THE GENETICS OF AGE-RELATED MACULAR DEGENERATION: EXPLORING PATHWAY AND EPISTATIC EFFECTS

By

Jacob B. Hall

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

Human Genetics

May, 2016

Nashville, TN

Approved:

Professor Milam A. Brantley

Professor William S. Bush

Professor John A. Capra

Professor Jonathan L. Haines

Professor Marylyn D. Ritchie

Professor David C. Samuels

This work is dedicated to all of my family and friends
for their love, support, encouragement, and guidance.

ACKNOWLEDGEMENTS

The work presented in this dissertation was supported in part by an ocular genomics training grant (1T32EY021453-01), originally awarded to Vanderbilt University's Center for Human Genetics Research, obtained by Jonathan Haines and now managed by Milam Brantley, as well as an Alzheimer's sequencing grant (1UF01AG07133-01), which partially helped fund methods development. I would like to thank the International Age-related Macular Degeneration Genetics Consortium (IAMDGC) for granting me access to an invaluable dataset, enabling me to carry out well-powered genetic studies of AMD. I would also like to express my thanks and appreciation to all of the individuals who volunteered to be a part of the IAMDGC cohort.

This work was guided and greatly improved by input from my dissertation committee: David Samuels (committee chair), William Bush (advisor), Milam Brantley (co-advisor), John Capra, Jonathan Haines, and Marylyn Ritchie. Insight from every committee member helped me ask and pursue the most interesting research questions. In particular, I would like to thank my advisor, William Bush, for being generous with his time, for his patience, scientific enthusiasm, positivity, confidence in me, and his constructive feedback and suggestions.

I have had the opportunity to work with and gain knowledge from many colleagues and friends. I would like to thank Dana Crawford for allowing me to participate in a summer rotation (before I officially started graduate school), where I was able to work with Janina Jeff and Logan Dumitrescu to learn the basics of Linux and genetic analyses. Previous and current members of the Bush Lab have provided much

expertise and friendship. Programmers Peter Straub and Jeremy Fondran provided immeasurable help with everything from small coding questions to expertise in developing complex code and software. Laura Wiley, my lab mate and friend of many years, provided countless hours of help and advice and was always willing to lend an ear.

Some of the other students and post-doctoral researchers who have been a positive influence include Jessica Cooke Bailey, Sabrina Mitchell, Matt Oetjens, Nicole Restrepo, Alex Fish, Mike Sivley, Corinne Simonti, Noemi Hall, Heather Voss-Hoynes, and many current and former members of the Ritchie Lab. Finally, I would like to thank the former Center for Human Genetics Research for bringing together such a great group of scientist together – all of whom never failed to treat students more like colleagues while still sharing their expertise and encouraging challenging and dynamic discussions.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF APPENDICES

Appendix **Page**

Overview

Age-related Macular Degeneration (AMD) is a highly heritable disease affecting millions of people worldwide. Because AMD — as its name implies — is an age-related disease and average life expectancies continue to increase, the prevalence of AMD also continues to increase, resulting in both decreased quality of life and financial burdens. Compared to many other diseases, the genetics of AMD has been well-studied, yet known risk variants only explain a portion of the heritability and risk for AMD. With much of the "low-hanging fruit" discovered, this work explores some of the many possible genetic factors that could contribute to the unexplained heritability. Specifically, I seek to better understand the genetics of AMD by answering the following question:

How much of AMD's heritability is explained by biologically-relevant pathways, by dominance effects, and by epistasis between plausible genomic regions?

Existing literature have used linkage analyses to narrow down regions of cosegregation as well as genome-wide association studies (GWAS) in large case-control cohorts to uncover specific genome-wide significant SNPs contributing to risk for AMD. These GWAS have confirmed statistically significant associations between AMD risk and several genes, including *CFH*, *C2*, *C3*, *CFB*, and *ARMS2*/*HTRA1*. Known risk SNPs only explain a portion of AMD's heritability, however. AMD-related pathways likely contain single nucleotide polymorphisms (SNPs) with small effects that individually do not reach genome-wide significance but still contribute to the unexplained heritability of AMD. Additionally, a portion of the unexplained heritability could potentially be explained by interactions between loci (epistasis), which is not taken into consideration in a typical GWAS. Within this work I will present studies that explore the impact that these AMD-related pathways and epistasis have on AMD risk.

Chapter 1 begins by giving an in-depth background on age-related macular degeneration (AMD). I focus on two primary components — known biological and genetic factors of AMD. First, I emphasize the global impact of AMD on health and financial burden. While not fully understood, much is known about the pathogenesis and progression of AMD, including the fact that there are two main subtypes — dry and wet. I also review related clinical trials and treatment options before reviewing the history and current state of knowledge regarding AMD genetic risk factors and heritability.

For many years, the primary way to estimate heritability was through twin studies, which take advantage of the difference in trait variance between monozygotic and dizygotic twins to determine the extent to which a trait has a genetic component. In **Chapter 2** I review mixed linear model (MLM) methods and software that estimate heritability by comparing genetic variance, generated from unrelated individuals, to phenotypic variance. Focusing on MLM software, particularly Genome-wide Complex Trait Analysis (GCTA), I describe advantages, disadvantages, and alternatives.

In **Chapter 3**, using a case-control dataset, I perform a unique pathway analysis of AMD by leveraging known biological information to test specific pathways for their contribution to risk for AMD. First, I estimate the amount of AMD risk explained by known, published risk SNPs. Then I analyze multiple pathways, taking into consideration possible gene-environment interactions due to smoking, differences in risk explained by

AMD subtype, linkage disequilibrium, and potential overestimates of heritability due to genes shared between pathways. For each pathway I carefully dissect the contribution to AMD risk for genic SNPs, nearby flanking, potentially regulatory SNPs, and more distant SNPs in open chromatin regions in ocular tissue, leveraging data from the Encyclopedia of DNA Elements (ENCODE) project. I show that the complement and inflammatory pathways harbor statistically significant genetic variation that contributes to AMD risk. I also show that genetic variation in the complement pathway, separate from known risk SNPs, contributes cumulatively to risk for AMD. This chapter is adapted from my peer-reviewed article "Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models" in *BMC Bioinformatics* [1].

Thus far, published studies of genetic interactions have focused primarily on finding individual significant SNP-SNP pairs. In **Chapter 4** I describe two novel methods (iSim and iGRM) that I developed to explore genetic interaction effects. Advances in computing now allow us to test all pairwise genotyped SNPs and, while this can uncover meaningful, real interaction effects, such methods require the pair of SNPs to have a very large effect to reach statistical significance. For some traits, many interactions with small effect sizes may contribute cumulatively to trait variation or risk. The methods I describe in Chapter 4 allow this to be tested and quantified. First, I describe a method (iSim) to simulate datasets with specific effects — including additive, dominant, and epistatic genetic components. Second, I describe a method (iGRM) to estimate the impact that those genetic effects — within specified genomic regions — have on trait variation, using simulations from the first method as validation. I conclude each method section by describing the current efficiency of the software, as well as provide suggestions for future potential changes to optimize the software.

In Chapter 3 I explored pathways contributing to risk for AMD, which helps uncover and localize some of the unexplained heritability for AMD due to additive effects, yet more research is needed to explore potential genetic interaction effects influencing AMD risk. In **Chapter 5** I apply the methods developed in Chapter 4 to a large International AMD Genetics Consortium (IAMDGC) dataset. In many studies of AMD *ARMS2* contains a statistically significant risk SNP, yet its biological mechanism and relation to disease progression and pathogenesis is not understood. For this analysis I test for cumulative interaction effects between *ARMS2* and multiple AMD-related pathways to search for a potential biological mechanism linking *ARMS2* and risk for AMD. I show that *ARMS2* does not contribute to AMD risk through cumulative interaction effects with the antioxidant, complement, oxidative phosphorylation, nicotine, or TCA pathways. Additionally, I show that a previously discovered interaction between AMD risk SNPs (rs10737680 - *CFH* and rs42908 - *C2*/*CFB*) replicate using our iGRM method, with small but statistically significant amounts of risk explained cumulatively by epistatic effects.

Finally, in **Chapter 6** I review the extent to which this work has contributed to our understanding of AMD and to novel statistical genetic analysis methods. I summarize and make final conclusions of my findings as well as discuss promising areas for future related research.

Chapter 1 — Age-related Macular Degeneration: What Is Known

Biology of AMD

Age-related macular degeneration (AMD) is a progressive, neurodegenerative disease affecting the central portion of the retina, called the macula, leading to a loss of central vision (Figures 1 and 2). Although the macula only makes up a small portion of the retina, it contains the 2 millimeter-wide fovea (Figure 3), which has the highest density of cone photoreceptors and is responsible for clear central vision.

Prevalence and burden

AMD is the leading cause of irreversible blindness in elderly individuals in developed countries, with thirty to fifty million people, world-wide, estimated to be affected [2]. The prevalence of AMD differs by age, gender, ethnicity, and type of AMD [3]. Overall prevalence rates are higher in females compared to males and are roughly seven times higher in European-descent individuals compared to African-descent individuals [3]. In individuals of European-descent, by age 80 the prevalence of early AMD is approximately 25% and prevalence of late/advanced AMD (geographic atrophy or neovascular AMD) is approximately 12% [3]. Prevalence worldwide is predicted to increase over time due to increased lifespans [3, 4]. AMD Alliance International (AMDAI) estimated the worldwide direct healthcare cost of AMD to be $255 billion in 2010 and predicted it to increase to $294 billion by 2020 [5]. While AMD was responsible for 9.5% of the prevalence of visual impairment, globally, in 2010, it accounted for 24.1% of direct healthcare costs because lifetime treatment is required after diagnosis [5].

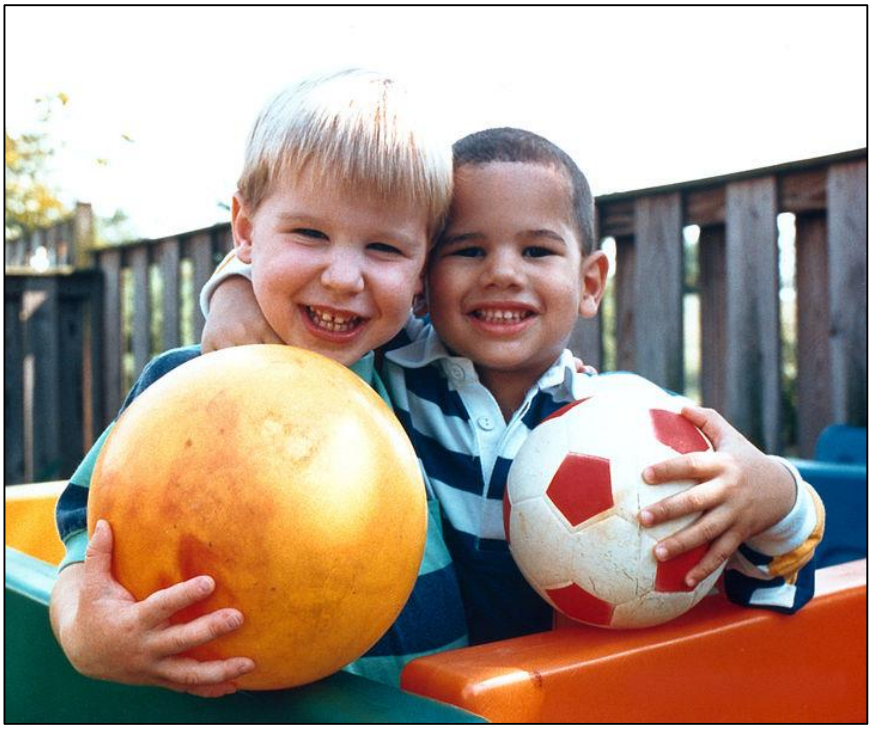


Figure 1. Example of normal vision.
Image created by the National Eye Institute and listed as public domain [6].

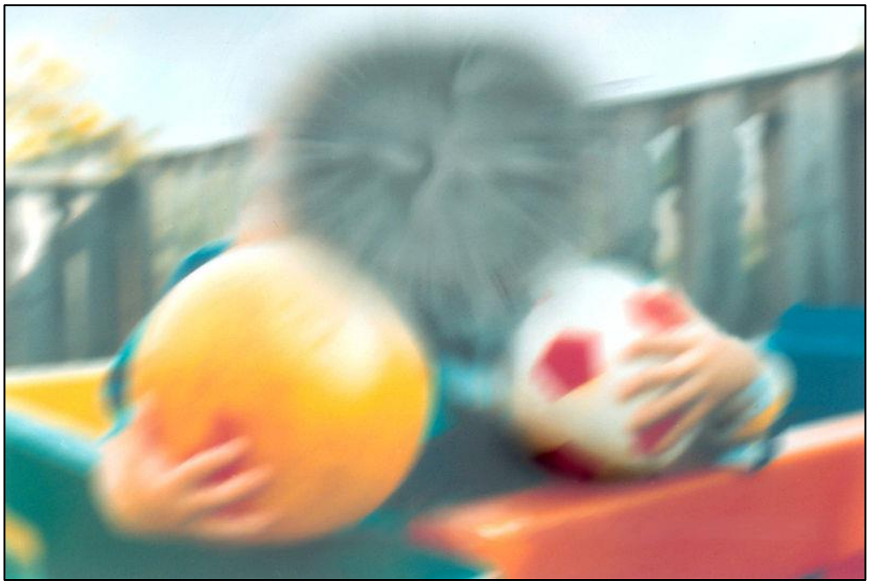


Figure 2. Example of vision with AMD.
Image created by the National Eye Institute and listed as public domain [6].

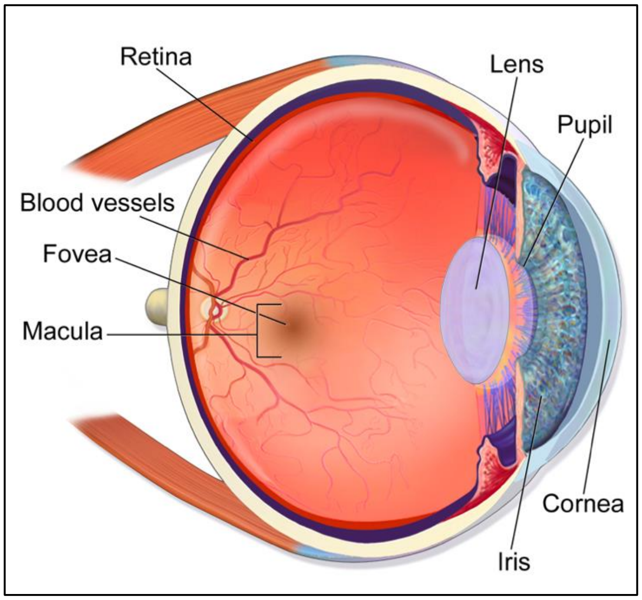


Figure 3. General anatomy of the eye.
Diagram from Blausen gallery 2014 [7]. Open access use permitted under the Creative Commons License.

Living with AMD can severely reduce quality of life in patients. Critical tasks such as driving can become dangerous due to difficulty seeing; once symptoms become severe patients often have to rely on others for transportation. Other important tasks such as walking, cooking, and reading may become difficult. Frustration from a loss of independence due to visual impairment can further lead to depression [8]. Patients still working may lose the ability to perform their job, leading to job loss and increased financial burden. These factors, combined with the current lack effective prevention methods or treatment options, justify the great need for further AMD research.

Pathogenesis and risk factors

Accumulation of drusen — tiny yellow or white lipid deposits in the Bruch's membrane layer of the retina — can be an early sign of AMD. Large, soft drusen are associated with a higher risk of developing AMD, while small, hard drusen are common in individuals over the age of 50 and not necessarily indicative of risk for AMD [9]. AMD can be classified as dry or wet. Geographic atrophy, the advanced form of dry AMD, results when the drusen affect the retinal pigment epithelial (RPE) layer, leading to a loss of photoreceptors [10]. Wet AMD, or neovascular (exudative) AMD, is caused by leakage of blood and protein below the macula due to abnormal blood vessel growth [10]. Neovascular AMD is roughly 2.5 times more prevalent than geographic atrophy in people over 75 years of age [11].

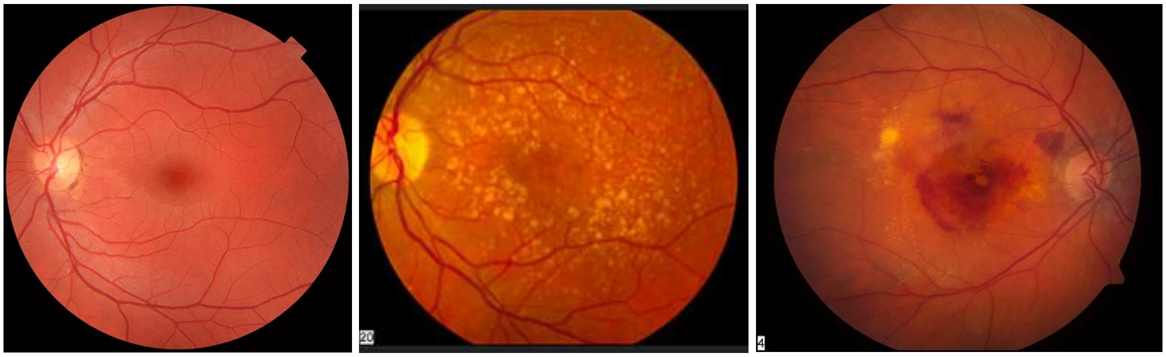


Figure 4. Fundus photographs of geographic atrophy and neovascular AMD. *Left* – normal eye; *Middle* – intermediate AMD; *Right* - neovascular/exudative AMD. All images from the National Institutes of Health [12] and listed as public domain.

Figure 4 shows fundus photos with and without AMD. The left frame shows a healthy eye with no signs of AMD. The middle frame shows numerous drusen, indicative of dry AMD (geographic atrophy). The right frame shows sub-retinal hemorrhaging indicative of neovascular, exudative (wet) AMD. Hallmarks of early/intermediate AMD additionally include a thickened Bruch's membrane and appearance macrophages in the choroid. For geographic atrophy, photoreceptor degeneration is typically due to drusen accumulation and altered blood flow in the choroid, limiting the supply of nutrients to the RPE and ultimately leading to visual loss. For neovascular (exudative) AMD, major changes to the RPE and surrounding areas typically occur, primarily due to abnormal blood vessel growth through the Bruch's membrane and subsequent leakage of blood or fluid, leading to photoreceptor cell degeneration and vision loss. Occurrence of dry and wet AMD is not mutually exclusive.

Because disease progression is painless, early signs and symptoms often go unnoticed until vision is affected. However, in some cases, drusen can affect the layers of the retina so that the retinal pigment epithelium (RPE) becomes detached [10].

Risk factors for AMD include age, use of tobacco products, family history, hypertension, cholesterol levels, obesity, ethnicity, oxidative stress, and exposure to sunlight [3, 10, 13, 14]. Familial risk is a particularly important factor; having just one first-degree relative with AMD increases the risk of developing AMD three fold [13]. Clinically, AMD is typically diagnosed using either fundoscopy with pupil dilation, fluorescein angiography, or optical coherence tomography (OCT) [15]. Visual loss and progression due to AMD can be tracked using the Amsler grid test (Figure 5).

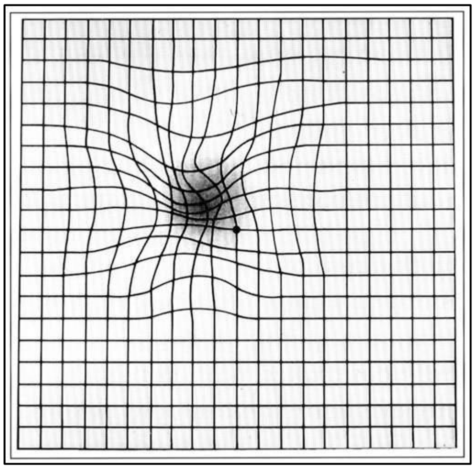


Figure 5. Amsler grid test.
The depiction here shows what an Amsler grid could look like to someone with AMD. The image is listed as public domain by the National Eye Institute, National Institutes of Health [16].

Treatment options

Treatment options are different for wet and dry AMD. Nutritional supplements from the Age-Related Eye Disease Study (AREDS) slow the progression of AMD and are recommended for people at high risk of developing advanced AMD [17]. The initial study, which included participants with no, early, intermediate, and advanced (in at most one eye, including wet and dry) AMD, assessed the effect of vitamin C (500 mg), vitamin E (400 IU), beta-carotene (15 mg), zinc oxide (80 mg), and copper (2 mg), none of which are naturally produced by the human body. The AREDS2 study [18] only included intermediate and advanced AMD patients (since no benefit was observed in the original AREDS study for people with no AMD or early AMD) and tested the effect of including additional dietary supplements for various combinations of lutein (10 mg), zeaxanthin (2 mg), and omega-3 fatty acids (docosahexaenoic acid [DHA] - 350 mg and eicosapentanoic acid [EPA] - 650 mg), as well as the effect of not including beta-carotene and/or reducing supplement zinc levels [18]. Results suggested that the omega-3 fatty acids (DHA and EPA) did not provide a significant benefit and that lutein and zeaxanthin could be used as substitutes for beta-carotene. Additionally, prophylactic laser treatment of drusen has been tested as a potential treatment for dry AMD but results showed that it was not beneficial [19].

Wet AMD is caused primarily by neovascularization through the Bruch's membrane, causing blood and protein to leak below the macula. Vascular endothelial growth factor inhibitor (anti-VEGF) treatments work by inhibiting angiogenesis — the formation of new blood vessels from preexisting vessels [20]. Two anti-VEGF drugs are

Avastin and Lucentis; both require monthly intraocular injections, although often treatment is only given upon signs of AMD progression. An older treatment option for wet AMD that is no longer used, photodynamic therapy (PDT), involves injecting a light sensitive drug (verteporfin) into the blood stream where it is picked up by lipoproteins, leading to accumulation specifically within abnormal vessels under the macula. A low power laser is then used to initiate a reaction that seals off leaky vessels [21].

While supplements may slow progression and treatment options exist for neovascular AMD, no treatment options exist for geographic atrophy and prevalence rates continue to increase, making it more important than ever to conduct research of AMD. In addition to better (for wet) or any (for dry) treatment options, better preventative measures are also needed — unlike in some species, once human photoreceptor cells die they cannot be regenerated [22].

GENETICS OF AMD

Early genetic associations

Familial aggregation of a trait is evidence that that trait is likely heritable. For AMD, the sibling recurrence risk is 2.95, indicating that AMD risk has a genetic component [23]. At-risk individuals can receive regular eye exams to help detect AMD before visual loss occurs. While a family history of AMD indicates potential risk, specific genetic variants can also be used to predict AMD risk. Many ocular diseases, including AMD, have had risk loci identified [24]. AMD was first reported to have a genetic component in 1966 [25]. The next major genetic finding occurred in 2005 with the simultaneous release of three *Science* papers, reporting the association between complement factor H (*CFH*) and risk for AMD [26–29]. Later in 2005, *ARMS2* (then referred to as *LOC387715*) was associated with risk for AMD [30]. In 2006, a haplotype in which two complement factor H-related genes were deleted (CFTR1 and CFTR2) was associated with decreased risk for AMD [31]. By 2007, complement factor B (*CFB*), complement factor 2 (*C2*), and complement factor 3 (*C3*) were also associated with AMD [32–34]. Not all associated variants necessarily increase risk for AMD, however. Variants in *C2* (rs547154: GT; Odds Ratio (OR): 0.57) and *ARMS2/HTRA1* (rs3750847: CC; OR: 0.47) can indicate reduced risk for AMD [35] and variants in *CFH* (rs1061147: AA; OR: 2.76), *C3* (rs2230199: CC; OR: 2.38), and *TIMP3* (rs96215532: AA; OR: 1.02) can indicate increased risk for AMD [35].

Recent genome-wide studies

Since 2007, additional variants have been associated with AMD and explain some of the heritability of AMD. Heritability of AMD has been estimated to be between 45% and 70% [36], with the most recent estimate from a large-scale AMD study being 46.7% for advanced AMD in European-ancestry individuals [37].

In one study, twelve known risk loci were found to explain approximately 39% of the total risk for advanced AMD (55% of the heritability) [38]. More recently, a meta-analysis confirmed association with seven new loci, bringing the total number of loci replicating at a genome-wide significant level to nineteen [36]. The nineteen loci (Table 1) explain 30% of the risk for AMD (65% of the heritability), based on an AMD prevalence of 10% [36]. While much progress has been made, genetic effects discovered thus far do not explain all of the genetic risk for AMD and more research is warranted to uncover additional genetic effects, with the ultimate goal of elucidating molecular mechanisms to aid in the develop of better treatment and prevention options.

Table 1. Nineteen known risk SNPs and nearby gene information.

RS Number	Chr.	Position	Nearby Genes	Distance to index SNP (kb)*	Location**
rs10490924	10	124214448	*ARMS2* *HTRA1*	0 6.6	Coding Upstream
rs10737680	1	194946078	*CFH*	0	Intronic
rs429608	6	32038441	*C2* *CFB* *SKIV2L*	17 10.6 0	Downstream Downstream Intronic
rs2230199	19	6718387	*C3*	0	Coding
rs5749482	22	31389665	*TIMP3* *SYN3*	137.1 0	Upstream Intronic
rs4420638	19	45422946	*APOE* *APOC1*	10.3 5	Downstream Downstream
rs1864163	16	55554734	*CETP*	0	Intronic
rs943080	6	43934605	*VEGFA*	72.4	Downstream
rs13278062	8	23082971	*TNFRSF10A*	0.3	Upstream
rs920915	15	58688467	*LIPC*	35.7	Upstream
rs4698775	4	110590479	*CFI* *CCDC109B*	71.4 0	Downstream Intronic
rs3812111	6	116443735	*COL10A1*	0	Intronic
rs13081855	3	99481539	*COL8A1*	0	Intronic
rs3130783	6	30774357	*IER3* *DDR1*	62 77.5	Upstream Upstream
rs8135665	22	38476276	*SLC16A8*	0	Intronic
rs334353	9	100948186	*TGFBR1*	0	Intronic
rs8017304	14	68785077	*RAD51B*	0	Intronic
rs6795735	3	64705365	*ADAMTS9* *ADAMTS9-AS2* *MIR548A2*	32 0 0.3	Upstream Intronic Upstream
rs9542236	13	30717325	*B3GALTL*	0	Intronic

Adapted from Fritsche 2013 [36]. Information on allele frequencies for each SNP is given in Appendix A. * Distance from SNP to nearest (or resident) gene. ** Location in regards to nearest (or resident) gene.

CONCLUSIONS

Much research has been dedicated to understanding the biology and genetics of AMD. More effective treatment options are needed for wet AMD, and essentially no treatment options exist for dry AMD. As the average age of the human population worldwide continues to increase the prevalence of AMD will also continue to increase unless better treatment and/or preventative measures are developed.

AMD is highly heritable and many genetic variants have been statistically associated. Cellular processes, such as complement activation, have been associated with AMD through genome-wide association studies, but how these processes contribute to pathogenesis is not fully understood. It may be the case that, while there are many genome-wide significant SNPs associated with risk for AMD, other genetic effects might be due to genetic variation with lower effect sizes localized to particular pathways. In particular, we focus on the gene *ARMS2*, which harbors a statistically significant SNP, but it's biological relevance and contribution to pathogenesis is unknown.

In the following chapters I seek to better elucidate the genetic factors of AMD to try to uncover plausible mechanisms of pathogenesis and better link genetics of AMD with the biology of AMD. **First**, I will describe how mixed linear models can be used to estimate heritability genome-wide or by assessing specific genomic regions (genomic partitioning). **Second**, I will use such modeling methods to partition heritability of AMD into potentially-relevant genetic pathways. This should confirm known mechanisms (e.g. complement system) as well as estimate the overall contribution to AMD risk from other plausible pathways, some of which harbor no independent genome-wide significant risk

SNPs. **Third**, with the goal of exploring epistasis in AMD, I will first describe new methods (iGRM and iSim) I developed to estimate cumulative effects on risk from dominance and epistatic genetic effects. **Fourth**, and lastly, I will use those developed methods to explore the possibility that *ARMS2* modulates risk for AMD through interactions with biologically relevant genetic pathways.

Chapter 2 — Mixed Model Approaches To Heritability Estimation and Genetic Association

Background

Before DNA was discovered or the study of genetics had begun investigators observed that some traits were heritable and seemed to be passed from parent to offspring. Height is a clear example that it is possible to observe heritability without the need for genetic information; a person's height is highly correlated with the height of their parents, implying high heritability. Similarly, diseases that tend to cluster within families (i.e. familial aggregation) likely have a genetic component, barring potential environmental effects.

The concepts of how traits were inherited began with Gregor Mendel, the father of genetics, who used experiments in pea plant hybridization to show that additive and dominant characters (now referred to as alleles) allowed for physical pea plant characteristics to be predicted based on characteristics from previous generations [39]. Fast forward to today and we have a *much* more detailed understanding of genes, DNA, and how they are inherited.

Heritability can be defined quantitatively as the proportion of phenotypic variation due to genotypic variation; so if a trait is 100% heritable, that trait can be perfectly predicted using genetic information alone. Twins provide a special case for allowing heritability to be estimated. Because twins are exposed to nearly the same environmental factors (including shared intrauterine environment, parenting style, wealth, culture, and time period), genetic factors can be better interpreted and quantified. Monozygotic (MZ, identical) twins are essentially genetically identical, so if a trait is heritable the twins should have similar trait characteristics. Dizygotic (DZ, fraternal) twins, however, only share about half of their genes, on average. Using Falconer's formula [40] (Equation 1) heritability can be calculated as twice the difference between monozygotic and dizygotic twin correlation (r_{mz} and r_{dz}, respectively).

$$h^2 = 2(r_{mz} - r_{dz}) \tag{1}$$

Heritability, can be further broken down into two types — broad-sense (H^2) and narrow-sense (h^2) heritability. Whereas broad-sense heritability is defined as the amount of heritability due to all genetic effects, narrow-sense heritability is the amount of heritability due to additive genetic effects only. Whether twin studies estimate broad or narrow-sense heritability is debatable because only monozygotic twins share dominant and epistatic (gene-gene) effects in addition to additive effects. Due to crossing over during meiosis, dizygotic twins are expected to share less than half of possible epistatic effects. However, some have proposed a simple weight that could account for some of this discrepancy and better estimate broad-sense heritability [41] from twin studies.

Twins are not the only level of familial relation that can be used to estimate heritability, though. Sibling recurrence risk is one way to gauge whether a trait has a genetic component. Another method is to use trios (mother, father, and child) to regress the difference, for a quantitative trait, between the child's phenotype and the mid-parent phenotype (average between mother and father) (Figure 6).

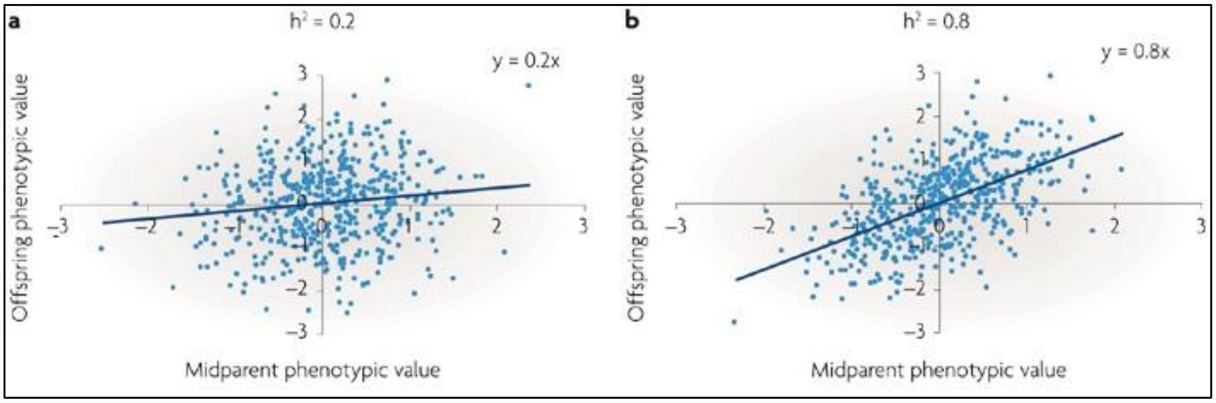


Figure 6. Estimating heritability from parent-offspring relationships. Example figure from [42]. In this example, an arbitrary phenotype is used, with values ranging from negative to positive three. Frame a shows lower heritability (20%) due to less correlation between the mid-parent value and the offspring value, while frame b shows more correlation and correspondingly higher heritability (80%), both calculated using linear regression. Each point represents one trio.

With the advent of DNA genotyping and sequencing and newer statistical methods, such as Genome-wide Complex Trait Analysis (GCTA) [43], it is now possible to estimate heritability for a trait using genotyped single nucleotide polymorphisms (SNPs) from groups of unrelated individuals. However, it is important to note that such methods can only provide heritability estimates, often referred to as "chip heritability" since estimates are derived from SNPs genotyped on a "SNP chip". Heritability, this way, is estimated as the proportion of trait variation due to measured genetic variation. Essentially, the more heritable a trait is the more correlated phenotypic similarity should be with genotypic similarity when comparing all pairs of individuals in a group. GCTA is often included as a component of genetic studies (Table 2) as an efficient and relatively easy way to determine the extent to which a trait has a genetic component. In addition, several similar, alternative methods exist, such as ACTA [44], EMMAX [45], FaST-LMM [46], GEMMA [47], and GRAMMAR-Gamma [48]. Below, we present scenarios where

GCTA can be used to solve certain questions. Then, we describe advantages and disadvantages of GCTA and briefly discuss alternatives.

Table 2. Example findings from studies that used GCTA.

Trait Studied	PVE*	PMID
Childhood Adiposity	30%	23528754
Drug Dependence	36%	25424661
Height	45%	20562875
Intelligence (From age 11)	62%	22258510
Multiple Myeloma	15.2%	26208354
Multiple System Atrophy	4.37%	23874384
Psoriasis (in Han Chinese)	45.7%	26172869
Pulmonary Function Measures	41.6 – 71.2%	25745850
Schizophrenia	39%	26198764

*Proportion of Variance Explained

ESTIMATING HERITABILITY WITH GENOME-WIDE COMPLEX TRAIT ANALYSIS

Overview

Genome-wide Complex Trait Analysis (GCTA) is a computer program that uses a mixed model approach to estimate heritability as the proportion phenotypic variance due to cumulative additive genetic variance [43]. The method was first applied to a study of human height where 45% of the variance was explained by the additive effect of ~300k common SNPs in 4,259 individuals [49]. A key feature of GCTA is its use of a genetic relationship matrix (GRM; Equation 2) to estimate the relatedness of all individuals by considering the additive effect of all SNPs simultaneously. Genetic effects can be partitioned multiple ways, such as by chromosome, by gene, or by pathway, by creating GRMs specific to certain genomic regions.

GRMs are calculated using a variance-covariance matrix, with the denominator being the SNP population variance and the numerator being the covariance between two individuals. A_{jk} is the calculated genetic relationship value for individuals j and k; N is the number of non-missing SNPs shared between two individuals; x_{ij} is the number of copies of the reference allele for the ith SNP of the jth individual; p_i is the frequency of the reference allele. The effects from each SNP are then summed (Σ) and weighted equally (1/N). The equal weighting assumes SNPs are independent, which is generally true for genome-wide SNP chips which typically have a wide, representative coverage of the genome. Unlike typical genome-wide association studies (GWAS), in which the effects of single SNPs are calculated independently, GCTA accounts for relative levels of genomic sharing between individuals.

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)} \qquad (2)$$

Using GCTA

Availability and versioning

To date, GCTA is still being improved and updated; download links and documentation are currently hosted by The University of Queensland's Centre for Neurogenetics and Statistical Genomics (http://cnsgenomics.com/software/gcta). As of November 2015, the latest version of GCTA is 1.25.0. In addition, the online GCTA forum can be used as a valuable resource when errors are encountered or questions arise (http://gcta.freeforums.net). GCTA options are often similar to PLINK [50], a widely used genetic data manipulation and analysis software. In the following sections we review various analyses that can be performed using GCTA, along with examples of corresponding options and commands.

Data management and manipulation

The two types of genetic data files that GCTA can use as input are PLINK binary files (bfile) and MACH output files (dosage-mach and dosage-mach-gz for uncompressed and compressed MACH files, respectively). Non-binary PLINK files (*.ped/*.map) should be converted to binary format using PLINK [50]. Genetic data loaded into GCTA can then be filtered in various ways, such as filtering in/out individuals (keep/remove), filtering by specific chromosome or all autosomes (chr/autosome), filtering in/out SNPs (extract/exclude), filtering by minor allele frequency (maf/max-maf), and filtering for imputed data (imput-rsq). Output file names are specified the same way for every command (out). Any missing data points should be coded as “-9” or “NA”. More detailed specifics on formatting input data can be found in GCTA’s online documentation (http://cnsgenomics.com/software/gcta).

Estimating genetic relationships via GRMs

Above, we described the calculations involved in generating genetic relationship matrices (GRMs). In combination with any SNP/individual inclusion/exclusion options, GRMs can be calculated and output in binary compressed format (make-grm). Binary compressed format is the default GRM option — the older compressed text file version can also be created using a slightly different option (make-grm-gz). Importantly, GRMs are intended for autosomal SNPs only, due to the statistics involved; however, GCTA has an option which uses slightly modified calculations (different for male-male, male-female, and female-female pairs) [43] to make a separate GRM for SNPs on the X-chromosome (make-grm-xchr). At this time, GCTA does not explicitly support creating GRMs for mitochondrial SNPs, however mitochondrial SNPs can be dummy coded as X-chromosome SNPs to generate a mitochondrial GRM, but it is important to remove actual X-chromosome SNPs so they are not included. Importantly, GCTA can utilize multiple processors simultaneously to greatly speed up GRM (thread-num). An example command to generate a GRM is:

```
gcta64 --bfile inputfile --make-grm --autosome --out outname --thread-num 10
```

Finally, an inbreeding coefficient (F) can be calculated for each individual as the average across all SNPs (ibc - inbreeding coefficient) using three different calculation methods based on (1) variance of additive genotype values, (2) excess homozygosity, and (3) the correlation between uniting gametes [43].

Estimating trait heritability and genetic effect significance

GCTA uses restricted maximum likelihood (REML; also called GREML) analysis to estimate heritability — or more precisely, the proportion of trait variance explained (PVE; for quantitative traits) or proportion of trait risk explained (PRE; for binary, case-control traits). REML is an iterative method that finds the best fit for a mixed linear model (Equation 3), where Y is the phenotype, X is any fixed variable, β is the fixed variable effect size, Z is the GRM, γ is the vector of random effects from the GRM, and ε is the residual random effect (representing environmental, non-genetic effects).

$$Y = X\beta + Z\gamma + \ \varepsilon \tag{3}$$

PLINK files (which usually contain the phenotype values) are not used directly for REML, instead, the phenotypes must be specified as a separate file (pheno). The disease prevalence for case-control datasets can be specified to provide a better estimate of PRE (prevalence). Both discrete and continuous covariates can be included in the analysis for adjustment (covar and/or qcovar, respectively). A newer bivariate REML analysis method [51] allows two traits to be simultaneously fit to detect potential pleiotropy between the two traits. To determine significance of any genetic components (i.e. GRM) GCTA performs a likelihood ratio test (LRT), comparing full and reduced models, where the reduced model is created by dropping the genetic variance component (GRM) of interest from the full model. If only one GRM is used, that component will be dropped by default to create the reduced model. If multiple GRMs are used, any components the user wishes to test for significance can be specified (reml-lrt). To perform REML for a case-control dataset with a disease prevalence of 5%, two GRMs, covariates, and a likelihood ratio test for the second GRM, the following example command could be used:

```
gcta64 --reml --reml-lrt 2 --mgrm grmlistfile --pheno phenotypefile
--prevalence 0.05 --qcovar covariatefile --out outname
```

Principal components analysis

Sometimes it is useful to adjust for genetic ancestry to avoid confounding due to population stratification. GCTA performs principal component analysis (pca) in the same way as EIGENSTRAT [52], another popular tool for calculating principal components. The output is an *.eigenvec file, which includes principal component values that can be included as covariates in any analyses.

Estimating LD structure

GCTA can be sensitive to linkage disequilibrium (LD) and heritability can be underestimated or overestimated in influential regions with high or low LD. Lee et al. [53] and Purcell et al. [54] suggest that LD has a relatively minimal effect and propose a minor allele frequency (MAF) stratification approach. Speed et al. [55, 56] shows, through simulations, that GCTA-type analyses are robust as long as LD is similar for causal and non-causal regions. In regions where there is high LD near causal variants, heritability is overestimated — the opposite is true in areas of low LD. A modified method, linkage disequilibrium adjusted kinships (LDAK; www.ldak.org), can be used as an alternative method of generating a GRM, which generates a modified kinship matrix by weighting SNPs based on local LD patterns [55]. Alternatively, PLINK's built-in LD pruning option can be used to filter SNPs based on LD, keeping only representative SNPs using a given LD threshold [50], lessening the potential for confounding due to LD.

Estimating individual SNP effects

GCTA has two primary ways to assess individual-level SNP effects — mixed linear model based association analysis (MLMA) and calculation of best linear unbiased predictions (BLUPs). MLMA provides an advantage over typical linear regression by conditioning the effect of a given SNP on other genotyped SNPs, simultaneously, using a GRM. MLMA provides the estimated effect size, standard error, and p-value for each SNP. An example command to perform MLM association analysis is:

```
gcta64 --mlma --bfile plinkfile --grm grmfile --pheno phenotypefile
--out outname --thread-num 10
```

Similarly, SNP effects can be estimated by first calculating BLUPs for individuals (reml-pred-rand), then transforming those BLUP solutions to estimate BLUPs for SNPs (blup-snp). The final output is a list of all SNPs, with the reference allele, residual effect, and one SNP effect for each GRM used.

Estimating power

A study by Visscher et al. [57] uses complex theory and simulations to compare genetic and phenotypic sampling variance for different population sizes to model / predict power based on dataset characteristics. The power calculator tool is hosted online (http://cnsgenomics.com/shiny/gctaPower). Assuming the use of the default type 1 error rate of 0.05, the two pieces of information needed to estimate power are sample size and the variance of SNP-derived genetic relationships, the latter of which can be found in output after generating a GRM. The variance is calculated as the variance of the off-diagonal genetic relationship values in the GRM (the diagonal values, which represent genetic relationship values for self-pairs, similar to inbreeding values, are excluded). Those factors are then used to calculate power. For example, with a sample size of 7,777 individuals and an off-diagonal variance (ODV) of 0.00321, there is 87.6% power to detect heritability as low as 1%.

Advantages and disadvantages

Overall, GCTA is a powerful tool that is relatively easy to use and provides several advantages over typical single SNP analyses. Genome-wide association studies (GWAS) can have false-positive results due to geographic population structure, family relatedness, or cryptic relatedness [58]. In addition, when thousands or millions of variants are individually tested for trait association false-positive results will inherently be introduced. GCTA avoids this by utilizing the genetic structure within a dataset and performing a single model test. In addition, mixed model analyses in GCTA condition on non-candidate loci, increasing power for datasets regardless of whether or not population structure is present [58].

One disadvantage, however, is that power is reduced when conducting mixed linear model association analysis (MLMA) when the candidate SNP is included in the GRM. Sometimes referred to as “proximal contamination” [59], the loss of power is due to the marker being fit both as a fixed and random effect, simultaneously. Analysis with the candidate marker included in the GRM is referred to as “MLMi”, while analysis excluding the candidate marker (the marker(s) of interest that is included as fixed effects) is referred to as “MLMe”. FaST-LMM [46] serves as an alternative, enabling MLMe, however GCTA now can perform MLMA-LOCO (leave one chromosome out), where the chromosome that the candidate marker is on is not included in the model, helping to avoid potential proximal contamination. While MLMe is more powerful, MLMi is usually used due to greater computational efficiency.

Selection of which SNPs to include in GRMs is important. Above, we described the risk in including candidate loci in GRMs. Several studies [46, 59, 60] have also suggested that MLMA can gain power by carefully choosing subsets of SNPs to include in GRMs. However, a more recent study [58] used simulations to show that limiting GRMs to a subset of SNPs can compromise correction for population stratification, negating one of the benefits of mixed model analysis. In the case where a subset of SNPs are used, it is important to include principal components as fixed effects to account for potential population stratification.

Finally, case-control studies can suffer from decreased power due to unintentional correlation between case status and collected, relevant covariates, due to the ascertainment process, especially when disease prevalence is low [61]. A recent study [62] proposed methods for liability-threshold mixed linear model (LTMLM) association analyses for case-control datasets in which χ^2 statistics are calculated from posterior mean liabilities (PMLs), conditioned on each individual’s case status, dataset case-control status information, and genetic relationship matrices (GRMs). Heritability is then estimated using Haseman-Elston regression and transformed using the adjusted liability scale.

SUMMARY

The abundant availability of genetic analysis software has helped progress the study of genetics for many diseases and traits, but such a wide range of options can be overwhelming when trying to decide what is best for a given study. Advancements in statistical genetics have made it possible to estimate heritability using genotype data from unrelated individuals, making study participant recruitment easier than for twin studies. GCTA is a valuable and increasingly used tool that can be extremely powerful, but careful study design and a sufficient understanding of the concepts behind a given analysis is needed to make appropriate conclusions and avoid misinterpretation of results. In the following chapters I use GCTA and new methods related to GCTA to explore genetic effects of age-related macular degeneration (AMD).

CHAPTER 3 — GENETIC PATHWAY ANALYSIS OF AGE-RELATED MACULAR DEGENERATION [1]

INTRODUCTION

Multiple mechanisms have been proposed as having a role in AMD pathogenesis. A recent review of AMD by Fritsche et al. describes many risk factors and mechanisms [63]. The discovery of the association between the Complement Factor H gene and AMD led to further associations between other genes related to complement activation [64]. Inflammation is highly related to complement activation and can lead to apoptosis of retinal pigment epithelial (RPE) cells and photoreceptors [65]. Using terminal deoxynucleotidyl transferase dUTP nick end-labeling (TUNEL), Dunaief et al. found that RPE cells, photoreceptors, and inner nuclear layer cells can die through apoptosis during AMD progression [66]. The Age-Related Eye Disease Study (AREDS) showed that antioxidant and zinc vitamin supplements were able to slow AMD progression [18], implicating antioxidant mechanisms as candidates in disease progression. These include intermediates of the tricarboxylic acid cycle (TCA cycle), which can alter the effectiveness of zeaxanthin (a component of AREDS2 supplements) [67]. Zhao and Vollrath showed that when mitochondria in RPE were ablated in mice, the lack of oxidative

[1] This chapter is adapted from my peer-reviewed articled titled “Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models”, published in *BMC Bioinformatics* [1].

phosphorylation (OxPhos) in the RPE led to photoreceptor degeneration [68]. Angiogenesis plays a significant role in choroidal neovascularization (CNV), and anti-VEGF treatments, which aim to inhibit angiogenesis, are used as treatment for wet AMD [18]. Finally, smoking is a well-known risk factor for AMD [69], and thus nicotine metabolism may plausibly play a role in AMD pathogenesis. While there is substantial evidence that complement activation plays a major role in AMD, the correlation between genetic and biological mechanisms for the others are less stablished.

Genetic variants with large effect sizes, several of which are localized to complement system genes, have been repeatedly associated with AMD [27, 28, 32, 36, 64]. However, AMD-associated SNPs that reach genome-wide significance only account for a portion of the known heritability [36]. SNPs with smaller effects likely contribute cumulatively to an additional portion of the heritability. While overall heritability estimates of AMD are known, the estimated contribution to heritability, separately, for many AMD-related pathways is unknown. Existing genetic pathway analysis methods typically annotate SNP associations using databases such as the Gene Ontology (GO) [70], Ingenuity Pathway Analysis (IPA) [71], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [72], or Reactome [73]. These methods then utilize analytical approaches, such as Gene Relationships Across Implicated Loci (GRAIL) [74] or Pathway Analysis by Randomization Incorporating Structure (PARIS) [75], to determine the significance of pathways, usually using gene or SNP p-values or genotype data to calculate a rank-based pathway statistic [76]. These methods, however, do not provide a scaled measure of the effect and thus do not offer estimates of heritability or the proportion of overall disease risk explained by an entire pathway. In this study, using a case-control AMD cohort we estimate the significance and proportion of risk explained by additive genetic effects within specific AMD-related pathways to prioritize them for future molecular and epidemiological studies.

METHODS

Dataset summary

Subjects in this study (Table 3) were recruited from the Duke University Eye Center (DUEC), the Vanderbilt Eye Institute (VEI), and the Bascom Palmer Eye Institute (BPEI) at the University of Miami Miller School of Medicine starting in 1995, 1999, and 2007, respectively. Individuals were recruited through retinal clinics, mostly via referrals for possible AMD; recruitment was performed under research protocols approved by the appropriate institutional review boards at each institution, and written informed consent was obtained from all participants. Original recruitment was performed for a previous study of AMD [77] and permission to use the dataset for this study was obtained. Controls were recruited either as friends or spouses of cases or through regular eye exams. Examination, imaging, and grading were performed prior to the start of this analysis. All subjects were examined by a retina specialist using slit-lamp biomicroscopy and dilated fundus examination, including indirect ophthalmoscopy. Additionally, fundus imaging was analyzed to confirm case status. For consistency between sites, images were scored by a retina specialist using a modified grading system based on the Age-Related Eye Disease Study (AREDS) [78]. The grading system was used to score individuals on a scale between 1 and 5. Subjects with grades 1 and 2 were considered controls and subjects with grades 3 through 5 were considered cases, with grade 3 representing early AMD (non-neovascular) and grades 4 and 5 representing late AMD (GA and CNV, respectively). Both eyes were scored and an individual's overall grade was determined using the eye with the higher grade.

Table 3. Study population characteristics.

Cohort	Age[*] (SD)	Males (%)	Smokers (%)
Primary subset - 1,813 (100%)	75.1 (8.4)	713 (39.3)	—
Cases - 1,145 (63.2%)	77.6 (7.9)	415 (36.2)	—
Controls - 668 (36.8%)	70.9 (7.7)	298 (44.6)	—
Smoking subset - 1,358 (100%)	75.0 (8.2)	560 (41.2)	790 (58.2)
Cases - 850 (62.6%)	77.3 (7.7)	323 (38.0)	516 (60.7)
Controls - 508 (37.4%)	71.2 (7.6)	237 (46.7)	274 (53.9)

[*] Mean age in years

Primary cohort contains all individuals after QC measures were applied. The proportion of smokers in the primary subset is not shown since smoking status was not available for everyone. The smoking subset excludes individuals with unknown smoking status.

Genotyping and quality control

Three genotyping platforms were used: the Affymetrix 1M array (906,600 SNPs), a custom Sequenom array (84 SNPs), and custom TaqMan assays (4 SNPs). The Sequenom array was designed to interrogate potential AMD-related SNPs, while the TaqMan assays were used later to validate SNPs that performed poorly on the Sequenom array. SNP quality control (QC) was performed separately for Affymetrix SNPs and for merged Sequenom/TaqMan SNPs and was applied simultaneously to cases and controls. For the Affymetrix genotyping chip, 38,443 non-autosomal SNPs were removed, 102,735 SNPs with genotyping efficiency less than 95% were removed, 104,695 SNPs with a minor allele frequency (MAF) less than 1% were removed, 1,475 SNPs with Hardy-Weinberg

Equilibrium (HWE) p-values less than 1×10-6 were removed, 121 SNPs not able to be converted from genome build 36 to 37 using liftOver [79] were removed, and 25 Affymetrix SNPs that were present [46] in post-QC Sequenom/TaqMan SNPs were removed, resulting in 659,106 post-QC Affymetrix SNPs. QC procedures were applied to 88 merged Sequenom/TaqMan SNPs for 1,911 individuals that also had Affymetrix data. Forty-five individuals were removed that had genotyping efficiency less than 90%, leaving 1,866 individuals. For the merged data, 4 non-autosomal SNPs were removed, no SNPs had a genotyping efficiency less than 95%, 7 SNPs with a MAF less than 1% were removed, and 2 SNPs with a HWE p-value less than 1×10-6 were removed, leaving 75 SNPs for analysis. All merged genotype platforms resulted in a total of 659,181 SNPs for analysis.

All 1,967 individuals in our dataset were observer-reported to be white (European American), however we performed principal components analysis using 71 ancestry informative markers, seeding with six distinct HapMap phase 3, release 3 populations [80], to confirm genetic ancestry (Figure 7; Appendix B). Twelve individuals with non-European American genetic ancestry were removed to avoid potential confounding by population stratification, including eleven with African genetic ancestry and one with Asian genetic ancestry (Figure 7). Additionally, five individuals were removed that had genotyping efficiency less than 90%, based on Affymetrix genotype data, 84 individuals were removed that did not have available Sequenom/TaqMan genotype data, and 53 individuals were removed that did not have age recorded at time of examination, leaving 1,813 individuals for analysis (1,145 cases and 668 controls). Finally, some of our analyses required individuals to have known smoking status, with individuals considered to be smokers if they had smoked 100 or more cigarettes in their life; 455 individuals did not have available smoking status information, leaving 1,358 individuals for smoking status adjusted analyses (Table 3). The distribution of age (Figure 8) was significantly different between cases and controls, based on a Kolmogorov-Smirnov 2-sample test of equal distributions (K-S Statistic: 0.364; P-value: 1.15 E-49). Due to this difference we adjusted for age (in years) for all analyses.

Table 4. Gene Ontology terms used to define pathways.

GO Term	GO ID	# Genes	PMID
Angiogenesis	GO:0001525	379	23642783
Antioxidant Activity	GO:0016209	69	23645227
Apoptotic Signaling	GO:0097190	1,635	12427055
Complement Activation	GO:0006956	187	20711704
Inflammatory Response	GO:0006954	534	17021323
Response to Nicotine	GO:0035094	31	8827967
Oxidative Phosphorylation	GO:0006119	78	21483039
Tricarboxylic Acid Cycle	GO:0006099	33	14962143

Number of overlapping SNP between pathways is presented in Figure 12.

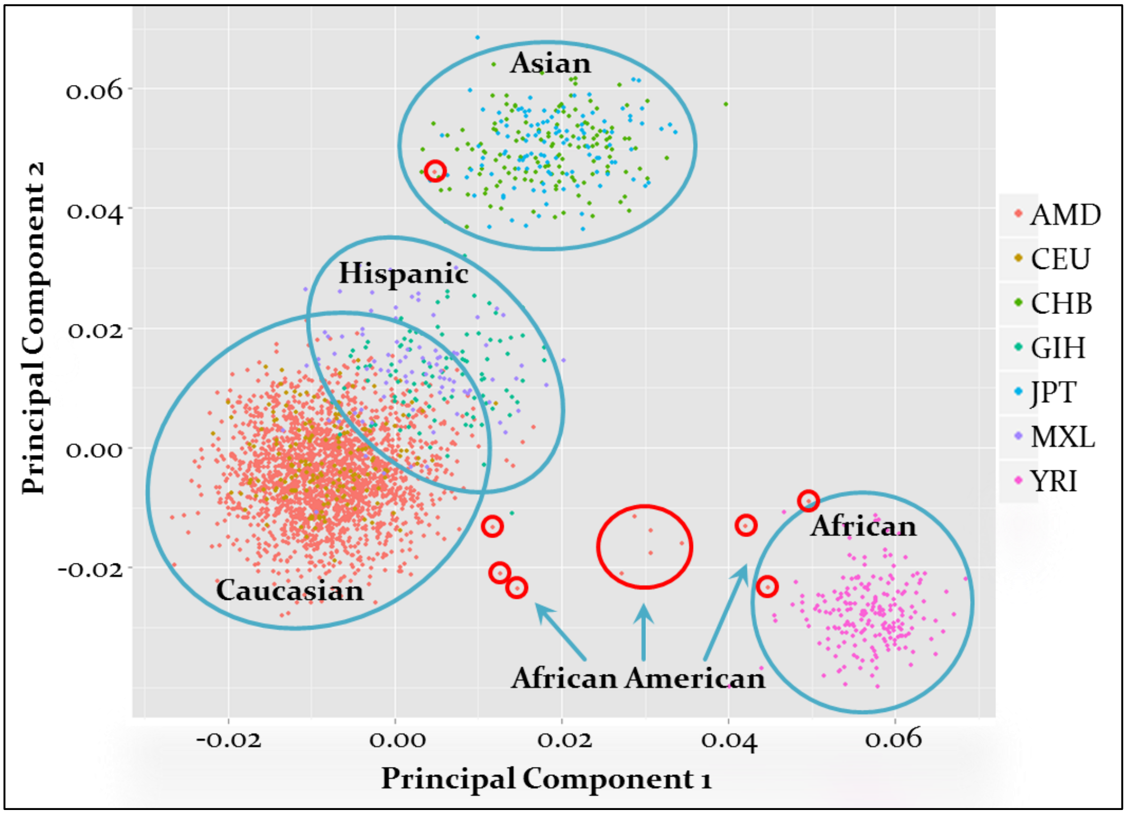


Figure 7. Genetic ancestry from principal component analysis.
AMD samples plotted with HapMap samples to confirm genetic ancestry. Ethnic group clusters circled in teal. Outliers circled in red.

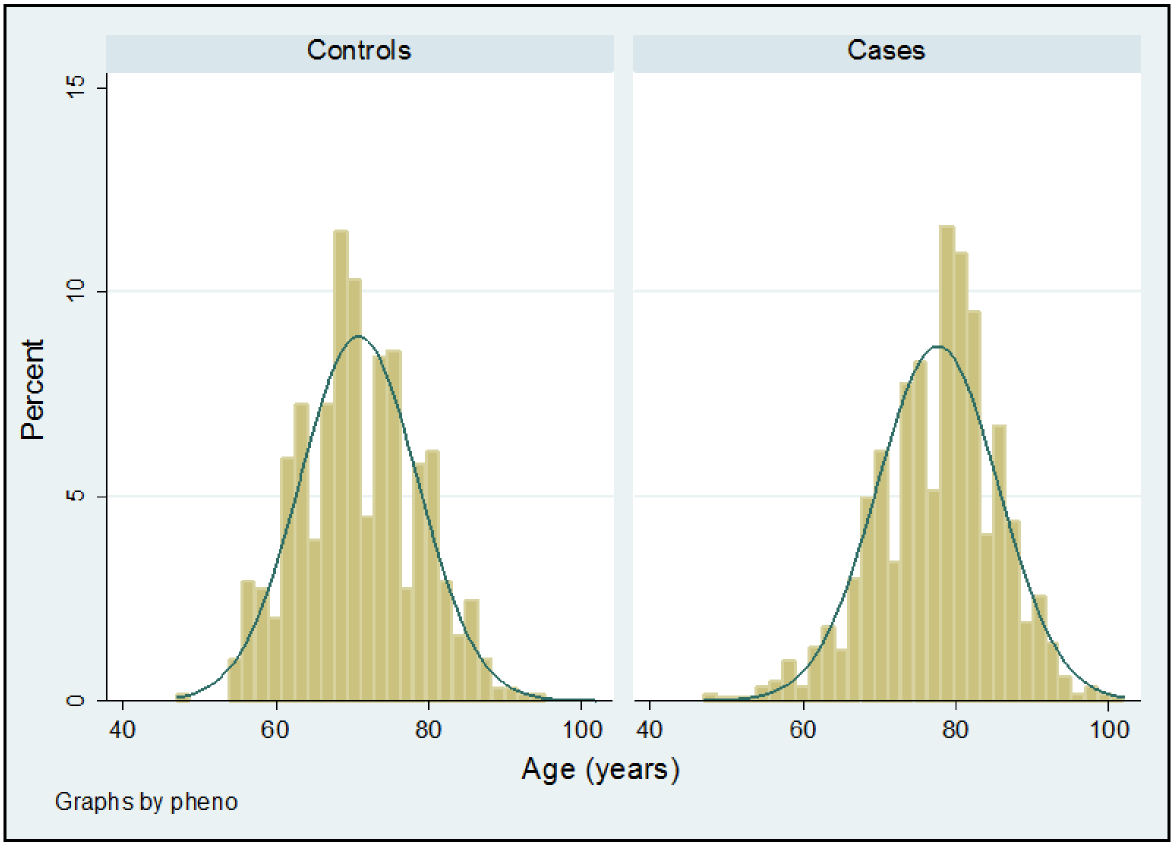
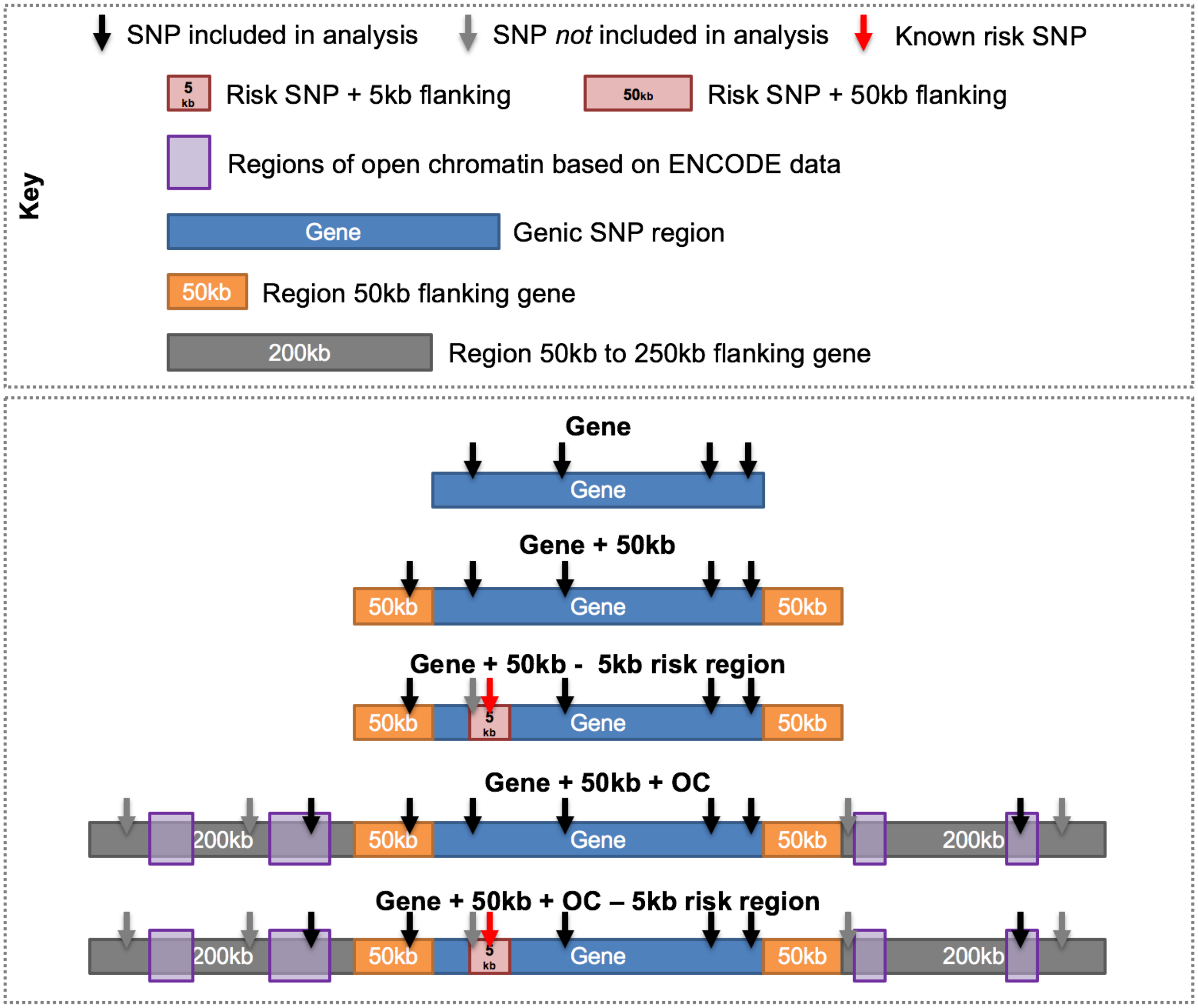


Figure 8. Histogram of age, by case status.
Age in years recorded at time of examination. Histogram for 668 controls and 1,145 cases. Individuals with no smoking status not excluded.



Key
SNP included in analysis
SNP *not* included in analysis
Known risk SNP
5 kb
Risk SNP + 5kb flanking
50kb
Risk SNP + 50kb flanking
Regions of open chromatin based on ENCODE data
Gene
Genic SNP region
50kb
Region 50kb flanking gene
200kb
Region 50kb to 250kb flanking gene
Gene
Gene + 50kb
Gene + 50kb - 5kb risk region
Gene + 50kb + OC
Gene + 50kb + OC – 5kb risk region
50kb
200kb

Figure 9. Details of subsets of gene regions analyzed.
Open chromatin regions determined using narrow peak windows from ENCODE DNaseI hypersensitivity analyses in human RPE cells.

Pathway selection and curation

For this study our goal was to determine the overall contribution of several pathways on AMD risk, to both confirm the importance of known mechanisms (e.g. complement activation) and to determine if some biological mechanisms contribute to cumulative AMD risk without harboring individual genome-wide significant, large-effect genetic variants. Based on an extensive literature search and advice from AMD experts, we chose eight mechanisms ranging from having plausible to extremely well-known AMD relation to test as pathways (Table 4) in our analysis.

The Gene Ontology (GO) [70] is a database of hierarchical gene relationships. To objectively determine genes related to each of the eight selected pathways we selected appropriate GO terms corresponding to each pathway (Table 4) and extracted all associated genes (Appendix C) falling under the hierarchy of that GO term using the November 2013 release of the GO database. Because GO is hierarchical, containing parent-child-type relationships, we included all descendants of the selected GO terms as to not omit directly related genes. For each assigned gene we tested three partitioned regions to represent the effect of that gene (Figure 9), including (1) SNPs within Ensembl-defined gene boundaries, (2) SNPs within 50 kb flanking each gene boundary (to capture cis-regulatory SNPs), and (3) SNPs within 50 kb and 250 kb flanking each gene that also lie within open chromatin regions based on ENCODE DNaseI hypersensitivity analyses of human retinal pigment epithelial cells (hRPEpiC) [81].

Mixed linear model analysis

To estimate the proportion of AMD risk explained by each pathway, we used Genome-wide Complex Trait Analysis (GCTA) [43], described in Chapter 2, to fit genetic relationship matrices (GRMs) using mixed linear models (MLMs) via the restricted maximum likelihood (REML) method. For many analyses we tested three different REML algorithms — average information (AI), Fisher-scoring, and expectation maximization (EM); here, we will only show results using the EM algorithm, which was computationally slower but, for our analyses, yielded models that converged more consistently than the other two algorithms. For all analyses we included age, sex, and the first two principal components as covariates. For case-control analyses, GCTA by default uses disease prevalence rates observed within a dataset; however, it is recommended to use prevalence rates from general populations based on literature. We used a prevalence rate (Table 5) of 5.07%, calculated by weighting all individuals in our dataset with U.S. prevalence rates, stratified by age [3]. The proportion of risk explained is then transformed from the observed scale to the specified prevalence scale. Linkage disequilibrium (LD) has a minimal effect on estimates from GCTA, with studies showing that cumulative estimates are stable and not necessarily over-inflated because both influential and non-influential SNPs in LD are considered and therefore possible confounding effects are neutralized [54, 82]. To explore potential LD effects within our study, we perform additional analyses on SNP sets pruned using LD.

We estimated the overall proportion of risk for AMD explained, as well as the proportion of risk explained by each pathway for various gene regions and exclusion criteria (Figure 9). We explored effects of LD, SNP overlap between pathways, smoking status, and stratification by AMD subtype on the proportion of AMD risk explained, either cumulatively or by pathway. The following are more detailed methods for each specific analysis.

Table 5. Weighted-by-age, expected population prevalence calculation.

Age Range	U.S. Prev.	Count	Count × Prev.
40-49	0.05	4	0.2
50-54	0.34	3	1.0
55-59	0.39	62	24.2
60-64	0.56	140	78.4
65-69	0.91	262	238.4
70-74	1.66	352	584.3
75-49	3.24	398	1289.5
80+	11.77	592	6967.8
Total		**1813**	**9183.9**

United States prevalence rates (percent of population for given age range) from [3].
[9183.9 / 1813] = **5.07%** weighted, expected AMD prevalence rate.

Genome-wide AMD risk explained

The first analysis we performed was to assess the overall proportion of AMD risk explained by all available genotyped SNPs in our dataset (often referred to as “chip heritability”). One GRM was created for all 659,181 SNPs and was included in a mixed linear model analysis using GCTA, adjusting for the covariates described previously.

Known Risk SNPs

A recent meta-analysis [36] of AMD, as described in Chapter 1, found 19 genome-wide significant AMD risk SNPs (Table 1). To determine the effect that those 19 known SNPs have in our dataset we created a GRM consisting of just those 19 SNPs, referred to as the risk GRM, and a GRM for all other SNPs (659,162), referred to here as the remainder GRM. Additionally, we created risk GRMs that included 5 kb and 50 kb flanking (and including) the 19 known risk SNPs, to capture effects of SNPs in LD with those known risk SNPs, resulting in a total of 83 and 566 risk SNPs, respectively, with the remainder GRM being all SNPs minus the given risk subset.

Risk explained by pathways

To estimate the effect of the eight selected AMD-related pathways, two GRMs were generated, unless otherwise specified, for each analysis of each pathway. Pathway GRMs consist of SNPs being assessed for a respective pathway and remainder GRMs contain all other SNPs being considered that are not in the respective pathway GRM and that are not excluded. Many pathways have overlapping genes and thus effects from all pathways, separately, could not be estimated in a single mixed linear model. We assessed the effect for several gene regions (Figure 9), starting with just genic SNPs, then subsequently adding SNPs within 50 kb flanking each gene, and then SNPs in open chromatin regions within 50 kb to 250 kb flanking each gene, based on the ENCODE DNaseI hypersensitivity data from human retinal pigment epithelial cells (hRPEpiC). Additionally, for each pathway we performed analyses excluding 5 kb risk regions around and encompassing each of the 19 known risk SNPs from the regions including genic SNPs, SNPs within 50 kb flanking, and more distant SNPs in open chromatin regions. When known risk regions were excluded from a pathway GRM, they were not included in the remainder GRM but were rather excluded entirely, so as to determine cumulative, additional risk explained by pathways. Finally, we calculated the risk explained for each pathway adjusting for the number of SNPs in each pathway to ensure that the amount of risk explained was not simply due to the number of SNPs included in a given pathway.

Gene overlap

For this study it was not feasible to allow all pathways to have unique, non-overlapping gene sets. Thus, we tested the overlap between all pairs of pathways to determine whether risk explained was unique to certain pathways or shared between pathways due to sharing of common genes. For each overlapping pathway we created a GRM using overlapping SNPs and a GRM using non-overlapping SNPs, based on genic SNPs and 50 kb flanking.

Linkage disequilibrium near known risk SNPs

While we assessed excluding risk SNPs and 5 kb flanking those risk SNPs from each pathway, SNPs in more distant LD with those risk SNPs could influence the calculation of pathway GRMs and inflate estimates of the proportion of risk explained. Thus, we used LD information from CEPH individuals in HapMap phase II to exclude all SNPs in LD with the 19 known risk SNPs. We used exclusion criteria of $R^2 \geq 0.10$, 0.05, and 0.01 (based on LD from HapMap Phase II and III, using CEU samples only), much more strict than the typically used R^2 cutoff of ≥ 0.80, therefore removing SNPs with even minimal LD to known risk SNPs. Each SNP had LD information for other SNPs within a 500 kb flanking region. To be even more conservative we also excluded 1MB regions flanking each risk SNP. For each threshold we created a remainder GRM for all SNPs minus any matching the exclusion criteria. Results were compared to previous estimates of AMD risk explained by known risk SNPs and all other genotyped SNPs to estimate risk explained due to LD near risk SNPs. Each analysis included a risk GRM and a remainder GRM.

Effect of smoking status

Smoking is a major risk factor for the development of AMD [69], so we also ran additional analyses for each pathway, including smoking status as a covariate, to detect any differences in significance or amount of risk explained per pathway, when adjusting for smoking. Genic SNPs plus 50 kb flanking were used to compare effects. Of the 1,813 individuals used in this study 455 did not have available smoking status.

Stratification by AMD subtype

We ran analyses stratifying by AMD subtype to confirm that our dataset exhibits no AMD-subtype effect, especially considering that some pathways analyzed are by definition more related to a particular AMD subtype (e.g. angiogenesis is highly related to neovascular AMD). For these analyses we excluded individuals with early AMD (grade 3) and considered only controls versus grade 4 (CNV) and controls versus grade 5 (CNV in at least one eye). We tested genic SNPs plus 50 kb flanking plus open chromatin for both subtypes of AMD for each pathway.

RESULTS AND DISCUSSION

Genome-wide AMD risk explained

In our first analysis we used all 659,181 genotyped SNPs that passed QC to estimate the heritability of AMD in our dataset. We found that 61.5% (p-value = 3.4×10^{-5}; S.E. = 16.9%) of the risk for AMD in our dataset was explained by those SNPs, in range of known AMD heritability estimates. This confirmatory step helps validate subsequent pathway analyses in this study, showing that there is substantial variation in our dataset that impacts AMD risk. When assessed separately, the 19 previously associated AMD risk SNPs explained 13.30% of the risk for AMD in our dataset (p = 1.35×10-61) while all other genotyped SNPs explained 36.72% of the risk. Regions flanking the risk SNPs were also considered in separate analyses and explained a total of 15.37% (p = 1.59×10^{-53}) when 5 kb flanking the risk SNPs were included, and 16.33% (p = 8.24×10^{-44}) when 50 kb flanking regions were included.

The difference between the heritability estimates using a single GRM (61.5%) and two GRMs (13.30% + 36.72% = 50.02%) is non-intuitive but not unexpected and is due to sample size and differences in how genetic variance happens to be partitioned in the GRMs and subsequently fit by GCTA. Theoretically, as sample size (number of individuals) approached infinity the two estimates would converge.

From this we see that known risk SNPs explain only a portion of the overall risk estimate, indicating that additional lower-effect SNPs may influence disease risk. Additionally, the increase in risk explained (from 13.3% to 16.33%) shows that the estimate of the risk explained by the remaining SNPs (36.72%) could be a slight overestimate, but only by about 3.0%.

Risk explained by pathways

We first assessed the effect of each pathway for three different gene region inclusion criteria without excluding any known risk SNPs (Figure 10). The complement and inflammatory pathways explained between approximately 10% ($p < 1\times10^{-25}$) and 17% ($p < 1\times10^{-7}$), respectively, of the risk for AMD, while the angiogenesis and apoptotic signaling pathways explained nearly 5% of the risk (non-significant), and the antioxidant, nicotine, oxidative phosphorylation, and tricarboxylic acid cycle pathways explained approximately 2% of the risk or less (non-significant). In general, we observed that inclusion of SNPs within 50 kb flanking pathway genes typically increased the amount of risk explained, while additional inclusion of more distant SNPs in open chromatin regions did not explain a great deal more risk (Figure 10), suggesting that local regulatory SNPs indeed modulate risk.

We also assessed each pathway, excluding known risk SNPs and 5 kb flanking (referred to as risk regions) from regions including genic SNPs plus 50 kb flanking plus open chromatin SNPs, to better estimate novel risk explained by each pathway (Figure 10, green bars). We observed little reduction in the amount of risk explained by each pathway when the risk regions (and SNPs in LD) were removed. This is likely due to the fact that each pathway, separately, contains only a subset of the known risk SNPs. The response to nicotine, oxidative phosphorylation, and tricarboxylic acid cycle pathways contained no SNPs within risk regions, while other pathways contained at most 10 SNPs within risk regions to be removed, indicating that risk explained by each pathway is in addition to the amount of risk explained by the 19 known risk SNPs.

Notably, the number of genes and SNPs differs significantly over the pathways we targeted. When we adjusted the proportion of risk explained from each pathway by the number of SNPs contained within each pathway, we observed results consistent with known genetic contributors to AMD (Figure 11). Unsurprisingly, after adjusting for the number of SNPs in each pathway, the complement pathway explains the highest amount of risk per SNP. The antioxidant, nicotine, and oxidative phosphorylation pathways, which each explain less 2% of the risk for AMD, have similar levels of per-SNP effects (about 0.02%), on the same order of magnitude as the complement pathway (0.05%) and inflammatory pathway (0.03%). Overall, we see little cumulative effect of SNPs outside the complement and inflammatory pathways, but identify additional risk from complement and inflammatory mechanisms, due in part to variation within the flanking regions of these genes that is likely to be regulatory.

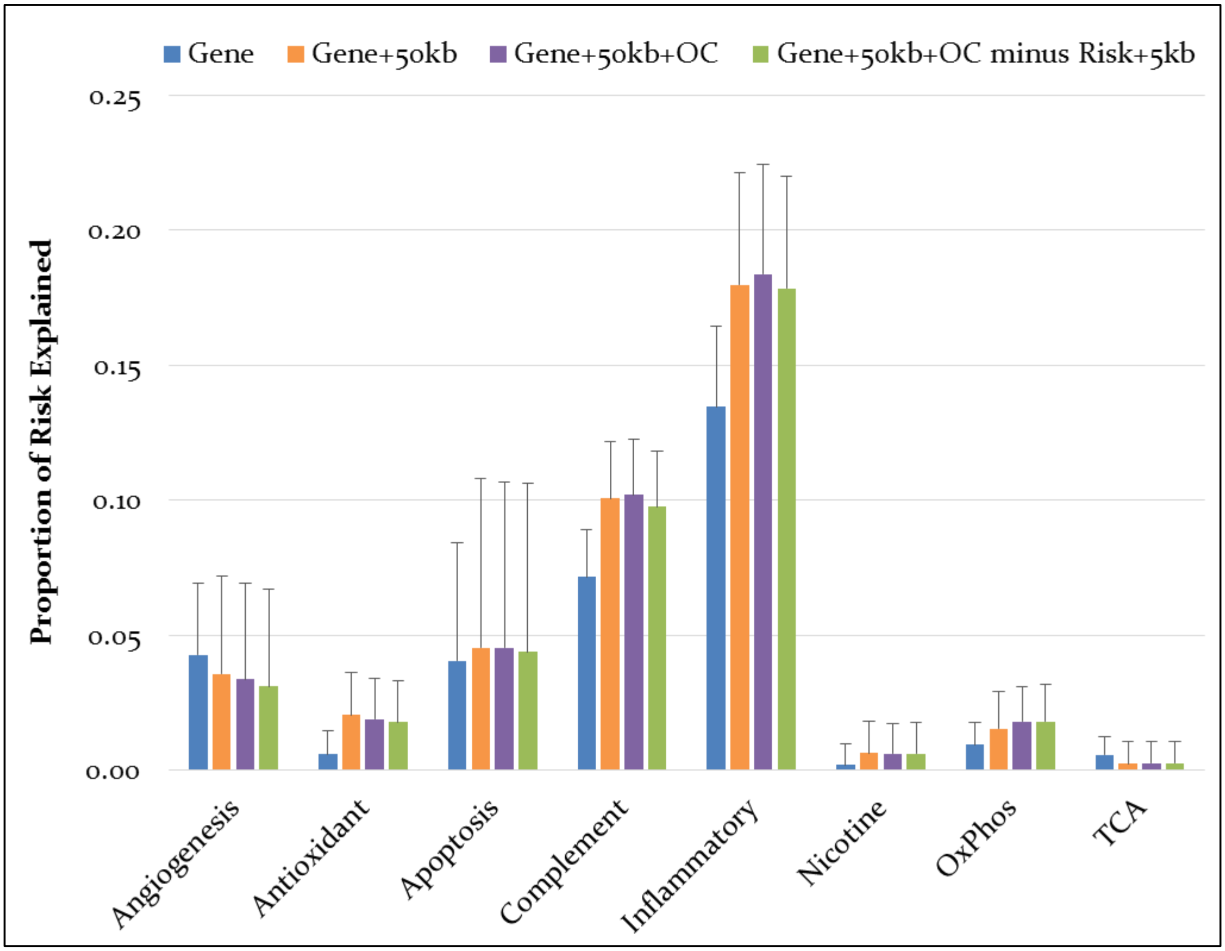


Figure 10. Risk explained by each pathway, by partitioning strategy.
Each bar represents the proportion of risk explained from a fitted mixed linear model using SNPs selected for each pathway for four different partitioning strategies. Error bars represent standard error (SE).

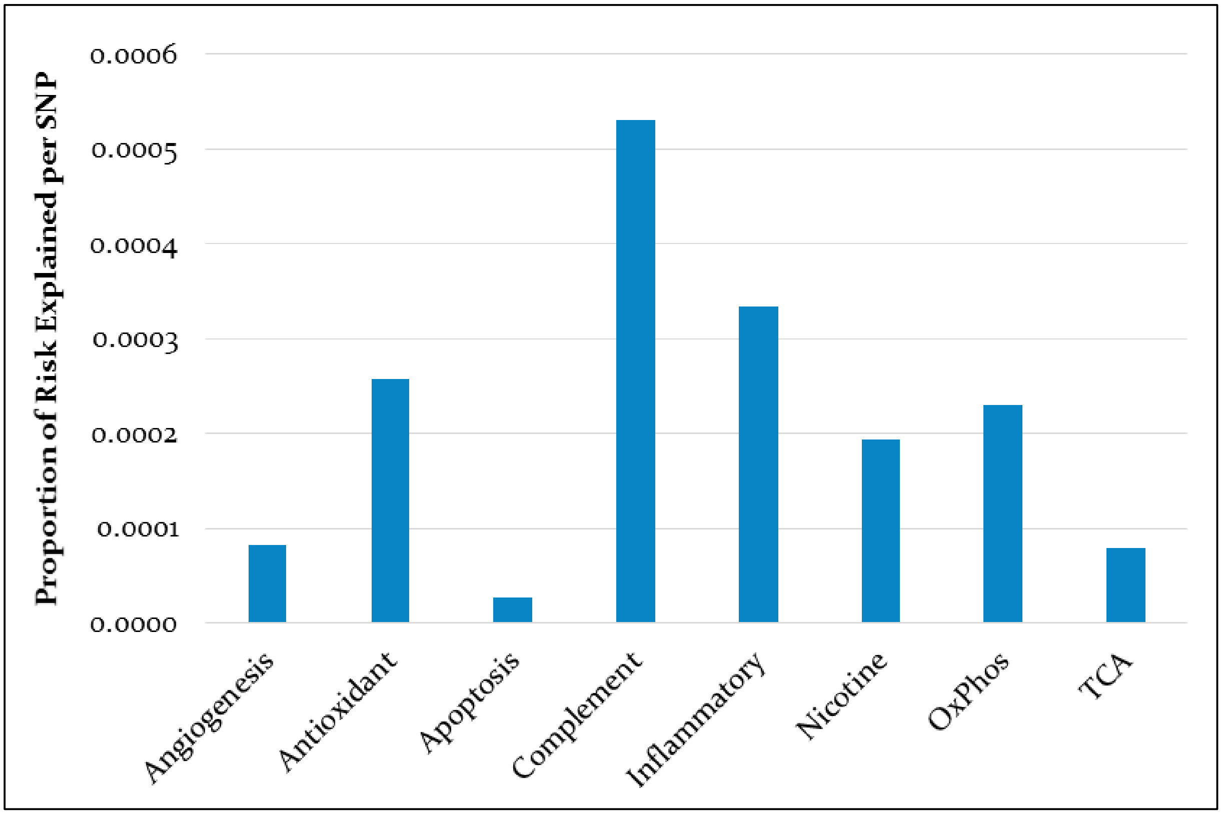
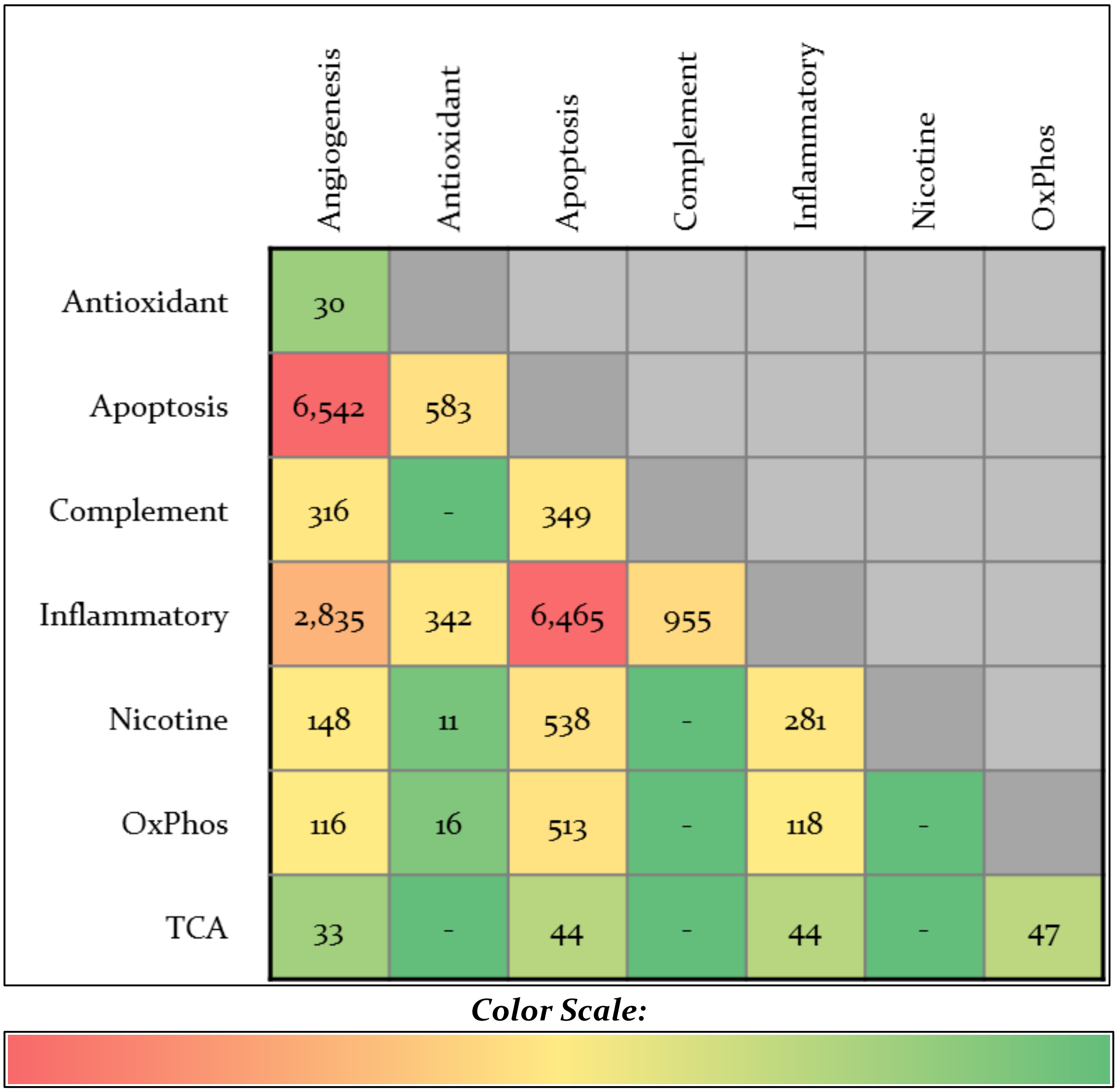


Figure 11. Average risk explained per SNP by pathway.
Each bar represents the proportion of risk explained divided by the number of SNPs per pathway. In this analysis, risk SNPs plus 5 kb regions were excluded.

<u>Gene overlap</u>

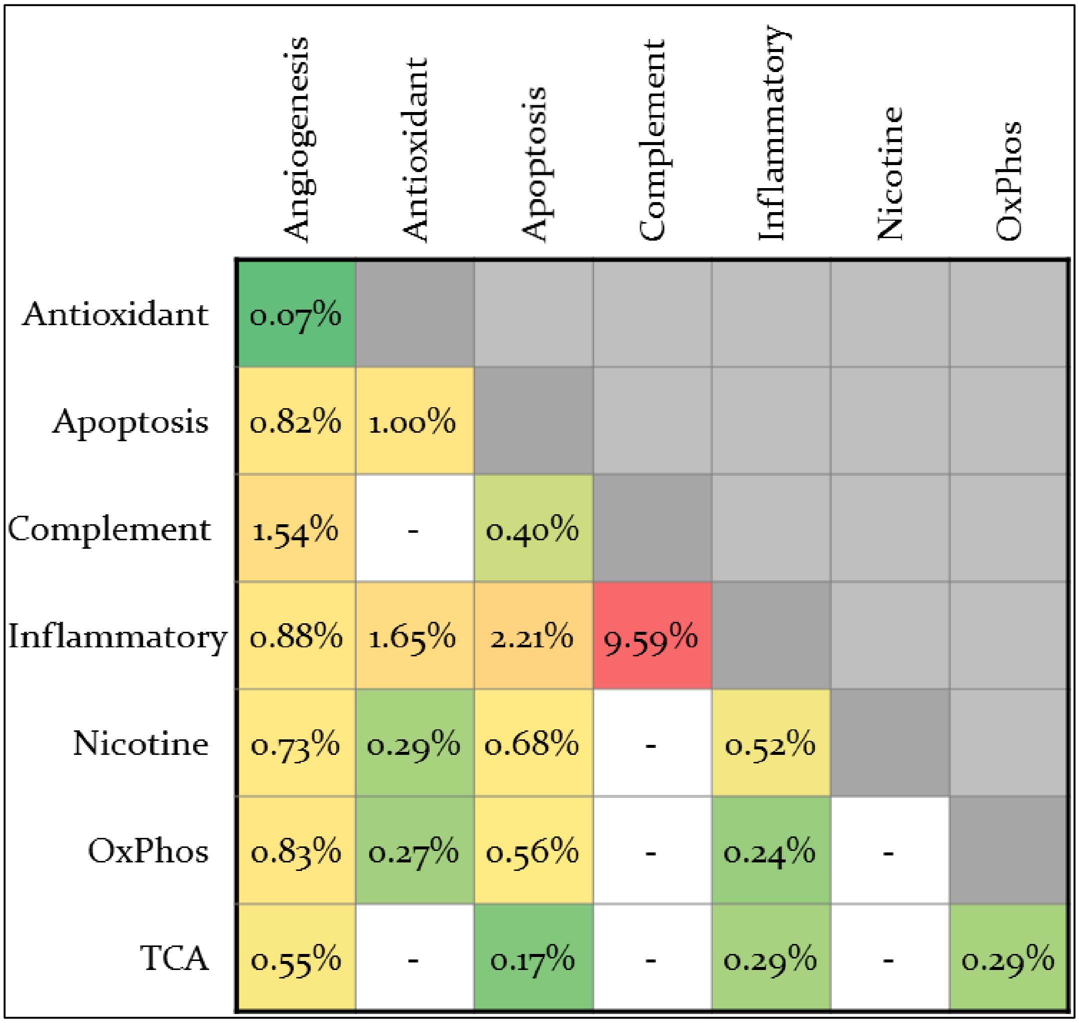
The pathways we selected to study for association to AMD risk were not all completely unrelated. For example, inflammation, apoptotic signaling, and angiogenesis are all biologically related and also have SNP overlap between pathways (Figure 12). We estimated the proportion of risk explained due to SNPs overlapping between pathways for each pathway pair where overlap was present and found that the overlap between most pathway pairs accounted for between 0.07% and 2.21% of the risk for AMD explained (Figure 13). The SNPs overlapping between the complement and inflammatory pathways, however, explained 9.59% of the risk for AMD; taking a closer look at SNPs shared provides a better understanding of the risk explained by the two pathways (Figure 15). Of the 1,343 SNPs in the complement pathway, 955 were also in the inflammatory pathway. The 15,038 SNPs unique to the inflammatory pathway, however, only explained 2.9% of the risk for AMD — a non-statistically significant amount. From this we observe that while the inflammatory pathway, at first glance, appears to explain more risk than the complement pathway, in reality, a large amount of the risk, but not all, is due to genes shared between the complement activation pathway.



	Angiogenesis	Antioxidant	Apoptosis	Complement	Inflammatory	Nicotine	OxPhos
Antioxidant	30						
Apoptosis	6,542	583					
Complement	316	-	349				
Inflammatory	2,835	342	6,465	955			
Nicotine	148	11	538	-	281		
OxPhos	116	16	513	-	118	-	
TCA	33	-	44	-	44	-	47

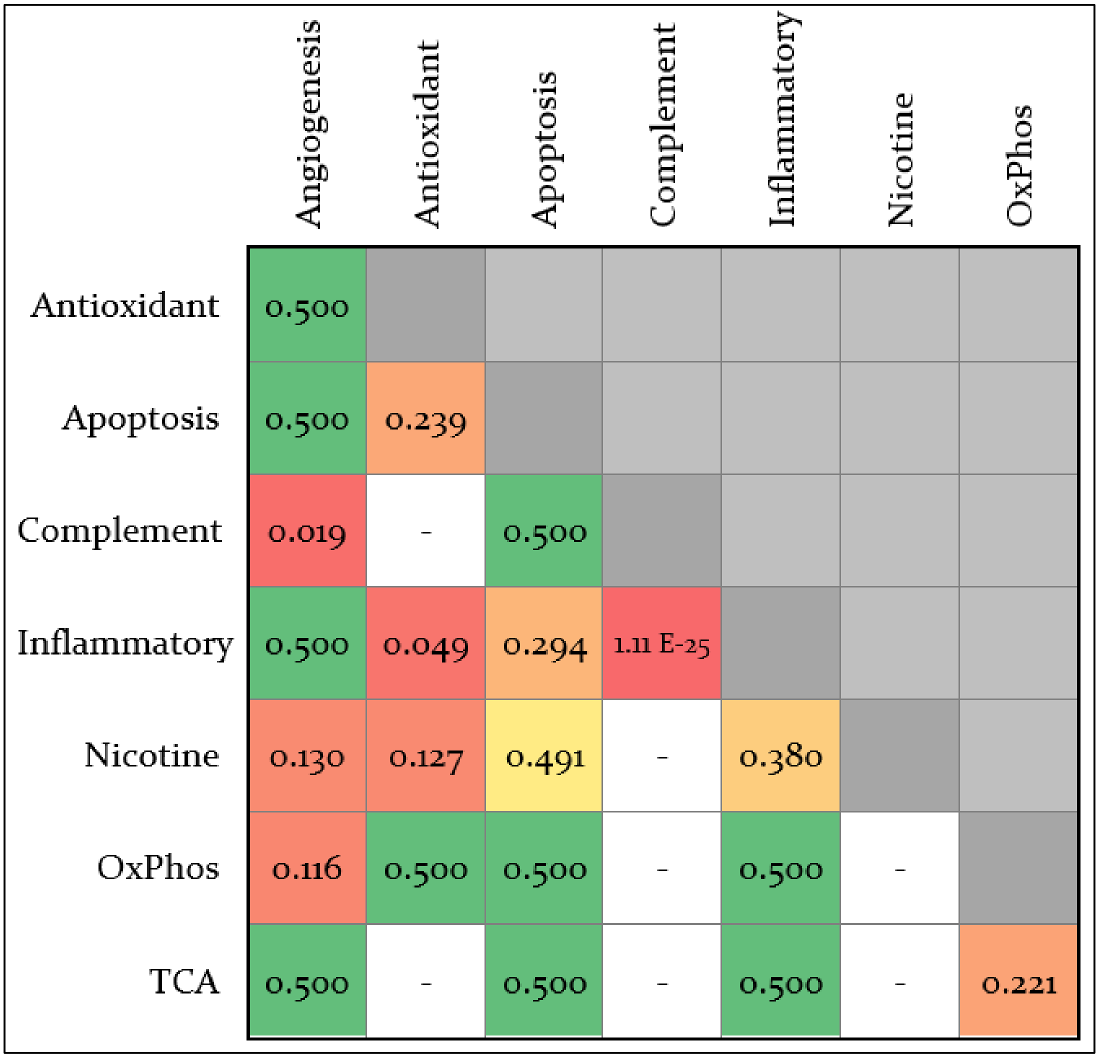
Color Scale:

Figure 12. Number of overlapping SNPs between pathway pairs.
Pathway pairs with little or no overlapping SNPs shown as green, fading to red for pathway pairs with the most overlapping SNPs. The same color scale will be used throughout the rest of this work, with red typically representing larger counts or more significant values.



	Angiogenesis	Antioxidant	Apoptosis	Complement	Inflammatory	Nicotine	OxPhos
Antioxidant	0.07%						
Apoptosis	0.82%	1.00%					
Complement	1.54%	-	0.40%				
Inflammatory	0.88%	1.65%	2.21%	9.59%			
Nicotine	0.73%	0.29%	0.68%	-	0.52%		
OxPhos	0.83%	0.27%	0.56%	-	0.24%	-	
TCA	0.55%	-	0.17%	-	0.29%	-	0.29%

Figure 13. Risk explained by overlapping SNPs between pathway pairs. Values represent the proportion of risk explained for SNPs contained in each pathway overlap. Pathway pairs with no overlapping SNPs shown as white boxes. Pathway pairs with less risk explained by overlap shown as green, fading to red for pathway pairs with more risk explained by overlap. Overlap was calculated using gene plus 50 kb regions.



	Angiogenesis	Antioxidant	Apoptosis	Complement	Inflammatory	Nicotine	OxPhos
Antioxidant	0.500						
Apoptosis	0.500	0.239					
Complement	0.019	-	0.500				
Inflammatory	0.500	0.049	0.294	1.11 E-25			
Nicotine	0.130	0.127	0.491	-	0.380		
OxPhos	0.116	0.500	0.500	-	0.500	-	
TCA	0.500	-	0.500	-	0.500	-	0.221

Figure 14. P-value for overlapping SNPs between pathway pairs.
Pathway pairs with overlapping SNPs contributing to smaller p-values shown as red, fading to green for pathway pairs with overlapping SNPs resulting large, non-significant p-values. Three pathway pairs shown have significant (< 0.05) p-values.

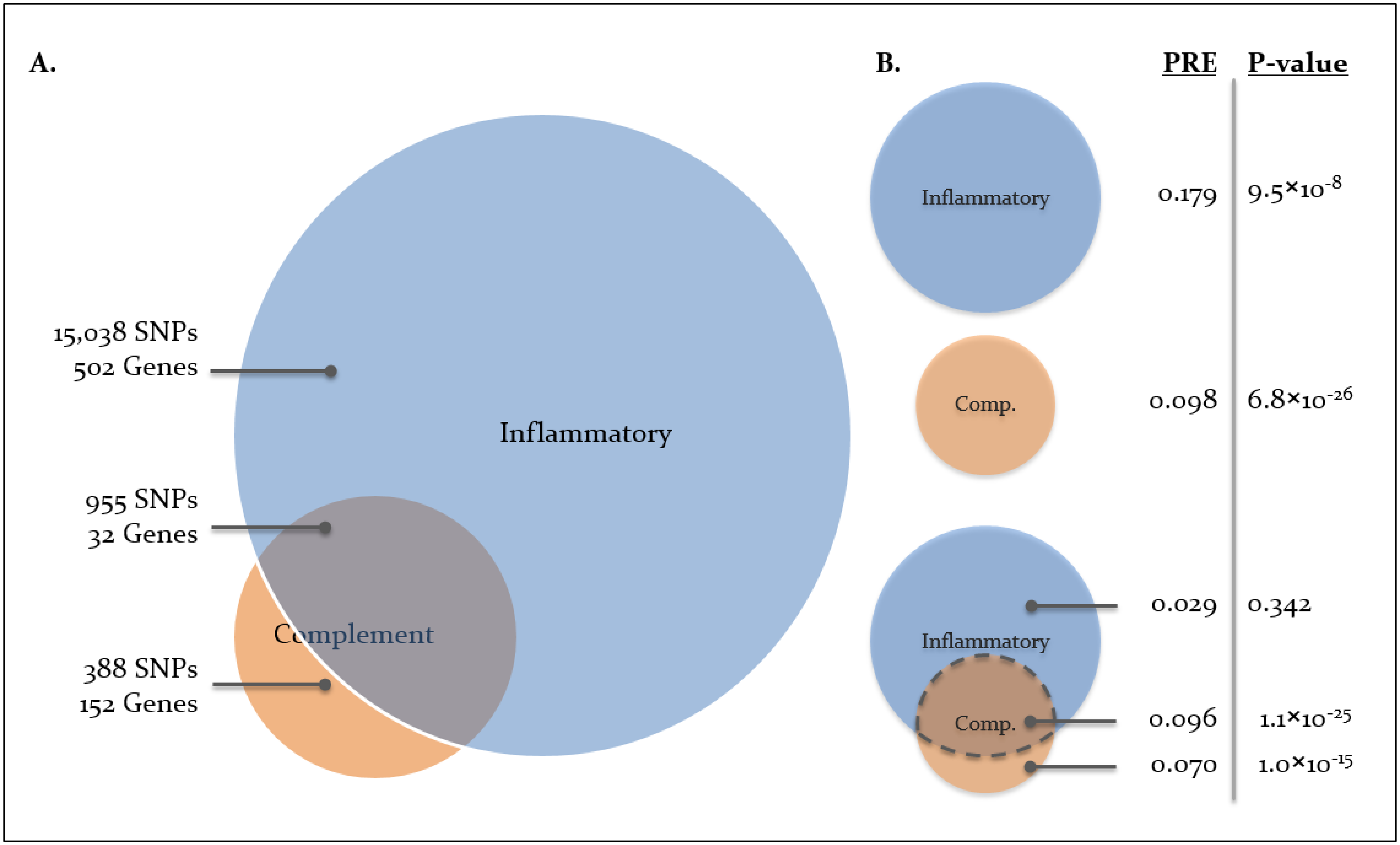


Figure 15. Overlap between complement and inflammatory pathways.
(**A**) Venn diagram of SNP and gene overlap between the complement and inflammatory pathways. (**B**) P-values and the proportion of risk explained (PRE) by complement and inflammatory pathways, separately and for overlapping regions. Overlapping SNPs were determined using regions including genic SNPs plus 50 kb flanking regions.

Linkage disequilibrium near known risk SNPs

To ensure that SNPs near the 19 known risk SNPs (Table 1) were not overinflating estimates of risk explained, we used LD information around the risk SNPs to exclude SNPs in LD and measure any changes in overall, genome-wide estimates of AMD risk explained (Table 6). As mentioned previously, the 19 risk SNPs alone explained 13.3% of risk for AMD while all other SNPs (included in a remainder GRM) explained 36.7% of the risk for AMD. Exclusion of SNPs using the threshold of $R^2 \geq 0.01$ only reduced this latter risk explained by 1.6%, to 35.1%. In an even more conservative case, we excluded 1MB flanking each side of the risk SNPs, regardless of LD, resulting in a reduction in risk explained of 5.4%, to 31.3% — unsurprising given the number of total SNPs excluded. Based on this we can assume that LD between risk SNPs and pathways SNPs would not confound estimates of AMD risk explained.

Table 6. Risk explained excluding known risk SNPs and SNPs in LD.

Exclusion Criteria	Number of SNPs Excluded	PVE (%)	SE (%)	p-val.
None	0	36.72	16.13	0.0042
$R^2 \geq 10\%$	1,183	35.25	14.75	0.0063
$R^2 \geq 5\%$	1,684	35.19	14.75	0.0064
$R^2 \geq 1\%$	1,925	35.12	14.75	0.0064
1MB flanking	9,938	31.33	14.70	0.0145

19 known risk SNPs: PVE ~13.3%, SE ~3.92%, p-val. < 1.35×10^{-61} for all.
Remainder SNPs with none excluded: 659,162.
PVE, SE, and p-values shown for non-known risk SNPs, excluding specified SNPs.
R^2: 0 = linkage equilibrium; 1 = perfect linkage disequilibrium.

Effect of smoking status

Smoking is a major risk factor for AMD; therefore, we assessed the impact of smoking status as a covariate in a sub-analysis of these data in samples where smoking status was available. After adjusting for smoking, the proportion of risk explained by each pathway did not change considerably (Figure 16). In fact, after adjustment, the angiogenesis, complement, and inflammatory pathways actually explained slightly more risk for AMD. All pathways exhibited little change and we conclude that adjusting for smoking status does not modulate the cumulative effect of SNPs within any of the targeted pathways.

Stratification by AMD subtype

Finally, we compared the effects on AMD risk, stratified by AMD subtype, for all pathways (Figure 17). There were 668 controls, 1,145 total AMD cases, 113 cases with GA (grade 4; advanced dry AMD), 667 cases with CNV (grade 5 in at least one eye; wet AMD), and 365 cases with grade 3 or an undocumented grade. We hypothesized that, because of the strong biological correlation between wet AMD and the angiogenesis pathway, a significant proportion of risk explained by the angiogenesis pathway would be observed when comparing CNV cases to controls. However, that was not the case; we observe slightly more risk explained by the angiogenesis pathway (and most other pathways) when comparing cases with GA to controls. We observe an unusual peak of risk explained by the apoptosis pathway when comparing GA versus controls, which is intriguing given possible associations with GA and cell death in literature [66, 83]. However, this signal may be an artifact of the more limited power within our GA subset.

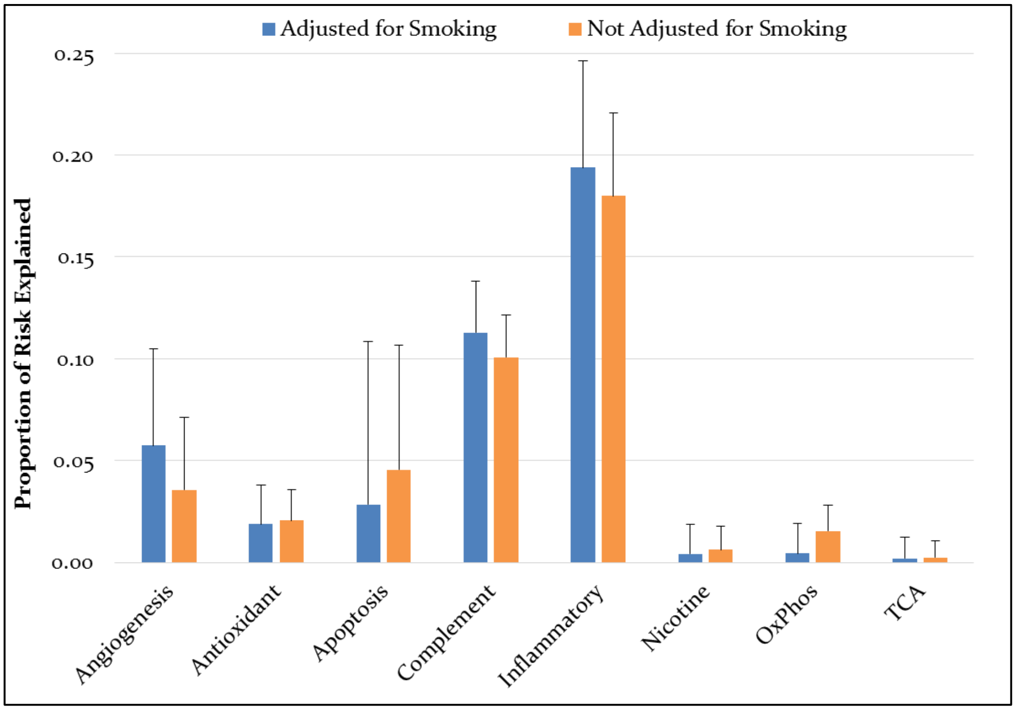


Figure 16. Effect of smoking adjustment on pathway risk explained.
Gene plus 5 kb flanking, minus risk plus 5 kb flanking regions excluded.

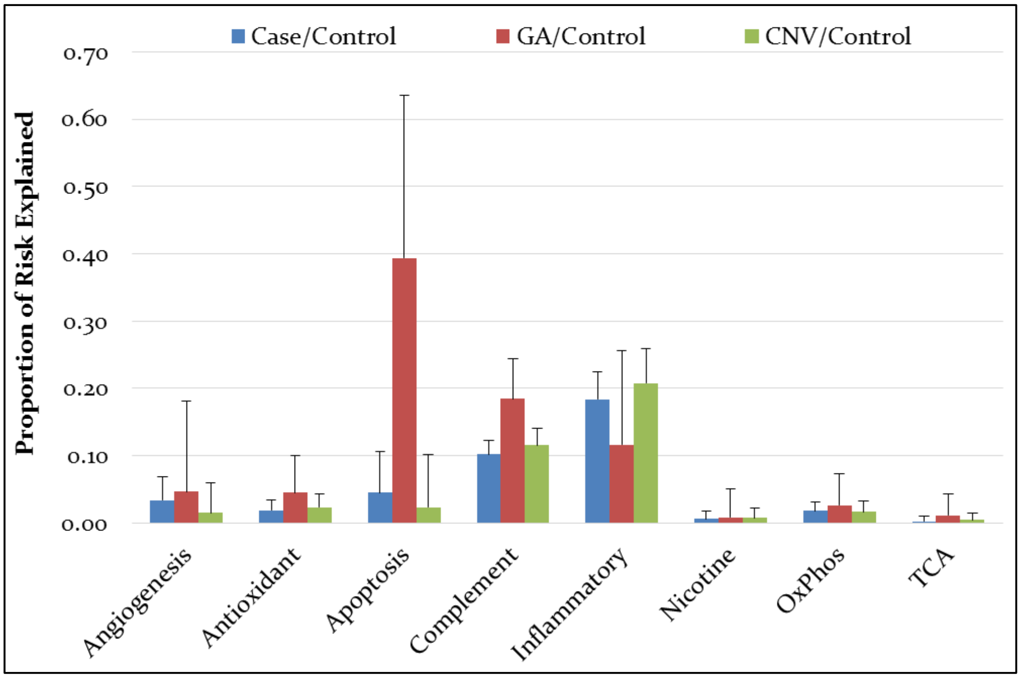


Figure 17. Risk explained per pathway by AMD subtype.
Genic SNPs plus 50 kb plus open chromatin SNPs included in analyses. Standard error is large for geographic atrophy (GA) cases versus controls because of the low sample size of cases with GA in our dataset.

Table 7. SNP counts, proportion of risk explained, and p-values for partitioned regions.

	GO Term Pathway	G	G+50	G+50+OC	G+50+OC minus R+5
Number of SNP	Angiogenesis	8,853	16,907	17,465	17,455
	Antioxidant Activity	616	2,014	2,115	2,114
	Apoptotic Signaling	26,682	54,417	56,136	56,130
	Complement Activation	478	1,343	1,374	1,370
	Inflammatory Response	5,883	15,993	16,635	16,630
	Response to Nicotine	426	1,101	1,150	1,150
	Oxidative Phosphorylation	350	1,451	1,548	1,548
	Tricarboxylic Acid Cycle	342	981	1,019	1,019
Proportion of Risk Explained	Angiogenesis	0.0427	0.0356	0.0337	0.0309
	Antioxidant Activity	0.0059	0.0205	0.0189	0.0178
	Apoptotic Signaling	0.0405	0.0454	0.0454	0.0438
	Complement Activation	0.0715	0.1006	0.1020	0.0976
	Inflammatory Response	0.1346	0.1799	0.1839	0.1786
	Response to Nicotine	0.0020	0.0063	0.0060	0.0060
	Oxidative Phosphorylation	0.0093	0.0154	0.0179	0.0179
	Tricarboxylic Acid Cycle	0.0057	0.0024	0.0026	0.0026
P-value	Angiogenesis	0.0797	0.3017	0.3592	0.4410
	Antioxidant Activity	0.5000	0.1081	0.1439	0.1631
	Apoptotic Signaling	0.2828	0.5000	0.5000	0.5000
	Complement Activation	1.2E-28	7.1E-27	6.3E-27	6.8E-26
	Inflammatory Response	6.3E-11	3.1E-08	4.4E-08	9.5E-08
	Response to Nicotine	0.5000	0.5000	0.5000	0.5000
	Oxidative Phosphorylation	0.2232	0.1103	0.0802	0.0804
	Tricarboxylic Acid Cycle	0.4602	0.5000	0.5000	0.5000

G=Gene; 50=50 kb flanking gene; OC=SNPs in open chromatin 50 kb to 250 kb flanking; R+5=Risk SNPs plus 5 kb flanking.

CONCLUSIONS AND FUTURE DIRECTIONS

In our analyses, we both confirm existing knowledge of AMD genetics and provide new, additional information on putative disease-associated pathways influencing risk for AMD. Our results show that SNPs in genes (and within 50 kb flanking) associated with complement activation and inflammation significantly contribute to AMD risk, separately from the risk explained by 19 known risk SNPs. We note, however, that the complement and inflammatory pathways are not discrete; we found that a large proportion of risk explained by the inflammatory and complement activation pathways are due to overlap of genes between the two. Other mechanisms thought to be involved in AMD pathogenesis do not appear to greatly influence disease risk through the cumulative action of common genetic variants. We also observe that while smoking is a known risk factor for AMD, inclusion as a model covariate does not significantly affect risk estimates from pathways. Overall, we show genes that interplay between the complement and inflammatory pathways explain additional risk, apart from the known, large-effect AMD risk SNPs, and that some portion of these are localized to the 50 kb flanking regions, indicating a regulatory role. As such, further targeted genomic or molecular studies should consider additional loci within the complement pathway in addition to the established risk SNPs.

In this study we found additional risk from the complement pathway not explained by known risk SNPs. Future studies could further partition complement genes into multiple components to allow for more specific localization of the effects contributing to AMD risk. Additionally, future analyses could expand the methods presented here to include additional pathways using either more exhaustive pathway catalogs or custom curated pathways of interest. Lastly, in this study we only assessed common SNPs (minor allele frequency greater than 5%); however, future studies could assess contributions to cumulative AMD risk, stratified by allele frequency, to estimate the importance of rare variation on risk for AMD.

This chapter, and most studies to date, only test for additive genetic effects on AMD. A better understanding of non-additive effects, if any, on AMD is needed. In the following chapters I describe methods we developed to estimate cumulative dominance and epistatic effects on traits and then I apply those methods to specific interactions (gene-gene and gene-pathway) to further explore the genetic architecture of AMD.

Chapter 4 — Detecting and Simulating Cumulative Epistatic Effects

Chapter background and introduction

Epistasis

Thus far we have primarily discussed genetic analyses involving methods that test for additive genetic effects, yet epistasis — the interaction between two or more loci resulting in a different phenotype than if the loci acted independently — is a known genetic phenomenon that is studied much less frequently than additive genetics. Statistically, epistasis refers to the deviation from additivity that results from the effects of alleles at different loci [84]. Epistasis is believed to play a role in many complex human diseases but the degree of impact is unknown [85–87]. Much of the genome was once referred to as junk DNA but evidence suggests that a large portion of what was once thought to serve little purpose is actually regulatory, some of which may be mediated through interactions [88]. Most studies focus on additive genetic effects and are not able to explain all of a trait's heritability. Heritability estimates may often include a small portion of effects from non-additive SNP interactions [89]. Methods used to detect SNP-SNP interactions include regression models, exhaustive two-locus interaction searches, recursive partitioning, multifactor dimensionality reduction (MDR), ReliefF, Tuned ReliefF, evaporative cooling, and Bayesian epistasis association mapping [87]. Most of the methods test for interactions between specific SNPs, often two-locus interactions requiring at least one locus to have a main effect.

Genome-wide complex trait analysis (GCTA) [43], discussed in depth in Chapter 2, can be used to estimate the cumulative, genome-wide, additive effect that genetic variance has on phenotypic variance, but no similar method currently exists for assessing the cumulative impact of SNP-SNP interactions on a trait. As our broader understanding of genetics increases, epistasis may prove to be a key component in more accurately predicting disease risk. Knowledge of the degree to which epistasis plays a role for a given trait would help direct resources and guide future studies.

Genetic encoding for additive, dominant, and epistatic effects

In 1954 C. Clark Cockerham published an article [90] in *Genetics* titled "An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present". It extended the concepts of additive and dominant genetic variance, introduced by Fisher [84], to epistatic (genetic interaction) genetic variance. Some of the statistical derivations in his paper account for inbreeding by including adjustments involving Wright's inbreeding coefficient F [91] — particularly necessary when considering livestock — but for the methods described here we assume a randomly mating population and do not account for inbreeding.

Similar to how genetic relationship values are calculated by GCTA using variance-covariance matrices, as described in Chapter 2, Cockerham describes orthogonal variance components that can be derived using genotype-specific variance and covariance calculations. In the tables below we describe these components and notations. Later in the methods we describe in more detail how these calculations are used to simulate specific genetic effects. The calculations and concepts described throughout this chapter depend on dividing SNPs into two groups for the sake of assessing genetic interactions (epistasis). Each group will typically be called group A and group B and can contain a varying number of SNPs. The total number of SNP pairs being assessed can be calculated by multiplying the number of SNPs in group A by the number of SNPs in group B. For notation purposes, when describing interactions between the two groups of SNPs, the SNP in group A being tested for potential interaction with a SNP in group B will be called SNP A (or SNP 1) and SNP B (or SNP 2), respectively. Table 8 shows the notation used for alleles and allele frequencies for each allele of two SNPs. Table 9 shows the notations for genotype and SNP pair frequencies, as described in the table legend. Table 10 and Table 11 list the variance and covariance calculations for each of the eight orthogonal variance components — we often refer to these as “Cockerham calculations”. The first four components are for additive and dominant effects from each group of SNPs. The last four components are for epistatic effects, derived as shown in Table 12. For the two components for group A, W_1 and W_2 (Table 11), the covariance values depend only on SNP 1 genotypes, so that for W_1 we see that 2v, v-u, and -2u correspond to AA, Aa, and aa, respectively, regardless of the genotype for SNP 2. The two components for group B, W_3 and W_4, exhibit similar patterns corresponding to the genotype for SNP 2, regardless of the genotype for SNP 1.

Table 8. SNP allele frequency notations.

Allele	Meaning	Frequency Notation
A	SNP 1 (Group A) Major Allele	u
a	SNP 1 (Group A) Minor Allele	v
B	SNP 2 (Group B) Major Allele	x
b	SNP 2 (Group B) Minor Allele	y

Table 9. Genotype and marginal frequency notations.

	BB	**Bb**	**bb**	
AA	f_{22}	f_{21}	f_{20}	f_{2*}
Aa	f_{12}	f_{11}	f_{10}	f_{1*}
aa	f_{02}	f_{01}	f_{00}	f_{0*}
	f_{*2}	f_{*1}	f_{*0}	f_{**}

Adapted from [90].

The f values (Table 9) represent SNP pair (in the box) and marginal (outside the box) frequencies. The first subscripted number represents the number of major alleles of SNP 1 and the second subscripted number represents the number of major alleles of SNP 2. Stars (*) are used to represent the marginal genotype frequency, where the number shown corresponds to the frequency of a particular genotype (e.g. f_{21} is the frequency of AABb and f_{*0} is the frequency of bb). f_{**} should equal 1.

Table 10. Cockerham variance calculations.

W^*	Component	Equation
W_1	Additive A	$2uv$
W_2	Dominant A	$1/(uv)^2$
W_3	Additive B	$2xy$
W_4	Dominant B	$1/(xy)^2$
W_5	A × A Interaction	$4uvxy$
W_6	A × D Interaction	$2uv/(xy)^2$
W_7	D × A Interaction	$2xy/(uv)^2$
W_8	D × D Interaction	$1/(uvxy)^2$

Adapted from [90]. *Cockerham notation. For W_5 - W_8, A = additive and D = dominant.

Table 11. Cockerham covariance calculations.

W^*	Component	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
W_1	Additive A	$2v$	$2v$	$2v$	$v-u$	$v-u$	$v-u$	$-2u$	$-2u$	$-2u$
W_2	Dominant A	$1/f_{2*}$	$1/f_{2*}$	$1/f_{2*}$	$-2/f_{1*}$	$-2/f_{1*}$	$-2/f_{1*}$	$1/f_{0*}$	$1/f_{0*}$	$1/f_{0*}$
W_3	Additive B	$2y$	$y-x$	$-2x$	$2y$	$y-x$	$-2x$	$2y$	$y-x$	$-2x$
W_4	Dominant B	$1/f_{*2}$	$-2/f_{*1}$	$1/f_{*0}$	$1/f_{*2}$	$-2/f_{*1}$	$1/f_{*0}$	$1/f_{*2}$	$-2/f_{*1}$	$1/f_{*0}$
W_5	A × A Interaction	$4vy$	$2v(y-x)$	$-4vx$	$2y(v-u)$	$(v-u)(y-x)$	$-2x(v-u)$	$-4uy$	$-2u(y-x)$	$-4ux$
W_6	A × D Interaction	$2v/f_{*2}$	$-4v/f_{*1}$	$2v/f_{*0}$	$(v-u)/f_{*2}$	$-2(v-u)/f_{*1}$	$(v-u)/f_{*0}$	$2u/f_{*2}$	$-4u/f_{*1}$	$2u/f_{*0}$
W_7	D × A Interaction	$2y/f_{*2}$	$(y-x)/f_{*2}$	$2x/f_{*2}$	$-4y/f_{*1}$	$-2(y-x)/f_{*1}$	$4x/f_{*1}$	$2y/f_{*0}$	$(y-x)/f_{*0}$	$-2x/f_{*0}$
W_8	D × D Interaction	$1/f_{22}$	$-2/f_{21}$	$1/f_{20}$	$-2/f_{12}$	$4/f_{11}$	$-2/f_{10}$	$1/f_{02}$	$-2/f_{01}$	$1/f_{00}$

Each column represents one of nine possible genotype pair combinations for two bi-allelic SNPs. Unlike variance calculations, covariance calculations are dependent on an individual's genotype combination at a particular SNP pair. Adapted from [90]. *Cockerham notation. Interaction component notations described in Table 12.

Table 12. Variance and covariance interaction derivations.

W*	Shorthand	Component	Derivation
W_5	A×A	Additive × Additive	$W_1 \times W_3$
W_6	A×D	Additive × Dominant	$W_1 \times W_4$
W_7	D×A	Dominant × Additive	$W_2 \times W_3$
W_8	D×D	Dominant × Dominant	$W_2 \times W_4$

*Cockerham notation.

Methods overview

This chapter consists of two related but distinct methods (Figure 18). The first method, which we refer to as **iGRM**, uses statistics derived by C. Clark Cockerham [90] to estimate the contribution of additive, dominant, and epistatic genetic effects from genomic regions. This is done by generating orthogonal genetic relationship matrices (GRMs); GCTA can currently only compute for additive genetic effects. The GRMs are then fit in a mixed linear model (MLM) using restricted maximum likelihood (REML) fitting procedures to estimate the heritability or proportion of phenotypic variance explained by each genetic component.

The second method, which we refer to as **iSim**, involves simulating datasets with cumulative additive, dominant, and epistatic effects, using the same variance components calculations, to partition genetic effects into orthogonal components. iSim is adapted using components from genome-wide complex trait analysis (GCTA), which currently only allows datasets to be simulated with user-specified levels of heritability due to additive genetic effects [43]. Simulations generated using iSim are also used to validate iGRM using varying levels of heritability and SNP minor allele frequencies.

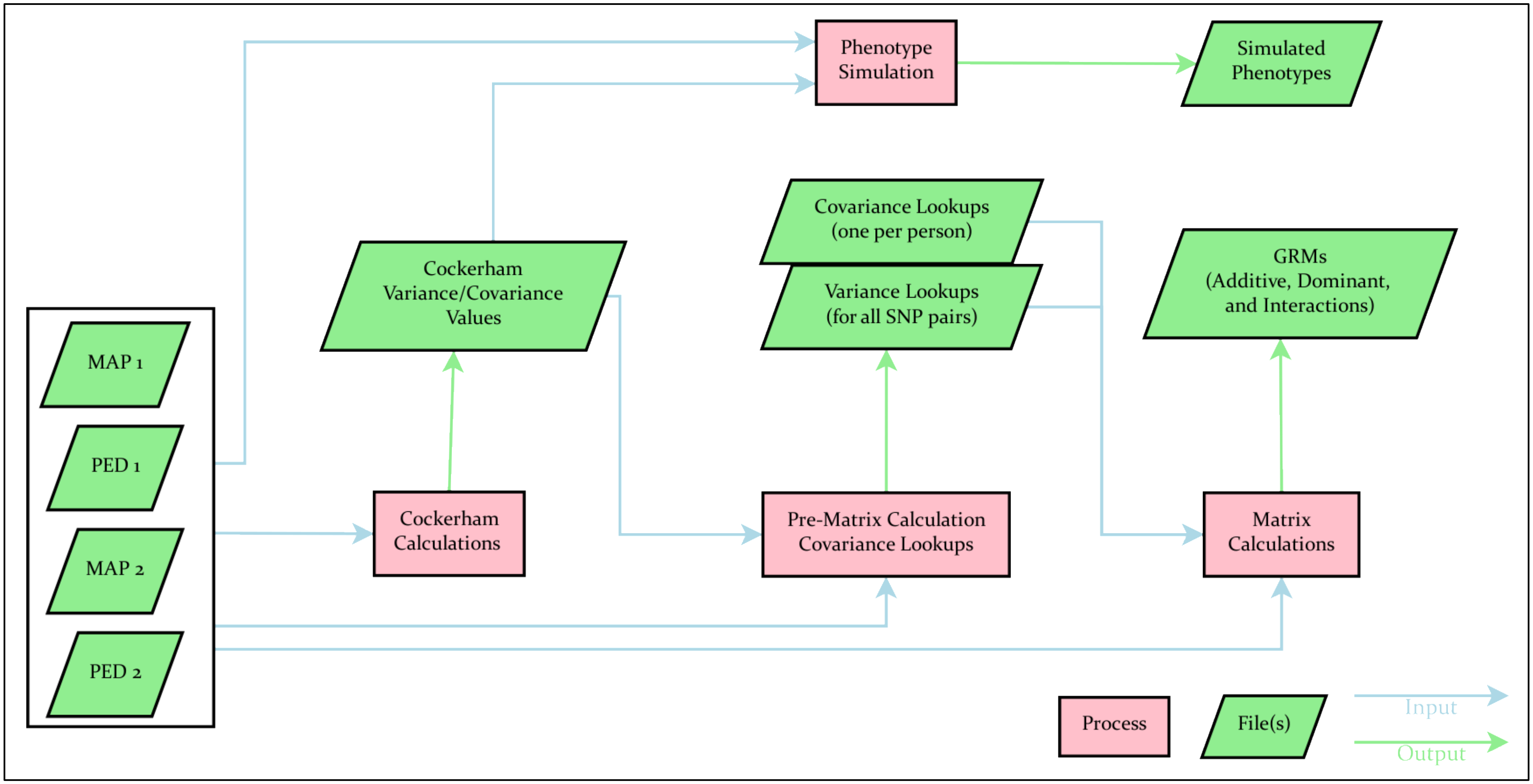

Phenotype Simulation
Simulated Phenotypes
Covariance Lookups (one per person)
Variance Lookups (for all SNP pairs)
GRMs (Additive, Dominant, and Interactions)
Cockerham Variance/Covariance Values
MAP 1
PED 1
MAP 2
PED 2
Cockerham Calculations
Pre-Matrix Calculation Covariance Lookups
Matrix Calculations
Process
File(s)
Input
Output

Figure 18. Software overview.
iGRM involves Cockerham calculations, pre-matrix calculations, and matrix calculations, along with various intermediate and final output files. **iSim** involves Cockerham calculations and simulations, along with intermediate files and final simulated phenotype files. Both methods require two sets of PLINK PED/MAP files as input.

iGRM — Estimating cumulative additive, dominant, and epistatic effects

Introduction

Most genetic studies focus mainly or solely on additive genetic effects of traits. Dominance effects, if present for a given trait, are often detected through additive-encoded statistical tests — another reason for the usual preference of only testing for additive effects. As was mentioned in the background to this chapter, methods are available to estimate cumulative, genome-wide effects from additive genetic effects, but no equivalent method exists for dominant or epistatic (interaction) effects. We present methods to simultaneously test for additive, dominant, and epistatic effects for a given trait to address this problem.

The statistics presented in the Cockerham article [90] (Tables 10 and 11), described in the background and introduction to this chapter, and the statistics used to calculate genetic relationship matrices (GRMs) [43], described in Chapter 2, follow the same ideology; GCTA notation and Cockerham notation are shown to be parallel through basic algebra, as is shown in Figure 19. The generalized form of the implementation of GCTA involves the covariance normalized by the variance for the additive (W_1) component.

The dominant and epistatic components are similarly derived in the Cockerham article; as such, we sought to modify the GRM equation to incorporate additive, dominant, and epistatic (interaction) components by substituting the additive variance and covariance calculations with the variances and covariances representing dominant and epistatic components. The existing method for calculating GRMs subsequently allows for the estimation of trait variance due to cumulative additive effects; we hypothesized that calculation of additional GRMs using the orthogonal variance components described by Cockerham would subsequently enable dominant and epistatic effects to be estimated by fitting those GRMs simultaneously in a mixed linear model (e.g. GCTA's restricted maximum likelihood estimation model fitting) to estimate the overall contribution of each component on trait variance or risk.

Cockerham | *Intermediate* | *GCTA*

Covariance

$$2v = 2(1-u) = 2-2u$$

$$v-u = (1-u)-u = 1-2u$$

$$-2u = -2u = 0-2u$$

$$(x-2u) = (x_{ij}-2p_i)$$

Variance

$$2uv = 2u(1-u) = 2p_i(1-p_i)$$

Figure 19. Cockerham-GCTA equivalency.
Letters u and v represent the major and minor allele (A and a, respectively) for a given SNP, where the sum of the two allele frequencies is equal to 1. Covariance and variance calculations are from the Additive A (W_1) portion of Cockerham statistics [90] and the GRM equation from the GCTA software publication [43].

Methods

Software design

After uncovering the parallel between Cockerham calculations and GCTA's GRM equation we formulated a plan to calculate and generate the new, additional GRMs, and ultimately implemented the software as a C++ package. First, SNP and SNP pair population variance and covariance values are calculated, as described in Tables 10 and 11. These calculations are derived from SNP allele frequencies and multi-locus genotype frequencies (Tables 8 and 9, respectively). As a note, it is important that multi-locus genotype frequencies be calculated directly from the population and not inferred using genotype frequencies, due to potential missingness issues. If a multi-locus genotype contains any missing alleles, that SNP pair is skipped and not used in frequency or GRM calculations. Also of note is that all SNPs should be bi-allelic. After frequency information is calculated that data is used to calculate GRMs based on modifying Equation 2 (Chapter 2) to incorporate additional variance and covariance components, instead of only additive components. Equations 4 and 5 show the modified GRM equations for single SNP effect GRM and interaction effect GRM calculations, respectively. The original GRM equations with additive-specific components are replaced with more generally applicable variance/covariance terms, where j and k refer to person 1 and 2, respectively. In Equation 4, N represents the number of SNPs in the genetic component being assessed. In Equation 5, "N" is replaced with "$N_1 \times N_2$" to reflect that calculations are for each SNP pair, rather than only for each SNP. For a given pair of individuals (j and k), their genetic relationship value (A_{jk}) is calculated as the weighted sum of each SNP or SNP pair effect, determined by the respective variance and covariance calculations for one of the eight genetic components.

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{(covar.j)(covar.k)}{(variance)} \tag{4}$$

$$A_{jk} = \frac{1}{N_1 \times N_2}\sum_{i=1}^{N_1 \times N_2} \frac{(covar.j)(covar.k)}{(variance)} \tag{5}$$

After a general pipeline was established for calculating additional GRMs, we developed software to load data, perform calculations, and generate GRMs as quickly and efficiently as possible. The overall program for iGRM was divided into three steps. All steps require at least three processors, with one processor assigned as the "manager", one processor assigned as the "accumulator", and the remaining processors assigned as the "workers". Essentially, the workers are the middle-men, receiving computational tasks from the manager then passing results to the accumulator. The accumulator then aggregates data from multiple workers and subsequently writes either intermediate or final data to files.

(1) The first step performs initial frequency calculations (Figure 20) derived from Cockerham statistics previously described. These statistics are population-based and are calculated for every SNP pair. **(2)** The second step performs pre-matrix calculations (Figure 21). Covariance values (calculated and stored in step 1) are looked up for every person for every SNP pair and are dependent on a given person's multi-locus genotype for each SNP pair; one file containing matched covariance values is written per person. Additionally, variance calculations are performed, per SNP pair, which are then written to a separate file. **(3)** The third step uses the pre-matrix values to calculate the genetic relationship value for every pair of individuals, using the corresponding GRM equation for each of the eight genetic components (Figure 22). The resulting 8 GRM files (one per genetic component) each contain a single genetic relationship value for every pair of individuals in the dataset.

Initially, variance and covariance values were performed “on the fly” for each genotype and genotype pair lookup. One improvement was to store pre-calculated variance and covariance values that could be matched, rather than matched and computed. Another improvement was in regard to how frequency calculations are stored. Initially, all frequency calculations were stored in a single file; however, when the size of the frequency file becomes greater than any single worker can load into memory, the program dies. Thus, the current version divides the frequency information in to separate files based on the number of workers being used, enabling workers the ability to load and perform calculations on smaller assigned groups of SNPs or SNP pairs. Before this change, each worker calculated a genetic relationship value (A_{jk}) for a single pair of individuals. Now, all workers simultaneously perform calculations for a pair of individuals until that genetic relationship value is complete.

Two more recent additions are functional but not optimized. First, a cache system was implemented to decrease the turnover rate of workers loading frequency files; the amount of memory available for each processor must be specified. Essentially, if a worker is calculating A_{jk} and the next pair is A_{jl}, there is no need to re-load the variance and covariance information for person j since it is already in memory. Currently, once the cache becomes full it is completely emptied and new data is loaded and re-used when needed. We have attempted making changes to enable a “smart cache” but initial tests showed no improvement. Second, a “resume mode” was implemented. In instances where the compute cluster kills the software after hours or days of runtime, resume mode can be used to recover partially-completed GRMs. Debugging is not complete, however; in some scenarios, such as partial line writes, resume mode is not able to recover data.

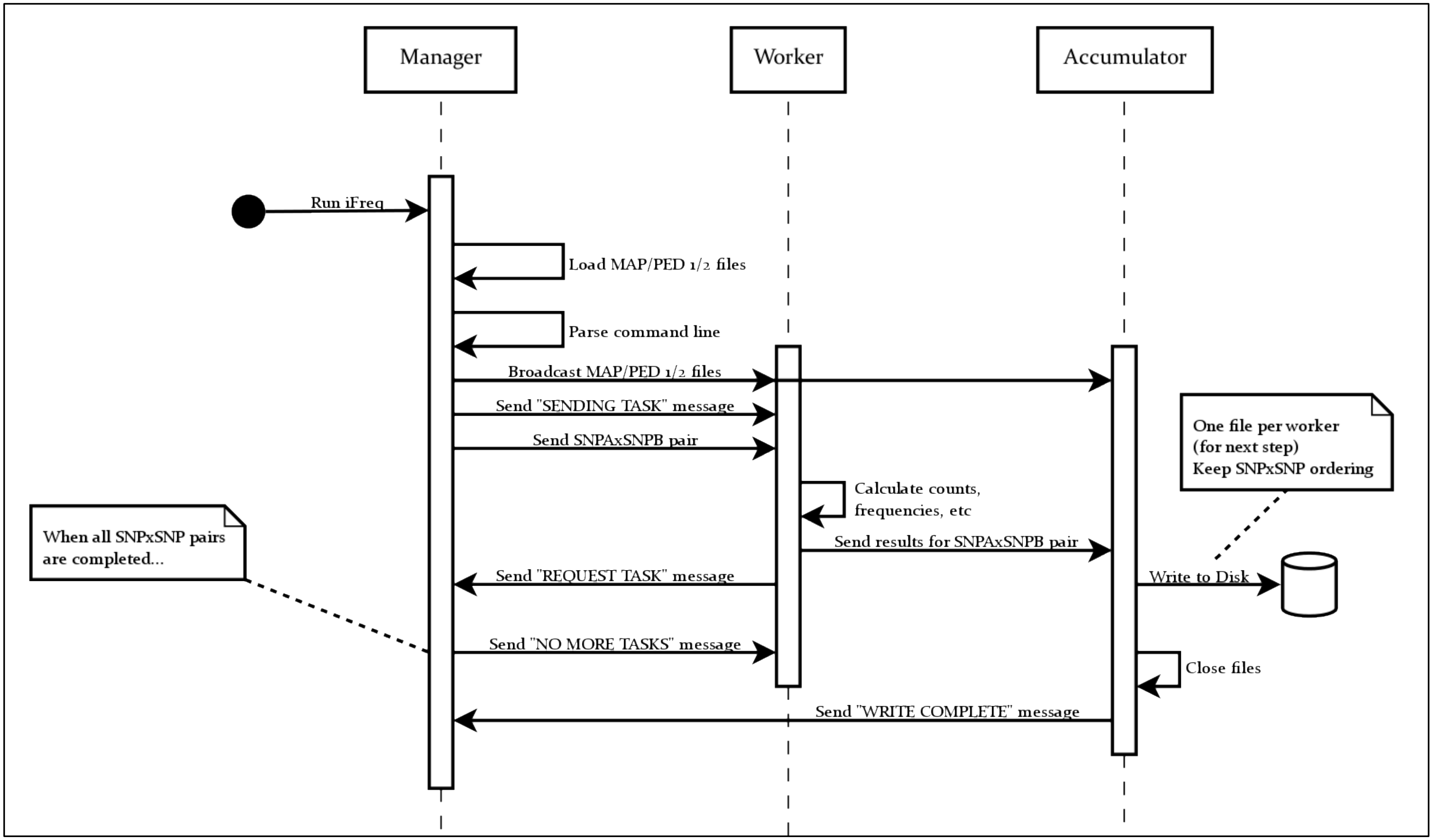


Figure 20. Frequency calculation processing flowchart.
All Cockerham values are calculated, including SNP frequencies, SNP pair (multi-locus genotype) frequencies, as well as variance and covariance values for each of the eight genetic components.

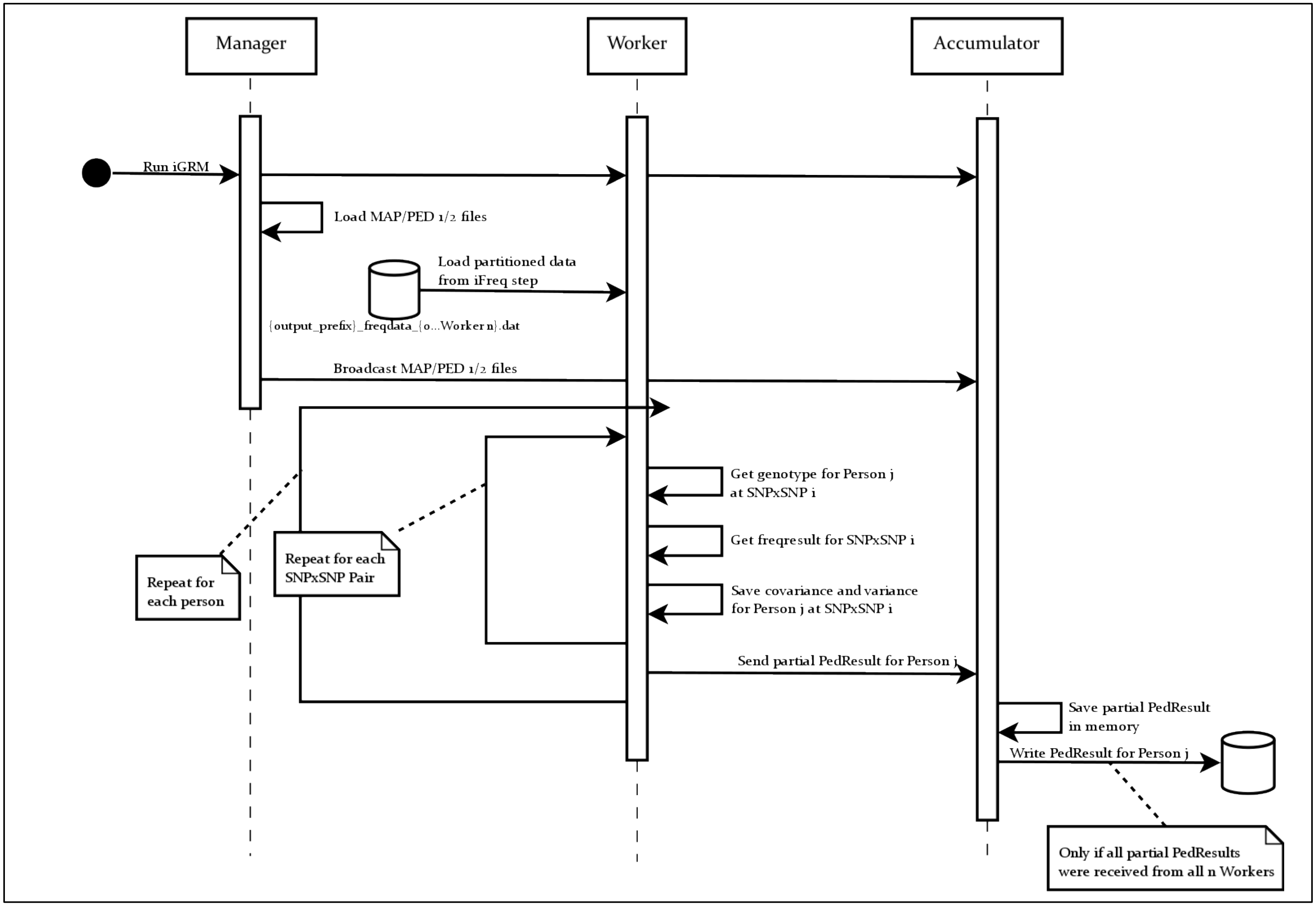

Manager
Worker
Accumulator
Run iGRM
Load MAP/PED 1/2 files
Load partitioned data
from iFreq step
{output_prefix}_freqdata_{0...Worker n}.dat
Broadcast MAP/PED 1/2 files
Get genotype for Person j
at SNPxSNP i
Get freqresult for SNPxSNP i
Save covariance and variance
for Person j at SNPxSNP i
Repeat for each
SNPxSNP Pair
Repeat for
each person
Send partial PedResult for Person j
Save partial PedResult
in memory
Write PedResult for Person j
Only if all partial PedResults
were received from all n Workers

Figure 21. Pre-matrix calculation processing flowchart.
For each individual, an intermediate file is written, containing covariance values corresponding with their genotypes. Additionally, a single file variance file is written (since variance values are based on population, not individual-level data).

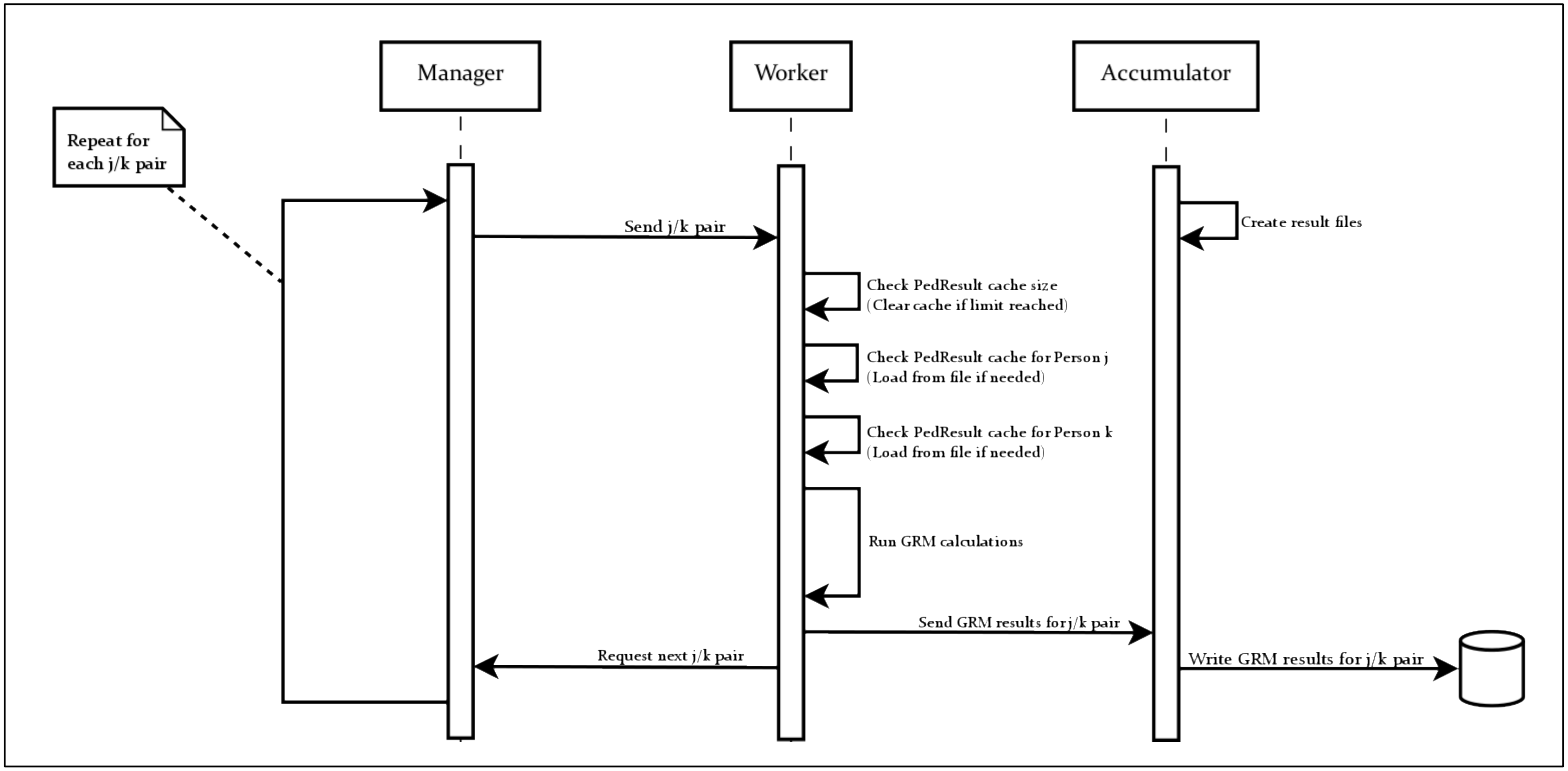

Manager
Worker
Accumulator
Repeat for each j/k pair
Send j/k pair
Create result files
Check PedResult cache size (Clear cache if limit reached)
Check PedResult cache for Person j (Load from file if needed)
Check PedResult cache for Person k (Load from file if needed)
Run GRM calculations
Send GRM results for j/k pair
Request next j/k pair
Write GRM results for j/k pair

Figure 22. Matrix calculation processing flowchart.
For each pair of individuals, pre-matrix values are loaded, then the genetic relationship value is calculated for each genetic component, using the weighted averages of variance-covariance calculates for respective genetic components.

Implementation

In previous sections we described the theory behind our additive, dominant, and interaction GRM calculation method and the details and steps involved in the software used to perform the calculations. An example of the runtime and changes in progress rate as the cache system fills is shown in Appendix H. Once the GRMs are created, minor processing is needed, including compressing the files using Gzip and creation of GRM ID files containing a list of all pairs of individuals, corresponding to the order in the GRM files. Finally, the orthogonal GRMs can be included in a mixed linear model REML analysis (described in Chapter 2), using GCTA, to estimate the contribution of each genetic component on trait variance or risk for a study phenotype.

Discussion and conclusions

In the previous sections of this chapter we describe the basis behind the need to develop a method to estimate additive, dominant, and interaction effects via calculation of GRMs. We then showed the concepts and math involved in calculating orthogonal GRMs representing each of the different genetic effects and described the details of how the method was implemented as software.

Describing the methods and implementation as software, alone, does not prove that results obtained using the method would be valid. In the next section we use simulations to validate iGRM and show that power to accurately detect effects from all eight genetic components is dependent on sample size, along with other factors.

To date, we have made substantial improvements to iGRM, compared to initial versions. Optimizations have been made along the way that now allow the software to be used for thousands of individuals at the gene or small pathway level. The current software is sufficient for validation using simulated data (see iSim in the next section) and application to study potential interactions in an AMD dataset with over 36,000 individuals and over 52,000 SNP pairs (see Chapter 5).

Future directions

Although some improvements to iGRM are logical and warranted, none are necessary for applying iGRM to the AMD interaction study presented in Chapter 5. Currently, a small amount of processing is needed to prepare the GRM files to be compatible with GCTA. Software modification to automatically compress (Gzip) GRM files and create GRM ID files would be useful for future updates. Additionally, an option to import binary PLINK files (either natively or through conversion) would be useful, as well as the option to output binary GRM files (the newer GCTA format) rather than compressed text files (the older GCTA format). Other improvements that are needed for future development include more user-friendly command line options for running iGRM, better GRM calculation progress tracking (e.g. estimated time to completion), better cache utilization, and possible implementation using graphics processing units (GPU's).

iSim — SIMULATING WIDE-SPREAD CUMULATIVE GENETIC INTERACTION EFFECTS

Introduction

A key component of validating new genetic analysis methods is testing them on datasets containing appropriate effects — either effects from real or simulated data. Depending on the study question and method goal, simulated data may or may not be appropriate. To date, much is unknown about the true genetic architecture and effects of epistasis on traits. Current methods for simulating genetic interaction effects, including EpiSIM [92], GAMETES [93], and GenomeSIMLA [94], embed effects by specifying penetrance functions modeling epistasis; however, they only simulate interactions between a limited number loci. Such methods are in line with current approaches of testing for interactions, which test individual SNP pairs for statistically significant effects.

However, if epistatic effects are exhibited through many small-effect interactions that contribute cumulatively to trait variance, as iGRM tests for, existing simulation strategies are not sufficient for embedding similar effects. Our iSim method seeks to address this gap and enable simulation of wide-spread, small-effect interactions that cumulatively contribute to user-specified levels of heritability resulting from many "causal" SNP pairs. In addition, simulations generated with iSim allow us to test iGRM with appropriately embedded additive, dominant, and interaction effects. We develop iSim software and compare accuracy and precision of simulating specific genetic effects for various levels of heritability. In addition, we perform tests of association for individual pairs of SNPs to better understand the architecture of epistatic effects generated by iSim.

Methods

Adaptation and software design

Just as we extended the way GCTA calculates GRMs for additive genetic effects to also calculate dominant and epistatic components, here we use a simulation framework provided by GCTA to develop iSim (also written as a C++ package). Simulation strategies often take the approach of embedding certain effects by simulating genotypes from scratch; GCTA, however, takes the reverse approach by using existing genotype data to simulate phenotype values that match the genotypes in a way that yields a particular genetic effect [43].

Here, we describe the statistics behind GCTA's simulation approach and explain how we modified it to allow for simulation of dominant and epistatic effects, in addition to additive effects. Equation 6 shows the general equation used by GCTA to simulate effects, where Y represents the simulated quantitative phenotype value. W is a standardized genotype matrix and β is a vector of causal SNP effects. (GCTA's notation for the vector of causal SNP effects is u, however we use β.) The residual effect is represented by ε, which we will describe more below.

$$Y = W\beta + \varepsilon \tag{6}$$

$$Y_j = \sum\left(w_{ij} \times \beta_i\right) + \varepsilon_j \tag{7}$$

The "simple" notation shown in Equation 6, and presented in GCTA's software publication [43], can be expanded to be more descriptive (Equation 7). Here, i represents a given SNP and j represents a given person. The genotype matrix, W, is calculated where each SNP value for an individual is equal to w_{ij}, as is shown in Equation 8, where x_{ij} is the number of copies of the reference allele (p_i) for the ith SNP of the jth individual. Each w_{ij} is weighted by a corresponding β value, which is drawn from a normal distribution. Notably, GCTA has conflicting documentation — while the original GCTA article [43] states that β effects are generated from a standard normal distribution (mean zero, variance one), newer documentation on its website reflects that β effects are generated from a normal distribution with a mean of zero and a variance equal to the variance of all calculated w's for the respective SNP across all individuals in a dataset. We use the newer method for iSim.

$$w_{ij} = \frac{(x_{ij} - 2p_i)}{\sqrt{(2p_i(1-p_i))}} \tag{8}$$

The residual effect (sometimes referred to as the error term) is what allows a specific level of heritability to be embedded. With zero error, genetic variation will perfectly predict phenotypic variation, resulting in 100% heritability. To simulate less than 100% heritability residual error is added, lessening the correlation between genetic and phenotypic variation. The residual effect ε_j is calculated as shown in Equation 9, where h^2 is the desired level of heritability to be simulated and $\sigma^2{}_g$ is the variance of the W matrix (Equation 10).

$$\varepsilon_j = \sigma^2{}_g\left(\frac{1}{h^2-1}\right) \tag{9}$$

$$\sigma^2{}_g = var\left(\sum w_{ij} \times \beta_i\right) \tag{10}$$

In summary, the previous equations take existing genotype data, assign random effect sizes based on SNP allele frequencies, then scale variation of phenotype assignments to simulate a specific level of trait heritability. The equations shown thus far, however, only allow for the simulation of heritability due to additive genetic effects. We modify these calculations to allow heritability to be simulated additionally for dominant and epistatic effects by taking advantage of the equivalency (described in Equation 11) between the W matrix formula (Equation 8) and the Cockerham encodings for variance and covariance components described in the introduction to this chapter (Tables 10 and 11).

$$w_{ij} = \frac{(x_{ij}-2p_i)}{\sqrt{(2p_i(1-p_i))}} = \frac{Covar.}{\sqrt{Var.}} = \frac{\sqrt{Covar.\times Covar.}}{\sqrt{Var.}} = \frac{Covar.\times Covar.}{Var.} \qquad (11)$$

For iSim we use the same frequency calculation step as iGRM (Figure 18). The variance and covariance values are then used in a new step which implements the formulas and calculations described above. Processing is not computationally intensive because calculations are at the SNP pair level and not for all pairs of individuals (as with iGRM) — thus only one worker (processor) is needed and used per replicate. Figure 24 shows a general outline for iSim and a more detailed overview of how the data is structured for each step of iSim is shown in Appendix D.

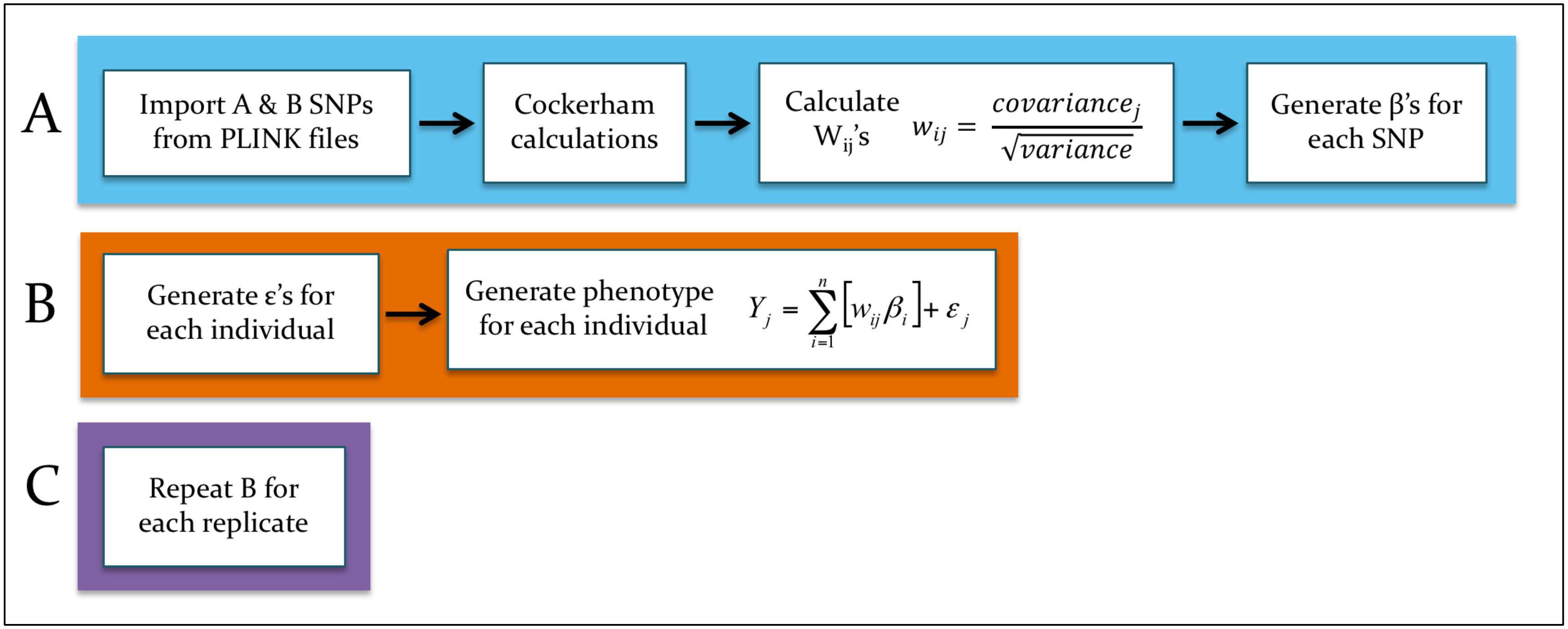

A
Import A & B SNPs from PLINK files
Cockerham calculations
Calculate W$_{ij}$'s $w_{ij} = \frac{covariance_j}{\sqrt{variance}}$
Generate β's for each SNP
B
Generate ε's for each individual
Generate phenotype for each individual $Y_j = \sum_{i=1}^{n}[w_{ij}\beta_i] + \varepsilon_j$
C
Repeat B for each replicate

Figure 23. iSim processing workflow.
(**A**) Frequency, variance, and covariance calculations. (**B**) Phenotypes are specified for every individual to match a user-specified level heritability for a given genetic component. (**C**) Step B is repeated for additional replicates, using the same calculations from Step A each time. Supplementary data (e.g. Beta effect sizes and residual values) are currently only output for the first replicate.

Simulation analyses

After developing iSim we simulated heritability using various parameters, estimated power based on GRM variance, tested for individual-level SNP-SNP interactions, and performed REML analysis to estimate the genetic effects observed from each simulation (Figure 24). Our strategy for simulating genetic effects and subsequently testing the ability of GRMs to capture those effects is modeled based on previously described methods [49].

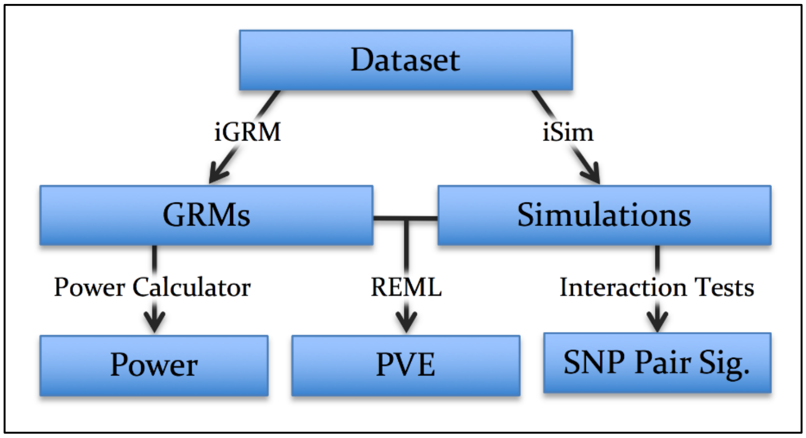


Figure 24. Simulation analysis and method validation workflow.
iSim is used to simulate heritability for genetic effects. iGRM is used to generate GRMs. GCTA's GREML power calculator is used to estimate power based on genetic variance. Interaction regression tests are used to test individual SNP pairs for significance. REML analysis is used to fit GRMs to estimate genetic effects embedded in each and calculate significance via likelihood ratio tests.

Dataset: Using the AMD dataset described in Chapter 5 we randomly selected 10,000 individuals and 200 bi-allelic SNPs with minor allele frequencies greater than 5%. We further divided the SNPs into two groups of 100 SNPs and flagged the first 50 SNPs in each group to be assigned as causal for simulating genetic effects.

iGRM and power: We then used the dataset of 10,000 individuals to generate GRMs for all 200 SNPs, using iGRM, and used the off-diagonal variances of each GRM to estimated power (as described in Chapter 2).

iSim: Next, we used iSim to embed multiple levels (1%, 5%, 10%, and 20%) of additive, dominant, and interaction genetic effects from the SNPs flagged as causal, as described above, for six replicates (4 heritability levels × 8 genetic components × 6 replicates = 192 simulations). Each SNP group (A and B) contains 50 SNPs assigned to be causal for simulation of genetic effects, thus there are 50 SNPs used for each respective additive and dominant simulation and 2,500 (50 × 50) total SNP pairs used for interaction simulations. For each simulation we only embed a genetic effect for one of the eight genetic components — subsequent analyses (i.e. ordinary least squares (OLS) regression or REML) estimate potential contributions from each of the eight effects, regardless of which effect was simulated. This approach allows us to better estimate the accuracy of embedding specific effects via simulations for different genetic effects and different levels of heritability.

SNP-SNP interaction tests: We hypothesized that simulated SNP pair effect sizes would be correlated with regression-determined (observed) effect size and thus that SNPs/SNP pairs assigned to be causal for a specific genetic effect would exhibit increased significance for tests of the corresponding genetic effect, compared to non-causal SNPs/SNP pairs. To calculate the effect size and significance of each individual SNP pair for all simulations we created a script to perform OLS regression using genotype encodings described by Cordell [87]. Additive encodings for AA, Aa, and aa are 1, 0, and -1, respectively. Dominant encodings for AA, Aa, and aa are -0.5, 0.5, and -0.5, respectively. Multi-locus genotype encodings are then calculated as products of combinations of respective genotype encodings, such that the multi-locus interaction encoding for Add. A × Dom. B, given the genotype pair AA/Bb, would be 0.5 (1 × 0.5). These genotype and multi-locus genotype encoding were determined for every SNP pair in our dataset (200 × 199 / 2 = 19,900). Association tests could have been limited to between-group interactions (100 × 100 = 10,000), but we tested all possible pairs to check for possible effects from SNP pairs in the same group, not assigned as causal for simulations. We performed OLS regression using the general linear regression equation (Y = a + bX), where Y is the phenotype and X is the respective genotypic encoding, with output from regression being the slope (b) and the y-intercept (a). The effect size of the SNP or SNP pair is then calculated as the Pearson's correlation coefficient (P_{xy}), where P_{xy} = [covariance(X,Y)]/[$\sigma_X \times \sigma_Y$]. The significance of the SNP or SNP pair is then calculated as a p-value calculated from a Student's t-distribution using a t-statistic (t), where t is equal to the slope divided by the standard error of the slope (t = b / SE). From these results we calculated summary statistics to compare average p-values between causal and non-causal SNP pairs, for each genetic component, stratified by the level of simulated heritability. Additionally, we calculated sensitivity of the individual SNP-SNP tests of association as the proportion of times SNP pairs assigned to be causal for a given genetic effect had significant p-values for tests of each genetic effect.

PVE: Finally, we used the eight generated GRMs to fit each simulated dataset (N=192) using REML analysis. Average information (AI) and Fisher's scoring REML algorithms would not allow for analysis due to matrix inversion issues. We performed REML analysis using the expectation maximization (EM) algorithm, however the likelihood of the models did not converge for any of the simulations — we assume this is due to the difficulty of model fitting using eight large GRMs. GCTA only performs a likelihood ratio test (LRT), fitting a reduced model, after the full model converges. Since the models never converged we used the final estimations to calculate the best estimates of the proportion of variance explained (PVE) for each genetic component. The PVE explained by each genetic component was calculated as V(#)/V(p), where V(#) is the genetic variance estimate for a given GRM (1-8) and V(p) is the phenotypic variance estimate [V(p) = V(1) + … + V(8) + V(e)], where V(e) is the residual variance. Finally, we performed likelihood ratio tests (LRTs) for select simulations (described in the results) by first manually calculating reduced models (e.g. to test the effect of Additive A for a simulation, run REML with 7 GRMs, excluding the Additive A GRM). GCTA calculates the likelihood of a model for every iteration — we manually calculated LRT test statistics as twice the difference between the full and the reduced model likelihoods. P-values then were calculated from chi-squared distributions using the LRT test statistics. For summary information and results interpretation we calculated PVE for each simulation, averaged across replicates, and formatted the data so that — separately for each level of simulated heritability — rows corresponded to simulated genetic components and columns corresponded to the average observed PVE for each genetic component. Within the aggregated data, table diagonals represent observed PVE for the same component which was simulated to have an effect while off-diagonals represent observed PVE for the components not simulated to exhibit a genetic effect.

Results

Using iGRM we generated eight orthogonal GRMs for each of the simulated datasets described in the methods. From each GRM we calculated the off-diagonal variances (ODVs) and subsequently the estimated power for each GRM (Table 13). The GRMs and minimum detectable heritabilities are independent of simulated phenotypes, so this only had to be performed once. Power, here, is representative of the minimum heritability that can be detected with 80% power, at an alpha level of 0.05, determined using GCTA's online GREML power calculator.

Table 13. Estimated power for simulation dataset using off-diagonal GRM variance.

Component	ODV†	Min. h^2‡
Additive A	1.01 e-02	0.395%
Dominant A	9.92 e-03	0.398%
Additive B	1.01 e-02	0.394%
Dominant B	9.88 e-03	0.399%
A × A	1.03 e-04	3.913%
A × D	1.00 e-04	3.962%
D × A	9.21 e-05	4.129%
D × D	9.97 e-05	3.969%

† ODV = off-diagonal variance. ‡ Minimum heritability that can be detected with at least 80% power. GRMs here were generated with a dataset of 10,000 individuals and 100 SNPs per group. For the interaction components, A = additive and D = dominant.

Using phenotypes simulated to contain heritability effects for each genetic component at 1%, 5%, 10%, and 20% heritability, as described in the methods section, we fit each simulation using REML to determine the proportion of simulated trait variance explained (PVE) by genetic variance from each genetic component. We then averaged the PVE's across all replicates for visualization and comparison (Table 14). Table 14 is colored to highlight the diagonals, to compare estimates between different levels of simulated heritability, and to compare "background noise" between single SNP (additive/dominant) and interaction components.

Table 14. Observed average proportion of variance explained in simulated datasets.

					Average PVE				
	Component	**Add A**	**Dom A**	**Add B**	**Dom B**	**A × A**	**A × D**	**D × A**	**D × D**
1% Simulated	**Add A**	0.0097	0.0013	0.0011	0.0008	0.0136	0.0091	0.0067	0.0069
	Dom A	0.0011	0.0095	0.0013	0.0011	0.0104	0.0102	0.0059	0.0060
	Add B	0.0014	0.0008	0.0104	0.0007	0.0080	0.0061	0.0069	0.0065
	Dom B	0.0011	0.0011	0.0012	0.0098	0.0092	0.0065	0.0083	0.0106
	A × A	0.0012	0.0011	0.0012	0.0013	0.0175	0.0109	0.0095	0.0095
	A × D	0.0010	0.0009	0.0010	0.0012	0.0054	0.0096	0.0056	0.0053
	D × A	0.0015	0.0008	0.0017	0.0013	0.0100	0.0078	0.0087	0.0064
	D × D	0.0011	0.0016	0.0012	0.0013	0.0089	0.0067	0.0060	0.0107
5% Simulated	**Add A**	0.0493	0.0010	0.0013	0.0012	0.0067	0.0103	0.0101	0.0074
	Dom A	0.0010	0.0503	0.0009	0.0008	0.0089	0.0099	0.0049	0.0086
	Add B	0.0009	0.0007	0.0490	0.0010	0.0098	0.0068	0.0042	0.0039
	Dom B	0.0019	0.0010	0.0008	0.0480	0.0091	0.0076	0.0096	0.0087
	A × A	0.0011	0.0014	0.0008	0.0012	0.0528	0.0079	0.0062	0.0042
	A × D	0.0010	0.0010	0.0010	0.0010	0.0081	0.0577	0.0071	0.0075
	D × A	0.0007	0.0009	0.0008	0.0009	0.0099	0.0123	0.0459	0.0069
	D × D	0.0008	0.0010	0.0012	0.0007	0.0131	0.0130	0.0065	0.0372
10% Simulated	**Add A**	0.1026	0.0005	0.0007	0.0009	0.0122	0.0077	0.0056	0.0061
	Dom A	0.0013	0.1021	0.0009	0.0011	0.0079	0.0072	0.0061	0.0068
	Add B	0.0009	0.0009	0.1005	0.0009	0.0101	0.0054	0.0057	0.0045
	Dom B	0.0011	0.0012	0.0008	0.1014	0.0073	0.0060	0.0120	0.0082
	A × A	0.0007	0.0010	0.0010	0.0012	0.0931	0.0081	0.0125	0.0072
	A × D	0.0009	0.0008	0.0019	0.0012	0.0105	0.1044	0.0053	0.0127
	D × A	0.0010	0.0011	0.0008	0.0011	0.0132	0.0113	0.1000	0.0079
	D × D	0.0013	0.0012	0.0010	0.0018	0.0094	0.0146	0.0079	0.0905
20% Simulated	**Add A**	0.1961	0.0007	0.0011	0.0010	0.0079	0.0100	0.0073	0.0052
	Dom A	0.0009	0.2014	0.0010	0.0008	0.0059	0.0054	0.0070	0.0063
	Add B	0.0007	0.0012	0.2053	0.0010	0.0080	0.0108	0.0065	0.0077
	Dom B	0.0010	0.0012	0.0007	0.2014	0.0077	0.0057	0.0063	0.0074
	A × A	0.0009	0.0008	0.0011	0.0014	0.2024	0.0063	0.0080	0.0074
	A × D	0.0009	0.0013	0.0010	0.0009	0.0092	0.1905	0.0054	0.0066
	D × A	0.0018	0.0013	0.0008	0.0006	0.0138	0.0097	0.2097	0.0071
	D × D	0.0012	0.0008	0.0011	0.0009	0.0059	0.0068	0.0055	0.1971

Values colored from red to green, with bright green representing the smallest values and bright red representing the largest values. Values for the standard error of the means are in Appendix E. 3-D depictions of 1% and 20% simulations are in Appendix F.

In Table 14 there is a very consistent trend that (A) the diagonals exhibit observed effects (average PVE) very close to their simulated effects and (B) the off-diagonal observed values for single SNP components (additive/dominant) and interaction components are relatively consistent within those two groupings. With this in mind, we summarized the PVEs based on these groupings (Table 15). We manually tested reduced models for representative simulations then calculated LRT test statistics and p-values (Table 15 and Appendix G).

As expected, we observed significant p-values (less than 0.05) for all simulated components (diagonal values), with the exception of the interaction component simulated to have a 1% effect size. Additionally, that interaction effect was the only one we tested that exhibited a significant value for an off-diagonal component (i.e. false-positive). P-values were calculated for select simulations from one replicate (Appendix G), though, so exhaustive significance tests across all simulations would likely yield significant average p-values for all diagonals and non-significant averages for all off-diagonals. At low simulated heritability effects, near 1%, we still would expect a higher number of false-positives and false-negatives than for simulations with higher levels of simulated effects.

GCTA's GREML power calculator, which calculates power using off-diagonal variances for each GRM, showed that, given our dataset, we would be powered to detect heritability greater than about 4% from any interaction component (Table 13). Notably, our results using simulations for the same dataset showed that 'background noise' could be distinguished from the true signal starting at around 5% for interaction components

(Table 14). These results show that power estimated from GCTA's GREML power calculator are in agreement with the trends observed from our simulations using iSim. Thus, GCTA's GREML power calculator can be used to most easily estimate basic statistical power for a dataset based on sample size.

Table 15. Proportion of variance explained summary and statistical significance.

	Simulated	Single SNP	Single SNP	Interaction	Interaction
	Effect	Diagonal	Off-diagonal	Diagonal	Off-diagonal
Average PVE	1%	0.99%	0.12%	1.16%	0.80%
	5%	4.92%	0.10%	4.84%	0.82%
	10%	10.17%	0.10%	9.70%	0.86%
	20%	20.10%	0.10%	19.99%	0.74%
Select P-values	1%	0.0014	0.4032	0.0521	0.0373
	5%	6.8 e-69	0.3274	0.0006	0.2919
	10%	1.3 e-173	0.1184	3.4 e-14	0.0756
	20%	< 1 e-300	0.1563	3.4 e-40	0.2312

Average PVE summarized from Table 14. Select representative p-values calculated manually as described in Appendix G.

Next, we performed regression for each SNP pair for all simulations. Genotype pairs were encoded using the Cordell model described above and regressed to the simulated phenotypes. We summarize the results first by comparing average significance for each component, stratified by causal versus non-causal SNP pair tests (Table 16). We expected to see enrichment in significance (smaller p-values) for causal SNP pairs versus non-causal SNP pairs when the simulated genetic component was tested using the same corresponding genetic encoding (i.e. the diagonals, which are outlined separately in black). For the off-diagonal components, representing tests for genetic effects that were not simulated (i.e. null effects), we expected to see roughly the same average p-values for causal and non-causal SNPs, for each separate component. We observed several expected trends. Generally, increased simulated heritability corresponded with greater differences in significance between causal and non-causal SNP pairs. Higher heritability corresponds with increased correlation between genotypic and phenotypic effects resulting in causal SNP pairs being more predictive of the phenotype and therefore more statistically significant. The simulation procedure introduces variability (error) to genotype-phenotype correlation to scale heritability lower, so the observed trend is expected.

We also observed a consistent trend of much larger differences in average significance between causal and non-causal SNP pairs for additive and dominant simulated effects compared to simulated interaction effects; we propose two potential explanations. First, additive and dominant simulations assign effect sizes per SNP (N), while interaction simulations are per SNP pair (N × N); thus, each causal SNP for additive and dominant effects contributes to a much greater proportion of the overall embedded level of heritability than any individual SNP pair does when interaction effects are simulated. Second, due to the nature of the regression and multi-locus genotype encoding used, when additive or dominant effects are simulated a given SNP assigned a specific effect size is tested multiple times — one test for every SNP pair that it is a member of. Due to this effect, in Table 16 we assigned colors for values separately for single SNP (additive/dominant) and interaction components (column-wise) to better distinguish the subtle enrichment in significance for the interaction effects simulated to be causal.

A final observation from Table 16 is that based on which additive or dominant effect is simulated, other effects see marginal enrichment in significance. For example, when 20% heritability was simulated for the additive B genetic component, we also see slight enrichment in significance from the additive A component but not from the dominant A or dominant B component. In general, when an additive effect is simulated for one group (A or B), additive effects from the other group are affected slightly, but dominant effects are not. On the other hand, when dominant effects are simulated, additive and dominant effects from both groups are affected. This is expected due to the nature of genotypic encoding for additive effects compared to dominant effects and is the reason most genetic analyses test only for additive effects — because additive tests are able to capture some existing dominant effects.

Table 16. Average interaction p-values for tests of simulated components.

		Average Observed p-values															
		C	NC	C	NC	C	NC	C	NC	C	NC	C	NC	C	NC	C	NC
	Component	**Add A**		**Dom A**		**Add B**		**Dom B**		**A × A**		**A × D**		**D × A**		**D × D**	
1% Simulation	**Additive A**	0.402	0.480	0.491	0.499	0.470	0.469	0.503	0.499	0.494	0.495	0.497	0.496	0.497	0.496	0.496	0.495
	Dominant A	0.457	0.502	0.377	0.492	0.475	0.475	0.456	0.457	0.498	0.496	0.499	0.496	0.500	0.496	0.500	0.496
	Additive B	0.470	0.468	0.507	0.509	0.394	0.503	0.513	0.511	0.498	0.494	0.497	0.495	0.498	0.496	0.499	0.495
	Dominant B	0.486	0.484	0.471	0.467	0.440	0.502	0.380	0.499	0.500	0.495	0.495	0.496	0.497	0.495	0.495	0.495
	A × A	0.499	0.502	0.492	0.491	0.479	0.492	0.486	0.488	0.497	0.494	0.499	0.495	0.497	0.495	0.498	0.496
	A × D	0.497	0.502	0.495	0.503	0.524	0.509	0.508	0.498	0.494	0.495	0.493	0.497	0.496	0.495	0.496	0.496
	D × A	0.486	0.498	0.496	0.497	0.487	0.493	0.499	0.506	0.504	0.497	0.500	0.498	0.495	0.496	0.494	0.497
	D × D	0.478	0.483	0.470	0.477	0.502	0.495	0.470	0.485	0.496	0.494	0.494	0.496	0.492	0.495	0.489	0.496
5% Simulation	**Additive A**	0.262	0.493	0.485	0.496	0.437	0.435	0.490	0.488	0.501	0.495	0.500	0.495	0.499	0.495	0.499	0.496
	Dominant A	0.360	0.500	0.210	0.493	0.462	0.461	0.432	0.429	0.497	0.497	0.497	0.495	0.496	0.496	0.498	0.495
	Additive B	0.437	0.435	0.491	0.491	0.293	0.497	0.498	0.503	0.498	0.496	0.501	0.498	0.499	0.497	0.502	0.496
	Dominant B	0.465	0.461	0.423	0.422	0.378	0.493	0.230	0.496	0.497	0.494	0.498	0.494	0.498	0.495	0.498	0.493
	A × A	0.490	0.491	0.482	0.482	0.500	0.498	0.493	0.493	0.479	0.495	0.496	0.496	0.495	0.495	0.499	0.496
	A × D	0.502	0.494	0.511	0.492	0.490	0.492	0.480	0.485	0.494	0.496	0.472	0.496	0.497	0.495	0.491	0.496
	D × A	0.491	0.491	0.493	0.495	0.482	0.489	0.486	0.489	0.496	0.493	0.495	0.493	0.472	0.495	0.496	0.495
	D × D	0.514	0.495	0.491	0.505	0.507	0.500	0.491	0.497	0.496	0.493	0.489	0.497	0.488	0.494	0.472	0.493
10% Simulation	**Additive A**	0.190	0.502	0.513	0.509	0.432	0.428	0.511	0.506	0.499	0.494	0.500	0.495	0.499	0.495	0.499	0.494
	Dominant A	0.333	0.490	0.164	0.498	0.447	0.449	0.402	0.403	0.494	0.496	0.498	0.495	0.497	0.495	0.499	0.493
	Additive B	0.417	0.414	0.495	0.492	0.237	0.499	0.491	0.492	0.500	0.495	0.500	0.496	0.499	0.495	0.496	0.497
	Dominant B	0.467	0.464	0.417	0.413	0.299	0.508	0.184	0.488	0.496	0.495	0.498	0.496	0.497	0.496	0.501	0.496
	A × A	0.488	0.493	0.492	0.499	0.467	0.492	0.478	0.487	0.464	0.492	0.491	0.495	0.502	0.494	0.500	0.496
	A × D	0.482	0.488	0.499	0.491	0.481	0.484	0.488	0.498	0.479	0.494	0.461	0.492	0.496	0.491	0.494	0.493
	D × A	0.459	0.481	0.467	0.481	0.496	0.478	0.488	0.472	0.485	0.491	0.495	0.491	0.456	0.493	0.495	0.495
	D × D	0.467	0.495	0.497	0.490	0.491	0.495	0.486	0.492	0.487	0.492	0.476	0.492	0.478	0.492	0.442	0.491
20% Simulation	**Additive A**	0.156	0.486	0.506	0.488	0.410	0.407	0.502	0.498	0.499	0.493	0.500	0.493	0.494	0.495	0.499	0.494
	Dominant A	0.293	0.488	0.133	0.500	0.440	0.439	0.419	0.413	0.497	0.496	0.496	0.495	0.500	0.497	0.499	0.496
	Additive B	0.409	0.406	0.494	0.493	0.155	0.501	0.493	0.496	0.496	0.495	0.498	0.495	0.496	0.496	0.497	0.495
	Dominant B	0.453	0.450	0.403	0.400	0.300	0.488	0.142	0.486	0.495	0.493	0.497	0.493	0.497	0.494	0.499	0.494
	A × A	0.475	0.505	0.494	0.491	0.457	0.498	0.474	0.496	0.443	0.489	0.493	0.495	0.501	0.492	0.495	0.496
	A × D	0.479	0.497	0.473	0.494	0.484	0.489	0.492	0.495	0.466	0.491	0.425	0.494	0.487	0.488	0.489	0.493
	D × A	0.417	0.470	0.445	0.488	0.498	0.474	0.514	0.486	0.466	0.486	0.487	0.487	0.424	0.492	0.489	0.493
	D × D	0.486	0.494	0.497	0.499	0.497	0.500	0.472	0.500	0.481	0.489	0.458	0.489	0.465	0.487	0.405	0.485

C = Causal (includes effects drawn from a normal distribution centered around zero); NC = Non-causal effects; Red = most significant; Green = least significant. Values are averages over all assessed SNPs (additive and dominant) and SNP pairs (interaction components).

Discussion and conclusions

One of the goals of iSim was to validate our method for calculating additive, dominant, and epistatic GRMs (iGRM). To do this we sought to show that for datasets containing specific levels of varying genetic effects (additive, dominant, or interaction) those effects could be quantified and captured via calculation of orthogonal GRMs. By simulating multiple replicates for multiple genetic effects using different levels of embedded heritability we were able to show that specific levels of heritability can be accurately estimated for each of the eight genetic components (Table 14), with accuracy depending on the level of heritability, among other factors, such as sample size. Additionally, we sought to show that datasets with simulated overall heritability (cumulative effects) contain individual-level, detectable effects. We observed both an enrichment in significance for SNP pairs simulated to be causal versus non-causal as well as a positive correlation between sensitivity and the level of heritability, noting that estimates are slightly skewed due to all SNP pairs with non-zero simulated effect sized being assigned as causal. When large levels of heritability due to interaction effects are embedded, individual-level SNP tests of those interactions were enriched for significance, supporting the need for, and power of, tests of cumulative effects.

Future directions

The simulation method presented here is a major advancement over current epistasis simulation methods. Additional analyses and improvements are warranted to make iSim more powerful and useful. For our analyses we used a custom script to perform linear regression using Cordell model-encoded genotype values. Additional comparisons of significance enrichment and sensitivity for simulated datasets using other association methods, such as INTERSNP and FastEpistasis, are also warranted.

In particular, iSim would benefit from the addition of two options — (1) the ability to simulate weighted levels of heritability for multiple genetic effects simultaneously and (2) the ability to simulate heritability for case-control datasets. We have already started development of methods to enable simulation of heritability due to multiple genetic components (Figure 25), but the statistics developed thus far are not fully worked out. We were able to simultaneously embed weighted effects across the four single SNP components (additive/dominate A & B) and four interaction components, separately, but not any weighted combinations that included at least one single SNP component and one interaction component. We believe that the limitation may be due to differences in variance introduced when simulating phenotypes using SNP pair data compared to single SNP data. Lastly, our simulation method functions by specifying continuous quantitative phenotypes as output. Adaptation to simulate heritability for case-control datasets would require implementing a liability threshold model, which assigns case status depending on whether the generated phenotype value is greater or less than a calculated threshold. While this may be somewhat straightforward, the power to accurately embed specific effects would likely be reduced compared to using continuous phenotypes.

$$\boxed{Y_{j_i} = \sum_{c_i \in c} \left[\varphi_i \sum_{n_i \in n} \left(W_{n_i j_i} \beta_{n_i} \right) \right] + \epsilon_{j_i}} \qquad h^2 = {h^2}_1 + \cdots + {h^2}_c$$

$$\varphi_i = \frac{{h^2}_i}{h^2}$$

$$Y_{j_i} = \sum_{c_i \in c} \left[Q_{c_i j_i} \right] + \epsilon_{j_i} \qquad Q_{c_i j_i} = \varphi_i \sum_{n_i \in n} \left(W_{n_i j_i} \beta_{n_i} \right)$$

$$W_{n_i j_i} = \frac{covariance}{\sqrt{variance}} \qquad Q_{j_i} = Q_{c_1 j_i} + \ldots + Q_{c_c j_i}$$

$$\beta_{n_i} \sim \left(\mu = 0; \sigma = \sqrt{var\left\{ W_{n_i j_1}, W_{n_i j_2}, \ldots, W_{n_i j_j} \right\}} \right)$$

$$\varepsilon_{j_i} \sim \left(\mu = 0;\ \sigma = \sqrt{var\left\{ Q_{j_1}, Q_{j_2}, \ldots, Q_{j_j} \right\} \left\{ \frac{1}{h^2} - 1 \right\}} \right)$$

Figure 25. Proposed scheme for weighted simulations.
Formulas here are similar to the formulas reviewed in the iSim methods section but here include the ability to specify different heritability values for multiple components, which are then used for weighting at multiple steps.

CONCLUSIONS TO METHODS

Previously, the primary way to search for potential epistatic effects for a trait was to perform many independent single SNP pair tests. This introduces an often inhibitory multiple-testing burden because of the large number of tests being performed. To reduce the number of tests and to make analyses more computationally feasible, studies often test interactions where at least one SNP in the SNP pair has a main effect, which may miss actual interactions, depending on the true effect. In this chapter, we introduced two interrelated methods — one method for estimating interaction effects (iGRM) as well as one method for simulating heritability due to interactions (iSim).

The development of iGRM involved modifying methods from existing software (GCTA) to scale genetic relatedness based on orthogonal genetic variance components, based on statistics proposed in 1954 by C. Clark Cockerham. We showed that the variance and covariance values used to calculate additive GRMs in GCTA are relatable to and substitutable with the variance and covariance value calculations described in Cockerham's article. We then described details of the development of C++ software to calculate additive, dominant, and epistatic GRMs as well as the degree to which iGRM becomes computationally difficult/impossible based on sample size (number of individuals and number of SNPs).

Finally, we developed a unique method (iSim) that allows heritability to be simulated due to either additive, dominant, or epistatic effects. In addition to being a useful tool for future studies of epistasis and future development of epistasis analysis methods, we used iSim to show that when epistasis is present (or simulated) those effects can be effectively captured using GRMs and subsequently estimated by fitting in a mixed linear model with GCTA. Lastly, we showed that simulations can be used to determine (*a priori*) the minimum level of heritability that can be accurately estimated given a particular dataset and that and that those estimates are in agreement with GCTA's power calculator. Ultimately, the significance of a given component is determined through model fitting (via GCTA's REML) and a likelihood ratio test.

Chapter 5 — Epistasis in Age-related Macular Degeneration

Background

For genetic studies of many diseases much of the "low-hanging fruit" has been discovered. Variants with large additive effects can often be uncovered relatively easily with modest sample sizes. Additional variants can then be associated by increasing sample size or by studying more diverse populations. In addition to additive effects, there are many other sources of potential genetic contributions to traits and diseases, such as dominant, recessive, and epistatic effects, epigenetic effects, and effects from rare variants and copy number variants (CNVs) [95].

Age-related macular degeneration (AMD) is an example of a disease that has little low-hanging fruit left, in respect to previously captured genetic variation; common variation, in particular, has been extensively studied. A recent well-powered meta-analysis of AMD [36], as discussed in Chapters 1 and 3, uncovered seven additional risk loci, in addition to twelve already-known significant risk loci, by leveraging a large sample size to achieve statistical power. The seven additional loci, however, only improved risk modeling marginally (Figure 26; AUC 0.745 versus 0.753), due to the small but statistically significant effect sizes. In the same study, in addition to additive effects, an interaction analysis between all pairs of the 19 risk loci found that nine SNP pairs had a p-value less than 0.05 (Table 17) [36]. The most significant interaction, and the only interaction that was still significant after a Bonferroni correction, was an interaction between loci in *CFH* and *C2/CFB* (Table 17).

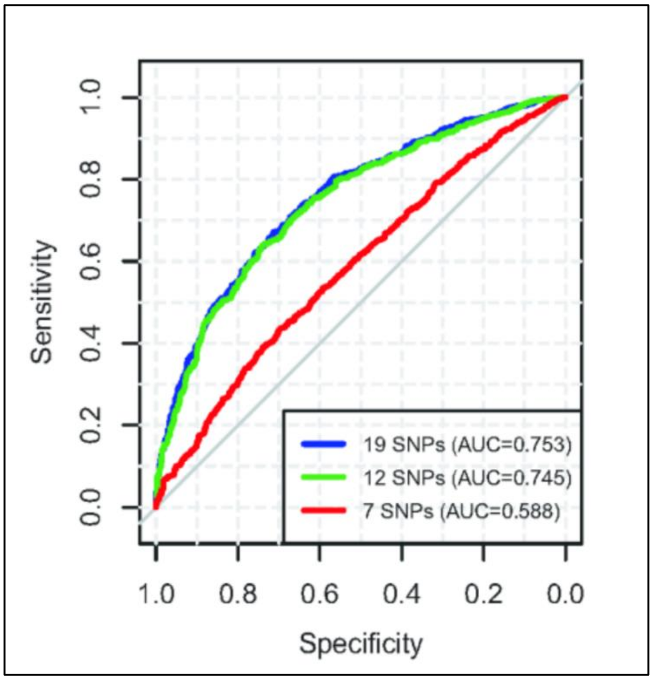


Figure 26. Risk score analysis comparing AMD risk SNP subsets.
The red curve represents risk explained by the seven newly associated loci. The area under the curve (AUC) for the combination of the seven new SNPs with 12 previously significantly associated SNPs (blue curve) differs little from the AUC when using only the 12 SNPs (green curve). Figure from supplemental material in [36].

Table 17. Interactions between known AMD risk SNPs.

Index SNP 1	Gene	Index SNP 2	Gene	Interaction p-value
rs10737680	CFH	rs429608	C2/CFB	0.000052*
rs10490924	ARMS2	rs5749482	TIMP3	0.0052
rs5749482	TIMP3	rs920915	LIPC	0.011
rs1864163	CETP	rs6795735	ADAMTS9	0.021
rs920915	CFI	rs4698775	LIPC	0.022
rs10490924	ARMS2	rs10737680	CFH	0.025
rs5749482	TIMP3	rs3130783	IER3/DDR1	0.034
rs920915	LIPC	rs9542236	B3GALTL	0.038
rs2230199	C3	rs5749482	TIMP3	0.041

Gene is resident gene or nearby gene. For the top interaction, the direction of effect was positive for nine study sites and negative for three. *Statistically significant after Bonferroni correction for 171 tests. Adapted from Fritsche et al. [36].

In this chapter we use iGRM to search for potential cumulative epistatic effects of risk for AMD, given that few individual-level SNP-SNP interactions were statistically significant among the 19 risk SNPs in the referenced study. In particular, we conduct tests for interaction effects between *ARMS2* and several pathways, as well as between regions flanking the top previously published interaction (rs10737680 x rs429608) to estimate its overall contribution to AMD risk, as well as to determine the specific nature of any observed interactions (e.g. additive × additive).

Methods

Dataset and quality control

The dataset used in this study was ascertained and curated by the International AMD Genetics Consortium (IAMDGC) and is composed of 26 studies in total [37]. After receiving approval from institutional review boards at each study site and informed consent from each participant, DNA was collected for cases (intermediate and advanced AMD) and controls (no intermediate or advanced AMD). A custom HumanCoreExome array by Illumina was used to genotype each participant; the array included GWAS-level data (tagging SNPs), exome-level data (from a catalog of protein-altering variants), and select additional SNPs based on previous AMD associations. In total, 55,720 individuals were genotyped for 569,645 SNPs. Individual quality control (QC) measures excluded technical controls, related individuals, all Beaver Dam Eye Study (BDES) participants (due to difficulty in obtaining study approval), participants with unclear phenotypes, and participants with an overall SNP call rate less than 95%. SNP QC excluded SNPs with a minor allele frequency (MAF) less than 5% as well as SNPs with a genotyping rate less than 95%. After QC, 33,603 individuals remained with 252,727 SNPs and a genotyping rate of 99.924%. The reduction in number of SNPs due to QC is high but expected due the large number of rare and low-frequency variants captured by the custom exome chip. Although rare variation is useful for many study questions, in this study we restrict analyses to common variation. Of the individuals who passed QC, 20,561 were cases (8,269 males/12,292 females) and 16,042 were controls (6,992 males/9,050 females).

***CFH-C2/CFB* interaction analysis**

As previously discussed, a statistically significant interaction between two loci in *CFH* (rs10737680) and near *C2/CFB* (rs42908) has been associated with risk for AMD [36]. In this study, we sought to apply our method to the corresponding regions near those loci to quantify the overall effect on risk for AMD of interactions between those regions.

Using the dataset of 36,603 individuals (after quality control) we extracted SNPs within 30 kb of the risk loci (based on chromosomal position) that were part of the top significant interaction for each subset, resulting in 30 SNPs for the *CFH* region and 40 SNPs for the *C2/CFB* region. We then used iGRM to calculate GRMs for additive, dominant, and epistatic effects for the *CFH* and *C2/CFB* regions and fit all eight GRMs in a mixed linear model using GCTA, adjusting for age and sex. All analyses were performed with the default REML algorithm (average information; AI).

Importantly, the genotyping platforms used in the referenced study [36] are different than the one used for this study (HumanCoreExome custom chip). The referenced study performed imputation across study sites, given that the same genotyping platform was not used for each site. The interaction index SNP for *CFH* (rs10737680) was not genotyped in our study; however, some SNPs in linkage disequilibrium with that SNP were genotyped. To confirm that the expected main (additive) effects from *CFH* were present, we performed basic logistic regression for the 30 SNPs in the *CFH* region and found that 29 of the 30 SNPs had a p-value less than 5 e-8. From this we felt confident in proceeding with testing for potential interactions.

ARMS2 interaction analysis

Interaction selection

With over 62 billion SNP pairs in our dataset it is currently computationally infeasible to calculate GRMs representative of all possible interactions, even given efficiently coded software and access to high performance computing clusters; runtime would be on the order of years, even with thousands of processors. Thus, we narrowed down specific interactions to test based on outstanding questions regarding the genetics of AMD.

One of the least understood genetic associations in AMD is *ARMS2* — its biological function and mechanism by which it contributes to AMD risk remains unknown. As described in Chapter 1, *ARMS2* (age-related maculopathy susceptibility 2), also referred to as *LOC387715*, was the second locus found to be associated with AMD, after *CFH* [30]. It is located on the long arm of chromosome 10 and codes for a protein of unknown function [97]. Originally, it was thought that products from *ARMS2* were found mostly in placental tissue and the retina; however, more recent studies have found that it is expressed in multiple human tissues [98], as well as the mitochondrial outer membrane [99]. Additionally, *ARMS2* is near another gene — *HTRA1* — that harbors a variant that has also been associated with risk for AMD; however, one study explicitly refutes *HTRA1's* association [99]. A recent book chapter titled "Gene Structure of the 10q26 Locus: A Clue to Cracking the *ARMS2/HTRA1* Riddle?" [100] re-affirms that *ARMS2's* association with AMD remains somewhat of a mystery. Thus, in this section we use epistasis analysis to search for possible clues about the biological mechanism underlying *ARMS2's* association with AMD risk.

ARMS2-pathway analysis details

To explore potential interactions between *ARMS2* and several AMD-related pathways (as described in Chapter 3) we used PLINK to extract SNPs from the *ARMS2* region and from each pathway using the dataset described in the methods above. First, we extracted SNPs in *ARMS2* (Chr. 10 — 124,187,179 to 124,246,868, including 30 kb flanking; genome build GRCh37). Then, we extracted SNPs from pathways, using pathway gene lists described in Chapter 3, including 30 kb flanking genes in each respective pathway. To generate a rough comparison of differences in computational difficulty for estimating interactions between *ARMS2* and each pathway we calculated the number of frequency and matrix calculations required for each (Table 18). Frequency calculations are based on SNPs and SNP pairs, while matrix calculations are based on SNP pairs and pairs of individuals.

After SNP extraction we had a total of 50 *ARMS2* SNPs and between 370 and 23,430 SNPs for pathways (Table 18). Notably, based on frequency and matrix calculations, TCA, nicotine, oxidative damage, complement, and antioxidant pathways are on roughly the same order of number of required calculations, while other pathways have much higher levels of computational difficulty based on required matrix calculations (Table 18).

Table 18. Comparison of *ARMS2*-pathway interaction scales.

Pathway	# Genes	Pathway SNPs	SNP Pairs	Frequency Calcs.	Matrix Calcs.
Tricarboxylic Acid Cycle	33	370	18,500	6.78 e+08	1.24 e+13
Nicotine	47	560	28,000	1.03 e+09	1.88 e+13
Oxidative Damage	353	743	37,150	1.36 e+09	2.49 e+13
Complement	61	893	44,650	1.64 e+09	2.99 e+13
Antioxidant	70	1,057	52,850	1.94 e+09	3.54 e+13
Angiogenesis	186	7,217	360,850	1.32 e+10	2.42 e+14
Inflammatory	591	7,314	365,700	1.34 e+10	2.45 e+14
Apoptosis	588	23,430	1,171,500	4.29 e+10	7.85 e+14

SNP pairs are calculated using 50 *ARMS2* SNPs. All pathway SNPs exclude *ARMS2* SNPs. Frequency calculations are the number of individuals multiplied by the number of SNP pairs. Matrix calculations are the number of pairs of individuals multiplied by the number of SNP pairs. Ordered by increasing computational difficulty.

After SNP extraction we used iGRM to generate additive and dominant GRMs for *ARMS2* and all pathways as well as interaction GRMs for five of the eight pathways (antioxidant, complement, nicotine, oxidative damage, and TCA). Last, we performed mixed linear model REML analysis using GCTA, adjusting for age and sex, to estimate the effect of each genetic component on risk for AMD. Importantly, for computational feasibility, each genetic component (GRM) was assessed in a separate mixed linear model to ensure model fitting and REML convergence. All analyses were performed with the default REML algorithm (average information; AI).

RESULTS

***CFH-C2/CFB* interaction analysis results**

We used iGRM to test additive, dominant, and epistatic genetic effects for two regions known to harbor a single pair of interacting SNPs (rs10737680 and rs429608, on chromosomes 1 and 6, respectively). These regions contained 30 and 40 SNPs, respectively, resulting in 1,200 pairs of SNPs between the two groups. Though neither of the two SNPs were directly genotyped in our dataset, both have proxy genotyped SNPs with in high linkage disequilibrium (rs10737680 – at least 8 of 30 SNPs genotyped with D' = 1; rs429608 – at least 12 of 40 SNPs genotyped with D' = 1), based on data from HapMap 3, release 2, so we have sufficient coverage of both regions. The additive component for the *CFH* region explained about 7.0% of the risk for AMD while the *C2/CFB* region explained about 3.7% (Figure 27; Table 19). Both additive components were very statistically significant (p = 2.2 e-159 and 3.7 e-22 for *CFH* and *C2/CFB*, respectively). Dominant effects of *CFH* and *C2/CFB* explained about 0.22% and 0.19%, respectively (Figure 27; Table 19) and neither components were statistically significant. The four interaction components together explained a total of 0.63% of risk for AMD (Figure 27; Table 19). Although the interaction effects were small, when a single LRT was performed to drop all interaction components, simultaneously, the interaction p-value was 0.01472 — statistically replicating the previously reported interaction.

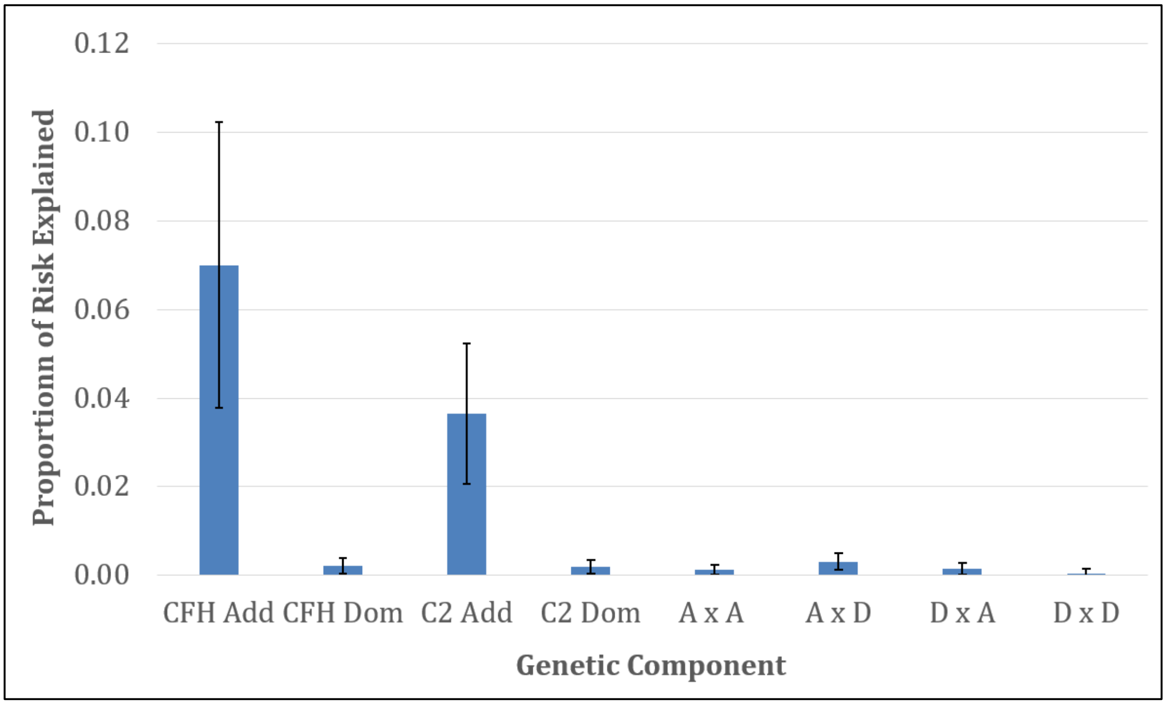


Figure 27. Risk for AMD explained by *CFH* and *C2/CFB* genetic components. *CFH* region = rs10737680 ± 30 kb; *C2/CFB* region = rs429608 (Chr. 6) ± 30 kb.

Table 19. Observed effects and significance of CFH and C2/CFB regions.

	CFH Add	CFH Dom	C2 Add	C2 Dom	A x A	A x D	D x A	D x D
PRE	0.0700	0.0022	0.0365	0.0019	0.0012	0.0031	0.0015	0.0005
SE	0.0323	0.0018	0.0158	0.0015	0.0011	0.0019	0.0014	0.0011
LRT	722.26	0.83	92.31	0.71	3.20	1.01	0.002	0.42
P-val.	2.2 e-159	0.1811	3.7 e-22	0.1997	0.2478	0.0614	0.1869	0.4812

Proportion of risk explained (PRE); Standard error (SE); Likelihood ratio test (LRT); P-value (P-val.). P-values less than 0.05 highlighted in red. C2 = C2/CFB. The sum of interaction effects was 0.63% of AMD risk explained. The overall interaction p-value when all four interaction components were dropped was 0.01472.

***ARMS2* interaction results**

For five pathways (antioxidant, complement, nicotine, oxidative damage, and TCA) we used iGRM to calculate eight orthogonal (additive, dominant, and interaction) GRMs using the same dataset described in the *CFH* x *C2*/*CFB* interaction analysis. For each set of *ARMS2*-pathway interactions, we used GCTA to fit GRMs (separately) in mixed linear models to estimate the amount of risk in AMD explained by additive and dominant genetic components (Figure 28; all eight pathways, plus the *ARMS2* region) as well as interaction components (Figure 29; five *ARMS2*-pathway interactions).

Additive effects from *ARMS2* explained roughly 2.5% of risk for AMD. Additive effects from the angiogenesis, apoptosis, complement, and inflammation pathways explained between 2.7% and 6.9% of the risk for AMD, while additive effects from the antioxidant, nicotine, oxidative phosphorylation, and TCA pathways, each, explained less than 1.0% of the risk for AMD (Figure 28). The additive effects on risk for each pathway (except for TCA) were statistically significant (Table 20). The results here are similar to the additive pathway results presented in Chapter 3, but here we use a larger multi-site cohort and are better powered to more accurately estimate effects from each pathway. Dominant effects from *ARMS2* and each pathway contributed to between near zero and 0.53%, with only the dominant effects from the antioxidant pathway being marginally significant (Table 20).

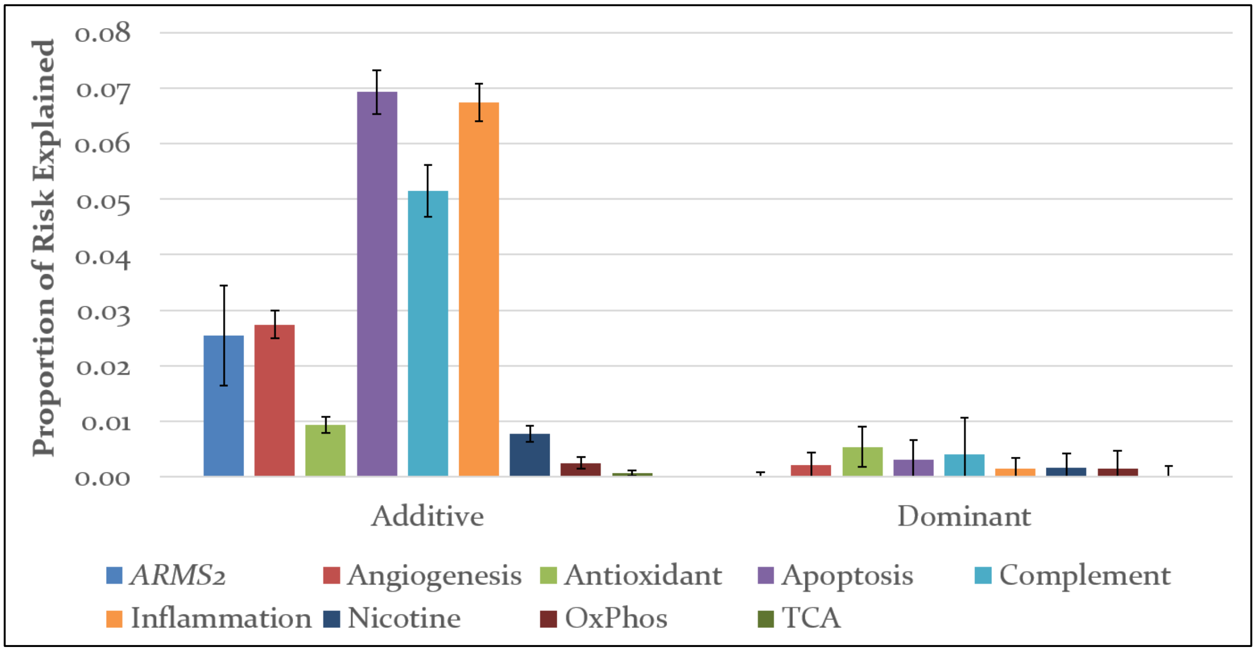


Figure 28. Heritability estimates for additive and dominant genetic effects of *ARMS2* and pathways.

Table 20. Significance of additive and dominant components.

P-val.	*ARMS2*	Angio.	Antiox.	Apop.	Comp.	Inflam.	Nico.	Ox.Phos.	TCA
Add.	3.0 e-102	9.0 e-60	4.0 e-46	4.5 e-147	< 1 e-300	< 1 e-300	1.8 e-20	9.7 e-05	0.1151
Dom.	0.413	0.398	0.043	0.251	0.500	0.317	0.218	0.318	0.500

P-values less than 0.05 are highlighted in red.

Interaction effects were all between near zero and 0.33%, with relatively large standard errors mostly crossing or nearly crossing 0% (Figure 29). The large (relative to the observed effect) standard errors are likely due in part to the fact that that each component was fit in a separate model using REML. Thus, if all eight genetic components, for each pathway, were fit in a single model, the effects from each component would be able to be estimated more precisely and lead to smaller standard errors. Due to limitations we were not able to use a single mixed linear model for each pathway interaction, thus significance (p-values from likelihood ratio tests) is determined for each component, separately. Three of the tested interactions — antioxidant, nicotine, and oxidative phosphorylation — had marginally significant effects (p < 0.05; Table 21).

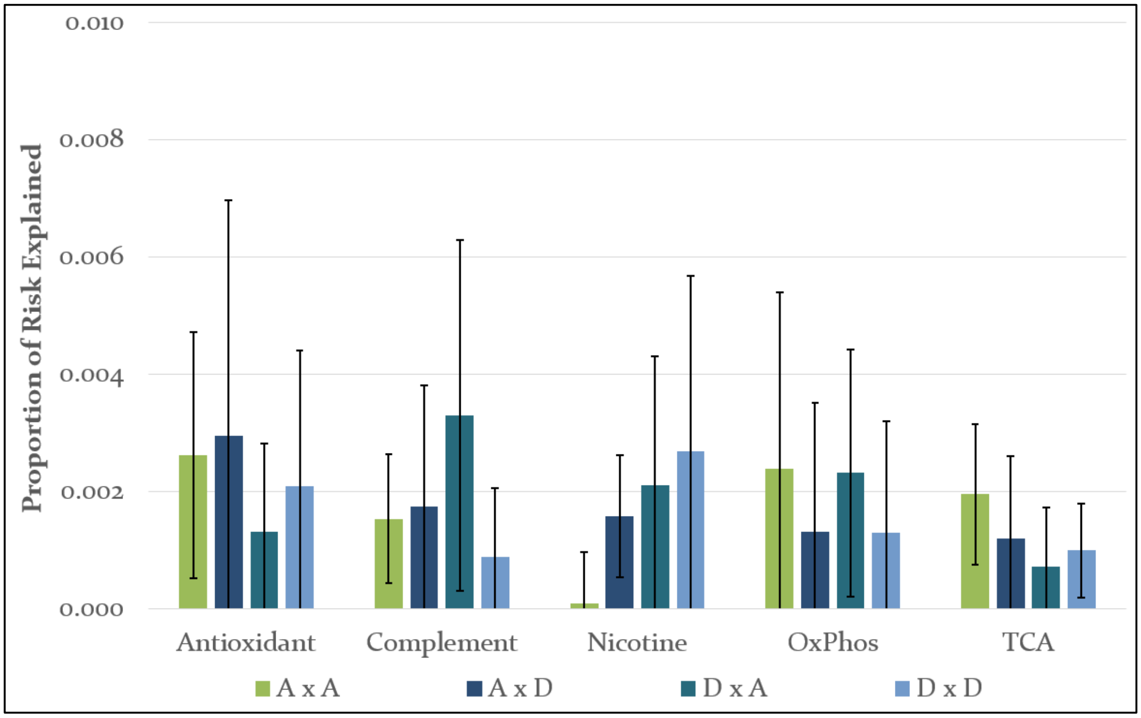


Figure 29. Heritability estimates for ARMS2-pathway interaction effects.
A = additive; D = dominant. The first interacting component listed is the effect from *ARMS2* while the second interacting component is the effect from the respective pathway (*ARMS2* x pathway). No interaction effect exceeded 0.4% risk explained. P-values for interaction components were between 0.0312 and 0.5000 (Table 21). Standard errors were generally large compared to the amount of risk explained, though it is important to note the overall scale. (Y-axis: Min. = 0.0% / Max = 1.0%).

Table 21. Significance of interaction components.

	Antiox.	Comp.	Nico.	Ox.Phos.	TCA
A x A	0.0398	0.0447	0.5000	0.5000	0.0312
A x D	0.5000	0.4430	0.0366	0.5000	0.3490
D x A	0.3100	0.0780	0.2950	0.1510	0.4020
D x D	0.3490	0.4180	0.4430	0.5000	0.1800

P-values from likelihood ratio tests. P-values less than 0.05 highlighted in red.

Discussion

CFH-C2/CFB interaction

For our first analysis of AMD interactions we tested regions near a previously-associated individual-level SNP-SNP interaction that affect AMD risk. We observed additive effects from the *CFH* and *C2/CFB* region that reflect previous knowledge of effects from those genes. Dominant effects explained less than 0.22% of the risk for AMD (not statistically significant). Although no large amounts of risk for AMD were explained by the *CFH* x *C2/CFB* interaction, the total effect explained 0.63% of risk for AMD, a statistically significant amount (LRT p-value = 0.0147), which replicates the previously reported interaction.

Although we observe small, but detectable, effects, there are multiple possible explanations as to why larger effects were not observed. **First**, the AMD interaction reported was from a meta-analysis and, importantly, not all cohorts had the same direction of effect (9 cohorts had positive directions of effect, while 3 cohorts had negative directions of effect) [36], which could confound results when assessing all cohorts simultaneously, as we did. **Second**, as we discussed in the methods and results sections, the primary index SNP (rs10737680) for *CFH* was not on our custom genotyping platform. Although we had multiple proxy SNPs in strong LD with that SNP, there could be subtle effects not captured without rs10737680 directly genotyped. **Third**, it is possible that only one moderate-effect interaction was present and that additional cumulative, additional interactions between those two regions do not exist, resulting in a small overall amount of risk for AMD explained by interactions.

ARMS2 interactions

For our second analysis of AMD interactions we assessed effects from and between *ARMS2* and other pathways. We observed significant results from additive genetic effects from *ARMS2* and all pathways, except TCA, on risk for AMD, which was not unexpected. Interestingly, the results here for additive genetic components are similar to, but not identical to, the trends observed in pathway analyses in Chapter 3. In this analysis we see the strongest signal from additive effects in apoptosis, whereas in Chapter 3, the apoptosis pathway ranked third in terms of the amount of risk for AMD explained.

We observed marginally significant dominant effects from the antioxidant pathway (p = 0.043) but no significant dominant effects from other pathways. When we tested interaction components for effects between *ARMS2* and five pathways, we observed marginally significant effects for three pathways (antioxidant, nicotine, and TCA). The largest total interaction effect for any pathway tested was 0.9% of total risk explained by the antioxidant pathway. However, because all GRMs were fit separately, this estimate is likely an overestimate. Future studies should find a way to fit all of the GRMs simultaneously to get a more precise estimate as well as a single test of interaction significance for each pathway. Future analyses could also assess the larger pathways (angiogenesis, apoptosis, and inflammation) once computational limits are lessened.

Conclusions and future directions

In this chapter we used iGRM to test for potential cumulative interaction effects between regions near a previously reported AMD interaction as well as for interactions between the *ARMS2* region and five pathways. We were able to significantly replicate the interaction between *CFH* and *C2/CFB*. Previous studies have reported statistical significance for specific *CFH* and *C2/CFB* variants, but here we quantify the overall additive effect (heritability estimate) from each region and show that they explain about 4% (*C2/CFB*) and 7% (*CFH*) of the risk for AMD. These estimates may be slight overestimates; however, since each additive GRM contained between 30 and 40 SNPs. To confirm the estimates, a "remainder" GRM could be additionally included in an additive-only mixed model so that genetic variation is more precisely partitioned.

Although we did confirm that cumulative interaction effects between *CFH* and *C2/CFB* were statistically significant, it is currently unknown whether the effect is from a single SNP pair or more cumulative effects. Future analyses could try to remove SNPs in LD with the reported interacting SNPs to see if the cumulative interaction effect become non-significant when not included in the model.

When we applied iGRM to assess interactions between *ARMS2* and multiple pathways, we were able to confirm expected additive effects from each pathway as well as notable dominant and epistatic effects. Additive effects from angiogenesis, apoptosis, complement, and inflammation (PRE 2.74%, 6.93%, 5.15%, and 6.74%, respectively) should be further dissected to narrow down unique and shared genetic effects. Although the overall effects from dominant and epistatic components were relatively small (individually all less than 0.54%), some components were statistically significant — a feat that would not have been possible without such a large sample size. Nonetheless, better estimates, and likely greater significance, would come from modeling all eight GRMs for each ARMS2-pathway analysis in single models, instead of separately for each GRM, as we had to do for now due to computational limitations.

Ultimately, we leveraged having a large case-control AMD dataset to confirm a previously reported AMD interaction with a new method — iGRM. Future studies could extend the methods presented here to test for all possible pairwise regions around known AMD risk SNPs to potentially uncover additional epistatic effects contributing to risk for AMD. Lastly, we applied iGRM to assess potential interactions between *ARMS2* and multiple pathways and found marginal evidence that non-additive effects exist and contribute in part to risk for AMD. The methods and results presented here provide a basis for a more detailed investigation of the detected interactions as well as further investigation of additional possible interactions. We can now say that non-additive effects for AMD are present and that future studies are warranted.

CHAPTER 6 — CONCLUSIONS AND FUTURE DIRECTIONS

Vision loss is the third most feared medical condition — only behind cancer and cardiovascular disease [63]. Age-related macular degeneration (AMD) is a major contributor to visual loss and blindness worldwide and in developed countries is the leading cause of blindness. As the average age of populations increase worldwide, so will the prevalence of AMD, making it is increasingly critical to pursue genetic and molecular research of AMD. No effective method exists to prevent the development of AMD, but AREDS supplements, taken when determined to be at risk or following early signs of AMD, were found to delay or slow progression in some patients [18]. Dry AMD, the more slowly progressing form, has no treatment options. Some treatment options exist for wet AMD, the more quickly progressing form; treatment is often required for the life of the patient after diagnosis and may involve regular injections of vascular endothelial growth factor inhibitors directly into the eye [101].

Much progress has been made towards uncovering genetic effects influencing risk for AMD. However, even the most significantly associated variants do not have a well-understood correlation between genetic and molecular processes affecting disease pathogenesis [63]. To gain a better understanding of the genetics of AMD, and to potentially uncover novel associations that could lead to a better molecular understanding, we performed novel pathway and interaction analyses of AMD. In addition to the work we presented here and the future directions we discussed, analyses of potential epigenetic and noncoding associations could yield additional insight.

With the knowledge that the complement system has been associated with risk for AMD, via several complement genes, we sought to better quantify the overall genetic effect of the complement pathway, as well as other potentially-related pathways. An advantage of our pathway analysis approach was that we were able to estimate cumulative effects, regardless of individual SNP-level significance, so that even non-statistically significant genetic variation contributing cumulatively to risk for AMD effects would be detectable. The pathways we tested for association were angiogenesis, antioxidant activity, apoptotic signaling, complement activation, inflammatory response, response to nicotine, oxidative phosphorylation, and the tricarboxylic acid cycle. We only observed statistically significant additive genetic effects from the complement activation and inflammatory response pathways; however, the significance from the inflammatory response pathway was primarily due to the large number of genes it had in common with the complement activation pathway. Genes in the inflammatory response pathway but not the complement pathway did not have a statistically significant effect. While our results recapitulated the importance of the complement pathway, we interestingly found that additional variation in the complement pathway, separate from known risk variants (and variants in LD with those variants), contributes a statistically significant amount of additional risk for AMD (proportion of risk explained = 7%; p-value = 1×10^{-15}).

Next, we sought to test for potential epistatic effects contributing to risk for AMD, but no method existed that would let us assess cumulative genetic interaction effects in the way that GCTA does for additive effects via genetic relationship matrices (GRMs). Thus, we developed a method (iGRM) to allow GRMs reflective of additive, dominant, and epistatic effects to be calculated by modifying the variance-covariance matrix to incorporate additional genetic effects based on variance calculations published by C. Clark Cockerham [90]. Software development included many modifications and optimizations to calculate GRMs for datasets of ~10,000 individuals and ~100,000 or less interacting SNP pairs. GRMs for datasets of similar sizes can be calculated in less than two days when using 100 processors.

Additionally, we developed a method (iSim) to simulate variable levels of heritability due to either additive, dominant, or epistatic effects. Genetic effects are embedded by assigning phenotypes to individuals in an existing dataset in a way that phenotypic variation is modified to have a specific level of correlation with genetic variation, depending on the specified level of desired heritability. We performed multiple simulations using iSim and tested those simulations using iGRM to show that additive, dominant, and epistatic effects can not only be simulated but can be estimated by creating respective GRMs using iGRM. Both iGRM and iSim serve as useful tools for future genetic studies wishing to assess effects beyond additive effects.

There are many possible pipelines that could be developed to study various traits, implementing iGRM and iSim as a backbone. Key components to the pipeline would be a knowledge-driven SNP selection process, using biological (or other) information to determine which SNPs are more likely to be interacting. For example, one could use chromatin accessibility to include only variants that are in open chromatin regions in a specific disease-related tissue or chromatin conformation data to select only variants potentially interacting due to physical proximity.

By applying iGRM to a large AMD dataset we were able to replicate a previously published interaction between *CFH* and *C2/CFB*; the sum of the four interaction components was small (0.63% PRE) but statistically significant (p = 0.0147). The gene *ARMS2* has been significantly associated with AMD risk in many studies, however, its function remains uncharacterized [100]. With the hypothesis that *ARMS2* could be impacting risk via epistatic effects between AMD-related pathways we applied iGRM to search for a potential missing link between *ARMS2* and AMD. We observed additive effects from the *ARMS2* region (2.5% PRE) and from each of the eight tested pathways (0.77% to 6.93%, excluding oxidative phosphorylation and TCA). Additionally, we observed one marginally significant dominant effect (from the antioxidant pathway) and three marginally significant epistatic effects between *ARMS2* and antioxidant (A x A), nicotine (A x D), and TCA (A x A).

Together, the novel analyses, methods, and results presented here help advance our understanding of the genetics of AMD and provide advancements in statistical genetics that are can be applied to essentially any trait/disease that might have a genetic component. Results shown here should guide future work. For example, the cumulative genetic effects from the complement pathway, that are separate from known risk SNPs, should be further partitioned and analyzed to localize contributing factors. Additionally, the pathway analysis described in Chapter 3 could be applied to a more exhaustive pathway list to potentially implicate a novel mechanism or to further prioritize subsets of genetic variation for additional analyses.

Future studies should investigate any differences in estimating power for additive genetic effects using GCTA's GREML online power calculator (which was designed specifically for additive effects) compared to estimating power for dominant and interaction effects (which was first done in this study). It may be possible that inclusion of all combinations of SNP pair genotypes to calculate interaction GRMs may inadvertently lead to an unexpectedly low genetic variance and subsequently affect power. Whereas SNPs are often filtered out using a minor allele frequency, perhaps it would be beneficial to exclude multi-locus combinations that do not have a sufficient amount of variability to contribute to calculating genetic differences between pairs of individuals. On the other hand, rare variation may contribute to potential genetic interaction effects and removal would reduce power to detect such effects.

In Chapter 4 we described several potential ways to improve iGRM, such as through the utilization of GPUs. Although currently sufficient for gene-level interaction analyses of relatively large sample sizes, further improvements are needed to accommodate larger datasets — a point that will become increasingly important as genetic data is shared and combined to create larger research cohorts. As computational limitations are lessened — likely through both software improvement and more advanced high performance computing clusters — more large-scale questions will be able to be investigated.

Appendices

Appendix A. Risk SNP allele frequencies.

SNP	Chr.	Pos.	Risk Allele	Alt. Allele	EUR*	AFR*
rs10490924	10	124214448	T	G	0.1948	0.2458
rs10737680	1	194946078	C	A	0.4235	0.4650
rs429608	6	32038441	A	G	0.1481	0.1929
rs2230199	19	6718387	G	C	0.2207	0.0333
rs5749482**	22	31389665	C	G	0.1083	0.6633
rs4420638	19	45422946	G	A	0.1978	0.2201
rs1864163	16	55554734	A	G	0.2704	0.2814
rs943080	6	43934605	C	T	0.4851	0.1672
rs13278062	8	23082971	G	T	0.4970	0.8654
rs920915	15	58688467	C	G	0.5030	0.3457
rs4698775	4	110590479	G	T	0.3052	0.0219
rs3812111	6	116443735	A	T	0.3956	0.7360
rs13081855	3	99481539	T	G	0.0964	0.0340
rs3130783	6	30774357	G	A	0.2008	0.4153
rs8135665	22	38476276	T	C	0.2187	0.3676
rs334353	9	100948186	G	T	0.2416	0.2337
rs8017304	14	68785077	G	A	0.4225	0.7027
rs6795735	3	64705365	T	C	0.4353	0.8661
rs9542236	13	30717325	C	T	0.4205	0.1815

* Population risk allele frequency (data from 1000 Genomes Project).

** Frequency data from HapMap (CEU/YRI), instead of 1000 genomes project (EUR/AFR).

Appendix B. Information for principal component analysis.

1,983 AMD dataset individuals (pre-QC)
805 HapMap individuals

- 165 CEU - Utah residents with Northern and Western European ancestry
- 137 CHB - Han Chinese in Beijing China
- 101 GIH - Gujarati Indians in Houston, Texas
- 113 JPT - Japanese in Tokyo, Japan
- 86 MXL - Mexican ancestry in Los Angeles, California
- 203 YRI - Yoruba in Ibadan, Nigeria

To calculate principal components, we used 71 ancestry-informative markers (AIMs) that were present in both HapMap and AMD individuals. 1 Asian-descent and 11 African American individuals were excluded from analysis (circled in red).

Reference SNP IDs (RS numbers) for 71 AIMs:
rs3845596, rs2007350, rs1229133, rs1409778, rs2291409, rs6426327, rs520354, rs1868092, rs975612, rs972881, rs1521527, rs1435850, rs1320131, rs737516, rs1996818, rs1479371, rs1461131, rs1147696, rs2686085, rs225160, rs999634, rs736201, rs173686, rs1807912, rs1560550, rs31251, rs2296412, rs169125, rs942150, rs839556, rs369643, rs1080085, rs3294, rs901170, rs4107736, rs1440369, rs1868280, rs4246828, rs6474795, rs2151065, rs878400, rs7860423, rs11813505, rs722317, rs540819, rs236919, rs1630675, rs916041, rs1548837, rs903770, rs310935, rs1372177, rs981270, rs4904574, rs2873, rs1648282, rs1030588, rs936013, rs461785, rs168206, rs2164062, rs1019977, rs1426311, rs959419, rs1981431, rs186659, rs354731, rs816943, rs2837956, rs756658, rs739096

Appendix C. Pathway gene lists.
Lists of all genes contained within each of the eight pathway assessed.

Angiogenesis: AAMP, ACKR3, ACVR1, ACVR2B, ACVRL1, ADAM15, ADAM8, ADD1, ADM, ADM2, ADRA2B, AGGF1, AIMP1, ALOX12, AMOT, ANG, ANGPT1, ANGPT2, ANGPT4, ANGPTL3, ANGPTL4, ANGPTL6, ANPEP, ANXA2, ANXA3, APOD, APOH, APOLD1, AQP1, ARHGAP22, ARHGAP24, ATP5B, ATPIF1, B4GALT1, BAI1, BAI2, BAI3, BMP4, BMPER, BMPR2, BTG1, C1GALT1, C3, C3AR1, C5, C6, CALCRL, CASP8, CAV1, CCBE1, CCL11, CCL2, CCL24, CCR2, CCR3, CD34, CDC42, CDH13, CEACAM1, CHI3L1, CHRNA7, CLIC4, CMA1, COL15A1, COL18A1, COL4A1, COL4A2, COL4A3, COL8A1, COL8A2, CRHR2, CSPG4, CTGF, CTNNB1, CTSH, CX3CL1, CX3CR1, CXCL10, CXCL12, CXCL13, CXCL17, CXCR3, CXCR4, CYP1B1, CYR61, CYSLTR1, CYSLTR2, DAB2IP, DDAH1, DICER1, DLL4, E2F7, E2F8, ECM1, ECSCR, EDN1, EDNRA, EFNA1, EFNB2, EGF, EGFL7, EGLN1, EGR3, ELK3, ENG, ENPEP, EPAS1, EPGN, EPHA1, EPHA2, EPHB1, EPHB2, EPHB3, EPHB4, ERAP1, ERBB2, EREG, ESM1, ETS1, F3, FAM105B, FASLG, FGF1, FGF10, FGF18, FGF2, FGF6, FGF8, FGF9, FGFR1, FGFR2, FIGF, FLT1, FLT4, FN1, FOXC2, FOXO4, FOXS1, FZD5, FZD6, GATA2, GATA4, GATA6, GBX2, GDF2, GHRL, GJA5, GNA13, GPI, GPLD1, GPR124, GPR56, GPX1, GREM1, GTF2I, H3BM21, HAND1, HAND2, HDAC5, HDAC7, HDAC9, HEY1, HHEX, HIF1A, HIPK1, HIPK2, HMOX1, HOXA3, HOXA5, HOXA7, HOXB13, HOXB3, HPSE, HRG, HS6ST1, HSPB1, HSPG2, HTATIP2, HYAL1, ID1, IHH, IL17F, IL18, IL1A, IL1B, IL6, IL8, ISL1, ITGA5, ITGAV, ITGB1, ITGB1BP1, ITGB3, JAG1, JAM3, JMJD6, JUN, KDR, KLF4, KLF5, KLK3, KRIT1, KRT1, LAMA5, LECT1, LEF1, LOXL2, MAP2K5, MAP3K7, MAPK14, MAPK7, MCAM, MED1, MEG3, MEIS1, MEOX2, MFGE8, MMP14, MMP19, MMP2, MMRN2, MTDH, MYH9, NAA15, NCL, NF1, NFATC3, NFATC4, NODAL, NOS3, NOTCH1, NOTCH4, NOX1, NOX5, NPPB, NPR1, NR2E1, NR4A1, NRARP, NRCAM, NRP1, NRP2, NRXN1, NRXN3, NTRK1, NUS1, OVOL2, PARVA, PDCD10, PDCD6, PDE3B, PDGFA, PDGFRB, PDPN, PF4, PGF, PIK3CA, PIK3CG, PIK3R6, PITX2, PKNOX1, PLCD1, PLCD3, PLCG1, PLXDC1, PLXND1, PML, PNPLA6, POFUT1, PRKCA, PRKCB, PRKD1, PRKD2, PRKX, PROK1, PROK2, PROX1, PTEN, PTGIS, PTGS2, PTK2, PTK2B, PTPN14, PTPRB, PTPRM, RAMP1, RAMP2, RAPGEF3, RASIP1, RBM15, RBPJ, RGCC, RHOB, RNH1, ROBO1, ROBO4, ROCK1, ROCK2, RRAS, RTN4, RUNX1, S100A7, S1PR1, SAT1, SCG2, SEMA3E, SEMA4A, SEMA5A, SERPINE1, SERPINF1, SETD2, SFRP1, SFRP2, SH2D2A, SHB, SHC1, SHH, SIRT1, SLC12A6, SLIT2, SOX17, SOX18, SP100, SPHK1, SPINK5, SRF, SRPK2, SRPX2, STAB1, STAB2, STAT1, STK4, SULF1, SYK, TAL1, TBX1, TBX20, TBX4, TBXA2R, TDGF1, TEK, TGFA, TGFB2, TGFBI, TGFBR1, TGFBR2, THBS1, THBS2, THBS4, THSD7A, THY1, TIE1, TMEM100, TMPRSS6, TNFAIP2, TNFAIP3, TNFRSF12A, TNFRSF1A, TNFSF12, TNMD, TSPAN12, TWIST1, TYMP, UBP1, UTS2, UTS2R, VASH1, VASH2, VAV2, VAV3, VEGFA, VEGFB, VEGFC, VEZF1, VHLL, WARS, WASF2, WNT5A, ZC3H12A

Antioxidant Activity: ALB, ALOX5AP, APOA4, APOE, APOM, CAT, CCS, CLIC2, COX-2, CYGB, DUOX1, DUOX2, EPX, FABP1, FAM213A, GPX1, GPX2, GPX3, GPX4, GPX5, GPX6, GPX7, GPX8, GSR, GSTK1, GSTO1, GSTO2, GSTT1, GSTZ1, HBA1, HBB, HP, IPCEF1, IYD, LPO, LTC4S, MGST1, MGST2, MGST3, MPO, MT3, NQO1, NXN, PARK7, PRDX1, PRDX2, PRDX3, PRDX4, PRDX5, PRDX6, PTGS1, PTGS2, PXDN, PXDNL, S100A9, SEPW1, SOD1, SOD2, SOD3, SRXN1, TP53INP1, TPO, TXNDC17, TXNDC2, TXNRD1, TXNRD2, TXNRD3, UBIAD1, VIMP

Apoptotic Signaling: AAMDC, AARS, AATF, AATK, ABL1, ABR, ACAA2, ACER2, ACIN1, ACKR3, ACSL5, ACTC1, ACTN1, ACTN2, ACTN3, ACTN4, ACVR1, ACVR1B, ACVR1C, ADA, ADAM17, ADAM8, ADAMTS20, ADAMTSL4, ADAR, ADCY10, ADD1, ADIPOQ, ADNP, ADORA1, ADORA2A, ADRA1A, ADRB1, AEN, AES, AGAP2, AGT, AGTR2, AHI1, AHR, AIFM1, AIFM2, AIFM3, AIM2, AIMP1, AIMP2, AIPL1, AKAP13, AKR1C3, AKT1, AKT1S1, AKT2, AKTIP, ALB, ALDH1A2, ALDH1A3, ALDOC, ALK, ALKBH1, ALOX12, ALOX15, ALOX15B, ALX3, ALX4, AMBRA1, AMIGO2, ANGPT1, ANGPT4, ANGPTL4, ANKRD1, ANKRD13C, ANXA1, ANXA4, ANXA5, APAF1, APBB1, APBB2, APC, APH1A, APH1B, API5, APIP, APLP1, APOE, APOH, APOPT1, APP, APPL1, AQP1, AQP2, AR, ARAF, AREL1, ARF6, ARHGAP10, ARHGAP4, ARHGDIA, ARHGEF11, ARHGEF12, ARHGEF16, ARHGEF17, ARHGEF18, ARHGEF2, ARHGEF3, ARHGEF4, ARHGEF6, ARHGEF7, ARHGEF9, ARNT2, ARRB1, ARRB2, ASAH2, ASCL1, ASIC2, ASNS, ATAD3A, ATF2, ATF4, ATF5, ATG3, ATG4D, ATG5, ATG7, ATM, ATN1, ATOH1, ATP2A1, ATP7A, ATPIF1, AURKB, AVEN, AVP, AXIN1, AXL, AZU1, B4GALT1, BAD, BAG1, BAG3, BAG4, BAG6, BAK1, BARD1, BARHL1, BAX, BBC3, BCAP29, BCAP31, BCAR1, BCL10, BCL11B, BCL2, BCL2A1, BCL2L1, BCL2L10, BCL2L11, BCL2L12, BCL2L13, BCL2L14, BCL2L15, BCL2L2, BCL2L2-PABPN1, BCL3, BCL6, BCL7C, BCLAF1, BDKRB2, BDNF, BECN1, BEX2, BFAR, BID, BIK, BIRC2, BIRC3, BIRC5, BIRC6, BIRC7, BIRC8, BLCAP, BLID, BLOC1S2, BMF, BMP2, BMP4, BMP5, BMP7, BMX, BNIP1, BNIP2, BNIP3, BNIP3L, BNIPL, BOK, BOP, BRAF, BRCA1, BRCA2, BRE, BRMS1, BRSK2, BTC, BTG1, BTG2, BTK, BUB1, BUB1B, C11orf82, C1D, C1QBP, C3orf38, C5AR1, C6orf120, C8orf4, CAAP1, CACNA1A, CADM1, CALR, CAMK1D, CAMK2B, CAPN10, CAPN3, CARD10, CARD11, CARD14, CARD16, CARD17, CARD18, CARD6, CARD8, CARD9, CASP1, CASP10, CASP12, CASP14, CASP2, CASP3, CASP4, CASP5, CASP6, CASP7, CASP8, CASP8AP2, CASP9, CAST, CAT, CAV1, CBL, CBS, CBX4, CCAR1, CCAR2, CCK, CCL19, CCL2, CCL21, CCL3, CCL5, CCNB1IP1, CCNG1, CCR7, CD14, CD2, CD24, CD248, CD27, CD28, CD36, CD38, CD3E, CD3G, CD40, CD40LG, CD44, CD5, CD59, CD5L, CD70, CD74, CDC42, CDCA7, CDH1, CDIP1, CDK1, CDK11A, CDK11B, CDK4, CDK5, CDK5R1, CDKN1A, CDKN1B, CDKN2A, CDKN2D, CEBPB, CECR2, CERKL, CFDP1, CFL1, CFLAR, CGB, CHAC1, CHD8, CHEK2, CHIA, CHL1, CHMP3, CHST11, CIAPIN1, CIB1, CIDEA, CIDEB, CIDEC, CITED1, CITED2, CKAP2, CLC, CLCF1, CLEC5A, CLIP3, CLN3, CLN8, CLPTM1L, CLSPN, CLU, CNR1, CNTF, CNTFR, COL18A1, COL2A1, COL4A3, COMP, CPEB4, CRADD, CRH, CRIP1, CRLF1, CRYAA, CRYAB, CSE1L, CSF2, CSNK2A1, CSNK2A2, CSRNP1, CSRNP2, CSRNP3, CSTB, CTGF, CTH, CTLA4, CTNNA1, CTNNB1, CTNNBL1, CTSB, CTSH, CTSL, CUL1, CUL2, CUL3, CUL4A, CUL5, CX3CL1, CX3CR1, CXCL12, CXCR2, CXCR3, CXCR4, CYCS, CYFIP2, CYLD, CYR61, DAB2, DAB2IP, DAD1, DAP, DAP3, DAPK1, DAPK2, DAPK3, DAPL1, DAXX, DBH, DBNL, DCC, DCUN1D3, DDAH2, DDIT3, DDIT4, DDX20, DDX3X, DDX41, DDX47, DDX5, DEDD, DEDD2, DEPTOR, DFFA, DFFB, DFNA5, DHCR24, DHODH, DHRS2, DIABLO, DICER1, DIDO1, DKFZp686L0365, DKFZp781B1423, DLC1, DLG5, DLX1, DMPK, DNAJA3, DNAJB6, DNAJC10, DNAJC5, DNASE1, DNASE1L3, DNASE2, DNM1L, DNM2, DOCK1, DPEP1, DPF1, DPF2, DRAM1, DRAM2, DSG1, DSG2, DSG3, DSP, DUSP1, DUSP2, DUSP22, DUSP6, DYNAP, DYNLL1, DYNLL2, DYRK2, E2F1, E2F2, EAF2, EBAG9, ECT2, EDAR, EDN1, EDNRA, EDNRB, EEF1A2, EEF1E1, EFNA5, EGFR, EGLN2, EGLN3, EGR1, EGR2, EGR3, EGR4, EI24, EIF2AK3, EIF5A, EIF5AL1, ELL3, ELMO1, ELMO2, ELMO3, EMC4, ENDOG, EP300, EPB41L3, EPHA2, EPHA7, EPO, ERBB3, ERBB4, ERCC1, ERCC2, ERCC3, ERCC5, ERCC6, ERN1, ERN2, ERO1L, ESPL1, ESR1, ESR2, estrogen receptor, ETS1, ETV2, EVA1A, EYA1, EYA2, F2R, F3, FABP1, FADD, FAF1, FAIM, FAIM2, FAIM3, FAM129B, FAM162A, FAM188A, FAM215A, FAM32A, FAM3B, FAS, FASLG, FASTK, FCER1G, FEM1B, FGD1, FGD2, FGD3, FGD4, FGF10, FGF2, FGF4, FGF8, FGFR1, FGFR2, FGFR3, FHIT, FHL2, FIGNL1, FIS1, FKBP8, FKSG2, FLCN, FLT3, FLT4, FMN2, FNDC1, FNIP1, FNIP2, FNTA, FOSL1, FOXB1, FOXC1, FOXC2, FOXL2, FOXO1, FOXO3, FOXO4, FOXS1, FRZB, FXN, FXR1, FYN, FZD5, G0S2, G2E3, GABARAP, GADD45A, GADD45B, GADD45G, GAL, GAPDH, GAS1, GAS2, GAS6, GATA1, GATA3, GATA6, GCG, GCLC, GCLM, GCM2, GDF5, GDF6, GDNF, GFRAL, GGCT, GHITM, GHRL, GIMAP5, GJA1, GLI2, GLI3, GLO1, GLRX2, GLS2, GML, GNB1, GNB2L1, GNGT1, GNRH1, GPAM, GPER1, GPI, GPLD1, GPR65, GPX1, GRAMD4, GREM1, GRID2, GRIK2, GRIK5, GRIN1, GRIN2A, GRK1, GRK5, GRM4, GSDMA, GSK3A, GSK3B, GSN, GSTP1, GULP1, GZMA, GZMB, GZMH, GZMM, H1F0, H3BNH8, HAND2, HCAR2, HCK, HCLS1, HDAC1, HDAC2, HDAC3, HELLS, HERPUD1, HEY2, HGF, HIC1, HIF1A, HIGD1A, HIGD2A, HINT1, HINT2, HIP1, HIPK1, HIPK2, HIPK3, HK2, HMGA2, HMGB1, HMGB2, HMGCR, HMOX1, HNF1B, HNRNPK, HOXA13, HOXA5, HPN, HRAS, HRG, HRK, HSP90AA1, HSP90AB1, HSP90B1, HSPA1A, HSPA5, HSPA9, HSPB1, HSPD1, HSPE1, HTATIP2, HTR2B, HTRA2, HTT, HYAL2, IAPP, ID1, ID3, IDO1, IER3, IER3IP1, IFI16, IFI27, IFI6, IFIH1, IFIT2, IFIT3, IFNA2, IFNB1, IFNG, IFT57, IGBP1, IGF1, IGF1R, IGFBP3, IKBKB, IKBKE, IKBKG, IKZF3, IL10, IL12A, IL12B, IL17A, IL18, IL19, IL1A, IL1B, IL1RN, IL2, IL24, IL2RA, IL2RB, IL31RA, IL4, IL6, IL6R, IL6ST, IL7, ILK, INCA1, ING2, ING3, ING4, ING5, INHBA, INHBB, INPP5D, INS, INSL3, INSL6, INTS1, IP6K2, IRAK1, IRF1, IRF3, IRF5, IRF7, IRS2, ISL1, ITCH, ITGA1, ITGA5, ITGA6, ITGAV, ITGB1, ITGB2, ITGB3BP, ITM2B, ITM2C, ITPR1, ITPRIP, ITSN1, IVNS1ABP, JAG2, JAK2, JAK3, JMJD6, JMY, JTB, JUN, KALRN, KANK2, KAT2A, KCNIP3, KCNMA1, KDM1A, KDM2B, KDR, KIAA0141, KIAA1324, KIF1B, KITLG, KLF11, KLF4, KLHL20, KLLN, KMT2A, KNG1, KPNA1, KPNB1, KRAS, KRIT1, KRT18, KRT20, KRT8, LALBA, LAMTOR5, LCK, LCMT1, LCN2, LEF1, LEP, LGALS1, LGALS12, LGALS13, LGALS14, LGALS16, LGALS7, LGMN, LHX3, LHX4, LIG4, LILRB1, LITAF, LMNA, LMNB1, LPAR1, LRP1, LRP6, LTA, LTBR, LTK, LY86, LYN, MAD2L1, MADD, MAEA, MAEL, MAGED1, MAGEH1, MAGI3, MAL, MALT1, MAP1S, MAP2K4, MAP2K5, MAP2K6, MAP2K7, MAP3K1, MAP3K10, MAP3K11, MAP3K5, MAP3K7, MAP3K9, MAPK1, MAPK14, MAPK3, MAPK7, MAPK8, MAPK8IP1, MAPK9, MAPT, MARCO, MAX, MBD4, MCF2, MCF2L, MCL1, mdm2, MDM4, MECOM, MECP2, MED1, MEF2A, MEF2C, MEF2D, MEGF10, MEIS3, MELK, MEN1, MERTK, MFF, MFGE8, MFSD10, MGMT, MICAL1, MIEN1, MIF, MITF, MKL1, MKNK2, MLH1, MLLT11, MLTK, MMP9, MNAT1, MNDA, MNT, MOAP1, MPO, MPV17L, MRPL41, MRPS30, MSH2, MSH6, MST4, MSX1, MSX2, MT-RNR2, MT3, MTCH1, MTDH, MTFP1, MUC1, MUC2, MUL1, MUSK, MX1, MYBBP1A, MYC, MYD88, MYO18A, MYOCD, MZB1, NACA, NACC2, NAE1, NAIF1, NAIP, NANOS3, NBN, NCF2, NCKAP1, NCKAP1L, NCOA1, NCSTN, NDNF, NDUFA13, NDUFS1, NDUFS3, NEK6, NES, NET1, NEURL, NEUROD1, NF1, NFATC4, NFE2L2, NFKB1, NFKBIA, NFKBID, NGB, NGEF, NGF, NGFR, NGFRAP1, NISCH, NKX2-5, NKX2-6, NKX3-1, NKX3-2, NLRC4, NLRP1, NLRP12, NLRP2, NLRP3, NME1, NME2, NME3, NME5, NME6, NMT1, NOA1, NOC2L, NOD1, NOD2, NODAL, NOG, NOL3, NOS1AP, NOS3, NOTCH1, NOTCH2, NOX4, NOX5, NOXIN, NPM1, NQO1, NR1H3, NR2E1, NR3C1, NR4A1, NR4A2, NR4A3, NRAS, NRBP2, NRG1, NSMF, NTF3, NTN1, NTRK1, NTRK2, NUAK2, NUDT2, NUP62, NUPR1, OBSCN, OCLN, OGT, OPA1, OSM, OSR1, P2RX1, P2RX4, P2RX7, P4HB, PACS2, PAFAH2, PAK1, PAK2, PAK4, PAK6, PAK7, PALB2, PARK2, PARK7, PARL, PAWR, PAX2, PAX3, PAX7, PAX8, PCBP4, PCGF2, PCID2, PCNT, PCSK9, PDCD1, PDCD10, PDCD2, PDCD4, PDCD5, PDCD6, PDCD6IP, PDCD7, PDCL3, PDE1B, PDE3A, PDIA3, PDK1, PDK2, PDK4, PDPK1, PEA15, PEAR1, PEG10, PEG3, PERP, PF4, PGAP2, PHB, PHF17, PHIP, PHLDA1, PHLDA2, PHLDA3, PHLPP1, PIAS4, PIDD, PIGT, PIK3CA, PIK3CG, PIK3R1, PIM1, PIM2, PIM3, PINK1, PKN2, PKP1, PLA2G6, PLAC8, PLAGL1, PLEC, PLEKHF1, PLEKHG2, PLEKHG5, PLG, PLK1, PLK2, PLK3, PLK5, PLSCR1, PLSCR3, PMAIP1, PML, PNMA1, PNMA2, PNMA3, PNMA5, POLB, POLR2G, POR, POU3F3, POU3F4, POU4F1, POU4F3, PPARD, PPARG, PPARGC1A, PPID, PPIF, PPM1F, PPP1R13B, PPP1R13L, PPP1R15A, PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R2B, PPP2R4, PPP2R5C, PPP3CC, PPP3R1, PPT1, PRAME, PRAMEF1, PRAMEF10, PRAMEF11, PRAMEF12, PRAMEF13, PRAMEF14, PRAMEF15, PRAMEF16, PRAMEF17, PRAMEF18, PRAMEF19, PRAMEF2, PRAMEF20, PRAMEF22, PRAMEF23, PRAMEF24, PRAMEF25, PRAMEF3, PRAMEF4, PRAMEF5, PRAMEF6, PRAMEF7, PRAMEF8, PRDX2, PRDX3, PRDX5, PRELID1, PRF1, PRKAA1, PRKAA2, PRKCA, PRKCB, PRKCD, PRKCE, PRKCG, PRKCH, PRKCI, PRKCQ, PRKCZ, PRKD1, PRKDC, PRLR, PRMT2, PRNP, PROC, PRODH, PROK2, PROKR1, PROP1, PRUNE2, PSEN1, PSEN2, PSENEN, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB11, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME3, PSME4, PSMF1, PSMG2, PTCRA, PTEN, PTGFR, PTGIS, PTGS2, PTH, PTK2, PTK2B, PTPN6, PTPRC, PTPRH, PTRH2, PUF60, PYCARD, QRICH1, RABEP1, RAC1, RAD21, RAD9A, RAF1, RAG1, RALB, RALDH2, RAMP2, RAPGEF2, RAPSN, RARA, RARB, RARG, RASA1, RASGRF2, RASSF5, RASSF6, RASSF7, RB1, RB1CC1, RBCK1, RBM10, RBM25, RBM5, RELA, REST, RET, RFFL, RGCC, RHBDD1, RHOA, RHOB, RHOT1, RHOT2, RIPK1, RIPK2, RIPK3, RMDN3, RNF130, RNF144B, RNF152, RNF216, RNF34, RNF41, RNF7, RNPS1, ROBO1, ROBO2, ROCK1, RPS27A, RPS27L, RPS3, RPS3A, RPS6, RPS6KA1, RPS6KA2, RPS6KA3, RPS6KB1, RRAGA, RRAGC, RRM2B, RRN3, RRP8, RTKN, RTN3, RTN4, RXFP2, RYBP, RYR2, S100A14, S100A8, S100A9, S100B, SAP18, SAP30BP, SART1, SATB1, SAV1, SCARB1, SCG2, SCIN, SCN2A, SCRIB, SCRT2, SCT, SCXA, SDIM1, SEMA3A, SEMA4D, SEMA5A, SEMA6A, SENP1, SEPT4, SERPINB10, SERPINB2, SERPINB9, SERPINE1, SET, SFN, SFRP1, SFRP2, SFRP4, SFRP5, SGK1, SGMS1, SGPL1, SGPP1, SH3GLB1, SH3KBP1, SH3RF1, SHARPIN, SHB, SHF, SHH, SHISA5, SHQ1, SIAH1, SIAH2, SIGMAR1, SIK1, SIN3A, SIRT1, SIVA1, SIX1, SIX4, SKI, SKIL, SLC11A2, SLC25A27, SLC25A4, SLC25A5, SLC25A6, SLC35F6, SLC40A1, SLC46A2, SLC5A11, SLC5A8, SLC9A3R1, SLIT2, SLIT3, SLK, SLTM, SMAD3, SMAD6, SMNDC1, SMO, SMPD1, SMPD2, SNAI1, SNAI2, SNCA, SNCB, SNW1, SOCS2, SOCS3, SOD1, SOD2, SON, SORT1, SOS1, SOS2, SOX10, SOX2, SOX4, SOX7, SOX8, SOX9, SP100, SPATA5L1, SPDEF, SPHK1, SPHK2, SPIN2B, SPN, SPRY2, SPTAN1, SQSTM1, SRA1, SRC, SRGN, SRPK2, SRPX, SSBP3, SST, SSTR3, ST20, STAMBP, STAT1, STAT5A, STAT5B, STEAP3, STIL, STK11, STK17A, STK17B, STK24, STK25, STK3, STK4, STPG1, STRADB, STXBP1, SUDS3, SULF1, SUPV3L1, SYCE3, SYCP2, SYNGAP1, SYVN1, TAF9, TAF9B, TAOK1, TAOK2, TATDN1, TAX1BP1, TBX1, TBX3, TBX5, TCF7, TCF7L2, TCHP, TCTN3, TDGF1, TEK, TERF1, TERT, TEX11, TFAP2A, TFAP2B, TFAP2D, TFAP4, TFPT, TGFA, TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, TGM2, THBS1, THEM4, THOC1, THOC6, THRA, TIA1, TIAF1, TIAL1, TIAM1, TIAM2, TICAM1, TIGAR, TIMM50, TIMP1, TJP1, TJP2, TLE1, TLR2, TLR3, TM2D1, TMBIM4, TMBIM6, TMEM102, TMEM109, TMEM161A, TMEM173, TMEM214, TMEM219, TMEM23, TNF, TNFAIP1, TNFAIP3, TNFAIP8, TNFRSF10A, TNFRSF10B, TNFRSF10C, TNFRSF10D, TNFRSF11B, TNFRSF12A, TNFRSF18, TNFRSF19, TNFRSF1A, TNFRSF1B, TNFRSF21, TNFRSF25, TNFRSF4, TNFRSF6B, TNFRSF8, TNFRSF9, TNFSF10, TNFSF12, TNFSF14, TNFSF15, TNFSF18, TNFSF8, TNFSF9, TNIP2, TNS4, TOP2A, TOPORS, TOX3, TP53, TP53AIP1, TP53BP2, TP53I3, TP53INP1, TP63, TP73, TPD52L1, TPT1, TPX2, TRADD, TRAF1, TRAF2, TRAF3, TRAF3IP2, TRAF4, TRAF5, TRAF6, TRAF7, TRAIP, TRIAP1, TRIB3, TRIM2, TRIM24, TRIM32, TRIM35, TRIM39, TRIM69, TRIO, TSC22D1, TSC22D3, TSPO, TWIST1, TWIST2, TXNDC5, TXNIP, TYRO3, UACA, UBA52, UBB, UBC, UBD, UBE2B, UBE2D3, UBE2M, UBE2V2, UBE2Z, UBE4B, UBQLN1, UCN, UNC13B, UNC5A, UNC5B, UNC5C, UNC5D, URI1, USP17L1P, USP17L2, USP17L24, USP17L3, USP17L5, USP28, USP47, UTP11L, VAV1, VAV2, VAV3, VCP, VDAC1, VDAC2, VDR, VEGFA, VEGFB, VHL, VIL1, VIM, VIMP, VIP, VNN1, WDR92, WFS1, WNK3, WNT1, WNT10B, WNT11, WNT3A, WNT4, WNT5A, WNT7A, WNT9A, WRN, WT1, WWOX, XAF1, XDH, XIAP, XKR8, XPA, XRCC2, XRCC4, XRCC5, YAP1, YARS, YBX3, YWHAB, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ, ZBTB16, ZC3H12A, ZC3H8, ZC3HC1, ZDHHC16, ZFAND6, ZGLP1, ZMAT3, ZMYND11, ZNF16, ZNF205, ZNF268, ZNF346, ZNF385A, ZNF385B, ZNF443, ZNF622, ZSWIM2

Complement Activation: A2M, C1QA, C1QB, C1QBP, C1QC, C1R, C1RL, C1S, C2, C3, C3AR1, C4A, C4B, C4BPA, C4BPB, C4B_2, C5, C5AR1, C6, C7, C8A, C8B, C8G, C9, CALR, CD46, CD55, CD59, CD93, CFB, CFD, CFH, CFHR1, CFHR2, CFHR3, CFHR4, CFHR5, CFI, CFP, CLU, CR1, CR2, CRP, FCN1, FCN2, FCN3, GPLD1, IGHG1, IGHG2, IGHG3, IGHG4, IGKC, IGKV1-5, IGKV4-1, IGLC1, IGLC2, IGLC3, IGLC6, IGLC7, KRT1, MASP1, MASP2, MBL2, P01593, P01594, P01595, P01596, P01597, P01598, P01599, P01600, P01601, P01603, P01604, P01605, P01606, P01607, P01608, P01609, P01610, P01611, P01612, P01613, P01614, P01615, P01616, P01617, P01619, P01620, P01621, P01622, P01623, P01624, P01625, P01699, P01700, P01701, P01702, P01703, P01704, P01705, P01706, P01707, P01708, P01709, P01710, P01711, P01712, P01713, P01714, P01715, P01716, P01717, P01718, P01719, P01720, P01721, P01722, P01742, P01743, P01744, P01760, P01761, P01762, P01763, P01764, P01765, P01766, P01767, P01768, P01769, P01770, P01771, P01772, P01773, P01774, P01775, P01776, P01777, P01778, P01779, P01780, P01781, P01782, P01814, P01815, P01816, P01817, P01818, P01824, P01825, P04206, P04207, P04208, P04209, P04211, P04430, P04431, P04432, P04433, P04434, P04438, P06309, P06310, P06311, P06313, P06314, P06315, P06316, P06317, P06318, P06319, P06326, P06331, P06887, P06888, P06889, P18135, P18136, P80362, P80748, P83593, PROS1, RGCC, SERPING1, VSIG4, VTN

Inflammatory Response: A2M, ABCF1, ABR, ACE2, ACKR2, ACP5, ACVR1, ADA, ADAM8, ADCYAP1, ADIPOQ, ADORA1, ADORA2A, ADORA2B, ADORA3, ADRA2A, AFAP1L2, AGER, AGT, AGTR1, AGTR2, AHSG, AIF1, AIM2, AIMP1, AK7, AKT1, ALOX15, ALOX5, ALOX5AP, ANKRD42, ANO6, ANXA1, AOAH, AOC3, AOX1, APCS, APOA1, APOA2, APOC3, APOD, APOE, APOL2, APOL3, ATRN, AXL, AZU1, B4GALT1, BCL6, BCR, BDKRB1, BDKRB2, BDNF, BIRC2, BIRC3, BLNK, BMP2, BMP6, BMPR1B, BRD4, C1QBP, C1QTNF3, C2, C3, C3AR1, C4A, C4B, C4BPA, C4BPB, C4B_2, C5, C5AR1, C6, C7, C8A, C8B, C8G, C9, CALCA, CALCRL, CAMK1D, CAMK4, CARD18, CCL11, CCL13, CCL16, CCL17, CCL18, CCL19, CCL2, CCL20, CCL21, CCL22, CCL23, CCL24, CCL25, CCL26, CCL3, CCL3L1, CCL4, CCL4L1, CCL5, CCL7, CCL8, CCR1, CCR2, CCR3, CCR4, CCR5, CCR7, CCRL2, CD14, CD163, CD180, CD276, CD28, CD40, CD40LG, CD44, CD46, CD47, CD55, CD59, CD97, CDKN2A, CDO1, CEBPA, CEBPB, CELA1, CFB, CFH, CFI, CFP, CHI3L1, CHIA, CHST1, CHST2, CHST4, CHUK, CLEC7A, CMA1, CNR1, CNR2, CR1, CRH, CRHBP, CRP, CSF1, CSF1R, CTNNBIP1, CTSL, CX3CL1, CXCL1, CXCL10, CXCL11, CXCL13, CXCL2, CXCL3, CXCL6, CXCL9, CXCR1, CXCR2, CXCR3, CXCR4, CXCR6, CYBA, CYBB, CYP4F11, CYSLTR1, DAB2IP, DARC, DEFB1, DUSP10, ECM1, EDNRA, ELANE, ELF3, EMR2, EPHX2, EVI1, F11R, F12, F2, F2R, F2RL1, F3, F8, FABP4, FAM105B, FAM132A, FAS, FASLG, FCER1A, FCER1G, FEM1A, FFAR4, FN1, FOS, FOXF1, FOXP3, FPR2, GAL, GATA3, GBA, GGT1, GGT5, GHRL, GHSR, GPER1, GPR68, GPX1, GPX2, GSTP1, H0Y858, HCAR2, HCK, HDAC4, HDAC5, HDAC9, HIF1A, HIST1H2BA, HLA-DRB1, HMGB1, HMOX1, HNRNPA0, HP, HRH1, HRH4, HYAL1, HYAL2, HYAL3, IDO1, IER3, IFI16, IFNA2, IFNG, IGFBP4, IKBKB, IKBKG, IL10, IL10RB, IL12B, IL13, IL15, IL17A, IL17B, IL17C, IL17D, IL17F, IL17RA, IL17RC, IL17RE, IL18, IL18RAP, IL1A, IL1B, IL1F10, IL1RAP, IL1RL1, IL1RN, IL2, IL20, IL20RB, IL21, IL22, IL23A, IL23R, IL25, IL27, IL2RA, IL33, IL34, IL36A, IL36B, IL36G, IL37, IL4, IL4R, IL5, IL5RA, IL6, IL6R, IL6ST, IL8, IL9, INS, IRAK2, IRF3, IRF7, IRG1, IRGM, ISL1, ITCH, ITGAL, ITGB2, ITGB6, ITIH4, JAK2, JAM3, KCNJ10, KDM6B, KIT, KL, KLF4, KLKB1, KLRG1, KNG1, LAT, LBP, LIAS, LIPA, LRRC32, LTA, LTA4H, LTB4R, LXN, LY75, LY86, LY96, LYN, LYZ, MAP2K3, MAPK13, MAPK7, MAPKAPK2, MAS1, MASP1, MBL2, MECOM, MEF2A, MEF2C, MEFV, MEP1B, MGLL, MIF, MMP25, MRGPRX1, MS4A2, MVK, MYD88, MYLK3, NAIP, NCF1, NCR3, NDFIP1, NDST1, NFAM1, NFATC3, NFATC4, NFE2L1, NFKB1, NFKBID, NFKBIZ, NFRKB, NFX1, NGF, NLRC4, NLRP1, NLRP12, NLRP3, NLRP6, NMI, NOD1, NOD2, NOS2, NOTCH1, NOX1, NOX4, NPFF, NPY5R, NR1D2, NR1H3, NT5E, NUPR1, OGG1, OLR1, ORM1, ORM2, OSM, OSMR, P2RX1, P2RX7, PARK7, PARP4, PDE2A, PDPN, PIK3AP1, PIK3CD, PIK3CG, PLA2G2A, PLA2G2D, PLA2G2E, PLA2G4B, PLA2G4C, PLA2G7, PLAA, PLGRKT, PLSCR1, PNMA1, PPARG, PRDX5, PRKCA, PRKCD, PRKCQ, PRKCZ, PRKD1, PROK2, PROS1, PSMA1, PSMA6, PSMB4, PTAFR, PTGDR, PTGER3, PTGER4, PTGES, PTGIS, PTGS2, PTPN2, PTX3, PXK, PYCARD, RAC1, RASGRP1, RBPJ, REG3A, REG3G, RELA, RIPK2, RPS19, RPS6KA4, RPS6KA5, S100A12, S100A8, S100A9, S1PR3, SAA1, SAA2, SAA4, SAAL1, SBNO2, SCG2, SCGB1A1, SCN9A, SCUBE1, SELE, SELP, SEMA7A, SERPINA1, SERPINA3, SERPINC1, SERPINE1, SERPINF2, SERPING1, SETD6, SGMS1, SHARPIN, SHPK, SIGIRR, SIGLEC1, SIRT1, SLC11A1, SLC7A2, SMAD1, SMAD3, SPHK1, SPN, SPP1, STAB1, STAT3, STAT5A, STAT5B, STK39, SYK, TAC1, TACR1, TBC1D23, TBK1, TBXA2R, TEK, TFF2, TGFB1, TGM2, THBS1, THEMIS2, TICAM1, TICAM2, TIRAP, TLR1, TLR10, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TMEM23, TNF, TNFAIP3, TNFAIP6, TNFAIP8L2, TNFRSF11A, TNFRSF1A, TNFRSF1B, TNFRSF4, TNFSF11, TNFSF4, TNIP1, TNIP2, TNIP3, TOLLIP, TP73, TPST1, TRIL, TSC2, TUSC2, TYRO3, UACA, UCN, UGT1A1, UNC13D, VCAM1, VIMP, VNN1, VTN, WNT5A, XCL1, XCR1, XIAP, YWHAZ, ZFP36, ZP3

Response to Nicotine: ABAT, ATP1A2, AVP, BAD, BCL2, CHRNA3, CHRNA4, CHRNA5, CHRNA7, CHRNB1, CHRNB2, CHRNB4, CNR1, DRD2, EDN1, GNAT3, HMOX1, HOMER1, IL13, KCNK1, NGF, NKX6-1, NTRK1, PDX1, PPARA, SLC6A3, SLC7A11, STAR, TACR1, TH, TNF

Oxidative Phosphorylation: AK2, ATP5C1, ATP5D, ATP7A, BDNF, CHCHD10, COQ7, COX10, COX15, DLD, FXN, GBAS, MECP2, MLXIPL, MSH2, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6, NDUFA1, NDUFA10, NDUFA2, NDUFA3, NDUFA4, NDUFA5, NDUFA6, NDUFA7, NDUFA8, NDUFA9, NDUFAB1, NDUFAF1, NDUFB1, NDUFB10, NDUFB2, NDUFB3, NDUFB4, NDUFB5, NDUFB6, NDUFB7, NDUFB8, NDUFB9, NDUFC1, NDUFC2, NDUFC2-KCTD14, NDUFS1, NDUFS2, NDUFS3, NDUFS4, NDUFS5, NDUFS6, NDUFS7, NDUFS8, NDUFV1, NDUFV2, NDUFV3, PARK7, PPIF, SDHAF2, SNCA, SURF1, TAZ, TEFM, TWIST1, UCP1, UCP3, UQCC2, UQCR10, UQCRB, UQCRC1, UQCRC2, UQCRH

Tricarboxylic Acid Cycle: ACO1, ACO2, BCKDHA, BCKDHB, BCKDK, CS, DBT, DHTKD1, DLAT, DLD, DLST, FH, IDH1, IDH2, IDH3A, IDH3B, IDH3G, MDH1, MDH1B, MDH2, NNT, OGDH, OGDHL, PDHA1, PDHA2, PDHB, SDHA, SDHB, SDHC, SDHD, SUCLA2, SUCLG1, SUCLG2

Appendix D. Simulation workflow data structuring.

Suppose: i is the number of SNPs (or SNP pairs), j is the number of individuals, k is the number of simulation replicates, and h^2 is the heritability estimate squared. Also suppose that:

$$\vec{\boldsymbol{M}}=\begin{bmatrix} m_1 & m_2 & \dots & m_i \end{bmatrix} \text{ where } m_i \text{ is the } i^{th} \text{ SNP}$$

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ p_{21} & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} \end{bmatrix} \text{ where } p_{ij} \text{ is the genotype for the } i^{th} \text{ SNP of the } j^{th} \text{ individual}$$

$$\vec{\boldsymbol{V}}_{Model}=\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_i \end{bmatrix} \qquad \boldsymbol{C}_{Model}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} \\ c_{21} & c_{22} & \cdots & c_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} \end{bmatrix}$$

$$where\, Model=\{one\, of\, Cockerham\, W_1 \dots W_8\},\ v_i=Var_{Model}(m_i),\ and\ c_{ij}=Cov_{Model}(p_{ij})$$

Calculate w_{ij} for each i and j:

$$\boldsymbol{W}=\begin{bmatrix} \frac{c_{11}}{\sqrt{v_1}} & \frac{c_{12}}{\sqrt{v_1}} & \cdots & \frac{c_{1j}}{\sqrt{v_1}} \\ \frac{c_{21}}{\sqrt{v_2}} & \frac{c_{22}}{\sqrt{v_2}} & \cdots & \frac{c_{2j}}{\sqrt{v_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{c_{i1}}{\sqrt{v_i}} & \frac{c_{i2}}{\sqrt{v_i}} & \cdots & \frac{c_{ij}}{\sqrt{v_i}} \end{bmatrix}=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1j} \\ w_{21} & w_{22} & \cdots & w_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} \end{bmatrix}$$

Proceed with simulation for k replicates:

$$\forall k:\ \vec{\mathrm{B}}=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \end{bmatrix}\ where\ \beta_i \sim N(\mu,\sigma^2) \text{ with } \mu=0 \text{ and } \sigma^2=\begin{cases} Var(W_{n*}) \\ or \\ 1 \end{cases}$$

$$\forall k:\ \vec{\boldsymbol{S}}=[s_1 \quad s_2 \quad \cdots \quad s_j]=\begin{bmatrix} \sum_{n=1}^{i} w_{n1}\beta_n & \sum_{n=1}^{i} w_{n2}\beta_n & \cdots & \sum_{n=1}^{i} w_{nj}\beta_n \end{bmatrix}$$

$$\forall k:\ \vec{\mathrm{E}}=[\epsilon_1 \quad \epsilon_2 \quad \cdots \quad \epsilon_j]\ where\ \epsilon_j \sim \boldsymbol{N}(\mu,\sigma^2)\ with\ \mu=0 \text{ and } \sigma^2=Var(\vec{\boldsymbol{S}})\times[\tfrac{1}{h^2}-1]$$

Finalize simulated phenotypes for j individuals and k replicates:

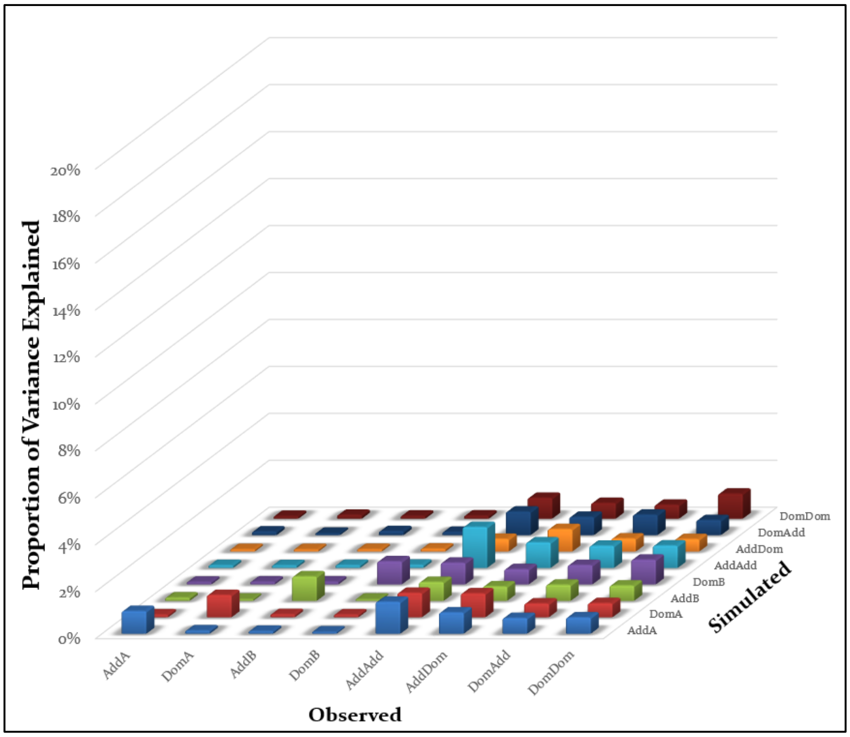
$$\mathrm{Y}=\begin{bmatrix} \overbrace{\qquad}^{1} & \overbrace{\qquad}^{2} & & \overbrace{\qquad}^{k} \\ s_1+\epsilon_1 & s_1+\epsilon_1 & \cdots & s_1+\epsilon_1 \\ s_2+\epsilon_2 & s_2+\epsilon_2 & \cdots & s_2+\epsilon_2 \\ \vdots & \vdots & \ddots & \vdots \\ s_j+\epsilon_j & s_j+\epsilon_j & \cdots & s_j+\epsilon_j \end{bmatrix}=\begin{bmatrix} \overbrace{}^{1} & \overbrace{}^{2} & & \overbrace{}^{k} \\ y_1 & y_1 & \cdots & y_1 \\ y_2 & y_2 & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ y_j & y_j & \cdots & y_j \end{bmatrix}$$

Appendix E. Standard error of the mean for simulations.

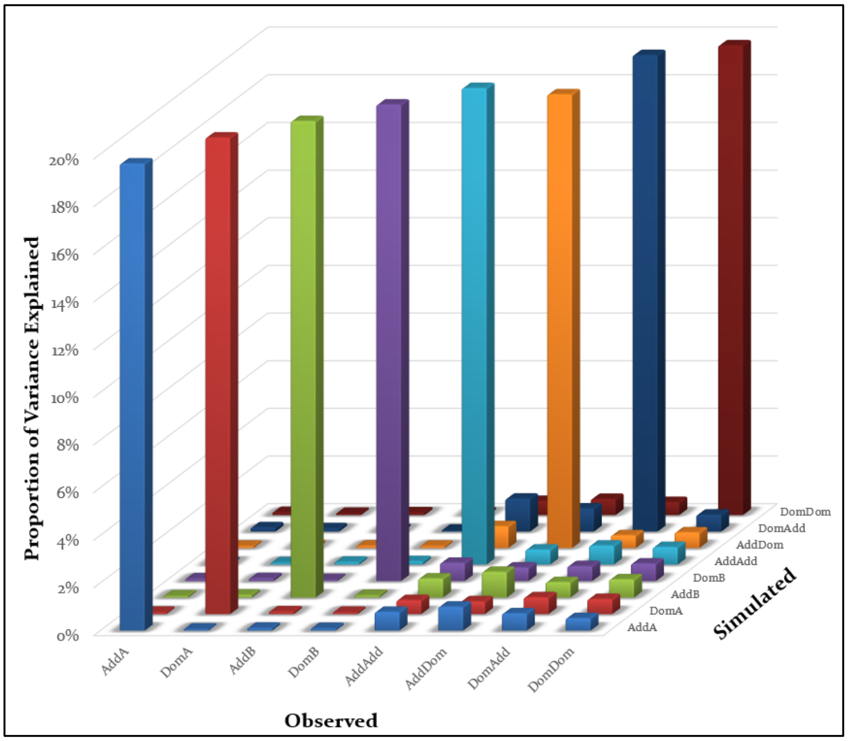
	Component	Add A	Dom A	Add B	Dom B	A × A	A × D	D × A	D × D
1% Simulated	Add A	0.0010	0.0003	0.0002	0.0001	0.0037	0.0014	0.0011	0.0015
	Dom A	0.0001	0.0007	0.0004	0.0002	0.0032	0.0029	0.0008	0.0007
	Add B	0.0003	0.0001	0.0010	0.0001	0.0013	0.0014	0.0016	0.0014
	Dom B	0.0003	0.0002	0.0002	0.0011	0.0015	0.0005	0.0010	0.0026
	A × A	0.0002	0.0003	0.0002	0.0003	0.0017	0.0033	0.0016	0.0035
	A × D	0.0001	0.0001	0.0002	0.0002	0.0005	0.0012	0.0008	0.0006
	D × A	0.0003	0.0001	0.0002	0.0003	0.0013	0.0012	0.0010	0.0014
	D × D	0.0002	0.0003	0.0002	0.0002	0.0009	0.0013	0.0008	0.0039
5% Simulated	Add A	0.0015	0.0001	0.0005	0.0003	0.0014	0.0023	0.0021	0.0035
	Dom A	0.0004	0.0018	0.0002	0.0002	0.0011	0.0039	0.0005	0.0018
	Add B	0.0002	0.0001	0.0029	0.0002	0.0021	0.0011	0.0008	0.0006
	Dom B	0.0005	0.0001	0.0003	0.0015	0.0030	0.0021	0.0023	0.0020
	A × A	0.0004	0.0003	0.0001	0.0003	0.0040	0.0017	0.0010	0.0002
	A × D	0.0002	0.0002	0.0002	0.0001	0.0016	0.0084	0.0020	0.0020
	D × A	0.0001	0.0001	0.0001	0.0001	0.0010	0.0035	0.0046	0.0015
	D × D	0.0002	0.0002	0.0003	0.0000	0.0046	0.0064	0.0010	0.0063
10% Simulated	Add A	0.0014	0.0001	0.0002	0.0001	0.0048	0.0013	0.0008	0.0013
	Dom A	0.0003	0.0026	0.0002	0.0002	0.0021	0.0017	0.0007	0.0010
	Add B	0.0001	0.0002	0.0017	0.0002	0.0019	0.0011	0.0016	0.0007
	Dom B	0.0002	0.0002	0.0001	0.0027	0.0015	0.0017	0.0048	0.0015
	A × A	0.0001	0.0001	0.0001	0.0002	0.0035	0.0023	0.0034	0.0009
	A × D	0.0001	0.0001	0.0006	0.0003	0.0037	0.0074	0.0002	0.0034
	D × A	0.0002	0.0001	0.0001	0.0002	0.0044	0.0032	0.0077	0.0017
	D × D	0.0003	0.0004	0.0001	0.0004	0.0012	0.0044	0.0015	0.0071
20% Simulated	Add A	0.0026	0.0002	0.0003	0.0002	0.0013	0.0021	0.0021	0.0005
	Dom A	0.0002	0.0032	0.0002	0.0001	0.0008	0.0012	0.0019	0.0016
	Add B	0.0001	0.0002	0.0017	0.0003	0.0021	0.0028	0.0012	0.0009
	Dom B	0.0001	0.0002	0.0001	0.0013	0.0013	0.0012	0.0009	0.0008
	A × A	0.0001	0.0002	0.0003	0.0006	0.0050	0.0008	0.0015	0.0023
	A × D	0.0001	0.0003	0.0002	0.0001	0.0026	0.0038	0.0006	0.0009
	D × A	0.0005	0.0005	0.0002	0.0000	0.0028	0.0015	0.0023	0.0020
	D × D	0.0002	0.0001	0.0003	0.0003	0.0015	0.0011	0.0004	0.0054

Colored ranging from yellow to green, with bright green representing the smallest values and bright red representing the largest values.

Appendix F. Observed variance explained for 1% and 20% heritability.



Observed variance explained for 1% simulated heritability.



Observed variance explained for 20% simulated heritability.

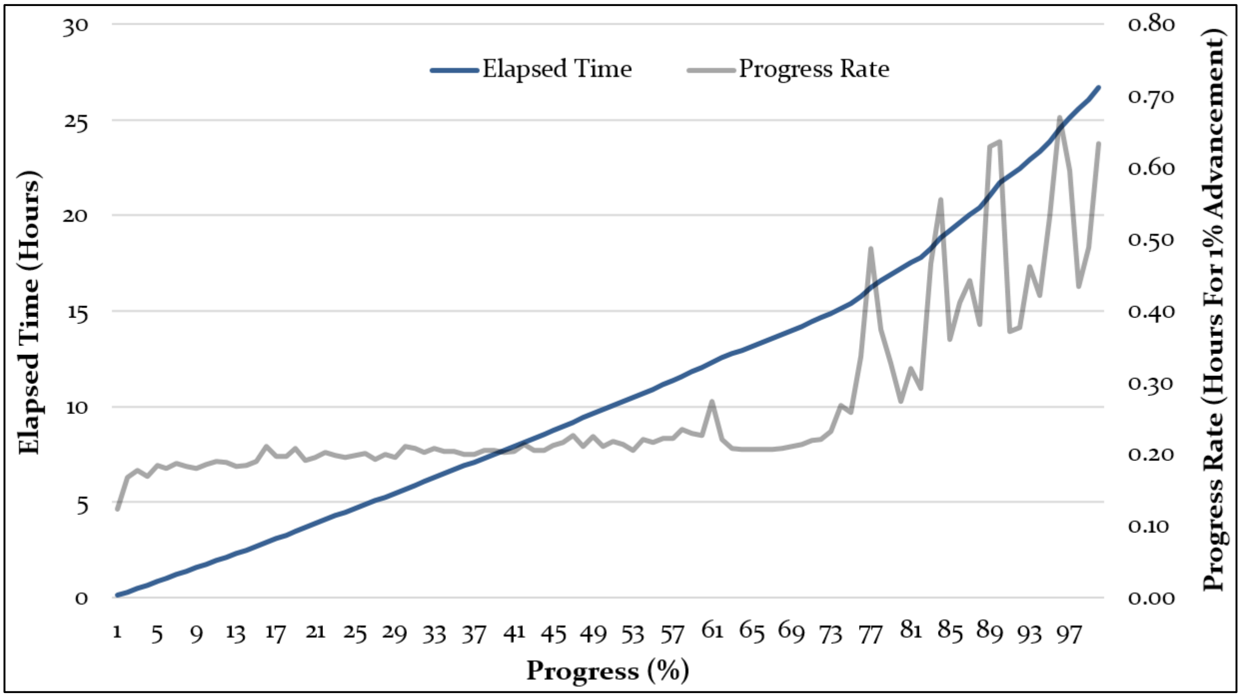
Both 3-D bar charts are scaled to 20% on the Y-axis, for comparison. The first figure shows ~1% PVE on the diagonal and the second figure shows ~20% — levels corresponding with the amount they were simulated to have. As the simulated effect size increases the "background noise" from null effects diminishes in comparison, making it easier to differentiate between real and spurious effects. As a note, results shown here are for a relatively small number of SNPs (100 per group, with 50 per group used for simulating effects). Based on other analyses we have performed (not shown), as the number of SNPs increases the amount of background noise correspondingly decreases.

Appendix G. Significance calculations for simulated effects.

Simulated Heritability	Observed PVE	Simulated Component	Excluded Component	Iteration	Likelihood (Full)	Likelihood (Reduced)	LRT	p-value
1%	0.51%	Add A	Add A	926	47,104.95	47,109.44	8.98	0.0014
	0.27%	Add A	Dom A	883	47,106.19	47,106.22	0.06	0.4032
	2.29%	A × A	A × A	1,404	67,150.15	67,151.47	2.64	0.0521
	0.44%	A × A	A × D	1,451	67,149.70	67,148.11	3.18	0.0373
5%	4.95%	Add A	Add A	890	37,792.53	37,945.71	306.36	6.8 e-69
	0.16%	Add A	Dom A	1,969	37,784.13	37,784.03	0.20	0.3274
	4.65%	A × A	A × A	1,737	58,607.59	58,612.81	10.44	0.0006
	0.93%	A × A	A × D	1,709	58,607.79	58,607.64	0.30	0.2919
10%	10.17%	Add A	Add A	1,923	35,074.42	35,468.25	787.66	1.3 e-173
	0.05%	Add A	Dom A	1,583	35,076.34	35,075.63	1.42	0.1184
	10.73%	A × A	A × A	1,679	55,458.43	55,486.49	56.12	3.4 e-14
	1.97%	A × A	A × D	1,673	55,458.47	55,459.50	2.06	0.0756
20%	20.42%	Add A	Add A	1,587	30,327.05	31,242.24	1830.38	< 1 e-300
	0.15%	Add A	Dom A	1,418	30,328.08	30,327.57	1.02	0.1563
	19.34%	A × A	A × A	1,873	52,211.18	52,298.55	174.74	3.4 e-40
	0.99%	A × A	A × D	1,680	52,212.35	52,212.08	0.54	0.2312

Data shown here is from the last replicate performed for each level of heritability. Based on results shown in Table 15 we selected representative single SNP and interaction simulations to test for significance, including one on-diagonal and one off-diagonal component from each. On-diagonals represent the significance of the simulated component, while off-diagonals represent the significance from a non-simulated component. After manually running REML for reduced models, excluding given components, we manually looked up likelihoods from the original (full) model and the reduced model, matching to use estimates from the same REML iteration. P-values were calculated using a Chi-square distribution from the calculated LRT test statistic.

Appendix H. Example iGRM calculation time and progress rate.



The graph here depicts data for the ARMS2-complement GRM set, however both other tested pathway sets exhibited similar elapsed time and progress rates.

Each of the three *ARMS*2-pathway GRM calculations required about the same amount of compute time and exhibited similar progress rate characteristics. Each of the three GRM set calculations required just over 24 hours, using 100 processors and had a decreased progress rate after about 75% progress. In this example, the decreased progress rate is not inhibitory but attempting the same analysis using the full dataset of 33,603 individuals greatly increases the number of pairs of individuals being assessed (~62.7 million pairs for 11,201 individuals versus ~564.6 million pairs for 33,603 individuals). Not only is compute time for GRM calculations increased, but REML analysis becomes increasingly difficult — so much so that it becomes the limiting step.

References

1. Hall JB, Cooke Bailey JN, Hoffman JD, Pericak-Vance MA, Scott WK, Kovach JL, Schwartz SG, Agarwal A, Brantley MA, Haines JL, Bush WS: **Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models.** *BMC Bioinformatics* 2015, **16**:329.

2. Gehrs KM, Anderson DH, Johnson L V, Hageman GS: **Age-related macular degeneration--emerging pathogenetic and therapeutic concepts.** *Ann Med* 2006, **38**:450–71.

3. Friedman DS, O'Colmain BJ, Muñoz B, Tomany SC, McCarty C, de Jong PTVM, Nemesure B, Mitchell P, Kempen J: **Prevalence of age-related macular degeneration in the United States.** *Arch Ophthalmol* 2004, **122**:564–72.

4. Klein R, Chou C-F, Klein BEK, Zhang X, Meuer SM, Saaddine JB: **Prevalence of age-related macular degeneration in the US population.** *Arch Ophthalmol* 2011, **129**:75–80.

5. **The Global Economic Cost of Visual Impairment** [http://www.icoph.org/dynamic/attachments/resources/globalcostofvi_finalreport.pdf]

6. **Eye disease simulations.** [https://nei.nih.gov/health/examples]

7. Blausen.com staff: **Blausen gallery 2014**. *Wikiversity J Med* , **1**.

8. Brody BL, Gamst AC, Williams RA, Smith AR, Lau PW, Dolnak D, Rapaport MH, Kaplan RM, Brown SI: **Depression, visual acuity, comorbidity, and disability associated with age-related macular degeneration**. *Ophthalmology* 2001, **108**:1893–1900.

9. Sarks JP, Sarks SH, Killingsworth MC: **Evolution of soft drusen in age-related macular degeneration.** *Eye (Lond)* 1994, **8 (Pt 3)**:269–83.

10. Fine SL, Berger JW, Maguire MG, Ho AC: **Age-related macular degeneration.** *N Engl J Med* 2000, **342**:483–92.

11. Klein R, Klein BE, Linton KL: **Prevalence of age-related maculopathy. The Beaver Dam Eye Study.** *Ophthalmology* 1992, **99**:933–43.

12. **Flickr Photostream** [https://www.flickr.com/photos/nationaleyeinstitute/]

13. Seddon JM, Ajani UA, Mitchell BD: **Familial aggregation of age-related**

maculopathy. *Am J Ophthalmol* 1997, **123**:199–206.

14. Hyman L, Schachat AP, He Q, Leske MC: **Hypertension, cardiovascular disease, and age-related macular degeneration. Age-Related Macular Degeneration Risk Factors Study Group.** *Arch Ophthalmol* 2000, **118**:351–8.

15. Krause L, Yousif T, Pohl K: **An epidemiological study of neovascular age-related macular degeneration in Germany.** *Curr Med Res Opin* 2013.

16. **Amsler grid for AMD.**
[http://www.nei.nih.gov/health/maculardegen/webAMD.pdf]

17. Zeng S, Hernández J, Mullins RF: **Effects of antioxidant components of AREDS vitamins and zinc ions on endothelial cell activation: implications for macular degeneration.** *Invest Ophthalmol Vis Sci* 2012, **53**:1041–7.

18. Chew EY, Sangiovanni JP, Ferris FL, Wong WT, Agron E, Clemons TE, Sperduto R, Danis R, Chandra SR, Blodi BA, Domalpally A, Elman MJ, Antoszyk AN, Ruby AJ, Orth D, Bressler SB, Fish GE, Hubbard GB, Klein ML, Friberg TR, Rosenfeld PJ, Toth CA, Bernstein P: **Lutein/Zeaxanthin for the Treatment of Age-Related Cataract: AREDS2 Randomized Trial Report No. 4.** *JAMA Ophthalmol* 2013, **131**:843–50.

19. Friberg TR, Brennen PM, Freeman WR, Musch DC: **Prophylactic treatment of age-related macular degeneration report number 2: 810-nanometer laser to eyes with drusen: bilaterally eligible patients.** *Ophthalmic Surg Lasers Imaging* , **40**:530–8.

20. Tranos P, Vacalis A, Asteriadis S, Koukoula S, Vachtsevanos A, Perganta G, Georgalas I: **Resistance to antivascular endothelial growth factor treatment in age-related macular degeneration.** *Drug Des Devel Ther* 2013, **7**:485–90.

21. Seidel G, Werner C, Weger M, Steinbrugger I, Haas A: **Combination treatment of photodynamic therapy with verteporfin and intravitreal ranibizumab in patients with retinal angiomatous proliferation.** *Acta Ophthalmol* 2013, **91**:e482–5.

22. Lamba DA, Karl MO, Reh TA: **Strategies for retinal repair: cell replacement and regeneration.** *Prog Brain Res* 2009, **175**:23–31.

23. Luo L, Harmon J, Yang X, Chen H, Patel S, Mineau G, Yang Z, Constantine R, Buehler J, Kaminoh Y, Ma X, Wong TY, Zhang M, Zhang K: **Familial aggregation of age-related macular degeneration in the Utah population.** *Vision Res* 2008, **48**:494–500.

24. Cooke Bailey JN, Sobrin L, Pericak-Vance MA, Haines JL, Hammond CJ, Wiggs

JL: **Advances in the genomics of common eye diseases.** *Hum Mol Genet* 2013, **22**:R59–65.

25. Braley AE: **Dystrophy of the macula.** *Am J Ophthalmol* 1966, **61**:1–24.

26. Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA: **Complement factor H polymorphism and age-related macular degeneration.** *Science* 2005, **308**:421–4.

27. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA: **Complement factor H variant increases the risk of age-related macular degeneration.** *Science* 2005, **308**:419–21.

28. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration.** *Science* 2005, **308**:385–9.

29. Hageman GS, Anderson DH, Johnson L V, Hancox LS, Taiber AJ, Hardisty LI, Hageman JL, Stockman HA, Borchardt JD, Gehrs KM, Smith RJH, Silvestri G, Russell SR, Klaver CCW, Barbazetto I, Chang S, Yannuzzi LA, Barile GR, Merriam JC, Smith RT, Olsh AK, Bergeron J, Zernant J, Merriam JE, Gold B, Dean M, Allikmets R: **A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration.** *Proc Natl Acad Sci U S A* 2005, **102**:7227–32.

30. Rivera A, Fisher SA, Fritsche LG, Keilhauer CN, Lichtner P, Meitinger T, Weber BHF: **Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk.** *Hum Mol Genet* 2005, **14**:3227–36.

31. Hughes AE, Orr N, Esfandiary H, Diaz-Torres M, Goodship T, Chakravarthy U: **A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration.** *Nat Genet* 2006, **38**:1173–7.

32. Gold B, Merriam JE, Zernant J, Hancox LS, Taiber AJ, Gehrs K, Cramer K, Neel J, Bergeron J, Barile GR, Smith RT, Hageman GS, Dean M, Allikmets R: **Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration.** *Nat Genet* 2006, **38**:458–62.

33. Maller JB, Fagerness JA, Reynolds RC, Neale BM, Daly MJ, Seddon JM:

Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet* 2007, **39**:1200–1.

34. Yates JRW, Sepp T, Matharu BK, Khan JC, Thurlby DA, Shahid H, Clayton DG, Hayward C, Morgan J, Wright AF, Armbrecht AM, Dhillon B, Deary IJ, Redmond E, Bird AC, Moore AT: **Complement C3 variant and the risk of age-related macular degeneration.** *N Engl J Med* 2007, **357**:553–61.

35. **Age-related Macular Degeneration** [https://www.23andme.com/health/Age-related-Macular-Degeneration]

36. Fritsche LG, Chen W, Schu M, Yaspan BL, Yu Y, Thorleifsson G, Zack DJ, Arakawa S, Cipriani V, Ripke S, Igo RP, Buitendijk GHS, Sim X, Weeks DE, Guymer RH, Merriam JE, Francis PJ, Hannum G, Agarwal A, Armbrecht AM, Audo I, Aung T, Barile GR, Benchaboune M, Bird AC, Bishop PN, Branham KE, Brooks M, Brucker AJ, Cade WH, et al.: **Seven new loci associated with age-related macular degeneration.** *Nat Genet* 2013, **45**:433–9, 439e1–2.

37. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, Burdon KP, Hebbring SJ, Wen C, Gorski M, Kim IK, Cho D, Zack D, Souied E, Scholl HPN, Bala E, Lee KE, Hunter DJ, Sardell RJ, Mitchell P, Merriam JE, Cipriani V, Hoffman JD, Schick T, Lechanteur YTE, Guymer RH, Johnson MP, Jiang Y, Stanton CM, Buitendijk GHS, et al.: **A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants.** *Nat Genet* 2015, **advance on**.

38. Yu Y, Bhangale TR, Fagerness J, Ripke S, Thorleifsson G, Tan PL, Souied EH, Richardson AJ, Merriam JE, Buitendijk GHS, Reynolds R, Raychaudhuri S, Chin KA, Sobrin L, Evangelou E, Lee PH, Lee AY, Leveziel N, Zack DJ, Campochiaro B, Campochiaro P, Smith RT, Barile GR, Guymer RH, Hogg R, Chakravarthy U, Robman LD, Gustafsson O, Sigurdsson H, Ortmann W, et al.: **Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration.** *Hum Mol Genet* 2011, **20**:3699–709.

39. Mendel GJ: **Experiments Concerning Plant Hybrids**. *Proc Nat Hist Soc Brunn* 1866:3–47.

40. Falconer DS, Mackay TF, Frankham R: **Introduction to quantitative genetics (4th edition)**. *Trends Genet* 1996, **12.7**:280.

41. **How to calculate heritability** [http://www.cureffi.org/2013/02/04/how-to-

calculate-heritability/]

42. Visscher PM, Hill WG, Wray NR: **Heritability in the genomics era--concepts and misconceptions.** *Nat Rev Genet* 2008, **9**:255–66.

43. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis.** *Am J Hum Genet* 2011, **88**:76–82.

44. Gray A, Stewart I, Tenesa A: **Advanced complex trait analysis.** *Bioinformatics* 2012, **28**:3134–6.

45. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet* 2010, **42**:348–54.

46. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies.** *Nat Methods* 2011, **8**:833–5.

47. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44**:821–4.

48. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS: **Rapid variance components-based method for whole-genome association analysis.** *Nat Genet* 2012, **44**:1166–70.

49. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42**:565–9.

50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559–75.

51. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR: **Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood.** *Bioinformatics* 2012, **28**:2540–2.

52. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904–9.

53. Lee SH, Wray NR, Goddard ME, Visscher PM: **Estimating missing heritability for disease from genome-wide association studies.** *Am J Hum Genet* 2011, **88**:294–305.

54. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** *Nature* 2009, **460**:748–52.

55. Speed D, Hemani G, Johnson MR, Balding DJ: **Improved heritability estimation from genome-wide SNPs.** *Am J Hum Genet* 2012, **91**:1011–21.

56. Speed D, Hemani G, Johnson MR, Balding DJ: **Response to Lee et al.: SNP-based heritability analysis with dense data.** *Am J Hum Genet* 2013, **93**:1155–7.

57. Visscher PM, Hemani G, Vinkhuyzen AAE, Chen G-B, Lee SH, Wray NR, Goddard ME, Yang J: **Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples**. *PLoS Genet* 2014, **10**:e1004269.

58. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL: **Advantages and pitfalls in the application of mixed-model association methods.** *Nat Genet* 2014, **46**:100–6.

59. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D: **Improved linear mixed models for genome-wide association studies.** *Nat Methods* 2012, **9**:525–6.

60. Lippert C, Quon G, Kang EY, Kadie CM, Listgarten J, Heckerman D: **The benefits of selecting phenotype-specific variants for applications of mixed models in genomics.** *Sci Rep* 2013, **3**:1815.

61. Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, Barton A, Bickeböller H, Bowden DW, Eyre S, Freedman BI, Friedman DJ, Field JK, Groop L, Haugen A, Heinrich J, Henderson BE, Hicks PJ, Hocking LJ, Kolonel LN, Landi MT, Langefeld CD, Le Marchand L, Meister M, Morgan AW, Raji OY, Risch A, Rosenberger A, Scherf D, Steer S, et al.: **Informed conditioning on clinical covariates increases power in case-control association studies.** *PLoS Genet* 2012, **8**:e1003032.

62. Hayeck TJ, Zaitlen NA, Loh P-R, Vilhjalmsson B, Pollack S, Gusev A, Yang J, Chen G-B, Goddard ME, Visscher PM, Patterson N, Price AL: **Mixed model with correction for case-control ascertainment increases association power.** *Am J Hum Genet* 2015, **96**:720–30.

63. Fritsche LG, Fariss RN, Stambolian D, Abecasis GR, Curcio CA, Swaroop A:

Age-related macular degeneration: genetics and biology coming together. *Annu Rev Genomics Hum Genet* 2014, **15**:151–71.

64. Zipfel PF, Lauer N, Skerka C: **The role of complement in AMD.** *Adv Exp Med Biol* 2010, **703**:9–24.

65. de Jong PTVM: **Age-related macular degeneration.** *N Engl J Med* 2006, **355**:1474–85.

66. Dunaief JL: **The Role of Apoptosis in Age-Related Macular Degeneration**. *Arch Ophthalmol* 2002, **120**:1435.

67. Bhosale P, Larson AJ, Bernstein PS: **Factorial analysis of tricarboxylic acid cycle intermediates for optimization of zeaxanthin production from Flavobacterium multivorum**. *J Appl Microbiol* 2004, **96**:623–629.

68. Zhao C, Vollrath D: **mTOR pathway activation in age-related retinal disease.** *Aging (Albany NY)* 2011, **3**:346–7.

69. Christen WG: **A Prospective Study of Cigarette Smoking and Risk of Age-Related Macular Degeneration in Men**. *JAMA J Am Med Assoc* 1996, **276**:1147.

70. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.

71. **Ingenuity IPA - Integrate and understand complex 'omics data** [www.ingenuity.com/products/ipa]

72. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29–34.

73. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Res* 2005, **33**(Database issue):D428–32.

74. Raychaudhuri S: **VIZ-GRAIL: visualizing functional connections across disease loci.** *Bioinformatics* 2011, **27**:1589–90.

75. Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, Sutcliffe JS, Haines JL: **Genetic analysis of biological pathway data through genomic randomization.** *Hum Genet* 2011, **129**:563–71.

76. Ramanan VK, Shen L, Moore JH, Saykin AJ: **Pathway analysis of genomic data: concepts, methods, and prospects for future development.** *Trends Genet* 2012, **28**:323–32.

77. Naj AC, Scott WK, Courtenay MD, Cade WH, Schwartz SG, Kovach JL, Agarwal A, Wang G, Haines JL, Pericak-Vance MA: **Genetic factors in nonsmokers with age-related macular degeneration revealed through genome-wide gene-environment interaction analysis.** *Ann Hum Genet* 2013, **77**:215–31.

78. Schmidt S, Saunders AM, De La Paz MA, Postel EA, Heinis RM, Agarwal A, Scott WK, Gilbert JR, McDowell JG, Bazyk A, Gass JD, Haines JL, Pericak-Vance MA: **Association of the apolipoprotein E gene with age-related macular degeneration: possible effect modification by family history, age, and gender.** *Mol Vis* 2000, **6**:287–93.

79. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34**(Database issue):D590–8.

80. **The International HapMap Project.** *Nature* 2003, **426**:789–96.

81. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA: **Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays.** *Nat Methods* 2006, **3**:511–8.

82. Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance MA, Haines JL: **Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come.** *Am J Hum Genet* 2010, **86**:621–5.

83. Yang Z, Stratton C, Francis PJ, Kleinman ME, Tan PL, Gibbs D, Tong Z, Chen H, Constantine R, Yang X, Chen Y, Zeng J, Davey L, Ma X, Hau VS, Wang C, Harmon J, Buehler J, Pearson E, Patel S, Kaminoh Y, Watkins S, Luo L, Zabriskie NA, Bernstein PS, Cho W, Schwager A, Hinton DR, Klein ML, Hamon SC, et al.: **Toll-like receptor 3 and geographic atrophy in age-related macular degeneration.** *N Engl J Med* 2008, **359**:1456–63.

84. Fisher RA: **The Correlation between Relatives on the Supposition of Mendelian Inheritance**. *Trans R Soc Edinburgh* 1919, **52**:399–433.

85. Moore JH: **The ubiquitous nature of epistasis in determining susceptibility to common human diseases.** *Hum Hered* 2003, **56**:73–82.

86. Phillips PC: **Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems**. *Nat Rev Genet* 2008, **9**:855–867.

87. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10**:392–404.

88. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

89. Zuk O, Hechter E, Sunyaev SR, Lander ES: **The mystery of missing heritability: Genetic interactions create phantom heritability.** *Proc Natl Acad Sci U S A* 2012, **109**:1193–8.

90. Cockerham CC: **An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present.** *Genetics* 1954, **39**:859–82.

91. Wright S: **Coefficients of inbreeding and relationship**. *Am Nat* 1922, **56**:330–338.

92. Sütterlin T, Kolb C, Dickhaus H, Jäger D, Grabe N: **Bridging the scales: semantic integration of quantitative SBML in graphical multi-cellular models and simulations with EPISIM and COPASI.** *Bioinformatics* 2013, **29**:223–9.

93. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH: **GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures.** *BioData Min* 2012, **5**:16.

94. Edwards T, Bush W, Turner S, Dudek S, Torstenson E, Schmidt M, Martin E, Ritchie M, Marchiori E, Moore J: **Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA**. *Lect Notes Comput Sci Evol Comput Mach Learn Data Min Bioinforma* 2008, **4973**:24–35.

95. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark

AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747–53.

96. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap.** *Bioinformatics* 2008, **24**:2938–9.

97. Kuman S: **Age related macular degeneration: a complex pathology**. *Austin J Genet Genomic Res* 2014, **1**:5–9.

98. Wang G, Spencer KL, Scott WK, Whitehead P, Court BL, Ayala-Haedo J, Mayo P, Schwartz SG, Kovach JL, Gallins P, Polk M, Agarwal A, Postel EA, Haines JL, Pericak-Vance MA: **Analysis of the indel at the ARMS2 3'UTR in age-related macular degeneration.** *Hum Genet* 2010, **127**:595–602.

99. Kanda A, Chen W, Othman M, Branham KEH, Brooks M, Khanna R, He S, Lyons R, Abecasis GR, Swaroop A: **A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration.** *Proc Natl Acad Sci U S A* 2007, **104**:16227–32.

100. Kortvely E, Ueffing M: **Gene Structure of the 10q26 Locus: A Clue to Cracking the ARMS2/HTRA1 Riddle?** *Adv Exp Med Biol* 2016, **854**:23–9.

101. Heier JS: **Neovascular age-related macular degeneration: individualizing therapy in the era of anti-angiogenic treatments.** *Ophthalmology* 2013, **120**(5 Suppl):S23–5.