999). "Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences." Proceedings of the Royal Society of London Series B-Biological Sciences **266**(1421): 825-833.

Butland, S. L., R. S. Devon, et al. (2007). "CAG-encoded polyglutamine length polymorphism in the human genome." Bmc Genomics **8**: 126.

Dieringer, D. and C. Schlotterer (2003). "Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species." Genome Research **13**(10): 2242-2251.

Drake, J. W., B. Charlesworth, et al. (1998). "Rates of spontaneous mutation." Genetics **148**(4): 1667-1686.

Eisenhaber, B., G. Schneider, et al. (2004). "A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and Schizosaccharomyces pombe." J Mol Biol **337**(2): 243-53.

Ellegren, H. (2004). "Microsatellites: Simple sequences with complex evolution." Nature Reviews Genetics **5**(6): 435-445.

Espagne, E., P. Balhadere, et al. (2002). "HET-E and HET-D belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus Podospora anserina." Genetics **161**(1): 71-81.

Fedorova, N. D., N. Khaldi, et al. (2008). "Genomic Islands in the Pathogenic Filamentous Fungus Aspergillus fumigatus." PLoS Genet **4**(4): e1000046.

Fidalgo, M., R. R. Barrales, et al. (2006). "Adaptive evolution by mutations in the FLO11 gene." Proceedings of the National Academy of Sciences of the United States of America **103**(30): 11228-11233.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-51.
Fondon, J. W. and H. R. Garner (2004). "Molecular origins of rapid and continuous morphological evolution." Proceedings of the National Academy of Sciences of the United States of America **101**(52): 18058-18063.
Frenkel, S. and E. Z. Blumenthal (2002). "Jmp in, Ver 4." Jama-Journal of the American Medical Association **287**(13): 1733-1734.
Galagan, J. E., S. E. Calvo, et al. (2005). "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." Nature **438**(7071): 1105-15.
Gatchel, J. R. and H. Y. Zoghbi (2005). "Diseases of unstable repeat expansion: Mechanisms and common principles." Nature Reviews Genetics **6**(10): 743-755.
Geiser, D. M., J. I. Pitt, et al. (1998). "Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus." Proceedings of the National Academy of Sciences of the United States of America **95**(1): 388-393.
Hamada, K., H. Terashima, et al. (1999). "Amino acid residues in the omega-minus region participate in cellular localization of yeast glycosylphosphatidylinositol-attached proteins." Journal of Bacteriology **181**(13): 3886-3889.
Hancock, J. M. and M. Simon (2005). "Simple sequence repeats in proteins and their significance for network evolution." Gene **345**(1): 113-118.
Huntley, M. A. and A. G. Clark (2007). "Evolutionary analysis of amino acid repeats across the genomes of 12 drosophila species." Molecular Biology and Evolution **24**(12): 2598-2609.
Iversen, M., C. M. Burton, et al. (2007). "Aspergillus infection in lung transplant patients: incidence and prognosis." European Journal of Clinical Microbiology & Infectious Diseases **26**(12): 879-886.
Jordan, P., L. A. S. Snyder, et al. (2003). "Diversity in coding tandem repeats in related Neisseria spp." Bmc Microbiology **3**: 23.
Karaoglu, H., C. M. Lee, et al. (2005). "Survey of simple sequence repeats in completed fungal genomes." Mol Biol Evol **22**(3): 639-49.
Kashi, Y. and D. G. King (2006). "Simple sequence repeats as advantageous mutators in evolution." Trends Genet **22**(5): 253-9.
Katti, M. V., P. K. Ranjekar, et al. (2001). "Differential distribution of simple sequence repeats in eukaryotic genome sequences." Mol Biol Evol **18**(7): 1161-7.
Katti, M. V., R. Sami-Subbu, et al. (2000). "Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications." Protein Science **9**(6): 1203-1209.
Kim, T. S., J. G. Booth, et al. (2008). "Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference." Bmc Genomics **9**: 31.
King, D. G., M. Soller, et al. (1997). "Evolutionary tuning knobs." Endeavour **21**(1): 36-40.
Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes." Journal of Molecular Biology **305**(3): 567-580.

Kumeda, Y. and T. Asao (2001). "Heteroduplex panel analysis, a novel method for genetic identification of Aspergillus Section Flavi strains." Applied and Environmental Microbiology **67**(9): 4084-4090.
Kurtzman, C. P., M. J. Smiley, et al. (1986). "DNA Relatedness among Wild and Domesticated Species in the Aspergillus-Flavus Group." Mycologia **78**(6): 955-959.
Kyte, J. and R. F. Doolittle (1982). "A Simple Method for Displaying the Hydropathic Character of a Protein." Journal of Molecular Biology **157**(1): 105-132.
Lai, Y. L., D. Shinde, et al. (2003). "The mutation process of microsatellites during the polymerase chain reaction." Journal of Computational Biology **10**(2): 143-155.
Lai, Y. L. and F. Z. Sun (2003). "The relationship between microsatellite slippage mutation rate and the number of repeat units." Molecular Biology and Evolution **20**(12): 2123-2131.
Levdansky, E., J. Romano, et al. (2007). "Coding tandem repeats generate diversity in *Aspergillus* fumigatus genes." Eukaryotic Cell **6**(8): 1380-1391.
Levinson, G. and G. A. Gutman (1987). "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." Mol Biol Evol **4**(3): 203-21.
Li, Y. C., A. B. Korol, et al. (2004). "Microsatellites within genes: Structure, function, and evolution." Molecular Biology and Evolution **21**(6): 991-1007.
Machida, M., K. Asai, et al. (2005). "Genome sequencing and analysis of Aspergillus oryzae." Nature **438**(7071): 1157-61.
Metzgar, D., J. Bytof, et al. (2000). "Selection against frameshift mutations limits microsatellite expansion in coding DNA." Genome Research **10**(1): 72-80.
Michael, T. P., S. Park, et al. (2007). "Simple sequence repeats provide a substrate for phenotypic variation in the Neurospora crassa circadian clock." PLoS ONE **2**(8): e795.
Mirkin, S. M. (2007). "Expandable DNA repeats and human disease." Nature **447**(7147): 932-40.
Montiel, D., M. J. Dickinson, et al. (2003). "Genetic differentiation of the Aspergillus section Flavi complex using AFLP fingerprints." Mycological Research **107**: 1427-1434.
Moreno-Hagelsieb, G. and K. Latimer (2008). "Choosing BLAST options for better detection of orthologs as reciprocal best hits." Bioinformatics **24**(3): 319-324.
Moxon, R., C. Bayliss, et al. (2006). "Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation." Annu Rev Genet **40**: 307-33.
Nierman, W. C., A. Pain, et al. (2006). "Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus (vol 438, pg 1151, 2005)." Nature **439**(7075): 1151-1156.
O'Dushlaine, C. T., R. J. Edwards, et al. (2005). "Tandem repeat copy-number variation in protein-coding regions of human genes." Genome Biology **6**(8): R69.
Oh, S. H., G. Cheng, et al. (2005). "Functional specificity of Candida albicans Als3p proteins and clade specificity of ALS3 alleles discriminated by the number of copies of the tandem repeat sequence in the central domain." Microbiology-Sgm **151**: 673-681.
Paoletti, M., S. J. Saupe, et al. (2007). "Genesis of a fungal non-self recognition repertoire." PLoS ONE **2**(3): e283.

Patterson, T. F., W. R. Kirkpatrick, et al. (2000). "Invasive aspergillosis. Disease spectrum, treatment practices, and outcomes. I3 Aspergillus Study Group." Medicine (Baltimore) **79**(4): 250-60.
Pearson, C. E., K. N. Edamura, et al. (2005). "Repeat instability: Mechanisms of dynamic mutations." Nature Reviews Genetics **6**(10): 729-742.
Pel, H. J., J. H. de Winde, et al. (2007). "Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88." Nat Biotechnol **25**(2): 221-31.
Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.
Rocha, E. P. C., I. Matic, et al. (2002). "Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions?" Nucleic Acids Research **30**(9): 1886-1894.
Rokas, A., Galagan, J. (2007). Aspergillus nidulans Genome and a Comparative Analysis of Genome Evolution in Aspergillus. The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods. G. H. Goldman, Osmani, S.A. New York, CRC Press**:** 43-56.
Rokas, A., G. Payne, et al. (2007). "What can comparative genomics tell us about species concepts in the genus Aspergillus?" Stud Mycol **59**: 11-7.
Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.
Sawyer, L. A., J. M. Hennessy, et al. (1997). "Natural variation in a Drosophila clock gene and temperature compensation." Science **278**(5346): 2117-2120.
Schilling, G., A. H. Sharp, et al. (1995). "Expression of the Huntingtons-Disease (It15) Protein Product in Hd Patients." Human Molecular Genetics **4**(8): 1365-1371.
Schlotterer, C. and D. Tautz (1994). "Chromosomal Homogeneity of Drosophila Ribosomal DNA Arrays Suggests Intrachromosomal Exchanges Drive Concerted Evolution." Current Biology **4**(9): 777-783.
Selkoe, K. A. and R. J. Toonen (2006). "Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers." Ecology Letters **9**(5): 615-629.
Sherman, S. L., P. A. Jacobs, et al. (1985). "Further Segregation Analysis of the Fragile X-Syndrome with Special Reference to Transmitting Males - Reply." Human Genetics **71**(2): 183-183.
Shinde, D., Y. L. Lai, et al. (2003). "Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites." Nucleic Acids Research **31**(3): 974-980.
Siwach, P., S. D. Pophaly, et al. (2006). "Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats." Mol Biol Evol **23**(7): 1357-69.
Sokal, R. R. and F. J. Rohlf (1995). Biometry : the principles and practice of statistics in biological research. New York, W.H. Freeman.
Sutherland, G. R. and R. I. Richards (1995). "Simple Tandem DNA Repeats and Human Genetic-Disease." Proceedings of the National Academy of Sciences of the United States of America **92**(9): 3636-3641.
Thomas, E. E. (2005). "Short, local duplications in eukaryotic genomes." Current Opinion in Genetics & Development **15**(6): 640-644.

Toth, G., Z. Gaspari, et al. (2000). "Microsatellites in different eukaryotic genomes: Survey and analysis." Genome Research **10**(7): 967-981.
van der Woude, M. W. and A. J. Baumler (2004). "Phase and antigenic variation in bacteria." Clinical Microbiology Reviews **17**(3): 581-611.
Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.
Verstrepen, K. J., A. Jansen, et al. (2005). "Intragenic tandem repeats generate functional variability." Nature Genetics **37**(9): 986-990.
Verstrepen, K. J., T. B. Reynolds, et al. (2004). "Origins of variation in the fungal cell surface." Nature Reviews Microbiology **2**(7): 533-540.
Weber, J. L. and C. Wong (1993). "Mutation of Human Short Tandem Repeats." Human Molecular Genetics **2**(8): 1123-1128.
Young, E. T., J. S. Sloan, et al. (2000). "Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae." Genetics **154**(3): 1053-68.
Yu, J., T. E. Cleveland, et al. (2005). "Aspergillus flavus genomics: gateway to human and animal health, food safety, and crop resistance to diseases." Rev Iberoam Micol **22**(4): 194-202.
Yu, J., C. A. Whitelaw, et al. (2004). "Aspergillus flavus expressed sequence tags for identification of genes with putative roles in aflatoxin contamination of crops." FEMS Microbiol Lett **237**(2): 333-40.

CHAPTER VI

ASSESSING THE GENOME-WIDE EFFECT OF PROMOTER REGION TANDEM REPEAT NATURAL VARIATION ON GENE EXPRESSION

John G. Gibbons[1*], Martha H. Elmore[1*] and Antonis Rokas[1]

[*] *These authors contributed equally to this work*

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA*

ABSTRACT

Small nucleotide tandem repeats (TRs), such as microsatellites and minisatellites, are both highly abundant and copy number variable in eukaryotic genomes. The mutational reversibility of TR copy number polymorphisms coupled with studies linking TR polymorphism to phenotypic variation, have led some to suggest that TR variation is a modulator of, and major contributor to, phenotypic variation. However, studies that assess the genome-wide impact of TR variation on phenotype are so far lacking. To address this question, we genotyped 143 TRs located in promoter regions genome-wide across 16 isolates of the closely related species *Aspergillus oryzae* and *A. flavus* and tested their relationship with the expression level of their downstream genes as measured by RNA-seq. Regression analysis between variation in TR copy number and gene expression showed a significant relationship in only 4.3% of comparisons. Significant relationships did not conform to a single pattern but to several different ones, and were consistent with the volume knob and tuning knob models but not with the expression switch model. Furthermore, we found no evidence that promoter region TRs contribute to greater levels of expression variability at a given locus. Although natural variation in TR copy number likely contributes, in some cases, to transcript abundance variation, our results suggest that the prevalence of TRs in facilitating the fine-tuning of phenotypes is unlikely to be as common as previously suggested.

INTRODUCTION

Short stretches of tandemly repeated nucleotides (TRs) are ubiquitous in eukaryotic genomes and occur in both coding and non-coding regions (Li, Korol et al. 2002). TRs can consist of short repeat units or copies made up of 1-9 bp, which are typically classified as microsatellites, as well as of longer copies that can be hundreds of base pairs long, which are classified as minisatellites. Microsatellites have a tendency to expand or contract during replication (slip-strand mispairing) (Levinson and Gutman 1987), whereas minisatellite variability is driven by recombination during meiosis (Hancock 1999). These underlying mutational events generate changes in the copy number of TRs and occur 100 to 10,000 times more often than single nucleotide substitutions (Lynch, Sung et al. 2008), creating local mutational hotspots in the genome. Importantly, TR mutational events typically result in the addition or subtraction of one or few copies and, in stark contrast to the more classic insertion, deletion and point mutation events, are often reversible (Kashi and King 2006).

Owing to their abundance, high variability, and presumed selective neutrality, in the last three decades TRs have been used as markers in countless evolutionary genetics and epidemiological studies (Goldstein and Schlötterer 1999). However, it has become increasingly evident that in some cases TR variation may not be neutral and can directly alter phenotype both in coding (Sawyer, Hennessy et al. 1997; Fondon and Garner 2004; Verstrepen, Jansen et al. 2005) and non-coding regions (Hammock, Lim et al. 2005; Rockman, Hahn et al. 2005; Vinces, Legendre et al. 2009). For example, TR length in the coding regions of different circadian rhythm genes in the non-migratory bird *Cyanistes*

caeruleus (Johnsen 2007), the fruit fly *Drosophila melanogaster* (Sawyer, Hennessy et al. 1997) and the filamentous fungi *Neurospora crassa* (Michael, Park et al. 2007), is directly involved in fine-tuning rhythm periodicity. Similarly, polymorphisms in TR length in *cis*-regulatory regions have been linked to the modification of gene expression levels in humans (Bennett, Lucassen et al. 1995), voles (Hammock and Young 2004), fish (Streelman and Kocher 2002) and fungi (Staib, Kretschmar et al. 2002; Vinces, Legendre et al. 2009). In one of the best documented examples, experimental modification of TR copy number in the promoter of the *MET3* and *SDT1* genes in *Saccharomyces cerevisiae* altered gene expression in a Gaussian-like pattern; expression was relatively low at short TR lengths, reached maximum level at intermediate TR lengths, and then reduced again as TR length increased (Vinces, Legendre et al. 2009).

The fact that TRs are a rich source of abundant, continuous and potentially reversible mutations coupled with their involvement in modulating phenotype in several case studies, has led to the hypothesis that TRs act as "evolutionary tuning knobs" of molecular (e.g., gene expression) and organismal (e.g., circadian rhythm periodicity) phenotypes (Kashi, King et al. 1997; King, Soller et al. 1997; Sawyer, Hennessy et al. 1997; Fondon and Garner 2004; Hammock and Young 2004; Vinces, Legendre et al. 2009). Although TR variation can affect phenotype in a variety of different ways, three different models have been proposed to explain the relationship between TR variation and gene expression; the "volume knob" model, the "tuning knob" model, and the "expression switch" model. The volume knob model predicts that the phenotype is negatively or positively associated with TR allele length (Schilling, Sharp et al. 1995),

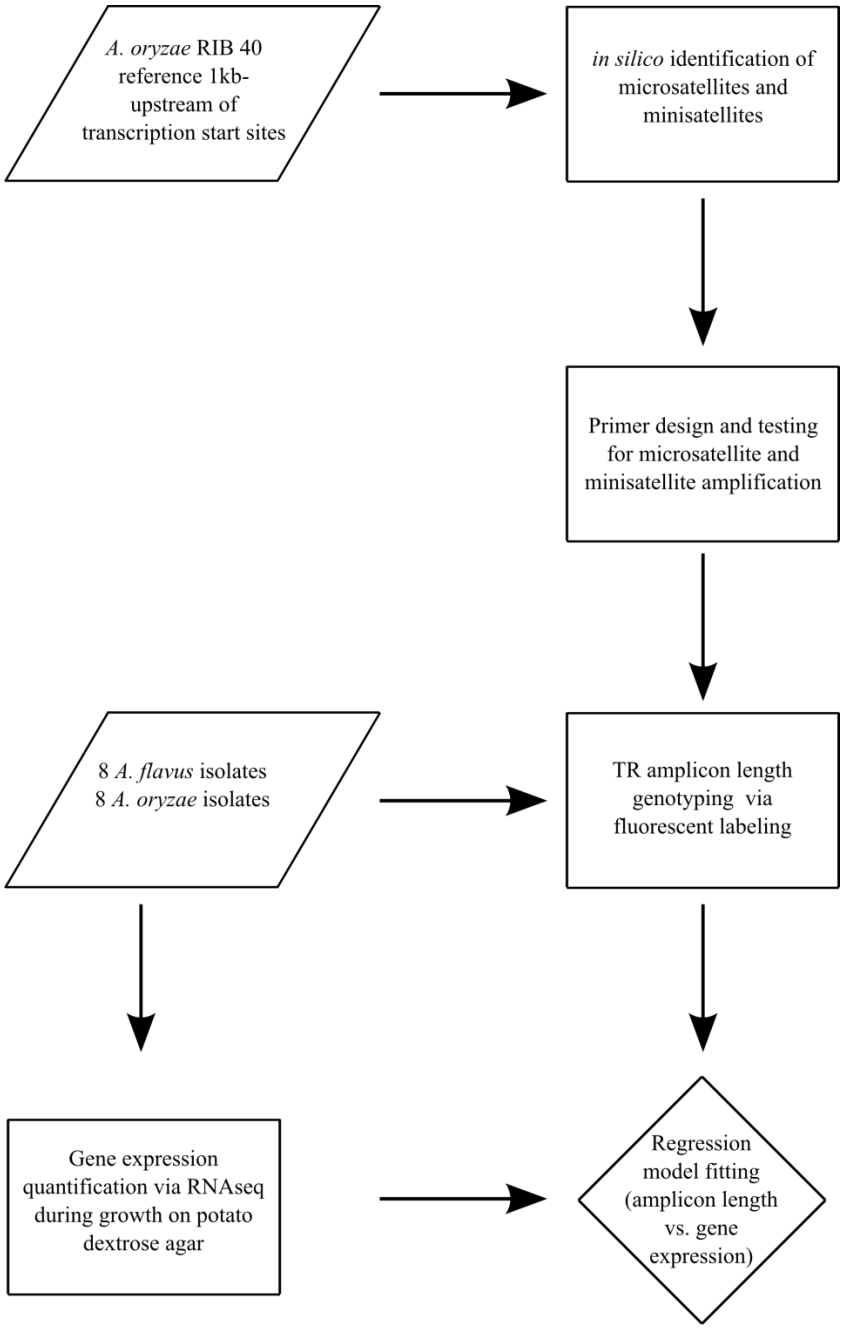
whereas in the tuning knob model predicts that TR allele length is associated with nonlinear changes to gene expression (Vinces, Legendre et al. 2009) analogous to a radio tuning dial. Finally, in the expression switch model, TR lengths beyond a certain threshold are predicted to act as on / off switches of gene expression (Chen, Liu et al. 2010).

To systematically evaluate the functional effect of TR variation on genome-wide gene activity, we measured TR allele length across 143 promoter regions in 16 isolates of the closely related fungal species *Aspergillus flavus* and its domesticated ecotype *A. oryzae* (Machida, Asai et al. 2005; Gibbons, Salichos et al. 2012) and compared these genotypes to their respective gene expression levels. Although we identified several significant patterns between promoter region TR length and gene expression, the vast majority of the loci examined did not show any relationship. Moreover, we found no evidence that genes containing promoter regions containing TRs have elevated levels of expression variance. Contrary to suggestions that TR variation is a major modulator of eukaryotic phenotypes, these results suggest that the vast majority of TRs exerts no measurable effect on phenotype.

MATERIALS AND METHODS

The experimental design is depicted in Figure 6.1.

Figure 6.1. Experimental design.



Identification of TRs in Promoter Regions

We defined promoter regions as the non-coding 1000 bp region upstream of annotated start codons. The EMBOSS etandem software (Rice, Longden et al. 2000) was used to identify TRs in the promoter regions of the *A. oryzae* RIB 40 reference genome (Machida, Asai et al. 2005). The conservation of A. oryzae TRs in the *A. flavus* genome was validated by checking for their presence in the corresponding orthologous region of the *A. flavus* reference strain NRRL 3357 genome. We defined microsatellites as sequence repeats between 2-9 bp and minisatellites as sequence repeats ≥ 10 bp. We considered TRs as significant if the repeat unit consensus sequence conservation was ≥ 90%.

Fungal Isolates and Nucleic Acid Extraction

We analyzed eight isolates of *A. oryzae* (RIB 333, RIB 642, RIB 331, RIB 302, RIB 40, RIB 537, RIB 632 and RIB 949) and eight isolates of *A. flavus* (SRRC 1357, SRRC 2112, SRRC 2632, SRRC 2524, SRRC 2653, SRRC 2114, SRRC 1273 and NRRL 3357) (Gibbons, Salichos et al. 2012). or DNA extraction, spores were inoculated in potato dextrose broth and grown for three days at room temperature in a tissue rotator at which point mycelium was harvested and ground in liquid nitrogen. gDNA was extracted using a basic CTAB protocol (Stewart and Via 1993). For RNA extraction, 500 µl of a water conidial suspension (10^7/ml) was spread onto a potato dextrose agar plate covered with a layer of sterile porous cellophane and grown at 30˚C for 24 hours. Mycelium was harvested with a metal spatula, flash frozen in liquid nitrogen and stored at -80˚C. Mycelium was ground with a mortar and pestle in liquid nitrogen. Total RNA was

extracted using TRIzol (Life Technologies), DNased then cleaned with an RNeasy column (Qiagen) according to the manufacturer's instructions. RNA-seq libraries were constructed and sequenced at the Vanderbilt Genome Sciences Resource as previously described (Gibbons, Beauvais et al. 2012; Gibbons, Salichos et al. 2012).

Primer Design

Primer pairs targeting the microsatellites and minisatellites were designed using Primer3 (Rozen and Skaletsky 2000). Forward and reverse primers were designed around the target region so that the amplicon would be between 150-450 bp. We incorporated a dinucleotide GC clamp into primer pairs when possible and a M13 tag to the 5' end of all forward primers for use in downstream fluorescent genotyping (Schuelke 2000).

TR Genotyping

We genotyped the 16 isolates across 72 microsatellite and 71 minisatellite loci following the fluorescent amplicon labeling approach of Schuelke (2000). A touchdown PCR protocol (Don, Cox et al. 1991) was implemented to limit nonspecific amplification and consisted of the following cycling profile: 95ºC for 3 min, 11 cycles of 94ºC for 30 s, 65ºC for 30 s (with annealing temperature dropping 1ºC per cycle) and 72ºC for 45s, followed by 29 cycles of 94ºC for 30 s, 53ºC for 30 s, 72ºC for 45 s, followed by a final extension of 72ºC for 20 min. PCR products were sized on an ABI 3730xl Genetic Analyzer at Genewiz (South Plainfield, NJ), using the LIZ500 size standard. Amplicon lengths were called using the Peak Scanner Software v1.0 (ABI). Because of the difficulties in extracting repeat unit number directly from amplicon length (Guichoux,

Lagache et al. 2011) we used Normalized Fragment Length (NAL) to distinguish TR alleles:

$$NAL = \text{amplicon length} / \text{length of smallest amplicon at locus}$$

Gene Expression Quantification

Gene expression levels were quantified in terms of Reads Per Kilobase of transcript per Million mapped reads (RPKM) (Mortazavi, Williams et al. 2008) using the rSeq package ((Jiang and Wong 2009) as previously described (Gibbons, Beauvais et al. 2012; Gibbons, Salichos et al. 2012).

Testing the Role of TRs as Functional "Knobs"

We investigated the "knob" relationship between TR allele length (NAL) and gene expression (RPKM) at each locus by testing their relationship for each species separately as well as combined. Specifically, we tested each comparison to untransformed *linear*, *quadratic* and *cubic* models as well as to linear models with four different RPKM (Y-axis) transformations (*logarithmic*: natural logarithm (RPKM); *square root*: $\sqrt{RPKM}$; *squared*: $RPKM^2$; and *reciprocal*: 1 / RPKM). The optimal regression model for each locus was assessed by choosing the model with the smallest sample size corrected Akaike's Information Criterion (AICc) value:

$$AICc = \ln(RSS) + 2K + (2K*(K+1))/(n-K-1)$$

where RSS is the regression residual sum of squares, K equals the number of parameters in the model and n is the number of observations (Akaike 1974). For each optimal regression model, *P-values* were calculated from the *F-ratio*, and results were reported

using a significance threshold of *P-value* < 0.01 as well as a Bonferroni multiple test corrected *P-value* cutoff of 0.000126.

Testing the Role of TRs as Functional "Switches"

We also tested whether TRs could function as expression "switches". If so, we would expect to see differences in expression levels between alleles of a given locus. We tested the utility of TRs as functional "switches" by comparing the expression levels of loci harboring multiple alleles with frequencies ≥ 0.25 in *A. oryzae* using either a T-test (for 2 alleles) or ANOVA (for 3 alleles). Results were reported using a significance threshold cutoff of *P-value* < 0.01 as well as a Bonferroni multiple test corrected *P-value* cutoff of 0.00135.

Testing the Role of TRs as Functional "Noise Makers"

To evaluate whether TR variation contributes to variation or "noise" in gene expression we tested whether genes containing TRs in their promoter regions had greater expression variance than two sets of background genes. First, to control for genome architecture effects, we compared the set of TR-containing genes against a background set comprised of genes lacking TRs in their respective promoters found two genes upstream and two genes downstream of TR-containing genes. Second, to control for functional effects, we compared the set of TR-containing genes against 10 different background sets of genes with the same functional classifications (according to the FunCat Annotation Database; (Ruepp, Zollner et al. 2004)). In cases where a single gene was classified in more than

one FunCat category, we randomly selected a single category. We tested the statistical significance of all comparisons using Tukey-Kramer post-hoc ANOVA tests.

TR Representation in Interspecific Differentially Expressed Genes

If TRs play an important role in rapid regulatory evolution(Fondon and Garner 2004), we would expect TRs to be overrepresented in the promoter regions of genes differentially expressed between populations or closely related species. To test this hypothesis, we compared the frequencies of promoter region TRs between the subset of *A. oryzae* and *A. flavus* differentially expressed and non-differentially expressed genes using a Fisher's exact test. The differentially expressed gene set was determined by comparing the expression levels of each gene between the 8 *A. oryzae* and 8 *A. flavus* isolates via T-tests with a *P-value* < 0.01 cutoff (185 total genes).

All statistical analyses were performed in JMP version 9 (http://www.jmp.com/).

RESULTS

Distribution and Genotyping of Promoter Region TRs

We identified 228 TRs in 190 general promoter regions (1 kb upstream of start codons) of the *A. oryzae* RIB 40 reference genome (Machida et al 2005), several of which contained both microsatellites and minisatellites. Specifically, 127 microsatellites and 101 minisatellites were identified in 125 and 99 promoter regions, respectively. 57% (72/127) and 70% (71/101) of these microsatellite and minisatellite containing regions were successfully genotyped, respectively.

Patterns of TR Allele Length and Expression Variance

From the 72 promoter region microsatellites and 71 minisatellites, 11 and 6 microsatellite loci and 6 and 1 minisatellite loci showed no variation in *A. oryzae* and *A. flavus*, respectively. Of the 72 microsatellite loci, 18 had significantly reduced allelic variance in *A. oryzae* and only 3 loci had reduced variance in *A. flavus* (F-test; $P < 0.0007$). Of the 71 minisatellite loci, 11 and 9 had reduced allelic variance in *A. oryzae* and *A. flavus*, respectively (F-test; $P < 0.0007$). The reduction of allelic variation in *A. oryzae* mirrors the overall reduction of genetic variation in these isolates compared to *A. flavus* (Gibbons, Salichos et al. 2012). Of the 123 genes for which we analyzed gene expression (several genes had multiple TRs in their promoter), 4 and 7 had reduced expression variance in *A. oryzae* and *A. flavus*, respectively (F-test; $P < 0.0004$).

Functional Associations of Genes Containing Promoter Region TRs

We investigated the functional associations of genes containing promoter region TRs by comparing their FunCat classification (Ruepp et al. 2004) to that of all genes lacking TRs in their promoter regions. Genes with regulatory region TRs were significantly overrepresented in the *phosphate metabolism* (Fisher's Exact Test (FET); $P = 0.036$), *stress response* ($P = 0.023$), *cell growth / morphogenesis* ($P = 0.041$) and *bud/ growth tip* ($P = 0.038$) FunCat categories and were under represented in the *protein binding* ($P = 0.008$) category. Furthermore, the subset of genes showing significant relationships between TR length and gene expression were overrepresented in the *nucleotide/nucleoside/nucleobase binding* category ($P = 0.047$).

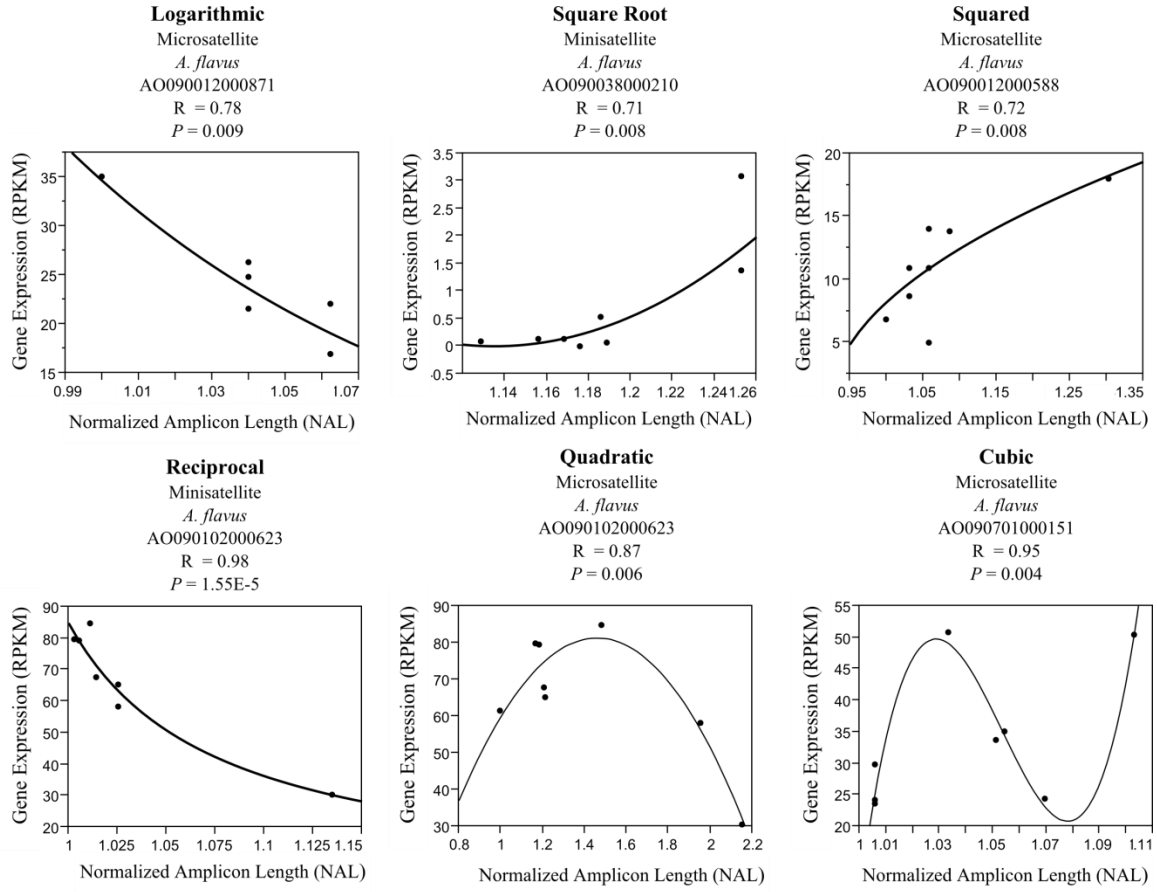
TRs are not Overrepresented in the Promoter Regions of Differentially Expressed Genes

It has been hypothesized that TR variation in cis-regulatory regions may be a major source of rapid phenotypic evolution (Fondon and Garner 2004). If this was a general trend, we would expect to TRs overrepresented in genes differentially expressed between populations or species. We tested this hypothesis by comparing the frequencies of promoter region TRs in differentially expressed and non-differentially expressed genes between *A. oryzae* and *A. flavus*. 2.2% (4 of 185) and 1.6% (186 of 11,877) of differentially expressed and non-differentially expressed genes, respectively, contained promoter region TRs. TRs were not significantly overrepresented in the differentially expressed gene set (Fisher's Exact Test; *P-value* = 0.55).

Promoter Region TRs are Infrequently Associated with Expression "Knob" Functions

We investigated the "knob" relationship between promoter region TR allele length (NAL) and gene expression (RPKM) both by species and combined, using regression models reflecting biologically relevant patterns. Specifically, the linear models correspond to a "volume knob" pattern in which TR copy number changes directly correlate with gene expression levels, the quadratic model to an "optimality knob" in which gene expression increases up to a point and then decreases with corresponding copy number increase, and the cubic model to a "tuning knob" in which gene expression oscillates in step with TR variation (Figure 6.2).

Figure 6.2. Examples of significant TR length vs. gene expression patterns. For each example, the regression fit, TR type (microsatellite or minisatellite), comparison type (*A. oryzae*, *A. flavus* or combined), locus identifier, regression R^2, and regression *P*-values are provided. Each plot shows the TR allele length (NAL) on the X-axis and gene expression level (RPKM) on the Y-axis.



In total, only 4.3% (17 of 396) of comparisons showed a significant relationship (Table 6.1) at the *P-value* < 0.01 level while only 1 comparison withstood the Bonferroni corrected *P-value* < 0.000126. We identified 6, 1 and 2 significant relationships within the microsatellite data and 3, 3 and 2 significant relationships within the minisatellite data in *A. flavus*, *A. oryzae* and the combined data, respectively. These relationships fit several

different regression models (1 *linear*, 3 *quadratic*, 8 *cubic*, 1 *logarithmic*, 1 *square root*, 1 *squared* and 2 *reciprocal*) (Figure 6.2).

Table 6.1. Genes containing promoter region TRs with significant expression patterns.

	Species	Gene Promoter	Function	Best Fit	R^2	*P*-value
Microsatellites	*A. flavus*	AO090012000588	SNF2 family helicase/ATPase	Squared	0.72	0.008
		AO090012000871	PAP2 superfamily	Logarithmic	0.78	0.009
		AO090102000623	HLH transcription factor	Quadratic	0.87	0.006
		AO090206000041	F-box domain	Cubic	0.94	0.007
		AO090701000151	Growth-arrest-specific protein 2 domain	Cubic	0.95	0.004
		AO090701000375	RhoGAP domain	Quadratic	0.96	0.002
	A. oryzae	AO090005000013	Uncharacterized protein	Cubic	0.96	1.00E-04
	Combined	AO090003000121	6-phosphogluconate dehydrogenase	Cubic	0.67	0.009
		AO090012000871	PAP2 domain-containing protein	Cubic	0.64	0.009
Minisatellites	*A. flavus*	AO090005000959	Hypothetical protein	Reciprocal	0.83	0.002
		AO090038000210	Polyketide synthase	Square Root	0.71	0.008
		AO090102000623	HLH transcription factor	Reciprocal	0.98	1.55E-05
	A. oryzae	AO090005000567	Hypothetical protein	Cubic	1.00	2.77E-27
		AO090009000040	Hypothetical protein	Cubic	0.98	0.001
		AO090010000582	Eukaryotic-type carbonic anhydrase	Cubic	0.95	0.004
	Combined	AO090005000567	Hypothetical protein	Quadratic	0.74	2.00E-04
		AO090102000623	HLH transcription factor	Linear	0.46	0.005

Promoter Region TR Alleles do not act as Expression "Switches" in *A. oryzae*

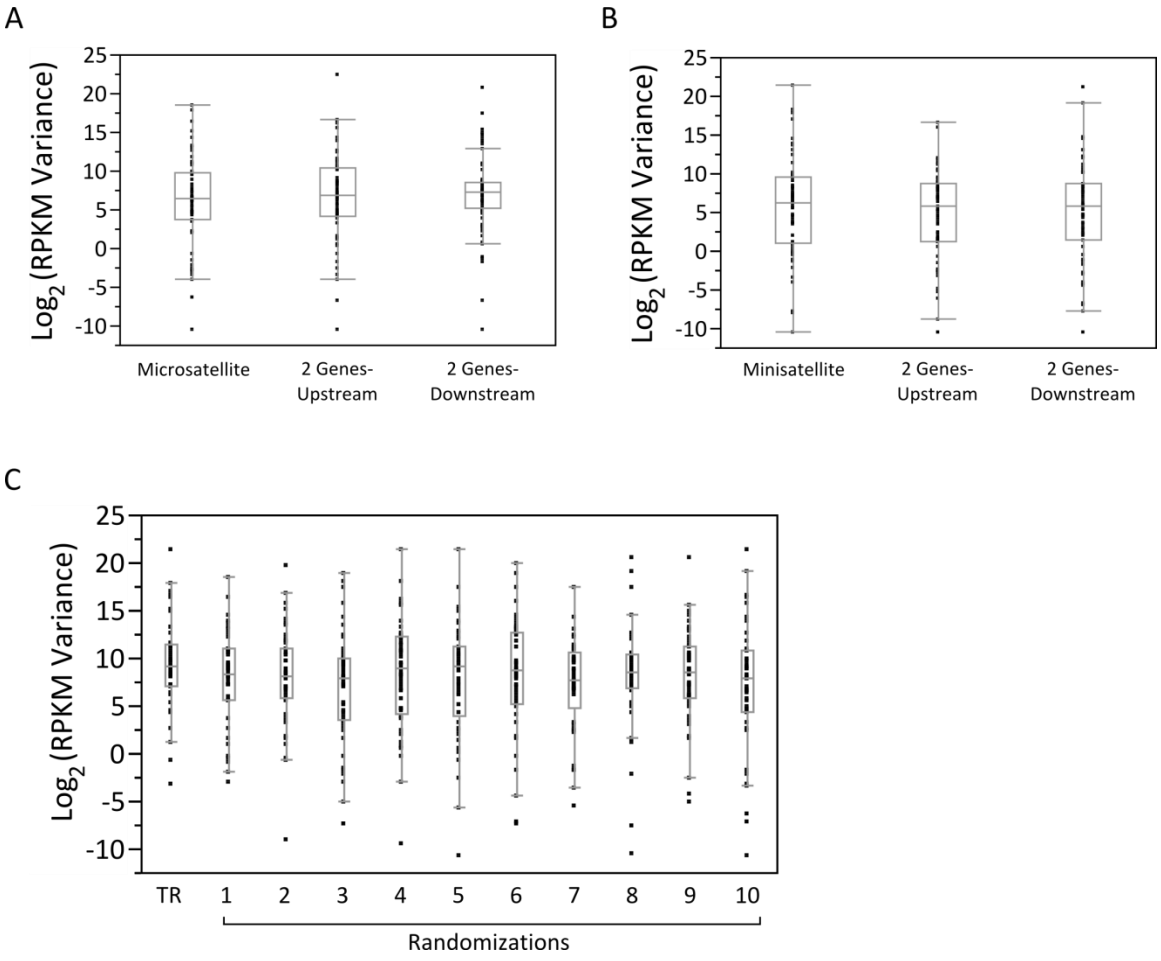
We investigated the "switch" function of TRs by comparing the expression values of genes with TR alleles occurring at frequencies ≥ 0.25 in *A. oryzae*. In total, 49 loci (12 microsatellite and 37 minisatellite) were analyzed. We found no significant differences in the expression levels of different alleles in any of our comparisons (at the *P-value* < 0.01 level).

Promoter Region TRs do not Generate Expression Noise

We also tested the hypothesis that TRs in promoter regions may increase expression noise by comparing the distributions of expression variance between genes containing

promoter region TRs against background sets of genes lacking promoter region TRs that are either in the same genomic region or have the same functional annotation. If TRs were acting as expression "noise makers" we would expect the average expression variance to be higher in genes containing promoter region TRs than in background genes. For both microsatellites and minisatellites, we found no significant difference between expression variance between genes with promoter region TRs and genes lacking promoter region TRs controlling for effects of genomic region (Tukey Kramer; microsatellite: vs. up-stream $P = 0.93$, vs. down-stream $P = 0.62$ and minisatellite: vs. up-stream $P = 0.52$, vs. down-stream $P = 0.75$) (Figure 6.3A, B) and functional annotation (Tukey Kramer; all $P > 0.65$) (Figure 6.3C).

Figure 6.3. Expression variance is not elevated in genes with promoter region TRs. Expression variance comparison between genes containing promoter region TRs and those without promoter region TRs but (A and B) are located in the same "genomic neighborhood" or (C) have similar functions. For each box plot, the horizontal line represents the sample median, the box extends from the first to the third quartile and the whiskers extend to the interquantile ranges. The grand mean, is given by the horizontal line running through each analysis. For (C), "TR" represents the TR dataset (combined microsatellite and minisatellite), while numbers 1-10 are the randomized datasets of genes with similar functional classifications lacking promoter region TRs.



DISCUSSION

To investigate the utility and prevalence of TRs as suppliers of robust genetic and phenotypic variation (Kashi, King et al. 1997; King, Soller et al. 1997) we examined the genome-wide patterns of promoter region TR polymorphisms and their downstream effects on gene expression in the closely related species *A. flavus* and *A. oryzae*. Although we found several cases of strong "knob-like" relationships between TR allele length and gene expression (Table 6.1, Figures 6.2), the vast majority of loci examined showed no such relationship. We also found no evidence in our dataset that TRs act as switches (i.e. particular alleles at a given locus did not have distinct expression levels). Interestingly, we found that genes containing promoter region TRs were significantly associated with functions involved in the response to environmental stimuli. This lead us to hypothesize that TRs may actually be acting as "noise makers" rather than "knobs" or "switches". TR derived noise may provide a source of transcriptional variation from which cells can respond to constantly changing environments. However, we found no evidence that expression variance of genes containing promoter region TRs is greater than in genes lacking promoter region TRs (Figure 6.3).

In the late 1990s King (1997) and Kashi (1997) raised the hypothesis that the abundant, reversible, continuous, and rarely deleterious genetic variation produced by the instability of TRs could be a potential source of rapid adaptation. This hypothesis states that subtle genetic changes in TR copy number can act as a phenotypic "tuning knob" on which selection can act upon (King, Soller et al. 1997). In recent years, many case studies have disovered TR-containinf loci whose effect on phenotype supports for this hypothesis

(Sawyer, Hennessy et al. 1997; Fondon and Garner 2004; Li, Korol et al. 2004; Verstrepen, Jansen et al. 2005; Kashi and King 2006; Vinces, Legendre et al. 2009; Gemayel, Vinces et al. 2010). For example, repeat polymorphisms in the coding regions of the *Alx-4* and *Runx-2* genes have been directly associated with skeletal morphology differences in the domestic dog breeds (Fondon and Garner 2004), whereas repeat polymorphisms in promoter regions modulate expression in the *S. cerevisiae* genes *MET3* and (Vinces, Legendre et al. 2009). Our results too suggest several examples of loci where repeat length variation affects gene expression phenotype in a specific manner (Table 6.1, Figures 6.2). However, the limited frequency of these relationships (only 4.3% of comparisons) suggests that the "evolutionary tuning knob" hypothesis may not be as prevalent as previously hypothesized. Moreover, if TRs played an important role in supplying rapid phenotypic variation to populations experiencing different selective pressures (Fondon and Garner 2004), we would expect to see TRs enriched in differentially expressed genes between *A. oryzae* and *A. flavus*; however we found no such relationship.

An alternative to the "tuning knob" and "switch" hypotheses is that TR variation in promoter regions may contribute to gene expression "noise". Under this hypothesis, TR variation may alter the regulatory landscape of the promoter region and generate a range of expression levels, in turn providing pliancy in the response to labile environmental conditions. Consistent with the "noise maker" hypothesis, Vinces *et al.* (2009) found elevated divergence in the transcriptional activity of genes with polymorphic TRs in their promoter regions. If TRs generate expression noise, one would expect particular classes

of genes to tolerate this variation better than others. In yeast, genes localized to the plasma membrane and functional in stress response have elevated expression divergence (Tirosh, Weinberger et al. 2006). Similarly, we found that genes containing promoter region TRs are associated with responding to environmental stimuli (signal transduction, stress response, growth and development). However, we did not find evidence for elevated levels of expression variance in genes containing promoter region TRs (Figure 6.3).

ACKNOWLEDGMENTS AND CONTRIBUTIONS

We thank members of the Rokas lab, Travis Clark and Chelsea Baker for constructing RNAseq libraries, Dr. Osamu Yamada and the National Research Institute of Brewing of Japan for *A. oryzae* isolates and Maren Klich of the USDA for *A. flavus* isolates. JGG is funded by the Graduate Program in Biological Sciences at Vanderbilt University and the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH, NIAID: F31AI091343-01). Research in A.R.'s lab is supported by the Searle Scholars Program and the National Science Foundation (DEB-0844968).

J.G.G. and A.R. designed the study. M.H.E. maintained fungal cultures, extracted nucleic acids, designed primers, and genotyped isolates. J.G.G. analyzed data.

REFERENCES

Bennett, S. T., A. M. Lucassen, et al. (1995). "Susceptibility to Human Type-1 Diabetes at Iddm2 Is Determined by Tandem Repeat Variation at the Insulin Gene Minisatellite Locus." Nature Genetics **9**(3): 284-292.

Chen, F., W. Q. Liu, et al. (2010). "mutL as a genetic switch of bacterial mutability: turned on or off through repeat copy number changes." Fems Microbiology Letters **312**(2): 126-132.

Don, R. H., P. T. Cox, et al. (1991). "'Touchdown' PCR to circumvent spurious priming during gene amplification." Nucleic Acids Res **19**(14): 4008.

Fondon, J. W. and H. R. Garner (2004). "Molecular origins of rapid and continuous morphological evolution." Proceedings of the National Academy of Sciences of the United States of America **101**(52): 18058-18063.

Gemayel, R., M. D. Vinces, et al. (2010). "Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences." Annual Review of Genetics, Vol 44 **44**: 445-477.

Gibbons, J. G., A. Beauvais, et al. (2012). "Global Transcriptome Changes Underlying Colony Growth in the Opportunistic Human Pathogen Aspergillus fumigatus." Eukaryotic Cell **11**(1): 68-78.

Gibbons, J. G., L. Salichos, et al. (2012). "The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*." Current Biology ***in press***.

Goldstein, D. B. and C. Schlötterer (1999). Microsatellites : evolution and applications. Oxford ; New York, Oxford University Press.

Guichoux, E., L. Lagache, et al. (2011). "Current trends in microsatellite genotyping." Molecular Ecology Resources **11**(4): 591-611.

Hammock, E. A., M. M. Lim, et al. (2005). "Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior." Genes Brain Behav **4**(5): 289-301.

Hammock, E. A. D. and L. J. Young (2004). "Functional microsatellite polymorphism associated with divergent social structure in vole species." Molecular Biology and Evolution **21**(6): 1057-1063.

Hancock, J. M. (1999). Microsatellites and other simple sequences: genomic context and mutational mechanisms. Microsatellites : evolution and applications. Oxford ; New York, Oxford University Press**:** 1-9.

Jiang, H. and W. H. Wong (2009). "Statistical inferences for isoform expression in RNA-Seq." Bioinformatics **25**(8): 1026-1032.

Kashi, Y., D. King, et al. (1997). "Simple sequence repeats as a source of quantitative genetic variation." Trends Genet **13**(2): 74-8.

Kashi, Y. and D. G. King (2006). "Simple sequence repeats as advantageous mutators in evolution." Trends Genet **22**(5): 253-9.

King, D. G., M. Soller, et al. (1997). "Evolutionary tuning knobs." Endeavour **21**(1): 36-40.

Levinson, G. and G. A. Gutman (1987). "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." Mol Biol Evol **4**(3): 203-21.

Li, Y. C., A. B. Korol, et al. (2002). "Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review." Molecular Ecology **11**(12): 2453-2465.
Li, Y. C., A. B. Korol, et al. (2004). "Microsatellites within genes: Structure, function, and evolution." Molecular Biology and Evolution **21**(6): 991-1007.
Lynch, M., W. Sung, et al. (2008). "A genome-wide view of the spectrum of spontaneous mutations in yeast." Proceedings of the National Academy of Sciences of the United States of America **105**(27): 9272-9277.
Machida, M., K. Asai, et al. (2005). "Genome sequencing and analysis of Aspergillus oryzae." Nature **438**(7071): 1157-61.
Michael, T. P., S. Park, et al. (2007). "Simple sequence repeats provide a substrate for phenotypic variation in the Neurospora crassa circadian clock." PLoS ONE **2**(8): e795.
Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-8.
Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.
Rockman, M. V., M. W. Hahn, et al. (2005). "Ancient and recent positive selection transformed opioid cis-regulation in humans." Plos Biology **3**(12): 2208-2219.
Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.
Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.
Sawyer, L. A., J. M. Hennessy, et al. (1997). "Natural variation in a Drosophila clock gene and temperature compensation." Science **278**(5346): 2117-2120.
Schilling, G., A. H. Sharp, et al. (1995). "Expression of the Huntingtons-Disease (It15) Protein Product in Hd Patients." Human Molecular Genetics **4**(8): 1365-1371.
Schuelke, M. (2000). "An economic method for the fluorescent labeling of PCR fragments." Nature Biotechnology **18**(2): 233-234.
Staib, P., M. Kretschmar, et al. (2002). "Host versus in vitro signals and intrastrain allelic differences in the expression of a Candida albicans virulence gene." Molecular Microbiology **44**(5): 1351-1366.
Stewart, C. N. and L. E. Via (1993). "A Rapid Ctab DNA Isolation Technique Useful for Rapid Fingerprinting and Other Pcr Applications." Biotechniques **14**(5): 748-&.
Streelman, J. T. and T. D. Kocher (2002). "Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia." Physiological Genomics **9**(1): 1-4.
Tirosh, I., A. Weinberger, et al. (2006). "A genetic signature of interspecies variations in gene expression." Nature Genetics **38**(7): 830-834.
Verstrepen, K. J., A. Jansen, et al. (2005). "Intragenic tandem repeats generate functional variability." Nature Genetics **37**(9): 986-990.
Vinces, M. D., M. Legendre, et al. (2009). "Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability." Science **324**(5931): 1213-1216.

CHAPTER VII

CONCLUSION

This dissertation demonstrates the utility of comparative genomics in addressing fundamental questions at the organismal, population and species levels. By combining genomics with functional experiments, population genetics and evolutionary analyses, I have revealed biological insight which no methodology could alone uncover. My dissertation addresses three central themes (1) The pathogenicity of *Aspergillus fumigatus*, (2) The domestication of *Aspergillus oryzae* and (3) The function and evolution of tandemly repeated DNA.

The Pathogenicity of *A. fumigatus*

During localized aspergilloma infections, *A. fumigatus* grows as a tightly adhered biofilm which confers increased virulence and drug resistance. Previously, our collaborators developed an *in vitro* biofilm model for *A. fumigatus* closely resembling the *in vivo* morphology (Beauvais, Schmidt et al. 2007). I compared the expression profiles of *in vitro* biofilm growth to that of planktonic growth to understand the transcriptional programming underlying this clinically important phenotype. These analyses identified genes likely involved in the adherence (cell surface encoding), increased drug resistance (multidrug resistance membrane transporters), and heightened virulence (secondary metabolism associated) phenotypes characteristic of the biofilm.

***Future Directions*.** The list of candidate genes generated from this analysis has provided a valuable resource for more detailed biofilm related studies. For example, the

genes encoding hydrophobins, a family of hydrophobic cell-surface proteins with adhesive properties, were substantially up-regulated during biofilm growth. It has been hypothesized, that these proteins play a major role in "gluing" the biofilm together. Our collaborators are currently generating sequential deletions of these genes to investigate their functional role in the extracellular matrix. Further, although our *in vitro* biofilm model closely mirrors the *in vivo* morphology, analyzing the transcriptional profile of *A. fumigatus* while in a mammalian host would be ideal. Recently, our collaborators have developed an *in vivo* mouse model for *A. fumigatus* infection. We are now in the early stages of analyzing the transcriptomes of both *A. fumigatus* while growing on a mouse lung, as well as the host response to infection the fungal infection.

Treating *A. fumigatus* infections has become increasingly challenging due to the emergence and rapid spread of multiple triazole resistance (Snelders, van der Lee et al. 2008). Understanding the population dynamics of *A. fumigatus* is critical in the control of resistance and treatment of infection. In the past, several efforts have concluded that *A. fumigatus* shows no detectable global population structure (Debeaupuis, Sarfati et al. 1997; Pringle, Baker et al. 2005; Rydholm, Szakacs et al. 2006). By focusing on a more defined geographical location, analyzing hundreds of isolates, and using a larger and more informative set of genetic markers, this study was the first to detect population differentiation in *A. fumigatus.* Importantly, all drug resistant isolates were confined to a single, predominantly asexually reproducing population indicating that sexual recombination had not contributed to the spread of drug resistance. However, several non-resistant isolates in this population had contributions from different genetic

backgrounds, suggesting that, although likely rare, resistant isolates may have the ability to recombine.

Future Directions. In the last several years, *A. fumigatus* isolates with the identical TR/L98H resistance allele have appeared elsewhere around the globe. Genotyping these isolates with our set of markers would enable the scientific and medical communities to determine if these isolates originated from the population we identified in the Netherlands, or if recombination has spread the resistance mechanism into new genetic backgrounds. Better understanding the population biology of resistant isolates will help to better control and treat infections.

The Domestication of *A. oryzae*

Plant and animal models of domestication have been extensively studied and have revealed that the majority of morphological phenotypes desirable to humans are the result of genetic changes that shape developmental pathways (Diamond 2002; Doebley, Gaut et al. 2006; Purugganan and Fuller 2009; Trut, Oskina et al. 2009). However, the genetic and functional alterations underlying microbial domestication remain scantly explored. Microbes differ from plants and animals with respect to their underlying reproductive biology, population dynamics as well as their purpose to humans. A specifically interesting example of microbial domestication is that of *A. oryzae*, which has been used for thousands of years as a metabolic workhorse in the production of traditional Asian foods and beverages. Despite this long history, very little is known about the effect domestication has had on the *A. oryzae* genome. I therefore used this system to

specifically address the genomic signatures of domestication in *A. oryzae* while more generally comparing the broad observed patterns to plant and animal models.

To date, this analysis is the most extensive study of microbial domestication as it combined whole genome sequencing, transcriptomics and proteomics. This work provided several interesting results likely reflecting human selective pressure. First, in comparison to its progenitor *A. flavus*, *A. oryzae* exhibited widespread down-regulation of secondary metabolism genes, including the clusters of genes responsible for synthesizing the toxins cyclopiazonic acid and aflatoxin. The implications of this phenotype are two-fold. First, reducing or eliminating toxins in food prevents human sickness. Perhaps just as importantly however, aflatoxin, and possibly other secondary metabolites, can kill yeast, in turn negatively impacting fermentation during sake, shoyu and miso production. This work also identified α-amylase, the primary starch metabolizing enzyme, as the most highly expressed transcript and abundant protein in the *A. oryzae* genome. This phenotype is essential in the saccharification of industrial substrates such as rice and soy. Finally, differences in several genes and pathways associated with flavor were revealed during this analysis, including a glutaminase and the presence of a polymorphic sesquiterpene encoding gene cluster which was fixed in *A. oryzae.* Taken together, these observations suggest that microbial domestication principally results in metabolic reshaping, rather than developmental modifications as is the case in plants and animals.

***Future Directions*.** One particularly interesting finding from this study was that genes involved in secondary metabolism were down-regulated in *A. oryzae*. We hypothesized that this phenotype was selected for over the course of domestication

because of its effect on yeast growth (and thus fermentation efficiency) rather than its negative impact on human health. To test this hypothesis, we have begun to develop a simple, yet informative growth assay in which sake yeast is grown in competition with *A. oryzae* and *A. flavus*. Supporting our hypothesis, our preliminary results suggest that yeast growth was higher in the *A. oryzae* and closely related *A. flavus* isolates compared the *A. flavus* isolates highly expressing secondary metabolism associated genes. To validate these results, I would like to expand this assay to screen many more isolates of *A. flavus* and *A. oryzae*.

The Function and Evolution of Tandemly Repeated DNA

The sequenced *Aspergillus* genomes offered a phenomenal opportunity to investigate more general evolutionary questions. Specifically, it was my goal to understand how tandemly repeated DNA evolves and functions in eukaryotes. I first surveyed the distribution and variation of tandem repeats across the coding regions of 10 *Aspergillus* genomes. Tandem repeats were generally unique between species and highly polymorphic within species. Furthermore, the genes containing tandem repeats were evolutionarily less conserved than background genes. Interestingly, tandem repeats were enriched in genes involved in key developmental and signaling processes critical to the fungal lifestyle while being underrepresented in genes involved in essential housekeeping functions. Notably, similar functional trends have been observed in other eukaryotes as well (Katti, Sami-Subbu et al. 2000; Young, Sloan et al. 2000; Alba and Guigo 2004; O'Dushlaine, Edwards et al. 2005; Huntley and Clark 2007). It is tempting to speculate that genes which interact with labile environments may better tolerate, or benefit from, variation generated by tandem repeats. The *Aspergillus* present an ideal substrate to experimentally

evaluate this hypothesis in the future, as we discovered several candidates involved in biologically interesting functions such as fungal mating pheromone reception and heterokaryon formation.

Future Directions. We have shown that particular types of tandem repeats are tolerated more so than others in coding regions (i.e. trinucleotide repeats). However we have not explored the interaction between tandem repeats and their impact upon protein function and structure. Specifically, are tandem repeats found more often in protein domain regions or non-domain regions and in ordered regions or non-ordered protein regions? Furthermore, this chapter solely focuses on the evolutionary patterns of tandem repeats within coding regions. For a more comprehensive genomic characterization of tandem repeats, similar analyses in non-coding regions would build upon this work. Specifically, do levels of tandem repeat variation and conservation vary between intergenic, intronic and coding regions?

Similarly, in non-coding regions variation in tandem repeat length can modulate gene expression (Bennett, Lucassen et al. 1995; Staib, Kretschmar et al. 2002; Streelman and Kocher 2002; Hammock, Lim et al. 2005; Vinces, Legendre et al. 2009) and has been hypothesized to act as a transcriptional "tuning knob". However, no study before the present one has evaluated this hypothesis on a genome-wide level in natural populations. This work identified several significant relationships between tandem repeat length and gene expression levels; however the majority of loci fit no such pattern. I also did not find evidence that genes containing promoter region tandem repeats have greater variance in expression levels. Collectively, these results suggest that although tandem repeats can affect gene expression in particular circumstances, the "tuning-knob" hypothesis may not be as general as once presumed.

***Future Directions*.** Our results strongly suggest that tandem repeat variation may not be a general source of rapid and continuous phenotypic variation, contrary previous hypotheses. However, this study only focused on gene expression as a phenotype. It would be worthwhile to investigate the functional consequences of a smaller subset of genes with known phenotypes; however presently, only few genes are well functionally annotated in *A. oryzae* and *A. flavus*. Furthermore, our analyses relied on inferences drawn from only two closely related species. To reliably verify our results, this experiment should be repeated in other eukaryotes.

Summary

Together, my dissertation work has contributed novel insights into lifestyle and genome evolution of the *Aspergillus*, a genus of filamentous fungi with extreme importance to human society. This research has helped to understand the pathogenicity of *A. fumigatus* by studying various aspects of its biology at both the organismal level (Chapter II) and population level (Chapter III). In addition, study of *A. oryzae* domestication using a multidisciplinary approach yielded a variety of findings; some of which may have applied industrial implications, while others provide a glimpse into how humans have utilized microorganisms for thousands of years (Chapter IV). Lastly, the *Aspergillus* genome resources have offered an ideal platform to study the evolution and function of tandemly repeated DNA (Chapters V and VI). The genetic and ecological diversity of the *Aspergillus* make this system a remarkably useful model for the study of functional and evolutionary genomics.

References

Alba, M. M. and R. Guigo (2004). "Comparative analysis of amino acid repeats in rodents and humans." Genome Res **14**(4): 549-54.

Beauvais, A., C. Schmidt, et al. (2007). "An extracellular matrix glues together the aerial-grown hyphae of Aspergillus fumigatus." Cellular Microbiology **9**(6): 1588-1600.

Bennett, S. T., A. M. Lucassen, et al. (1995). "Susceptibility to Human Type-1 Diabetes at Iddm2 Is Determined by Tandem Repeat Variation at the Insulin Gene Minisatellite Locus." Nature Genetics **9**(3): 284-292.

Debeaupuis, J. P., J. Sarfati, et al. (1997). "Genetic diversity among clinical and environmental isolates of Aspergillus fumigatus." Infection and Immunity **65**(8): 3080-3085.

Diamond, J. (2002). "Evolution, consequences and future of plant and animal domestication." Nature **418**(6898): 700-707.

Doebley, J. F., B. S. Gaut, et al. (2006). "The molecular genetics of crop domestication." Cell **127**(7): 1309-21.

Hammock, E. A., M. M. Lim, et al. (2005). "Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior." Genes Brain Behav **4**(5): 289-301.

Huntley, M. A. and A. G. Clark (2007). "Evolutionary analysis of amino acid repeats across the genomes of 12 drosophila species." Molecular Biology and Evolution **24**(12): 2598-2609.

Katti, M. V., R. Sami-Subbu, et al. (2000). "Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications." Protein Science **9**(6): 1203-1209.

O'Dushlaine, C. T., R. J. Edwards, et al. (2005). "Tandem repeat copy-number variation in protein-coding regions of human genes." Genome Biology **6**(8): R69.

Pringle, A., D. M. Baker, et al. (2005). "Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus Aspergillus fumigatus." Evolution **59**(9): 1886-1899.

Purugganan, M. D. and D. Q. Fuller (2009). "The nature of selection during plant domestication." Nature **457**(7231): 843-848.

Rydholm, C., G. Szakacs, et al. (2006). "Low genetic variation and no detectable population structure in Aspergillus fumigatus compared to closely related Neosartorya species." Eukaryotic Cell **5**(4): 650-657.

Snelders, E., H. A. L. van der Lee, et al. (2008). "Emergence of Azole Resistance in Aspergillus fumigatus and Spread of a Single Resistance Mechanism." Plos Medicine **5**(11): 1629-1637.

Staib, P., M. Kretschmar, et al. (2002). "Host versus in vitro signals and intrastrain allelic differences in the expression of a Candida albicans virulence gene." Molecular Microbiology **44**(5): 1351-1366.

Streelman, J. T. and T. D. Kocher (2002). "Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia." Physiological Genomics **9**(1): 1-4.

Trut, L., I. Oskina, et al. (2009). "Animal evolution during domestication: the domesticated fox as a model." Bioessays **31**(3): 349-360.

Vinces, M. D., M. Legendre, et al. (2009). "Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability." Science **324**(5931): 1213-1216.

Young, E. T., J. S. Sloan, et al. (2000). "Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae." Genetics **154**(3): 1053-68.