

**GENETICS OF AGE-RELATED MACULOPATHY &  
SCORE STATISTICS FOR X-LINKED  
QUANTITATIVE TRAIT LOCI**

by

**Jóhanna Jakobsdóttir**

BS, University of Iceland, Reykjavík, Iceland, 2004

Submitted to the Graduate Faculty of  
the Department of Biostatistics  
Graduate School of Public Health in partial fulfillment  
of the requirements for the degree of  
**Doctor of Philosophy**

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

**Jóhanna Jakobsdóttir**

It was defended on

**April 1, 2009**

and approved by

Dissertation Advisor: Daniel E. Weeks, Ph.D., Professor, Depts. of Human Genetics and  
Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Eleanor Feingold, Ph.D., Associate Professor, Depts. of Human Genetics  
and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Sati Mazumdar, Ph.D., Professor, Department of Biostatistics, Graduate  
School of Public Health, University of Pittsburgh

Committee Member: Lisa Weissfeld, Ph.D., Professor, Department of Biostatistics, Graduate  
School of Public Health, University of Pittsburgh

Copyright © by Jóhanna Jakobsdóttir  
2009

**GENETICS OF AGE-RELATED MACULOPATHY & SCORE STATISTICS FOR  
X-LINKED QUANTITATIVE TRAIT LOCI**

Jóhanna Jakobsdóttir, PhD

University of Pittsburgh, 2009

Age-related maculopathy (ARM) is a common cause of irreparable vision loss in industrialized countries. The disease is characterized by progressive loss of central vision making everyday tasks challenging. The etiology is complex and has both an environmental and a strong genetic components. The public health relevance of the work is to improve the understanding genetic causes in the disease etiology and ultimately to lead to better disease management and prevention. From my ARM work, I present four papers covering range of statistical approaches. The first paper presents fine-mapping efforts, using both linkage and association methods, under previously identified linkage peaks on chromosomes 1q31 and 10q26. We replicate the discovery of the complement factor H (*CFH*) gene on 1q31 and identify a novel locus, harboring three closely linked genes (*PLEKHA1*, *LOC387715*, and *HTRA1*), on 10q26. Both discoveries have been widely replicated. In the next paper I present meta-analysis of 11 *CFH* and 5 *LOC387715* data sets. We also replicate these findings in two independent case-control cohorts, including one cohort, where ARM status was not a factor in the ascertainment. In the third paper we replicate discoveries of new complement related loci (*C2* and *CFB*) on chromosome 19p13 as well as developing classification models based on SNPs from *CFH*, *LOC387715*, and *C2*. The last paper focuses on applying statistical techniques from the diagnostic medicine literature to ARM. We comment on the importance of understanding the difference and similarities between different goals of genetic studies: improving etiological understanding or finding variants that discriminate well between cases and controls. This work is particularly relevant today when there has been explosion in the availability of direct-to-consumer DNA tests.

In addition to carrying out linkage and association analysis, I also have extended the statistical theory behind score-based linkage analyses for X chromosomal markers. This work has public health relevance because many complex common diseases have sex-specific differences, such as prevalence

and age of onset. Modeling those appropriately with powerful and robust methods will bring an improved understanding of their genetic basis.

## TABLE OF CONTENTS

<b>PREFACE</b> . . . . .	xvii
<b>1.0 BASICS IN GENETICS AND GENE MAPPING</b> . . . . .	1
<b>2.0 BACKGROUND ON THE AGE-RELATED MACULOPATHY PROJECT</b> . . . . .	5
2.1 Epidemiology and genetics of Age-related maculopathy . . . . .	5
2.2 Statistical methods used in the project . . . . .	8
2.2.1 Linkage analysis . . . . .	8
2.2.1.1 Parametric LOD scores and HLOD scores . . . . .	8
2.2.1.2 Model-free LOD scores . . . . .	9
2.2.2 Association analysis . . . . .	10
2.2.2.1 $\chi^2$ and Fisher's exact tests and logistic regression . . . . .	10
2.2.2.2 CCREL and MQLS . . . . .	11
2.2.3 Joint linkage and association . . . . .	12
2.2.3.1 GIST . . . . .	12
2.2.4 Meta-analysis . . . . .	12
2.2.4.1 Fixed effects modeling . . . . .	13
2.2.4.2 Random effects modeling . . . . .	13
2.2.5 Model building . . . . .	13
2.2.5.1 Logistic regression . . . . .	13
2.2.5.2 GMDR . . . . .	14
2.2.5.3 ROC curves . . . . .	14
2.2.6 Practical issues . . . . .	14
2.2.6.1 Samples vs. individuals . . . . .	14
2.2.6.2 Check your files, scripts, and results . . . . .	15

<b>3.0 SUSCEPTIBILITY GENES FOR AGE-RELATED MACULOPATHY ON CHROMOSOME 10Q26</b>	17
3.1 Abstract	17
3.2 Introduction	18
3.3 Material and Methods	19
3.3.1 Families and Case-Control Cohort	19
3.3.2 Affection-Status Models	19
3.3.3 Pedigree and Genotyping Errors and Data Handling	21
3.3.4 Allele Frequencies and Hardy-Weinberg Equilibrium	21
3.3.5 Genetic Map	21
3.3.6 LD Structure	21
3.3.7 Linkage Analysis	22
3.3.7.1 Two-point analysis	22
3.3.7.2 Multipoint analysis ignoring LD	22
3.3.7.3 Multipoint analysis using htSNPs	22
3.3.8 Association Analysis	23
3.3.9 GIST Analysis	23
3.3.10 Tripartite Analyses	24
3.3.11 Part I: Analysis of CIDR SNPs	24
3.3.12 Part II: Analysis of Locally Genotyped SNPs	26
3.3.13 Part III: Interaction and OR Analysis	26
3.3.13.1 Unrelated cases	26
3.3.13.2 Analysis of interaction with <i>CFH</i>	29
3.3.13.3 Magnitude of association	29
3.3.13.4 Multiple-testing issues	30
3.4 Results	30
3.4.1 Part I: Analysis of CIDR SNPs	30
3.4.1.1 CIDR linkage results	30
3.4.1.2 CIDR association results	31
3.4.1.3 CIDR GIST results	35
3.4.2 PART II: Analysis of Locally Genotyped SNPs	35
3.4.2.1 Local association results	35

3.4.2.2	Local GIST results . . . . .	35
3.4.3	Part III: Interaction and OR Analyses . . . . .	37
3.4.3.1	GIST results . . . . .	37
3.4.3.2	Logistic regression results . . . . .	37
3.4.3.3	OR and AR . . . . .	37
3.4.3.4	Subphenotype analyses . . . . .	38
3.5	Discussion . . . . .	38
<b>4.0</b>	<b><i>CFH</i>, <i>ELOVL4</i>, <i>PLEKHA1</i> AND <i>LOC387715</i> GENES AND SUSCEPTI-</b>	
	<b>BILITY TO AGE-RELATED MACULOPATHY: AREDS AND CHS CO-</b>	
	<b>HORTS AND META-ANALYSES . . . . .</b>	<b>45</b>
4.1	Abstract . . . . .	45
4.2	Introduction . . . . .	46
4.3	Results . . . . .	48
4.3.1	Association analyses . . . . .	48
4.3.1.1	<i>CFH</i> . . . . .	49
4.3.1.2	<i>ELOVL4</i> . . . . .	52
4.3.1.3	<i>PLEKHA1</i> and <i>LOC387715</i> . . . . .	53
4.3.2	Interaction analyses . . . . .	54
4.3.3	<i>APOE</i> results . . . . .	55
4.3.4	Meta-analyses . . . . .	55
4.3.4.1	Meta-analysis of <i>CFH</i> . . . . .	55
4.3.4.2	Meta-analysis of <i>LOC387715</i> . . . . .	57
4.4	Discussion . . . . .	57
4.5	Materials and Methods . . . . .	62
4.5.1	Cardiovascular health study (CHS) participantssampling and phenotyping .	62
4.5.2	Age-related eye disease study (AREDS) participantssampling and pheno-	
	typing . . . . .	63
4.5.3	Genotyping . . . . .	64
4.5.4	Association analyses . . . . .	64
4.5.5	Distinguishing between <i>PLEKHA1</i> and <i>LOC387715</i> . . . . .	65
4.5.6	Interaction analyses . . . . .	65
4.5.7	<i>APOE</i> analyses . . . . .	66



4.5.8	Meta-analyses . . . . .	67
<b>5.0</b>	<b>C2 AND CFB GENES IN AGE-RELATED MACULOPATHY AND JOINT ACTION WITH CFH AND LOC387715 GENES . . . . .</b>	<b>71</b>
5.1	Abstract . . . . .	71
5.2	Introduction . . . . .	72
5.3	Materials and methods . . . . .	74
5.3.1	Phenotyping, study participants and quality control . . . . .	74
5.3.2	Genotyping . . . . .	75
5.3.3	Association analyses and LD estimation . . . . .	77
5.3.3.1	Case-Control data . . . . .	77
5.3.3.2	Family data . . . . .	77
5.3.4	Multifactor and gene-gene interaction analyses . . . . .	78
5.3.4.1	Logistic regression . . . . .	78
5.3.4.2	GMDR . . . . .	79
5.3.5	Interaction with cigarette smoking . . . . .	79
5.4	Results . . . . .	80
5.4.1	Results of association analyses . . . . .	80
5.4.2	Results of multifactor analyses . . . . .	80
5.4.2.1	Logistic regression . . . . .	80
5.4.2.2	GMDR . . . . .	82
5.4.2.3	Logistic regression vs. GMDR . . . . .	86
5.4.3	Results of gene-cigarette smoking interaction analysis . . . . .	86
5.5	Discussion . . . . .	86
<b>6.0</b>	<b>INTERPRETATION OF GENETIC ASSOCIATION STUDIES: MARKERS WITH REPLICATED HIGHLY SIGNIFICANT ODDS RATIOS MAY BE POOR CLASSIFIERS . . . . .</b>	<b>91</b>
6.1	Abstract . . . . .	91
6.2	Introduction . . . . .	92
6.3	Two statistical methods . . . . .	93
6.4	The Odds Ratio, Classification, Calibration, and Prediction . . . . .	95
6.5	The Odds Ratio, Relative Risk, and Risk . . . . .	96
6.6	Clinical Validity and Utility of Predictive Genetic Testing . . . . .	98

6.7	Reclassification	99
6.8	Examples	99
6.8.1	Risk of Cardiovascular Events	99
6.8.2	Risk of Type 2 Diabetes	100
6.8.3	Risk of Prostate Cancer	100
6.8.4	Risk of Inflammatory Bowel Disease	100
6.8.5	Risk of Age-Related Macular Degeneration	101
6.9	Discussion of the AMD Example	105
6.10	Conclusions	107
<b>7.0</b>	<b>SCORE STATISTICS FOR X-LINKED QTLS</b>	<b>109</b>
7.1	X-linked inheritance	109
7.2	Overview: QTL Linkage methods	110
7.3	Model	112
7.4	Allelic and genotypic effects	112
7.5	Variances-Covariances	113
7.5.1	X-linked kinship coefficient and variances-covariances	113
7.5.2	Female-Female relative pairs	116
7.5.3	Male-Male relative pairs	116
7.5.4	Female-Male relative pairs	117
7.5.5	Marker loci	117
7.6	Score functions and likelihood	119
7.6.1	Score for $\sigma_{a,f}^2$	120
7.6.2	Score for $\sigma_{d,f}^2$	123
7.6.3	Score for $\sigma_{X,m}^2$	123
7.6.4	Score for $\sigma_{X,fm}$	123
7.7	Summary of scores	124
7.8	Variance of the scores	125
7.9	Score statistics and asymptotic distributions	126
7.9.1	Derivation of the statistic when parameter estimates are out of bounds	128
7.9.1.1	Condition $\sigma_{a,f}^2 \geq 0$ fails	128
7.9.1.2	Condition $\sigma_{X,m}^2 \geq 0$ fails	128
7.9.1.3	Both $\sigma_{a,f}^2 \geq 0$ and $\sigma_{X,m}^2 \geq 0$ fail	129

7.9.1.4	Condition $\sigma_{a,f}^2 \geq 0$ fails	129
7.9.1.5	If both $\sigma_{a,f}^2 \geq 0$ and $\sigma_{X,m}^2 \geq 0$ hold but	129
7.9.2	Distribution of the statistic	130
7.9.3	Simpler model: Assuming equal allelic effects in both sexes	131
7.9.3.1	No dominance in females	132
7.10	Discussion and future work	132
7.10.1	Selected sampling, small samples, and choice of variance	132
7.10.2	Inactivation in females and the pseudoautosomal regions	132
7.10.3	Other genetic and environmental effects	133
7.10.4	Asymptotic and empirical distributions	133
7.10.5	Score statistics for X-linked association analysis	133
7.10.6	General properties of score statistics	134
<b>8.0</b>	<b>CONCLUSIONS</b>	<b>135</b>
8.1	Synthesis of the ARM work	135
8.2	My future aims	139
<b>APPENDIX A. FOR CHAPTER 3</b>		<b>141</b>
<b>APPENDIX B. FOR CHAPTER 4</b>		<b>150</b>
B.1	Distinguishing between <i>PLEKHA1</i> and <i>LOC387715</i>	150
B.1.1	Distinguishing between <i>PLEKHA1</i> and <i>LOC387715</i> –Results	152
B.2	HapMap populations	154
<b>APPENDIX C. FOR CHAPTER 5</b>		<b>172</b>
C.1	Logistic regression analyses	172
C.1.1	Coding in two-factor models	172
C.1.2	Coding in three-factor models	173
C.2	Association analyses– <i>CFH</i> and <i>LOC387715</i>	173
<b>APPENDIX D. FOR CHAPTER 6</b>		<b>178</b>
D.1	Application of classification-based methods to AMD data	178
D.1.1	AMD data	179
D.1.2	Methods	179
D.1.3	Accounting for covariates	180
D.2	Estimating the AUC from meta-data	180
D.3	Details on data in other real data examples	181

D.3.1 Cardiovascular events . . . . .	181
D.3.2 Type 2 diabetes . . . . .	181
D.3.3 Prostate cancer . . . . .	181
D.3.4 Inflammatory bowel diseases . . . . .	183
<b>APPENDIX E. FOR CHAPTER 7 . . . . .</b>	<b>184</b>
<b>BIBLIOGRAPHY . . . . .</b>	<b>186</b>

## LIST OF TABLES

3.1	Distribution of Subphenotypes in Patients with Advanced ARM . . . . .	20
3.2	Summary of Statistical Analyses and Sample Sizes in Parts I–III . . . . .	25
3.3	CCREL, GIST, and Allele-Frequency Estimation for Families and Controls Typed at CIDR . . . . .	34
3.4	CCREL, GIST, and Allele-Frequency Estimation for Locally Typed Families and Controls . . . . .	36
3.5	ORs, ARs, and Simulated $P$ Values from $\chi^2$ Test with 10,000 Replicates . . . . .	39
4.1	Characteristics of the study populations . . . . .	49
4.2	Genotype distributions by ARM status . . . . .	50
4.3	Results of allele- and genotype-association tests . . . . .	50
4.4	ORs and PAR% for subjects who are hetero- and homozygous for $Y402H$ in <i>CFH</i> and $S69A$ in <i>LOC387715</i> . . . . .	51
4.5	Results of fitting two-factor models by logistic regression . . . . .	56
5.1	Samples sizes and other characteristics of the data. . . . .	76
5.2	Association results for $C2/CFB$ variants, $Y402H$ in <i>CFH</i> , and $S69A$ in <i>LOC387715</i> . . . . .	81
5.3	Results of fitting two-factor logistic regression models. . . . .	83
5.4	Results of fitting three-factor logistic regression models. . . . .	84
5.5	Results of GMDR analyses. . . . .	87
6.1	AUC, risk allele frequency in cases and controls in an additive model . . . . .	98
6.2	Results of logistic regression and ROC analysis . . . . .	103
6.3	Positive predictive values for different prevalence values . . . . .	105
A1	Allele labeling . . . . .	141
A2	Primers, Annealing Conditions, and Restriction Endonucleases Used for Genotype Data Collection . . . . .	144

A3	Results of Fitting Two-Locus Models by Logistic Regression . . . . .	145
A4	ORs, ARs, and Simulated $P$ Values from $\chi^2$ Test with 10,000 Replicates . . . . .	146
B1	Genotype distributions in AREDS and CHS cohorts, by ARM status . . . . .	155
B2	Estimated crude ORs, corresponding 95% CIs, and PARs, unadjusted for age and gender . . . . .	156
B3	Estimated ORs, corresponding 95% CIs, and PARs, adjusted for age and gender . . . . .	157
B4	Joint ORs and 95% CIs at $Y402H$ in $CFH$ and $S69A$ in $LOC387715$ . . . . .	158
B5	Joint genotype distribution at $Y402H$ in $CFH$ and $S69A$ in $LOC387715$ in the AREDS cohort . . . . .	159
B6	Joint genotype distribution at $Y402H$ in $CFH$ and $S69A$ in $LOC387715$ in the CHS cohort . . . . .	159
B7	Joint ORs and 95% CIs at $Y402H$ in $CFH$ and smoking, and $S69A$ in $LOC387715$ and smoking . . . . .	160
B8	Genotype distribution at $Y402H$ in $CFH$ and $S69A$ in $LOC387715$ in the AREDS cohort, by smoking history (ever vs. never smoked) . . . . .	161
B9	Genotype distribution at $Y402H$ in $CFH$ and $S69A$ in $LOC387715$ in the CHS cohort, by smoking history (ever vs. never smoked) . . . . .	161
B10	Characteristics of studies included in meta-analysis of $Y402H$ in $CFH$ . . . . .	162
B11	Results of meta-analysis of $Y402H$ in $CFH$ . . . . .	163
B12	Characteristics of studies included in meta-analysis of $S69A$ in $LOC387715$ . . . . .	164
B13	Results of meta-analysis of $S69A$ in $LOC387715$ . . . . .	165
C1	Genotype counts for $C2/CFB$ variants, $Y402H$ in $CFH$ , and $S69A$ in $LOC387715$ . . . . .	174
C2	Joint and relative genotype frequencies . . . . .	175
D1	Association results of 9 SNPs associated with LDL and HDL cholesterol . . . . .	182
D2	Association results of 12 type 2 diabetes SNPs. . . . .	182
D3	Association results of two prostate cancer disease SNPs. . . . .	183
D4	Association results of five Crohns disease SNPs. . . . .	183

## LIST OF FIGURES

2.1	Anatomy of the eye . . . . .	6
2.2	Vision loss due to ARM . . . . .	7
3.1	Location of CIDR SNPs and locally genotyped SNPs with respect to candidate genes . . . . .	27
3.2	LD patterns around the candidate genes . . . . .	28
3.3	Two-point and multipoint linkage results on chromosome 10 . . . . .	32
3.4	Two-point (2pt) and multipoint (mpt) linkage results on chromosome 1 . . . . .	33
4.1	Estimated crude ORs and 95% CIs for <i>CFH</i> , <i>ELOVL4</i> , <i>PLEKHA1</i> and <i>LOC387715</i> genes . . . . .	52
4.2	Estimated ORs and 95% CIs from meta-analysis of <i>Y402H</i> in <i>CFH</i> . . . . .	69
4.3	Estimated ORs and 95% CIs from meta-analysis of <i>S69A</i> in <i>LOC387715</i> . . . . .	70
5.1	Linkage disequilibrium (LD) across the <i>C2/CFB</i> region . . . . .	82
5.2	Results of unadjusted GMDR analysis . . . . .	85
6.1	Accuracy curves for binary markers . . . . .	96
6.2	AUC for additive risk models . . . . .	97
6.3	ROC curves for AMD classification models . . . . .	102
6.4	Integrated predictiveness and classification plot . . . . .	104
7.1	Detailed identity states and condensed states . . . . .	115
7.2	The condensed identity states for X-linked loci . . . . .	115
A1	LD patterns on chromosome 10 . . . . .	148
A2	LD patterns on chromosome 1 . . . . .	149

B1	Estimated ORs and 95% CIs for <i>CFH</i> . A: $OR_{dom}$ for evaluation of dominance effects (CT+CC vs. TT). B: $OR_{het}$ for evaluation of the risk for heterozygotes (CT vs. TT). C: $OR_{rec}$ for evaluation of recessive effects (CC vs. CT+TT). D: $OR_{hom}$ for evaluation of the risk for homozygotes (CC vs. TT). The dotted vertical line marks the null value of OR of 1. . . . .	166
B2	Estimated ORs and 95% CIs for <i>ELOVL4</i> . . . . .	167
B3	Estimated ORs and 95% CIs for <i>PLEKHA1</i> . . . . .	168
B4	Estimated ORs and 95% CIs for <i>LOC387715</i> . . . . .	169
B5	Genotype frequencies in meta-analysis of Y402H in <i>CFH</i> . . . . .	170
B6	Genotype frequencies in meta-analysis of S69A in <i>LC387715</i> . . . . .	171
C1	Minor allele frequency (MAF) of SNPs typed for the HapMap CEU population . . .	176
C2	GMDR sensitivity analyses for the three-factor unadjusted model . . . . .	177



## PREFACE

*To Palli, Gréta Björg and my parents*

I am grateful to my advisor, Dr. Daniel E. Weeks. I could not have asked for a better mentor. He was always supportive and encouraged me to keep going even when I felt as I would never finish. By example he motivated me to work hard. I also am sincerely thankful to Dr. Weeks and Dr. Eleanor Feingold for all their help when I first applied to the University of Pittsburgh. It is fair to say that without their help and kindness I would not have come here. I would also like to thank Dr. Sati Mazumdar and Dr. Lisa Weissfeld for serving on my proposal and dissertation committees. Additionally, both of them are excellent teachers and taught two of my favorite and most useful biostatistics courses. I thank Dr. Michael B. Gorin for supporting me through his NIH funded grant and giving me the opportunity to work on such a fun and important project. I am also very thankful to my mathematics teachers at the Reykjavk Junior College (MR) for teaching me real mathematics. I am grateful for the strong basic education I received while at MR and all the friends I made.

## 1.0 BASICS IN GENETICS AND GENE MAPPING

The genetic material in humans is stored in 23 pairs of chromosomes: 22 pairs of autosomes, and 1 pair of sex chromosomes. Each chromosome is a double helix of deoxyribonucleic acid (DNA), which is composed of a chain of nucleotides. Each nucleotide consists of three components, one of which is the nitrogenous base. The bases that make up the DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). A sequence of three bases codes for an amino acid, the sub-components of proteins (THOMAS 2004).

Genetic differences between people arise from variant genes along the chromosomes. Specific positions along the chromosomes are often referred to as loci and the variant genes are known as alleles. Two alleles at the same locus constitute a specific genotype. A locus can span from a single base pair to thousands of megabases. Single nucleotide polymorphisms (SNPs) are genetic variants whose alleles consist of differences at a single specific base pair position. Opposed to genotype, the term phenotype denotes an observable characteristic or trait. For a simple Mendelian diseases it may be possible to infer the genotype underlying a specific observed phenotype but in complex diseases this is not necessarily the case (LANGE 2003). In genetics we generally distinguish between simple Mendelian traits and complex traits. Mendelian traits exhibit a simple pattern of inheritance of major genes (usually only one gene) that are both necessary and sufficient to infer the outcome. Complex traits do not exhibit the same simple pattern of inheritance, and even genes with strong effects on the complex trait are usually neither necessary nor sufficient to infer the phenotype. Typically, more than one gene, and often many genes with small effects (so called polygenes), along with environmental factors and gene-environmental interactions determine complex traits (THOMAS 2004).

The old fashioned definition of a gene as a physical unit of heredity, given above, has been expanded and the term is commonly used to refer to the part of the DNA (i.e., deoxyribonucleic acid) that codes for protein. The term locus is, on the other hand, used for any physical unit of the

DNA. For example if a chromosomal region is identified via linkage study, one often uses ‘locus’ to refer to that region rather than ‘gene’ (LANGE *et al.* 2001).

Mendel’s first law, or the law of segregation, states that each parent transmits one copy of each of his/her chromosomes randomly to his/her offspring. The two parental chromosomes are termed grandpaternal and grandmaternal chromosomes and are said to be homologous. Rarely an entire chromosome is transmitted, usually the chromosome inherited through the gametes is made up of recombined segments of the grandpaternal and grandmaternal chromosomes. Thus the child’s chromosomes usually differ from both chromosomes of each parental homologous pair. The transmitted chromosomes are formed during a process called meiosis, which is the cell division that results in the production of gametes (THOMAS 2004).

During meiosis, the chromosomes in the cell duplicate and homologous pairs of chromosomes come together and genetic material is exchanged through a process process called chromosomal crossover or recombination. Recombination is the process that ensures genetic diversity and linkage analysis relies on it. The recombination fraction is the rate of recombination events (odd number of crossovers) between two loci.  $\theta$  traditionally denotes the recombination fraction.  $\theta = 0.5$  between two loci means that the two loci segregate independently and the loci are said to be unlinked; the two loci are considered linked if  $\theta < 0.5$  (LANGE 2003; THOMAS 2004).

The recombination fraction defines the genetic map of the genome. The genetic map is in the unit centi-Morgan (cM) derived from the recombinations fractions through a specific mapping function ( $\theta \mapsto \text{cM}$ ). Numerous map functions have been proposed but the Haldane (HALDANE 1919) and Kosambi (KOSAMBI 1944) map functions are the most commonly used ones. As opposed to the genetic map, the physical map defines the base pair location and physical distance between loci (THOMAS 2004; OTT 1999).

Association analysis relies on the presence of linkage disequilibrium (LD). Statistically, two loci are in linkage equilibrium if the genotypes at those two loci are independently distributed in the population, otherwise they are said to be in LD; i.e., LD defines population associations between the alleles at two loci. The closer two loci are together the less likely it is that recombination between them occurs, and so they are transmitted together through many generations. Therefore, LD typically occurs over small distances and significant associations should imply proximity to the gene of interest. However, LD can occur for number of reasons, including random drift in allele frequencies in finite populations, natural selection for or against a combination of alleles (which

may or may not be linked), nonrandom mating, mutations, and founder effects. Therefore, LD can result in spurious associations (THOMAS 2004).

The strategies for gene-mapping may be classified into two broad classes, linkage and association methods. Linkage methods attempt to localize disease genes by using the co-segregation in families of the disease gene and a nearby genotyped marker. The farther away the marker is from the disease locus the more likely recombination is to occur and so the likelihood of co-segregation drops with distance. Association mapping on the other hand tries look for associations that could reflect linkage disequilibrium between a causal disease allele and the marker allele. Association mapping has the most power when the associated marker allele is the same allele in all affected individuals. In contrast, linkage mapping only requires the marker loci to co-segregate (with high probability) with the disease locus – the actual risk allele can be different in different lines of descent. Since LD generally operates over small distances, association mapping may be perceived as the more powerful approach for localizing disease genes. However, due to confounding factors, such as admixture or founder effects, LD (or association) can occur between markers spaced far apart and even on different chromosomes resulting in spurious associations. Presence of strong linkage, on the other hand, is directly related to proximity with the disease locus (BHATTACHARJEE 2008).

Association and linkage are well-powered under different conditions. Linkage generally has better power than association to detect rarer variants of larger effects, while association has more power than linkage to detect common variants of smaller effects (CLERGET-DARPOUX and ELSTON 2007; BOURGAIN *et al.* 2007). Intuitively, it makes sense that linkage methods lose power as the risk variant becomes more common, as then the risk variant is more likely to enter the pedigree more than once.

The underlying principle of all statistical genetic mapping methods is that people who have similar traits should have higher than expected levels of sharing of genetic material near the genes that influence those traits. The genetic sharing among family members is referred to as identity by descent (IBD). Sharing of two alleles are IBD occurs if one is a physical copy of the other (e.g. parent-offspring sharing) or if they are both physical copies of the same ancestral allele (e.g. sibling sharing). IBD is distinguished from two alleles being identical by state (IBS), which simply means the alleles are the same but not necessarily IBD. Any two unrelated individuals share zero alleles IBD. Siblings can share 0, 1, or 2 alleles IBD with probabilities 1/4, 1/2, and 1/4, respectively. A child shares one allele IBD with each of its parents (LANGE 2003; SZATKIEWICZ 2004).

To do genetic mapping, the IBD sharing at each locus is estimated from marker data. In two-point linkage analysis, the IBD sharing is estimated at a single marker and then the likelihood that the trait locus is at the marker locus (or that there is no recombination between them) is evaluated. Estimates of IBD sharing at each locus can be improved by incorporating marker information from all the typed markers along the chromosome, which is called multipoint analysis. Multipoint analysis is based on a probability model for recombination between markers along the chromosome and require external information about the genetic map ([OTT 1999](#)).

When mapping complex traits, researchers face many challenges. Incomplete penetrance, phenocopies, genetic heterogeneity, or polygenic inheritance may for example cause major difficulties. Not only are those factors often present but each one is acting to an unknown and different degree. The penetrance is the probability of disease given the mutant disease genotype. If the penetrance is incomplete (i.e.  $< 1$ ) then an individual who carries the risk genotype of interest may be unaffected. On the other hand individuals who don't carry the risk genotype but are nevertheless affected by the disease are said to be phenocopies. Genetic heterogeneity in the context of monogenic diseases arises from either allelic heterogeneity, when different mutations at the same locus cause the same phenotype, or from locus heterogeneity when mutations at different loci cause the same phenotype. Allelic heterogeneity is generally not of concern in linkage analysis but can cause severe problems for association analysis. Locus heterogeneity complicates linkage analysis and association analysis. Polygenic inheritance is typically defined by the phenotype requiring the simultaneous presence of mutations in multiple genes. Isolating pieces of the polygenic inheritance puzzle can be particularly challenging as each single gene may have only a very small effect on the overall disease risk. Moreover, different non-overlapping subgroups may be involved in different individuals ([STRAUCH \*et al.\* 2003](#)).

## 2.0 BACKGROUND ON THE AGE-RELATED MACULOPATHY PROJECT

### 2.1 EPIDEMIOLOGY AND GENETICS OF AGE-RELATED MACULOPATHY

Age-related macular degeneration (AMD) is the leading cause of irreparable vision loss in the elderly in industrialized countries and the third leading cause of blindness world-wide ([FRIEDMAN \*et al.\* 2004](#)). Age-related maculopathy (ARM) refers to the full spectrum of the disease from mild to the advanced forms (exudative and atrophic AMD) ([GORIN 2007](#)). The disease is a progressive, chronic, and degenerative condition of the eye and primarily, but not exclusively, affects the central macular region of the retina (Figure 2.1). Thus ARM results in blurred vision and loss of central vision (Figure 2.2) making daily activities challenging and clearly affects quality of life.

ARM is a complex disease caused by a combination of genetic predisposition and environmental factors ([SEDDON and CHEN 2004](#)). Old age and cigarette smoking are well-established risk factors for ARM. Among other factors, exposure to sunlight, increased body mass index, and hypertension have been found associated with increased risk of ARM. Dietary behaviors, such as antioxidant, zinc, vitamin D and E, and omega-3 fatty acid consumption, have been found to be associated with decreased risk of ARM ([HADDAD \*et al.\* 2006](#)).

The genetic basis for ARM was established by twin and familial aggregation studies. In a recent twin study, the roles of environment and heredity were investigated by comparing ARM concordance rates between monozygotic and dizygotic twins. Genetic factors explained 46% to 71% of the variation in the overall severity, unique environmental exposures (residuals) accounted for 19% to 37%, but shared environmental exposures were not statistically significant and accounted for 5% to 17% of the variation ([SEDDON \*et al.\* 2005](#)). Familial aggregation studies have also found that the prevalence of ARM among relatives of cases is higher than among relatives of controls.

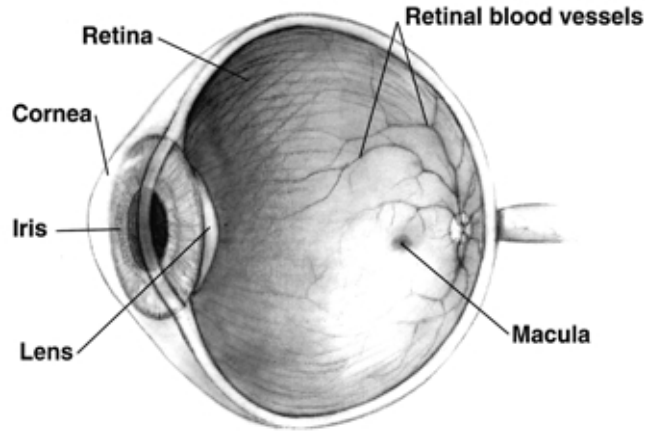


Figure 2.1: Anatomy of the eye. The retina is a multi-layered sensory tissue that lines the back of the eye. The macula is a small area in the center of the retina at the back of the eye. The macula is responsible for sharp, clear central vision and the ability to perceive color. Photo credit: The National Eye Institute.

A number of linkage studies as well as candidate gene association studies have been done in the search for ARM susceptibility loci and in 2005 the first genome-wide association (GWA) study (KLEIN *et al.* 2005) for complex disease was done for ARM. The linkage studies consistently found linkage signals at chromosomes 1q and 10q (FISHER *et al.* 2005). Other regions identified through linkage studies are on chromosomes 2p, 3p, 4q, 12q and 16q (FISHER *et al.* 2005). The earlier (prior to 2005) candidate gene association studies most consistently suggested that the *APOE* gene might harbor protective and risk alleles for ARM. However, numerous studies have failed to demonstrate an association between *APOE* variants and ARM. Other candidate genes with conflicting evidence of association are *ABCR*, *CX3CR1*, *HLA*, *VEGF*, *ELOVL4*, and *FBLN5* (HADDAD *et al.* 2006).

In 2005, three studies, including the first GWA study, identified the complement factor H (*CFH*) gene, a susceptibility gene under the linkage peak on 1q (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005) and two studies identified a cluster of three genes, *PLEKHA1*, *LOC387715*, and *HTRA1*, under the linkage peak on 10q (JAKOBSDOTTIR *et al.* 2005; RIVERA *et al.* 2005). Note that the first part of my ARM project resulted in the publication of one of the (and the first) 10q studies (JAKOBSDOTTIR *et al.* 2005). The association of the *Y402H* variant in

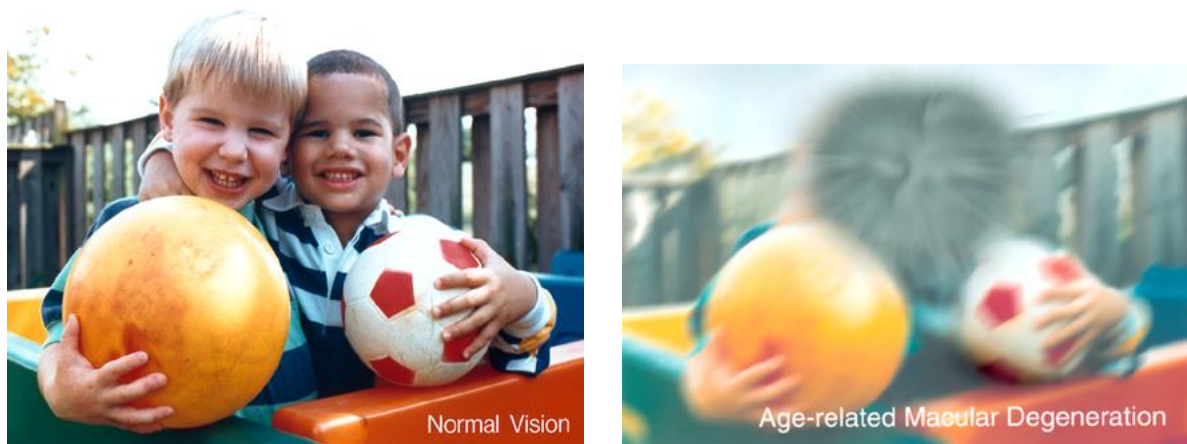


Figure 2.2: Vision loss due to ARM. These pictures shows a central blind spot with the surrounding area appearing fairly clear. Unfortunately this does not mean that if a person just moves the blind spot out of the way, everything will be clear. The peripheral vision provides vision under low light conditions and is not capable of the same sharp and clear vision as the macula. Photo credit: The National Eye Institute.

*CFH* and *S69A* variant in *LOC387715* have been widely replicated and negative results are rare. In fact now *ARMS1* and *ARMS2* (for ARM susceptibility 1 and 2) are also the official names for the *CFH* and *LOC387715* genes, respectively.

The *CFH* gene is a regulator of the complement system and since its discovery, a number of other complement genes (*CFHR1*, *CFHR3*, *CFB*, *C2*, and *C3*) have been found associated with ARM. The function of *LOC387715* is not yet understood and so its discovery has not yet resulted in as fruitful studies as the *CFH* discovery has (HADDAD *et al.* 2006).

The first component of my ARM project (chapter 3) presents our paper on the fine-mapping efforts under previously identified linkage peak on chromosomes 1q31 and 10q26. We successfully replicate the discovery of the *CFH* gene on 1q31 and identify a novel locus, harboring three closely linked genes (*PLEKHA1*, *LOC387715*, and *HTRA1*), on 10q26. Both discoveries have now been widely replicated and in the next part of the ARM project (chapter 4) I present our paper on the meta-analysis of 11 *CFH* and 5 *LOC387715* data sets. There we also replicate these findings in two independent case-control cohorts, including the first cohort, where ARM status was not



a factor in the ascertainment of study participants. In our third paper on ARM (chapter 5) we replicate discoveries of new complement related loci (*C2* and *CFB*) on chromosome 19p13 as well as developing crude classification models based on SNPs from the three loci (*CFH*, *LOC387715*, and *C2*). Finally, the last component on my ARM project (chapter 6) specifically focuses on applying statistical techniques from the diagnostic medicine literature to our ARM data as well as providing a review for the genetics community. We comment on the importance of understanding the difference and similarities between different goals of genetic studies, namely improving the etiological understanding or finding variants that discriminate well between cases and controls. This work is quite timely and particularly relevant today when there has been explosion in the availability of direct-to-consumer DNA tests.

## 2.2 STATISTICAL METHODS USED IN THE PROJECT

In this section I explain the basic principles behind most important statistical techniques used in the ARM project component of this dissertation.

### 2.2.1 Linkage analysis

Linkage methods attempt to localize disease genes by using the co-segregation of the gene and a nearby genotyped marker in families. The farther away the marker is from the disease locus the more likely recombination is to occur between them and so the likelihood of co-segregation drops with distance. Within the context of linkage analysis of qualitative traits there are two approaches to gene mapping; parametric or model-based analysis and the non-parametric or model-free analysis, which evaluate allele-sharing between affected relative pairs.

**2.2.1.1 Parametric LOD scores and HLOD scores** Parametric linkage analysis, also known as LOD score analysis, dates back to [MORTON \(1955\)](#) and originated from the idea of counting recombinant and non-recombinant offsprings as systematically described by [OTT \(1999\)](#). The parameters of the genetic model need be explicitly specified prior to analysis. The penetrances for each genotype combination and the disease allele frequencies always need to be specified. If not all founders are genotyped or some were not successfully typed, the population allele frequencies

of the marker data may also be needed in the calculations of the pedigree likelihood. The LOD, or the likelihood of the odds, score is defined as

$$LOD(\theta) = \log_{10} \frac{L(\theta)}{L(0.5)}$$

where  $L(\theta)$  is the likelihood of observing the marker genotypes and disease phenotypes in a family given the disease model and  $\theta$  is the recombination fraction between the unmeasured disease locus and the marker locus. The LOD score for multiple families is simply the sum of individual family-specific LOD scores. The  $LOD(\theta)$  is maximized as function of  $\theta$  (OTT 1999; THOMAS 2004).

In complex disease it is not unrealistic for some proportion ( $\alpha$ ) of families to be linked and some proportion unlinked; in the unlinked families the true recombination fraction is  $\theta = 0.5$  and in the linked families the true recombination fraction is equal and  $< 0.5$ . Then the heterogeneity LOD (HLOD) score is calculated over all families  $i$  per formula

$$HLOD(\theta, \alpha) = \sum_i [\log_{10}(\alpha L_i(\theta) + 1 - \alpha)]$$

where  $L_i(\theta)$  is now the conditional likelihood of family  $i$  given that this is a linked family.  $HLOD(\theta, \alpha)$  is maximized as function of  $\theta$  and  $\alpha$  (OTT 1983).

A single marker can be uninformative or partially uninformative for linkage if the observed genotypes give no or limited information on the IBD sharing in the pedigree. However, there may be other markers nearby that are either informative or partially informative. In multipoint linkage analysis the underlying genetic map is used to infer the IBD status at locations along the whole chromosome using information from all typed markers on that chromosome. The multipoint IBD estimation is done via a sophisticated statistical technique called hidden Markov models (SIEGMUND and YAKIR 2007).

**2.2.1.2 Model-free LOD scores** Parametric linkage analysis can be sensitive to misspecification of the inheritance model and so non-parametric linkage analysis is an important alternative for all but the simplest of traits (CLERGET-DARPOUX *et al.* 1986). Non-parametric linkage methods are based on allele sharing and use of scoring function,  $S$ . A locus that is linked to a disease-susceptibility gene is expected to show higher number of alleles IBD among the affecteds, relative to null of no linkage. Testing for linkage becomes testing for excess sharing and  $S$  will put increased weight on affected relative pairs who share alleles IBD (SHIH and WHITTEMORE 2001). A number of scoring functions have been proposed, including  $S_{pairs}$  and  $S_{all}$ .  $S_{pairs}$  is simply the number of

pairs of alleles that are IBD from distinct affected relatives combined in a pairwise manner. The  $S_{all}$  score function considers all affected relatives simultaneously and more heavily weights three or more affected relatives sharing an allele IBD (WHITTEMORE and HALPERN 1994). The NPL score (KRUGLYAK *et al.* 1996) for  $n$  pedigrees, each equally weighted and with score  $S_i$ , is  $\sum \bar{Z}_i$  where  $\bar{Z}_i = (S_i - \mu_i)/\sigma_i$  are the standardized scores.

KONG and COX (1997) extended the model-free NPL score to a semi-parametric LOD score, that is a function of one parameter  $\delta$  and defines the likelihood, which is the basis of the Kong-Cox  $S_{all}$  LOD score used in our analysis, as

$$\ell(\delta) = C + \sum \ln(1 + \delta \bar{Z}_i)$$

where  $C$  is a constant that depends on the data but not on  $\delta$  and  $\delta = 0$  corresponds to the null hypothesis and  $\delta \geq 0$  corresponds to the alternative hypothesis of excess sharing.

## 2.2.2 Association analysis

Association mapping tries look for associations that could reflect linkage disequilibrium between a causal disease allele and the marker allele. Association mapping has the most power when the associated risk allele is the same in all individuals. Since LD generally operates over small distances, association mapping may be perceived as a powerful approach to localize disease genes.

**2.2.2.1  $\chi^2$  and Fisher's exact tests and logistic regression** Contingency tables, which are composed of  $R \times C$  rows and columns, are an appropriate way of displaying categorical data; for example for a binary trait one may display the genotypic distribution among cases and controls in a  $2 \times 3$  contingency table. In the contingency table approach, we fix the column and row total on  $(R - 1) \times (C - 1)$  degrees of freedom. Given the row, column, and grand totals we can calculate the expected table and perform a  $\chi^2$  test of the deviation of the observed vs. the expected table and estimate odds ratios. The  $\chi^2$  test, however, rests on the assumption of approximate normality of binomial proportions and therefore may not always be appropriate especially when samples sizes are small. For smaller contingency tables like those of allelic ( $2 \times 2$ ) and genotypic tests ( $2 \times 3$ ) the Fisher's exact test may be used instead where all tables possible for the observed row and column sums need to be evaluated (ROSNER 2000).

In the logistic model the binary outcome ( $y$ ) is modeled as

$$\text{logit}(p) = \ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \dots + \beta_n x_n$$

where  $p$  is the probability of  $y$  being diseased (coded as 1), and the  $x_i$ 's are the exposure variables, which in our case are the genotypes and possible covariates; typically for ARM the covariates are age, sex, cigarette smoking and possibly genotypes at other loci than the primary loci. The advantage of using the logistic model over the contingency table approach is that it allows for easy adjustment of continuous covariates and estimation of odds ratios adjusted for those covariates. If  $x_1$  is our genotype variable then the odds ratio relating the genotype to the phenotype after controlling for all the covariates in the model is estimated with  $e^{\hat{\beta}_1}$ ; the logistic regression approach can also be applied without the covariates and will then give the same results as the contingency table approach and  $\chi^2$  tests (ROSNER 2000).

**2.2.2.2 CCREL and MQLS** When testing for association using related individuals the correlation between individuals due to their relationships needs to be modeled appropriately to avoid excess of false-positive findings. The CCREL and M<sub>QLS</sub> tests are both test statistics designed for collection of related and unrelated individuals allowing families ascertained for linkage to be appropriately pooled with unrelated cases and controls ascertained for association studies (BROWNING *et al.* 2005; THORNTON and MCPEEK 2007). The CCREL methods, however, advises against using familial controls, while the M<sub>QLS</sub> test is applicable to association testing in completely general combinations of family and case-control designs and allows cases to be related to controls. Furthermore, M<sub>QLS</sub> distinguishes between unaffected controls and controls of unknown phenotype and makes use of phenotype data about relatives who have missing genotype data at a given SNP. Those are the reasons that, even though M<sub>QLS</sub> has only been developed for allelic tests while the CCREL method has been developed for allelic, genotypic, and haplotypic tests, we now prefer the M<sub>QLS</sub> over the CCREL test.

The CCREL tests accounts correlations between related individuals by calculating a weight for each person. “The weights are used in constructing a composite likelihood, which is maximized iteratively to form likelihood ratio tests for single-marker and haplotypic associations” (BROWNING *et al.* 2005). On the other hand, the CCREL genotypic test is a standard  $2 \times 3 \chi^2$  contingency table test of weighted genotype counts (BROWNING *et al.* 2005).

The  $M_{QLS}$  is derived as a quasi-likelihood score test of a mean model for the expected allele frequencies. The mean model of  $M_{QLS}$  is an improvement over a mean model that puts similar weights on individuals as the CREL test by explicitly taking into account the fact that affected individuals who have affected relatives (whether genotyped or not) have a higher expected frequency of the susceptibility alleles for a genetic trait than do individuals with no affected relatives. The mean model of  $M_{QLS}$  is therefore a function of both the relationships and phenotypes (THORNTON and MCPEEK 2007).

### 2.2.3 Joint linkage and association

Linkage analysis tends to identify broad genomic susceptibility regions that often contain many candidate genes. For many complex diseases, before the era of genome-wide association studies, the study design of choice was linkage analysis followed by fine-mapping in identified regions of linkage (THOMAS 2004). There has been interest in the literature to identify polymorphisms that may be responsible for an observed linkage peak.

**2.2.3.1 GIST** We used the genotype IBD sharing test (GIST) (Li *et al.* 2004) to explore which alleles of some SNPs under our linkage peaks of chromosomes 1q31 and 10q26 for ARM accounts in part the observed linkage signals. The GIST performs weighted analysis of nonparametric linkage (NPL) scores, in which each family is weighted according to the genotype distribution of members of the pedigree. The correlation between the family weight variable and the NPL score forms the basis of the test statistic. The weights are unbiased under the null hypothesis of no disease-marker association in sibship data. Therefore the GIST is currently applicable only to affected sib pair families and we split our families into their component nuclear families before computing the NPL scores and weights. Since the majority of our ARM families are sibships, this was unlikely to cause problems with the analysis (JAKOBSDOTTIR *et al.* 2005).

### 2.2.4 Meta-analysis

Meta-analysis is powerful tool to pool summary data, such as odds ratios, from many studies into a single estimate, which often has narrower confidence intervals than any single study. The methods

for performing meta-analysis fall into two broad classes, fixed effects modeling and random effects modeling (VAN HOUWELINGEN *et al.* 2002; BROWN 2006)

**2.2.4.1 Fixed effects modeling** When assuming the between-study variation is due to chance, a fixed-effects model may be employed. Under the fixed-effect model, the maximum likelihood estimator of the pooled OR is an average of individual estimates, weighted by the inverse of their variances, and the variance of the pooled OR is estimated by the inverse of the sum of individual weights (VAN HOUWELINGEN *et al.* 2002).

**2.2.4.2 Random effects modeling** Rather than assuming the between-study variation is due to chance, heterogeneity can be modeled by incorporating a label for the study variable as a random covariate in the mixed model. The idea behind this approach is to assume that the studies at hand are random sample of studies, and thus that the sample of studies cannot measure all the factors contributing to the variance of the pooled OR (VAN HOUWELINGEN *et al.* 2002; BROWN 2006; DERSIMONIAN and LAIRD 1986).

## 2.2.5 Model building

I broadly split model building into two parts. First many different logistic regression models may be fit to any particular data set in order to investigate which factors are important and how they act together without asking whether the model can be used as a classification model. Secondly, classification models or screening or diagnostic tests that assign individuals into categories of testing affected or unaffected may be developed.

**2.2.5.1 Logistic regression** For ARM two very strongly associated variants were discovered and it was of particular interest to shed some light on their joint contribution to the phenotype. Various two-locus logistic regression models can be fit to find a most parsimonious model (which for example may be a model of additive effects of both loci with or without interaction) (NORTH *et al.* 2005). Nested models can be compared statistically using a likelihood ratio test but non-nested models can be compared using the Akaike's information criterion (AIC), which penalizes the models as number of parameters increases.

**2.2.5.2 GMDR** The generalized multifactor dimensionality reduction (GMDR) method (LOU *et al.* 2007) is an extension of the original MDR method (RITCHIE *et al.* 2001). Both methods attempt to classify multi-locus genotypes into “high” and “low” risk groups. The MDR methods simply uses the ratio of the number of cases to the number of controls (within the multi-factor genotype) as the basis for labeling each multi-factor genotype. Therefore the method is only applicable to data with equal number of cases and controls. In the GMDR method individual-level score statistics for the genotype effect (unadjusted or adjusted for covariates) are defined based on logistic regression model:

$$S_i = \sum_j \frac{x_{ij}(y_i - \hat{p}_i)}{\sqrt{\text{Var}(\hat{y}_i)}}$$

where  $y_i$  is the phenotype of individual  $i$ ,  $E[y_i] = p_i$ , and  $x_{ij}$  is the genotype of marker  $j$  for individual  $i$ . Each multi-locus genotype is labeled “high-risk” or “low-risk” based on the average score over individuals harboring the specific multi-locus genotype, and on the preassigned threshold, which typically is set to zero as the scores are both standardized and normalized. Both methods use cross-validation to guard against over-fitting.

**2.2.5.3 ROC curves** As is the case with multi-locus genotypes a screening test may provide several categories or be reported as a continuous variable rather than simply “test positive” or “test-negative”. When evaluating such a test it may be inadequate to only estimate the true positive (TPF) and false positive fractions (FPF) corresponding to one threshold. The ROC curves (i.e., the receiver operating characteristic curves) plot all pairs of TPF and FPF on single plot and estimate the area under the curve (AUC). The larger the AUC the better the test is overall (ZHOU *et al.* 2002).

## 2.2.6 Practical issues

**2.2.6.1 Samples vs. individuals** The most important practical issue in data analysis is the data preparation. Even though I have worked with the ARM data for almost 5 years and we now have a working database for the genotype data, it still takes considerable amount of time and script writing to prepare the data for analysis. Once the data files are all set up then the analysis are relatively easy to do. By preparation of the genotype data for analysis, as I am not considering analysis quality checks such as running PedCheck to check for Mendelian inconsistencies

and testing for Hardy-Weinberg equilibrium, I consider those steps part of the analysis. In the laboratory, samples of DNA are the important unit, not the individual, and in a project like ARM, which has been ongoing for over 15 years, many individuals have donated more than one sample of DNA. Hence, for different genotypings those individuals may come back from the lab with different identifiers (IDs). Therefore, merging the data with the phenotype data may become a huge issue and actually we could not find a database that could handle multiple person IDs and non-numeric IDs, so we built our own database.

**2.2.6.2 Check your files, scripts, and results** It is easy to make mistakes in any tiny step along the way from the raw data files, to the formatted data files, to the script writing, both for reformatting data files and for analysis, to pulling out the results, and so on. I cannot explain all the mistakes that I have made in detail but will focus on a few and how they were discovered and in some cases how I could have discovered them sooner.

**Reformatting of files** I once got SNP genotype data files where the SNP alleles were presented using the codes for the nitrogenous bases (A, C, G, T) instead of numeric codes (1 and 2) used before. Since our database only allows numeric allele labels and we had previously imported data for those same SNPs using numeric allele labels, I had to recode the genotypes before importing them into the database. I started by setting up a recode table that linked the A, C, G, and T to 1 and 2. I used arrays in the awk scripting language, using A, C, G, and T as keys and 1 and 2 as values. I forgot include the missing to missing (i.e. 0 to 0) link in the array. Since awk does not give any warning if ones tries to look up the value for non-existing key, then missing genotypes were set equal to the non-missing genotype of the marker in the two columns before the marker of interest. This was discovered when doing Hardy-Weinberg checks on the final data files but could have been discovered earlier, and before the data were imported into the database, by noting that after running the awk script, all but the first marker had no missing genotypes.

**Your scripts for analysis** The genotypes are usually listed in two side-by-side columns for each marker. I once had a file with number of markers and created a allele tables in R by looping through the SNPs. However, I mistakenly set my starting column as the one in front of the first marker (i.e. the column containing sex). This meant that the association tests performed on the allele tables were wrong: the first marker's first allele was the sex code and the second allele was the first allele, the second marker's first allele was the first marker's second allele and so on. I noticed this by noting that the allele tables showed an excess of half-typed individuals.



**Programs by others** Even though a program is published as a part of a peer-reviewed paper, it can still contain errors. We once used programs by others to look at various two locus models. This program, however, mixed up the names of the loci in the output. The only reason I noticed the mix-up was that we had already done single-locus analysis and those suggested that one marker was much more strongly associated with ARM than the other. We then wrote our own R script to double check our suspicion.

### 3.0 SUSCEPTIBILITY GENES FOR AGE-RELATED MACULOPATHY ON CHROMOSOME 10Q26

This chapter has been published in American Journal of Human Genetics, volume 77, issue 3, pages 389 - 407 (JAKOBSDOTTIR *et al.* 2005). The journal grants the authors rights to include the article in full in a thesis or a dissertation. No changes have been made to the published version of the paper, except that tables and figures have been renumbered and the supporting information published online is in Appendix A. My contribution to this paper was in writing all components of the paper, data analysis and script writing for method implementation.

#### 3.1 ABSTRACT

On the basis of genomewide linkage studies of families affected with age-related maculopathy (ARM), we previously identified a significant linkage peak on 10q26, which has been independently replicated by several groups. We performed a focused SNP genotyping study of our families and an additional control cohort. We identified a strong association signal overlying three genes, *PLEKHA1*, *LOC387715*, and *PRSS11*. All nonsynonymous SNPs in this critical region were genotyped, yielding a highly significant association ( $P < 0.00001$ ) between *PLEKHA1/LOC387715* and ARM. Although it is difficult to determine statistically which of these two genes is most important, SNPs in *PLEKHA1* are more likely to account for the linkage signal in this region than are SNPs in *LOC387715*; thus, this gene and its alleles are implicated as an important risk factor for ARM. We also found weaker evidence supporting the possible involvement of the *GRK5/RGS10* locus in ARM. These associations appear to be independent of the association of ARM with the Y402H allele of complement factor H, which has previously been reported as a major susceptibility factor for ARM. The combination of our analyses strongly implicates *PLEKHA1/LOC387715* as primar-

ily responsible for the evidence of linkage of ARM to the 10q26 locus and as a major contributor to ARM susceptibility. The association of either a single or a double copy of the high-risk allele within the *PLEKHA1/LOC387715* locus accounts for an odds ratio of 5.0 (95% confidence interval 3.27.9) for ARM and a population attributable risk as high as 57%.

### 3.2 INTRODUCTION

Age-related maculopathy (ARM), or age-related macular degeneration (ARMD-1 [MIM 603075]), is a leading cause of central blindness in the elderly population, and numerous studies support a strong underlying genetic component to this complex disorder. Genomewide linkage scans performed using large pedigrees, affected sib pairs, and, more recently, discordant sib pairs have identified a number of potential susceptibility loci (KLEIN *et al.* 1998; WEEKS *et al.* 2000; MAJEWSKI *et al.* 2003; SCHICK *et al.* 2003; SEDDON *et al.* 2003; ABECASIS *et al.* 2004; IYENGAR *et al.* 2004; KENEALY *et al.* 2004; SCHMIDT *et al.* 2004; WEEKS *et al.* 2004; SANTANGELO *et al.* 2005). Our genomewide linkage screen strongly implicated the 10q26 region as likely to contain an ARM gene (WEEKS *et al.* 2004); this region has also been supported by many other studies and was the top-ranked region in a recent meta-analysis (FISHER *et al.* 2005). Recently, three studies (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005) identified an allelic variant in the complement factor H gene (*CFH* [MIM 134370]) as responsible for the linkage signal seen on chromosome 1 and as the variant accounting for a significant attributable risk (AR) of ARM in both familial and sporadic cases. We and others have confirmed these findings (CONLEY *et al.* 2005; HAGEMAN *et al.* 2005; ZAREPARSI *et al.* 2005). *CFH* has previously been suspected of playing a role in ARM, as a result of the work of Hageman and Anderson (HAGEMAN and MULLINS 1999; JOHNSON *et al.* 2000; JOHNSON *et al.* 2001; MULLINS *et al.* 2000; HAGEMAN *et al.* 2001), who have shown that the subretinal deposits (drusen) that are observed in many patients with ARM contain complement factors. However, until other genes that contribute to ARM are identified, *CFH* remains an isolated piece of the puzzle, implicating the alternative pathway and inflammation as part of the ARM pathogenesis but failing to fully account for the unique pathology that is observed in the eye.

We have expanded our family linkage studies and have also undertaken a case-control association study, using a high-density SNP panel in two regions of linkage on 1q31 and 10q26 that

we had previously reported. Our SNP linkage and association results for chromosome 1q31 yielded the same findings as others, confirming that the peak of linkage and the strongest associations with ARM were localized over the *CFH* gene. We have analyzed both our family data and the case-control data on chromosome 10q26 to identify the next major ARM susceptibility-related gene.

### 3.3 MATERIAL AND METHODS

#### 3.3.1 Families and Case-Control Cohort

A total of 612 ARM-affected families and 184 unrelated controls were sent to the Center for Inherited Disease Research (CIDR) for genotyping. Because of possible population substructure, we restricted our analysis to the subset of data from white subjects; we were not able to analyze the set of data from nonwhites separately, because it was too small. The white subset had 594 ARM-affected families, containing 1,443 genotyped individuals, and 179 unrelated controls. The white families contained 430 genotyped affected sib pairs, 38 genotyped affected avuncular pairs, and 52 genotyped affected first-cousin pairs.

A total of 323 white families, 117 unrelated controls, and 196 unrelated cases were also genotyped locally for additional SNPs. The local subset contained 824 genotyped individuals, 298 genotyped affected sib pairs, 23 genotyped affected avuncular pairs, and 38 genotyped affected first-cousin pairs. We used PedStats from the Merlin package ([ABECASIS \*et al.\* 2002](#)) to easily get summary counts on the family data.

#### 3.3.2 Affection-Status Models

We have defined three classification models (types A, B, and C) for the severity of ARM status ([WEEKS \*et al.\* 2004](#)). For simplicity, we have restricted our attention here to individuals affected with “type A” ARM, our most stringent and conservative diagnosis. We used only unrelated controls who were unaffected under all three diagnostic models. Unaffected individuals were those for whom eye-care records and/or fundus photographs indicated either no evidence of any macular changes (including drusen) or a small number ( $< 10$ ) of hard drusen ( $\leq 50\mu m$  in diameter) without

Table 3.1: Distribution of Subphenotypes in Patients with Advanced ARM

Subphenotype	No. of Patients from CIDR Families <sup>a</sup>		No. of Patients from Local families <sup>a</sup>		No. of Local Unrelated Patients	
	With GA	Without GA	With GA	Without GA	With GA	Without GA
With CNV	220 (76)	187	130 (45)	106	71 (17)	59
Without CNV	108	62	57	28	40	26

NOTE.—The numbers in parentheses are the numbers of individuals with both CNV and GA who were also included in GA group (see section 3.3.2 for selection criteria) for OR and AR estimation and association tests.

<sup>a</sup> Counts are based on the set of unrelated cases generated by selecting one type A-affected person from each family (see section 3.3.13.1).

any other retinal pigment epithelial (RPE) changes. Individuals with evidence of large numbers of extramacular drusen were not coded as unaffected.

In our efforts to examine specific ARM subphenotypes, we chose to look at only patients with end-stage disease, either those with evidence of choroidal neovascular membrane (CNV) in either eye or those with geographic atrophy (GA) in either eye. There are a significant number of individuals who have been described as having both GA and CNV, though this is problematic, since, in these cases, it is often difficult to determine whether the GA is secondary to the damage from the CNV or is from the treatment given to limit the CNV growth (i.e., laser, surgery, or photodynamic therapy). Because it is often difficult to discern from photographs or records whether a person had GA in an eye prior to the development of CNV, we included the patients who had both pathologies in the CNV group. However, we allowed only a subset of this overlapping group to be included in the GA group, specifically those who reportedly had GA in one eye that did not have evidence of CNV. Table 3.1 shows the numbers of individuals in each of our three sets. This approach may have excluded a small proportion of individuals from the GA group who had asymmetric GA prior to the development of CNV in the same eye or who may have had bilateral GA but developed CNV in both eyes.

### 3.3.3 Pedigree and Genotyping Errors and Data Handling

We used the program PedCheck (O'CONNELL and WEEKS 1998) to check for Mendelian inconsistencies. Since it can be extremely difficult to determine which genotypes within small families are erroneous (Mukhopadhyay et al. 2004), we set all genotypes at each problematic marker to missing within each family containing a Mendelian inconsistency. Mega2 (MUKHOPADHYAY *et al.* 2005) was used to set up files for linkage analysis and for allele-frequency estimation by gene counting.

### 3.3.4 Allele Frequencies and Hardy-Weinberg Equilibrium

The allele frequencies used in the linkage analyses were estimated, by direct counting, from the unrelated and unaffected controls. All controls were unaffected under all three affection status models. Genotyped spouses who had no children or who had children who were not yet part of the study were combined with the controls for this study. The exact test of Hardy-Weinberg equilibrium (HWE), implemented in Mega2 (MUKHOPADHYAY *et al.* 2005), was performed on our SNPs.

We also used Mendel (version 5) (LANGE *et al.* 2001) to estimate allele frequencies directly from the family data, because Mendel properly accounts for relatedness of the subjects while estimating the allele frequencies. Since the majority of the genotyped family members were affected, these estimates were quite close to estimates obtained using our unrelated affected cases.

### 3.3.5 Genetic Map

We used linear interpolation on the Rutgers combined linkage-physical map (version 2.0) (KONG *et al.* 2004) to predict the genetic position of the SNPs that were not already present in the Rutgers map. Since the distribution of our SNPs was very dense in the regions of interest, the estimated recombination between several SNPs was zero; for these, we set the recombination to 0.000001. We obtained the physical positions for all our SNPs from the National Center for Biotechnology Information (NCBI) dbSNP database (human build 35).

### 3.3.6 LD Structure

Ignoring high linkage disequilibrium (LD) between SNPs when performing linkage analysis can result in false-positive findings (SCHAID *et al.* 2002; HUANG *et al.* 2004). Our efforts to take high

SNP-SNP LD into account included the following measures. (1) We used the H-clust method (RINALDO *et al.* 2005), which is implemented in R (R Development Core Team 2004; see R Project for Statistical Computing Web site), to determine haplotype-tagging SNPs (htSNPs) for linkage analysis. The method uses hierarchical clustering to cluster highly correlated SNPs. After the clustering, the H-clust method chooses a htSNP from each cluster; the htSNP chosen is the SNP that is most correlated with all other SNPs in the cluster. We chose to cluster the SNPs so that each SNP had a correlation coefficient ( $r^2$ )  $> 0.5$  with at least one htSNP; we used HaploView (BARRETT *et al.* 2005) to get a graphical view of SNP-SNP LD along both chromosomes, and we compared LD estimates of htSNPs with SNPs omitted by H-clust. (2) We performed haplotype-based association analyses using two- and three-SNP moving windows (see Association Analysis section).

### 3.3.7 Linkage Analysis

**3.3.7.1 Two-point analysis** As in our previous study (WEEKS *et al.* 2004), we computed LOD scores under a single simple dominant model (with disease-allele frequency of 0.0001 and penetrance vector of [0.01, 0.90, 0.90]). Because of the complexities and late onset of the ARM phenotype, only two disease phenotypes were used: “affected under model A” (i.e., “type A-affected”) and “unknown”. Parametric LOD scores were computed under heterogeneity (HLOD), whereas model-free LOD scores were computed with the linear  $S_{all}$  statistic. Both scores were computed using Allegro (GUDBJARTSSON *et al.* 2000).

**3.3.7.2 Multipoint analysis ignoring LD** Since intermarker distances are often very small, LD between SNPs can be high and thus violate the assumption of no LD made by most linkage analysis programs. Multipoint analyses ignoring LD were performed using Allegro (GUDBJARTSSON *et al.* 2000). Both HLOD scores and  $S_{all}$  statistics were computed. Our main goal in estimating the multipoint linkage curve without properly accounting for LD was not to predict the position of ARM-associated loci but to compare the results with those from analyses in which LD was taken into account.

**3.3.7.3 Multipoint analysis using htSNPs** When only htSNPs were used for LOD score calculation, the number of SNPs decreased from 679 to 533 on chromosome 1 and from 196 to

159 on chromosome 10. Multipoint linkage analyses were done as described elsewhere ([WEEKS \*et al.\* 2004](#)). The SNPs that were omitted fit well into the SNP-SNP LD structure estimated by HaploView ([BARRETT \*et al.\* 2005](#)).

### 3.3.8 Association Analysis

To incorporate all cases from the families, we used the new CCREL program ([BROWNING \*et al.\* 2005](#)), which permits testing for association with the use of related cases and unrelated controls simultaneously. CCREL was used to analyze SNPs under the linkage peak on chromosomes 1 and 10, to test for association. The CCREL test accounts for biologically related subjects by calculating the effective number of cases and controls. For these analyses, type A-affected family members were assigned the phenotype “affected”, unrelated controls were assigned the phenotype “normal”, and family members that were not affected with type A ARM were assigned the phenotype “unknown”. (The CCREL approach has not yet been extended to permit the simultaneous use of both related cases and related controls.) The effective number of controls for each SNP used for association testing is therefore the number of controls genotyped for that SNP. An allelic test, a haplotype test with a two-SNP sliding window, a haplotype test with a three-SNP sliding window, and a genotype test were performed. We used the CCREL R package for analysis, as provided by ([BROWNING \*et al.\* 2005](#)).

### 3.3.9 GIST Analysis

To explore which allele/SNP contributes the most to the linkage signal, we performed the genotype-identity by descent (IBD) sharing test (GIST) using our locally genotyped SNPs and significant SNPs from the CCREL test that are located around the linkage peaks on chromosomes 1 and 10. GIST determines whether an allele, or another allele in LD with it, accounts in part for the observed linkage signal ([LI \*et al.\* 2004](#)). Weights were computed for each affected sibship under three different disease models (recessive, dominant, and additive); these weights are unbiased under the null hypothesis of no disease-marker association. The correlation between the family weight variable and the nonparametric linkage (NPL) score is the basis of the test statistic. Since the GIST method is currently applicable only to affected sib pair families, we split our families into their component nuclear families before computing the NPL scores. Since we do not know the underlying disease



model, we performed tests using three different disease models (recessive, dominant, and additive) and then took the maximum result, using a  $P$  value that was adjusted for multiple testing over the three models.

### 3.3.10 Tripartite Analyses

Our analyses were performed in three sequential steps. First, we analyzed the set of data that had been genotyped at CIDR. Second, after locally genotyping eight additional SNPs in the *PLEKHA1/LOC387715/PRSS11* region on chromosome 10, we then analyzed the locally genotyped data set. Note that all of the known nonsynonymous SNPs in the region from *PLEKHA1* (MIM 607772) through *PRSS11* (MIM 602194) were investigated. Because these two data sets differ in size and composition, it is most straightforward to analyze them separately (table 3.2). Allele-frequency estimation, CCREL association testing, and GIST analysis were performed on both of these (overlapping) data sets, as described above. Third, we tested for interaction between the chromosome 1 and chromosome 10 regions and examined whether or not the risk differed as a function of the presence of either GA or CNV.

### 3.3.11 Part I: Analysis of CIDR SNPs

To identify the responsible gene on chromosome 10q26, the CIDR performed high-density custom SNP genotyping of 612 ARM-affected families and 184 unrelated controls with the use of 199 SNPs spanning 13.4 Mbp (26.7 cM), from *rs7080289* through *rs6597818* (nucleotide position: 115094788–128517320 bp), which spans our region of interest. For our analysis, we used 196 SNPs; 3 were skipped because of a lack of polymorphism in the controls (when this was checked within the family data, the less common allele was extremely rare and was only present in heterozygotes). In addition, 684 SNPs spanning 45.7 Mbp (47.1 cM) on chromosome 1q31, from *rs723858* through *rs653734* (nucleotide position: 169749920–215409007 bp), were also genotyped; 5 SNPs were skipped because of a lack of polymorphism in the controls—the less common allele was either not present or very rare and, in the family data, was only present in heterozygotes. Table A1 shows the correspondence between our allele labels and the actual alleles, and, for nonsynonymous SNPs, the amino acid change.

Table 3.2: Summary of Statistical Analyses and Sample Sizes in Parts I–III

Part and Analysis	Set of SNPs, Method, and Sample Used	Results Shown in
I:		
htSNP selection	CIDR SNPs, 179 controls	
SNP-SNP LD	CIDR SNPs, 179 controls	Figs. <a href="#">A1</a> and <a href="#">A2</a>
Linkage	CIDR SNPs and htSNPs, 594 ARM-affected families	Figs. <a href="#">3.3</a> and <a href="#">3.4</a>
Allele frequencies	Mendel v5 for 594 ARM-affected families; counting for 179 controls	Table <a href="#">3.3</a>
CCREL	CIDR SNPs, 594 ARM-affected families and 179 controls	Table <a href="#">3.3</a>
GIST	594 ARM-affected families split into 734 typed nuclear families	Table <a href="#">3.3</a>
II:		
Allele frequencies	All SNPs (CIDR and local); Mendel v5 for 323 ARM-affected families; counting for 117 controls	Table <a href="#">3.4</a>
CCREL	CIDR SNPs and local SNPs, 323 families and 117 controls	Table <a href="#">3.4</a>
GIST	323 ARM-affected families split into 407 typed nuclear families	Table <a href="#">3.4</a>
SNP-SNP LD	CIDR and local SNPs, 117 unrelated controls	Fig. <a href="#">3.2</a>
III:		
Interaction by GIST	See GIST in I and II above	Tables <a href="#">3.3</a> and <a href="#">3.4</a>
Logistic regression	CIDR SNPs, 577 cases and 179 controls	Table <a href="#">A3</a>
OR and AR	CIDR SNPs, 577 cases and 179 controls; local SNPs, 517 cases (321 familial, 196 sporadic) and 117 controls	Table <a href="#">3.5</a>
OR and AR of subtypes:		
CIDR SNPs	For CNV, 407 cases and 179 controls; for GA, 184 cases and 179 controls	Table <a href="#">A4</a>
Local SNPs	For CNV, 366 cases and 117 controls; for GA, 159 and 117 controls	Table <a href="#">A4</a>

### 3.3.12 Part II: Analysis of Locally Genotyped SNPs

We genotyped eight additional SNPs on chromosome 10 that overlie three susceptibility genes, *PLEKHA1* (*rs12258692*, *rs4405249*, and *rs1045216*), *LOC387715* (*rs10490923*, *rs2736911*, and *rs10490924*), and *PRSS11* (*rs11538141* and *rs1803403*). This genotyping effort included all of the nonsynonymous SNPs that have been reported for these genes in the NCBI databases (see fig. 3.1). As part of another study (Conley et al. 2005), we genotyped two CFH variants (*rs10922093* and *rs1061170*), which we have used here as well. Genotyping of additional SNPs under the *GRK5/RGS10* (MIM 600870/MIM 602856) locus is in process. Genotype data for *rs12258692*, *rs1803403*, and the newly characterized SNP *rs4405249* (which is 1 base 3' of *rs12258692*) were collected by sequencing (Rexagen) and were analyzed using Sequencher software (Gene Codes). Genotype data for *rs11538141*, *rs2736911*, *rs10490923*, and *rs10490924* were collected using RFLP. The primers, amplification conditions, and restriction endonucleases, where appropriate, for SNPs that were genotyped by sequencing or RFLP can be found in table A2. Genotype data for *rs1045216* were collected using a 5' exonuclease Assay-on-Demand TaqMan assay (Applied Biosystems). Amplification and genotype assignments were conducted using the ABI 7000 and SDS 2.0 software (Applied Biosystems). Two unrelated CEPH samples were genotyped for each variant and were included on each gel and in each TaqMan tray, to assure internal consistency in genotype calls. Additionally, double-masked genotyping assignments were made for each variant and were compared, and each discrepancy was addressed using raw data or resequencing. Table A1 shows the correspondence between our allele labels and the actual alleles and, for nonsynonymous SNPs, the amino acid change.

### 3.3.13 Part III: Interaction and OR Analysis

**3.3.13.1 Unrelated cases** No unrelated cases were genotyped by CIDR, but 196 unrelated cases were genotyped locally for our additional SNPs. For computation of odds ratios (ORs) and for interaction analyses (see below), we chose to generate a set of unrelated cases by drawing one type A-affected person from each family. A total of 321 locally genotyped families had at least one type A-affected person. If a family had more than one type A-affected person, we chose the person who had the most complete genotyping at the *Y402H* variant (*rs1061170*) and three CIDR SNPs representative of *CFH*, *GRK5*, and *PLEKHA1*: *rs800292* (*CFH*), *rs1537576*

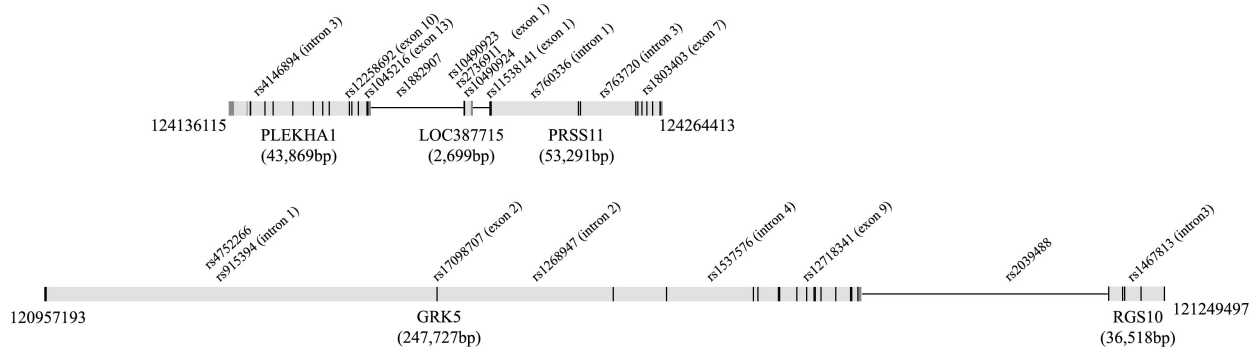


Figure 3.1: Location of CIDR SNPs and locally genotyped SNPs with respect to candidate genes. Positions, distances, and nucleotide positions along chromosome 10 are derived from NCBI Entrez Gene and dSNP databases.

(*GRK5*), and *rs4146894* (*PLEKHA1*; *rs4146894* also represents *LOC387715*, because of high LD with *rs10490924*) (see fig. 3.2). If they could not be distinguished by the number of genotyped SNPs, we chose the person who developed the disease at the youngest age, or, if more than one shared the earliest age at onset, we selected one type A-affected individual at random from those with the most SNPs genotyped and the earliest age at onset. A total of 577 CIDR families had at least one type A-affected person; 321 of these families were also genotyped locally, and the type A-affected person was the same one chosen for the local set. For the remaining 256 families, we based our selection on the same criteria described above, except that only *rs800292* (*CFH*), *rs1537576* (*GRK5*), and *rs4146894* (*PLEKHA1*) were used to identify the person with the most complete genotyping.

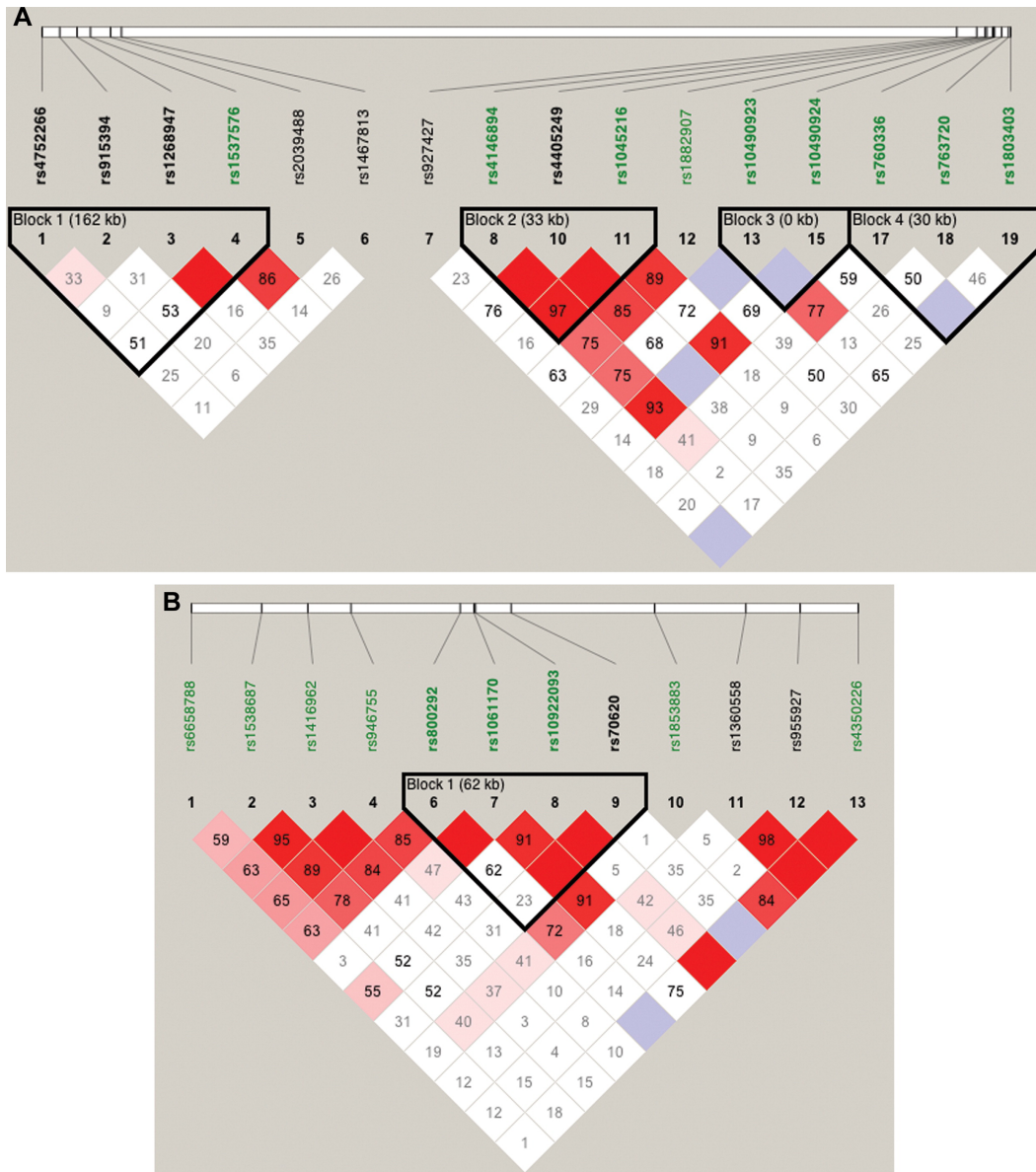


Figure 3.2: A, LD patterns in *GRK5* (Block 1), *RGS10* (SNP 6), *PLEKHA1* (Block 2), *LOC387715* (Block 3), and *PRSS11* (Block 4). B, LD patterns in *CFH* (Block 1). Squares shaded pink or red indicate significant LD between SNP pairs (bright red indicates pairwise  $D' = 1$ ), white squares indicate no evidence of significant LD, and blue squares indicate pairwise  $D' = 1$  without statistical significance. Significant SNPs from the CCREL allele test are highlighted in green (see table 3.4). Three SNPs (*rs6428352*, *rs12258692*, and *rs11538141*) were not included, because of very low heterozygosity, and one SNP (*rs2736911*) was not included, because it was uninformative. Note that the blocks were drawn to show clearly the position of the genes and do not represent haplotype blocks.

**3.3.13.2 Analysis of interaction with *CFH*** We investigated possible interaction between *CFH* on chromosome 1 and the genes on chromosome 10 by using GIST to test whether SNPs in *CFH* are associated with the linkage signal on chromosome 10 and whether SNPs on chromosome 10 are associated with the linkage signal on chromosome 1. We did this by using weights from SNPs on one chromosome and family-based NPLs from the other.

We also used logistic regression to evaluate different interaction models and to test for interaction by use of the approach described by [NORTH \*et al.\* \(2005\)](#). In this approach, many different possible models of the interactions, allowing simultaneously for additive and dominant effects at both of the loci, are fit, and relative likelihoods of the different models are compared to draw inferences about the most likely and parsimonious model. As described elsewhere ([NORTH \*et al.\* 2005](#)), the fitted models include a MEAN model, in which only the mean term is estimated; ADD1, ADD2, and ADD models, which assume an additive effect at one or both loci; DOM1, DOM2, and DOM models, which additionally incorporate dominance effects; and three further models, ADDINT, ADDDOM, and DOMINT, which allow for interactive effects (for more details, see the work of ([NORTH \*et al.\* 2005](#))). Since some pairs of these models are not nested, we compared them by using the Akaike information criteria (AIC); in this approach, the model with the lowest AIC is considered to be the best fitting and the most parsimonious. For these analyses, we used the program provided by North and colleagues (2005), after some bugs that we discovered had been fixed; we double-checked our results with our own R program. To maximize the sample size, we chose CIDR SNPs in high LD with a highly significant nonsynonymous SNP within each gene. The CIDR SNP *rs800292* was chosen to represent *rs10611710* (the *Y402H* variant of *CFH*), and the CIDR SNP *rs4146894* represented *rs1045216* in *PLEKHA1*. Similarly, we also selected a representative CIDR SNP in *GRK5*, *RGS10*, and *PRSS11*.

**3.3.13.3 Magnitude of association** We calculated crude ORs and estimated ARs for SNPs in each gene. The allele that was least frequent in the controls was considered to be the risk allele. AR was estimated using the formula  $AR = 100 \times P \times (OR - 1) / [1 + P \times (OR - 1)]$ , where OR is the OR and  $P$  is the frequency of the risk factor (genotype) in the population, as estimated from the controls. We did this by comparing type A-affected subjects with controls, comparing subjects who had CNV with controls, and comparing subjects who had GA with controls. To have the maximum possible sample size, we used different but overlapping samples for CIDR and locally typed SNPs. A total of 577 cases selected from the families and 179 unrelated cases were used for

calculating OR and AR of CIDR SNPs, but 517 cases (of those, 321 are within the 577 CIDR SNP cases) and 117 controls (all within the 179 CIDR SNP controls) were used for calculating the OR and AR on the locally genotyped SNPs.

**3.3.13.4 Multiple-testing issues** Since we have very strong evidence from previous studies that there is an ARM-susceptibility locus in the chromosome 10q26 region, the analyses performed here were aimed at estimating the location of the susceptibility gene, rather than testing a hypothesis. Multiple-testing issues are most crucial and relevant in the context of hypothesis testing. In estimation, we are simply interested in determining where the signal is strongest. In any event, any correction for multiple testing would not alter the rank order of the results. A Bonferroni correction, which does not account for any correlation between tests due to LD, for 196 tests at the 0.05 level would lead to a significance threshold of  $0.05/196=0.00026$ ; correlations due to LD would lead to a larger threshold.

## 3.4 RESULTS

Our analyses were performed in three sequential steps. First, we analyzed the set of data that had been genotyped at CIDR. Second, after locally genotyping eight additional SNPs in the *PLEKHA1/LOC387715/PRSS11* region on chromosome 10, we then analyzed the locally genotyped data set. Allele-frequency estimation, testing for HWE (table A1), CCREL association testing, and GIST testing were performed on both of these (overlapping) data sets, as described above. Third, we tested for interaction between the chromosome 1 and chromosome 10 regions and examined whether or not the risk differed as a function of the presence of GA or CNV.

### 3.4.1 Part I: Analysis of CIDR SNPs

**3.4.1.1 CIDR linkage results** The narrow peak of our  $S_{all}$  linkage curve obtained using the 159 htSNPs on chromosome 10 suggests that there might be an ARM gene in the *GRK5* region (marked 'G' in fig. 3.3, right panel); *rs1537576* in *GRK5* had a two-point  $S_{all}$  of 1.87, whereas the largest (across our whole region) two-point  $S_{all}$  of 3.86 occurred at *rs555938*, 206 kb centromeric of *GRK5*. Several elevated two-point nonparametric  $S_{all}$  LOD scores and our highest HLOD score



drew attention to the *PLEKHA1/LOC387715/PRSS11* region (marked ‘P’ in fig. 3.3). In this region, SNP *rs4146894* in *PLEKHA1* had a two-point  $S_{all}$  of 3.34 and the highest two-point HLOD of 2.66, whereas SNPs *rs760336* and *rs763720* in *PRSS11* had two-point  $S_{all}$  values of 2.69 and 2.23, respectively. However, the 1–unit support interval is large (10.06 cM) (fig. 3.3), and so localization from the linkage analyses alone is rather imprecise.

We also explored the effect of failing to take SNP-SNP LD into account, by comparing the multipoint scores computed using all SNPs (fig. 3.3, left panel) with those computed using only the htSNPs (fig. 3.3, right panel). Two of the peaks found using all SNPs (referred to as “false peaks”; marked ‘F’ in fig. 3.3, left panel) almost vanish completely when using only htSNPs; interestingly, these two peaks lie within haplotype blocks (Figure 4 and Figure 4), whereas the LD around our highest multi- and two-point LOD scores is low (fig. 4C), indicating the importance of taking LD into account when performing linkage analysis.

Our linkage results on chromosome 1 gave three peaks with  $S_{all} > 2$ , and only one of those peaks was observed when we restricted our analysis to htSNPs (fig. 3.4). This remaining peak overlies *CFH* and includes two SNPs with very high two-point  $S_{all}$  and HLOD scores: *rs800292*, a nonsynonymous SNP in *CFH*, had an  $S_{all}$  of 1.53 and an HLOD of 2.11, whereas SNP *rs1853883*, 165 kb telomeric of *CFH*, had an  $S_{all}$  of 4.06 and an HLOD of 3.49. These results strongly support earlier findings of *CFH*s involvement in ARM (CONLEY *et al.* 2005; EDWARDS *et al.* 2005; HAGEMAN *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005; ZAREPARSI *et al.* 2005). The vanishing peaks (marked ‘F’ in fig. 3.4, left panel) that we saw when we used all of our SNPs in the linkage analysis are located within strong haplotype blocks (figs. A2A and A2B), whereas the LD under the *CFH* peak is relatively low (fig. A2C).

**3.4.1.2 CIDR association results** For finer localization than can be obtained by linkage, we turned to association analyses (which were very successful in discovering *CFH* on chromosome 1). Here, we performed association analyses using the CCREL approach (BROWNING *et al.* 2005), which permitted the simultaneous use of our unrelated controls and all of our related familial cases by appropriately adjusting for the relatedness of the cases. In the CIDR sample on chromosome 10, within our linkage peak, we found a cluster of four adjacent SNPs with very small  $P$  values (*rs4146894*, *rs1882907*, *rs760336*, and *rs763720*) that overlies three genes: *PLEKHA1*, *LOC387715*, and *PRSS11*. Our strongest CCREL results on chromosome 10 were for



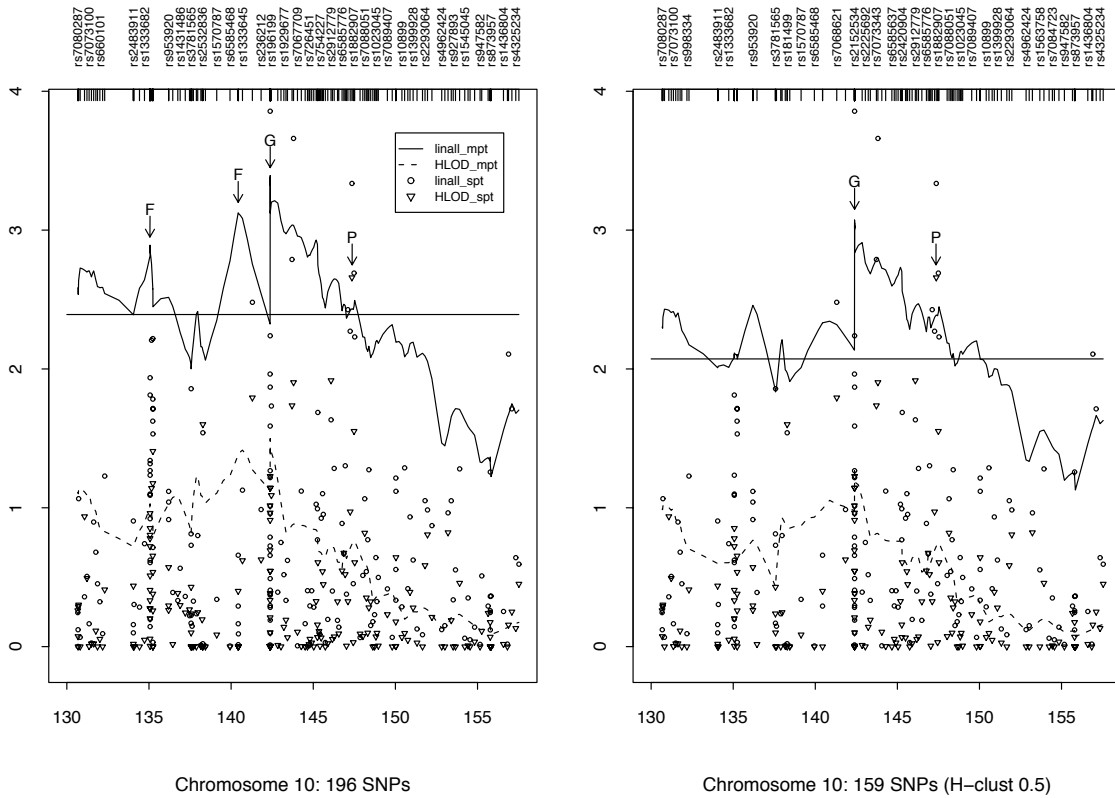


Figure 3.3: Two-point (2pt) and multipoint (mpt) linkage results on chromosome 10. The panel on the left summarizes the results when all SNPs were used for analysis. The panel on the right summarizes the results when only htSNPs were used. The peaks marked ‘F’ represent likely false peaks due to high SNP-SNP LD, whereas the peaks marked ‘G’ and ‘P’ correspond to the loci containing *GRK5* and *PLEKHA1*, respectively. The horizontal lines indicate the 1–unit support interval of multipoint  $S_{all}$  (i.e., maximum  $S_{all} - 1$ ).

SNP *rs4146894* in *PLEKHA1* (table 3.3). The moving-window haplotype analyses using three SNPs at a time resulted in very small  $P$  values across the whole *PLEKHA1* to *PRSS11* region (table 3.3). The association testing also generated some moderately small  $P$  values in the *GRK5* region, which is where our highest evidence of linkage occurred.

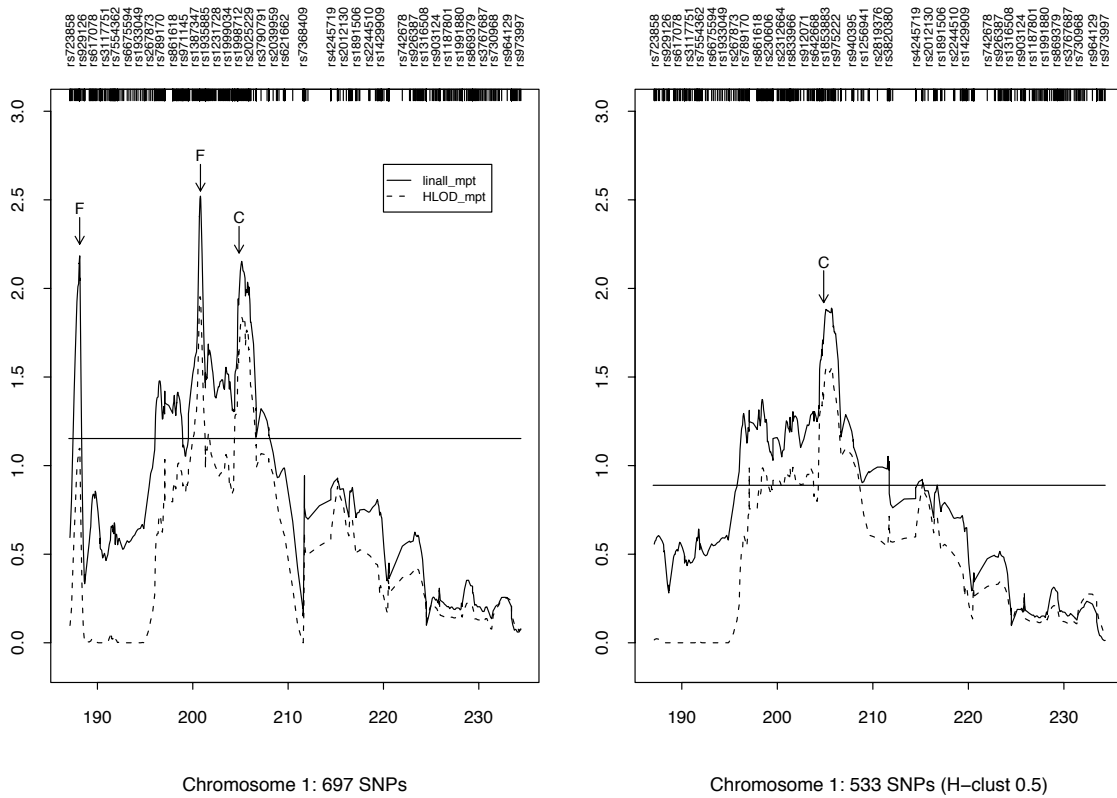


Figure 3.4: Two-point (2pt) and multipoint (mpt) linkage results on chromosome 1. The panel on the left summarizes the results when all SNPs were used for analysis. The panel on the right summarizes the results when only htSNPs were used. The peaks marked ‘F’ represent likely false peaks due to high SNP-SNP LD, whereas the peak marked ‘C’ corresponds to the *CFH* gene. The horizontal lines indicate the 1–unit support interval of multipoint  $S_{all}$  (i.e., maximum  $S_{all}$  over  $CFH-1$ ).

We performed the CCREL on 56 SNPs spanning the linkage peak on chromosome 1 and found two highly significant SNPs (*rs800292* and *rs1853883*) that overlie *CFH* (table 3.3). The moving-window haplotype analyses, performed using two and three SNPs at a time, resulted in

Table 3.3: CCREL, GIST, and Allele-Frequency Estimation for Families and Controls Typed at CIDR

SNP	Gene	P value for Test							
		Frequency in		Allele Test	Moving-Window Haplotype Test		Genotype Test	GIST	
		Families ( $n = 594$ )	Controls ( $n = 179$ )		With 2 SNPs	With 3 SNPs		NPL 10	NPL1
<i>rs6658788</i>		0.460	0.489	0.37312	0.01616	0.00778	0.44415	0.106	0.055
<i>rs1538687</i>		0.234	0.307	0.00178	0.00206	0.00674	0.0054	0.781	0.129
<i>rs1416962</i>		0.321	0.352	0.16378	0.39256	0.4157	0.38009	0.566	0.019
<i>rs946755</i>		0.317	0.344	0.20073	0.20147	<0.00001	0.37434	0.513	0.012
<i>rs6428352</i>		0.001	0.003	...	<0.00001	<0.00001	...	...	...
<i>rs800292</i>	<i>CFH</i>	0.132	0.232	<0.00001	<0.00001	<0.00001	<0.00001	0.437	0.001
<i>rs70620</i>	<i>CFH</i>	0.147	0.173	0.15602	<0.00001	<0.00001	0.33122	0.893	0.333
<i>rs1853883</i>		0.630	0.489	<0.00001	<0.00001	<0.00001	<0.00001	0.521	<0.001
<i>rs1360558</i>		0.425	0.397	0.34842	0.60377	0.01118	0.63012	0.183	0.296
<i>rs955927</i>		0.416	0.391	0.36201	0.00833	...	0.65613	0.065	0.145
<i>rs4350226</i>		0.055	0.095	0.00182	...	...	0.00183	0.171	0.242
<i>rs4752266</i>	<i>GRK5</i>	0.220	0.223	0.84131	0.23223	0.28973	0.03802	0.088	0.475
<i>rs915394</i>	<i>GRK5</i>	0.214	0.187	0.15214	0.19235	0.00309	0.35594	0.028	0.643
<i>rs1268947</i>	<i>GRK5</i>	0.112	0.117	0.97426	0.00969	0.01031	0.97976	0.052	0.345
<i>rs1537576</i>	<i>GRK5</i>	0.507	0.433	0.01881	0.01354	0.03257	0.0295	0.006	0.251
<i>rs2039488</i>		0.078	0.115	0.01339	0.07877	...	0.05075	0.004	0.609
<i>rs1467813</i>	<i>RGS10</i>	0.286	0.293	0.63177	...	...	0.71857	0.539	0.582
<i>rs927427</i>		0.514	0.464	0.06936	0.00003	0.00002	0.05976	0.198	0.577
<i>rs4146894</i>	<i>PLEKHA1</i>	0.598	0.466	<0.00001	<0.00001	0.00001	<0.00001	0.008	0.802
<i>rs1882907</i>		0.127	0.187	0.00261	0.00013	0.00006	0.00521	0.169	0.172
<i>rs760336</i>	<i>PRSS11</i>	0.395	0.480	0.00469	0.00126	...	0.02036	0.232	0.581
<i>rs763720</i>	<i>PRSS11</i>	0.295	0.212	0.00053	...	...	0.00290	0.198	0.021

NOTE.—The minor-allele frequency is reported for controls (estimated by counting) and families (estimated by Mendel, version

5). The moving-window haplotype  $P$  values correspond to the SNPs in the same row as the  $P$  value and the next one or two SNPs for the two- and three-SNP moving window, respectively. For GIST, with the use of NPL scores from chromosome 1 (NPL 1) and chromosome 10 (NPL 10),  $P$  values  $\leq 0.05$  are in bold italics and  $P$  values  $\leq 0.001$  are underlined. Blank spaces separate the three chromosomal regions corresponding to SNPs in and around *CFH*, *GRK5/RGS10*, and *PLEKHA1/LOC687715/PRSS1*.

extremely low  $P$  values across the whole *CFH* gene (table 3.3), which supports earlier findings of strong association between *CFH* and ARM.

**3.4.1.3 CIDR GIST results** When GIST was performed on the CIDR data set, the two smallest  $P$  values in chromosome 10q26 (0.006 and 0.004) occurred in the *GRK5/RGS10* region, whereas the third smallest  $P$  value (0.008) occurred in *PLEKHA1* (table 3.3). All four SNPs in the *GRK5* gene have small GIST  $P$  values. The GIST results suggest that both *GRK5* and *PLEKHA1* contribute significantly to the linkage signal on chromosome 10 and that *CFH* contributes to the linkage signal on chromosome 1. Neither of the two SNPs in *PRSS11* contributes significantly to the linkage signal on chromosome 10. There was no evidence that the genes on chromosome 10 were related to the linkage signal seen on chromosome 1.

## 3.4.2 PART II: Analysis of Locally Genotyped SNPs

**3.4.2.1 Local association results** After additional SNPs were typed locally, the allele and genotype test generated extremely small  $P$  values for each of the three genes *PLEKHA1*, *LOC387715*, and *PRSS11* (table 3.4). The moving-window haplotype analyses with three SNPs resulted in very small  $P$  values across the entire *PLEKHA1/LOC387715/PRSS11* region (table 3.4). Thus, although association implicates the *PLEKHA1/LOC387715/PRSS11* region, it does not distinguish between these genes.

**3.4.2.2 Local GIST results** Of the three genes *PLEKHA1*, *LOC387715*, and *PRSS11*, GIST most strongly implicated *PLEKHA1* (table 3.4). It also generated a small  $P$  value for *rs10490924* in *LOC387715*, but this SNP is in high LD with the *PLEKHA1* SNPs (see fig. 3.2A). When the locally typed data set was used, GIST did not generate any significant results for *PRSS11*, similar to the nonsignificant results observed in the larger CIDR sample. This implies that *PLEKHA1* (or a locus in strong LD with it) is the most likely to be involved in ARM, and therefore *LOC387715* remains a possible candidate locus.

For a fair assessment of which SNP accounts for the linkage signal across the region, the NPLs were computed using only the locally genotyped families. This permitted us to compare the *PLEKHA1/LOC387715/PRSS11* results (table 3.4) directly with the *GRK5/RGS10* results. For the locally typed data set, the GIST results for *GRK5* are also interesting, with modest  $P$  values

Table 3.4: CCREL, GIST, and Allele-Frequency Estimation for Locally Typed Families and Controls

SNP	Gene	P value for Test							
		Frequency in		Allele Test	Moving-Window Haplotype Test		Genotype Test	GIST	
		Families ( $n = 323$ )	Controls ( $n = 117$ )		With 2 SNPs	With 3 SNPs		NPL 10	NPL 1
<i>rs6658788</i>		0.563	0.483	0.02200	0.00052	0.00162	0.04920	0.319	0.244
<i>rs1538687</i>		0.213	0.342	0.00004	0.00043	0.00066	0.00014	0.652	0.302
<i>rs1416962</i>		0.299	0.393	0.00597	0.02623	0.02051	0.01819	0.442	0.041
<i>rs946755</i>		0.295	0.380	0.01234	0.01243	<0.00001	0.04531	0.409	0.040
<i>rs6428352</i>		0.001	0.004	...	<0.00001	<0.00001	...	...	...
<i>rs800292</i>	<i>CFH</i>	0.120	0.269	<0.00001	<0.00001	<0.00001	<0.00001	0.315	0.014
<i>rs1061170</i>	<i>CFH</i>	0.609	0.310	<0.00001	<0.00001	<0.00001	<0.00001	0.895	0.132
<i>rs10922093</i>	<i>CFH</i>	0.210	0.295	0.00693	0.00175	<0.00001	0.01723	0.360	0.327
<i>rs70620</i>	<i>CFH</i>	0.148	0.150	0.91163	<0.00001	<0.00001	0.56770	0.737	0.356
<i>rs1853883</i>		0.633	0.432	<0.00001	<0.00001	<0.00001	<0.00001	0.776	0.011
<i>rs1360558</i>		0.437	0.389	0.18014	0.43576	0.02079	0.37993	0.975	0.488
<i>rs955927</i>		0.433	0.385	0.15343	0.01037	...	0.36087	0.017	0.585
<i>rs4350226</i>		0.050	0.103	0.00312	...	...	0.00373	0.228	0.174
<i>rs4752266</i>	<i>GRK5</i>	0.223	0.226	0.81772	0.27748	0.64917	0.08279	0.107	0.453
<i>rs915394</i>	<i>GRK5</i>	0.228	0.209	0.34489	0.83219	0.05560	0.62183	0.049	0.320
<i>rs1268947</i>	<i>GRK5</i>	0.117	0.115	0.81975	0.02748	0.02192	0.78965	0.049	0.689
<i>rs1537576</i>	<i>GRK5</i>	0.497	0.419	0.02604	0.02232	0.05636	0.06334	0.012	0.023
<i>rs2039488</i>		0.083	0.115	0.11177	0.42428	...	0.42399	0.025	0.358
<i>rs1467813</i>	<i>RGS10</i>	0.293	0.295	0.86608	...	...	0.85954	0.506	0.492
<i>rs927427</i>		0.506	0.487	0.56710	0.00056	0.00083	0.42264	0.306	0.625
<i>rs4146894</i>	<i>PLEKHA1</i>	0.611	0.474	0.00004	0.00012	0.00053	0.00024	0.006	0.737
<i>rs12258692</i>	<i>PLEKHA1</i>	0.008	0.000	...	0.54750	0.00018	...	...	...
<i>rs4405249</i>	<i>PLEKHA1</i>	0.139	0.158	0.39378	0.00026	0.00280	0.33118	0.003	0.345
<i>rs1045216</i>	<i>PLEKHA1</i>	0.289	0.427	0.00004	0.00036	0.00001	0.00026	0.068	0.825
<i>rs1882907</i>		0.131	0.184	0.01761	0.00140	0.01099	0.04401	0.017	0.372
<i>rs10490923</i>	<i>LOC387715</i>	0.089	0.141	0.02112	0.05024	<0.00001	0.03415	0.086	0.251
<i>rs2736911</i>	<i>LOC387715</i>	0.121	0.119	0.71668	<0.00001	<0.00001	0.64230	0.312	0.968
<i>rs10490924</i>	<i>LOC387715</i>	0.475	0.193	<0.00001	<0.00001	<0.00001	<0.00001	0.018	0.327
<i>rs11538141</i>	<i>PRSS11</i>	0.004	0.005	...	0.00726	0.01676	...	...	...
<i>rs760336</i>	<i>PRSS11</i>	0.373	0.474	0.00527	0.01386	0.00036	0.01396	0.479	0.683
<i>rs763720</i>	<i>PRSS11</i>	0.296	0.226	0.01645	0.00016	...	0.03899	0.305	0.451
<i>rs1803403</i>	<i>PRSS11</i>	0.118	0.030	0.00009	...	...	0.00022	0.714	0.778

NOTE.—The minor-allele frequency is reported for controls (estimated by counting) and families (estimated by Mendel, version

5). The moving-window haplotype P values correspond to the SNPs in the same row as the P value and the next one or two SNPs for the two- and three-SNP moving window, respectively. For GIST, with the use of NPL scores from chromosome 1 (NPL 1) and chromosome 10 (NPL 10), P values  $\leq 0.05$  are in bold italics and P values  $\leq 0.001$  are underlined. Blank spaces separate the three chromosomal regions corresponding to SNPs in and around *CFH*, *GRK5/RGS10*, and *PLEKHA1/LOC687715/PRSS1*.

of the same magnitude as the  $P$  values we got from applying GIST to *CFH* (table 3.4). However, note that the  $P$  values are not as small as those seen when the CIDR data set was analyzed. Since all of the SNPs in the *GRK5* region are CIDR SNPs, this difference is solely a function of sample size, because the locally typed data set is smaller than the CIDR data set (see table 3.2).

### 3.4.3 Part III: Interaction and OR Analyses

**3.4.3.1 GIST results** We did not see any strong evidence of an interaction between the chromosome 1 and chromosome 10 regions, by use of GIST. When the CIDR data set was used to test whether SNPs on chromosome 10 contribute to the linkage signal on chromosome 1 (see GIST, NPL 1, in table 3.3), only *rs763720* in *PRSS1* gave a  $P$  value  $< 0.05$ ; however, *rs763720* does not contribute significantly to the linkage signal on chromosome 10, which makes this  $P$  value less convincing. When we used the local data set, one *GRK5* variant (*rs1537576*), which was not significant in the larger CIDR data set, gave a  $P$  value  $< 0.05$ . Similarly, we did not see evidence that SNPs within *CFH* contribute to the linkage signal on chromosome 10; only one SNP (*rs955927*) gave a  $P$  value  $< 0.05$ —this SNP, however, is not in the *CFH* gene and is not in strong LD (see fig. 3.2B) with any SNPs in *CFH*.

**3.4.3.2 Logistic regression results** The logistic regression results (table A3) suggest that an additive model including the variants from *CFH* and *PLEKHA1* is the best model for predicting case-control status; this indicates that both genes are important to the ARM phenotype. The AIC criteria also suggest that an additive model including an additive interaction term is the next best model (table A3); however, the interaction term is not significant ( $P = 0.71$ ). We obtain similar results for interaction between *CFH* and *PRSS11*, where the additive model including both variants appears to be the best model. Within the *GRK5/RGS10* region, a model with the *CFH* SNP alone is the best-fitting model, which suggests that the prediction of case-control status with *CFH* genotype does not improve by the addition of either the *GRK5* or *RGS10* variant to the model.

**3.4.3.3 OR and AR** We estimated the magnitude of association by calculating OR and AR values; the significant associations we saw (table 3.5) are, not surprisingly, consistent with the results from the CCREL tests in parts I and II. Our two most significant SNPs in the *PLEKHA1/LOC387715*

region are SNPs *rs4146894* (*PLEKHA1*) and *rs10490924* (*LOC387715*); the two tests are highly correlated because the LD between those SNPs is very high ( $D' = 0.93$ ) (see fig. 3.2A). The third most significant SNP (*rs1045216*) in the chromosome 10 region is a nonsynonymous SNP in *PLEKHA1* and in high LD with both *rs4146894* ( $D' = 0.97$ ) and *rs10490924* ( $D' = 0.91$ ).

We obtained results and OR and AR values (table 3.5) similar to those that others have reported for the *CFH* gene. The three most significant SNPs were *rs1061170* (*Y402H* variant), *rs800292* (in *CFH*), and *rs1853883* (in strong LD with *rs1061170*;  $D' = 0.91$ ).

The magnitude of the association we saw within *PLEKHA1/LOC387715* is very similar to the level of association seen between *CFH* and ARM; both loci result in extremely low  $P$  values ( $P < 0.0001$ ). The OR and AR values were also similar: the dominant OR was 5.29 (95% CI 3.358.35) within *CFH* and 5.03 (95% CI 3.27.91) within *PLEKHA1/LOC387715*, and the dominant AR for *CFH* and *PLEKHA1/LOC387715* was 68% and 57%, respectively.

**3.4.3.4 Subphenotype analyses** We estimated ORs and ARs for patients with exudative disease versus controls and for patients with GA versus controls (table A4). ORs and corresponding  $P$  values yielded similar findings to those of the allele test of *CCREL* (tables 3.3 and 3.4). We found no major differences between the ORs for the presence of either GA or CNV.

## 3.5 DISCUSSION

Our linkage studies of families with ARM have consistently identified the chromosome 1q31 and chromosome 10q26 loci, in addition to several other loci. Multiple linkage studies have replicated this finding; thus, we undertook a focused SNP analysis of both regions, using ARM-affected families as well as unrelated affected individuals and controls. We confirmed the strong association of chromosome 1q31 with *CFH* that has been reported by others (see also (CONLEY *et al.* 2005)), and we have shown, for the first time, that SNPs in *CFH* significantly account for the linkage signal. Interestingly, our smallest GIST  $P$  value ( $< 0.001$ ) was for *rs1853883* (which has a high  $D'$  of 0.91 with the *Y402H* variant) and not for the presumed “disease-associated” *Y402H* variant itself. This raises the possibility that we may still have to consider other possible ARM-related variants within the *CFH* gene and that these may be in high LD with *Y402H*.

Table 3.5: ORs, ARs, and Simulated  $P$  Values from  $\chi^2$  Test with 10,000 Replicates

SNP (allele)	Gene	Dominant ([RR+RN] vs. NN)				Heterozygotes (RN vs. NN)		Recessive (RR vs [RN+NN])				Homzygotes RR vs. NN	
		OR	95% CI	AR	P	OR	AR	OR	95% CI	AR	P	OR	AR
<i>rs6658788</i> (2)		0.83	0.57-1.22	-14.04	0.3909	1.09	2.69	1.01	0.68-1.5	0.21	1.0	0.88	-5.92
<i>rs1538687</i> (2)		0.68	0.49-0.95	-19.38	0.023	0.5	-11.74	0.42	0.23-0.78	-6.52	0.0068	0.38	-12.42
<i>rs1416962</i> (2)		0.84	0.6-1.18	-10.02	0.3418	0.89	-2.57	0.82	0.49-1.38	-2.31	0.5002	0.77	-5.74
<i>rs946755</i> (2)		0.8	0.57-1.13	-12.52	0.232	1.0	0.04	0.9	0.53-1.52	-1.24	0.7816	0.8	-4.34
<i>rs6428352</i> (2)		...	...	...	...	...	...	...	...	...	...	...	...
<i>rs800292</i> (1)	<i>CFH</i>	0.43	0.3-0.62	-30.01	<0.0001	0.48	-23.85	0.15	0.05-0.45	-4.98	0.0001	0.12	-8.19
<i>rs1061170</i> (2)	<i>CFH</i>	5.29	3.35-8.35	68.2	<0.0001	2.66	28.55	4.57	2.48-8.42	30.06	<0.0001	10.05	63.72
<i>rs10922093</i> (1)	<i>CFH</i>	0.59	0.39-0.88	-25.61	0.0111	0.63	-19.65	0.5	0.24-1.04	-4.98	0.0736	0.41	-10.14
<i>rs70620</i> (1)	<i>CFH</i>	0.83	0.57-1.19	5.64	0.3366	0.85	-4.29	0.67	0.27-1.68	-1.3	0.4525	0.64	-1.93
<i>rs1853883</i> (2)		2.67	1.78-4.01	54.41	<0.0001	1.65	19.21	2.08	1.43-3.02	22.06	0.0003	3.55	55.04
<i>rs1360558</i> (1)		1.16	0.82-1.65	9.12	0.414	1.1	5.39	1.25	0.8-1.96	3.94	0.3774	1.32	9.01
<i>rs955927</i> (2)		1.13	0.79-1.6	7.5	0.5303	1.28	6.35	1.31	0.83-2.08	4.53	0.2588	1.36	9.38
<i>rs4350226</i> (2)		0.51	0.32-0.81	-9.68	0.0038	0.27	-4.76	0.16	0.01-1.74	-0.95	0.142	0.14	-1.16
<i>rs4752266</i> (2)	<i>GRK5</i>	0.88	0.62-1.23	-5.57	0.4325	3.27	10.68	2.81	0.98-8.04	3.89	0.0457	2.56	5.51
<i>rs915394</i> (2)	<i>GRK5</i>	1.28	0.9-1.82	8.91	0.1543	1.35	2.73	1.56	0.58-4.14	1.53	0.3892	1.68	2.72
<i>rs1268947</i> (2)	<i>GRK5</i>	1.05	0.7-1.57	1.06	0.841	1.24	1.82	1.27	0.35-4.55	0.45	0.7761	1.28	0.58
<i>rs1537576</i> (2)	<i>GRK5</i>	1.59	1.11-2.29	27.95	0.0109	0.89	-3.74	1.08	0.71-1.62	1.59	0.7579	1.47	15.14
<i>rs2039488</i> (2)		0.7	0.45-1.07	-6.5	0.1067	0.23	-11.98	0.19	0.04-0.79	-2.33	0.0242	0.18	-2.85
<i>rs1467813</i> (1)	<i>RGS10</i>	0.96	0.69-1.35	-1.84	0.8645	1.01	0.42	0.77	0.42-1.38	-2.27	0.4265	0.77	-3.76
<i>rs927427</i> (1)		1.09	0.74-1.62	6.57	0.6172	0.94	-4.66	1.67	1.09-2.56	10.73	0.0201	1.6	19.91
<i>rs4146894</i> (1)	<i>PLEKHA1</i>	2.22	1.49-3.31	46.78	0.0002	1.77	33.08	2.21	1.49-3.29	20.46	<0.0001	3.31	49.88
<i>rs12258692</i> (2)	<i>PLEKHA1</i>	...	...	...	...	...	...	...	...	...	...	...	...
<i>rs4405249</i> (1)	<i>PLEKHA1</i>	0.62	0.33-1.15	-12.96	0.1692	0.61	-12.69	0.87	0.1-7.56	-0.23	1.0	0.77	-0.57
<i>rs1045216</i> (2)	<i>PLEKHA1</i>	0.48	0.32-0.74	-51.23	0.0005	0.49	-18.27	0.37	0.21-0.65	-14.3	0.0003	0.28	-35.68
<i>rs1882907</i> (2)		0.58	0.4-0.84	-16.73	0.0026	0.44	-5.79	0.31	0.1-0.97	-2.37	0.0438	0.27	-3.65
<i>rs10490923</i> (2)	<i>LOC387715</i>	0.53	0.31-0.9	-13.27	0.0239	0.34	-9.01	0.22	0.04-1.09	2.51	0.0809	0.2	-3.32
<i>rs2736911</i> (2)	<i>LOC387715</i>	0.72	0.42-1.21	-6.92	0.2552	1.47	1.99	1.1	0.13-9.53	0.1	1.0	1.03	0.04
<i>rs10490924</i> (2)	<i>LOC387715</i>	5.03	3.2-7.91	57.11	<0.0001	2.72	22.76	5.75	2.46-13.46	21.2	<0.0001	10.57	42.71
<i>rs11538141</i> (2)	<i>PRSS11</i>	...	...	...	...	...	...	...	...	...	...	...	...
<i>rs760336</i> (2)	<i>PRSS11</i>	0.64	0.44-0.93	-35.37	0.013	0.8	-6.95	0.69	0.46-1.03	-7.95	0.0773	0.55	-26.43
<i>rs763720</i> (1)	<i>PRSS11</i>	1.69	1.2-2.38	21.24	0.0018	1.55	16.95	2.63	1.1-6.25	5.17	0.0277	3.16	10.14
<i>rs1803403</i> (1)	<i>PRSS11</i>	2.98	1.25-7.06	10.51	0.0093	2.98	10.51	...	...	...	...	...	...

NOTE.—Type A-affected individuals are compared with controls. Allele denotes the risk allele (minor allele in controls). RR = homozygotes for the risk allele; RN = heterozygotes for the risk allele; NN = homozygotes for the normal allele. Locally typed SNPs are in bold italics. Blank spaces separate the three chromosomal regions corresponding to SNPs in and around *CFH*, *GRK5/RGS10*, and *PLEKHA1/LOC687715/PRSS11*.



Our studies of chromosome 10q26 have implicated two potential loci: (1) a very strongly implicated locus that includes three tightly linked genes, *PLEKHA1*, *LOC387715*, and *PRSS11*, and (2) a less strongly implicated locus comprising two genes, *GRK5* and *RGS10* (fig. 3.1). The GIST analysis does not support *PRSS11* as the ARM-related gene, but it does not completely exclude it as a potential candidate. *PLEKHA1* has the lowest GIST-derived *P* values, whereas *LOC387715* harbors the SNP with the strongest association signal and the highest ORs. With the high LD between SNPs in *LOC387715* and *PLEKHA1*, one cannot clearly distinguish between these genes by statistical analyses alone. However, it is clear that the magnitude of the impact of the *PLEKHA1/LOC387715* locus on ARM is comparable to that which has been observed for the *CFH* locus. As in recent studies (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005), we have found, in our case-control population, that the *CFH* allele (either heterozygous or homozygous) accounts for an OR of 5.3 (95% CI 3.48.4) and a significant population AR of 68%. In the same fashion, the high-risk allele within the *PLEKHA1/LOC387715* locus accounts for an OR of 5.0 (95% CI 3.27.9) and an AR of 57% when both heterozygous and homozygous individuals are considered. As noted by KLEIN *et al.* (2005), unless the disease is very rare, the OR determined from a case-control study will usually overestimate the equivalent relative risk. Estimates of AR based on ORs for common genetic disorders can misrepresent the extent to which a variant accounts for the population AR. However, if this caution is kept in mind, it is still useful for us to present AR values to allow for relative comparisons and to allow the reader to appreciate that the potential impact of the *CFH Y402H* variant on ARM is comparable to that of the variants observed in the *PLEKHA1/LOC387715* locus.

In the case of *CFH* on chromosome 1, the association data were extremely compelling for a single gene, even though *CFH* is within a region of related genes. In addition to the association data found by multiple independent groups, there is additional biological data to implicate *CFH*, including localization of the protein within drusen deposits of patients with ARM. Thus, we also must consider the biological relevance of the potential ARM-susceptibility genes identified by our studies of chromosome 10q26.

As noted above, the GIST analysis most strongly implicated *PLEKHA1*, particularly when we included the additional nonsynonymous SNPs that we genotyped locally. *PLEKHA1* encodes the protein TAPP1, which is a 404-aa protein with a putative phosphatidylinositol 3,4,5-trisphosphate-binding motif (PPBM), as well as two plectstrin homology (PH) domains. The last three C-terminal amino acids have been predicted to interact with one or more of the 13 PDZ domains

of MUPP1 (similar to the PDZ domain within PRSS11). Dowler and colleagues ([DOWLER \*et al.\* 2000](#)) have shown that the entire TAPP1 protein, as well as the C-terminal PH domain, interacts specifically with phosphatidylinositol 3,4-bisphosphate (PtdIns(3,4)P2) but not with any other phosphoinositides. TAPP1, which has 58% identity with the first 300 aa of TAPP2, shows a fivefold higher affinity for PtdIns(3,4)P2 than does TAPP2, and this binding is nearly eliminated by mutation of the conserved arginine 212 to leucine within the PPBM region (which is part of the second PH domain). The most well-defined role for TAPP1 (and its relatives, Bam32 and TAPP2) has been as an activator of lymphocytes. PtdIns(3,4)P2 is preferentially recruited to cell membranes when lipid phosphatase (SHIP) is activated along with PI3K (phosphatidyl inositol 3-kinase). SHIP is responsible for the dephosphorylation of PIP3 to PtdIns(3,4)P2. SHIP is a negative regulator of lymphocyte activation, and thus TAPP1 and TAPP2 may be crucial negative regulators of mitogenic signaling and of the PI3K signaling pathway. Thus, one can envision a role in the eye for *PLEKHA1* and its protein, TAPP1, in modifying local lymphocyte activation, consistent with the hypothesis that ARM is closely linked to an inflammatory process.

However, we need to still consider the biological plausibility of the other two candidate genes within this locus, *LOC387715* and *PRSS11*. Little is known regarding the biology of *LOC387715*, except that its expression appears to be limited to the placenta. Our own reverse transcription experiments with human retinal RNA have confirmed the expression of *PLEKHA1* and *PRSS11*, but we have not detected *LOC387715* transcripts in the retina under standard conditions, even though we confirmed its expression with placental RNA (data not shown). However, we cannot exclude the possibility that *LOC387715* is expressed at very low levels in the retina or retinal pigment epithelium or that its expression in nonocular tissues, such as dendritic cells or migrating macrophages, could be a factor in the pathogenesis of ARM.

*PRSS11* is one of the genes of the mammalian high temperature requirement A (HtrA) serine protease family, which has a highly conserved C-terminal PDZ domain ([OKA \*et al.\* 2004](#)). These secretory proteases were initially identified because of their homologies to bacterial forms that are required for survival at high temperatures and molecular chaperone activity at low temperatures. The ATP-independent serine protease activity is thought to degrade misfolded proteins at high temperatures. The mammalian form, HtrA1, has been shown to be selectively stimulated by type III collagen alpha 1 C propeptide, in contrast to HtrA2 ([MURWANTOKO \*et al.\* 2004](#)). Type III collagen is a major constituent (35%–39% of the total collagen) in Bruch membrane and is also present in small amounts in the retinal microvascular basement membranes. Developmental studies

have reported ubiquitous expression of HtrA1, but with temporal and spatial specificities that coincide with those regions in which Tgf $\beta$  proteins play a regulatory role (DE LUCA *et al.* 2004). Oka and colleagues (OKA *et al.* 2004) have shown that HtrA1 is capable of inhibiting the signaling of a number of Tgf $\beta$  family proteins, including Bmp4, Bmp2, and Tgf $\beta$ 1, presumably by preventing receptor activation with a requirement for protease activity of the HtrA1 molecule. One clue as to the potential importance of these relationships for ARM comes from the studies of HOLLBORN *et al.* (2004), who found that human RPE cells in vitro experienced reduced proliferation in the presence of Tgf $\beta$ 1 and Tgf $\beta$ 2 and an increase in levels of collagen III and collagen IV transcripts. Normally, a rise in collagen III would activate HtrA1 and would lead to secondary inhibition of the effects of Tgf $\beta$ 1. However, if the serine protease is less effective (because of either reduced synthesis or a nonfunctional mutation), then this regulatory pathway would be disrupted, leading to an overall reduction in the proliferation potential of the RPE cells, perhaps contributing to RPE atrophy or further changes that could lead to the development of ARM. The gradual reduction in solubility of type III collagen in Bruch membrane that has been observed with aging could also, in part, account for a general reduction in HtrA1 activity as an individual ages.

Both *PRSS11* and *PLEKHA1* are expressed in the retina, and a SAGE analysis of central and peripheral retina (Gene Expression Omnibus [GEO] expression data) indicates higher levels of transcripts of both genes in the central macula (more so for *PLEKHA1* than for *PRSS11*). Multiple studies (reported in GEO profiles) have shown that *PLEKHA1* expression is significantly induced in a variety of cell types in response to exposure to specific inflammatory cytokines. *PRSS11* has also been investigated as part of a microarray expression analysis of dermal fibroblasts that have been oxidatively challenged, in a comparison between normal individuals and patients with ARM. In that study, half of the ARM samples (9 of 18) had lower Htra1 expression levels than any of the normal samples. The lower levels of Htra1 in nonocular tissues of patients with ARM would suggest that this is an intrinsic difference in the biology of these patients, compared with that of normal individuals, and is not a consequence of degenerative changes in the eye.

Several lines of evidence support the *GRK5/RGS10* locus. The peak of our  $S_{all}$  multipoint curve is directly over *GRK5*, and our largest two-point  $S_{all}=3.86$  (*rs555938*) is only 206 kb centromeric of *GRK5*. The  $P$  values for the GIST analysis of the *GRK5/RGS10* CIDR data were 0.004 and 0.006, which are even smaller than the  $P$  value for the SNP within *PLEKHA1* ( $P = 0.008$ ). By use of our locally genotyped sample, the GIST  $P$  value for the *GRK5* locus was 0.012, which is comparable to the  $P$  value that we found for the *Y402H* variant in *CFH* ( $P = 0.011$ ). However,

the CCREL analyses were not very significant for the *GRK5* SNPs, and the ORs were mostly nonsignificant.

On the basis of biological evidence, *GRK5* is a reasonable ARM candidate gene, given its role in modulating neutrophil responsiveness to chemoattractants and its interactions with the Toll 4 receptor (HARIBABU and SNYDERMAN 1993; FAN and MALIK 2003), which has also been implicated in ARM (ZAREPARSI *et al.* 2005). The retinal or RPE expression of *GRK5* is not especially relevant to the argument of causality, because it would be the expression and function of *GRK5* in migrating lymphocytes and macrophages that would be crucial to its role in the immune and/or inflammatory pathways that may be pathogenic in ARM. The strongest GIST results occur at *rs2039488*, which is located between *GRK5* and *RGS10*, 3' of the ends of both genes. Several other SNPs within *GRK5* also have small GIST P values, whereas the *RGS10* SNP has a nonsignificant GIST P value. However, we cannot completely exclude the possibility that there is a SNP within *RGS10* that is in strong LD with *rs2039488*.

*RGS10* is one of a family of G protein-coupled receptors that has been implicated in chemokine-induced lymphocyte migration (MORATZ *et al.* 2004) and whose expression in dendritic cells (which have been identified in ARM-related drusen deposits) is modified by the Toll-like signaling pathway (SHI *et al.* 2004). *RGS10* and *GRK5* expression in the same microarray study of oxidatively stressed dermal fibroblasts in patients with ARM and control subjects showed minor fluctuations among the samples but no clear differences between the controls and affected individuals. This does not necessarily lower the potential for these genes being involved in ARM, since the dermal fibroblasts lack the cell populations that would be expected to have modulation of *RGS10*- and/or *GRK5*-related proteins.

We have attempted to look at potential interactions between the high-risk alleles within the *PLEKHA1/LOC387715* and *GRK5/RGS10* loci with respect to *CFH* on chromosome 1. This is perhaps the first report to use GIST to examine these interactions, and we found no evidence that the NPL data on chromosome 1 could be accounted for by the SNP data on chromosome 10. Conversely, we found no such associations between the NPL data on chromosome 10 and the SNP data from the *CFH* alleles. Logistic regression analysis also failed to identify an interaction, and it appears that a simple additive risk model is the most parsimonious. We have performed some initial logistic analyses that include exposure to smoking. These analyses were initiated because of the previous suggestion of an interaction between smoking and the biology of complement factor

H (ESPARZA-GORDILLO *et al.* 2004) and because of our prior studies, which found an interaction between smoking and the locus on chromosome 10q26 (WEEKS *et al.* 2004). To date, we have found no strong interaction between smoking and either *CFH* or *PLEKHA1/LOC387715*, but we are still exploring a possible interaction with the *GRK5/RGS10* locus and different modeling strategies.

We also examined the associations of ARM subphenotypes with the SNPs on chromosomes 1 and 10 (table A4). We found no major differences in the ORs for the presence of either GA or CNV, which suggests that these ARM loci contribute to a common pathogenic pathway that can give rise to either end-stage form of the disease. This does not exclude the possibility that there are other as-yet-undescribed genetic loci that may confer specific risk of GA or CNV development separately.

In summary, these SNP-based linkage and association studies illustrate both the power and the limitation of such methods to identify the causative alleles and genes underlying ARM susceptibility. These genetic approaches allow us to consider genes and their variants that may contribute to disease, whether or not there is tissue-specific expression. Through high-density SNP genotyping, we have narrowed the list of candidate genes within the linkage peak found on chromosome 10q26, from hundreds of genes to primarily *GRK5* and *PLEKHA1*, but we cannot completely exclude possible roles for *RGS10* and/or *PRSS11* and *LOC387715*. Additional genotyping of nonsynonymous 3 SNPs within the *GRK5* gene may help to further discriminate between *GRK5* and *RGS10*, but it may not establish a definitive assignment of causality. Replication by other studies (as done in the case of *CFH*) may allow the attention to be focused on a single gene in future studies of ARM pathology, but there is also the distinct possibility that we will be unable to achieve further resolution with association studies or to clearly establish whether there are more than two genes responsible for ARM susceptibility on chromosome 10q26. However, it is now well within the capabilities of molecular biologists to investigate the potential role of each of these candidate genes in mouse models of ARM and to address the issue of a causal role in disease pathogenesis. Association studies are an incredibly powerful means of testing hypotheses of genetic contributions to disease, but, except in the most extreme cases, they cannot provide definitive answers, even when there are impressive *P* values.

#### 4.0 *CFH*, *ELOVL4*, *PLEKHA1* AND *LOC387715* GENES AND SUSCEPTIBILITY TO AGE-RELATED MACULOPATHY: AREDS AND CHS COHORTS AND META-ANALYSES

This section has been published in Human Molecular Genetics volume 15, issue 21, pages 3206-3218 (CONLEY *et al.* 2006). The authors retain the rights to include the article in full or in part in a thesis or dissertation, provided that this not published commercially. No changes have been made to the published version of the paper, except that tables and figures have been renumbered, the citations are of different style and some minor formatting has been made to keep this chapter coherent to the remainder of the thesis. The supporting information published online is in Appendix B. My contribution to this paper was in writing all components of the paper, data analysis, script writing for method implementation.

#### 4.1 ABSTRACT

Age-related maculopathy (ARM) is an important cause of visual impairment in the elderly population. It is of crucial importance to identify genetic factors and their interactions with environmental exposures for this disorder. This study was aimed at investigating the *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715* genes in independent cohorts collected using different ascertainment schemes. The study used a casecontrol design with subjects originally recruited through the Cardiovascular Health Study (CHS) and the Age-Related Eye Disease Study (AREDS). *CFH* was significantly associated with ARM in both cohorts ( $P \leq 0.00001$ ). A meta-analysis confirmed that the risk allele in the heterozygous or homozygous state (OR, 2.4 and 6.2; 95% CI, 2.2–2.7 and 5.4–7.2, respectively) confers susceptibility. *LOC387715* was also significantly associated with ARM in both cohorts ( $P \leq 0.00001$ ) and a meta-analysis confirmed that the risk allele in the heterozygous and homozy-

gous state (OR, 2.5 and 7.3; 95% CI, 2.2–2.9 and 5.7–9.4, respectively) confers susceptibility. Both *CFH* and *LOC387715* showed an allele-dose effect on the ARM risk, individuals homozygous at either locus were at more than two-fold risk compared to those heterozygous. *PLEKHA1*, which is closely linked to *LOC387715*, was significantly associated with ARM status in the AREDS cohort, but not the CHS cohort and *ELOVL4* was not significantly associated with ARM in either cohort. Joint action of *CFH* and *LOC387715* was best described by independent multiplicative effect without significant interaction in both cohorts. Interaction of both genes with cigarette smoking was insignificant in both cohorts. This study provides additional support for the *CFH* and *LOC387715* genes in ARM susceptibility via the evaluation of cohorts that had different ascertainment schemes regarding ARM status and through the meta-analyses.

## 4.2 INTRODUCTION

Age-related maculopathy (ARM) is a leading cause of central blindness in the elderly of industrialized nations. The prevalence of ARM is expected to increase because of the aging of these populations (1). The etiology of ARM is complex, with environmental as well as genetic susceptibility playing a role. Association-based analyses are generally more sensitive to small genetic effects than linkage-based analyses and are extremely valuable for fine mapping of disease-related genes (CORDELL and CLAYTON 2005). Case-control association studies with the use of unrelated individuals may have advantages over family-based studies, especially when a multilocus genetic model is anticipated (HOWSON *et al.* 2005; RISCH 2001), however, such studies are potentially sensitive to the ascertainment scheme for the case and control cohorts. For this reason, there is value in assessing candidate genes in populations from projects with different study designs. This current study investigates the complement factor H (*CFH*) gene, the elongation of very long chain fatty acid-like 4 (*ELOVL4*) gene, the pleckstrin homology domain-containing protein (*PLEKHA1*) gene, and the hypothetical *LOC387715* gene in two such distinct cohorts.

The association of the *CFH* gene with ARM susceptibility has been established in samples of European American descent (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005; HAGEMAN *et al.* 2005; CONLEY *et al.* 2005; ZAREPARSI *et al.* 2005) as well as in samples from the United Kingdom (SEPP *et al.* 2006), Germany (RIVERA *et al.* 2005), France (SOUIED *et al.* 2005), Iceland (MAGNUSSON *et al.* 2006) and Japan (OKAMOTO *et al.* 2006).



Three studies support the *PLEKHA1/LOC387715* locus on chromosome 10q26 (JAKOBSDOTTIR *et al.* 2005; RIVERA *et al.* 2005; SCHMIDT *et al.* 2006). The study by Jakobsdottir *et al.* (JAKOBSDOTTIR *et al.* 2005) reported that the *PLEKHA1/LOC387715* locus was significantly associated with ARM status, however, strong linkage disequilibrium between *PLEKHA1* and *LOC387715* in the independent family-based and casecontrol populations utilized for the study meant that a role for one gene over the other could not be determined (JAKOBSDOTTIR *et al.* 2005). Evidence that the hypothetical *LOC387715* gene was more likely to be the gene accounting for susceptibility to ARM came from a study by RIVERA *et al.* (2005) that utilized two independent casecontrol samples and a study by Schmidt *et al.* that utilized both family-based and case-control studies (SCHMIDT *et al.* 2006). All three studies indicated that the association of this region on chromosome 10q26 with ARM status was independent of the association with *CFH* that had been previously reported in all three populations (HAINES *et al.* 2005; CONLEY *et al.* 2005; RIVERA *et al.* 2005). In addition, based on the Schmidt *et al.* study, the effect of the *LOC387715* locus appears to be modified by smoking history (SCHMIDT *et al.* 2006).

Two studies have evaluated a potential role for *ELOVL4* in ARM in humans. AYYAGARI *et al.* (2001) evaluated the gene and found no significant association with ARM status in their sporadic case-control analysis. However, Conley *et al.* found a significant association of *ELOVL4* and ARM status in familial and sporadic case-control analyses (CONLEY *et al.* 2005). The difference in findings between these studies may be related to the proportion of cases with exudative ARM in each population, since Conley *et al.* found that *ELOVL4* was especially associated with the exudative sub-phenotype (CONLEY *et al.* 2005). These results indicate that additional studies are needed to establish or refute a relationship between *ELOVL4* and ARM.

The two cohorts utilized for this study were the Cardiovascular Health Study (CHS), a population-based cohort of individuals 65 years and older at baseline for which ARM status was not a factor for ascertainment (FRIED *et al.* 1991), and the Age-Related Eye Disease Study (AREDS), a cohort of individuals aged 55-80 years participating in a randomized controlled clinical trial of antioxidant and zinc intervention for which ARM status was a factor for ascertainment (AGE RELATED EYE DISEASE STUDY GROUP 1999). These cohorts have been previously described (KLEIN *et al.* 2003; AGE RELATED EYE DISEASE STUDY GROUP 2000).

This study was designed to evaluate the *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715* genes in two independent cohorts with very different ascertainment schemes in relation to ARM status



and then to incorporate the findings into meta-analyses. Association of a gene with susceptibility to ARM regardless of ascertainment scheme would further increase the evidence that the association is real and would enhance the likelihood that evaluation of the gene(s) would accurately identify at risk individuals.

### 4.3 RESULTS

To further evaluate *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715* in ARM, we genotyped previously reported SNPs within all four genes in samples from the AREDS and CHS studies. Separate analyses were performed on each data set, using a total of 701 non-Hispanic white ARM patients and 175 controls from the AREDS study, and a total of 126 non-Hispanic white ARM patients and 1051 controls from the CHS study (see Table 4.1 for sample sizes and other characteristics of the data, and Table 4.2 for genotype frequencies). The disease status of subjects at their last follow-up visit was the primary endpoint evaluated for AREDS subjects. The AREDS subjects include controls of grade 1 and cases (grades 3-5) with moderate ARM and advanced ARM in one or both eyes. The ARM disease status of CHS subjects was evaluated by Dr. Gorin, using monocular, non-mydratic fundus photographs taken at the 8-year follow-up visit. The majority of CHS cases had moderate ARM including multiple drusen with and without pigment epithelial changes (equivalent to AREDS grade 3) with a small number of cases having geographic atrophy (GA) or choroidal neovascular membranes (CNV) and the CHS controls are of AREDS grade 1 with the exclusion of those cases with significant extramacular drusen.

#### 4.3.1 Association analyses

For each gene, *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715*, association of one non-synonymous SNP with ARM was assessed by a 2 statistic. The magnitude of the effect of each variant was estimated by odds ratios (ORs) and population attributable risks (PARs). To evaluate whether the variants confer risk similarly to mild/moderate and advanced ARM, ORs were calculated for each grade and subtype (GA and CNV) separately using the AREDS data.

Table 4.1: Characteristics of the study populations

	Mean age (SD)	Clinical subtypes				Total	No. Males (%)
		Neither	GA only	CNV only	Both		
AREDS data							
Controls (1)	76.53 (4.44)	175	...	...	...	175	86 (49)
Cases (345)	79.46 (5.23)	123	147	278	153	701	293 (42)
Cases (45)	79.54 (5.23)	27	147	278	153	605	253 (42)
Cases (3)	78.93 (5.22)	96	0	0	0	96	40 (42)
Cases (4)	78.83 (5.23)	24	59	149	34	266	124 (47)
Cases (5)	80.10 (5.17)	3	88	129	119	339	129 (38)
CHS data							
Controls	70.27 (3.92)	1051	...	...	...	1051	455 (43)
Cases	73.22 (4.84)	100	15	9	2	126	55 (44)

NOTE.—In the AREDS cohort, mean age and phenotypic classification is based on age at last fundus photography. The number in the parentheses denotes the disease severity according the AREDS grading method. In the CHS cohort mean age is based on age at baseline visit, but retinal evaluation was done at 8-year follow-up visit.


**4.3.1.1 CFH** The association of the *Y402H* variant in *CFH* with ARM is extremely significant ( $P \leq 0.00001$ ) in both the AREDS and CHS cohorts (Table 4.3), confirming earlier findings by ourselves (CONLEY *et al.* 2005; JAKOBSDOTTIR *et al.* 2005) and others (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005; RIVERA *et al.* 2005). The estimated ORs for *Y402H* in *CFH* suggest that the variant confers similar risk to all stages of ARM and both forms of advanced ARM, GA and CNV (Fig. 4.1 and Table B1). An allele-dose effect appears to be present, with carriers of two C alleles at higher risk of ARM than carriers of one C allele (Table 4.4) and . Despite the increased risk in carriers of two C alleles, the PAR is similar for the two risk genotypes, owing to relatively high frequency of the CT genotype compared to the CC genotype in the general population. PAR estimates derived from the CHS data set suggest that the CT and CC genotypes explain 27% and 25% of ARM in the non-Hispanic white population, respectively.

Table 4.2: Genotype distributions by ARM status

Gene (Variant) and genotypes	Genotype frequencies in				HapMap (CEU)
	AREDS cases ( <i>n</i> = 701)	CHS cases ( <i>n</i> = 126)	AREDS controls ( <i>n</i> = 175)	CHS controls ( <i>n</i> = 1051)	
<i>CFH</i> ( <i>Y402H</i> )					
TT	0.170	0.264	0.434	0.448	...
CT	0.435	0.482	0.416	0.450	...
CC	0.395	0.255	0.150	0.103	...
<i>ELOVL4</i> ( <i>M299V</i> )					
AA	0.781	0.742	0.711	0.802	0.717
AG	0.195	0.250	0.259	0.174	0.233
GG	0.024	0.008	0.030	0.024	0.050
<i>PLEKHA1</i> ( <i>A320T</i> )					
GG	0.474	0.411	0.339	0.346	0.317
AG	0.443	0.460	0.464	0.476	0.467
AA	0.084	0.129	0.196	0.178	
<i>LOC387715</i> ( <i>S69A</i> )					
GG	0.313	0.442	0.645	0.604	0.583
GT	0.492	0.408	0.331	0.353	0.400
TT	0.195	0.150	0.023	0.043	0.017

NOTE.—AREDS cases are of grades 3–5 and AREDS controls of grade 1. Genotype counts are available by each grade and subphenotype in Table S1 of the Supplementary Material. Description of the HapMap CEU populations is given in the Supplementary Material.

Table 4.3: Results of allele- and genotype-association tests

Evaluated contrast in AREDS or CHS	<i>CFH</i> <i>P</i> -value for test		<i>ELOVL4</i> <i>P</i> -value for test		<i>PLEKHA1</i> <i>P</i> -value for test		<i>LOC387715</i> <i>P</i> -value for test	
	Allele	Genotype	Allele	Genotype <sup>a</sup>	Allele	Genotype	Allele	Genotype
AREDS								
1 vs 345	≤ 0.00001	≤ 0.00001	0.06775	0.13963	0.00004	0.00004	≤ 0.00001	≤ 0.00001
1 vs 5	≤ 0.00001	≤ 0.00001	0.20518	0.32438	≤ 0.00001	≤ 0.00001	≤ 0.00001	≤ 0.00001
1 vs 5 (GA) <sup>b</sup>	≤ 0.00001	≤ 0.00001	0.10465	0.21869	0.04131	0.03862	≤ 0.00001	≤ 0.00001
1 vs 5 (CNV) <sup>c</sup>	≤ 0.00001	≤ 0.00001	0.03445	0.04851	≤ 0.00001	≤ 0.00001	≤ 0.00001	≤ 0.00001
CHS	≤ 0.00001	≤ 0.00001	0.33832	0.07819	0.07626	0.22544	≤ 0.00001	≤ 0.00001

NOTE.—<sup>a</sup> Two-sided *P*-values from Fishers exact test.

<sup>b</sup> ARM cases have GA in both eyes.

<sup>c</sup> ARM cases have CNV in both eyes.

Table 4.4: ORs and PAR% for subjects who are hetero- and homozygous for *Y402H* in *CFH* and *S69A* in *LOC387715*

Evaluated contrast in AREDS or CHS	<i>Y402H</i> in <i>CFH</i>				<i>S69A</i> in <i>LOC387715</i>			
	Heterozygotes (CT vs TT)		Homozygotes (CC vs TT)		Heterozygotes (GT vs GG)		Homozygotes (TT vs GG)	
	ORhet	PAR%	ORhom	PAR%	ORhet	PAR%	ORhom	PAR%
AREDS								
1 vs 345	2.66 (1.81,3.92)	43 (29,54)	6.69 (4.08,10.98)	37 (24,48)	3.06 (2.13,4.39)	42 (33,50)	17.26 (6.22,47.89)	41 (36,46)
1 vs 45	2.82 (1.89,4.19)	45 (31,56)	7.06 (4.27,11.70)	38 (24,50)	3.18 (2.20,4.60)	43 (34,52)	18.30 (6.57,50.93)	43 (37,48)
1 vs 3	1.93 (1.04,3.60)	30 (-3,52)	4.95 (2.46,9.95)	29 (-2,50)	2.45 (1.42,4.23)	34 (13,49)	11.89 (3.70,38.19)	32 (18,43)
1 vs 4	2.67 (1.67,4.27)	43 (24,57)	6.33 (3.60,11.16)	35 (16,50)	2.34 (1.55,3.53)	32 (19,43)	8.19 (2.80,24.00)	24 (16,31)
1 vs 5	2.94 (1.87,4.63)	47 (29,60)	7.71 (4.46,13.34)	41 (23,54)	4.32 (2.85,6.57)	54 (43,63)	32.07 (11.30,91.01)	57 (50,64)
1 vs 45 (GA)	2.54 (1.44,4.48)	41 (15,59)	7.04 (3.69,13.41)	38 (12,56)	2.81 (1.74,4.52)	39 (23,52)	10.14 (3.28,31.31)	28 (17,38)
1 vs 4 (GA)	1.68 (0.78,3.61)	23 (-21,51)	5.55 (2.48,12.41)	32 (-8,57)	2.74 (1.46,5.17)	38 (13,56)	7.57 (1.97,29.06)	22 (5,36)
1 vs 5 (GA)	3.47 (1.69,7.14)	53 (20,72)	8.65 (3.92,19.09)	44 (7,66)	2.86 (1.63,5.02)	40 (19,55)	12.02 (3.65,39.57)	32 (17,45)
1 vs 45 (CNV)	2.48 (1.57,3.93)	40 (21,55)	5.60 (3.21,9.78)	32 (13,47)	3.30 (2.17,5.01)	45 (33,55)	15.34 (5.32,44.25)	38 (30,46)
1 vs 4 (CNV)	2.78 (1.61,4.80)	44 (20,61)	5.24 (2.74,10.01)	30 (4,50)	2.44 (1.53,3.90)	34 (17,47)	6.58 (2.07,20.90)	19 (9,28)
1 vs 5 (CNV)	2.17 (1.22,3.86)	34 (6,54)	6.00 (3.12,11.53)	34 (7,53)	5.24 (3.02,9.10)	60 (43,72)	35.22 (11.47,108.17)	60 (46,70)
CHS	1.82 (1.13,2.92)	27 (1,46)	4.22 (2.39,7.42)	25 (3,42)	1.58 (1.05,2.39)	17 (21,32)	4.75 (2.56,8.80)	14 (1,25)

NOTE.—95% CIs are given in the parentheses. Results for the *ELOVL4* and *PLEKHA1* genes are given in table B2. Results for evaluations of dominance and recessive effects are given in table B2.

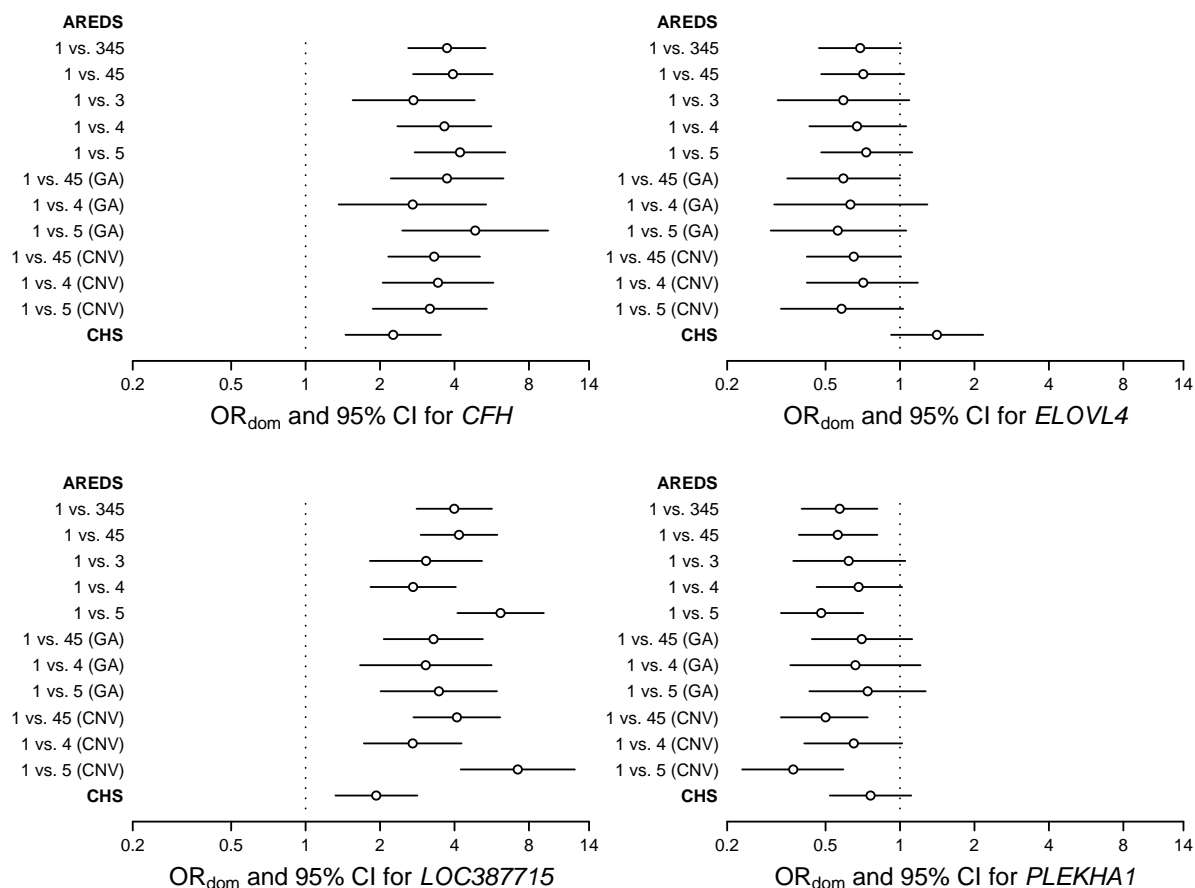


Figure 4.1: Estimated crude ORs and 95% CIs for *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715* genes. Carriers of one or two risk alleles (RR+RN) are compared with those subjects homozygous for the non-risk allele (NN). The solid lines denote the 95% CI corresponding to an OR (open circle). The dotted vertical line marks the null value of an OR of 1. The contrasts that were evaluated in AREDS and CHS cohorts are given on the vertical axis.

**4.3.1.2 *ELOVL4*** The *M299V* variant in *ELOVL4* is significantly associated ( $P = 0.034$ ) with exudative ARM in the AREDS sample (Table 4.3), in agreement with our previous findings (CONLEY *et al.* 2005). However, no ORs are statistically significant at 95% significance level (Fig. 4.1 and Table B2 and Fig. B2). These results do not exclude the potential role of *ELOVL4* in

ARM, but do not strongly support it. The small number of individuals with exudative ARM did not allow for subphenotype analysis in the CHS cohort.

**4.3.1.3 PLEKHA1 and LOC387715** The association of the *S69A* variant in *LOC387715* with all presentations of ARM is extremely significant ( $P \leq 0.00001$ ) in both the AREDS and CHS data sets (Table 4.3), confirming earlier findings by ourselves (JAKOBSDOTTIR *et al.* 2005) and others (RIVERA *et al.* 2005; SCHMIDT *et al.* 2006). The *A320T* variant in *PLEKHA1*, which is located on the same haplotype block as *LOC387715*, is highly significant ( $P = 0.00004$ ) in the AREDS sample but only borderline significant ( $P = 0.08$ ) in the CHS sample. The degree of linkage disequilibrium between *A320T* and *S69A* is statistically significant in both AREDS ( $D' = 0.66$ ) and CHS ( $D' = 0.65$ ) controls. In order to identify which gene, *PLEKHA1* or *LOC387715*, more likely harbors the true ARM-predisposing variant, we applied the haplotype method (VALDES and THOMSON 1997). According to the haplotype method, the relative frequency of alleles at neutral variants is expected to be the same in cases and controls for a haplotype containing all the predisposing variants. The results based on applying the method suggest that *S69A* in *LOC387715*, and not *A320T* in *PLEKHA1*, is an ARM-predisposing variant (Appendix B., Distinguishing between *PLEKHA1* and *LOC387715*–Results). Further, by permutation testing, the null hypothesis:  $H_0$  : the *S69A* variant in *LOC387715* fully accounts for the ARM predisposition to the *PLEKHA1*–*LOC387715* haplotype block, is not rejected ( $P = 0.92$  in the AREDS data,  $P = 0.45$  in the CHS data), while a similar hypothesis for *A320T* is rejected ( $P \leq 0.0001$  in the AREDS data,  $P = 0.0002$  in the CHS data).

The *S69A* variant in *LOC387715* shows different risk patterns than *Y402H* in *CFH*. The variant appears to increase the risk of severe ARM substantially more than the risk of mild ARM (fig. 4.1, table B2 and fig. B4) in the AREDS data where severity of disease is differentiated. For example, the OR for AREDS cases of grade 3, who carry one or two T alleles, is 3.07 (95% CI 1.82–5.17), while the OR for AREDS cases, with CNV in both eyes, who carry one or two T alleles, is 7.21 (95% CI 4.24–12.27). Similar to *CFH*, *S69A* shows an allele-dose effect without dramatic differences in the PAR of the GT and TT genotypes (table 4.4 and fig. B4). Since only four AREDS controls are TT homozygous at *S69A*, point estimates and CIs, for recessive and homozygote contrasts, derived from regular logistic regression were compared with estimates from exact regression [models fitted in SAS software release 8.2 (SAS Institute Inc., Cary, NC, USA)]. These quality checks revealed no major differences in point estimates (which is the basis of the

PAR estimates) and lower confidence limits (which is the basis of comparison with the ORs), but the upper confidence limits were higher (results not shown).

### 4.3.2 Interaction analyses

We used logistic regression modeling to build a model of the joint contribution of *CFH* and *LOC387715*, *CFH* and cigarette smoking and *LOC387715* and cigarette smoking. A series of models were fitted in order to draw inferences about the most likely and most parsimonious model(s). As described by [NORTH \*et al.\* \(2005\)](#), models were compared using the Akaike information criterion (AIC). When the most parsimonious model had been identified we estimated joint ORs of the risk factors. Separate estimates were calculated from each cohort. In order to maximize the AREDS sample size, no subphenotype or subgrade analyses were performed; AREDS cases of grade 3–5 were compared with AREDS controls of grade 1.

In a previous article ([JAKOBSDOTTIR \*et al.\* 2005](#)), we found no evidence of interacting effects of the *CFH* and *PLEKHA1/LOC387715* loci; the joint action of the two loci was best described by independent multiplicative effects (additive on a log-scale). [RIVERA \*et al.\* \(2005\)](#) reported that *S69A* in *LOC387715* acted independently of *Y402H* in *CFH*. [SCHMIDT \*et al.\* \(2006\)](#) also arrived at the same most parsimonious model. The AREDS and CHS data also suggest that the two genes contribute independently to disease risk. The best fitting model (the model with the smallest AIC) derived from the AREDS data is an additive model with an interaction term. This model, with AIC of 721.4, does however not provide a significantly better fit (AIC difference  $< 2$ ) than a simpler additive model with AIC of 723.0. The additive model is the most parsimonious model (AIC=635.1) derived from the CHS data and is also the best fitting model (Table 4.5). Joint ORs for combinations of risk genotypes at *Y402H* and *S69A* were computed to further understand the joint action of the two loci (Table B4). Using all cases regardless of severity, the AREDS data suggest that individuals heterozygous for the risk allele at one of the loci and homozygous for the non-risk allele at the other are more susceptible to ARM than individuals with no-risk allele at both loci (for the CT–GG joint genotype, OR 2.8, 95% CI 1.6–5.0; for the TT–GT joint genotype, OR 3.2, 95% CI 1.7–6.0). The ARM risk more than doubles if a person is heterozygous at both loci (for the CT–GT joint genotype, OR 7.2, 95% CI 3.8–13.5) and being homozygous for the risk allele for at least one of the loci further increases the risk. The joint ORs estimated from the CHS data show a similar pattern, but having only one risk allele is not sufficient to increase the risk (for

the CT–GG joint genotype, OR 1.3, 95% CI 0.6–2.7; for the TT–GT joint genotype, OR 1.2, 95% CI 0.5–2.8).

A recent study (SCHMIDT *et al.* 2006) reported a strong statistical interaction between genotypes at *S69A* and smoking, both on binary (ever versus never smoked) and continuous scale (pack-years of smoking). We fail to replicate this finding in both the AREDS and CHS data sets (table 4.5). Results from the AREDS sample suggests that the joint effects of *Y402H* and smoking are best described by independent multiplicative effects, without significant dominance or interacting effects. On the other hand, the model that best describes the CHS data includes only additive effects of *Y402H*. Results from the AREDS data suggest that the joint effects of *S69A* and smoking are best described by independent multiplicative effects, without significant dominance or interacting effects. The CHS data implicate a model with only *S69A*. When smoking exposure is a continuous variable (pack-years of smoking) and the *S69A* genotypes are coded in additive fashion, the interaction term is not significant ( $P = 0.40$ ) in the CHS data. Pack-years of cigarette smoking were not available for participants in the AREDS study. To further understand the combined effect of the genes and cigarette smoking, joint ORs of risk genotypes at each gene and smoking were estimated from the AREDS data (table B7). The results suggest that, while the risk of ARM due to any of the risk genotypes (at *Y402H* and *S69A*) is elevated in smokers, both genes have substantially more influence on ARM risk than cigarette smoking. Both the model fitting approach and a simple 2 test ( $P = 0.71$ ) show that the main effects of cigarette smoking are insignificant (on binary scale) in the CHS data.

### 4.3.3 *APOE* results

Main effects of the *APOE* gene in ARM were tested using the CHS data. Neither the distribution of *APOE* –  $\epsilon 4$  carriers ( $P = 0.41$ ) nor *APOE* –  $\epsilon 2$  ( $P = 0.42$ ) carriers was significantly different between cases and controls, when compared to *APOE* –  $\epsilon 3/\epsilon 3$ .

### 4.3.4 Meta-analyses

**4.3.4.1 Meta-analysis of *CFH*** We used a meta-analysis approach to pool estimated ORs for *Y402H* from 11 independent data sets [including the CHS and AREDS cohorts reported here (Table B10)]. This resulted in the analysis of 5451 cases and 3540 controls all of European or European



Table 4.5: Results of fitting two-factor models by logistic regression

Two-Factor Model	AREDS data		CHS data	
	AIC	AIC difference	AIC	AIC difference
Y402H (Factor 1) and S69A (Factor 2)				
ADD1	799.3	77.9	652.7	17.6
ADD2	786.1	64.7	656.0	21.0
ADD-BOTH	723.0	1.7	635.1	0.0
DOM1	801.2	79.8	654.4	19.3
DOM2	786.9	65.5	656.0	21.0
DOM-BOTH	726.5	5.1	636.3	1.3
ADD-INT	721.4	0.0	635.8	0.8
ADD-DOM	724.3	3.0	638.8	3.8
DOM-INT	...	...	637.8	2.8
Y402H (Factor 1) and Smoking (ever vs. never)				
ADD1	787.3	6.0	677.3	0.0
SMOKE	848.3	67.0	700.6	23.3
ADD1-SMOKE	781.3	0.0	679.1	1.8
DOM1	789.3	8.0	679.0	1.7
ADD1-SMOKE-INT	783.2	1.8	678.3	1.0
DOM1-SMOKE-INT	786.6	5.3	681.9	4.6
S69A (Factor 2) and Smoking (ever vs. never)				
ADD2	774.0	6.1	745.6	0.1
SMOKE	842.9	75.0	765.2	19.8
ADD2-SMOKE	767.9	0.0	747.3	1.8
DOM2	774.7	6.7	745.5	0.0
ADD2-SMOKE-INT	769.7	1.8	749.1	3.7
DOM2-SMOKE-INT	772.4	4.4	748.9	3.4

NOTE.—95% CIs are given in the parentheses. Results for the *ELOVL4* and *PLEKHA1* genes are given in table B2.

Results for evaluations of dominance and recessive effects are given table B2.

American descent. The results confirm the increased ARM risk due to the C allele in the non-Hispanic white population (fig. 4.2 and table B11). The pooled estimates have narrower CI than any individual study, and non-overlapping CI for hetero- and homozygote ORs:  $OR_{\text{het}}=2.43$  (95% CI 2.17–2.72) and  $OR_{\text{hom}}=6.22$  (95% CI 5.38–7.19), when assuming homogeneity across studies. When the analysis is performed under heterogeneity, the point estimates are essentially the same and the CIs are slightly wider. Leave-one-out sensitivity analyses, under a fixed effect model, show that no study has dramatic influence on the pooled estimates (table B11). The study by RIVERA *et al.* (2005) changes the estimates more than any other study; when the study is excluded, the  $OR_{\text{dom}}$  and  $OR_{\text{het}}$  are approximately 0.2 higher, while the  $OR_{\text{rec}}$  and  $OR_{\text{hom}}$  are lowered by approximately 0.2. The Rivera *et al.* study is the only study where the genotype distribution, in the control group, deviates from Hardy–Weinberg equilibrium [HWE ( $P = 0.03$ )]. The allele and genotype distributions, in cases and controls, are strikingly similar across studies. However, the genotype distribution in CHS cases differs from the other studies and the frequency of the CC risk genotype is lower compared to other cohorts (fig. B5).

**4.3.4.2 Meta-analysis of *LOC387715*** Meta-analysis of the risk associated with *S69A* in ARM included five independent data sets [including the CHS and AREDS cohorts reported here (Table B12)]. This resulted in the analysis of 3147 cases and 2381 controls all of European or European American descent. The studies of *LOC387715* are more heterogeneous than the studies of *CFH*;  $OR_{\text{dom}}$  and  $OR_{\text{het}}$  differ significantly across studies ( $P < 0.01$  and 0.02, respectively). The results support earlier findings of the association of the T allele with increased ARM risk (table B13). Carriers of two T alleles are at substantially higher risk than are carriers of one T allele; when accounting for between-study variation, the  $OR_{\text{het}}$  and  $OR_{\text{hom}}$  are 2.48 (95% CI 1.67–3.70) and 7.33 (95% CI 4.33–12.42), respectively. The genotype distribution is similar across all control populations and across all ARM populations, except the CHS ARM population (fig. B6).

## 4.4 DISCUSSION

During the past year, major discoveries of associations of the *CFH* and *PLEKHA1/LOC387715* genes with ARM were published. A number of reports established a strong association of the *Y402H* coding change in *CFH* with ARM and three reports found an association, of similar magnitude as

the association of *Y402H*, of the *S69A* coding change in *LOC387715* with ARM. Both of those genes lie within chromosomal regions, *CFH* on 1q31 and *LOC387715* on 10q26, consistently identified by family-based linkage studies (SEDDON *et al.* 2003; MAJEWSKI *et al.* 2003; IYENGAR *et al.* 2004; WEEKS *et al.* 2001; WEEKS *et al.* 2004; KLEIN *et al.* 1998; KENEALY *et al.* 2004).

Because the majority of the studies of *Y402H* and all three studies of *S69A* were specially designed to search for (and find) genes involved in ARM complex etiology, it is possible that they overestimate the effect size of the risk alleles at *Y402H* and *S69A*. Therefore, we analyzed two independent case-control cohorts with varying inclusion and exclusion criteria based on ARM status, the AREDS and CHS cohorts. The AREDS cohort did have inclusion and exclusion criteria relevant to severity of ARM and both affected and non-affected individuals were enrolled (AGE RELATED EYE DISEASE STUDY GROUP 1999). In contrast, the CHS cohort is a population-based cohort that utilized community-based recruitment of individuals 65 years and older with no inclusion and exclusion criteria relevant to ARM status (FRIED *et al.* 1991). Retinal assessments in the CHS cohort were not conducted until the 8-year follow-up visit. Given the difference in ascertainment of subjects into the two studies, replication of association of a candidate gene in both cohorts greatly strengthens the support for its causal involvement in ARM pathogenesis.

We evaluated previously reported associations of four genes, *CFH* (1q31), *ELOVL4* (6q14), *PLEKHA1* (10q26) and *LOC387715* (10q26). Variants in both *CFH* and *LOC387715* are extremely significantly ( $P \leq 0.00001$ ) associated with ARM in both AREDS and CHS cohorts. Both variants show an allele-dose effect on the ARM risk and a model of independent multiplicative contribution of the two genes is most parsimonious in both AREDS and CHS cohorts. The *A320T* coding change in the *PLEKHA1* gene, adjacent to and in linkage disequilibrium with *LOC387715* on 10q26, is significantly associated with ARM in the AREDS cohort ( $P = 0.00004$ ), but not in the CHS cohort ( $P = 0.08$ ). Because of extensive linkage disequilibrium between *PLEKHA1* and *LOC387715* in our initial study population we could not, with reasonable certainty, distinguish between their association signals. Our results based on applying the haplotype method to both the AREDS and CHS cohorts, combined with the findings of RIVERA *et al.* (2005), who used conditional haplotype analysis and detected, for the first time, a weak expression of *LOC387715* in the retina, and SCHMIDT *et al.* (2006), who detected only a weak association signal at *PLEKHA1*, indicate that *S69A* in *LOC387715* is most likely the major ARM-predisposing variant on 10q26. The results of the haplotype method show that *PLEKHA1* is not sufficient to account for the ARM-

predisposition at 10q26; however, we cannot exclude the possibility that *A320T* in *PLEKHA1* may be on a causative haplotype with *S69A* and other unknown variants.

The replication of associations of *CFH* and *LOC387715* genes with ARM in AREDS and CHS cohorts, two cohorts with different study designs, continues to provide strong support for their involvement in ARM. Variable findings for *PLEKHA1* in AREDS and CHS cohorts do however need to be considered in the light of differences between the two cohorts. In addition to differences in ascertainment of the case and control populations, the evaluation of retinal changes, documentation of retinal findings and prevalence of advanced ARM differed between the two cohorts. In the CHS study, fundus photography was only available for one randomly selected eye and the photography was performed with non-dilated pupils and these limitations could certainly influence the sensitivity to detect disease pathology, although this is more likely to influence the detection of early retinal changes. The proportion of advanced ARM in the entire CHS cohort that was evaluated at the 8-year follow-up evaluation was 1.3% (KLEIN *et al.* 2003) compared to 17% in the AREDS (AGE RELATED EYE DISEASE STUDY GROUP 2000) and the variation in the proportion of advanced ARM disease pathology between the two cohorts could lead to variation in findings, especially if a gene is more likely to influence progression of the disease. In addition, one important difference between these two cohorts is the timing of the retinal evaluations. AREDS participants had retinal evaluations conducted at baseline as well as during follow-up evaluations, whereas CHS participants had retinal evaluations done eight or more years after enrollment, when they would have been at least 73 years old. It is possible that survival to the retinal evaluation for the CHS participants could bias the population available for this particular type of study. Potential confounding issues related to the use of the AREDS cohort are that subjects in categories other than the unaffected group were randomized into a clinical trial using vitamin and mineral supplements to evaluate the impact of these on ARM progression and there is some evidence indicating that unaffected subjects in category 1 have different demographic characteristics than affected subjects in the other categories (AGE RELATED EYE DISEASE STUDY GROUP 2000). It is not clear whether these could impact the results of our study, but it should be considered when findings are interpreted.

As mentioned previously, most studies that have investigated the genetic etiology of ARM were designed to optimize identification of regions of the genome housing susceptibility genes for ARM and for ARM candidate gene testing. Published attributable risks range from 43 to 68% (EDWARDS *et al.* 2005; HAINES *et al.* 2005; JAKOBSDOTTIR *et al.* 2005; SCHMIDT *et al.* 2006) for the *Y402H* variant in *CFH* and from 36 to 57% (JAKOBSDOTTIR *et al.* 2005; SCHMIDT *et al.*

2006) for the *S69A* variant in *LOC387715*. Interestingly, the PARs for the CHS population are lower than those previously published: 41% for the *Y402H* variant in *CFH* and 27% for the *S69A* variant in *LOC387715* (table B2). Because the majority of the CHS cases have moderate ARM the PAR estimates derived from the CHS data are not completely comparable with estimates from previous studies in which the proportion of patients with advanced ARM was considerably higher. However, they are comparable to estimates derived from using AREDS cases of grade 3. Those estimates are within the previously published range of PARs: 49% for *Y402H* in *CFH* and 45% for *S69A* in *LOC387715*. These findings may indicate that the ARM attributed to these two susceptibility variants may be lower than previously thought, given that the CHS cohort was not ascertained based on ARM status. A prospective design is needed to more precisely estimate the relative risks, which are approximated by ORs estimated from retrospective case-control designs, and corresponding PARs.

We were not able to replicate the association of *ELOVL4* with overall ARM (CONLEY *et al.* 2005). The number of individuals with exudative ARM allowed us to perform subphenotype analysis in the AREDS, but not the CHS cohort. Subphenotype analysis was especially important with regard to *ELOVL4*, where our previous findings indicated a role for *ELOVL4* in exudative ARM; this is trending towards significance in the AREDS cohort. Given the lack of strong association and significant ORs for *ELOVL4* in ARM susceptibility in both cohorts and the lack of association reported by Ayyagari *et al.*, it is very unlikely that *ELOVL4* plays a substantial role in ARM susceptibility. The power to detect an OR of 0.6 for overall ARM is reasonable, with type I error rate 5%, minor allele frequency 0.15 and population prevalence 6% the power is  $\sim 81\%$  in AREDS and  $\sim 69\%$  in CHS. The power to detect the same effect in exudative ARM is only 53% in AREDS data, under the same conditions. Therefore, the possibility that *ELOVL4* plays a role in overall ARM is unlikely but mild effect in exudative ARM cannot be refuted. These power estimates were performed using Quanto (32).

We also used the CHS cohort to test whether the  $\epsilon 4$  or  $\epsilon 2$  alleles of the *APOE* gene are associated with ARM. In several studies, the  $\epsilon 2$  allele is suggested to contribute to disease risk and the  $\epsilon 4$  allele has been found to protect from ARM. Our results do not reach statistical significance and do not support the hypothesized role of the gene in ARM pathogenesis.

The AREDS and CHS data support the independent contribution of *Y402H* in *CFH* and *S69A* in *LOC387715* to ARM susceptibility. A multiplicative risk model for these two variants is

the most parsimonious based on evaluation of the AREDS and CHS cohorts; this model was also supported by our previous paper (JAKOBSDOTTIR *et al.* 2005) as well as data presented by RIVERA *et al.* (2005) and SCHMIDT *et al.* (2006). The ARM risk appears to increase as the total number of risk alleles at *Y402H* and *S69A* increases (table B4).

Prior to the discovery of *CFH* and *LOC387715* cigarette smoking was one of the more important known ARM-related risk factors. Cigarette smoking is generally accepted as a modifiable risk factor for ARM; van Leeuwen *et al.* provide a review of the epidemiology of ARM and discuss the support of smoking as ARM risk factor (VAN LEEUWEN *et al.* 2003). SCHMIDT *et al.* (2006) recently reported statistically significant interaction between *LOC387715* and cigarette smoking in ARM. Their data suggested that the association of *LOC387715* with ARM was primarily driven by the gene effect in heavy smokers. Our own analyses of interaction do not support this finding and the AREDS data suggest that the joint action of *S69A* and smoking is multiplicative.

A role for *CFH* and *LOC387715* in ARM susceptibility is further supported via the results of our meta-analysis. The meta-analysis, which include the CHS and AREDS cohorts reported in this article, indicates that having one or two copies of the risk allele at *CFH* or *LOC387715* increases the risk of ARM, and those who have two copies are at higher risk. The combined results from all studies as well as the results from each independent study were remarkably tight (figs. 4.2 and 4.3). One known limitation of meta-analysis is the susceptibility to publication bias. Generally, such bias is a result of non-publication of negative findings (NORMAND 1999). In the case of *CFH* and *LOC387715*, all published studies have reported strong association with ARM in the same direction, with the risk allele for *CFH* being the allele that codes for histidine and the risk allele for *LOC387715* being the allele that codes for serine. We expect the preferential publication of statistically significant associations to show random directionality if the significant association is a false-positive result (LOHMUELLER *et al.* 2003). It is therefore unlikely that the consistency of the association of *CFH* and *LOC387715* with ARM is a result of publication bias.

While the results of our statistical analyses are in agreement with *LOC387715* being the major ARM-related gene on 10q26, they do not prove causality. The possible causal role of *CFH* in ARM pathogenesis has been further supported by the localization of its protein within drusen deposits of ARM patients and involvement in activation of the complement pathway. Regarding *LOC387715*, little is currently known about the biology of the gene and nothing about how its

protein may affect ARM susceptibility. Until recently the expression of *LOC387715* appeared limited to the placenta, but recently weak expression was reported in the retina ([RIVERA \*et al.\* 2005](#)), which opens up the possibility of a tissue-specific role of the gene.

In summary, our results continue to support a role of both *CFH* and *LOC387715* in etiology of ARM, given that both genes harbor variants highly associated with ARM, regardless of how the subjects were ascertained. Evaluation of *PLEKHA1* and *ELOVL4* in the AREDS and CHS cohorts demonstrates that these genes are much less likely to play role in ARM susceptibility. The *CFH* and *LOC387715* genes appear to act independently in a multiplicative way in ARM pathogenesis and individuals homozygous for the risk alleles at either locus are at highest risk. The continued support for these genes in ARM susceptibility will hopefully bring us closer to being able to utilize the information in these genes to identify at risk individuals and provide a rational basis for future clinical trials to test preventive therapies in high-risk cohorts as well as to provide insights into the basic pathogenesis of this condition.

## 4.5 MATERIALS AND METHODS

### 4.5.1 Cardiovascular health study (CHS) participant sampling and phenotyping

CHS is a population-based, longitudinal study primarily designed to identify factors related to cardiovascular disease in those aged 65 and older. Retinal assessments were performed at the 8-year follow-up visit. Community-based recruitment took place in Forsyth County, NC; Sacramento County, CA; Washington County, MD; and Pittsburgh, PA. Medicare eligibility lists of the Health Care Financing Administration were utilized to identify individuals who were aged 65 and older. Individuals aged 65 years and older living in the households of list members were also eligible. Inclusion criteria were minimal and included being non-institutionalized, expected to remain in the area for at least 3 years, able to give informed consent, not wheelchair-bound, not receiving hospice care and not receiving radiation or chemotherapy for cancer ([FRIED \*et al.\* 1991](#)). DNA samples from the CHS from participants who consented for genetic studies were used for this research. Only DNA samples from subjects who had a retinal examination where the findings fit our criteria of a case or control were included in this study.

CHS subjects usually had the retina of one randomly selected eye photographed and the photographs were graded by Dr. Gorin using the same classification model that was described in prior publications ([WEEKS \*et al.\* 2004](#)). Only Caucasian individuals are included in the analysis, as the sample size of other groups with ARM is too small for reasonable results: there were 180 black controls but only three cases, and five controls of other races. All CHS cases (n=126) used for analyses are “Type A” which falls into our most stringent model for clinical classification ([WEEKS \*et al.\* 2004](#)). Individuals in this category are clearly affected with ARM based on extensive and/or coalescent drusen, pigmentary changes (including pigment epithelial detachments) and/or the presence of end-stage disease (GA and/or CNV membranes). Very few CHS cases had end-stage ARM, GA or CNV (Table 4.1); therefore, analyses of specific subtypes of ARM were not conducted. All CHS controls (n=1051) were of AREDS grade 1. A few potential controls (n=22) had unclear signs of GA or CNV and were excluded from analyses.

#### **4.5.2 Age-related eye disease study (AREDS) participantssampling and phenotyping**

AREDS is a prospective, multicenter study of the natural history of ARM and age-related cataract with a clinical trial of high-dose vitamin and mineral supplementation embedded within the study. Individuals recruited into the AREDS study were men and women aged 55–80 years at enrollment; these individuals were required to be free of any condition or illness that would hinder long-term follow-up. Inclusion criteria were minimal and included having ocular media clear enough to allow for fundus photography and either no evidence of ARM in either eye or having ARM in one eye while the other maintained good vision (20/30 or better) ([AGE RELATED EYE DISEASE STUDY GROUP 1999](#)). DNA samples from subjects who consented for genetic studies from the NEI-AREDS Genetic Repository were used for this research.

ARM status was assigned using the AREDS ARM grading system and based on phenotypes assigned at the most recent follow-up visit. Again, only Caucasian individuals are included in the analysis, as the sample size of other groups is too small for reasonable results: there are only 15 African American, two Hispanic and three individuals of other races. AREDS cases (n=701) consisted of grade 3, 4 and 5. AREDS subjects of grade 3 (n=96) have ARM but do not suffer from end-stage ARM, subjects of grade 4 (n=266) have end-stage ARM in one eye and subjects of grade 5 (n=339) have end-stage ARM in both eyes. AREDS controls (n=175) have AREDS grade 1 (grade 2 individuals were excluded prior to analyses).



### 4.5.3 Genotyping

The *M299V* variant in *ELOVL4* (*rs3812153*), the *Y402H* variant in *CFH* (*rs1061170*) and the *S69A* variant in *LOC387715* (*rs10490924*) were genotyped using RFLP techniques. The primers, annealing temperatures and restriction endonuclease for each assay were: 5'-AGATGCCGATGTTGTTAAAAG-3' (F), 5'-CATCTGGGTATGGTATTAAC-3' (R), 50 °C and *BspHI* for *ELOVL4*; 5'-TCTTTTTGTGCAAACCTTTGTTAG-3' (F), 5'-CCATTGGTAAAACAAGGTGACA-3' (R), 52 °C and *NlaIII* for *CFH*; 5'-GCACCTTTGTCCACCACATTA-3' (F), 5'-GCCTGATCATCTGCA TTTCT-3' (R), 54 °C and *PvuII* for *LOC387715*.

The *A320T* variant in *PLEKHA1* (*rs1045216*) was genotyped using 5' exonuclease Assay-on-Demand TaqMan assays (Applied Biosystems Incorporated). Amplification and genotype assignments were conducted using the ABI7000 and SDS 2.0 software (Applied Biosystems Incorporated). For all genotyping conducted for this research, double-masked genotyping assignments were made for each variant, compared and each discrepancy addressed using raw data or by re-genotyping.

### 4.5.4 Association analyses

SNP-disease association was measured with allele- and genotype  $\chi^2$  tests, and P-values were simulated using 100,000 replicates; in cases with one or more expected cell numbers less than five, the Fisher's exact test was used. The strength of the association was estimated by crude OR and PAR. A general formula was used to calculate the PAR:  $PAR = P_r(OR - 1) / (1 + P_r(OR - 1))$ , where  $P_r$  is the prevalence of the risk factor in the general population. Estimates of  $P_r$  were derived from the CHS controls; this is reasonable, because the CHS subjects were not selected on the basis of ARM disease status, and the number of CHS controls is large ( $n=1,051$ ). Confidence intervals for the PARs were derived using asymptotic normal distribution of  $\log(1 - PAR)$  and transforming to an interval for the PAR. The CIs derived in this way are likely to be too narrow when the risk factor is rare ( $P_r < 0.1$ ) and sample sizes are small (WALTER 1975). For comparison purposes, ORs adjusted ( $OR_{adj}$ ) for age and gender were estimated. Logistic regression models were used to calculate both crude and adjusted ORs, using R (37). The less frequent allele in the control group was considered the risk allele, and the OR and  $OR_{adj}$  were calculated by comparing those homozygous for the risk allele (RR) to the baseline group [those homozygous for the normal allele (NN)] and comparing

those heterozygous for the risk allele (RN) to the baseline group. The contrasts for dominance (RR and RN versus NN) and recessive (RR versus RN and NN) effects were also evaluated.

#### 4.5.5 Distinguishing between *PLEKHA1* and *LOC387715*

We employed the haplotype method (VALDES and THOMSON 1997) to identify which one of the two loci, *A320T* in *PLEKHA1* or *S69A* in *LOC387715*, is more likely the actual disease predisposing variant in the 10q26 region. The basis of the haplotype method is simple and elegant [for a mathematical proof, see VALDES and THOMSON (1997)]. If all predisposing variants are included on a haplotype, then the neutral variants are expected to be in the same ratio in cases and controls on a particular disease-predisposing haplotype, although the actual frequencies may differ. On the other hand, if not all predisposing variants have been identified, equality in the ratios of haplotype frequencies of non-predisposing variants is not expected. The expected ratios for the *A320T-S69A* haplotype are formulated in Appendix B.. Two null hypotheses were tested: one that *A320T* fully accounts for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block, and the other that *S69A* fully accounts for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block (for details on the hypotheses and permutation procedure to generate *P* values, see the Appendix B.). The program SNP HAP (CLAYTON ) was used to estimate haplotype frequencies and individual haplotypes. SNP HAP uses the EM algorithm to calculate a maximum likelihood estimate of haplotype frequencies given the unphased genotype data. The posterior probabilities of individual haplotype assignments exceed 87% for every individual typed at both *A320T* and *S69A*. For 80% of the haplotype assignments the underlying genotype at one or both loci is homozygous and hence the posterior probability is 100%.

#### 4.5.6 Interaction analyses

The analyses of interaction were three-fold: first, we tested for interacting genetic effects of *Y402H* in *CFH* and *S69A* in *LOC387715* in both CHS and AREDS samples, then we tested for interaction of both *Y402H* and *S69A* with smoking history in both CHS and AREDS samples and finally we calculated joint ORs of the three risk factors.

We followed a modeling strategy proposed by NORTH *et al.* (2005). Series of logistic regression models are fitted to the AREDS and CHS data sets in order to find the model that best describes

the joint effects of *CFH* and *LOC387715*. For each genotype, models allowing for additive effects (ADD1, ADD2 and ADD-BOTH), and models which incorporate dominance effects (DOM1, DOM2 and DOM-BOTH) are fitted. The ADD1 model includes only the term  $x_1$  for additive effects of *CFH*, coded as  $-1$  for genotype TT at *Y402H*, as 0 for genotype CT and as 1 for genotype CC. The ADD2 includes only model term  $x_2$  for additive effects of *LOC387715*, coded as  $-1$  for genotype GG at *S69A*, as 0 for genotype GT and as 1 for genotype TT. The ADD-BOTH models the joint additive effects of *CFH* and *LOC387715*. The DOM1 incorporates dominance effects to ADD1, and includes  $x_1$  and  $z_1$ , coded as 0.5 for genotype CT and  $-0.5$  for genotypes TT and CC at *Y402H*. The DOM2 model similarly incorporates dominance effects to ADD2, and includes  $x_2$  and  $z_2$ , coded as 0.5 for genotype GT and  $-0.5$  for genotypes GG and TT at *S69A*. DOM-BOTH models the joint dominance effects of *CFH* and *LOC387715*. Three further models, that model the interaction between *CFH* and *LOC387715* are fitted: ADD-INT includes the product term  $x_1 * x_2$ , ADD-DOM includes  $x_1 * x_2$ ,  $x_1 * z_2$  and  $z_1 * x_2$  and DOM-INT includes  $x_1 * x_2$ ,  $x_1 * z_2$ ,  $z_1 * x_2$  and  $z_1 * z_2$ .

The above modeling strategy was modified to investigate the joint effects of *CFH* and smoking, and the joint effects of *LOC387715* and smoking. The modified approach is the same as used by [SCHMIDT \*et al.\* \(2006\)](#) to test for interaction between *LOC387715* and smoking. The coding scheme is the same, as above, except that smoking is coded as 0 for never smokers and 1 for ever smokers. The models fitted for the effects of *CFH* and smoking are: ADD1, SMOKE, ADD1-SMOKE, DOM1, ADD1-SMOKE-INT and DOM1-SMOKE-INT, and the models fitted for the effects of *LOC387715* and smoking are: ADD2, SMOKE, ADD2-SMOKE, DOM2, ADD2-SMOKE-INT and DOM2-SMOKE-INT.

All models were compared by the AIC. Models for which the AIC differed by  $< 2$  are considered indistinguishable ([NORTH \*et al.\* 2005](#)), and the model with fewer parameters was chosen as the most parsimonious model. Since adjusting for age and gender did not affect the estimates of ORs for *Y402H* nor *S69A* (table [B3](#)), and to keep number of parameters as small as possible, no adjustment was made for these covariates when modeling interaction. Based on the results of the above interaction analyses, joint ORs were calculated.

#### 4.5.7 *APOE* analyses

Previous studies have reported possible protective and harmful effects of the apolipoprotein E (*APOE*) gene in ARM. The  $\epsilon 4$  allele may have protective effects ([KLAVER \*et al.\* 1998](#); [SCHMIDT](#)

*et al.* 2000; SCHMIDT *et al.* 2002; BAIRD *et al.* 2004; ZAREPARSI *et al.* 2004), whereas the least frequent allele,  $\epsilon 2$ , may increase the risk of ARM (KLAVER *et al.* 1998; ZAREPARSI *et al.* 2004). The *APOE* variant was genotyped by CHS and its association with ARM was assessed in this study. Individuals were classified by *APOE* genotype into individuals with *APOE*- $\epsilon 3/\epsilon 3$  genotype, and *APOE*- $\epsilon 2$  and *APOE*- $\epsilon 4$  carriers (denoted *APOE*- $\epsilon 2/*$  and *APOE*- $\epsilon 4/*$ , respectively); individuals with *APOE*- $\epsilon 2/\epsilon 4$  genotype were included in both the *APOE*- $\epsilon 2/*$  and *APOE*- $\epsilon 4/*$  groups.  $\chi^2$  tests were used to test for differences in distributions of *APOE*- $\epsilon 3/\epsilon 3$  and *APOE*- $\epsilon 2/*$ , and *APOE*- $\epsilon 3/\epsilon 3$  and *APOE*- $\epsilon 4/*$ , genotypes in controls and cases.

#### 4.5.8 Meta-analyses

We undertook a meta-analysis approach to pool estimated OR from previously published reports on *CFH* and *LOC387715* and the two reports presented here. Initially data were analyzed, assuming the between-study variation is due to chance, and fixed-effects model was employed. Under the fixed-effect model, the maximum likelihood estimator of the pooled OR is an average of individual estimates, weighted by the inverse of their variances, and the variance of the pooled OR is estimated by the inverse of the sum of individual weights. Meta-analyses under homogeneity were performed in R (R DEVELOPMENT CORE TEAM 2005). The assumption of homogeneity was checked using a  $\chi^2$  test. However, tests of homogeneity tend to have low power, and therefore, for comparison, we also pooled the OR in a random effects setting. Meta-analyses under heterogeneity were performed using the method of restricted maximum likelihood (REML), as implemented in SAS Proc Mixed [SAS software release 8.2 (SAS Institute Inc.)]. The pooled REML estimator is identical to the DerSimonian-Laird estimator (DERSIMONIAN and LAIRD 1986; VAN HOUWELINGEN *et al.* 2002). The SAS codes by van Houwelingen *et al.* (VAN HOUWELINGEN *et al.* 2002) were modified to perform the analyses under heterogeneity. A literature search was performed in PubMed in May 2006 and was limited to the English language. *CFH* studies were found by entering the search phrase: (CFH or ‘Complement Factor H) and (‘Age-related macular degeneration’ or ‘Age-related maculopathy or AMD or ARM). Similarly, *LOC387715* studies were found using the search phrase: LOC387715 and ‘Age-related macular degeneration or ‘Age-related maculopathy or AMD or ARM. The only inclusion criterion was that the research participants were Caucasian.

The *Y402H* variant within *CFH* has been found strongly associated with ARM in 11 studies (EDWARDS *et al.* 2005; HAINES *et al.* 2005; KLEIN *et al.* 2005; HAGEMAN *et al.* 2005; CONLEY *et al.*

2005; ZAREPARSI *et al.* 2005; SEPP *et al.* 2006; RIVERA *et al.* 2005; SOUIED *et al.* 2005; MAGNUSSON *et al.* 2006; JAKOBSDOTTIR *et al.* 2005; SCHMIDT *et al.* 2006); two of these 11 studies are ours, so only the results from our JAKOBSDOTTIR *et al.* (2005) paper, that evaluated all contrasts, were used in meta-analysis. The KLEIN *et al.* (2005) study used a small subset of the AREDS sample, and the MAGNUSSON *et al.* (2006) paper only reported allele-based ORs and no genotype counts. Therefore, these two studies were not included. Results from the HAINES *et al.* (2005) study were included in pooled estimates of ORs for hetero- and homozygotes; genotype counts were not available to evaluate contrasts for dominance and recessive effects. Three studies have reported highly associated variant, *S69A*, within the hypothetical *LOC387715* (JAKOBSDOTTIR *et al.* 2005; RIVERA *et al.* 2005; SCHMIDT *et al.* 2006). All three reports on *LOC387715* were included in the meta-analysis. Research participants in all studies of *CFH* and *LOC387715* are non-Hispanic whites of European and European American descent. Tables B10 and B12 summarize the studies included in the meta-analyses of *CFH* and *LOC387715*, respectively.

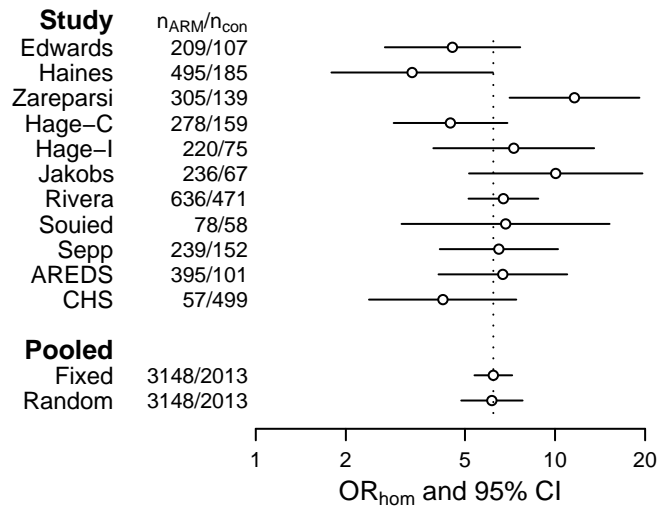
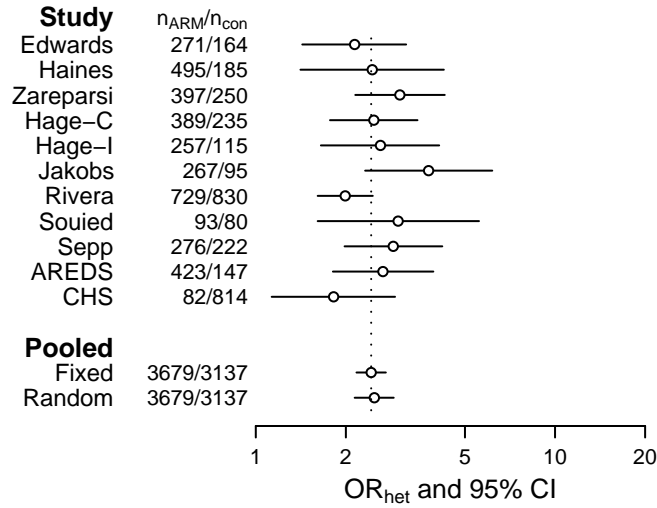


Figure 4.2: Estimated ORs and 95% CIs, derived from data sets included in meta-analysis of *Y402H* in *CFH*, and pooled estimates from fixed and random effect models. The top figure shows OR<sub>het</sub> (OR for CT heterozygotes compared to TT) and the bottom figure shows OR<sub>hom</sub> (OR for CC homozygotes compared to TT). ‘Hage-C’ and ‘Hage-I’ denote estimates derived from the Columbia and Iowa cohorts of Hageman et al., respectively, and Jakobs denotes estimates from the Jakobsdottir et al. paper. ‘Fixed’ denotes pooled estimates derived from all the studies assuming the between-study variability is due to chance. ‘Random’ denotes pooled estimates derived from all the studies allowing for heterogeneity across studies. ‘ $n_{ARM}$ ’ is the total number of ARM cases included in the estimates and ‘ $n_{con}$ ’ is the total number of controls without ARM included in the estimates. For the Haines et al. study ‘ $n_{ARM}$ ’ and ‘ $n_{con}$ ’ refer to the whole sample (individuals of all genotypes). The dotted vertical line marks the point estimate of the pooled OR under homogeneity (‘Fixed’).

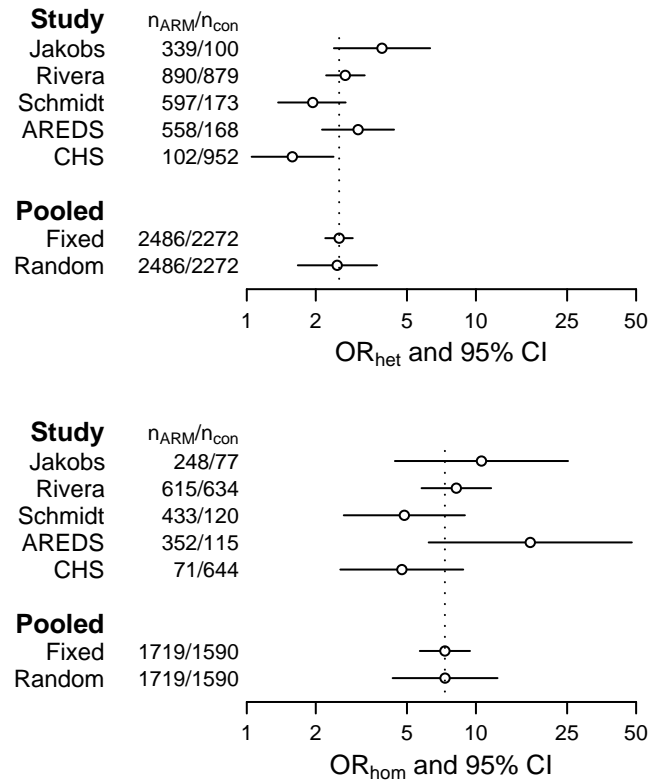


Figure 4.3: Estimated ORs and 95% CIs, derived from data sets included in meta-analysis of *S69A* in *LOC387715*, and pooled estimates from fixed and random effect models. The top figure shows  $OR_{het}$  (OR for GT heterozygotes compared to GG) and the bottom figure shows  $OR_{hom}$  (OR for TT homozygotes compared to GG). ‘Jakobs denote estimates from the Jakobsdottir et al. paper. ‘Fixed denotes pooled estimates derived from all the studies assuming the between-study variability is due to chance. ‘Random denotes pooled estimates derived from all the studies allowing for heterogeneity across studies. ‘ $n_{ARM}$  is the total number of ARM cases included in the estimates and ‘ $n_{con}$  is the total number of controls without ARM included in the estimates. For the Haines et al. study, ‘ $n_{ARM}$  and ‘ $n_{con}$  refer to the whole sample (individuals of all genotypes). The dotted vertical line marks the point estimate of the pooled OR under homogeneity (‘Fixed).

## 5.0 C2 AND CFB GENES IN AGE-RELATED MACULOPATHY AND JOINT ACTION WITH CFH AND LOC387715 GENES

This section has been published in PLoS ONE volume 3, issue 5, pages e2199 ([JAKOBSDOTTIR \*et al.\* 2008](#)). The copyright of the article belongs to the authors and the article is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. No changes have been made to the published version of the paper, except that tables and figures have been renumbered, the citations are of different style and some minor formatting has been made to keep this chapter coherent to the remainder of the thesis. My contribution to this paper was in writing all components of the paper, data analysis and script writing for method implementation.

### 5.1 ABSTRACT

**Background** Age-related maculopathy (ARM) is a common cause of visual impairment in the elderly populations of industrialized countries and significantly affects the quality of life of those suffering from the disease. Variants within two genes, the complement factor H (CFH) and the poorly characterized LOC387715 (ARMS2), are widely recognized as ARM risk factors. CFH is important in regulation of the alternative complement pathway suggesting this pathway is involved in ARM pathogenesis. Two other complement pathway genes, the closely linked complement component receptor (C2) and complement factor B (CFB), were recently shown to harbor variants associated with ARM. **Methods/Principal Findings** We investigated two SNPs in C2 and two in CFB in independent case-control and family cohorts of white subjects and found rs547154, an intronic SNP in C2, to be significantly associated with ARM in both our case-control (P-value



0.00007) and family data (P-value 0.00001). Logistic regression analysis suggested that accounting for the effect at this locus significantly (P-value 0.002) improves the fit of a genetic risk model of CFH and LOC387715 effects only. Modeling with the generalized multifactor dimensionality reduction method showed that adding C2 to the two-factor model of CFH and LOC387715 increases the sensitivity (from 63% to 73%). However, the balanced accuracy increases only from 71% to 72%, and the specificity decreases from 80% to 72%. **Conclusions/Significance** C2/CFB significantly influences AMD susceptibility and although accounting for effects at this locus does not dramatically increase the overall accuracy of the genetic risk model, the improvement over the CFH-LOC387715 model is statistically significant.

## 5.2 INTRODUCTION

Age-related maculopathy (ARM), also known as age-related macular degeneration (AMD), is a devastating disorder and a major public health issue. ARM poses one of the greatest threats to vision in the elderly of developed countries and an estimated 1.75 million individuals over 40 years old in the United States suffer vision loss from the disease with an estimated increase to 2.95 million individuals by 2020 (FRIEDMAN *et al.* 2004). ARM is a degenerative disorder primarily, but not exclusively, affecting the central macular region of the retina. It is characterized by formation of drusen, pigment epithelial changes, atrophic degenerative changes, and formation of choroidal neovascularization.

The etiology of ARM is complex and the disease susceptibility is influenced by both environmental and genetic components (SEDDON and CHEN 2004; VAN LEEUWEN *et al.* 2003). Of modifiable risk factors, the most recognized one is cigarette smoking (THORNTON *et al.* 2005). In the past couple of years, a light has been shed on our understanding of the genetic susceptibility of the disease (GORIN 2007). A genome-wide association scan (KLEIN *et al.* 2005) and two targeted searches (HAINES *et al.* 2005; EDWARDS *et al.* 2005) identified variants in the complement factor H (*CFH*, Entrez GeneID 3075) gene on chromosome 1q32 and two targeted searches (JAKOBSDOTIR *et al.* 2005; RIVERA *et al.* 2005) identified variants in the poorly characterized *LOC387715* (also known as *ARMS2*, GeneID 387715) gene, as well as in the closely linked *PLEKHA1* (GeneID 59338) and *HTRA1* (GeneID 5654) genes, on chromosome 10q26. Both findings have proven to be

robust and the associations of *CFH* and *LOC387715* variants and haplotypes, especially *Y402H* and *S69A*, respectively, have been replicated in multiple cohorts of various nationalities and ethnic backgrounds. This includes mostly samples of white European (BAIRD *et al.* 2006; DESPRIET *et al.* 2006; ENNIS *et al.* 2007; HUGHES *et al.* 2006; MAGNUSSON *et al.* 2006; SEITSONEN *et al.* 2006; SEPP *et al.* 2006; SIMONELLI *et al.* 2006; SOUIED *et al.* 2005; WEGSCHEIDER *et al.* 2007; WANG *et al.* 2007) and white European American (CONLEY *et al.* 2006; CONLEY *et al.* 2005; FRANCIS *et al.* 2007; HAGEMAN *et al.* 2005; LI *et al.* 2006; MALLER *et al.* 2006; ROSS *et al.* 2007; SCHAUMBERG *et al.* 2006; SCHMIDT *et al.* 2006; SEDDON *et al.* 2007; THOMPSON *et al.* 2007; TUO *et al.* 2006; ZAREPARSI *et al.* 2005; DEANGELIS *et al.* 2007; NARAYANAN *et al.* 2007; SCHAUMBERG *et al.* 2007) descent, but also samples of Hispanic origin (TEDESCHI-BLOK *et al.* 2007) and samples from Russia (FISHER *et al.* 2007), India (KAUR *et al.* 2006), China (LAU *et al.* 2006; CHEN *et al.* 2006), and Japan (OKAMOTO *et al.* 2006; TANIMOTO *et al.* 2007; MORI *et al.* 2007). Negative findings have, however, been reported for the role of *CFH* in Japanese ARM cohorts (FUSE *et al.* 2006; GOTOH *et al.* 2006; UKA *et al.* 2006). Two more recent studies (DEWAN *et al.* 2006; YANG *et al.* 2006) identified an additional variant (*rs1120638*) in the promoter region of *HTRA1*. This variant is in extremely strong linkage disequilibrium (LD) with the *S69A* variant in *LOC387715*, keeping the debate on the true susceptibility gene in the 10q26 region ongoing (CAMERON *et al.* 2007; MORI *et al.* 2007; YOSHIDA *et al.* 2007; KANDA *et al.* 2007). In the present study, we do not try to distinguish between the genes and variants in this region but use *S69A* as a tagging SNP; given the extensive LD in the region, especially between *S69A* and the *HTRA1* promoter variant, *S69A* can serve as a reasonable proxy for the genetic risk contributed by this region. In fact, a recent fine-mapping effort in this region does suggest that *S69A* is more likely, than the *HTRA1* promoter variant, to be causally responsible for the impact of this locus on ARM (KANDA *et al.* 2007).

*CFH* is now widely accepted as an important ARM susceptibility gene, harboring variants and haplotypes associated with increased and reduced disease risk. Functional studies suggest that *CFH* inhibits the activation of the alternative complement cascade and complements have been found in the drusen of ARM patients (HAGEMAN *et al.* 2001; HAGEMAN and MULLINS 1999; JOHNSON *et al.* 2001; JOHNSON *et al.* 2000; MULLINS *et al.* 2000). It is therefore logical to ask whether other genes involved in the alternative complement pathway may influence the risk. This task was partly tackled by Gold *et al.* (GOLD *et al.* 2006) who found ARM-associated variants in the complement component receptor B (*CFB*, GeneID 629) gene and the adjacent complement component 2 (*C2*, GeneID 717) gene on chromosome 6p21. Both genes play a role in complement

pathways: *CFB* in the alternative pathway and *C2* in the classical pathway. As was the case for *CFH* and *LOC387715*, this finding also seems robust and has been replicated in two case-control cohorts (MALLER *et al.* 2006; SPENCER *et al.* 2007) and one family cohort (SPENCER *et al.* 2007). However, because of the strong LD across the *C2/CFB* region, distinguishing between the genes and identifying true functional variants has proven challenging. Recently two studies (YATES *et al.* 2007; MALLER *et al.* 2007) reported significant associations between ARM and variants in the complement component 3 (*C3*, GeneID 718) gene on chromosome 19p13. *C3* plays an important role in activation of both the classical and the alternative complement pathways and the plasma complement C3a des Arg levels are significantly elevated in ARM cases compared to controls (SIVAPRASAD *et al.* 2007). A fourth recent study (DINU *et al.* 2007) also found ARM associated variants in the *C7* (GeneID 730) and *MBL2* (GeneID 4153) complement pathway genes by complement pathway focused analysis of an earlier genome-wide association scan (KLEIN *et al.* 2005).

In the present study, we investigated four SNPs in the *C2/CFB* region, *rs9332739* and *rs547154* in *C2* and *rs4151667* and *rs2072633* in *CFB*, in case-control and family cohorts of white subjects. Only *rs547154*, an intronic SNP in *C2*, was significantly associated with ARM in our data. Subsequently, *rs547154* was used as a tag for this region in multifactor analyses of the joint effect of the three genomic regions (*CFH*, *LOC387715*, and *C2/CFB*) on ARM susceptibility.

## 5.3 MATERIALS AND METHODS

### 5.3.1 Phenotyping, study participants and quality control

Because of the complexity and ambiguity in the ARM phenotype, we have previously defined three affection status models (types A, B, and C) (WEEKS *et al.* 2000; WEEKS *et al.* 2004). For clarity we restrict our analyses here to unaffected (or normal) individuals and type A affected individuals. The type A model is our most stringent and conservative diagnostic model and individuals classified as type A ARM affected are clearly affected with ARM based on extensive and/or coalescent drusen, pigmentary changes (including pigment epithelial detachments) and/or the presence of end-stage disease (geographic atrophy [GA] and/or choroidal neovascular [CNV] membranes). Unaffected individuals were those for whom eye-care records and/or fundus photographs indicated either no

evidence of any macular changes (including drusen) or a small number ( $< 10$ ) of hard drusen (50 m in diameter) without any other retinal pigment epithelial (RPE) changes. Individuals with evidence of large numbers of extramacular drusen were not classified as unaffected and therefore not included in the analyses. No family member was considered unaffected but was considered of unknown phenotype if not affected with type A ARM.

Using only the subset of white participants, our data include 611 ARM families, 187 unrelated cases and 168 unrelated controls. The ARM families consist of 1,524 genotyped individuals (569 males and 955 females) and, in terms of genotyped affected relative pairs, the families include total of 501 sib pairs, 7 half sib pairs, 60 cousin pairs, 13 parent-child pairs, and 38 avuncular pairs; Pedstats (version 0.6.8) (WIGGINTON and ABECASIS 2005) was used to get summary counts of the family data. See Table 5.1 for other characteristics of the subjects. Before analyzing the family data, PedCheck (version 1.1) (O'CONNELL and WEEKS 1998) was used to check for Mendelian inconsistencies. Since it can be extremely difficult to determine who exactly has the erroneous genotype within small families (MUKHOPADHYAY *et al.* 2004), we set genotypes of problematic markers to missing for every individual within each family containing a Mendelian inconsistency. Informed consent was obtained from all participants under research protocols that have been reviewed and approved in accordance with the Declaration of Helsinki and the Guidelines for Human Subjects Protection issued by the Office of Human Subjects Research (National Institutes of Health) by the University of Pittsburgh IRB (#9506133) and the University of CaliforniaLos Angeles IRB (#10-06-096-01).

### 5.3.2 Genotyping

The variants: *rs9332739* (*E318D*) and *rs547154* (*IVS10*) in *C2*, and *rs4151667* (*L9H*) and *rs2072633* (*IVS17*) in *CFB*, were genotyped using 5' exonuclease Assay-on-Demand TaqMan assays (Applied Biosystems Incorporated). Amplification and genotype assignments were conducted using the ABI7000 and SDS 2.0 software (Applied Biosystems Incorporated, Foster City, CA). The variant *rs1061170* (*Y402H*) in *CFH* and the variant *rs10490924* (*S69A*) in *LOC387715* were genotyped using RFLP techniques. The primers, annealing temperatures and restriction endonuclease for each assay were: 5'-TCTTTTTGTGCAAACCTTTGTTAG-3' (F), 5'-CCATTGGTAAAACAA GGTGACA-3' (R), 52 °C, *Nla*III for *Y402H* in *CFH*; 5'-GCACCTTTGTCCACCACATTA-3' (F), 5'-GCCTGATCATCTGCATTTCT-3' (R), 54 °C, *Pvu*II for *S69A* in *LOC387715*. For all genotyp-

Table 5.1: Samples sizes and other characteristics of the data.

	Family data		Case-control data	
	Type A	not Type A	Cases (Type A)	Controls
Number of genotyped individuals				
Females	690	265	113	87
Males	405	164	74	81
Total	1095	429	187	168
Mean age (SD)				
Females	77.7 (7.3)	73.4 (12.9)	78.6 (7.0)	71.3 (10.2)
Males	77.0 (7.1)	73.3 (11.5)	79.8 (6.0)	74.6 (9.4)
Total	77.4 (7.2)	73.4 (12.4)	79.1 (6.6)	72.9 (9.9)
Cigarette smokers (%)				
Females	37	35	43	34
Males	61	50	55	42
Total	46	41	48	38
GA (%)				
Females	56	...	55	...
Males	52	...	58	...
Total	54	...	56	...
CNV (%)				
Females	70	...	64	...
Males	71	...	69	...
Total	70	... 66	...	...

GA=geographic atrophy

CNV=choroidal neovascular membranes

SD=standard deviation

ing conducted for this research, double-masked genotyping assignments were made for each variant and compared; each discrepancy was addressed using raw data or by re-genotyping. Genotype efficiency for the *C2/CFB* SNPs ranged from 93% to 96% and 88% to 90% for the two previously published *CFH* and *LOC387715* SNPs.

### 5.3.3 Association analyses and LD estimation

**5.3.3.1 Case-Control data** Using the set of unrelated cases and controls, SNP-disease allelic and genotypic associations were tested using the Fisher’s exact test as implemented in R (version 2.2.1) (R DEVELOPMENT CORE TEAM 2005). For significantly associated SNPs the strength of the association was estimated by crude odds ratios (ORs) and population attributable risks (PARs). To calculate the PARs we used the general formula:  $PAR = P_f(OR - 1)/(1 + P_f(OR - 1))$ , where  $P_f$  is the prevalence of the risk or protective factor (genotype) in the general population as estimated from the controls. The ORs were calculated using logistic regression models in R. Confidence intervals (CIs) for the ORs and PARs were derived using the asymptotic normal distribution of  $\ln(OR)$  and  $\ln(1-PAR)$ , respectively. Haplotypic associations of 2- and 3-SNP moving window haplotypes in the *C2/CFB* locus were evaluated using the *haplo.cc* function of the *haplo.stats* package (version 1.2.2) (SCHAID *et al.* 2002) of R. This function implements a score test for global test of association between binary traits and haplotypes and accounts for ambiguous linkage phase by the EM algorithm; empirical P-values were generated using 10,000 replicates. Allele and genotype frequencies were estimated by direct counting and deviations from Hardy-Weinberg equilibrium (HWE) were tested, in cases and controls separately, using the exact test as implemented in R Genetics package (version 1.2.1) (WARNES and FRIEDRICH 2006). Haploview (version 3.32) (BARRETT *et al.* 2005) was used to estimate the LD across the *C2/CFB* region, both  $D'$  and  $r^2$  were calculated separately in cases and controls.

**5.3.3.2 Family data** When incorporating cases from the families into the analyses, the CCREL method (version 0.3) (BROWNING *et al.* 2005) was used to test SNP-disease allelic, genotypic and 2- and 3-SNP haplotypic associations. The CCREL method permits testing for association with the use of related cases and unrelated controls simultaneously and, briefly, it accounts for biologically related subjects by calculating an effective number of cases such that individuals are assigned weights that are used to construct a composite likelihood, which is then maximized iteratively to

form likelihood ratio tests. For the CCREL analyses, type A-affected family members were assigned the phenotype affected, unrelated controls the phenotype normal and family members not affected with type A ARM the phenotype unknown.

### 5.3.4 Multifactor and gene-gene interaction analyses

To build predictive models of the genetic risk of ARM contributed by the *CFH*, *LOC387715*, and *C2/CFB* loci, we applied both logistic regression and the new generalized multifactor dimensionality reduction (GMDR) method (version 0.7) (LOU *et al.* 2007). The GMDR method, unlike the original MDR method (RITCHIE *et al.* 2001), permits adjustment for covariates and better handles data with unequal numbers of cases and controls, and can be used to analyze both qualitative (e.g. binary) and quantitative traits via different link functions. Both methods only handle unrelated individuals. Therefore, to make use of more of our data, we combined one type A affected person picked at random from each of the 611 ARM families with the data of unrelated cases and controls. We consider this to be appropriate to do since the association results suggest the effects of the genes to be similar in both groups.

**5.3.4.1 Logistic regression** For each pair of loci, we first followed the modeling strategy proposed by North *et al.* (NORTH *et al.* 2005) for two-factor genetic risk models. A series of logistic regression models were fitted to the data in order to find a parsimonious model for the joint effects of each pair of loci. Models allowing for additive effects (ADD1, ADD2, and ADD-BOTH), models incorporating dominance effects (DOM1, DOM2, and DOM-BOTH), and three interaction models (ADD-INT, ADD-DOM, and DOM-INT) were fitted. We fit three-factor models of the joint effect of all three loci and test, using a likelihood ratio test (LRT), whether accounting for the protective effects at *C2/CFB* significantly improves the fit of a model with *CFH* and *LOC387715* effects only. Since, for each pair of loci, the two-factor analyses implicated additive models as the most parsimonious and to keep the number of parameters as small as possible we only fit three-factor additive models without interaction (ADD1, ADD2, ADD3, ADD12, ADD13, ADD23, and ADD123). The models are compared by the Akaike information criterion (AIC); the most parsimonious model has the lowest AIC and a model is considered to provide a significantly better fit to the data if it has AIC more than 2 units lower than the comparison model (NORTH *et al.* 2005). Details regarding coding of genotypes in the models are available in appendix C.

**5.3.4.2 GMDR** Just as in the case of logistic regression, when using the GMDR method, one needs to be aware of the risk of overfitting, especially in the case of small sample sizes. The GMDR method, however, uses cross-validation to guard against overfitting. We applied the method to our data in order to identify three-locus genotypes associated with increased and decreased disease risk. For comparison we also present and discuss the *CFH* and *LOC387715* two-factor model. We performed both crude analysis and analyzed the data while adjusting for age, gender, and cigarette smoking. We used 5-fold leave-one-out cross-validation and exhaustive search of all possible one- to three-locus models in the GMDR analyses. In the adjusted analysis age (in years) was the age at the time blood was drawn (i.e. DNA donated), and cigarette smoking was a binary variable (ever vs. never smoked). The smokers smoked on averaged 40.45 (standard deviation [SD] 32.96; range 0.23207.00) pack-years (years?packs/day smoked) of cigarettes. The sample in the adjusted analysis includes fewer observations (557 cases and 118 controls fully typed at all three SNPs) than the sample in the unadjusted analysis (640 cases and 142 controls fully typed at all three SNPs) because of missing information. We compared both the sensitivity= $TP/(TP+FN)$  and the specificity= $TN/(TN+FP)$  of the models, where TP=number of true positives, TN=number of true negatives, FP=number of false positives, and FN=number of false negatives. As a single measure of the accuracy of the models we used the balanced accuracy= $(\text{sensitivity}+\text{specificity})/2$  rather than the accuracy= $(TP+TN)/(TP+TN+FP+FN)$  because number of cases and controls is unequal. The average sensitivity, specificity, and balanced accuracy over the testing sets of all five cross-validations are reported. As a measure for the appropriateness of the models, the sensitivity, specificity, balanced accuracy, and P-value are reported for all models when applied to the whole dataset.

### 5.3.5 Interaction with cigarette smoking

In a logistic regression framework we tested, using a LRT and the combined data of unrelateds and one type A affected from each family, whether cigarette smoking interacts with the SNPs at the three genes. The genotypes were coded in additive way, as in the logistic regression analysis above, and cigarette smoking as ever vs. never smoked.



## 5.4 RESULTS

### 5.4.1 Results of association analyses

The genotype distributions of the 4 SNPs typed in *C2* and *CFB* and the *Y402H* variant in *CFH* and the *S69A* variant in *LOC387715* are in HWE in both our cases and controls (Table 5.2). Of the 4 SNPs typed in the *C2/CFB* region, only *rs547154*, an intronic SNP in *C2*, is significantly associated with ARM (Table 5.2) in both our case-control (P-value of genotypic test 0.00007) and family data (P-value of genotypic test 0.00001), which is also significant after adjusting for multiple testing of 4 tests (Bonferroni corrected 0.05 significance level is 0.0125). The haplotypic association tests show that haplotypes spanning the entire *C2/CFB* locus are significantly associated with ARM (Table 5.2). Although LD between *rs547154* and the SNPs in *CFB* (Figure 5.1) is not strong, in neither cases nor controls, these results are not sufficient to rule out either *C2* or *CFB* as an ARM candidate gene, because of limited number of SNPs investigated. Individuals carrying the protective allele at *C2* are at 0.22 (95% CI 0.10 to 0.48) times less risk of having ARM compared to controls as estimated with a crude OR. The corresponding PAR is 18% (95% CI 28% to 8%). Detailed results of marginal association of *Y402H* in *CFH* and *S69A* in *LOC387715* are in Table 5.2 and the supporting information (Text S2).

### 5.4.2 Results of multifactor analyses

**5.4.2.1 Logistic regression** First we fitted two-factor genetic risk models for each pair of loci and found that an additive model without interaction was the most parsimonious in all cases (Table 5.3). Three-factor additive model was then fitted in order to test whether the three-factor model provided better fit to the data than any two-factor models (Table 5.4). The three-factor model of *CFH*, *LOC387715*, and *C2* SNPs coded in additive fashion was the most parsimonious and fitted significantly better (P-value of LRT 0.002) than the next-best model (which modeled *CFH* and *LOC387715* additive effects only).

Table 5.2: Association results for *C2/CFB* variants, *Y402H* in *CFH*, and *S69A* in *LOC387715*.

SNP (Location)	Gene	MA	P-value for test															
			MAF in				HWE in				Single SNP in				Moving window haplotypic test			
			Cases	Controls	Cases	Controls	CCREL		Exact test in unrelateds		CCREL		Global test in unrelateds					
							Allelic	Genotypic	Allelic	Genotypic	With 2 SNPs	With 3 SNPs	With 2 SNPs	With 3 SNPs				
<i>rs9332739 (E318D)</i>	<i>C2</i>	C	0.027	0.033	1.000	0.157	0.26542	0.37187	0.66583	0.90278	0.00088	0.00076	0.00020	0.00000				
<i>rs547154 (IVS10)</i>	<i>C2</i>	T	0.025	0.096	1.000	0.365	0.00010	0.00001	0.00011	0.00007	0.00071	0.00131	0.00020	0.00030				
<i>rs4151667 (L9H)</i>	<i>CFB</i>	A	0.028	0.036	1.000	0.185	0.19863	0.27610	0.66609	0.81796	0.38067	...	0.27480	...				
<i>rs2072633 (IVS17)</i>	<i>CFB</i>	A	0.330	0.393	0.499	0.104	0.64767	0.07003	0.09299	0.05780	...	...	...	...				
<i>rs1061170 (Y402H)</i>	<i>CFH</i>	C	0.621	0.348	0.615	0.288	<0.00001	<0.00001	$6.3 \times 10^{-12}$	$7.7 \times 10^{-11}$	...	...	...	...				
<i>rs10490924 (S69A)</i>	<i>LOC387715</i>	T	0.470	0.200	0.272	0.075	<0.00001	<0.00001	$4.2 \times 10^{-13}$	$8.2 \times 10^{-11}$	...	...	...	...				

MA = minor allele

MAF = minor allele frequency

HWE = Hardy-Weinberg equilibrium

The haplotypic P-values correspond to the haplotypes of the SNP in the same row as the P-value and the next one or two SNPs for the 'With 2 SNPs' and 'With 3 SNPs'

P-values, respectively

Genotype counts in unrelated cases and controls are available in Table S1

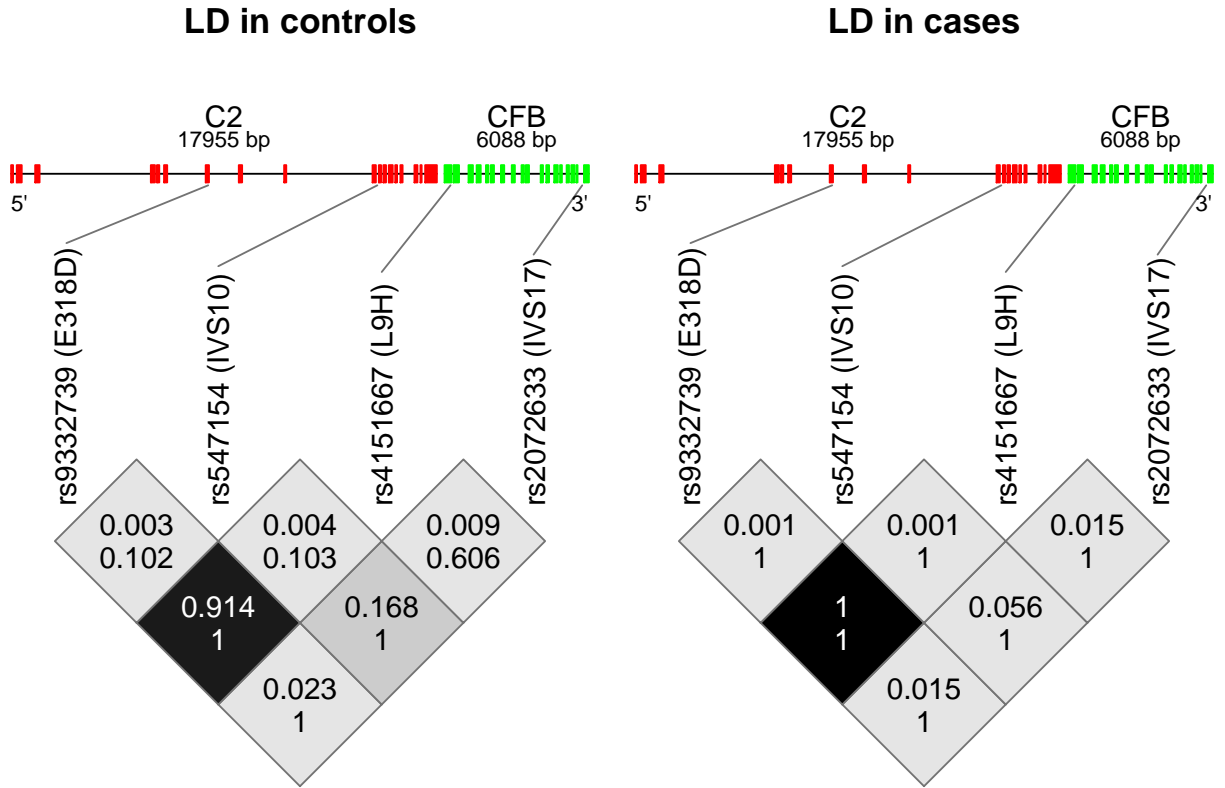


Figure 5.1: Linkage disequilibrium (LD) across the *C2/CFB* region in unrelated cases and controls. The darker the boxes the higher the  $r^2$ . The top number in each box is  $r^2$  and the bottom number is  $D'$ . Locations of the SNPs within the genes are shown. Red lines/boxes show the locations of exons in *C2* and green lines/boxes the locations of exons in *CFB*.

**5.4.2.2 GMDR** The two-factor GMDR unadjusted and adjusted models (Figure 5.2A and B) classify everyone with a homozygous (*TT*) *LOC387715* risk genotype as cases and everyone with the homozygous (*GG*) *LOC387715* non-risk genotype as controls. On the other hand, individuals heterozygous (*GT*) at *LOC387715* need to have at least one *CFH* risk allele (*C*) to be classified as cases. When comparing the unadjusted two-factor model (Figure 5.2A) to the unadjusted three-factor model (Figure 5.2C), the most dramatic change is in the upper left most cell (*CC-GG* *CFH-LOC387715* joint genotype): 76 cases in that cell that were wrongly classified as controls by the two-factor model while 71 are correctly (5 wrongly) classified in the three-factor model. This

Table 5.3: Results of fitting two-factor logistic regression models.

Two-factor model	AIC	AIC difference
<i>CFH</i> (Factor 1) and <i>LOC387715</i> (Factor 2)		
ADD1	702.6	68.2
ADD2	699.0	64.5
ADD-BOTH	634.5	0.0
DOM1	704.0	69.5
DOM2	698.6	64.2
DOM-BOTH	634.9	0.5
ADD-INT	636.1	1.6
ADD-DOM	634.5	0.0
DOM-INT	636.4	1.9
<i>CFH</i> (Factor 1) and <i>C2</i> (Factor 2)		
ADD1	716.3	8.7
ADD2	764.9	57.3
ADD-BOTH	707.6	0.0
DOM1	717.6	10.0
DOM2	764.9	57.3
DOM-BOTH	709.0	1.5
ADD-INT	707.7	0.1
ADD-DOM	709.9	2.4
DOM-INT	709.9	2.4
<i>LOC387715</i> (Factor 1) and <i>C2</i> (Factor 2)		
ADD1	729.1	13.2
ADD2	783.7	67.8
ADD-BOTH	715.9	0.0
DOM1	729.2	13.3
DOM2	783.7	67.8
DOM-BOTH	716.0	0.1
ADD-INT	717.9	2.0
ADD-DOM	718.3	2.4
DOM-INT	718.3	2.4

Detailed model definitions are given in the Materials and Methods - Multifactor and interaction analyses section.

AIC difference is the difference from the AIC of the best fitting model. Most parsimonious model is in bold. Model with best fit (lowest AIC) has AIC difference = 0.

Table 5.4: Results of fitting three-factor logistic regression models.

Model	AIC	AIC difference
ADD1	685.5	71.2
ADD2	682.8	68.6
ADD3	728.3	114.1
ADD12	622.1	7.9
ADD13	677.8	63.6
ADD23	669.2	55.0
ADD123	614.2	0.0

Factor 1 is *Y402H* in *CFH*, Factor 2 is *S69A* in *LOC387715*, and Factor 3 is *rs547154* in *C2*. Detailed model definitions are given in the Materials and Methods - Multifactor and interaction analyses section. AIC difference is the difference from the AIC of the best fitting model. Most parsimonious model is in bold. Model with best fit (lowest AIC) has AIC difference = 0.

increases the sensitivity from 63% to 73%, but comes at a cost of decreased specificity (80% to 72%).

Now looking more closely at the three-factor models (Figures 2C and D), the results of the GMDR analyses suggest that having at least one copy of the protective allele (*T*) at *C2/CFB* may reduce the risk contributed by *CFH* and *LOC387715* risk genotypes. For example in the unadjusted model (Figure 5.2C), individuals with the *CT-GT* and *CT-GG* two-locus genotypes at *CFH* and *LOC387715* and without the *C2/CFB* protective allele are classified as cases while those with the protective allele are classified as controls. In the adjusted model (Figure 5.2D), however, individuals with the *CT-GT* and *CC-GG* as well as *TT-TT* and *CC-GT* two-locus genotypes at *CFH* and *LOC387715*, are classified as controls if they carry the *C2/CFB* protective allele but cases otherwise. Note that the difference between the three-factor unadjusted and adjusted models is not due to the smaller dataset used in the adjusted analysis. To make sure this was not the case, we ran unadjusted analysis on the smaller dataset and arrived at the same model as in the unadjusted analysis. The predictive models presented in Figure 5.2 seem sensible as the predicted high-risk two- and three-locus genotypes group together. The predictive accuracy of the three-factor model measured by sensitivity, specificity, and balanced accuracy is > 70% of both the unadjusted and adjusted models (Table 5.5). In the unadjusted analysis all five cross-validations suggest that the classification scheme classifies individuals significantly better than random (P-values < 0.05)

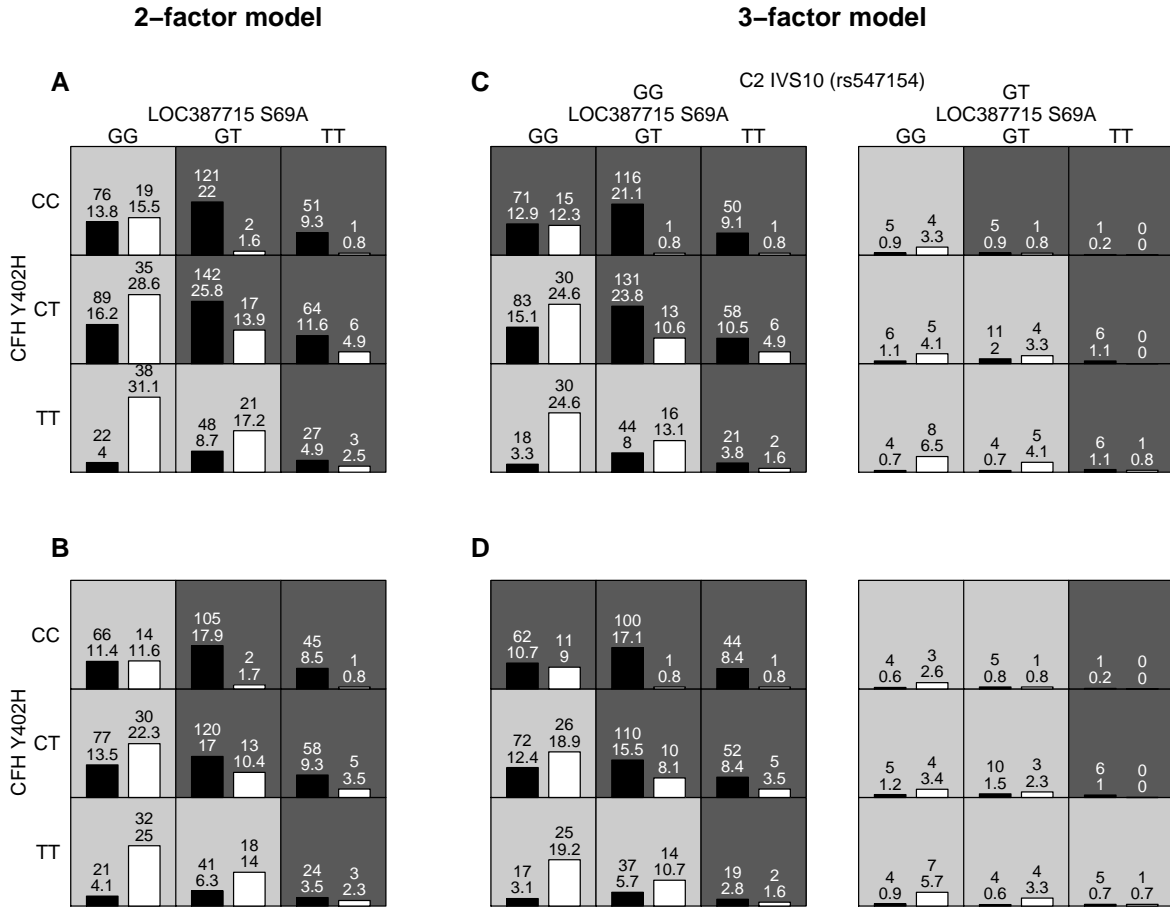


Figure 5.2: **A:** Results of unadjusted GMDR analysis for the best two-factor model. **B:** Results of adjusted GMDR analysis for the best two-factor model. **C:** Results of unadjusted GMDR analysis for the three factor model. **D:** Results of adjusted GMDR analysis for the three-factor model. Dark grey and light grey boxes correspond to the high- and low-risk genotype combinations, respectively. The black and white bars within each box correspond to cases and controls, respectively. The top number above each bar is number of individuals and the bottom number is the sum of scores for the corresponding group of individuals (cases or controls with particular three-locus genotype). The heights of the bars are proportional to the sum of scores in each group.

and in the adjusted analysis the classification is significantly better in all but one cross-validation experiment (Table 5.5). Both models provide excellent fit to the whole data (P-values < 0.0001).

In Table C2, we present the joint genotype and relative genotype frequencies in cases and controls, which provides a complementary view of the same findings as in Figure 5.2.

**5.4.2.3 Logistic regression vs. GMDR** In both the logistic regression and GMDR analyses, the best fitting one-factor model is the model with *LOC387715* only (Tables 5.4-5.5); in the logistic regression, the *LOC387715* model has the lowest AIC of all one-factor models and, in the GMDR results the *LOC387715* model has the highest balanced accuracy (in both the unadjusted and adjusted analyses). However, the difference between the *CFH* and *LOC387715* one-factor models is, very small, and, as the GMDR analyses show, the difference lies in the sensitivity and specificity rather than the overall balanced accuracy measure (Table 5.5). The three-factor model of *CFH*, *LOC387715*, and *C2/CFB* effects is implicated as the best model in both the regression and GMDR analyses (Tables 5.4-5.5). The logistic regression analyses suggest that accounting for *C2/CFB* effects significantly improves the two-factor model of *CFH* and *LOC387715* only (P-value 0.002). The GMDR analyses show that this improvement is due the increases sensitivity but the balanced accuracy increases only from 71% to 72% (Table 5.5). The GMDR analyses also suggest that adjusting for age, gender, and cigarette smoking does not dramatically improve the fit of the models. In fact, all models (1-, 2-, and 3-factor) have approximately the same balanced accuracy irrespective of whether adjustment is made (Table 5.5).

### 5.4.3 Results of gene-cigarette smoking interaction analysis

Cigarette smoking does not significantly interact with any of the three variants investigated in our data. The p-values of LRTs are 0.24, 0.99, and 0.43 for *Y402H* in *CFH*, *S69A* in *LOC387715*, and *IVS10* in *C2*, respectively. For all three genes the most parsimonious models, according to AIC, are the models with only the additive gene effect and no smoking effect (results not shown).

## 5.5 DISCUSSION

We have replicated the association of one *C2* variant (*rs547154*) with ARM in both our case-control and family datasets and we have shown that accounting for the effects of *C2/CFB* significantly improves the fit of the logistic regression model in comparison to the two-factor model of joint

Table 5.5: Results of GMDR analyses.

Model	Unadjusted				Adjusted			
	P-value	Sensitivity	Specificity	Balanced Accuracy	P-value	Sensitivity	Specificity	Balanced Accuracy
<i>CFH, LOC387715, and C2</i>								
Testing 1	0.0079	0.76	0.62	0.69	0.0029	0.70	0.79	0.74
Testing 2	0.0140	0.55	0.79	0.67	0.0047	0.70	0.77	0.73
Testing 3	0.0027	0.73	0.71	0.72	0.0172	0.75	0.64	0.70
Testing 4	0.0001	0.65	0.93	0.79	0.0237	0.49	0.86	0.68
Testing 5	0.0298	0.71	0.61	0.66	0.0823	0.75	0.54	0.64
Average	...	0.68	0.73	0.71	...	0.68	0.72	0.70
Whole data	<0.0001	0.73	0.72	0.72	<0.0001	0.70	0.74	0.72
<i>CFH and LOC387715</i>								
Testing 1	0.0079	0.63	0.76	0.69	0.0026	0.66	0.83	0.74
Testing 2	0.0140	0.55	0.79	0.67	0.0038	0.76	0.72	0.74
Testing 3	0.0087	0.77	0.61	0.69	0.0320	0.62	0.73	0.68
Testing 4	0.0003	0.67	0.86	0.76	0.0165	0.52	0.86	0.69
Testing 5	0.0117	0.66	0.71	0.69	0.2298	0.75	0.45	0.60
Average	...	0.66	0.75	0.70	...	0.66	0.72	0.69
Whole data	<0.0001	0.63	0.80	0.71	<0.0001	0.61	0.80	0.71
<i>CFH</i>								
Testing 1	0.0317	0.89	0.38	0.64	0.0276	0.88	0.45	0.66
Testing 2	0.0341	0.78	0.52	0.65	0.0764	0.41	0.85	0.63
Testing 3	0.1653	0.85	0.32	0.59	0.1413	0.83	0.39	0.61
Testing 4	0.0011	0.83	0.64	0.74	0.1484	0.81	0.41	0.61
Testing 5	0.0794	0.89	0.32	0.61	0.0400	0.89	0.41	0.65
Average	...	0.85	0.44	0.64	...	0.77	0.50	0.63
Whole data	<0.0001	0.85	0.44	0.64	<0.0001	0.85	0.45	0.65
<i>LOC387715</i>								
Testing 1	0.0132	0.67	0.69	0.68	0.0564	0.71	0.60	0.66
Testing 2	0.0447	0.67	0.62	0.65	0.0074	0.71	0.73	0.72
Testing 3	0.0012	0.73	0.75	0.74	0.1565	0.72	0.51	0.61
Testing 4	0.0305	0.74	0.57	0.66	0.0140	0.60	0.81	0.70
Testing 5	0.0224	0.73	0.61	0.67	0.1102	0.68	0.59	0.63
Average	...	0.71	0.65	0.68	...	0.68	0.65	0.66
Whole data	<0.0001	0.71	0.65	0.68	<0.0001	0.68	0.64	0.66

Each testing set corresponds to 1/5 of the data. The same individuals are in each testing set across models and within type of analysis (unadjusted or adjusted). The individuals are not necessarily the same in the testing sets across type of analysis because of the smaller number of individuals that were available in the adjusted analyses compared to the unadjusted analysis (see the text for details). The average is the average over the five testing sets and the P-value corresponds to  $\chi^2$  tests of fitting the models to the testing sets or the whole data.



additive effects of *CFH* and *LOC387715* (Table 5.2 and 5.4). Interestingly, both of the non-synonymous coding changes, *E318D* (*rs9332739*) in *C2* and *L9H* (*rs415667*) in *CFB*, identified by Gold et al. (GOLD *et al.* 2006) are insignificant in both of our datasets. However, as these variants (*rs9332739* and *rs415667*) are quite rare, power to detect these variants is low. Even so, our independent confirmation of the statistically significant effect of this locus in ARM in two datasets, including family-based data, further supports the contribution of this locus to the genetic susceptibility of ARM.

As mentioned above, accounting for the effect of the *C2/CFB* locus significantly improved the fit of a logistic regression model of additive effects of *CFH* and *LOC387715* variants. To further understand this, we built predictive models of these three loci using the new generalized multifactor dimensionality reduction method (GMDR), and found that addition of *C2/CFB* to the model increased sensitivity (from 63% to 73%). However, the specificity is lowered (from 80% to 72%) and so the balanced accuracy only increases from 71% for the two-factor *CFH-LOC387715* model to 72% for the three-factor *CFH-LOC387715-C2* model in the unadjusted analysis (Table 5.5). If it were considered more important to identify cases than controls correctly, while maintaining a reasonable specificity, the three-factor model would be the better choice.

Since our associated variant (*rs547154*) in *C2/CFB* is rare, it is expected that accounting for the effect of this locus, using *rs547154* as a tag, would not markedly improve the overall prediction accuracy of the genetic risk model with *CFH* and *LOC387715* effects only, even though the effect may be strong. Although, positive associations in the *C2/CFB* region have been found and replicated primarily for rare variants (GOLD *et al.* 2006; MALLER *et al.* 2006; SPENCER *et al.* 2007) we cannot exclude the possibility that the true causal variant(s) in this region may be common, especially since not all known common SNPs have been typed in *C2/CFB* studies (Figure S1). Obviously, a genetic risk model of *CFH*, *LOC387715*, and *C2/CFB* effects could be quite different from our model presented here if the *C2/CFB* causal variant(s) were common, as then the rare *rs547154* would be a bad proxy.

Another concern regarding correctness of the three-factor model is the small sample size for the 'protective' *GT* genotype at *IVS10* (*rs547154*) in *C2* (Figure 5.2 and Table ??), although it is important to remember that cross-validation does guard against over-fitting due to small sample sizes or a large number of parameters. The least stable classifications in Figure 5.2C are those cells in which the height of the bars is similar or number of individuals is low. In such cases, the

classification rule can change if only a few individuals were added to that cell. For example, if we had only one additional control with a *CC-GG-GG CFH-LOC387715-C2* genotype (upper left most cell, left panel in Figure 5.2C), then individuals with this genotype combination would have been classified as controls instead of cases. To construct our original unrelated data set, we picked one case at random from each of the families. Figure S2 examines the sensitivity of our three-factor analyses when we randomly re-pick one case from each family. We created 10 other combined data-sets (overlap among cases from the families ranges from 57% to 66%) and ran the GMDR method. The figure clearly shows that only classifications corresponding to the rare *GT* genotype at *IVS10 (rs547154)* in *C2* are changed across samples, while the classifications corresponding to the common *GG* genotype are robust.

Accounting for covariates (age, gender, and cigarette smoking) failed to improve the prediction accuracy of the genetic risk models (Table 5.5). In fact, for the one- and two-factor models, the adjusted analyses arrived at the same high-risk (and low-risk) genotype combinations as the unadjusted analyses. The difference in sensitivity, specificity, and balanced accuracy between the two analyses is solely due to different number of individuals used in each set of analyses (because of incomplete smoking information). In the three-factor model, genotypes were grouped differently depending on whether unadjusted or adjusted analyses were performed (Figure 5.2) and, as mentioned in the results section, this difference is not solely due to different number of individuals used in each set of analyses.

The one-factor models of *CFH* and *LOC387715* did worse than the higher-factor models (balanced accuracy 64% and 68%, respectively), although, when considering they only model genetic effects at one locus, both models perform amazingly well. The GMDR method selected the *LOC387715* model as the best of all the one-factor models. However, depending on what the goals of using a prediction model are, one could easily choose the *CFH* model as the best one-factor model. For example, the sensitivity of the *CFH* model is much higher than of the *LOC387715* model (85% vs. 71%), but this increased sensitivity comes at a cost of low specificity (44% vs. 65%).

In their original report on the *C2/CFB* locus in ARM, Gold et al. (GOLD *et al.* 2006) did not include *LOC387715* variants and, using a genetic algorithm search approach, they arrived at a genetic risk model of two *CFH* variants and three *C2/CFB* variants. The sensitivity and specificity of their model were 74% and 56%, respectively (which results in balanced accuracy of

65%). Interestingly, our three-factor model, which includes *LOC387715* effects, provides a better prediction accuracy (balanced accuracy 72%), similar specificity (73%), and better specificity (72%) than their more complicated five-factor model. Furthermore, even our simpler two-factor model of *CFH* and *LOC387715* effects also provides better prediction accuracy (balanced accuracy 71%).

We believe that a word of caution must be provided with regard to the possible use of these predictive models in clinical situations. It must be understood that the models presented in this paper and by others are based on comparison of extreme phenotypes (those with advanced forms of ARM and age-matched controls with minimal or no clinical findings). This does not address the determination of ARM risk for individuals for whom mild to moderate retinal findings are present. Secondly, odds ratios based on case-control association studies are not comparable to prospective, population-based relative risk assessments that still need to be done for ARM. Finally, one must always consider the composition of the population that may be subjected to molecular genetic screening. If we are considering the general population for whom the risk of ARM-related vision loss is less than 1% over their lifetime, then the current genetic models have inadequate levels of specificity to avoid a high percentage of false positive results. However, for individuals from high-risk cohorts for whom the prevalences of the high-risk variants are known, molecular diagnostic testing may be sufficiently discriminating of relative risk, though it is unclear how such knowledge would affect individual behavior or preventive treatments at this time.

In summary, we have confirmed the likely influence of the *C2/CFB* locus on ARM and shown that accounting for the effects at this locus can likely further stratify individuals as being at high or low risk of developing ARM. The important role the classical and/or alternative complement pathways seem to have in the disease-pathology of ARM should now encourage investigators to not only look at more complement pathway genes, but also to establish the biological mechanism behind the influence of *LOC387715* (or *HTRA1*) on the development of the disorder. Then, once either *LOC387715* or *HTRA1* has been convincingly shown to be the true ARM susceptibility gene on 10q26, it is likely that we will see similar trends in discoveries of genes involved in the same pathway as either of those genes.

## 6.0 INTERPRETATION OF GENETIC ASSOCIATION STUDIES: MARKERS WITH REPLICATED HIGHLY SIGNIFICANT ODDS RATIOS MAY BE POOR CLASSIFIERS

This section has been published in PLoS Genetics volume 2, issue 2, pages e1000337 ([JAKOBSDOTTIR \*et al.\* 2009](#)). The copyright of the article belongs to the authors and the article is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. No changes have been made to the published version of the paper, except that tables and figures have been renumbered, the citations are of different style and some minor formatting has been made to keep this chapter coherent to the remainder of the thesis. My contribution to this paper was in writing all components of the paper, data analysis, and script writing for method implementation.

### 6.1 ABSTRACT

Recent successful discoveries of potentially causal single nucleotide polymorphisms (SNPs) for complex diseases hold great promise, and commercialization of genomics in personalized medicine has already begun. The hope is that genetic testing will benefit patients and their families, and encourage positive lifestyle changes and guide clinical decisions. However, for many complex diseases, it is arguable whether the era of genomics in personalized medicine is here yet. We focus on the clinical validity of genetic testing with an emphasis on two popular statistical methods for evaluating markers. The two methods, logistic regression and receiver operating characteristic (ROC) curve analysis, are applied to our age-related macular degeneration dataset. By using an additive model of the CFH, LOC387715, and C2 variants, the odds ratios are 2.9, 3.4, and 0.4, with  $P$

values of  $10^{-13}$ ,  $10^{-13}$ , and  $10^{-3}$ , respectively. The area under the ROC curve (AUC) is 0.79, but assuming prevalences of 15%, 5.5%, and 1.5% (which are realistic for age groups 80 y, 65 y, and 40 y and older, respectively), only 30%, 12%, and 3% of the group classified as high risk are cases. Additionally, we present examples for four other diseases for which strongly associated variants have been discovered. In type 2 diabetes, our classification model of 12 SNPs has an AUC of only 0.64, and two SNPs achieve an AUC of only 0.56 for prostate cancer. Nine SNPs were not sufficient to improve the discrimination power over that of nongenetic predictors for risk of cardiovascular events. Finally, in Crohn’s disease, a model of five SNPs, one with a quite low odds ratio of 0.26, has an AUC of only 0.66. Our analyses and examples show that strong association, although very valuable for establishing etiological hypotheses, does not guarantee effective discrimination between cases and controls. The scientific community should be cautious to avoid overstating the value of association findings in terms of personalized medicine before their time.

## 6.2 INTRODUCTION

Recent successes in the discoveries of potentially causal single nucleotide polymorphisms (SNPs) for complex diseases hold great promise, and commercialization of genomics in personalized medicine has already begun. A number of companies now offer, for relatively modest fees, personalized genomics services that provide individualized disease-risk estimates based on genome-wide SNP genotyping. Most companies offering such profiling make it clear that they are not a clinical service and that their calculations are not intended for diagnostic or prognostic purposes. They typically advise their clients to consult their health care provider for more information. In most cases, people would turn to their general physician (MITKA 1998). However, as noted by others (FEERO 2008; GOETZ 2007), few doctors currently have enough genetics training to actually make sense of the risk calculations now commercially offered. Many physicians seem to feel the same way. In surveys in five European countries, physicians ranked the disciplines in which they felt they needed more training to overcome future challenges (CALEFATO *et al.* 2008; JULIAN-REYNIER *et al.* 2008). In all countries, the top ranked area was “genetics of common disease”, and ranked second was “approaching genetic risk assessment in clinical practice”.

Not only are risk results likely to be often poorly understood by the tested individuals and their physicians, but also these results are often based on risk models, such as logistic regression models, that may not be good classification models (PEPE *et al.* 2004). Therefore, the disclaimer made by the companies that their services are not intended as medical advice cannot be overemphasized. Current knowledge of the role of most genes in complex diseases is at the group level of correlations of disease status with SNPs. Most of these SNPs were discovered via genetic association studies aimed at finding variants correlated with disease risk. It is hoped that these discoveries will provide insights into the pathogenesis and etiology, and ultimately lead to developments of new treatments or preventive therapies. Assuming these SNPs will also be effective classifiers, they are now being used in individual-level risk estimation, classification, and clinical decision-making. However, for many complex diseases, such as the ones discussed here (age-related macular degeneration [AMD], type II diabetes, inflammatory bowel disease [Crohn’s disease], and cardiovascular disease), it is arguable whether the era of genomics in personalized medicine is here yet. In this article, we discuss and explore how useful highly associated SNPs might be for individual-level risk estimation and prediction. Our focus will be on the classification accuracy of genetic testing, with an emphasis on two popular statistical methods for evaluating biomarkers. We give realistic real-data examples that illustrate that, currently, the genetic information is of limited value for personalized medicine. We also discuss and apply risk-based and classification-based analysis approaches to our AMD data.

### 6.3 TWO STATISTICAL METHODS

There are two basic statistical approaches for evaluating markers. The risk-based approach models the risk as a function of marker(s), often with adjustment for covariates, and is commonly applied in genetic studies. In case-control studies, this is done with logistic regression, and the markers with the strongest effect on disease risk are those associated with the smallest p-values and most extreme odds ratios (ORs). The other method, the classification-based approach, evaluates markers based on how well they can discriminate between cases and controls. The performance is evaluated by various measures, such as the proportion of positive test results among cases or the true positive fraction (TPF, or sensitivity) and the proportion of positive test results among controls or the false positive fraction (FPF, or 1–specificity). A perfect classifier will assign a positive test result to

everyone with the condition (TPF = sensitivity = 1) and a negative test result to everyone without the condition (FPF = 0, specificity = 1). Often more than one possible grouping into cases and controls is possible based on a classifier. The receiver operating characteristic (ROC) curve is a plot of all (FPF, TPF) pairs for each possible grouping. The area under the ROC curve (AUC) is a popular measure of the discrimination power of a classifier. It is the probability that given two random individuals, one who will develop the disease and the other who will not, the classifier will assign the former a positive test result and the latter a negative result. Theoretically, the AUC can take values between 0 and 1, but the practical lower bound is 0.5; a perfect classifier has an AUC of 1. Classifiers with an AUC significantly greater than 0.5 have at least some ability to discriminate between cases and controls. However, for screening of individuals with an increased risk of disease, it is suggested that the AUC be  $> 0.75$ , and for presymptomatic diagnosis of the general population, the AUC should be  $> 0.99$  (JANSSENS *et al.* 2007). When prognosis is the goal, one typically also evaluates the classification model by two additional measures: (1) the proportion of individuals who will develop the disease among those with a positive test result, or the positive predictive value (PPV), and (2) the proportion of individuals who will not develop the disease among those with negative test result, or the negative predictive value (NPV) (Box 1). We note in passing that there are other methods that model classification performance and have been applied in genetic studies, including, for example, genetic algorithms, generalized multifactor dimensionality reduction, and random forests (DUNAI *et al.* 2008; GOLD *et al.* 2006; JAKOBSDOTTIR *et al.* 2008). However, to keep our discussion focused, we do not discuss these other methods here.

Although the risk-based (logistic regression) and classification-based (ROC theory) methods do not yield contradictory results in terms of directionality, they can and often will differ in terms of size or importance. For example, a marker strongly related to risk may very well be a poor classifier; and vice versa, a good classifier may only be weakly associated with risk (PEPE *et al.* 2004). Furthermore, neither method directly measures calibration, which is how well the predicted risks agree with the underlying true risks (COOK 2007) (Box 2).

In a diagnostic setting in which discrimination between cases and controls is most important, it only matters that the cases have higher estimated risk, accurate or not, than the controls. However, when prognosis or risk stratification is the goal, both discrimination and calibration are important. We then need a model that both discriminates well between future cases and those who will remain controls, and also accurately estimates the exact risk of developing disease in the future.

## 6.4 THE ODDS RATIO, CLASSIFICATION, CALIBRATION, AND PREDICTION

The OR is widely used to evaluate markers, and it is assumed the markers associated with the most extreme OR are effective predictors. However, as we mentioned above, a marker strongly related to risk may very well be a poor classifier, and vice versa, a good classifier may only be weakly associated with risk (PEPE *et al.* 2004). In addition, a marker associated with risk may be well or poorly calibrated, that is, the predicted risk may agree well or poorly with the true risk (COOK 2007).

For a strongly associated marker to be effective in classification, the associated OR must be of an extreme magnitude rarely (if ever) seen in genetic association studies. As illustrated in fig. 6.1, if one wants to be able to detect 80% of cases with a binary marker, such as the presence or absence of a risk allele, with ORs of 1.5, 10, or 50, then about 73%, 29%, and 7% of the controls would be mislabeled as cases, and the AUC achieved by the binary marker would be 0.54, 0.76, and 0.86, respectively. Even a huge OR of 50 does not guarantee that a marker will have acceptable prediction accuracy; for example, the TPF may be unacceptably low (TPF = 55%, FPF = 2.4%, and AUC = 0.76) or the FPF unacceptably high (TPF = 97.6%, FPF = 45%, and AUC = 0.76) (fig. 6.1).

Let us examine the achievable AUC as a function of risk allele frequency under an additive genetic model in which the genotypes are coded 0, 1, and 2 (fig. 6.2 and table 6.1). In fig. 6.2, we have plotted the AUC for fixed values of the OR, as a function of risk allele frequency in cases (pca) under the assumption of Hardy-Weinberg equilibrium in both cases and controls. We clearly see that markers with a reasonably high OR of 3 have a maximum possible AUC of less than 0.70, and markers with an OR of 5 do not even reach an AUC of 0.80. For each OR, the risk allele frequency in controls (pco) corresponding to the maximum possible AUC is given on the plot, and not surprisingly, to reach the maximum possible AUC for each OR, the risk allele frequency difference between cases and controls has to be quite large (table 6.1). For example, to reach an AUC of 0.80 using a marker with an OR of 10, the allele frequencies in cases and controls would be quite different (pca = 0.49 and pco = 0.09) (table 6.1).



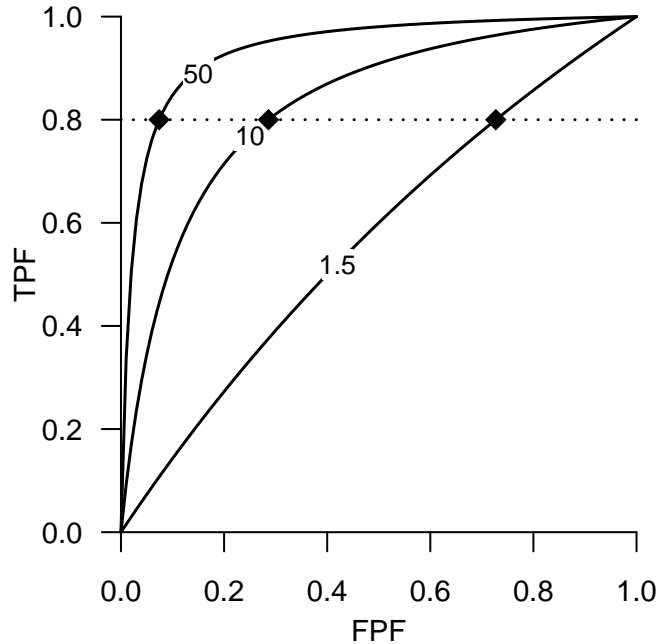


Figure 6.1: Accuracy curves for binary markers. The curves of accuracy points (FPF, TPF) for binary markers with ORs 1.5, 10, and 50 are plotted. The black diamonds and horizontal dotted line highlight the points  $(FPF, TPF)=(FPF, 80\%)$  on the accuracy curves. The ORs are marked on the curves.

## 6.5 THE ODDS RATIO, RELATIVE RISK, AND RISK

In retrospective studies, the relative risk or risk ratio (RR) cannot be estimated unless the prevalence is known, and therefore, the OR is used as a proxy. Theoretically, the OR will give a good approximation for the RR if the prevalence is low, but otherwise it tends to overestimate the RR (DAVIES *et al.* 1998; DEEKS 1998). RRs, which are the ratio of two risks (probabilities), are correctly interpreted as an estimate of how much more likely people sharing the same genotype combination are to develop the condition of interest when compared to a group without this genotype combination. The numerator of the RR is the risk of the condition given the genotype combination of interest, but clearly, the RR (or the OR) itself is not an estimate of individual-level risk and certainly not a diagnostic test or classifier.

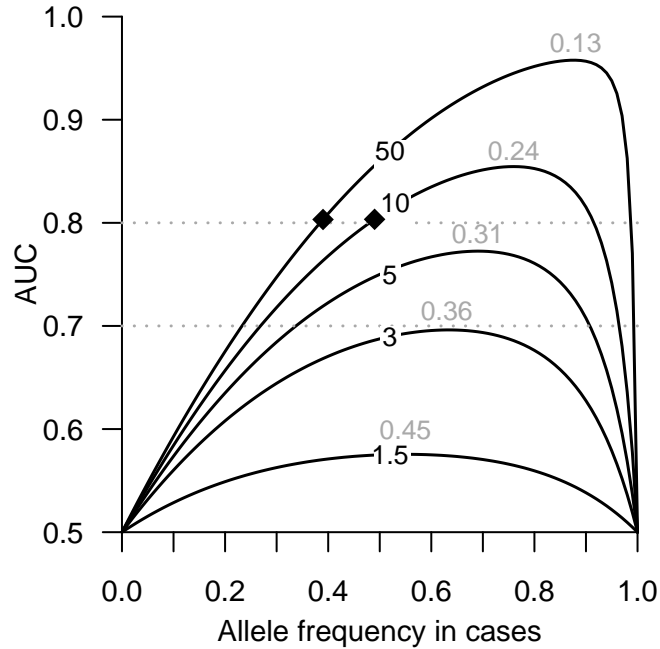


Figure 6.2: AUC for additive risk models of SNP markers as function of risk allele frequency in cases. The AUC is estimated for all risk allele frequencies in controls assuming additive ORs 1.5, 3, 5, 10, and 50 (the ORs are marked on the curves). The numbers in gray are the risk allele frequencies in controls corresponding to the maximum AUC for each OR. The dotted horizontal line in gray marks an AUC of 0.7 and 0.8. The black diamonds highlight the points  $(p_{ca}, AUC) = (p_{ca}, 0.80)$  for markers with additive ORs 10 and 50 (see table 6.1).

Statisticians should easily understand this relationship between OR, RR, and risk, but a person not trained in statistics (or science in general) may not make the same distinction as easily. Numerous studies in the genetic counseling literature have investigated what people make of risk estimates. For example, in a study of women’s perceived risk of breast cancer, 98% of women overestimated their risk of dying from breast cancer in 10 y by half to 8-fold when asked to quantify risk as a number out of 1,000. Interestingly, only 10% of those women thought they were at higher risk than an average woman their age (WOLOSHIN *et al.* 1999).

Table 6.1: AUC, risk allele frequency in cases ( $p_{ca}$ ) and controls ( $p_{co}$ ) for specific ORs in an additive model (genotypes coded 0-1-2 according to number of risk alleles).

OR	Maximum AUC	$p_{ca}$	$p_{co}$	AUC = 0.80	
				$p_{ca}$	$p_{co}$
1.5	0.58	0.55	0.45	NP	NP
3	0.70	0.63	0.36	NP	NP
5	0.77	0.69	0.31	NP	NP
10	0.85	0.76	0.24	0.49	0.09
50	0.96	0.88	0.13	0.39	0.01

NP=Not possible

## 6.6 CLINICAL VALIDITY AND UTILITY OF PREDICTIVE GENETIC TESTING

The clinical validity is measured by the discrimination ability of the marker, or its ability to classify people as cases or controls. The AUC, though imperfect, is a popular and easily interpretable measure of classification accuracy. It can be interpreted as the probability that predicted risk is higher for a case than a control. Various TPF and FPF pairs and various values of the AUC can correspond to the same OR (fig. 6.1). Thus, the OR by itself cannot give a meaningful indication of the probability of being correctly classified as case (TPF) or of the probability of being wrongly classified as a case (FPF), and alone its value is essentially useless to the individual.

The clinical utility of predictive genetic profiling for complex diseases rests on at least two conditions: (1) preventive means with high efficacy in the general population are available, and (2) these preventive means will also be effective in the genetically high-risk cohorts. Additionally, it is worth noting that for many complex diseases, known preventive lifestyle changes are broadly beneficial: weight loss, smoking cessation, blood pressure control, regular exercise, diets enriched with fruits and vegetables, etc., so to many individuals, it might be wasteful to spend 1,000 to find out they are genetically at increased risk for some condition only to have their doctor tell them all they can do is to lose weight and stop smoking. On the other hand, if the person is more likely to make lifestyle changes and stick to them, then the benefits can be great, both for the individual and the population as whole. Of course, the flip side is what the actions will be if the genetic test suggests lower than average risk for one or more specific conditions.

## 6.7 RECLASSIFICATION

The AUC attempts to measure the ability of a model to discriminate between cases and controls for a set of cutoff values that separate the two groups. However, on an individual basis, we also want the model to provide the best possible estimation of that person’s risk. One way to compare the accuracy of individual-level risk estimates of different risk models is to use the reclassification table approach (COOK 2007; PENCINA *et al.* 2008). In this approach, one measures how often subjects are estimated to be in different risk strata when different risk models are applied and whether the reclassification more accurately stratifies individuals into higher or lower risk strata. A marker that has a modest or no effect on the AUC can improve risk classification (COOK 2007). For example, suppose we are comparing two risk models that differ regarding a single individual’s membership in the 20%–30% risk stratum versus the 10%–20% risk stratum. If both models achieve the best discrimination by classifying everyone below the 40% risk threshold as controls and everyone above as cases, then the TPF and FPF will not be altered due to this person’s reclassification, but one model is more accurate than the other in terms of the true value of the individual’s risk estimate.

## 6.8 EXAMPLES

We now provide several examples, from the literature as well as from our own data, illustrating that although a set of SNPs can be strongly associated with disease risk with extremely small p-values, that same set of SNPs may not necessarily have high discrimination ability or may not dramatically improve the discrimination ability of a classification model constructed using “conventional” nongenetic risk factors without the SNPs.

### 6.8.1 Risk of Cardiovascular Events

In a recent replication study of nine SNPs associated with levels of either low-density lipoprotein (LDL) or high-density lipoprotein (HDL) cholesterol, KATHIRESAN *et al.* (2008) created a genotype score on the basis of the total number of unfavorable alleles at these risk SNPs, and investigated the classification accuracy of the genotype score and the effect on reclassification beyond standard risk factors for cardiovascular events. The authors found that accounting for the effect of the nine SNPs

did not improve the classification accuracy of their model. The ROC curves with and without the genotype score lined up almost perfectly, and both had an AUC of 0.80 despite the SNPs having  $P$  values as low as  $10^{-29}$ , with six out of nine SNPs having  $P$  values  $< 10^{-6}$  (Appendix D and Table D1). Adding the genotype score to the model did, however, modestly improve the reclassification. Unfortunately for this dataset, the classification accuracy of the genotype score alone was not estimated. Nevertheless, these data provide an example of highly associated variants that do not markedly improve the discrimination ability of a model, yet at the same time, they give hope that genetic variants may become valuable prognostic tools.

### 6.8.2 Risk of Type 2 Diabetes

In type 2 diabetes, 12 SNPs (SCOTT *et al.* 2007; SLADEK *et al.* 2007; WEEDON *et al.* 2006) with  $P$  values as low as  $10^{-34}$  (Appendix D and Table D2) reach an AUC of 0.64, suggesting only fair discrimination power. We arrived at this AUC of 0.64 using only published allele frequencies; we did this using the method of LU and ELSTON (2008) (Appendix D, Estimating the AUC from meta-data). LU and ELSTON (2008) also applied their method to a model of the same 12 SNPs and four additional environmental factors, and got a slightly improved AUC of 0.67.

### 6.8.3 Risk of Prostate Cancer

A genetic classification model of two prostate cancer risk SNPs in low linkage disequilibrium with each other (YEAGER *et al.* 2007) has an AUC of 0.56, based on the method of LU and ELSTON (2008). An AUC of this magnitude suggests that the model has a very poor discrimination power. The SNPs have  $P$  values of  $10^{-13}$  and  $10^{-14}$ , but the genotype-specific ORs are not extreme and range from 1.3 to 2.2 (D and Table D3).

### 6.8.4 Risk of Inflammatory Bowel Disease

A genetic classification model of five well-replicated genetic associations (CUMMINGS *et al.* 2007; CUMMINGS *et al.* 2007; DUERR *et al.* 2006; PARKES *et al.* 2007; RIOUX *et al.* 2007) in inflammatory bowel disease (Crohn's disease) has an AUC of only 0.66. This suggests only fair discrimination power for Crohn's disease despite the variants being highly significant ( $P$  values range from  $10^{-7}$

to  $10^{-14}$ ) and one SNP having quite an extreme OR of 0.26 (1/4). Again, the method of Lu and Elston was used to estimate the AUC [20]. For more details, see Appendix D and Table D4.

### 6.8.5 Risk of Age-Related Macular Degeneration

Using our previous published AMD data (JAKOBSDOTTIR *et al.* 2008) on the *CFH*, *LOC387715*, and *C2* variants, we plotted the ROC curves and estimated the AUC and positive predictive values of one-, two-, and three-factor models (detailed methods are in Appendix D). Fig. 6.3 displays the ROC curves for the null model and for five genetic risk models: the three-factor model of *CFH*, *LOC387715*, and *C2* SNPs, the two-factor model of *CFH* and *LOC387715*, and all of the one-factor models. We see that to correctly identify about 74% of the cases using the three-factor model, we would wrongly classify 31% of the controls, and for the TPF to be around 80%, the FPF needs to be unacceptably high (>40%). The AUC for the three-factor model is quite high, 0.79, and significantly different from 0.5 (95% confidence interval [CI] 0.74–0.83) (table 6.2). Table 6.2 also gives the results of logistic regression analysis: the ORs for additive inheritance of *CFH* and *LOC387715* risk alleles are about 3 with *P* values of around  $10^{-13}$ .

We also plotted the integrated predictiveness and classification plot, which combines information from both the risk- and classification-based analysis approaches discussed above (PEPE *et al.* 2008). In the integrated plot (fig. 6.4), there are two aligned plots: in the top plot, ordered individual risks are plotted as function of the risk percentile, and in the bottom plot, the TPF and FPF are plotted as a function of the risk percentile such that at each point, the TPF and FPF are calculated for the risk threshold equal to the risk associated with the corresponding risk percentile. If we now look at the integrated predictiveness and classification plot for the three-factor model, we see that the TPF and FPF pair 74% and 31% corresponds to the 35% risk percentile (fig. 6.4, bottom panel), which then corresponds to choosing an AMD risk of 4% as the cutoff point for classifying individuals (fig. 6.4, top panel). Those with risk greater than 4% are assumed to be at high risk and are classified as cases, and those with lower risk are classified as controls. To illustrate this, suppose we have a population of size 1,000 and a prevalence of 5.5% (which is the prevalence of advanced AMD in the U.S. in white individuals 65 y or older according to FRIEDMAN *et al.* (2004) and the U.S. 2000 census data—see Appendix D for further details). If the prevalence is 5.5%, there would be 55 cases in our population. Of those 55 cases, 74%, or 41, would be correctly considered to be at high risk of AMD, and 31%, or 293, of the true 945 controls would be wrongly

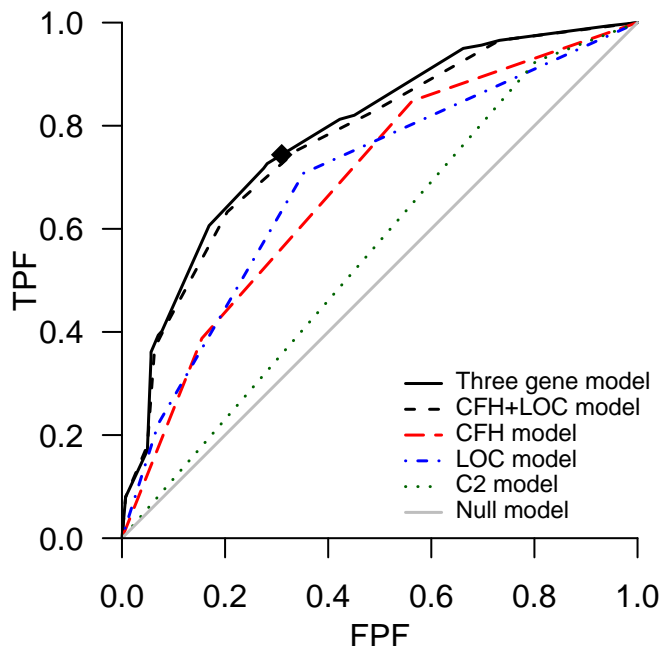


Figure 6.3: ROC curves for AMD classification models. The black diamond highlights the point  $(\text{FPF}, \text{TPF}) = (31\%, 74\%)$  on the ROC curve of the three-factor model of *CFH*, *LOC387715*, and *C2*. The gray line for reference gives the “chance” classification rule: the farther the ROC curve is from the chance line, the better the classification rule.

assumed to be at high risk. Therefore, out of the 334 (41+293) individuals in the high-risk group, 88% should actually be in the low-risk group, or in other words, the PPV would be only 12% (i.e., 100%–88%). When designing a clinical trial to test preventive therapies in high-risk cohorts based on genotyping alone, it may or may not be cost effective to have 12% (instead of 5.5%) of the study cohort as true cases. However, as a clinical test, it may be considered unethical to needlessly alarm 88% of the high-risk cohort, especially when limited treatment and preventive options are available (YOUNG 2007).

To lower the proportion of controls in the high-risk cohort, a more stringent threshold for calling someone high risk, say 25%, can be used instead of the 4% threshold used above. However, using this higher risk threshold only lowers the proportion of controls in the high-risk group from 88% to 84%, as can be seen in this manner: the plot (fig. 6.4, top panel) shows that the risk threshold of 25% corresponds to the 85% risk percentile. Looking at the classification curve (fig.

Table 6.2: Results of logistic regression and ROC analysis.

Model Factors	Logistic regression		ROC analysis	
	OR	<i>P</i> value	AUC	95% CI
Model 1			0.79	0.74–0.83
<i>CFH</i>	2.89	$9.1 \times 10^{-13}$		
<i>LOC387715</i>	3.42	$2.3 \times 10^{-13}$		
<i>C2</i>	0.39	$1.3 \times 10^{-3}$		
Model 2			0.77	0.73–0.82
<i>CFH</i>	3.00	$9.1 \times 10^{-14}$		
<i>LOC387715</i>	3.38	$2.5 \times 10^{-13}$		
Model 3			0.69	0.64–0.73
<i>CFH</i>	2.77	$2.1 \times 10^{-13}$		
Model 4			0.69	0.65–0.74
<i>LOC387715</i>	3.11	$6.2 \times 10^{-13}$		
Model 5			0.56	0.53–0.60
<i>C2</i>	0.33	$1.9 \times 10^{-5}$		

The ORs for each variant is for an additive model in which the genotypes are coded 0-1-2.

The confidence intervals (CIs) for the AUC are asymptotic and derived using the DeLong’s estimator (ZHOU *et al.* 2002) for the variance

6.4, bottom panel), we see that the 85% risk percentile corresponds to a TPF of 17% and FPF of 5%. Again, to put these numbers in perspective, let us again assume we have a population of size 1,000. Nine (17%) out of 55 true cases would then be correctly classified as “high risk”, and 47 (5%) out of 945 controls would be incorrectly classified as high risk. Therefore 84% ( $47/56 = 47/(9+47)$ ) of those classified as “high risk” would actually be controls (PPV =  $100\% - 84\% = 16\%$ ).

When applied to case–control data, the integrated predictiveness and classification plot depends on the assumed prevalence of the disease, which may not be known with precision or may, as in the case of AMD, depend strongly on age. Note that as the prevalence changes, the bottom plot does not change, only the top plot does: although it still will look essentially the same, the risks will be more spread out between 0 and 1 as the prevalence gets higher and less spread out otherwise.

Second, it is worth noting how the results of our AMD example change if different values for the prevalence are used. The prevalence of AMD is highly age-dependent, and in table 6.3, we calculate the PPV using prevalence estimates for different age groups. If the prevalence increases,



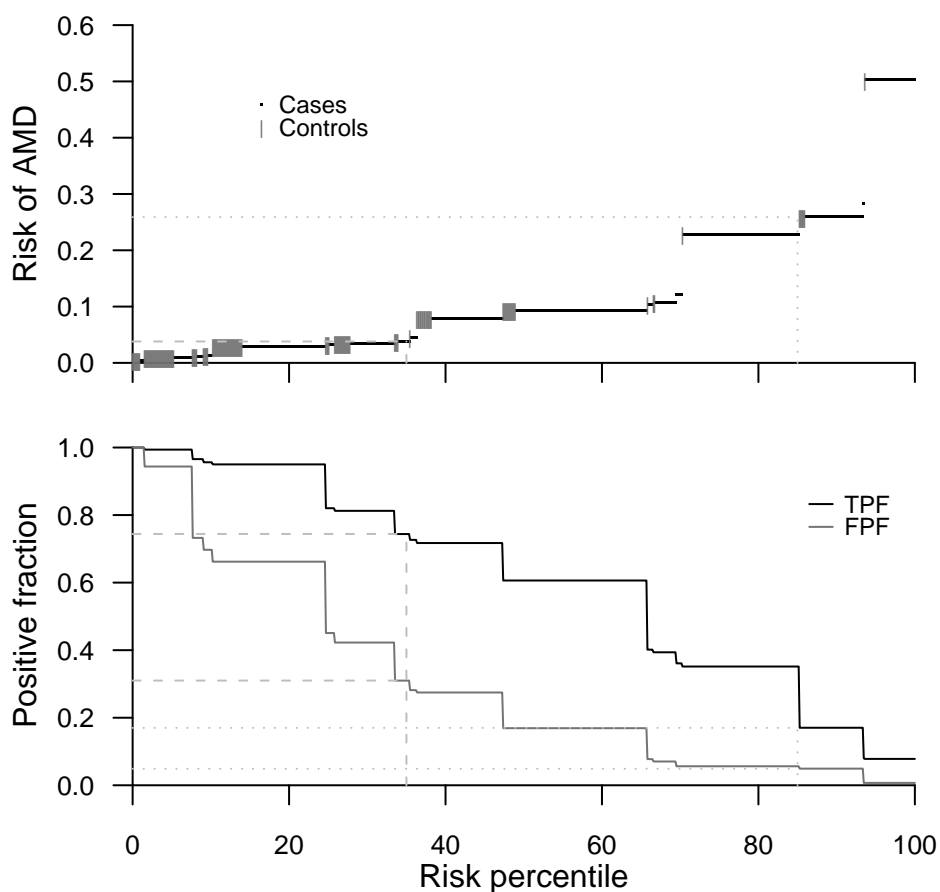


Figure 6.4: Integrated predictiveness and classification plot for the three-factor model. The light gray lines show how the plots are used in the examples given in the text: the dashed lines are for the first example with  $TPF = 74\%$ ,  $FPF = 31\%$ , risk percentile = 35%, and AMD risk threshold = 4%; and the dotted lines are for the second example with AMD risk threshold = 25%, risk percentile = 85%,  $TPF = 17\%$  and  $FPF = 5\%$ . On the top panel, the risks for cases are marked with a dot in black while the risks for controls are marked with a vertical line segment in dark-gary.

the results are less disappointing (PPV increases) but are even worse if it decreases (table 6.3). Clearly, the ability to discriminate between current cases and controls, based on genotype data from *CFH*, *LOC387715*, and *C2* alone, changes with age. A crude estimate of the lifetime risk at age 80 y, given a genetically high-risk score based on the three variants, is 30% compared to 15% baseline lifetime risk at age 80 (table 6.3).

Table 6.3: Positive predictive values (PPV) for different values of the prevalence.

Prevalence	Age group	Risk threshold	PPV
15%	80 y and older	10%	30%
5.5%	65 y and older	4%	12%
1.5%	40 y and older	1%	3%

The risk threshold corresponds to  $TPF = 74\%$  and  $FPF = 31\%$ .  
(as in the first example in the text).

PPV = proportion of cases in the high-risk group.

1-PPV = proportion of controls in the high-risk group.

## 6.9 DISCUSSION OF THE AMD EXAMPLE

If the primary goal of genetic diagnostic tests for AMD were to identify those who are at high risk before they show irreversible degenerative changes to maximize the effectiveness of long-term preventive strategies, then we would want to test individuals 40–55 y old (or younger) to predict whether they will develop AMD before age 80 y. Our case-control data presented here do not fully measure the ability of genetic data to predict future disease status (prognosis) for several reasons: (1) AMD prevalence increases with age, (2) females have higher prevalence in all age groups compared to males, (3) females live longer, (4) the FPF derived from case-control data is overestimated because some controls will develop AMD as the cohort ages, (5) the case/control counts are unbalanced, so our sample may not be optimal for estimating the classification accuracy of the markers (JANES and PEPE 2006), and (6) the estimates of the ORs, and estimates from most other AMD case-control studies, are based on the comparison of extreme phenotypes: a group of individuals with advanced AMD are contrasted with a control group of individuals with no or very minimal clinical findings. Therefore, they very likely overestimate the RR and the discrimination power for individuals with intermediate clinical findings. Even accounting for all these issues in an optimistic manner, the overall conclusions of our analysis are unlikely to change dramatically. Proper analyses of longitudinal cohort data using survival analysis techniques could lead to a more precise assessment of the potential value of genetic data in predicting lifetime AMD status (MOSKOWITZ and PEPE 2004; PEPE *et al.* 2008).

The major achievements that have been made in understanding the genetics of AMD are well known, and the AMD discoveries ([EDWARDS \*et al.\* 2005](#); [HAGEMAN \*et al.\* 2005](#); [HAINES \*et al.\* 2005](#); [JAKOBSDOTTIR \*et al.\* 2005](#); [KLEIN \*et al.\* 2005](#); [RIVERA \*et al.\* 2005](#)) are widely mentioned as the first “proof” that genome-wide association analysis works (although the majority of the AMD studies were not genome-wide association studies, but rather targeted searches following up regions of linkage). The results have been so exciting that perhaps all of us who study AMD are guilty of overstating our results. Here are just a few examples:

“Nevertheless, with all the genetic findings, it may soon be possible to provide pre-symptomatic diagnosis with reasonable accuracy, leading to better disease management strategies for high-risk individuals.”—[SWAROOP \*et al.\* \(2007\)](#)

“The continued support for these genes in ARM susceptibility will hopefully bring us closer to being able to utilize the information in these genes to identify at risk individuals and provide a rational basis for future clinical trials to test preventive therapies in high-risk cohorts.”—[CONLEY \*et al.\* \(2006\)](#)

“Expressed another way, these genotypes apparently identify individuals whose lifetime risk of AMD ranges from less than 1% to more than 50%; however, longitudinal studies are needed to define the true risk attributable to these loci and the ways in which these might interact with the known environmental and lifestyle risk factors.”—[MALLER \*et al.\* \(2006\)](#)

All these statements are scientifically valid, they are carefully worded, and it is clear the investigators are talking about “potential”, “future”, and “hope”. Nevertheless, they can and have been overinterpreted. For example, a recent review ([ROSS \*et al.\* 2007](#)) cites [MALLER \*et al.\* \(2006\)](#) and states:

“SNPs in complement factor H (CFH) and PLEKHA1/ARMS2/HtrA1 capture a substantial fraction of AMD risk and permit the identification of individuals at high risk of developing AMD.”

Even *Nature Genetics* appears to also overstate the potential impact of AMD genetics. In the December 2007 issue ([EDITORIAL 2007](#)), the editors discuss the new hype about personalized genomics and ask: With the possible exception of age-related macular degeneration, how much can we say with confidence about the spectrum of risk? However, as we have shown here, we cannot yet make an exception for AMD. We should, however, not let this discourage us. The discoveries of the AMD risk genes are truly amazing, and they should of course encourage and guide future research. In fact, the discovery of the likely involvement of the CFH gene gave firmer footing to the hypothesis that the abnormal function of complement pathway can cause AMD and has resulted in discoveries of other AMD genes in this pathway ([GOLD \*et al.\* 2006](#); [DINU \*et al.\* 2007](#); [MALLER \*et al.\* 2007](#); [YATES \*et al.\* 2007](#)).

## 6.10 CONCLUSIONS

Genetic association studies have identified many susceptibility variants for complex diseases and, in many cases, added to the understanding of the etiology of the diseases. However, as we discuss here using real data and theoretical examples, strong association does not necessarily guarantee good classification or discrimination ability. Before using association results for classification and risk estimation purposes, we need to establish their effectiveness formally using appropriate measures and, ideally, appropriate study designs. Additionally, when evaluating the improvement in the predictive value by adding a marker to a prediction model, we may need to use additional measures besides the AUC, such as reclassification tables.

In our examples, we saw that the addition of nine highly significant risk SNPs to the risk model could not improve the discrimination power for cardiovascular events beyond standard risk factors. For type 2 diabetes, the classification rule based on 12 SNPs gave an AUC of only 0.64, a value that is well below the guidelines of 0.75 and 0.99 cutoffs for screening and prognosis purposes, respectively. For Crohn's disease, a classification model based on five SNPs gave an AUC of only 0.66, and for prostate cancer, a model of two SNPs achieves an AUC of only 0.56. Both values are well below the 0.75 and 0.99 cutoffs. For AMD, the AUC of a model with three SNPs was 0.80, but the proportion of positive test results among affected individuals was only 30%, 12%, and 3%, depending on assumed prevalence (15%, 5.5%, and 1.5%, respectively). The results of these four examples, although somewhat disappointing, are not surprising given the theoretical results of [JANSSENS \*et al.\* \(2007\)](#), [JANSSENS \*et al.\* \(2006\)](#) that indicate that achieving a high AUC requires a much larger number of genetic variants than we have to date. For example, Janssens *et al.* demonstrated that for genetic profiling, on average 80 common variants with ORs of 1.25 each were needed to develop a model useful for identification of high-risk individuals (AUC<sub>c</sub>0.80).

Even though our examples illustrate that highly associated SNPs may not be effective as classifiers, it should not be concluded that the association findings are not important nor that association studies are not valuable. In many cases, the association discoveries have and will continue to result in new etiological hypotheses previously not considered. For example, in the case of AMD, the CFH discovery ([EDWARDS \*et al.\* 2005](#); [HAGEMAN \*et al.\* 2005](#); [HAINES \*et al.\* 2005](#); [KLEIN \*et al.\* 2005](#)) resulted in a new focus on the complement pathway and subsequent identification of additional novel disease genes in that pathway ([GOLD \*et al.\* 2006](#); [DINU \*et al.\*](#)

2007; MALLER *et al.* 2007; YATES *et al.* 2007). The scientific community should be very cautious to avoid overhyping association findings in terms of their “personalized medicine” value before their time, lest we lose the goodwill and support of the general public.

## 7.0 SCORE STATISTICS FOR X-LINKED QTLs

This part of the dissertation focuses on linkage mapping methods for X-linked QTLs. In the next sections I explain issues specific to X-linked inheritance and give an overview of linkage methods for QTLs and.

### 7.1 X-LINKED INHERITANCE

As mentioned previously the genetic material in humans is stored in 23 pairs of chromosomes and most human cells contain 46 chromosomes. Two of these are the sex chromosomes, two paired X's in females and an X and a Y in males. The remaining 22 pairs are the homologous pairs called autosomes. One chromosome of each homologous pair is maternally transmitted via an egg and the other is paternally transmitted via a sperm. The eggs contain 22 autosomes and an X but the sperm contain 22 autosomes and an X or a Y.

Most female cells contain two X chromosome but a simple dosage model (such as an additive model) may not be realistic for all X-linked loci. In each female cell 75% of loci are believed to be expressed only from one of the X chromosomes while the other chromosome is inactivated. About 15% of loci are estimated to escape inactivation (or have incomplete inactivation) and 10% of loci exhibit rates of inactivation that vary widely among individual females (CARREL and WILLARD 2005). Typically either the whole maternally or the whole paternally derived X chromosome is inactivated (randomly) during mitosis so that in one cell all the genes are expressed from the same chromosome. However, between cells it varies which chromosome is inactivated. In humans the timing of inactivation occurs early on in the development of the embryo. During the very early stages in the development of the placenta and other organs that support the embryo, inactivation is imprinted such that paternally derived X chromosome is always inactivated. Therefore, in both

female and male embryos the placenta only expresses genes from the mother's X chromosome. Later in the development this is reversed and random inactivation occurs in all cells that form other tissues. Since all the daughter cells have the same chromosome inactivated and after a certain time-point the inactivation cannot be reversed, cell patches are formed such that all the cells in any particular patch have the same X chromosome inactivated.

The pseudoautosomal regions, PAR1 and PAR2, are homologous sequences on the X and Y chromosomes. The regions are called pseudoautosomal because any genes located within them are inherited similarly to the autosomal genes. Males have two copies of these genes, one in the pseudoautosomal region on Y and the other on the corresponding region on the X chromosome. Females also have two copies of the pseudoautosomal genes, as each of their two X chromosomes contains a pseudoautosomal region. The pseudoautosomal regions allow the X and Y chromosomes to pair and properly segregate during meiosis in males. Crossing over can occur between the pseudoautosomal regions on the X and Y chromosomes in males. Therefore, females can inherit an allele originally from their father's Y chromosome and likewise males can inherit an allele originally from their father's X chromosome. Most mapping studies support at least one obligatory crossover in the PAR1 and it has been suggested that sons have 50% chance of receiving the Y PAR1 haplotype from their fathers as a whole without recombination ([FLAQUER \*et al.\* 2008](#)).

## 7.2 OVERVIEW: QTL LINKAGE METHODS

An early method for investigation of linkage of QTLs is the Haseman-Elston Regression (HE-regression) ([HASEMAN and ELSTON 1972](#)). In HE-regression the squared trait differences of sib pairs  $y_D$  are regressed on the estimated proportion of alleles at a locus shared identical by descent. Various extensions have been made to improve the original HE-regression and an X-linked version has also been developed ([WIENER \*et al.\* 2003](#)). An important extension of the HE-regression was based on the observation that the squared trait sum adds independent information to the regression model ([DRIGALENKO 1998](#)). Thus, the method has been improved by changing the dependent variable from the squared difference to the mean-corrected product of the sib-pair trait values ([ELSTON \*et al.\* 2000](#)).

Later, variance components (VC) methods were developed. The basic idea of the VC model is to fit a multivariate normal model to the trait values of relatives with a variance-covariance matrix that expresses the covariance between relatives as function of the IBD sharing at a marker locus (AMOS 1994; GOLDGAR 1990; SCHORK 1993). The VC model led to development of powerful statistics for both linkage and association analyses. The major strengths of the model are that the framework allows models to be easily extended to incorporate household effects, individual-specific covariates, and interactions (BLANGERO and ALMASY 1997). Since the power of the method is proportional to the heritability, power can often be increased by reducing the residual variance by adjusting for covariates (ZEEGERS *et al.* 2004). Naive incorporation of covariates can, however, inflate the type I error (ZEEGERS *et al.* 2004; PURCELL and SHAM 2002). The main disadvantage of the VC model is its sensitivity to violations of normality. This is of particular importance as the assumption of normality is often violated under the alternative. The VC models have been extended for X-linked traits; EKSTRØM (2004) assumed an additive model, ignoring the possibility of X-inactivation in females, while LANGE and SOBEL (2006) and KENT *et al.* (2005) allowed for X-inactivation.

The latest statistics for QTL mapping are regression-based score statistics (or score tests) proposed by numerous authors (LEBREC *et al.* 2004; TANG and SIEGMUND 2001; PUTTER *et al.* 2002; WANG and HUANG 2002b; WANG and HUANG 2002a; WANG 2002). The statistics attempt to achieve the power of the VC models while being less sensitive to non-normality and selected samples. The score test is a locally most powerful test and can be computationally fast, and much faster than the VC methods on large pedigrees. Score tests are simply the partial derivative of the likelihood with respect to the linkage parameter(s) evaluated under the null hypothesis of no linkage and standardized (by the information or an estimate of the standard error). The theoretical background of score tests for autosomal QTL mapping is more or less complete and recently investigation of appropriate methods to estimate the standardization factor was published (BHATTACHARJEE *et al.* 2008). The choice of an appropriate standardization factor directly affects power and so is very important. To my knowledge the score statistics have not been extended for X-linked traits, which is the goal of this part of the dissertation.



### 7.3 MODEL

In order to develop score statistics for X-linked traits, we assume the quantitative phenotypic value  $y_i$  of individual  $i$  is influenced by a X-linked locus in the following manner:

$$y_i = \mu_s^i + g_{kl}^i + e_i \quad (7.3.1)$$

where  $\mu_s$  is the sex-specific mean ( $s = m, f$ ),  $g_{kl}$  represents the genotypic value of the X-linked QTL for genotype  $k/l$  in females or  $k$  in males, and  $e$  represents the residuals, i.e., the environmental effect that is unique to each individual, and which is assumed to be uncorrelated between individuals. Assuming the genotypic values and the residuals are independent, the phenotypic variance  $\sigma_y^2$  can be partitioned as

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2 \quad (7.3.2)$$

where  $\sigma_g^2$  is the variance attributed to the X-linked QTL and  $\sigma_e^2$  the random noise.

For an additive model in females we define  $g_{kl} = \alpha_k + \alpha_l + \delta_{kl}$  and for males we define  $g_{kl} = g_k = \beta_k$ . We interpret  $\alpha_k$  as the effect of allele  $k$  in females and  $\beta_k$  is the effect in males and  $\delta_{kl}$  is the dominance deviation from an additive model. The inactivation of one of the X chromosomes in females suggests that dosage compensation needs to be modeled. However, we start by assuming that all loci escape inactivation and use the additive model. The main reason to start with this simplifying assumption is the fact that the same X chromosome needs not to be inactivated in all cells. Therefore, to model inactivation the most appropriate model is probably a model where the contribution of each allele is weighted by 0.5 in females. Hence, the derivation of the score statistics would be essentially the same as for an additive model in females. We denote the population allele frequency as  $p_k$  and assume Hardy-Weinberg equilibrium and equal allele frequencies in females and males.

### 7.4 ALLELIC AND GENOTYPIC EFFECTS

No generality is lost if all the trait values are standardized so that  $E[y] = 0$  and  $E[e] = 0$ . In this section we therefore work with the model  $y_i = g_{kl}^i + e_i$ . Note, that the means  $\mu_s^i$  we have subtracted from each side may be different for males and females. However, the global mean remains zero as

this sex-specific standardization gives

$$\begin{aligned} E[y] &= E[y|\text{female}]P[\text{female}] + E[y|\text{male}]P[\text{male}] \\ &= 0P[\text{female}] + 0P[\text{male}] = 0 \end{aligned}$$

In reality, the strictly additive model in females may not fit perfectly and so the allelic effects  $\alpha_k$  and the dominance deviation  $\delta_{kl}$  are estimated such that  $|\delta_{kl}| = |g_{kl} - \alpha_k - \alpha_l|$  are minimized. Minimizing  $\delta_{kl}$  for all  $k, l$  simultaneously is equivalent to minimizing  $\sum_{k,l} p_k p_l \delta_{kl}^2$ , where  $p_k$  is the population allele frequency of allele  $k$ . Now under our assumption of standardized trait values  $E[y] = 0$  we get

$$\begin{aligned} 0 &= E[y|\text{female}] = E[\alpha_k + \alpha_l + \delta_{kl}] \\ &= \sum_{k,l} p_k p_l \alpha_k + \sum_{k,l} p_k p_l \alpha_l + \sum_{k,l} p_k p_l \delta_{kl} \\ &= 2 \sum_k p_k \alpha_k + \sum_{k,l} p_k p_l \delta_{kl} \end{aligned}$$

So  $2 \sum_k p_k \alpha_k = - \sum_{k,l} p_k p_l \delta_{kl}$ . To minimize  $\sum_{k,l} p_k p_l \delta_{kl}^2$  we take the partial derivatives and set to zero

$$\frac{\partial}{\partial \delta_{kl}} \sum_{k,l} p_k p_l \delta_{kl}^2 = 2 \sum_{k,l} p_k p_l \delta_{kl} = 0$$

Therefore  $\sum_{k,l} p_k p_l \delta_{kl} = 0$  and  $2 \sum_k p_k \alpha_k = - \sum_{k,l} p_k p_l \delta_{kl} = 0$  and we have  $E[\alpha] = \sum_k p_k \alpha_k = 0$  and  $E[g] = 0$ . The effect of allele  $k$  in males is  $\beta_k = g_k$  and so  $E[\beta] = E[g] = E[y|\text{male}] = 0$ . Hence we can assume that  $E[g] = 0$ , which will simplify our derivations in later sections.

## 7.5 VARIANCES-COVARIANCES

### 7.5.1 X-linked kinship coefficient and variances-covariances

Genetic identity coefficients are powerful tools for genetic analysis. Here we define the kinship coefficient and condensed identity coefficients for non-inbred individuals.

The kinship coefficient  $\phi_{ij}$  is a simple measurement of relationship between two relatives  $i$  and  $j$ . Namely  $\phi_{ij}$  is the probability that allele selected randomly from  $i$  and an allele selected randomly from the same autosomal locus of  $j$  are IBD. Thus the kinship coefficient takes into

account the common ancestry of  $i$  and  $j$  as defined by the pedigree structure but ignoring their genotypes (LANGE 2003; MALCOT 1948).

The IBD relation partitions the four alleles of ordered genotypes,  $a_i^f|a_i^m$  and  $a_j^f|a_j^m$ , of individuals  $i$  and  $j$  into equivalent classes or identity states; superscript  $f$  and  $m$  denote the allele inherited from the mother and father, respectively. The detailed identity states keep track of the IBD status according to the ordered genotypes (JACQUARD 1966). The condensed identity states, however, keep track of how many alleles are IBD within and between individuals. At an autosomal locus there are a total of 15 detailed identity states and 9 condensed identity states. The condensed identity coefficient  $\Delta_c$  is the probability of the condensed identity state  $S_c$  (Figure 7.1) .

For two non-inbred individuals, the (autosomal) kinship coefficient and the (autosomal) condensed identity coefficients are related according to

$$\phi_{ij} = \frac{1}{2}\Delta_7 + \frac{1}{4}\Delta_8$$

If we let  $\psi_{ij}$  be the X-linked version of the kinship coefficient then for two non-inbred pairs we have the relation

$$\psi_{ij} = \begin{cases} \frac{1}{2}\Delta_{7,X} + \frac{1}{4}\Delta_{8,X} & \text{for female-female pair} \\ P_1 & \text{for male-male pair} \\ \frac{1}{2}\Lambda_3 & \text{for female-male pair} \end{cases}$$

where  $\Delta_{c,X}$ ,  $P_c$ , and  $\Lambda_c$  are the sex-specific condensed identity coefficients for female-female, male-male, and female-male relative pairs, respectively (Figure 7.2). Note that  $\Delta_{c,X}$  will generally be different from  $\Delta_c$ .

In next section we will derive the covariance formula for all three types of relative pairs (female–female pairs, male–male pairs and female–male pairs). For comparison the covariance for autosomal loci is

$$Cov(y_i, y_j) = 2\psi_{ij}\sigma_a^2 + \Delta_{7,X}^{ij}\sigma_d^2 \tag{7.5.1}$$

where  $\sigma_a^2$  is the variance component accounted for by the additive genetic effect of autosomal loci and  $\sigma_d^2$  is the variance component accounted for by the dominance genetic effect.

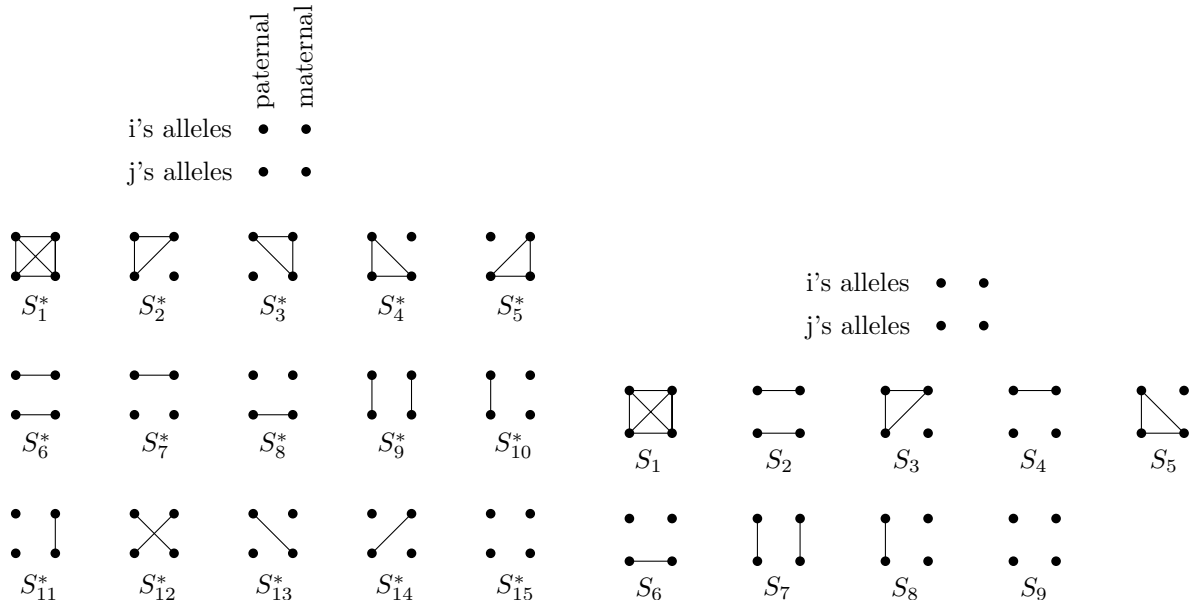


Figure 7.1: Detailed identity states on the left and condensed states on the right. The autosomal and female-female sex-specific states are the same.  $\Delta_c$  is the probability of state  $S_c$  at autosomal loci and  $\Delta_{c,X}$  the probability at X-linked loci. Alleles have lines drawn between them if they are IBD. Therefore condensed states  $S_1, S_2, S_3, S_4, S_5$  and  $S_6$  all have probability zero in non-inbred individuals.  $S_1 = S_1^*$ ,  $S_3 = S_2^* \cup S_3^*$ ,  $S_5 = S_4^* \cup S_5^*$ ,  $S_7 = S_9^* \cup S_{12}^*$ , and  $S_8 = S_{10}^* \cup S_{11}^* \cup S_{13}^* \cup S_{14}^*$ . The graphical presentation of identity states is adapted from Jacquard (JACQUARD 1974) and Lange (LANGE 2003).

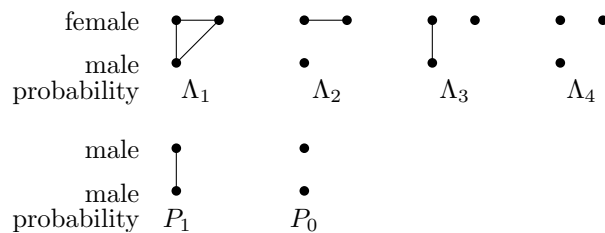


Figure 7.2: The condensed identity states for X-linked loci of female-male and male-male pairs. The probability of each state is given in the figure. For non-inbred individuals  $\Lambda_1 = \Lambda_2 = 0$ . Adapted after PAN *et al.* (2007), JACQUARD (1974), and LANGE (2003).

### 7.5.2 Female-Female relative pairs

In this section we derive the covariance between two non-inbred females by conditioning on various identity states (Figure 7.2) and using the facts  $\sum_k p_k = 1$ ,  $E[\alpha] = \sum_k p_k \alpha_k = 0$ ,  $\sum_j p_j \delta_{kj} = 0$ , and  $\alpha_k = \sum_j p_j g_{kj}$ . Since the trait values  $y$  have been standardized, we have

$$\begin{aligned}
Cov(y_i, y_j) &= Cov(g_{kl}^i, g_{kl}^j) && \text{trait values standardized} \\
&= E[g_{kl}^i g_{uv}^j] - E[g_{kl}^i] E[g_{uv}^j] && \text{general covariance formula} \\
&= E[g_{kl}^i g_{uv}^j] && E[g] = 0 \\
&= E[g_{kl}^i g_{uv}^j | 2 \text{ alleles IBD}] \Delta_{7,X}^{ij} \\
&\quad + E[g_{kl}^i g_{uv}^j | 1 \text{ allele IBD}] \Delta_{8,X}^{ij} && \text{condition on IBD status} \\
&\quad + E[g_{kl}^i g_{uv}^j | 0 \text{ alleles IBD}] \Delta_{9,X}^{ij} \\
&= \Delta_{7,X}^{ij} \sum_{k,l} (\alpha_k + \alpha_l + \delta_{kl})^2 p_k p_l && k = u; l = v \text{ if both alleles IBD} \\
&\quad + \Delta_{8,X}^{ij} \sum_{k,l,v} (\alpha_k + \alpha_l + \delta_{kl})(\alpha_k + \alpha_v + \delta_{kv}) p_k p_l p_v && k = u \text{ if one allele IBD} \\
&\quad + \Delta_{9,X}^{ij} \sum_{k,l,u,v} (\alpha_k + \alpha_l + \delta_{kl})(\alpha_u + \alpha_v + \delta_{uv}) p_k p_l p_u p_v \\
&= \Delta_{7,X}^{ij} [2 \sum_k \alpha_k^2 p_k + \sum_{k,l} \delta_{kl}^2 p_k p_l] + \Delta_{8,X}^{ij} \sum_k \alpha_k^2 p_k \\
&= 2[\frac{1}{2} \Delta_{7,X}^{ij} + \frac{1}{4} \Delta_{8,X}^{ij}] 2 \sum_k \alpha_k^2 p_k + \Delta_{7,X}^{ij} \sum_{kl} \delta_{kl}^2 p_k p_l \\
&= 2[\frac{1}{2} \Delta_{7,X}^{ij} + \frac{1}{4} \Delta_{8,X}^{ij}] 2E[\alpha^2] + \Delta_{7,X}^{ij} E[\delta_{kl}^2] \\
&= 2[\frac{1}{2} \Delta_{7,X}^{ij} + \frac{1}{4} \Delta_{8,X}^{ij}] 2Var[\alpha] + \Delta_{7,X}^{ij} Var[\delta] && E[\alpha] = E[\delta] = 0 \\
&= 2\psi_{ij} \sigma_{a,f}^2 + \Delta_{7,X}^{ij} \sigma_{d,f}^2 && \sigma_{a,f}^2 \equiv 2Var[\alpha]; \sigma_{d,f}^2 \equiv Var[\delta]
\end{aligned}$$

Note, that this derivation is essentially identical to derivations for covariances of traits due to an autosomal QTL (LANGE 2003)- the kinship coefficient  $\phi_{ij}$  is replaced by its X-linked version  $\psi_{ij}$  and the condensed identity coefficients  $\Delta_c$  are similarly replaced by its X-linked version  $\Delta_{c,X}$ .

### 7.5.3 Male-Male relative pairs

We derive the covariance between two non-inbred males by conditioning on various identity states (Figure 7.2) and using the facts  $\sum_k p_k = 1$  and  $E[\beta] = \sum_k p_k \beta_k = 0$ . Since the trait values are standardized, we have

$$\begin{aligned}
Cov(y_i, y_j) &= Cov(g_k^i, g_u^j) && \text{trait values standardized} \\
&= E[g_k^i g_u^j] - E[g_k^i] E[g_u^j] && \text{general covariance formula} \\
&= E[g_k^i g_u^j] && E[g] = 0 \\
&= E[g_k^i g_u^j | 1 \text{ allele IBD}] P_1 + E[g_k^i g_u^j | 0 \text{ alleles IBD}] P_0 && \text{condition on IBD status} \\
&= P_1 \sum_k \beta_k^2 p_k + P_0 \sum_{k,u} \beta_k \beta_u p_k p_u && k = u \text{ if alleles IBD} \\
&= P_1 \sum_k \beta_k^2 p_k && E[\beta] = \sum_k p_k \beta_k = 0 \\
&= P_1 E[\beta^2] \\
&= P_1 Var[\beta] && E[\beta] = 0 \\
&= P_1 \sigma_{X,m}^2 && \sigma_{X,m}^2 \equiv Var[\beta] \\
&= \psi_{ij} \sigma_{X,m}^2
\end{aligned}$$

### 7.5.4 Female-Male relative pairs

We derive the covariance between non-inbred females and males by conditioning on various identity states (Figure 7.2) and using the facts  $\sum_k p_k = 1$ ,  $E[\alpha] = \sum_k p_k \alpha_k = 0$ ,  $\alpha_k = \sum_j p_j g_{kj}$ , and  $E[\beta] = \sum_k p_k \beta_k = 0$ . Since the trait values are standardized, we have

$$\begin{aligned}
Cov(y_i, y_j) &= Cov(g_{kl}^i, g_u^j) && \text{trait values standardized} \\
&= E[g_{kl}^i g_u^j] - E[g_{kl}^i] E[g_u^j] && \text{general covariance formula} \\
&= E[g_{kl}^i g_u^j] && E[g] = 0 \\
&= E[g_{kl}^i g_u^j | 1 \text{ allele IBD}] \Lambda_3 + E[g_{kl}^i g_u^j | 0 \text{ alleles IBD}] \Lambda_4 && \text{condition on IBD status} \\
&= \Lambda_3 \sum_{k,l} (\alpha_k + \alpha_l + \delta_{kl}) \beta_k p_k p_l + \Lambda_4 \sum_{k,l,u} (\alpha_k + \alpha_l + \delta_{kl}) \beta_u p_k p_l p_u && k = u \text{ if alleles IBD} \\
&= \Lambda_3 \sum_k \alpha_k \beta_k p_k \\
&= \sqrt{2} \frac{1}{2} \Lambda_3 \sqrt{2} \sum_k \alpha_k \beta_k p_k \\
&= \sqrt{2} \frac{1}{2} \Lambda_3 \sqrt{2} E[\alpha \beta] \\
&= \sqrt{2} \frac{1}{2} \Lambda_3 \sqrt{2} Cov(\alpha, \beta) && E[\alpha] = E[\beta] = 0 \\
&= \sqrt{2} \frac{1}{2} \Lambda_3 \sigma_{X, fm} && \sigma_{X, fm} \equiv \sqrt{2} Cov(\alpha, \beta) \\
&= \sqrt{2} \psi_{ij} \sigma_{X, fm}
\end{aligned}$$

The scaling of the covariance by  $\sqrt{2}$  is the one chosen by BULMER (1985), KENT *et al.* (2005), and PAN *et al.* (2007) and leads to the typical covariance constraint under an additive model  $|\sigma_{X, fm}| \leq \sqrt{\sigma_{X, f}^2 \sigma_{X, m}^2}$ . The constraint is a result of the Cauchy-Schwarz inequality:

$$\sigma_{X, fm}^2 = 2E[\alpha\beta]^2 \leq 2E[\alpha^2]E[\beta^2] = \sigma_{a, f}^2 \sigma_{X, m}^2$$

Therefore

$$|\sigma_{X, fm}| \leq \sqrt{\sigma_{a, f}^2 \sigma_{X, m}^2}$$

From the derivations in sections 7.5.2, 7.5.3, and 7.5.4 we have the covariances due to an X-linked gene for each type of relative pairs

$$Cov(y_i, y_j) = \begin{cases} \sigma_{a, f}^2 + \sigma_{d, f}^2 + \sigma_e^2 & i = j \text{ female} \\ 2\psi_{ij} \sigma_{a, f}^2 + \Delta_{7, X}^{ij} \sigma_{d, f}^2 & i \neq j \text{ female-female} \\ \sigma_{X, m}^2 + \sigma_e^2 & i = j \text{ male} \\ \psi_{ij} \sigma_{X, m}^2 & i \neq j \text{ male-male} \\ \sqrt{2} \psi_{ij} \sigma_{X, fm} & i \neq j \text{ female-male} \end{cases} \quad (7.5.2)$$

### 7.5.5 Marker loci

Now consider a candidate marker locus. If either the marker is at the trait locus or there is no recombination between them then we derive the covariance between two related individuals

conditional on number alleles shared IBD at the marker locus. Let  $IBD(i, j)$  denote number of alleles individuals  $i$  and  $j$  share IBD. From sections 7.5.2, 7.5.3, and 7.5.4 we can quite easily see

$$Cov(y_i, y_j | IBD(i, j) = \tau) = \begin{cases} (1_{\{\tau=1\}}/2 + 1_{\{\tau=2\}})\sigma_{a,f}^2 + 1_{\{\tau=2\}}\sigma_{d,f}^2 & \text{female-female} \\ 1_{\{\tau=1\}}\sigma_{X,m}^2 & \text{male-male} \\ 1_{\{\tau=1\}}\sqrt{2}\sigma_{X,fm} & \text{female-male} \end{cases} \quad (7.5.3)$$

Now if we add and subtract  $Cov(y_i, y_j)$  from the right hand side and the substitute the second part according to formula 7.5.2

$$Cov(y_i, y_j | IBD(i, j) = \tau) = \begin{cases} Cov(y_i, y_j) + (1_{\{\tau=1\}}/2 + 1_{\{\tau=2\}} - 2\psi_{ij})\sigma_{a,f}^2 + (1_{\{\tau=2\}} - \Delta_{i,X}^{ij})\sigma_{d,f}^2 & \text{female-female} \\ Cov(y_i, y_j) + (1_{\{\tau=1\}} - \psi_{ij})\sigma_{X,m}^2 & \text{male-male} \\ Cov(y_i, y_j) + (1_{\{\tau=1\}} - \sqrt{2}\psi_{ij})\sigma_{X,fm} & \text{female-male} \end{cases} \quad (7.5.4)$$

These expressions have the well-known desirable form which separates the segregation parameters in  $Cov(y_i, y_j)$  and the linkage parameters ( $\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}^2$ ) in the remaining terms (TANG and SIEGMUND 2001; WANG 2002). This means that the partial derivatives of  $Cov(y_i, y_j)$  w.r.t the linkage parameters equals zero. Furthermore the covariance matrix is linear with respect to the linkage parameters. We like to point out that this separation of segregation and linkage parameters is not quite right, under our model, since the variance components due to the trait contribute to  $Cov(y_i, y_j)$  as can be seen in equation 7.5.2. However allowing the partial derivative of  $Cov(y_i, y_j) = 0$  w.r.t to linkage parameters can be defended in at least three ways:

1.  $Cov(y_i, y_j)$  can be estimated from the observations  $y_i$  themselves without any assumption of linkage or inheritance. If the underlying model is good and the ascertainment and sample size are such that the estimators are consistent then this is reasonable to do. This is the reason behind similar derivations in the autosomal case (TANG and SIEGMUND 2001; WANG 2002)
2. Generally a number of QTLs, each with relatively small effect on the trait, contribute to the trait variance and so  $Cov(y_i, y_j)$  will be approximately constant w.r.t to the trait linkage parameters. Hence the derivative should be approximately zero. Note, that for simplicity we have, as TANG

and SIEGMUND (2001) did, assumed only one major QTL in their derivations. WANG (2002), on the other hand, modeled multiple QTLs.

3. The test statistic resulting from assuming that the partial derivative of  $Cov(y_i, y_j) = 0$  w.r.t to linkage parameters is intuitively sensible as then we are looking at how far the IBD sharing deviates from its expected values.

## 7.6 SCORE FUNCTIONS AND LIKELIHOOD

The genetic similarity among individuals is characterized by their IBD configurations. Let  $\gamma_n$  be the probability of the  $n$ th IBD configuration in a pedigree. Assume that conditional on the  $n$ th IBD configuration among the pedigree members, the phenotype  $y$  follows a multivariate normal distribution with mean vector  $\mu_s$  of sex-specific global means and variance-covariance matrix  $\Sigma_n$ ; the elements of  $\Sigma_n$  are the conditional covariances in formula 7.5.4. Additionally assume random (or unselected) sampling. The density is

$$\phi(y; \mu_s, \Sigma_n) \propto |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\}$$

and the log-likelihood of one pedigree is

$$\ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) \propto \ln\left[\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n\right]$$

In the next subsections we derive the scores of the log-likelihood for each parameter.



### 7.6.1 Score for $\sigma_{a,f}^2$

The partial derivative of  $\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n$  with respect to  $\sigma_{a,f}^2$  is

$$\begin{aligned}
& \frac{\partial}{\partial \sigma_{a,f}^2} \sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n \\
&= \sum_n \frac{\partial}{\partial \sigma_{a,f}^2} \left\{ |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} \right\} \gamma_n \\
&= \sum_n \left\{ -\frac{1}{2} |\Sigma_n|^{-3/2} \left( \frac{\partial}{\partial \sigma_{a,f}^2} |\Sigma_n| \right) \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} \right. \\
&\quad \left. - \frac{1}{2} |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} \left( \frac{\partial}{\partial \sigma_{a,f}^2} (y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s) \right) \right\} \gamma_n \quad \Sigma_n \text{ include } \sigma_{a,f}^2 \\
&= \sum_n \left\{ -\frac{1}{2} |\Sigma_n|^{-3/2} |\Sigma_n| \text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} \right. \quad (*) \\
&\quad \left. - \frac{1}{2} |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} (y - \mu_s)^T (-\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \gamma_n \quad (**) \\
&= \sum_n \left\{ \frac{1}{2} |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\} \left\{ \right. \quad (***) \\
&\quad \left. - \text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) + (y - \mu_s)^T (\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \right\} \gamma_n \\
&= \sum_n \left\{ \frac{1}{2} \phi(y; \mu_s, \Sigma_n) \left\{ - \text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) + (y - \mu_s)^T (\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \right\} \gamma_n \quad (***)
\end{aligned}$$

In the above derivation we used

(\*) Here we used  $\frac{\partial |A|}{\partial x} = |A| \text{tr}(A^{-1} \frac{\partial A}{\partial x})$  for matrix  $A = \Sigma_n$  and scalar  $x = \sigma_{a,f}^2$

(\*\*) Here we used  $\frac{\partial}{\partial x} v^T A^{-1} v = v^T \frac{\partial A^{-1}}{\partial x} v$  if  $v$  independent of  $x$ , and  $\frac{\partial A^{-1}}{\partial x} = -A^{-1} \frac{\partial A}{\partial x} A^{-1}$  for matrix

$A = \Sigma_n$ , vector  $v = y - \mu_s$ , and scalar  $x = \sigma_{a,f}^2$

(\*\*\*) Here we factorize  $\frac{1}{2} |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\}$  out

(\*\*\*\*) Here we have  $\phi(y; \mu_s, \Sigma_n) \equiv \frac{1}{2} |\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu_s)^T \Sigma_n^{-1} (y - \mu_s)\right\}$

Then we can derive the partial derivative of the log-likelihood,  $\ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm})$

$\propto \ln[\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n]$ , w.r.t  $\sigma_{a,f}^2$

$$\begin{aligned}
\frac{\partial}{\partial \sigma_{a,f}^2} \ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) &= \frac{1}{\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n} \frac{\partial}{\partial \sigma_{a,f}^2} \sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n \\
&= \frac{\sum_n \left\{ \frac{1}{2} \phi(y; \mu_s, \Sigma_n) \left\{ - \text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) + (y - \mu_s)^T (\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \right\} \gamma_n}{\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n} \\
&= \frac{1}{2} \frac{\sum_n \left\{ \phi(y; \mu_s, \Sigma_n) \left\{ - \text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) + (y - \mu_s)^T (\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \right\} \gamma_n}{\sum_n \phi(y; \mu_s, \Sigma_n) \gamma_n}
\end{aligned}$$

(7.6.1)

Now let  $\theta = (\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm})$ . If we evaluate each partial derivative at  $\theta = 0$  then we note that all the terms  $\phi(y; \mu_s, \Sigma_n) \gamma_n$  are equal  $\Sigma_n|_{\theta=0} \equiv \Sigma(0)$ . Therefore we factorize  $\phi(y; \mu_s, \Sigma_n) \gamma_n$  out of the sums in the numerator and denominator in equation 7.6.1 and use  $\sum_n \gamma_n = 1$  to get

$$\begin{aligned}
b_{a,f} &\equiv \frac{\partial}{\partial \sigma_{a,f}^2} \ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) |_{\boldsymbol{\theta}=0} \\
&= \frac{1}{2} \sum_n \left\{ -\text{tr}(\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2}) + (y - \mu_s)^T (\Sigma_n^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \Sigma_n^{-1}) (y - \mu_s) \right\} \gamma_n |_{\boldsymbol{\theta}=0} \\
&= \frac{1}{2} \sum_n \left\{ -\text{tr}(\Sigma(0)^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0}) + (y - \mu_s)^T (\Sigma(0)^{-1} \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \Sigma(0)^{-1}) (y - \mu_s) \right\} \gamma_n \\
&= \frac{1}{2} \left\{ -\text{tr} \left\{ \Sigma(0)^{-1} \left( \sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n \right) \right\} + (y - \mu_s)^T (\Sigma(0)^{-1} \left( \sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n \right) \Sigma(0)^{-1}) (y - \mu_s) \right\}
\end{aligned}$$

Now define matrix  $\Pi$ , a symmetric matrix whose elements  $\pi_{i,j} = \pi_{ij}^{(1)}/2 + \pi_{ij}^{(2)}$ , where  $\pi_{ij}^{(k)}$  is the probability of individuals  $i$  and  $j$  sharing  $k$  alleles IBD at the marker locus. Note, that  $\pi_{i,j}$  can be interpreted as the averaged number of alleles shared IBD. We then split the matrix up according to the sexes of the relative pairs:  $\Pi = \Pi^f + \Pi^m + \Pi^{fm}$ , where  $\Pi^f$  are the same as  $\Pi$  except that the elements corresponding to non-female-female pairs are zero. The other matrices are similarly defined. We also define matrix  $\Omega$ , a symmetric matrix whose diagonal elements ( $\omega_{i,i}$ ) are all zero and the  $ij^{\text{th}}$  off-diagonal element is  $\omega_{i,j} = \pi_{i,j}$ . We then split the matrix up according to the sexes of the relative pairs:  $\Omega = \Omega^f + \Omega^m + \Omega^{fm}$ , where  $\Omega^f$  are the same as  $\Omega$  except that the elements corresponding to non-female-female pairs are zero. The other matrices are similarly defined.

Now we need to derive the form of  $\sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n$ . We assume that the partial derivatives of  $Cov(y_i, y_j)$  equals zero. Then we get

$$\frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} = \begin{cases} 0 & \text{if } i = j \text{ female (diagonal element)} \\ 1_{\{\tau=1\}}/2 + 1_{\{\tau=2\}} - 2\psi_{ij} & \text{if } i \neq j \text{ female-female (off-diagonal element)} \\ 0 & \text{o.w.} \end{cases}$$

and since  $\sum \gamma_n = 1$ ,  $\pi_{ij}^{(\tau)} = \sum_{IBD(i,j)=\tau} \gamma_n$ ,  $\pi_{ij} = \pi_{ij}^{(1)}/2 + \pi_{ij}^{(2)}$ , and  $E[\pi_{ij}] = 2\psi_{ij}$  we have

$$\sum \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} \gamma_n = \begin{cases} 0 & \text{if } i = j \text{ female (diagonal element)} \\ \sum (1_{\{\tau=1\}}/2 + 1_{\{\tau=2\}} - 2\psi_{ij}) \gamma_n \\ = \pi_{ij}^{(1)}/2 + \pi_{ij}^{(2)} - 2\psi_{ij} \\ = \pi_{ij} - E[\pi_{ij}] & \text{if } i \neq j \text{ female-female (off-diagonal element)} \\ 0 & \text{o.w.} \end{cases}$$

So  $\sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n = \Omega^f - E[\Omega^f]$

Since  $y \propto \phi(y; \mu_s, \Sigma_n)$  then under the null hypothesis of no linkage  $y \propto \phi(y; \mu_s, \Sigma(0))$  and  $w \equiv \Sigma(0)^{-1}(y - \mu_s) \propto \phi(w; 0, \Sigma(0)^{-1})$ . Additionally  $E[w^T w] = \Sigma(0)^{-1}$ . Therefore for any matrix  $M$

$$\begin{aligned} \text{tr}(\Sigma(0)^{-1}M) &= \sum_{i,j} (\Sigma(0)^{-1})_{i,j} m_{j,i} \\ &= \sum_{i,j} (E[w^T w])_{i,j} m_{j,i} \\ &= E[w^T M w | M] \end{aligned}$$

Then we may write the score for  $\sigma_{a,f}^2$  under the null as

$$\begin{aligned} b_{a,f} &= \frac{1}{2} \left\{ -\text{tr} \left\{ \Sigma(0)^{-1} \left( \sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n \right) \right\} + (y - \mu_s)^T \left( -\Sigma(0)^{-1} \left( \sum_n \frac{\partial \Sigma_n}{\partial \sigma_{a,f}^2} |_{\boldsymbol{\theta}=0} \gamma_n \right) \Sigma(0)^{-1} \right) (y - \mu_s) \right\} \\ &= \frac{1}{2} \left\{ -\text{tr}(\Sigma(0)^{-1}(\Omega^f - E[\Omega^f])) + w^T (\Omega^f - E[\Omega^f]) w \right\} \\ &= \frac{1}{2} \left\{ w^T (\Omega^f - E[\Omega^f]) w - E[w^T (\Omega^f - E[\Omega^f]) w | \Omega^f] \right\} \\ &= \frac{1}{2} w^T (\Omega^f - E[\Omega^f]) w \\ &= \frac{1}{2} \sum_{\substack{i,j \\ i,j \text{ females}}} (\omega_{i,j} - E[\omega_{i,j}])(w_i w_j - E[w_i w_j]) \\ &= \sum_{\substack{i,j \\ i > j \text{ females}}} (\pi_{i,j} - E[\pi_{i,j}])(w_i w_j - E[w_i w_j]) \end{aligned} \tag{7.6.2}$$

where we take the last step by using  $\omega_{i,i} = 0$  for all  $i$  and  $\omega_{i,j} = \pi_{i,j}$  for all  $i \neq j$ , as well as the fact that the  $\Pi$  matrices are symmetric ( $\pi_{i,j} = \pi_{j,i}$ ).

### 7.6.2 Score for $\sigma_{d,f}^2$

In a similar way as in subsection 7.6.1 we get

$$\sum \frac{\partial \Sigma_n}{\partial \sigma_{d,f}^2} \gamma_n = \begin{cases} 0 & \text{if } i = j \text{ female (diagonal element)} \\ \pi_{ij}^{(2)} - E[\pi_{ij}^{(2)}] & \text{if } i \neq j \text{ female-female (off-diagonal element)} \\ 0 & \text{o.w.} \end{cases}$$

Therefore

$$\begin{aligned} b_{d,f} &\equiv \frac{\partial}{\partial \sigma_{d,f}^2} \ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) \\ &= \sum_{\substack{i,j \\ i>j \text{ females}}} (\pi_{i,j}^{(2)} - E[\pi_{i,j}^{(2)}]) (w_i w_j - E[w_i w_j]) \end{aligned} \quad (7.6.3)$$

### 7.6.3 Score for $\sigma_{X,m}^2$

In a similar way as in subsection 7.6.1 we get

$$\sum \frac{\partial \Sigma_n}{\partial \sigma_{X,m}^2} \gamma_n = \begin{cases} 0 & \text{if } i = j \text{ male (diagonal element)} \\ \pi_{ij} - E[\pi_{ij}] & \text{if } i \neq j \text{ male-male (off-diagonal element)} \\ 0 & \text{o.w.} \end{cases}$$

Therefore

$$\begin{aligned} b_{X,m} &\equiv \frac{\partial}{\partial \sigma_{X,m}^2} \ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) \\ &= \sum_{\substack{i,j \\ i>j \text{ males}}} (\pi_{i,j} - E[\pi_{i,j}]) (w_i w_j - E[w_i w_j]) \end{aligned} \quad (7.6.4)$$

### 7.6.4 Score for $\sigma_{X,fm}$

In a similar way as in subsection 7.6.1 we get

$$\sum \frac{\partial \Sigma_n}{\partial \sigma_{X,fm}^2} \gamma_n = \begin{cases} \pi_{ij} - E[\pi_{ij}] & \text{female-male} \\ 0 & \text{o.w.} \end{cases}$$

Therefore

$$\begin{aligned}
b_{X,fm} &= \frac{\partial}{\partial \sigma_{X,fm}} \ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) \\
&= \sum_{\substack{i,j \\ i>j \text{ female,male}}} (\pi_{i,j} - E[\pi_{i,j}])(w_i w_j - E[w_i w_j])
\end{aligned} \tag{7.6.5}$$

## 7.7 SUMMARY OF SCORES

Remember that  $w = \Sigma(0)^{-1}(y - \mu_s)$  and that  $\Pi^{(2)}$  is the probability of two females sharing two alleles IBD and the other  $\Pi$ 's are the proportion of alleles shared IBD between relatives. Now we can write the scores (see equations 7.6.2, 7.6.3, 7.6.4, and 7.6.5) in matrix format

$$\begin{aligned}
b_{a,f} &= \text{vec}(w^T w - E[w^T w])^T \text{vec}(\Pi^f - E[\Pi^f]) \\
b_{d,f} &= \text{vec}(w^T w - E[w^T w])^T \text{vec}(\Pi^{(2)} - E[\Pi^{(2)}]) \\
b_{X,m} &= \text{vec}(w^T w - E[w^T w])^T \text{vec}(\Pi^m - E[\Pi^m]) \\
b_{X,fm} &= \text{vec}(w^T w - E[w^T w])^T \text{vec}(\Pi^{fm} - E[\Pi^{fm}])
\end{aligned}$$

where we let  $\text{vec}$  be an operator that vectorizes the lower diagonal elements in column-wise order.

In the derivations above we used a sample size of one family. If we had larger data set of  $K$  families our score would be a sum of the scores for each family  $k$

$$\begin{aligned}
b_{a,f} &= \sum_k b_{a,f,k} = \sum_k \text{vec}(w_k^T w_k - E[w_k^T w_k])^T \text{vec}(\Pi_k^f - E[\Pi_k^f]) \\
b_{d,f} &= \sum_k b_{d,f,k} = \sum_k \text{vec}(w_k^T w_k - E[w_k^T w_k])^T \text{vec}(\Pi_k^{(2)} - E[\Pi_k^{(2)}]) \\
b_{X,m} &= \sum_k b_{X,m,k} = \sum_k \text{vec}(w_k^T w_k - E[w_k^T w_k])^T \text{vec}(\Pi_k^m - E[\Pi_k^m]) \\
b_{X,fm} &= \sum_k b_{X,fm,k} = \sum_k \text{vec}(w_k^T w_k - E[w_k^T w_k])^T \text{vec}(\Pi_k^{fm} - E[\Pi_k^{fm}])
\end{aligned}$$

## 7.8 VARIANCE OF THE SCORES

We estimate the Fisher information from our data set of  $K$  pedigrees as

$$I = \begin{bmatrix} \text{Var}(b_{a,f}) & \text{Cov}(b_{a,f}, b_{d,f}) & \text{Cov}(b_{a,f}, b_{X,m}) & \text{Cov}(b_{a,f}, b_{X,fm}) \\ \text{Cov}(b_{a,f}, b_{d,f}) & \text{Var}(b_{d,f}) & \text{Cov}(b_{d,f}, b_{X,m}) & \text{Cov}(b_{d,f}, b_{X,fm}) \\ \text{Cov}(b_{a,f}, b_{X,m}) & \text{Cov}(b_{d,f}, b_{X,m}) & \text{Var}(b_{X,m}) & \text{Cov}(b_{X,m}, b_{X,fm}) \\ \text{Cov}(b_{a,f}, b_{X,fm}) & \text{Cov}(b_{d,f}, b_{X,fm}) & \text{Cov}(b_{X,m}, b_{X,fm}) & \text{Var}(b_{X,fm}) \end{bmatrix}$$

where conditional on the IBD sharing, the variances and the covariances have the following form

$$\begin{aligned} \text{Var}(b_{a,f}) &= \sum_k \text{vec}(\Pi_k^f - E[\Pi_k^f])^T \text{var}(\text{vec}(w_k^T w_k - E[w_k^T w_k])^T) \text{vec}(\Pi_k^f - E[\Pi_k^f]) \\ &= \sum_k \text{vec}(\Pi_k^f - E[\Pi_k^f])^T \text{var}(\text{vec}(w_k^T w) ) \text{vec}(\Pi_k^f - E[\Pi_k^f]) \\ \text{Cov}(b_{a,f}, b_{d,f}) &= \sum_k \text{vec}(\Pi_k^f - E[\Pi_k^f])^T \text{var}(\text{vec}(w_k^T w_k - E[w_k^T w_k])^T) \text{vec}(\Pi_k^{(2)} - E[\Pi_k^{(2)}]) \\ &= \sum_k \text{vec}(\Pi_k^f - E[\Pi_k^f])^T \text{var}(\text{vec}(w_k^T w_k) ) \text{vec}(\Pi_k^{(2)} - E[\Pi_k^{(2)}]) \end{aligned}$$

and similarly for the remaining scores.

We note that the covariances between scores for parameters corresponding to different types of sex-sex pairs are always zero, that is  $\text{Cov}(b_{a,f}, b_{X,m}) = \text{Cov}(b_{a,f}, b_{X,fm}) = \text{Cov}(b_{d,f}, b_{X,m}) = \text{Cov}(b_{d,f}, b_{X,fm}) = \text{Cov}(b_{X,m}, b_{X,fm}) = 0$  and so the Fisher information matrix is much simplified:

$$I = \begin{bmatrix} \text{Var}(b_{a,f}) & \text{Cov}(b_{a,f}, b_{d,f}) & 0 & 0 \\ \text{Cov}(b_{a,f}, b_{d,f}) & \text{Var}(b_{d,f}) & 0 & 0 \\ 0 & 0 & \text{Var}(b_{X,m}) & 0 \\ 0 & 0 & 0 & \text{Var}(b_{X,fm}) \end{bmatrix}$$

The variances may be estimated empirically. Generally, in the autosomal case, the empirical variance is preferred over theoretical variance ([BHATTACHARJEE \*et al.\* 2008](#)) and I suspect this is also the case in the X-linked case, though extensive simulations studies are required to confirm that.

## 7.9 SCORE STATISTICS AND ASYMPTOTIC DISTRIBUTIONS

Depending on our hypothesis of interest, we can define a number of score statistics. One obvious choice is a ‘global’ null hypothesis of no female or male linkage

$$H_0 : \sigma_{a,f}^2 = \sigma_{d,f}^2 = \sigma_{X,m}^2 = \sigma_{X,fm} = 0$$

versus

$$H_A : \sigma_{a,f}^2 > 0 \text{ or } \sigma_{d,f}^2 > 0 \text{ or } \sigma_{X,m}^2 > 0 \text{ or } |\sigma_{X,fm}| > 0$$

under the constraint  $|\sigma_{X,fm}| \leq \sqrt{\sigma_{a,f}^2 \sigma_{X,m}^2}$

Due to  $|\sigma_{X,fm}| \leq \sqrt{\sigma_{a,f}^2 \sigma_{X,m}^2}$  all four parameters are tested at the boundary of the parameter space; note that the point  $(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) = (\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, 0)$  is a boundary point only in the special case when  $\sigma_{a,f}^2 = \sigma_{d,f}^2 = \sigma_{X,m}^2 = 0$ . Deriving the statistic in a closed form and its distribution for testing this general hypothesis in general pedigrees is quite challenging (if not impossible). We follow similar procedure as [WANG \(2002\)](#), however the constraint  $|\sigma_{X,fm}| \leq \sqrt{\sigma_{a,f}^2 \sigma_{X,m}^2}$  causes problems. Briefly, the Wang procedure forms a standard likelihood ratio test which is then written as a score tests by applying theorem 16.7 of [VAN DER VAAR \(1998\)](#). This theorem states that under normality or the quadratic approximation of the log-likelihood the likelihood ratio and scores tests are equivalent.

We write the Fisher information matrix per pedigree as

$$I_0 \equiv \begin{bmatrix} I_{11} & I_{12} & I_{13} & I_{14} \\ I_{21} & I_{22} & I_{23} & I_{24} \\ I_{31} & I_{32} & I_{33} & I_{34} \\ I_{41} & I_{42} & I_{43} & I_{44} \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & 0 & 0 \\ I_{12} & I_{22} & 0 & 0 \\ 0 & 0 & I_{33} & 0 \\ 0 & 0 & 0 & I_{44} \end{bmatrix}$$

$$= \lim_{K \rightarrow \infty} K^{-1} \begin{bmatrix} \text{Var}(b_{a,f}) & \text{Cov}(b_{a,f}, b_{d,f}) & 0 & 0 \\ \text{Cov}(b_{a,f}, b_{d,f}) & \text{Var}(b_{d,f}) & 0 & 0 \\ 0 & 0 & \text{Var}(b_{X,m}) & 0 \\ 0 & 0 & 0 & \text{Var}(b_{X,fm}) \end{bmatrix}$$

From asymptotic theory ([PAWITAN 2001](#)) we have for a random vector  $\mathbf{a}$

$$K^{-1/2} S(\boldsymbol{\theta}) = K^{-1/2} (b_{a,f}, b_{d,f}, b_{X,m}, b_{X,fm})^T \xrightarrow{d} \mathbf{a} \sim N(\mathbf{0}, I_0)$$

where  $\boldsymbol{\theta} = (\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm})^T$  and we let  $\Theta_A = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in H_A\}$  be the set of parameters that correspond to the alternative hypothesis. Then we have the likelihood ratio statistic

$$\begin{aligned}\Lambda_K &= \sup_{\boldsymbol{\theta} \in \Theta_A} 2[\ell(\sigma_{a,f}^2, \sigma_{d,f}^2, \sigma_{X,m}^2, \sigma_{X,fm}) - \ell(0, 0, 0, 0)] \\ &\xrightarrow{d} 2 \sup_{\boldsymbol{\theta} \in \Theta_A} (\boldsymbol{\theta}^T \mathbf{a} - \frac{1}{2} \boldsymbol{\theta}^T I_0 \boldsymbol{\theta})\end{aligned}$$

Now we take the derivative of  $\boldsymbol{\theta}^T \mathbf{a} - \frac{1}{2} \boldsymbol{\theta}^T I_0 \boldsymbol{\theta}$  w.r.t  $\boldsymbol{\theta}$  and get

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \{ \boldsymbol{\theta}^T \mathbf{a} - \frac{1}{2} \boldsymbol{\theta}^T I_0 \boldsymbol{\theta} \} \\ &= \mathbf{a} - \frac{1}{2} (I_0 + I_0^T) \boldsymbol{\theta} \\ &= \mathbf{a} - I_0 \boldsymbol{\theta}\end{aligned}$$

If we set  $\mathbf{a} - I_0 \boldsymbol{\theta} = 0$  and solve for  $\boldsymbol{\theta}$  then we get the unrestricted solution  $\boldsymbol{\theta}^* = I_0^{-1} \mathbf{a}$ , which may return parameter estimates that are out of bounds, hence we refer to it as the unrestricted solution and the formula of the corresponding statistic as the unrestricted statistic, which is

$$\begin{aligned}\Lambda &= 2 \sup_{\boldsymbol{\theta} \in \Theta_1} (\boldsymbol{\theta}^T \mathbf{a} - \frac{1}{2} \boldsymbol{\theta}^T I_0 \boldsymbol{\theta}) \\ &= 2(I_0^{-1} \mathbf{a})^T \mathbf{a} - (I_0^{-1} \mathbf{a})^T I_0 I_0^{-1} \mathbf{a} \\ &= 2\mathbf{a}^T (I_0^{-1})^T \mathbf{a} - \mathbf{a}^T (I_0^{-1})^T \mathbf{a} \\ &= \mathbf{a}^T I_0^{-1} \mathbf{a}\end{aligned}$$

where

$$I_0^{-1} = \begin{bmatrix} \frac{I_{22}}{I_{11}I_{22}-I_{12}^2} & -\frac{I_{12}}{I_{11}I_{22}-I_{12}^2} & 0 & 0 \\ -\frac{I_{12}}{I_{11}I_{22}-I_{12}^2} & \frac{I_{11}}{I_{11}I_{22}-I_{12}^2} & 0 & 0 \\ 0 & 0 & \frac{1}{I_{33}} & 0 \\ 0 & 0 & 0 & \frac{1}{I_{44}} \end{bmatrix}$$

Now we need to derive the composite statistic, that is the form of the statistic in all situations, as function of the scores. To do that we start by writing the unrestricted solution  $\boldsymbol{\theta}^* = I_0^{-1} \mathbf{a}$  explicitly as



$$I_0^{-1} \mathbf{a} = \begin{bmatrix} a_1 \frac{I_{22}}{I_{11}I_{22}-I_{12}^2} - a_2 \frac{I_{12}}{I_{11}I_{22}-I_{12}^2} \\ -a_1 \frac{I_{12}}{I_{11}I_{22}-I_{12}^2} + a_2 \frac{I_{22}}{I_{11}I_{22}-I_{12}^2} \\ a_3 \frac{1}{I_{33}} \\ a_4 \frac{1}{I_{44}} \end{bmatrix} = \begin{bmatrix} x_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - x_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}} \\ x_2 \frac{1}{\sqrt{I_{22}(1-r^2)}} - x_1 \frac{I_{12}}{I_{22}\sqrt{I_{11}(1-r^2)}} \\ x_3 \frac{1}{\sqrt{I_{33}}} \\ x_4 \frac{1}{\sqrt{I_{44}}} \end{bmatrix} \equiv \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \theta_3^* \\ \theta_4^* \end{bmatrix}$$

where  $x_i = a_i/\sqrt{I_{ii}}$  and  $r = I_{12}/\sqrt{I_{11}I_{22}}$ .

If all conditions hold and no parameter estimates are out of bounds then the statistic has the form  $\mathbf{a}^T I_0^{-1} \mathbf{a}$ , which we may formulate as

$$\Lambda = \mathbf{a}^T I_0^{-1} \mathbf{a} = \frac{x_1^2 - 2rx_1x_2 + x_2^2}{1-r^2} + x_3^2 + x_4^2 \quad (7.9.1)$$

if all the following conditions hold

1.  $\theta_1^* = x_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - x_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}} \geq 0$  or  $x_1 \geq x_2 \frac{I_{12}}{\sqrt{I_{11}I_{22}}} = x_2 r$
2.  $\theta_2^* = x_2 \frac{1}{\sqrt{I_{22}(1-r^2)}} - x_1 \frac{I_{12}}{I_{22}\sqrt{I_{11}(1-r^2)}} \geq 0$  or  $x_2 \geq x_1 r$
3.  $\theta_3^* = x_3 \frac{1}{\sqrt{I_{33}}} \geq 0$  or  $x_3 \geq 0$
4.  $|x_4 \frac{1}{\sqrt{I_{44}}}| \leq \sqrt{x_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - x_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}}} x_3 \frac{1}{\sqrt{I_{33}}}$

### 7.9.1 Derivation of the statistic when parameter estimates are out of bounds

Now look at how the statistic given in equation 7.9.1 looks when the each condition fails, that is when the unrestricted formula returns parameter estimates that are out of bounds.

**7.9.1.1 Condition  $\sigma_{a,f}^2 \geq 0$  fails** The first condition  $\sigma_{a,f}^2 \geq 0$  fails when  $\theta_1^* = x_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - x_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}} < 0$  then  $\theta_1^* = 0$  and  $\theta_4^* = 0$  due to condition  $|\theta_4| \leq \sqrt{\theta_1 \theta_3}$ . Then we optimize the likelihood ratio statistic  $\Lambda_{2,3} = 2 \sup_{\theta \in \Theta_{1,3}} (\theta^T \mathbf{a} - \frac{1}{2} \theta^T I_0 \theta) = 2 \sup_{\theta \in \Theta_1} (\theta_{2,3}^T \mathbf{a}_{2,3} - \frac{1}{2} \theta_{2,3}^T I_{0,2,3} \theta_{2,3}) = \mathbf{a}_{2,3}^T I_{0,2,3}^{-1} \mathbf{a}_{2,3} = \frac{a_2^2}{I_{22}} + \frac{a_3^2}{I_{33}} = x_2^2 + x_3^2$  if  $x_2 \geq 0$  and  $x_3 \geq 0$ , where the subscripts 2, 3 indicate sub-matrix (with rows and columns 2 and 3), vector  $(\mathbf{a}_{2,3} = (a_2, a_3)^T)$ , and set  $(\Theta_{2,3} = \{\theta : \theta_1 = \theta_4 = 0\})$ .

**7.9.1.2 Condition  $\sigma_{X,m}^2 \geq 0$  fails** The third condition  $\sigma_{X,m}^2 \geq 0$  fails when  $\theta_3^* < 0$  then  $\theta_3^* = 0$  and  $\theta_4^* = 0$  due to condition  $|\theta_4| \leq \sqrt{\theta_1 \theta_3}$ . We get the optimized likelihood ratio statistic  $\Lambda_{2,4} = x_2^2 + x_4^2$  if  $x_2 \geq 0$  and  $x_4 \geq 0$ .

**7.9.1.3 Both  $\sigma_{a,f}^2 \geq 0$  and  $\sigma_{X,m}^2 \geq 0$  fail** If both  $\theta_1^* < 0$  and  $\theta_3^* < 0$  then  $\theta_4^* = 0$  due to condition  $|\theta_4| \leq \sqrt{\theta_1\theta_3}$ . We get the optimized likelihood ratio statistic  $\Lambda_2 = x_2^2$  if  $x_2 \geq 0$ .

**7.9.1.4 Condition  $\sigma_{d,f}^2 \geq 0$  fails** If  $\theta_2^* < 0$  then  $\theta_2^* = 0$  and so  $\Lambda_{1,3,4} = x_1^2 + x_3^2 + x_4^2$  if  $x_1 \geq 0$ ,  $x_3 \geq 0$ ,  $x_4 \geq 0$ , and  $|\frac{x_4}{\sqrt{I_{44}}}| \leq \sqrt{\frac{x_1}{I_{11}} \frac{x_3}{I_{33}}}$

**7.9.1.5 If both  $\sigma_{a,f}^2 \geq 0$  and  $\sigma_{X,m}^2 \geq 0$  hold but** If  $\theta_1^* \geq 0, \theta_3^* \geq 0$ , and  $|\theta_4^*| > \sqrt{\theta_1^*\theta_3^*}$  then we set  $\theta_4^{*2} = \theta_1^*\theta_3^*$  and look at

$$\begin{aligned} \theta^T \mathbf{a} - \frac{1}{2} \theta^T I_0 \theta &= \theta_1 a_1 + \theta_2 a_2 + \theta_3 a_3 + \theta_4 a_4 - \frac{1}{2} (\theta_1^2 I_{11} + 2\theta_1 \theta_2 I_{12} + \theta_2^2 I_{22} + \theta_3^2 I_{33} + \theta_4^2 I_{44}) \\ &= \theta_1 a_1 + \theta_2 a_2 + \theta_3 a_3 + \sqrt{\theta_1 \theta_3} a_4 - \frac{1}{2} (\theta_1^2 I_{11} + 2\theta_1 \theta_2 I_{12} + \theta_2^2 I_{22} + \theta_3^2 I_{33} + \theta_1 \theta_3 I_{44}) \end{aligned}$$

which needs to be maximized. By taking the partial derivatives w.r.t to each parameter, we quickly see that the solutions, as function of only the  $x_i$ , and  $I_{ij}$ , will be quite challenging to write out in a simple formula. However, the solution exists (and can be found with the help of Mathematica) and so we can get a legitimate score statistic when the fourth condition fails. However, the closed form solution is too long and cumbersome to write down and not usable in practice. In appendix E we show which equations need to be solved numerically and we give the Hessian matrices necessary the solutions corresponding to the maxima. In section 7.9.3 we show how we can arrive at a statistic described by a simple formula if we assume that the allelic effects are the same in females and males, i.e.  $\alpha_k = \beta_k$  for all alleles  $k$ .

Let  $Z_1 = b_{a,f}/\sqrt{Var(b_{a,f})}$ ,  $Z_2 = b_{d,f}/\sqrt{Var(b_{d,f})}$ ,  $Z_3 = b_{X,m}/\sqrt{Var(b_{X,m})}$ , and  $Z_4 = b_{X,fm}/\sqrt{Var(b_{X,fm})}$ , i.e. the scores are standardized by their respective variances. We replace the  $x_i$  with  $Z_i$  in equation 7.9.1 and the derivations in sections 7.9.1.1–7.9.1.5, and have the composite

score statistic

$$\Lambda = \begin{cases} \frac{Z_1^2 - 2rZ_1Z_2 + Z_2^2}{1-r^2} + Z_3^2 + Z_4^2 & \text{if } Z_1 \geq Z_2r, Z_2 \geq Z_1r, Z_3 \geq 0, \\ & |Z_4 \frac{1}{\sqrt{I_{44}}}| \leq \sqrt{(Z_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - Z_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}})Z_3 \frac{1}{\sqrt{I_{33}}}} \\ Z_1^2 + Z_3^2 + Z_4^2 & \text{if } Z_2 < Z_1r, Z_1 \geq 0, Z_3 \geq 0, \\ & |Z_4 \frac{1}{\sqrt{I_{44}}}| \leq \sqrt{Z_1 \frac{1}{I_{11}} Z_3 \frac{1}{\sqrt{I_{33}}}} \\ Z_2^2 + Z_3^2 & \text{if } Z_1 < Z_2r, Z_2 \geq 0, Z_3 \geq 0, \\ & |Z_4 \frac{1}{\sqrt{I_{44}}}| \leq \sqrt{(Z_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - Z_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}})Z_3 \frac{1}{\sqrt{I_{33}}}} \\ Z_2^2 + Z_4^2 & \text{if } Z_2 \geq 0, Z_4 \geq 0 \\ Z_2^2 & \text{if } Z_2 \geq 0 \\ W & \text{if } |Z_4 \frac{1}{\sqrt{I_{44}}}| > \sqrt{(Z_1 \frac{1}{\sqrt{I_{11}(1-r^2)}} - Z_2 \frac{I_{12}}{I_{11}\sqrt{I_{22}(1-r^2)}})Z_3 \frac{1}{\sqrt{I_{33}}}} \\ 0 & \text{o.w.} \end{cases} \quad (7.9.2)$$

where  $W$  is evaluated by numerical methods (see appendix E).

### 7.9.2 Distribution of the statistic

In developing the distribution of the test statistic from variance component models for the null hypothesis of no female or male linkage assuming that the female dominance is zero (i.e.  $H_0 : \sigma_{a,f}^2 = \sigma^2X, m = \sigma_{X,fm} = 0$ ) EKSTRØM (2004) notes that the variance terms are being tested at the boundary of their parameter space ( $\sigma_{a,f}^2 \geq 0, \sigma_{X,m}^2 \geq 0$ ) while the covariance ( $\sigma_{X,fm}$ ) is being tested inside its parameter space. However, due to the condition  $|\sigma_{X,fm}| \leq \sqrt{\sigma_{a,f}^2 \sigma_{X,m}^2}$ ,  $\sigma_{X,fm}$  is also being tested at the boundary of its parameters space. This suggests that EKSTRØM (2004) is using the wrong case from Self and Liang to arrive at the distribution of the test statistic. However, even if EKSTRØM (2004) had noticed this he would not be able to use the theoretical results of SELF and LIANG (2007) to derive the distribution. SELF and LIANG (2007) showed that the asymptotic distribution of the likelihood ratio test in multi-parameter cases is a mixture of  $\chi^2$  distributions, but in all the cases considered the parameter space could be described, in the case of  $p$  parameters, as a simple product space:  $\Omega = \Omega_1 \times \dots \times \Omega_p$ , where the  $\Omega_i$ 's are either closed, half-open, or open intervals in  $\mathbb{R}$ . Our parameter space cannot be approximated by such a simple space and therefore it is most appropriate to establish the distribution of the test statistic using simulations under the null hypothesis.

### 7.9.3 Simpler model: Assuming equal allelic effects in both sexes

If instead of allowing the allelic effects in the sexes to be different (i.e.  $\alpha_k$  in females and  $\beta_k$  in males) we can assume that they are equal, which is perfectly reasonable to do. If we go through the derivations in sections 7.5.2, 7.5.3, and 7.5.4 assuming  $\alpha_k = \beta_k$  for all alleles  $k$  we see that  $\sigma^2 \equiv \sigma_{a,f}^2 = \sigma_{X,m}^2 = \sigma_{X,fm}$  and we can write

$$Cov(y_i, y_j) = \begin{cases} \sigma^2 + \sigma_{d,f}^2 + \sigma_e^2 & i = j \text{ female} \\ 2\psi_{ij}\sigma^2 + \Delta_{7,X}^{ij}\sigma_{d,f}^2 & i \neq j \text{ female-female} \\ \sigma^2 + \sigma_e^2 & i = j \text{ male} \\ \psi_{ij}\sigma^2 & i \neq j \text{ male-male} \\ \sqrt{2}\psi_{ij}\sigma^2 & i \neq j \text{ female-male} \end{cases}$$

and

$$\begin{aligned} & Cov(y_i, y_j | IBD(i, j) = \tau) \\ &= \begin{cases} Cov(y_i, y_j) + (1_{\{\tau=1\}}/2 + 1_{\{\tau=2\}} - 2\psi_{ij})\sigma^2 + (1_{\{\tau=2\}} - \Delta_{7,X}^{ij})\sigma_{d,f}^2 & \text{female-female} \\ Cov(y_i, y_j) + (1_{\{\tau=1\}} - \psi_{ij})\sigma^2 & \text{male-male} \\ Cov(y_i, y_j) + (1_{\{\tau=1\}} - \sqrt{(2)}\psi_{ij})\sigma^2 & \text{female-male} \end{cases} \end{aligned}$$

Then the condition  $|\sigma_{X,fm}| \leq \sqrt{\sigma_{a,f}^2 \sigma_{X,m}^2}$  is no longer a problem and we can go through all the same derivations as above and get the score statistic for testing  $H_0 : \sigma^2 = \sigma_{d,f}^2 = 0$

$$\Lambda = \begin{cases} \frac{Z_1^2 - 2rZ_1Z_2 + Z_2^2}{1-r^2} & \text{if } Z_1 \geq Z_2r, Z_2 \geq Z_1r \\ Z_1^2 & \text{if } Z_2 < Z_1r, Z_1 \geq 0 \\ Z_2^2 & \text{if } Z_1 < Z_2r, Z_2 \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

whose distribution depends on  $r = Corr(Z_1, Z_2)$  and is (CHERNOFF 1954)

$$\left(\frac{1}{2} - \frac{1}{2\pi} \arccos(r)\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{2\pi} \arccos(r)\chi_2^2$$

which should be verified by simulations under the null hypothesis. Interestingly, this statistic has the same form as the autosomal statistic  $S_{2n}$  developed by WANG (2002), they only differ in the way the  $\Pi$ 's are estimated based on either the rules for autosomal or X-linked inheritance.

**7.9.3.1 No dominance in females** If we assume that the dominance in females is negligible (i.e.  $\sigma_{d,f}^2 = 0$ ) then our test statistic becomes  $Z_1^2 \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ , which is of the same form as the statistic  $S_{1n}$  developed by [WANG \(2002\)](#).

## 7.10 DISCUSSION AND FUTURE WORK

### 7.10.1 Selected sampling, small samples, and choice of variance

An extensive simulation study has investigated the optimal choice of variance under various conditions in the autosomal case ([BHATTACHARJEE \*et al.\* 2008](#)). When the theoretical basis for the X-linked score statistics is complete a similar study should be performed. I suspect that similar results will hold for the X-linked case: the estimators for the scores themselves are same whether samples are small or ascertainment is selective (i.e., sampling based on phenotype) ([LEBREC \*et al.\* 2004](#)) but the correct estimators for the variances of the scores are not the same ([BHATTACHARJEE \*et al.\* 2008](#)). In the above derivations we have used the “conditional on IBD” approach to decompose the unconditional variance; but we could easily replace the variance estimators by the unconditional and fully empirical estimators. However, the unconditional variance can also be decomposed by conditioning on the trait values (i.e., “conditional on trait” variance formula). Conditioning on the trait is done as a surrogate for conditioning on the ascertainment scheme, which may not be very well documented or overly complicated, and results in more robust variance estimators than the “conditional on IBD” approach. The choice of variance directly affects power of the statistics and so very important theoretical work remains to be done ([BHATTACHARJEE \*et al.\* 2008](#)).

### 7.10.2 Inactivation in females and the pseudoautosomal regions

While it should be relatively straightforward to change the above derivation to appropriately model X-inactivation in females, the work remains to be done. Additionally, the model needs to be extended to appropriately model the pseudoautosomal regions.

### 7.10.3 Other genetic and environmental effects

I suspect that for a complex trait with considerable genetic contribution one major gene on the X chromosome is unlikely to account for all the genetic variation in the complex trait. Therefore, it may be important to extend the model to incorporate effects of autosomal genes, which may or may not show sex-specific effects and may or may not interact with the X-linked genes. One of the advantages of the variance component model, underlying our derivations, is how easily extendable it is. For example to add the effect of a major autosomal QTL and household (i.e. shared environmental) effects to the model of X-linked QTL effect, we assume that the autosomal and X-linked QTLs are in linkage equilibrium and write

$$y = \mu_s + g + a + h + e$$

where  $g$  is the effect of the X-linked gene,  $a$  the effect of the major autosomal gene,  $h$  the household effect and  $e$  the residuals. Then the covariance matrix may be written as

$$\Sigma = 2\Psi^f \sigma_{a,f}^2 + \Delta_{\gamma,X} \sigma_{d,f}^2 + \Psi^m \sigma_{X,m}^2 + \sqrt{2}\Psi^{fm} \sigma_{X,fm} + \Phi \sigma_a^2 + \Delta_{\gamma} \sigma_d^2 + H \sigma_h^2 + I \sigma_e^2$$

where  $H$  is a matrix whose elements are 1 if the relative pair shares the environmental exposure but 0 otherwise,  $I$  is the identity matrix,  $\Psi^f$  is a matrix of X-linked kinship coefficients whose elements corresponding to non-female-female pairs are 0 (the other  $\Psi$ 's are similarly defined).

### 7.10.4 Asymptotic and empirical distributions

I mentioned that the asymptotic distribution of the test statistics may be impossible to derive, especially in the general case when allowing for unequal allelic effects in females and males. However, for the theoretical derivations, above, to be useful in practice, the distributions of the test statistics need to empirically evaluated.

### 7.10.5 Score statistics for X-linked association analysis

It remains part of my future goals to build on this work and develop score statistics for family-based association analysis of X-linked markers.

### 7.10.6 General properties of score statistics

Generally score statistics have many good properties such as being locally most powerful and robust to non-normality of the dependent variable. They are asymptotically equivalent to the likelihood ratio test but simpler and faster to compute, especially when explicit expressions are known. We were able to derive explicit formulas under the simplifying, but perfectly reasonable, assumption of equal allelic effects in females and males. Before we can assume that the general properties of score statistics hold in the more general case that allows for unequal effects, more work and extensive simulations need to be done. The complexity of the parameter space in the general case makes intuitive guesses about the properties of the statistic harder to make.

## 8.0 CONCLUSIONS

### 8.1 SYNTHESIS OF THE ARM WORK

When I started working on ARM, about 5 years ago, the project had been ongoing for over 10 years. At that time, there was strong epidemiologic evidence that implicated heredity in the ARM pathogenesis. However, there was considerable doubt that common genetic variants influencing ARM could be identified and that they even existed. After all, ARM is a late onset disorder, typically affecting people over 65 years of age. Clearly, that alone leaves plenty of opportunities and time for environmental exposures to influence the phenotype. Add to that the complexity of the phenotype itself, for example the two advanced forms of ARM, the dry (GA) and the wet (CNV) forms seem to manifest themselves quite differently even though the end result is, in both cases, loss of central vision and damaged macula. There is, however, no clear evidence or examples that the phenotypic variations of advanced ARM reflect genetic heterogeneity and many individuals have both forms of the disease, even in the same eye. We have focused on establishing which individuals are truly affected with ARM and collectively analyzed all individuals, rather than focussing on subtype-specific analysis.

Three microsatellite linkage studies had already been published when I joined the team ([WEEKS \*et al.\* 2000](#); [WEEKS \*et al.\* 2001](#); [WEEKS \*et al.\* 2004](#)). The initial linkage study using our data, which at that time included over 200 families and 386 markers, identified susceptibility regions on chromosomes 5, 9, 10, and 12. After adding over 100 families and typing 18 additional markers in those regions, only the signals on chromosome 5 and 10 remained. However, no signal reached the genome-wide significance level of a LOD score of 3 or greater, the LOD scores observed were in the 1–1.5 range ([WEEKS \*et al.\* 2000](#)). The second study, was an expanded collaborative study of almost 400 families, that identified four regions, 1q31, 9p13, 10q26, and 17q25, with LOD scores over 2 and some over 3 ([WEEKS \*et al.\* 2001](#)). The third study, also an expanded collaborative study



of over 500 families, found continued evidence of LOD scores close to 3 within the 1q31, 10q26, and 17q25 susceptibility regions ([WEEKS et al. 2004](#)).

In 2004–2005, high-density focused SNP genotyping in those regions (1q31, 10q26, and 17q25) was done on an expanded data-set of 594 ARM-affected families and 179 unrelated controls. The analysis successfully replicated the discovery of the *CFH* gene under the 1q31 linkage peak, which was independently published by three groups shortly after we received our genotype data. Using those data, we were the first to report a locus of three closely linked genes (*PLEKHA1*, *LOC387715*, and *HTRA1*) under the 10q26 linkage peak (see chapter 3). Those studies particularly pinpointed the nonsynonymous SNP *rs1061170* (*Y402H*) in *CFH* and the nonsynonymous SNPs *rs1045216* (*A320T*) and *rs10490924* (*S69A*) in *PLEKHA1* and *LOC387715*, respectively ([JAKOBSDOTTIR et al. 2005](#)).

The *CFH* region harbors number of other genes in the same biological pathway, the alternative complement pathway. Regulation plays of this pathway plays a central role in innate immunity and inflammation. While it is outside of the scope of this thesis to discuss the pathway in detail, it is interesting and important to note that the widely replicated *Y402H* variant is not the only variant in the *CFH* gene, nor in the whole region, showing strong replicable association with ARM. Following the *CFH* discovery, there has been a considerable amount of work done in order to understand the genetic contribution of the gene, and the surrounding complement related genes, to the etiology of ARM. Those follow-up studies have identified number of haplotypes, spanning the region, and other variants in *CFH* and nearby genes. Conditional analysis have shown that there are at least two LD blocks with variants and haplotypes contributing susceptibility ([LI et al. 2006](#)).

The LD in the *LOC387715* region appears to be much more extensive than in the *CFH* region. However, we were able to perform conditional analysis in two independent cohorts where the LD between *A320T* in *PLEKHA1* and *S69A* in *LOC387715* is much lower than in our own cohort. In both cohorts we demonstrated that the association of this region with ARM is more likely attributed to the *LOC387715* SNP than the *PLEKHA1* SNP, and that *A320T* does not explain significant amount of the variation after accounting for *S69A* ([CONLEY et al. 2006](#)). These findings, do not exclude the possibility that there is another variant in strong LD with *S69A*, that is the true disease causing variant, nor do they imply that there are no other variants in the region contributing significantly, beyond *S69A*, to the susceptibility. In fact, another SNP in the promoter region of *HTRA1* has been discovered ([DEWAN et al. 2006](#); [YANG et al. 2006](#)). This variant

is, however, in almost complete LD with *S69A*. While the putative role *HTRA1* in extracellular matrix homeostasis makes it the obvious causal candidate within the locus (DEWAN *et al.* 2007), no study convincingly suggests either the *HTRA1* gene or the *LOC387715* gene to be a better ARM candidate gene than the other, and two studies even suggest that *LOC387715* and the *S69A* variant specifically is more likely to be the causal gene in this locus. Nevertheless, the strong LD across the two genes (*LOC387715* and *HTRA1*) makes statistical methods alone insufficient to distinguish between them; comprehensive analysis and characterization of the molecular and functional relevance of the variants in the region is warranted.

The *CFH* discovery has been quite fruitful. A few complement pathway based candidate genes studies have been done for ARM and resulted in the discoveries of two novel loci, one harboring the closely linked *C2* and *CFB* genes (GOLD *et al.* 2006) and the other the *C3* gene (YATES *et al.* 2007). We have published our replication effort of the *C2/CFB* locus (see chapter 5) but not the *C3* locus. However, we have data on that locus now and our not yet published analysis show that we also replicate the *C3* association signal. We also showed that the *C2/CFB* locus significantly confers susceptibility after accounting for the effect of the *CFH* and *LOC387715* genes. Neither of those association signals were observed in any of the family-based linkage studies.

It is interesting to note that linkage and association studies are thought to be optimally powered under different conditions: linkage being more powerful than association to detect rare variants with strong effect on disease risk and association being more powerful than linkage to detect common variants of smaller effect. However, even though only the rarer variants of *C2/CFB* seem to be associated with ARM, no linkage signals have been observed near this locus. This apparent lack of correspondence between linkage and association, for rarer variants, can perhaps be explained by the observations that the most strongly associated SNPs could be protective and would therefore not be detected in family studies based mostly on affected sib pairs (GORIN 2007). In fact, the original report on the *C2/CFB* locus stated that the variants were protective (GOLD *et al.* 2006); although it is impossible to establish the direction of potential causality from descriptive frequency data alone.

The *LOC387715* gene is an interesting candidate gene. Unlike the *CFH* gene, there is little known about molecular function of the *LOC387715* gene and the biological properties of its putative protein explaining its role in the pathobiology of ARM. Studying the molecular function of the gene specific to the ARM will undoubtedly prove challenging for number of reasons, including: 1) the

gene is an evolutionary recent gene with conservation restricted to the primate lineage (RIVERA *et al.* 2005), 2) even though ARM in the rhesus monkeys is also associated with *LOC387715/HTRA1* variants, the LD appears to be just as strong (FRANCIS *et al.* 2008), and 3) the typical experimental animals, mice, do not have a macula (RAKOCZY *et al.* 2006).

In the past couple of years many novel potentially causal SNPs and genes for common complex disease have been discovered in genetic association studies. For ARM the *CFH* and *LOC387715* findings are particularly remarkable and it is only natural to ask how powerful these might be in discriminating between those with and without the disease and those who will and will not develop the disease in the future (see chapter 6). While association statistics, such as odds ratios and *P* values, used during the discovery phase measure these properties indirectly, they are not the most appropriate for evaluating the predictive value of genetic profiles and can exaggerate the potential predictive power of the genetic data. Other measures, such as sensitivity, specificity, positive predictive values, negative predictive values, and area under the ROC curves, are more useful if the goal is to identify a genetic profile for classification or risk prediction. We argue, in our study, that the “scientific community should be very cautious to avoid overhyping association findings in terms of their personalized medicine value” and while robust statistical associations are important to establish etiological hypothesis, which in turn may enhance our understanding of the causes of the disease and ultimately lead to development of new therapeutic targets for treatment and prevention, they do not, alone, guarantee the clinical validity for the use of genetic profiles in medical or public health practice.

Among ARM geneticists, there seems to be consensus that there are more genes than *CFH*, *LOC387715*, and the other complement related genes predisposing individuals to ARM. I agree. Given that environmental risk factors also play significant role in ARM and that the effect of *CFH* and *LOC387715* appears strong, it will be challenging to find novel genes. It seems intuitively unlikely that many common variants remain to be discovered, there are perhaps some undiscovered variants of very weak effects or in regions poorly covered by genome-wide SNP panels. In the current phase of the study, we have attempted to increase our chances of finding novel disease genes via an association study using a candidate gene approach. The candidate genes were selected based on extensive literature search of biological targets and pathways and then subjected to a selection process based on prior statistical evidence regarding which regions are most likely to contain ARM related loci. However, even though we have reduced the burden of multiple testing by performing a candidate gene study our study is still limited regarding the number of unaffected controls, so

that the weaker effects and rarer variants will not necessarily reach statistical significance.

It is easy to say that now that the genome-wide association studies have found ‘all’ the common variants we should begin to focus on rarer variants. First, the studies to date may have found all the common variants of strong enough effect to be possibly found with the sample sizes available. Second, to find the rarer variants it may be most powerful to use family data, but for many diseases, like ARM, linkage studies using family data have already been done. For ARM, fine-mapping under the linkage peaks led to the discoveries of *CFH* and *LOC387715*. Using the same family data again will unlikely results in new peaks, even if a genome-wide SNP linkage panel is used instead of a microsatellite panel, as was used in the past. To find new replicable linkage signals more powerful larger data sets are needed. I am very hopeful that the approach we have taken with our ARM families will result in new findings; the goal is to extend the genome-wide linkage analyses of our ARM families by extending the pedigrees downward to include the next generation. The majority of our families have consisted of a single generation of siblings but the third generation children are now reaching the age to have presymptomatic findings for ARM and a small percentage will also develop advanced disease during the study.

All the ARM genes discovered to date were found by investigating individuals with advanced disease. Association has also been observed for milder disease and some early onset clinical features. I think it could improve the understanding of the biophysiology of ARM by not only demonstrating an association with early onset clinical features but also to prospectively evaluate the effect of a combination of *CFH* and *LOC387715* genotypes and specific early onset clinical features on the incidence and the progression to ARM. I think the evidence for the potential causal effect those genes have on ARM are strong enough to warrant such a study, albeit expensive. In fact, the third generation children ascertained in the next phase of our study will contribute to such an investigation.

## 8.2 MY FUTURE AIMS

I have worked quite extensively on ARM. I am excited to continue working on dissecting the genetics of complex diseases and hope to do so in the future. However, for the next couple of years, I plan on focusing on strengthening my methodology development skills and use both my

theoretical background, from my steps taken in developing score statistics and as a mathematician, and my applied and invaluable data analysis experience from working on the ARM for almost 5 years, to tackle theoretical problems in complex disease and population genetics. I think there is room for basic methodology development in genetics, for example for association analysis of X-linked and mitochondrial markers, especially when using family data. Genetics is more than just SNP genotypes and so I am also interested in methodology development for analyzing genetic data other than SNP data, such as CNV, expression, sequence or proteomics data.

## APPENDIX A

### FOR CHAPTER 3

The supplementary material published online as a part of the paper ([JAKOBSDOTTIR \*et al.\* 2005](#)) presented in chapter 3 is given here.

Table A1: Allele labeling

SNP and Allele	Label	Amino Acid	CIDR controls	Local controls	HWE P-value.
rs6658788					0.580
A	1		0.511	0.483	
G	2		0.489	0.517	
rs1538687					0.410
A	1		0.693	0.658	
G	2		0.307	0.342	
rs1416962					0.440
T	1		0.648	0.607	
C	2		0.352	0.393	
rs946755					0.700
T	1		0.656	0.62	
C	2		0.344	0.38	
rs6428352					1.000
T	1		0.997	0.996	
C	2		0.003	0.004	
rs800292					0.820
T	1		0.232	0.269	
C	2		0.768	0.731	
rs1061170		Tyr402His			0.260
T	1 = Tyr		...	0.69	
C	2 = His		...	0.31	
rs10922093					0.660
G	1		...	0.295	
A	2		...	0.705	
rs70620					0.280

Continued on next page

Table A1 – continued from previous page

T	1		0.173	0.15	
C	2		0.827	0.85	
rs1853883					0.450
G	1		0.511	0.568	
C	2		0.489	0.432	
rs1360558					0.700
A	1		0.397	0.389	
G	2		0.603	0.611	
rs955927					0.850
T	1		0.609	0.615	
A	2		0.391	0.385	
rs4350226					0.340
A	1		0.905	0.897	
G	2		0.095	0.103	
rs4752266					0.180
A	1		0.777	0.774	
G	2		0.223	0.226	
rs915394					1.000
T	1		0.813	0.791	
A	2		0.187	0.209	
rs1268947					0.650
G	1		0.883	0.885	
C	2		0.117	0.115	
rs1537576					0.350
G	1		0.567	0.581	
C	2		0.433	0.419	
rs2039488					0.010
T	1		0.885	0.885	
C	2		0.115	0.115	
rs1467813					0.660
T	1		0.293	0.295	
C	2		0.707	0.705	
rs927427					0.100
A	1		0.464	0.487	
G	2		0.536	0.513	
rs4146894					1.000
A	1		0.466	0.474	
G	2		0.534	0.526	
rs12258692		Pro233Arg			...
C	1 = Pro		...	1	
G	2 = Arg		...	0	
rs4405249					1.000
T	1		...	0.158	
C	2		...	0.842	
rs1045216		Ala320Thr			0.460
G	1 = Ala		...	0.573	
A	2 = Thr		...	0.427	

Continued on next page

Table A1 – continued from previous page

rs1882907					0.760
A	1		0.813	0.816	
G	2		0.187	0.184	
rs10490923		His3Arg			0.390
G	1 = Arg		...	0.859	
A	2 = His		...	0.141	
rs2736911		Arg38Ter			1.000
C	1 = Arg		...	0.881	
T	2 = Ter		...	0.119	
rs10490924		Ser69Ala			0.210
G	1 = Ala		...	0.807	
T	2 = Ser		...	0.193	
rs11538141		Gly54Glu			1.000
A	1 = Glu		...	0.995	
G	2 = Gly		...	0.005	
rs760336					0.580
T	1		0.52	0.526	
C	2		0.48	0.474	
rs763720					0.790
A	1		0.212	0.226	
G	2		0.788	0.774	
rs1803403		Cys384Gly			1.000
T	1 = Cys		...	0.03	
G	2 = Gly		...	0.97	



Table A2: Primers, Annealing Conditions, and Restriction Endonucleases Used for Genotype Data Collection

Variant	Primer Sequences		Annealing Temperature (°C)	Restriction Enzyme
	Forward	Reverse		
<i>rs11538141</i>	CAG AGT CGC CAT GCA GAT CC	CCC GAA GGG CAC CAC GCA CT	58	<i>MnII</i>
<i>rs2736911</i>	GCA CCT TTG TCA CCA CAT TA	GCC TGA TCA TCT GCA TTT CT	54	<i>DraIII</i>
<i>rs10490923</i>	GCA CCT TTG TCA CCA CAT TA	GCC TGA TCA TCT GCA TTT CT	54	<i>HhaI</i>
<i>rs10490924</i>	GCA CCT TTG TCA CCA CAT TA	GCC TGA TCA TCT GCA TTT CT	54	<i>PvuII</i>
<i>rs1803403</i>	TGC TGT CCC TTT GTT GTC TC	AGA CAC AGA CAC GCA TCC TG	55	NA
<i>rs12258692</i> (and <i>rs4405249</i> )	GCC AGG AAA AGG AAC CTC	GCC AGG CAT CAA GTC AGA	54	NA

Table A3: Results of Fitting Two-Locus Models by Logistic Regression

Locus 2 and Model	AIC	AIC Difference
<i>rs1537576 (GRK5):</i>		
MEAN	822.5	23.65
ADD1	798.8	0.00
ADD2	821.2	22.35
ADD	799.1	0.26
DOM1	799.7	0.91
DOM2	820.1	21.24
DOM	799.2	0.37
ADDINT	800.9	2.07
ADDDOM	802.1	3.25
DOMINT	803.9	5.07
<i>rs1467813 (RGS10):</i>		
MEAN	821.9	23.53
ADD1	798.4	0.00
ADD2	823.6	25.25
ADD	800.3	1.92
DOM1	799.3	0.91
DOM2	825.2	26.79
DOM	802.6	4.23
ADDINT	801.3	2.93
ADDDOM	804.9	6.54
DOMINT	805.2	6.83
<i>rs4146894 (PLEKHA1)</i>		
MEAN	823.02	49.26
ADD1	799.24	25.49
ADD2	801.47	27.71
ADD	773.76	0.00
DOM1	800.16	26.41
DOM2	803.44	29.68
DOM	776.44	2.68
ADDINT	775.62	1.87
ADDDOM	779.85	6.09
DOMINT	778.26	4.50
<i>rs760336 (PRSS11)</i>		
MEAN	821.9	27.32
ADD1	798.4	3.78
ADD2	817.1	22.54
ADD	794.6	0.00
DOM1	799.3	4.69
DOM2	819	24.37
DOM	796.7	2.14
ADDINT	796	1.43
ADDDOM	802.1	7.46
DOMINT	803.4	8.75

Table A4: ORs, ARs, and Simulated  $P$  Values from  $\chi^2$  Test with 10,000 Replicates

SNP (allele)	Gene	Subtype	Dominant ([RR+RN] vs. NN)				Heterozygotes (RN vs. NN)		Recessive (RR vs [RN+NN])				Homzygotes RR vs. NN	
			OR	95% CI	AR	P	OR	AR	OR	95% CI	AR	P	OR	AR
rs6658788 (2)		CNV	0.84	.56-1.25	-13.41	0.36706	1.21	6.19	1.11	.73-1.68	2.5	0.68123	0.95	-2.26
rs6658788 (2)		GA	0.88	.55-1.4	-9.92	0.63064	1.07	2.18	1.01	.62-1.66	0.35	1.0	0.92	-3.97
rs1538687 (2)		CNV	0.71	.5-1.02	-17.04	0.07499	0.54	-10.68	0.47	.25-.9	-5.97	0.0202	0.42	-11.38
rs1538687 (2)		GA	0.62	.41-.94	-23.86	0.0317	0.56	-10.14	0.45	.2-1.01	-6.25	0.07239	0.38	-12.32
rs1416962 (2)		CNV	0.88	.61-1.25	-7.7	0.41676	1.02	0.46	0.95	.56-1.62	-.62	0.89111	0.89	-2.63
rs1416962 (2)		GA	0.77	.51-1.17	-15.12	0.24708	0.69	-7.53	0.62	.31-1.24	-5.07	0.22948	0.57	-11.11
rs946755 (2)		CNV	0.84	.59-1.2	-9.81	0.37326	1.14	2.86	1.03	.6-1.78	0.37	1.0	0.94	-1.39
rs946755 (2)		GA	0.73	.48-1.11	-18.18	0.17258	0.79	-4.46	0.69	.34-1.38	-3.8	0.37606	0.6	-9.26
rs6428352 (2)		CNV	...	...	...	...	...	...	...	...	...	...	...	...
rs6428352 (2)		GA	...	...	...	...	...	...	...	...	...	...	...	...
rs800292 (1)	CFH	CNV	0.48	.33-.7	-26.97	0.0002	0.53	-21.4	0.21	.07-.64	-4.59	0.0053	0.18	-7.64
rs800292 (1)	CFH	GA	0.39	.25-.62	-33.02	0.0002	0.44	-26.29	0.09	.01-.75	-5.33	0.0113	0.08	-8.66
rs1061170 (2)	CFH	CNV	5.25	3.22-8.55	68.0	< 0.0001	2.37	24.74	4.11	2.2-7.69	27.24	< 0.0001	9.35	61.82
rs1061170 (2)	CFH	GA	5.76	3.17-10.47	70.42	< 0.0001	3.31	35.78	5.66	2.9-11.04	35.95	< 0.0001	12.26	68.61
rs10922093 (1)	CFH	CNV	0.56	.37-.85	-28.05	0.0083	0.61	-20.84	0.4	.18-.91	-5.96	0.0327	0.33	-11.72
rs10922093 (1)	CFH	GA	0.51	.31-.84	-32.2	0.0089	0.58	-23.06	0.26	.08-.85	-7.43	0.032	0.21	-14.08
rs70620 (1)	CFH	CNV	0.77	.52-1.14	-7.45	0.23338	0.8	-5.9	0.63	.24-1.69	-1.46	0.42256	0.6	-2.2
rs70620 (1)	CFH	GA	0.72	.45-1.15	-9.48	0.18978	0.78	-6.4	0.28	.06-1.36	-2.9	0.17068	0.26	-4.1
rs1853883 (2)		CNV	2.52	1.64-3.89	52.14	0.0002	1.5	15.44	1.88	1.28-2.78	18.84	0.0014	3.2	51.28
rs1853883 (2)		GA	3.54	1.97-6.36	64.51	< 0.0001	1.95	25.93	2.57	1.65-4	29.15	0.0003	5.12	66.42
rs1360558 (1)		CNV	1.1	.76-1.59	5.96	0.64364	1.04	2.29	1.24	.78-1.98	3.75	0.41376	1.27	7.61
rs1360558 (1)		GA	1.16	.75-1.79	9.09	0.57904	1.13	6.6	1.17	.68-2.02	2.67	0.67873	1.25	7.14
rs955927 (2)		CNV	1.12	.78-1.63	7.31	0.51105	1.32	7.01	1.34	.83-2.17	4.9	0.20048	1.38	9.84
rs955927 (2)		GA	1.08	.7-1.67	5.0	0.74163	1.18	4.02	1.2	.68-2.1	2.86	0.57564	1.22	6.06
rs4350226 (2)		CNV	0.55	.34-.91	-8.65	0.0209	...	...	...	...	...	...	...	...
rs4350226 (2)		GA	0.52	.28-.96	-9.46	0.0462	...	...	...	...	...	...	...	...
rs4752266 (2)	GRK5	CNV	0.93	.65-1.34	-2.87	0.71243	3.13	10.08	2.82	.96-8.24	3.9	0.06229	2.63	5.74
rs4752266 (2)	GRK5	GA	0.78	.51-1.19	-10.33	0.27667	3.67	12.31	2.88	.9-9.23	4.04	0.06909	2.51	5.33
rs915394 (2)	GRK5	CNV	1.39	.96-2.01	11.91	0.08469	1.28	2.23	1.56	.57-4.3	1.54	0.48645	1.74	2.96
rs915394 (2)	GRK5	GA	1.09	.7-1.67	2.88	0.74493	1.38	2.94	1.42	.44-4.58	1.17	0.57824	1.45	1.81
rs1268947 (2)	GRK5	CNV	1.15	.75-1.75	3.15	0.52415	1.23	1.72	1.35	.36-5.05	0.58	0.76382	1.39	0.8
rs1268947 (2)	GRK5	GA	0.78	.46-1.32	-5.0	0.42146	1.24	1.82	1.0	.2-5.02	0.0	1.0	0.95	-.1
rs1537576 (2)	GRK5	CNV	1.57	1.07-2.3	27.1	0.0211	0.88	-4.02	1.06	.69-1.63	1.3	0.83192	1.44	14.26

Continued on next page

Table A4 – continued from previous page

rs1537576 (2)	GRK5	GA	1.84	1.15-2.94	35.48	0.0143	1.17	5.17	1.44	.89-2.34	8.51	0.17778	2.04	28.32
rs2039488 (2)		CNV	0.76	.48-1.2	-5.05	0.28877	0.2	-12.5	0.18	.03-.91	-2.36	0.0318	0.17	-2.88
rs2039488 (2)		GA	0.62	.35-1.09	-8.3	0.11599	0.27	-11.29	0.2	.02-1.7	-2.3	0.21758	0.19	-2.83
rs1467813 (1)	RGS10	CNV	0.95	.67-1.36	-2.31	0.85551	0.98	-.68	0.83	.45-1.54	-1.61	0.63004	0.83	-2.79
rs1467813 (1)	RGS10	GA	0.85	.56-1.29	-7.69	0.52905	0.88	-5.73	0.81	.39-1.69	-1.85	0.70453	0.76	-3.85
rs927427 (1)		CNV	1.08	.72-1.63	5.82	0.75722	0.91	-6.99	1.76	1.13-2.74	11.97	0.0107	1.65	21.15
rs927427 (1)		GA	1.1	.68-1.78	6.67	0.81152	0.98	-1.63	1.5	.9-2.5	8.16	0.15618	1.47	16.43
rs4146894 (1)	PLEKHA1	CNV	2.53	1.64-3.91	52.45	< 0.0001	1.94	37.72	2.46	1.63-3.71	23.64	< 0.0001	3.95	56.0
rs4146894 (1)	PLEKHA1	GA	2.09	1.24-3.51	44.0	0.0069	1.77	33.08	1.92	1.2-3.08	16.31	0.0084	2.87	44.63
rs12258692 (2)	PLEKHA1	CNV	...	...	...	...	...	...	...	...	...	...	...	...
rs12258692 (1)	PLEKHA1	GA	...	...	...	...	...	...	...	...	...	...	...	...
rs4405249 (1)	PLEKHA1	CNV	0.53	.27-1.02	-16.43	0.06989	0.51	-16.36	0.97	.11-8.85	-.05	1.0	0.83	-.41
rs4405249 (1)	PLEKHA1	GA	0.63	.3-1.33	-12.27	0.17898	0.64	-11.43	0.57	.04-9.31	-.76	1.0	0.51	-1.2
rs1045216 (2)	PLEKHA1	CNV	0.5	.32-.78	-48.35	0.0026	0.4	-22.18	0.31	.17-.58	-15.72	0.0002	0.25	-38.01
rs1045216 (2)	PLEKHA1	GA	0.44	.26-.72	-58.72	0.001	0.45	-19.85	0.32	.15-.7	-15.46	0.0017	0.24	-38.67
rs1882907 (2)		CNV	0.52	.35-.77	-19.38	0.0024	0.7	-3.08	0.44	.14-1.38	-1.91	0.20808	0.38	-3.11
rs1882907 (2)		GA	0.6	.38-.95	-15.9	0.035	0.22	-8.27	0.16	.02-1.37	-2.89	0.12039	0.14	-4.32
rs10490923 (2)	LOC387715	CNV	0.48	.28-.85	-14.84	0.0114	0.17	-11.63	0.1	.01-.98	-2.89	0.0413	0.09	-3.78
rs10490923 (2)	LOC387715	GA	0.74	.39-1.38	-7.07	0.41496	0.58	-5.49	0.48	.08-2.91	-1.66	0.65244	0.45	-2.24
rs2736911 (2)	LOC387715	CNV	0.71	.41-1.22	-7.12	0.24548	1.22	0.96	0.92	.09-8.92	-.08	1.0	0.86	-.18
rs2736911 (2)	LOC387715	GA	0.62	.32-1.19	-9.43	0.13179	2.2	4.96	1.43	.13-15.97	0.42	1.0	1.3	0.38
rs10490924 (2)	LOC387715	CNV	5.64	3.52-9.06	60.52	< 0.0001	2.81	23.7	6.18	2.62-14.59	22.67	< 0.0001	12.11	46.39
rs10490924 (2)	LOC387715	GA	3.43	2.02-5.84	44.55	< 0.0001	2.63	21.83	4.74	1.9-11.84	17.47	0.0003	7.05	32.05
rs11538141 (2)	PRSS11	CNV	...	...	...	...	...	...	...	...	...	...	...	...
rs11538141 (2)	PRSS11	GA	...	...	...	...	...	...	...	...	...	...	...	...
rs760336 (2)	PRSS11	CNV	0.63	.43-.92	-37.33	0.0178	0.71	-10.35	0.61	.39-.95	-10.03	0.0348	0.49	-30.78
rs760336 (2)	PRSS11	GA	0.63	.4-.98	-36.73	0.0322	0.84	-5.52	0.71	.42-1.19	-7.3	0.23778	0.56	-25.69
rs763720 (1)	PRSS11	CNV	1.77	1.24-2.54	23.25	0.0031	1.69	20.43	2.1	.85-5.18	3.55	0.12829	2.64	7.87
rs763720 (1)	PRSS11	GA	1.74	1.14-2.65	22.5	0.0107	1.4	12.86	4.71	1.88-11.79	11.06	0.0001	5.41	18.69
rs1803403 (1)	PRSS11	CNV	3.33	1.39-8.02	12.17	0.0055	3.33	12.17	...	...	...	...	...	...
rs1803403 (1)	PRSS11	GA	3.85	1.53-9.72	14.49	0.0039	3.85	14.49	...	...	...	...	...	...

NOTE.—Type A-affected individuals are compared with controls. Allele denotes the risk allele (minor allele in controls). RR = homozygotes for the risk allele; RN = heterozygotes for the risk allele; NN = homozygotes for the normal allele. Locally typed SNPs are in bold italics. Blank spaces separate the three chromosomal regions corresponding to SNPs in and around *CFH*, *GRK5/RGS10*, and *PLEKHA1/LOC687715/PRSS11*.

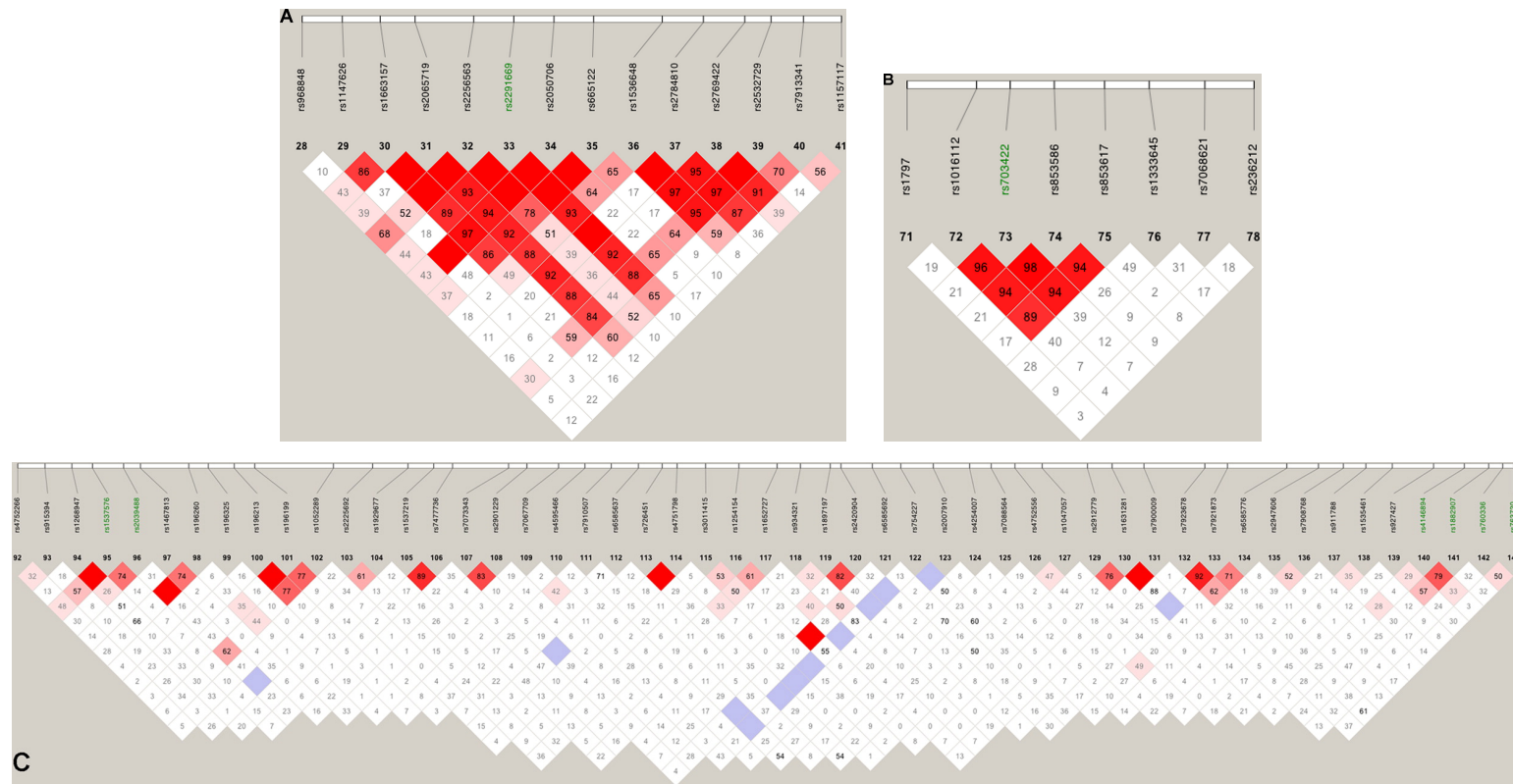


Figure A1: LD patterns on chromosome 10 based on analysis of 196 CIDR SNPs and 179 unrelated controls. A, The false peak at 135 cM (see fig. 3.3); the SNP with the largest  $S_{all}$  in the peak is highlighted in green. B, The false peak at 142 cM (see fig. 3.3); the SNP with the largest  $S_{all}$  in the peak is highlighted in green. C, Linkage peak. Significant SNPs, from CCREL (table 5), that overlie the five genes (*GRK5*, *RGS10*, *PLEKHA1*, *LOC387715*, and *PRSS11*) are highlighted in green. Squares shaded pink or red indicate significant LD between SNP pairs (bright red indicates pairwise  $D' = 1$ ), white squares indicate no evidence of significant LD, and blue squares indicate pairwise  $D' = 1$  without statistical significance. LD is measured using  $D'$ , and the values within the squares give pairwise LD in  $D'/100$ .

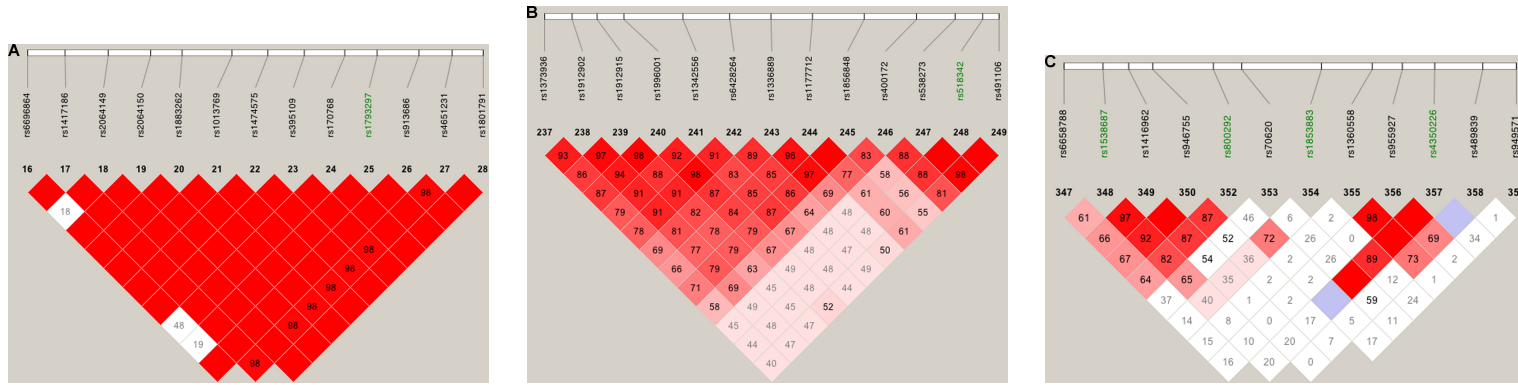


Figure A2: LD patterns on chromosome 1 based on analysis of 679 CIDR SNPs and 179 unrelated controls. A, The false peak at 188 cM (see fig. 3.4); the SNP with the largest  $S_{all}$  in the peak is highlighted in green. B, The false peak at 202 cM (see fig. 3.4); the SNP with the largest  $S_{all}$  in the peak is highlighted in green. C, Linkage peak. Significant SNPs, from CCREL (table 5), that overlie CFH are highlighted in green. Squares shaded pink or red indicate significant LD between SNP pairs (bright red indicates pairwise  $D' = 1$ ), white squares indicate no evidence of significant LD, and blue squares indicate pairwise  $D' = 1$  without statistical significance. LD is measured using  $D'$ , and the values within the squares give pairwise LD in  $D'/100$ .

## APPENDIX B

### FOR CHAPTER 4

Here the supplementary material published online as a part of the paper (CONLEY *et al.* 2006) presented in chapter 4 is given.

#### B.1 DISTINGUISHING BETWEEN *PLEKHA1* AND *LOC387715*

We employed the haplotype method (VALDES and THOMSON 1997) to identify which one of the two loci, *A320T* in *PLEKHA1* or *S69A* in *LOC387715*, is more likely the actual disease predisposing variant in the 10q26 region. If all predisposing variants are included on a haplotype, then the neutral variants are expected to be in the same ratio on a particular disease predisposing haplotype, in cases and controls, although the actual frequencies may differ. On the other hand, if not all predisposing variants have been identified, equality in the ratios of haplotype frequencies of non-predisposing variants is not expected.

The expected ratios for the *A320T-S69A* haplotypes are formulated below, assuming one variant is ARM-predisposing and the other is a neutral variant. We assume that *A320T* and *S69A* are all the ARM predisposing variants in the *PLEKHA1-LOC387715* haplotype block on chromosome 10q26. Four possible *A320T-S69A* haplotypes exist: G-G, A-G, G-T, and A-T. If *A320T* is the causal locus and *S69A* the neutral locus, we expect:

$$\left[ \begin{array}{c} f(G - G) \\ f(G - T) \end{array} \right]_{controls} = \left[ \begin{array}{c} f(G - G) \\ f(G - T) \end{array} \right]_{cases} \quad (1a)$$

$$\left[ \begin{array}{c} f(A - G) \\ f(A - T) \end{array} \right]_{controls} = \left[ \begin{array}{c} f(A - G) \\ f(A - T) \end{array} \right]_{cases} \quad (1b)$$

but, if *S69A* is the causal locus and *A320T* the neutral locus, we expect:

$$\left[ \begin{array}{c} f(G - G) \\ f(A - G) \end{array} \right]_{controls} = \left[ \begin{array}{c} f(G - G) \\ f(A - G) \end{array} \right]_{cases} \quad (2a)$$

$$\left[ \begin{array}{c} f(G - T) \\ f(A - T) \end{array} \right]_{controls} = \left[ \begin{array}{c} f(G - T) \\ f(A - T) \end{array} \right]_{cases} \quad (2b)$$

where  $f$  denotes frequencies of a particular haplotype in controls or cases.

The hypotheses of interest are:

$H_{0P}$ : The *A320T* variant in *PLEKHA1* fully accounts for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block.

$H_{0L}$ : The *S69A* variant in *LOC387715* fully accounts for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block.

Rejecting either of these hypotheses means that the tested variant is not sufficient to account for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block, alone. Four  $2 \times 2$  tables can be derived from equations 1a, 1b, 2a, and 2b:

<b>Table 1a</b>	Unexposed	Exposed	<b>Table 1b</b>	Unexposed	Exposed
Controls	$f(G - G)$	$f(G - T)$	Controls	$f(A - G)$	$f(A - T)$
Cases	$f(G - G)$	$f(G - T)$	Cases	$f(A - G)$	$f(A - T)$
<b>Table 2a</b>	Unexposed	Exposed	<b>Table 2b</b>	Unexposed	Exposed
Controls	$f(G - G)$	$f(A - G)$	Controls	$f(G - T)$	$f(A - T)$
Cases	$f(G - G)$	$f(A - G)$	Cases	$f(G - T)$	$f(A - T)$



Under  $H_{0_P}$  we expect homogeneity in contingency tables 1a and 1b, and under  $H_{0_L}$  we expect homogeneity in contingency tables 2c and 2d. Regular  $\chi^2$  statistic may be calculated from each contingency table to generate a combined statistic. For  $H_{0_P}$  the statistic is the maximum  $\chi^2$  from tables 1a and 1b, and for  $H_{0_L}$  the statistic is the maximum  $\chi^2$  from tables 2a and 2b. However, due to dependency of the statistics derived from each set of contingency tables, the distribution of the combined statistics is not clear. The lack of independence arises from (1) combining measurements corresponding to various alleles at predisposing loci, and (2) linkage disequilibrium between predisposing and non-predisposing loci. Both of these conditions are inevitable, (1) because variant always has more than one allele, and (2) because, if the variants are in complete linkage equilibrium, there is no need to distinguish between their independent association signals.

As a result of the dependency in the data a permutation testing needs to be done conditionally on the allele at the predisposing locus (under the null hypotheses). We start by grouping the haplotypes (two for each person) according to the allele at the predisposing locus. Then the case-control labels are permuted within each group and a combined statistic is calculated for each pair of replicate. This permutation procedure is similar to the procedure proposed by (Li 2001).

### B.1.1 Distinguishing between *PLEKHA1* and *LOC387715*—Results

Haplotype frequencies were estimated separately in controls and cases. The program SNPHAP (CLAYTON) was used to estimate the haplotype frequencies and phased haplotypes at each subject. The estimated haplotype frequencies are:

<i>A320T-S69A</i>	AREDS		CHS	
	Controls	Cases	Controls	Cases
<i>G – G</i>	0.3928	0.2802	0.3909	0.3188
<i>G – T</i>	0.1790	0.4149	0.1924	0.3294
<i>A – G</i>	0.4186	0.2792	0.3894	0.3337
<i>A – T</i>	0.0096	0.0257	0.0272	0.0180

We calculate the ratio of haplotypes under  $H_{0P}$  and compare the value estimated from controls and cases. From the AREDS data we have:

$$\left[ \frac{f(G-G)}{f(G-T)} \right]_{controls} / \left[ \frac{f(G-G)}{f(G-T)} \right]_{cases} = \left[ \frac{0.3928}{0.1790} \right] / \left[ \frac{0.2802}{0.4149} \right] = 3.25,$$

$$\left[ \frac{f(A-G)}{f(A-T)} \right]_{controls} / \left[ \frac{f(A-G)}{f(A-T)} \right]_{cases} = \left[ \frac{0.4186}{0.0096} \right] / \left[ \frac{0.2792}{0.0257} \right] = 4.01,$$

and  $H_{0P}$  is rejected ( $P \leq 0.0001$ ) in the AREDS data. Similarly, from the CHS data we have:

$$\left[ \frac{f(G-G)}{f(G-T)} \right]_{controls} / \left[ \frac{f(G-G)}{f(G-T)} \right]_{cases} = \left[ \frac{0.3909}{0.1924} \right] / \left[ \frac{0.3188}{0.3294} \right] = 2.10,$$

$$\left[ \frac{f(A-G)}{f(A-T)} \right]_{controls} / \left[ \frac{f(A-G)}{f(A-T)} \right]_{cases} = \left[ \frac{0.3894}{0.0272} \right] / \left[ \frac{0.3337}{0.0180} \right] = 0.77,$$

and  $H_{0P}$  is rejected ( $P=0.0002$ ) in the CHS data.

Now we calculate the ratio of haplotypes under  $H_{0L}$  and compare the value estimated from controls and cases. From the AREDS data we have:

$$\left[ \frac{f(G-G)}{f(A-G)} \right]_{controls} / \left[ \frac{f(G-G)}{f(A-G)} \right]_{cases} = \left[ \frac{0.3928}{0.4186} \right] / \left[ \frac{0.2802}{0.2792} \right] = 0.94,$$

$$\left[ \frac{f(G-T)}{f(A-T)} \right]_{controls} / \left[ \frac{f(G-T)}{f(A-T)} \right]_{cases} = \left[ \frac{0.1790}{0.0096} \right] / \left[ \frac{0.4149}{0.0257} \right] = 1.15,$$

and  $H_{0L}$  is not rejected ( $P=0.92$ ) in the AREDS data. Similarly, from the CHS data we have:

$$\left[ \frac{f(G-G)}{f(A-G)} \right]_{controls} / \left[ \frac{f(G-G)}{f(A-G)} \right]_{cases} = \left[ \frac{0.3909}{0.3894} \right] / \left[ \frac{0.3188}{0.3337} \right] = 1.05,$$

$$\left[ \frac{f(G-T)}{f(A-T)} \right]_{controls} / \left[ \frac{f(G-T)}{f(A-T)} \right]_{cases} = \left[ \frac{0.1924}{0.0272} \right] / \left[ \frac{0.3294}{0.0180} \right] = 0.39,$$

and  $H_{0L}$  is not rejected ( $P=0.45$ ) in the CHS data.

In conclusion,  $H_{0P}$  is rejected both when the hypothesis is tested using the AREDS and CHS data. Therefore, it is unlikely that *A320T* in *PLEKHA1* is sufficient to account for the ARM predisposition to the *PLEKHA1-LOC387715* haplotype block on chromosome 10q26. On the other hand we fail to reject similar hypothesis for *S69A* in *LOC387715*. Our results of applying the haplotype method support *LOC387715* as a major susceptibility gene for ARM.

## B.2 HAPMAP POPULATIONS

Individuals recruited into the CEPH population (CEU) of the International HapMap project ([THE INTERNATIONAL HAPMAP CONSORTIUM 2003](#); [THE INTERNATIONAL HAPMAP CONSORTIUM 2005](#)) were residents of Utah with ancestry from northern and western Europe. Individuals recruited from Ibadan in Yoruba, Nigeria (YRI), were individuals who identified themselves as having four Yoruba grandparents. 90 individuals (30 parent-offspring trios) were sampled from each population.

The following table compares the genotype frequencies in different populations without ARM. In addition to the 1051 white CHS controls and 126 white CHS ARM cases, a total of 180 African American CHS controls were genotyped for *Y402H* in *CFH*, *M299V* in *ELOVL4*, *A320T* in *PLEKHA1*, and *S69A* in *LOC387715*. Genotype frequencies for each of the two HapMap populations (CEU and YRI) are estimated from 60 individuals. Note that the HapMap populations have not been genotyped for *Y402H*.

Gene (Variant) and Genotypes	AREDS Whites	CHS Whites	HapMap CEU	CHS Blacks	HapMap YRI	<i>P</i> <sup>a</sup>
<i>CFH</i> ( <i>Y402H</i> )						
TT	0.434	0.448	...	0.367	...	
CT	0.416	0.450	...	0.528	...	0.14
CC	0.150	0.103	...	0.106	...	
<i>ELOVL4</i> ( <i>M299V</i> )						
AA	0.711	0.802	0.717	0.695	0.533	
AG	0.259	0.174	0.233	0.264	0.333	0.01
GG	0.030	0.030	0.050	0.040	0.133	
<i>PLEKHA1</i> ( <i>A320T</i> )						
GG	0.339	0.346	0.317	0.633	0.729	
AG	0.464	0.476	0.467	0.311	0.254	<0.01
AA	0.196	0.178	0.217	0.057	0.017	
<i>LOC387715</i> ( <i>S69A</i> )						
GG	0.645	0.604	0.583	0.561	0.467	
GT	0.331	0.353	0.400	0.368	0.450	0.25
TT	0.023	0.043	0.017	0.070	0.083	

<sup>a</sup> *P* value from 2 df Pearson's  $\chi^2$  test of difference in genotype frequencies between CHS

Whites (non-Hispanic whites) and CHS Blacks (African Americans)

Table B1: Genotype distributions in AREDS and CHS cohorts, by ARM status

Gene (Variant) Genotype	AREDS									CHS	
	Controls			Cases ( $n = 701$ )						Controls ( $n = 1051$ )	Cases ( $n = 126$ )
	Grade 1 ( $n = 175$ )	Grade 2 ( $n = 63$ )	Grade 3 ( $n = 96$ )	Grade 4 ( $n = 266$ )			Grade 5 ( $n = 399$ )				
All				GA only	CNV only	All	GA only	CNV only			
<b>CFH</b> (Y402H)											
TT	75	19	21	46	13	27	52	12	25	406	29
CT	72	35	39	118	21	72	147	40	52	408	53
CC	26	9	36	101	25	49	139	36	52	93	28
All	173	63	96	265	59	148	338	88	129	907	110
<b>ELOVL4</b> (M299V)											
AA	118	55	75	204	47	115	249	70	97	826	92
AG	43	7	17	50	10	30	65	14	23	179	31
GG	5	0	1	6	2	3	9	2	0	25	1
All	166	62	93	260	59	148	323	86	120	1030	124
<b>PLEKHA1</b> (A320T)											
GG	57	24	42	111	25	65	169	34	73	355	51
AG	78	31	45	114	25	61	142	43	46	489	57
AA	33	7	6	34	7	21	17	6	6	183	16
All	168	62	93	259	57	147	328	83	125	1027	124
<b>LOC387715</b> (S69A)											
GG	111	40	35	105	22	59	77	30	26	601	53
GT	57	19	44	126	31	74	171	44	70	351	49
TT	4	3	15	31	6	14	89	13	33	43	18
All	172	62	94	262	59	147	337	87	129	995	120

NOTE.—Genotypes are ordered: NN - RN - RR, where N is the normal allele and R is the risk allele. The risk allele is defined as the least frequent allele in controls. GA = geographic atrophy. CNV = choroidal neovascular membranes.

a Grade 2 AREDS subjects were not included in the analysis.

Table B2: Estimated crude ORs, corresponding 95% CIs, and PARs, unadjusted for age and gender

Gene (Variant) and Comparison in AREDS or CHS	DOMINANCE (RN+RR vs. NN)		RECESSIVE (RR vs. RN+NN)		HETEROZYGOTES (RN vs. NN)		HOMOZYGOTES (RR vs. NN)	
	OR <sub>dom</sub> (95% CI)	PAR	OR <sub>rec</sub> (95% CI)	PAR	OR <sub>het</sub> (95% CI)	PAR	OR <sub>hom</sub> (95% CI)	PAR
<i>CFH</i> (Y402H)								
1 vs. 345	3.73 (2.60, 5.34)	0.60	3.69 (2.37, 5.75)	0.22	2.66 (1.81, 3.92)	0.43	6.69 (4.08, 10.98)	0.37
1 vs. 45	3.94 (2.72, 5.71)	0.62	3.74 (2.39, 5.85)	0.22	2.82 (1.89, 4.19)	0.45	7.06 (4.27, 11.70)	0.38
1 vs. 3	2.73 (1.55, 4.83)	0.49	3.39 (1.89, 6.10)	0.20	1.93 (1.04, 3.60)	0.30	4.95 (2.46, 9.95)	0.29
1 vs. 4	3.64 (2.35, 5.64)	0.59	3.48 (2.14, 5.66)	0.20	2.67 (1.67, 4.27)	0.43	6.33 (3.60, 11.16)	0.35
1 vs. 5	4.21 (2.76, 6.42)	0.64	3.95 (2.47, 6.32)	0.23	2.94 (1.87, 4.63)	0.47	7.71 (4.46, 13.34)	0.41
1 vs. 45 (GA)	3.73 (2.21, 6.31)	0.60	4.01 (2.36, 6.82)	0.24	2.54 (1.44, 4.48)	0.41	7.04 (3.69, 13.41)	0.38
1 vs. 4 (GA)	2.71 (1.36, 5.37)	0.49	4.16 (2.14, 8.07)	0.24	1.68 (0.78, 3.61)	0.23	5.55 (2.48, 12.41)	0.32
1 vs. 5 (GA)	4.85 (2.46, 9.56)	0.68	3.91 (2.16, 7.10)	0.23	3.47 (1.69, 7.14)	0.53	8.65 (3.92, 19.09)	0.44
1 vs. 45 (CNV)	3.31 (2.16, 5.07)	0.56	3.24 (2.00, 5.26)	0.19	2.48 (1.57, 3.93)	0.40	5.60 (3.21, 9.78)	0.32
1 vs. 4 (CNV)	3.43 (2.05, 5.74)	0.57	2.80 (1.63, 4.80)	0.16	2.78 (1.61, 4.80)	0.44	5.24 (2.74, 10.01)	0.30
1 vs. 5 (CNV)	3.18 (1.87, 5.41)	0.55	3.82 (2.21, 6.59)	0.22	2.17 (1.22, 3.86)	0.34	6.00 (3.12, 11.53)	0.34
CHS	2.26 (1.45, 3.53)	0.41	2.99 (1.85, 4.83)	0.17	1.82 (1.13, 2.92)	0.27	4.22 (2.39, 7.42)	0.25
<i>ELOVL4</i> (M299V)								
1 vs. 345	0.69 (0.47, 1.01)	-0.07	0.78 (0.28, 2.16)	-0.01	0.69 (0.46, 1.02)	-0.06	0.72 (0.26, 1.99)	-0.01
1 vs. 45	0.71 (0.48, 1.04)	-0.06	0.85 (0.30, 2.38)	0.00	0.70 (0.46, 1.04)	-0.06	0.78 (0.28, 2.19)	-0.01
1 vs. 3	0.59 (0.32, 1.09)	-0.09	0.35 (0.04, 3.04)	-0.02	0.62 (0.33, 1.17)	-0.07	0.31 (0.04, 2.75)	-0.02
1 vs. 4	0.67 (0.43, 1.06)	-0.07	0.76 (0.23, 2.53)	-0.01	0.67 (0.42, 1.07)	-0.06	0.69 (0.21, 2.32)	-0.01
1 vs. 5	0.73 (0.48, 1.12)	-0.06	0.92 (0.30, 2.80)	0.00	0.72 (0.46, 1.12)	-0.05	0.85 (0.28, 2.60)	0.00
1 vs. 45 (GA)	0.59 (0.35, 1.00)	-0.09	0.91 (0.24, 3.47)	0.00	0.56 (0.32, 0.99)	-0.08	0.81 (0.21, 3.08)	0.00
1 vs. 4 (GA)	0.63 (0.31, 1.29)	-0.08	1.13 (0.21, 5.99)	0.00	0.58 (0.27, 1.26)	-0.08	1.00 (0.19, 5.36)	0.00
1 vs. 5 (GA)	0.56 (0.30, 1.06)	-0.10	0.77 (0.15, 4.04)	-0.01	0.55 (0.28, 1.07)	-0.09	0.67 (0.13, 3.57)	-0.01
1 vs. 45 (CNV)	0.65 (0.42, 1.01)	-0.07	0.36 (0.09, 1.55)	-0.02	0.69 (0.43, 1.09)	-0.06	0.33 (0.08, 1.42)	-0.02
1 vs. 4 (CNV)	0.71 (0.42, 1.18)	-0.06	0.67 (0.16, 2.84)	-0.01	0.72 (0.42, 1.22)	-0.05	0.62 (0.14, 2.64)	-0.01
1 vs. 5 (CNV)	0.58 (0.33, 1.03)	-0.09	...	...	0.65 (0.37, 1.15)	-0.06	...	...
CHS	1.41 (0.92, 2.17)	0.07	0.33 (0.04, 2.43)	-0.02	1.55 (1.00, 2.41)	0.09	0.36 (0.05, 2.68)	-0.02
<i>PLEKHAI</i> (A320T)								
1 vs. 345	0.57 (0.40, 0.81)	-0.39	0.37 (0.23, 0.60)	-0.13	0.68 (0.47, 0.99)	-0.18	0.31 (0.18, 0.51)	-0.14
1 vs. 45	0.56 (0.39, 0.81)	-0.40	0.39 (0.24, 0.63)	-0.12	0.67 (0.46, 0.98)	-0.19	0.31 (0.19, 0.53)	-0.14
1 vs. 3	0.62 (0.37, 1.05)	-0.33	0.28 (0.11, 0.70)	-0.15	0.78 (0.46, 1.35)	-0.12	0.25 (0.09, 0.64)	-0.16
1 vs. 4	0.68 (0.46, 1.02)	-0.26	0.62 (0.37, 1.04)	-0.07	0.75 (0.49, 1.15)	-0.13	0.53 (0.30, 0.94)	-0.09
1 vs. 5	0.48 (0.33, 0.71)	-0.51	0.22 (0.12, 0.42)	-0.16	0.61 (0.41, 0.92)	-0.23	0.17 (0.09, 0.34)	-0.17
1 vs. 45 (GA)	0.70 (0.44, 1.12)	-0.24	0.42 (0.21, 0.83)	-0.12	0.84 (0.52, 1.37)	-0.08	0.38 (0.18, 0.80)	-0.12
1 vs. 4 (GA)	0.66 (0.36, 1.21)	-0.29	0.57 (0.24, 1.38)	-0.08	0.73 (0.38, 1.40)	-0.15	0.48 (0.19, 1.24)	-0.10
1 vs. 5 (GA)	0.74 (0.43, 1.27)	-0.20	0.32 (0.13, 0.79)	-0.14	0.92 (0.53, 1.63)	-0.04	0.30 (0.12, 0.80)	-0.14
1 vs. 45 (CNV)	0.50 (0.33, 0.74)	-0.49	0.45 (0.26, 0.78)	-0.11	0.57 (0.37, 0.87)	-0.26	0.34 (0.19, 0.61)	-0.13
1 vs. 4 (CNV)	0.65 (0.41, 1.02)	-0.30	0.68 (0.37, 1.24)	-0.06	0.69 (0.42, 1.12)	-0.18	0.56 (0.29, 1.07)	-0.09
1 vs. 5 (CNV)	0.37 (0.23, 0.59)	-0.71	0.21 (0.08, 0.51)	-0.16	0.46 (0.28, 0.76)	-0.35	0.14 (0.06, 0.36)	-0.18
CHS	0.76 (0.52, 1.11)	-0.19	0.68 (0.39, 1.18)	-0.06	0.81 (0.54, 1.21)	-0.10	0.61 (0.34, 1.10)	-0.07
<i>LOC387715</i> (S69A)								
1 vs. 345	3.99 (2.81, 5.67)	0.54	10.16 (3.70, 27.88)	0.28	3.06 (2.13, 4.39)	0.42	17.26 (6.22, 47.89)	0.41
1 vs. 45	4.17 (2.92, 5.96)	0.56	10.52 (3.83, 28.93)	0.29	3.18 (2.20, 4.60)	0.43	18.30 (6.57, 50.93)	0.43
1 vs. 3	3.07 (1.82, 5.17)	0.45	7.97 (2.56, 24.81)	0.23	2.45 (1.42, 4.23)	0.34	11.89 (3.70, 38.19)	0.32
1 vs. 4	2.72 (1.83, 4.05)	0.41	5.64 (1.95, 16.27)	0.17	2.34 (1.55, 3.53)	0.32	8.19 (2.80, 24.00)	0.24
1 vs. 5	6.14 (4.11, 9.19)	0.67	15.07 (5.43, 41.82)	0.38	4.32 (2.85, 6.57)	0.54	32.07 (11.30, 91.01)	0.57
1 vs. 45 (GA)	3.29 (2.07, 5.21)	0.48	6.28 (2.09, 18.93)	0.19	2.81 (1.74, 4.52)	0.39	10.14 (3.28, 31.31)	0.28
1 vs. 4 (GA)	3.06 (1.66, 5.65)	0.45	4.75 (1.29, 17.49)	0.14	2.74 (1.46, 5.17)	0.38	7.57 (1.97, 29.06)	0.22
1 vs. 5 (GA)	3.46 (2.01, 5.94)	0.49	7.38 (2.33, 23.38)	0.22	2.86 (1.63, 5.02)	0.40	12.02 (3.65, 39.57)	0.32
1 vs. 45 (CNV)	4.09 (2.73, 6.12)	0.55	8.62 (3.05, 24.39)	0.25	3.30 (2.17, 5.01)	0.45	15.34 (5.32, 44.25)	0.38
1 vs. 4 (CNV)	2.71 (1.72, 4.27)	0.40	4.42 (1.42, 13.74)	0.13	2.44 (1.53, 3.90)	0.34	6.58 (2.07, 20.90)	0.19
1 vs. 5 (CNV)	7.21 (4.24, 12.27)	0.71	14.44 (4.96, 41.99)	0.37	5.24 (3.02, 9.10)	0.60	35.22 (11.47, 108.17)	0.60
CHS	1.93 (1.32, 2.83)	0.27	3.91 (2.17, 7.03)	0.11	1.58 (1.05, 2.39)	0.17	4.75 (2.56, 8.80)	0.14

NOTE - N denotes the normal allele and R denotes the risk allele. The risk allele is defined as the least frequent allele in controls. The OR for dominance effects compares those who carry one risk allele (RN and RR genotypes) to individuals homozygote for the normal allele (NN), the OR for recessive effects compares individuals with RR genotype to those who carry one normal allele (NN and RN genotypes). Hetero- and homozygote ORs compare individuals with one (RN) and two (RR) risk alleles to individuals with NN genotype, respectively. GA = geographic atrophy. CNV = choroidal neovascular membranes.

Table B3: Estimated ORs, corresponding 95% CIs, and PARs, adjusted for age and gender

Gene (Variant) and Comparison in AREDS or CHS	DOMINANCE (RN+RR vs. NN)		RECESSIVE (RR vs. RN+NN)		HETEROZYGOTES (RN vs. NN)		HOMOZYGOTES (RR vs. NN)	
	OR <sub>dom</sub> (95% CI)	PAR	OR <sub>rec</sub> (95% CI)	PAR	OR <sub>het</sub> (95% CI)	PAR	OR <sub>hom</sub> (95% CI)	PAR
<i>CFH</i> (Y402H)								
1 vs. 345	3.73 (2.60, 5.34)	0.60	3.69 (2.37, 5.75)	0.22	2.66 (1.81, 3.92)	0.43	6.69 (4.08, 10.98)	0.37
1 vs. 45	3.94 (2.72, 5.71)	0.62	3.74 (2.39, 5.85)	0.22	2.82 (1.89, 4.19)	0.45	7.06 (4.27, 11.70)	0.38
1 vs. 3	2.73 (1.55, 4.83)	0.49	3.39 (1.89, 6.10)	0.20	1.93 (1.04, 3.60)	0.30	4.95 (2.46, 9.95)	0.29
1 vs. 4	3.64 (2.35, 5.64)	0.59	3.48 (2.14, 5.66)	0.20	2.67 (1.67, 4.27)	0.43	6.33 (3.60, 11.16)	0.35
1 vs. 5	4.21 (2.76, 6.42)	0.64	3.95 (2.47, 6.32)	0.23	2.94 (1.87, 4.63)	0.47	7.71 (4.46, 13.34)	0.41
1 vs. 45 (GA)	3.73 (2.21, 6.31)	0.60	4.01 (2.36, 6.82)	0.24	2.54 (1.44, 4.48)	0.41	7.04 (3.69, 13.41)	0.38
1 vs. 4 (GA)	2.71 (1.36, 5.37)	0.49	4.16 (2.14, 8.07)	0.24	1.68 (0.78, 3.61)	0.23	5.55 (2.48, 12.41)	0.32
1 vs. 5 (GA)	4.85 (2.46, 9.56)	0.68	3.91 (2.16, 7.10)	0.23	3.47 (1.69, 7.14)	0.53	8.65 (3.92, 19.09)	0.44
1 vs. 45 (CNV)	3.31 (2.16, 5.07)	0.56	3.24 (2.00, 5.26)	0.19	2.48 (1.57, 3.93)	0.40	5.60 (3.21, 9.78)	0.32
1 vs. 4 (CNV)	3.43 (2.05, 5.74)	0.57	2.80 (1.63, 4.80)	0.16	2.78 (1.61, 4.80)	0.44	5.24 (2.74, 10.01)	0.30
1 vs. 5 (CNV)	3.18 (1.87, 5.41)	0.55	3.82 (2.21, 6.59)	0.22	2.17 (1.22, 3.86)	0.34	6.00 (3.12, 11.53)	0.34
CHS	2.26 (1.45, 3.53)	0.41	2.99 (1.85, 4.83)	0.17	1.82 (1.13, 2.92)	0.27	4.22 (2.39, 7.42)	0.25
<i>ELOVL4</i> (M299V)								
1 vs. 345	0.69 (0.47, 1.01)	-0.07	0.78 (0.28, 2.16)	-0.01	0.69 (0.46, 1.02)	-0.06	0.72 (0.26, 1.99)	-0.01
1 vs. 45	0.71 (0.48, 1.04)	-0.06	0.85 (0.30, 2.38)	0.00	0.70 (0.46, 1.04)	-0.06	0.78 (0.28, 2.19)	-0.01
1 vs. 3	0.59 (0.32, 1.09)	-0.09	0.35 (0.04, 3.04)	-0.02	0.62 (0.33, 1.17)	-0.07	0.31 (0.04, 2.75)	-0.02
1 vs. 4	0.67 (0.43, 1.06)	-0.07	0.76 (0.23, 2.53)	-0.01	0.67 (0.42, 1.07)	-0.06	0.69 (0.21, 2.32)	-0.01
1 vs. 5	0.73 (0.48, 1.12)	-0.06	0.92 (0.30, 2.80)	0.00	0.72 (0.46, 1.12)	-0.05	0.85 (0.28, 2.60)	0.00
1 vs. 45 (GA)	0.59 (0.35, 1.00)	-0.09	0.91 (0.24, 3.47)	0.00	0.56 (0.32, 0.99)	-0.08	0.81 (0.21, 3.08)	0.00
1 vs. 4 (GA)	0.63 (0.31, 1.29)	-0.08	1.13 (0.21, 5.99)	0.00	0.58 (0.27, 1.26)	-0.08	1.00 (0.19, 5.36)	0.00
1 vs. 5 (GA)	0.56 (0.30, 1.06)	-0.10	0.77 (0.15, 4.04)	-0.01	0.55 (0.28, 1.07)	-0.09	0.67 (0.13, 3.57)	-0.01
1 vs. 45 (CNV)	0.65 (0.42, 1.01)	-0.07	0.36 (0.09, 1.55)	-0.02	0.69 (0.43, 1.09)	-0.06	0.33 (0.08, 1.42)	-0.02
1 vs. 4 (CNV)	0.71 (0.42, 1.18)	-0.06	0.67 (0.16, 2.84)	-0.01	0.72 (0.42, 1.22)	-0.05	0.62 (0.14, 2.64)	-0.01
1 vs. 5 (CNV)	0.58 (0.33, 1.03)	-0.09	...	...	0.65 (0.37, 1.15)	-0.06	...	...
CHS	1.41 (0.92, 2.17)	0.07	0.33 (0.04, 2.43)	-0.02	1.55 (1.00, 2.41)	0.09	0.36 (0.05, 2.68)	-0.02
<i>PLEKHAI</i> (A320T)								
1 vs. 345	0.57 (0.40, 0.81)	-0.39	0.37 (0.23, 0.60)	-0.13	0.68 (0.47, 0.99)	-0.18	0.31 (0.18, 0.51)	-0.14
1 vs. 45	0.56 (0.39, 0.81)	-0.40	0.39 (0.24, 0.63)	-0.12	0.67 (0.46, 0.98)	-0.19	0.31 (0.19, 0.53)	-0.14
1 vs. 3	0.62 (0.37, 1.05)	-0.33	0.28 (0.11, 0.70)	-0.15	0.78 (0.46, 1.35)	-0.12	0.25 (0.09, 0.64)	-0.16
1 vs. 4	0.68 (0.46, 1.02)	-0.26	0.62 (0.37, 1.04)	-0.07	0.75 (0.49, 1.15)	-0.13	0.53 (0.30, 0.94)	-0.09
1 vs. 5	0.48 (0.33, 0.71)	-0.51	0.22 (0.12, 0.42)	-0.16	0.61 (0.41, 0.92)	-0.23	0.17 (0.09, 0.34)	-0.17
1 vs. 45 (GA)	0.70 (0.44, 1.12)	-0.24	0.42 (0.21, 0.83)	-0.12	0.84 (0.52, 1.37)	-0.08	0.38 (0.18, 0.80)	-0.12
1 vs. 4 (GA)	0.66 (0.36, 1.21)	-0.29	0.57 (0.24, 1.38)	-0.08	0.73 (0.38, 1.40)	-0.15	0.48 (0.19, 1.24)	-0.10
1 vs. 5 (GA)	0.74 (0.43, 1.27)	-0.20	0.32 (0.13, 0.79)	-0.14	0.92 (0.53, 1.63)	-0.04	0.30 (0.12, 0.80)	-0.14
1 vs. 45 (CNV)	0.50 (0.33, 0.74)	-0.49	0.45 (0.26, 0.78)	-0.11	0.57 (0.37, 0.87)	-0.26	0.34 (0.19, 0.61)	-0.13
1 vs. 4 (CNV)	0.65 (0.41, 1.02)	-0.30	0.68 (0.37, 1.24)	-0.06	0.69 (0.42, 1.12)	-0.18	0.56 (0.29, 1.07)	-0.09
1 vs. 5 (CNV)	0.37 (0.23, 0.59)	-0.71	0.21 (0.08, 0.51)	-0.16	0.46 (0.28, 0.76)	-0.35	0.14 (0.06, 0.36)	-0.18
CHS	0.76 (0.52, 1.11)	-0.19	0.68 (0.39, 1.18)	-0.06	0.81 (0.54, 1.21)	-0.10	0.61 (0.34, 1.10)	-0.07
<i>LOC387715</i> (S69A)								
1 vs. 345	3.99 (2.81, 5.67)	0.54	10.16 (3.70, 27.88)	0.28	3.06 (2.13, 4.39)	0.42	17.26 (6.22, 47.89)	0.41
1 vs. 45	4.17 (2.92, 5.96)	0.56	10.52 (3.83, 28.93)	0.29	3.18 (2.20, 4.60)	0.43	18.30 (6.57, 50.93)	0.43
1 vs. 3	3.07 (1.82, 5.17)	0.45	7.97 (2.56, 24.81)	0.23	2.45 (1.42, 4.23)	0.34	11.89 (3.70, 38.19)	0.32
1 vs. 4	2.72 (1.83, 4.05)	0.41	5.64 (1.95, 16.27)	0.17	2.34 (1.55, 3.53)	0.32	8.19 (2.80, 24.00)	0.24
1 vs. 5	6.14 (4.11, 9.19)	0.67	15.07 (5.43, 41.82)	0.38	4.32 (2.85, 6.57)	0.54	32.07 (11.30, 91.01)	0.57
1 vs. 45 (GA)	3.29 (2.07, 5.21)	0.48	6.28 (2.09, 18.93)	0.19	2.81 (1.74, 4.52)	0.39	10.14 (3.28, 31.31)	0.28
1 vs. 4 (GA)	3.06 (1.66, 5.65)	0.45	4.75 (1.29, 17.49)	0.14	2.74 (1.46, 5.17)	0.38	7.57 (1.97, 29.06)	0.22
1 vs. 5 (GA)	3.46 (2.01, 5.94)	0.49	7.38 (2.33, 23.38)	0.22	2.86 (1.63, 5.02)	0.40	12.02 (3.65, 39.57)	0.32
1 vs. 45 (CNV)	4.09 (2.73, 6.12)	0.55	8.62 (3.05, 24.39)	0.25	3.30 (2.17, 5.01)	0.45	15.34 (5.32, 44.25)	0.38
1 vs. 4 (CNV)	2.71 (1.72, 4.27)	0.40	4.42 (1.42, 13.74)	0.13	2.44 (1.53, 3.90)	0.34	6.58 (2.07, 20.90)	0.19
1 vs. 5 (CNV)	7.21 (4.24, 12.27)	0.71	14.44 (4.96, 41.99)	0.37	5.24 (3.02, 9.10)	0.60	35.22 (11.47, 108.17)	0.60
CHS	1.93 (1.32, 2.83)	0.27	3.91 (2.17, 7.03)	0.11	1.58 (1.05, 2.39)	0.17	4.75 (2.56, 8.80)	0.14

NOTE - N denotes the normal allele and R denotes the risk allele. The risk allele is defined as the least frequent allele in controls. The OR for dominance effects compares those who carry one risk allele (RN and RR genotypes) to individuals homozygote for the normal allele (NN), the OR for recessive effects compares individuals with RR genotype to those who carry one normal allele (NN and RN genotypes). Hetero- and homozygote ORs compare individuals with one (RN) and two (RR) risk alleles to individuals with NN genotype, respectively. GA = geographic atrophy. CNV = choroidal neovascular membranes.

Table B4: Joint ORs and 95% CIs at *Y402H* in *CFH* and *S69A* in *LOC387715*

Analyzed cohort and size of sample	OR (95% CI) for				
	S69A	Main effects	Y402H		
			TT	CT	CC
<b>AREDS</b>					
$n_{\text{controls}} = 171$					
$n_{\text{cases}} = 693$		OR <sub>Y402H</sub>	1.00 (Ref)	2.70 (1.83, 3.98)	6.64 (4.04, 10.91)
		OR <sub>S69A</sub>	Joint effects		
	GG	1.00 (Ref)	1.00 (Ref)	2.82 (1.59, 5.03)	...
	GT	3.03 (2.11, 4.36)	3.17 (1.68, 5.96)	7.16 (3.80, 13.49)	...
	TT	17.11 (6.17, 47.47)	...	...	15.79 (8.74, 28.54) <sup>a</sup>
<b>CHS</b>					
$n_{\text{controls}} = 871$					
$n_{\text{cases}} = 106$		OR <sub>Y402H</sub>	1.00 (Ref)	1.81 (1.12, 2.93)	4.12 (2.32, 7.33)
		OR <sub>S69A</sub>	Joint effects		
	GG	1.00 (Ref)	1.00 (Ref)	1.31 (0.64, 2.69)	...
	GT	1.59 (1.03, 2.47)	1.22 (0.53, 2.83)	2.90 (1.47, 5.73)	...
	TT	4.86 (2.55, 9.26)	...	...	4.82 (2.52, 9.23) <sup>a</sup>

NOTE -  $n_{\text{controls}}$  = number of controls fully typed at both loci,  $n_{\text{cases}}$  = number of cases fully typed at both loci. OR<sub>Y402H</sub> =

OR for Y402H averaged across S69A genotypes, OR<sub>S69A</sub> = OR for S69A averaged across Y402H genotypes.

<sup>a</sup> OR for individuals homozygous at least at one of the loci.

Table B5: Joint genotype distribution at  $Y402H$  in  $CFH$  and  $S69A$  in  $LOC387715$  in the AREDS cohort

Genotype at $S69A$ in $LOC387715$	Genotype at $Y402H$ in $CFH$					
	Controls ( $n = 171$ )			Cases ( $n = 693$ )		
	TT	CT	CC	TT	CT	CC
GG	42 (0.246)	46 (0.269)	22 (0.129)	32 (0.046)	99 (0.143)	86 (0.124)
GT	29 (0.170)	24 (0.140)	4 (0.023)	70 (0.101)	131 (0.189)	140 (0.202)
TT	3 (0.018)	1 (0.006)	0 (0.000)	15 (0.022)	73 (0.105)	47 (0.068)

Table B6: Joint genotype distribution at  $Y402H$  in  $CFH$  and  $S69A$  in  $LOC387715$  in the CHS cohort

Genotype at $S69A$ in $LOC387715$	Genotype at $Y402H$ in $CFH$					
	Controls ( $n = 871$ )			Cases ( $n = 106$ )		
	TT	CT	CC	TT	CT	CC
GG	231 (0.265)	227 (0.261)	57 (0.065)	14 (0.132)	18 (0.170)	13 (0.123)
GT	135 (0.155)	148 (0.170)	33 (0.038)	10 (0.094)	26 (0.245)	8 (0.075)
TT	23 (0.026)	16 (0.018)	1 (0.001)	4 (0.038)	7 (0.066)	6 (0.057)



Table B7: Joint ORs and 95% CIs at *Y402H* in *CFH* and smoking, and *S69A* in *LOC387715* and smoking

<b>AREDS cohort</b>			
OR (95% CI) for			
Gene (Variant) and Genotypes	Main effects	Smoking history	
		Never	Ever
<i>CFH</i> (Y402H)			
$n_{\text{controls}} = 170$	OR <sub>smk</sub>	1.00 (Ref)	1.59 (1.13, 2.23)
$n_{\text{cases}} = 682$			
	OR <sub>Y402H</sub>	Joint effects	
TT	1.00 (Ref)	1.00 (Ref)	1.65 (0.91, 2.98)
CT	2.65 (1.79, 3.90)	2.53 (1.43, 4.48)	4.77 (2.66, 8.54)
CC	7.27 (4.37, 12.09)	8.65 (4.03, 18.55)	10.55 (5.14, 21.66)
<i>LOC387715</i> (S69A)			
$n_{\text{controls}} = 169$	OR <sub>smk</sub>	1.00 (Ref)	1.57 (1.12, 2.20)
$n_{\text{cases}} = 676$			
	OR <sub>S69A</sub>	Joint effects	
GG	1.00 (Ref)	1.00 (Ref)	1.77 (1.11, 2.83)
GT	2.98 (2.07, 4.29)	3.19 (1.87, 5.41)	5.06 (2.99, 8.55)
TT	17.02 (6.13, 47, 26)	21.15 (4.96, 90.22)	25.74 (6.06, 109.34)
<b>CHS cohort</b>			
OR (95% CI) for			
Gene (Variant) and Genotypes	Main effects	Smoking history	
		Never	Ever
<i>CFH</i> (Y402H)			
$n_{\text{controls}} = 907$	OR <sub>smk</sub>	1.00 (Ref)	0.89 (0.60, 1.32)
$n_{\text{cases}} = 110$			
	OR <sub>Y402H</sub>	Joint effects	
TT	1.00 (Ref)	1.00 (Ref)	0.62 (0.29, 1.33)
CT	1.82 (1.13, 2.92)	1.52 (0.80, 2.91)	1.33 (0.69, 2.58)
CC	4.22 (2.39, 7.42)	2.52 (1.10, 5.79)	4.16 (1.95, 8.86)
<i>LOC387715</i> (S69A)			
$n_{\text{controls}} = 995$	OR <sub>smk</sub>	1.00 (Ref)	0.89 (0.61, 1.30)
$n_{\text{cases}} = 120$			
	OR <sub>S69A</sub>	Joint effects	
GG	1.00 (Ref)	1.00 (Ref)	0.96 (0.55, 1.68)
GT	1.58 (1.05, 2.39)	1.86 (1.04, 3.31)	1.28 (0.70, 2.34)
TT	4.75 (2.56, 8.80)	3.37 (1.32, 8.63)	6.09 (2.63, 14.14)

Table B8: Genotype distribution at *Y402H* in *CFH* and *S69A* in *LOC387715* in the AREDS cohort, by smoking history (ever vs. never smoked)

Gene (Variant) and Genotypes	Smoking history and ARM status			
	Never	Ever	Never	Ever
<i>CFH</i> ( <i>Y402H</i> )	Controls ( <i>n</i> = 170)		Cases ( <i>n</i> = 682)	
TT	36 (0.212)	38 (0.224)	42 (0.062)	73 (0.107)
CT	40 (0.235)	32 (0.188)	118 (0.173)	178 (0.261)
CC	11 (0.065)	13 (0.076)	111 (0.163)	160 (0.235)
<i>LOC387715</i> ( <i>S69A</i> )	Controls ( <i>n</i> = 169)		Cases ( <i>n</i> = 676)	
GG	55 (0.325)	53 (0.314)	78 (0.115)	133 (0.197)
GT	29 (0.172)	28 (0.166)	131 (0.194)	201 (0.297)
TT	2 (0.012)	2 (0.012)	60 (0.089)	73 (0.108)

Table B9: Genotype distribution at *Y402H* in *CFH* and *S69A* in *LOC387715* in the CHS cohort, by smoking history (ever vs. never smoked)

Gene (Variant) and Genotypes	Smoking history and ARM status			
	Never	Ever	Never	Ever
<i>CFH</i> ( <i>Y402H</i> )	Controls ( <i>n</i> = 907)		Cases ( <i>n</i> = 110)	
TT	176 (0.194)	230 (0.254)	16 (0.145)	13 (0.118)
CT	202 (0.223)	206 (0.227)	28 (0.255)	25 (0.227)
CC	48 (0.053)	45 (0.050)	11 (0.100)	17 (0.155)
<i>LOC387715</i> ( <i>S69A</i> )	Controls ( <i>n</i> = 995)		Cases ( <i>n</i> = 120)	
GG	277 (0.278)	324 (0.326)	25 (0.208)	28 (0.223)
GT	161 (0.162)	190 (0.191)	27 (0.225)	22 (0.183)
TT	23 (0.023)	20 (0.020)	7 (0.058)	11 (0.092)

Table B10: Characteristics of studies included in meta-analysis of *Y402H* in *CFH*

Study and sample	Sample size <sup>a</sup>	Mean age (±SD) <sup>b</sup>	% Males	HWE <sup>c</sup> <i>P</i> -value	Frequency of the C allele
Edwards et al. <sup>d</sup>					
Discovery sample					
Controls	131	67.6 (7.6)	42	0.99	0.340
Cases	225	72.7 (10.1)	58	0.42	0.553
Replication sample					
Controls	59	68.1 (9.0)	35	0.28	0.390
Cases	170	78.2 (7.9)	65	0.64	0.544
Haines et al. <sup>e</sup>					
Controls	185	≥ 55	...	...	...
Cases	495	≥ 55	...	...	...
Zarepars et al.					
Controls	275	≥ 68	...	0.11	0.338
Cases	616	...	...	0.15	0.608
Hageman et al. <sup>f</sup>					
Columbia sample					
Controls	272	68.8 (8.6)	...	0.23	0.344
Cases	549	71.3 (8.9)	...	0.86	0.538
Iowa sample					
Controls	131	78.4 (7.4)	...	0.70	0.336
Cases	403	79.5 (7.8)	...	0.22	0.589
Jakobsdottir et al.					
Controls	108	72.6 (8.9)	47	0.26	0.310
Cases	434	68.9 (8.8)	39	0.42	0.613
Rivera et al. <sup>g</sup>					
Original sample					
Controls	611	76.2 (5.3)	38	<0.01	0.382
Cases	793	76.3 (6.9)	36	0.30	0.595
Replication sample					
Controls	335	68.3 (8.1)	45	0.48	0.358
Cases	373	75.0 (7.5)	35	0.13	0.617
Souied et al.					
Controls	91	74.6 (6.3)	42	0.21	0.302
Cases	141	74.3 (8.0)	38	0.30	0.564
Sepp et al.					
Controls	262	75.8 (7.8)	42	0.14	0.363
Cases	443	80.3 (6.9)	45	0.49	0.607
AREDS (1 vs. 345)					
Controls	173	76.5 (4.4)	49	0.25	0.358
Cases	699	79.5 (5.2)	42	0.03	0.612
CHS					
Controls	907	70.3 (3.9)	43	0.55	0.327
Cases	110	73.2 (4.8)	44	0.71	0.495

<sup>a</sup> Sample sizes based on total number of genotyped persons when genotype counts are available other wise on total sample size, not accounting for missing data.

<sup>b</sup> Mean age and corresponding standard deviation, or other summary statistic available from the original paper.

<sup>c</sup> When genotype counts are available *P*-value, derived from the exact test (implemented in R Genetics package), given.

<sup>d</sup> The two data sets of Edwards et al. paper are combined in the meta-analysis. HWE *P*-values for the combined controls and cases are 0.53 and 0.36, respectively.

<sup>e</sup> Results from Haines et al. paper are included in meta-analysis of ORs for hetero- and homozygous individuals. Sample sizes are based on total number of individuals, not accounting for missing genotype data at *Y402H* in *CFH*.

<sup>f</sup> The two data sets of Hageman et al. paper are not combined, following the original paper.

<sup>g</sup> The two data sets of Rivera et al. paper are combined in the meta-analysis. HWE *P*-values for the combined controls and cases are 0.03 and 0.09, respectively.

Table B11: Results of meta-analysis of *Y402H* in *CFH*. ORs (95% CIs) estimated from individual studies and all studies pooled. Results of leave-one-out sensitivity analysis are shown

OR for	DOMINANCE (CT+CC vs. TT)		RECESSIVE (CC vs. CT+TT)		HETEROZYGOTES (CT vs. TT)		HOMOZYGOTES (CC vs. TT)	
	OR <sub>dom</sub> (95% CI)		OR <sub>rec</sub> (95% CI)		OR <sub>het</sub> (95% CI)		OR <sub>hom</sub> (95% CI)	
<b>Individual study</b>								
Edwards et al.	2.71 (1.86, 3.94)		2.89 (1.82, 4.60)		2.14 (1.44, 3.18)		4.54 (2.70, 7.65)	
Haines et al.	...		...		2.45 (1.41, 4.25)		3.33 (1.79, 6.20)	
Zarepari et al.	4.36 (3.13, 6.08)		5.52 (3.54, 8.59)		3.03 (2.15, 4.28)		11.61 (7.05, 19.14)	
Hageman et al. (Columbia)	2.97 (2.17, 4.07)		2.61 (1.76, 3.87)		2.48 (1.77, 3.47)		4.47 (2.89, 6.93)	
Hageman et al. (Iowa)	3.64 (2.38, 5.58)		4.08 (2.33, 7.16)		2.61 (1.66, 4.10)		7.28 (3.92, 13.51)	
Jakobsdottir et al.	5.29 (3.35, 8.35)		4.57 (2.48, 8.42)		3.78 (2.32, 6.17)		10.05 (5.16, 19.59)	
Rivera et al.	2.92 (2.39, 3.57)		4.29 (3.42, 5.39)		1.99 (1.61, 2.46)		6.72 (5.14, 8.79)	
Souied et al.	3.95 (2.22, 7.03)		3.75 (1.83, 7.71)		2.99 (1.61, 5.57)		6.84 (3.07, 15.21)	
Sepp et al.	3.85 (2.71, 5.47)		3.36 (2.28, 4.95)		2.88 (1.98, 4.20)		6.49 (4.12, 10.23)	
AREDS (1 vs. 345)	3.73 (2.60, 5.34)		3.69 (2.37, 5.75)		2.66 (1.81, 3.92)		6.69 (4.08, 10.98)	
CHS	2.26 (1.45, 3.53)		2.99 (1.85, 4.83)		1.82 (1.13, 2.92)		4.22 (2.39, 7.42)	
<b>All studies pooled</b>								
Fixed effects	3.33 (2.99, 3.71)	$P^a$	3.75 (3.29, 4.27)	$P^a$	2.43 (2.17, 2.72)	$P^a$	6.22 (5.38, 7.19)	$P^a$
Random effects	3.40 (2.88, 4.00)	...	3.70 (3.09, 4.42)	...	2.49 (2.14, 2.89)	...	6.15 (4.86, 7.79)	...
<b>Study excluded (Fixed effects)</b>								
		$\Delta^b$		$\Delta^b$		$\Delta^b$		$\Delta^b$
Edwards et al.	3.39 (3.03, 3.80)	-0.06	3.83 (3.35, 4.39)	-0.08	2.46 (2.19, 2.77)	-0.03	6.39 (5.49, 7.42)	-0.17
Haines et al.	...	...	...	...	2.43 (2.17, 2.73)	0.00	6.45 (5.56, 7.48)	-0.23
Zarepari et al.	3.22 (2.87, 3.61)	0.11	3.62 (3.16, 4.14)	0.13	2.37 (2.10, 2.67)	0.06	5.88 (5.05, 6.83)	0.35
Hageman et al. (Columbia)	3.38 (3.01, 3.79)	-0.05	3.92 (3.42, 4.50)	-0.17	2.43 (2.15, 2.74)	0.01	6.48 (5.56, 7.55)	-0.26
Hageman et al. (Iowa)	3.31 (2.96, 3.70)	0.02	3.73 (3.27, 4.26)	0.02	2.42 (2.16, 2.72)	0.01	6.16 (5.31, 7.15)	0.06
Jakobsdottir et al.	3.24 (2.89, 3.62)	0.09	3.72 (3.25, 4.24)	0.03	2.37 (2.11, 2.67)	0.06	6.08 (5.24, 7.05)	0.15
Rivera et al.	3.51 (3.09, 3.99)	-0.18	3.52 (3.00, 4.12)	0.23	2.63 (2.30, 3.00)	-0.20	6.03 (5.08, 7.16)	0.19
Souied et al.	3.31 (2.96, 3.69)	0.02	3.75 (3.29, 4.28)	0.00	2.42 (2.15, 2.71)	0.02	6.20 (5.35, 7.18)	0.02
Sepp et al.	3.28 (2.92, 3.67)	0.05	3.80 (3.31, 4.36)	-0.05	2.39 (2.13, 2.69)	0.04	6.19 (5.32, 7.21)	0.03
AREDS (1 vs. 345)	3.29 (2.94, 3.68)	0.04	3.76 (3.28, 4.30)	-0.01	2.41 (2.14, 2.72)	0.02	6.18 (5.31, 7.19)	0.04
CHS	3.41 (3.05, 3.81)	-0.08	3.82 (3.34, 4.37)	-0.07	2.48 (2.20, 2.78)	-0.04	6.39 (5.50, 7.42)	-0.17

<sup>a</sup>  $P$ -value for test of homogeneity of ORs across studies.

<sup>b</sup> Difference ( $\Delta$ ) of pooled point estimate when a study is excluded from the pooled estimate of all studies (under fixed effects model)

Table B12: Characteristics of studies included in meta-analysis of *S69A* in *LOC387715*

Study and sample	Sample size <sup>a</sup>	Mean age (±SD) <sup>b</sup>	% Males	HWE <sup>c</sup> <i>P</i> -value	Frequency of the T allele
Jakobsdottir et al.					
Controls	106	72.6 (8.9)	47	0.21	0.193
Cases	456	68.9 (8.8)	39	0.06	0.485
Rivera et al. <sup>d</sup>					
Original sample					
Controls	594	76.2 (5.3)	38	0.30	0.196
Cases	759	76.3 (6.9)	36	0.14	0.417
Replication sample					
Controls	328	68.3 (8.1)	45	0.75	0.215
Cases	361	75.0 (7.5)	35	0.01	0.460
Schmidt et al. <sup>e</sup>					
Controls	186	66.7 (8.1)	43	0.55	0.247
Cases	758	76.8 (7.7)	35	<0.01	0.427
AREDS (1 vs. 345)					
Controls	172	76.5 (4.4)	49	0.45	0.189
Cases	693	79.5 (5.2)	42	0.99	0.441
CHS					
Controls	995	70.3 (3.9)	43	0.41	0.220
Cases	120	73.2 (4.8)	44	0.24	0.354

<sup>a</sup> Sample sizes based on total number of genotyped persons.

<sup>b</sup> Mean age and corresponding standard deviation, or other summary statistic available from the original paper.

<sup>c</sup> When genotype counts are available *P*-value, derived from the exact test (implemented in R Genetics package), given.

<sup>d</sup> The two data sets of Rivera et al. paper are combined in the meta-analysis. HWE *P*-values for the combined controls and cases are 0.31 and 0.01, respectively.

<sup>e</sup> In the meta-analysis only grade 1 subjects are classified as controls (grade 2 subjects are dropped). The original study by Schmidt et al. classified grade 2 individuals as controls. The mean age and % males of controls is taken from the paper and based on both grade 1 and 2.

Table B13: Results of meta-analysis of *S69A* in *LOC387715*. ORs (95% CIs) estimated from individual studies and all studies pooled. Results of leave-one-out sensitivity analysis are shown

OR for	DOMINANCE (GT+TT vs. GG)		RECESSIVE (TT vs. GT+GG)		HETEROZYGOTES (GT vs. GG)		HOMOZYGOTES (TT vs. GG)	
	OR <sub>dom</sub> (95% CI)	<i>P</i> <sup>a</sup>	OR <sub>rec</sub> (95% CI)	<i>P</i> <sup>a</sup>	OR <sub>het</sub> (95% CI)	<i>P</i> <sup>a</sup>	OR <sub>hom</sub> (95% CI)	<i>P</i> <sup>a</sup>
Individual study								
Jakobsdottir et al.	5.03 (3.20, 7.91)		5.75 (2.46, 13.46)		3.89 (2.40, 6.31)		10.57 (4.43, 25.22)	
Rivera et al.	3.41 (2.84, 4.09)		5.28 (3.76, 7.41)		2.69 (2.22, 3.27)		8.21 (5.79, 11.65)	
Schmidt et al. (1 vs. 345)	2.42 (1.75, 3.35)		3.59 (1.99, 6.47)		1.94 (1.37, 2.74)		4.87 (2.65, 8.95)	
AREDS (1 vs. 345)	3.99 (2.81, 5.67)		10.16 (3.70, 27.88)		3.06 (2.13, 4.39)		17.26 (6.22, 47.89)	
CHS	1.93 (1.32, 2.83)		3.91 (2.17, 7.03)		1.58 (1.05, 2.39)		4.75 (2.56, 8.80)	
All studies pooled								
Fixed effects	3.19 (2.80, 3.63)	<0.01	4.91 (3.85, 6.27)	0.41	2.53 (2.20, 2.90)	0.02	7.32 (5.69, 9.42)	0.11
Random effects	3.15 (2.02, 4.90)	...	4.91 (3.48, 6.94)	...	2.48 (1.67, 3.70)	...	7.33 (4.33, 12.42)	...
Study excluded (Fixed effects)								
		$\Delta^b$		$\Delta^b$		$\Delta^b$		$\Delta^b$
Jakobsdottir et al.	3.06 (2.67, 3.51)	0.13	4.84 (3.76, 6.24)	0.07	2.43 (2.11, 2.81)	0.09	7.08 (5.44, 9.21)	0.24
Rivera et al.	2.98 (2.48, 3.58)	0.21	4.54 (3.20, 6.45)	0.37	2.37 (1.95, 2.88)	0.16	6.48 (4.51, 9.31)	0.84
Schmidt et al. (1 vs. 345)	3.36 (2.92, 3.87)	-0.17	5.24 (4.01, 6.85)	-0.33	2.66 (2.29, 3.08)	-0.13	7.97 (6.04, 10.51)	-0.65
AREDS (1 vs. 345)	3.08 (2.68, 3.54)	0.11	4.70 (3.65, 6.04)	0.22	2.45 (2.11, 2.84)	0.08	6.93 (5.34, 8.98)	0.40
CHS	3.41 (2.97, 3.91)	-0.22	5.15 (3.94, 6.73)	-0.24	2.68 (2.32, 3.10)	-0.15	7.99 (6.06, 10.52)	-0.66

<sup>a</sup> *P*-value for test of homogeneity of ORs across studies.

<sup>b</sup> Difference ( $\Delta$ ) of pooled point estimate when a study is excluded from the pooled estimate of all studies (under fixed effects model)

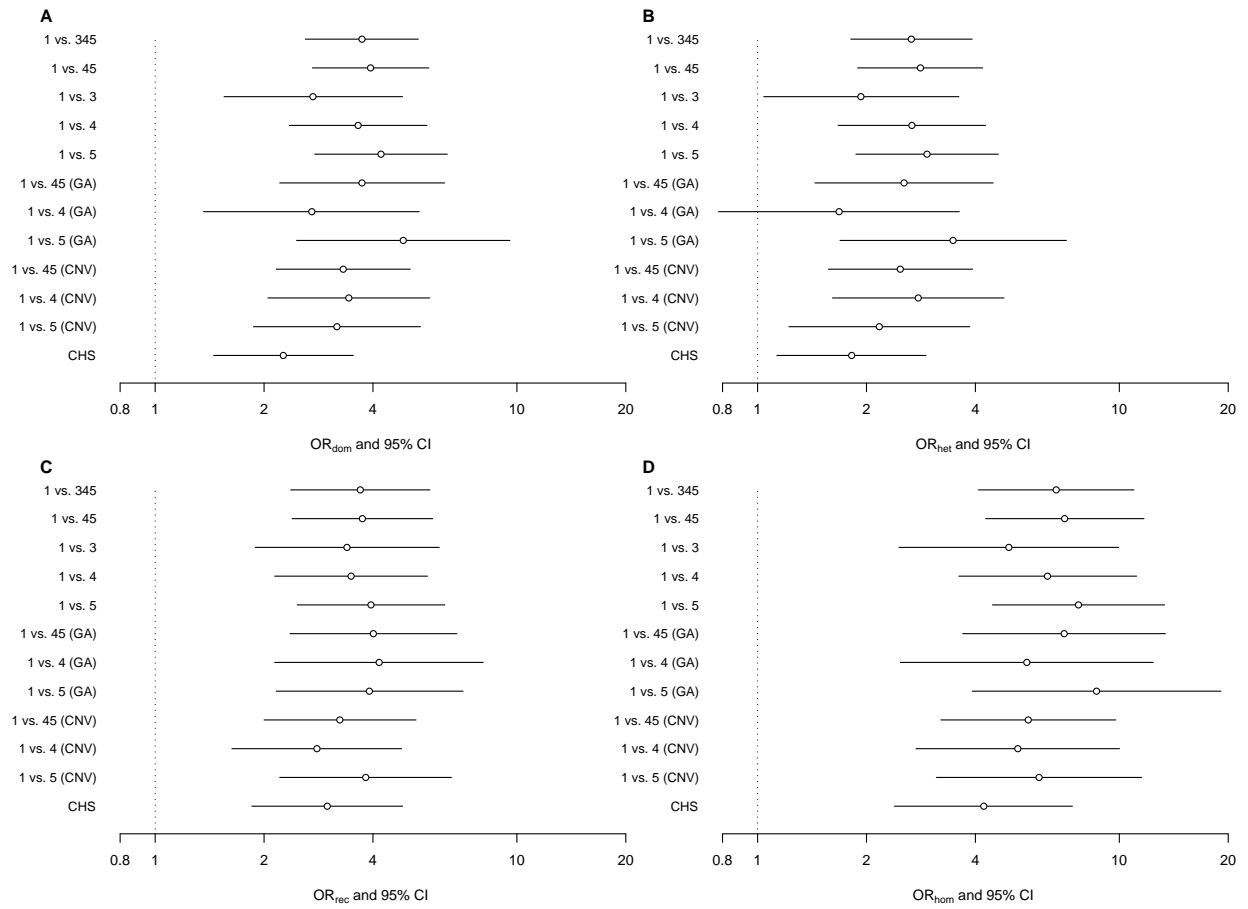


Figure B1: Estimated ORs and 95% CIs for *CFH*. A:  $OR_{dom}$  for evaluation of dominance effects (CT+CC vs. TT). B:  $OR_{het}$  for evaluation of the risk for heterozygotes (CT vs. TT). C:  $OR_{rec}$  for evaluation of recessive effects (CC vs. CT+TT). D:  $OR_{hom}$  for evaluation of the risk for homozygotes (CC vs. TT). The dotted vertical line marks the null value of OR of 1.

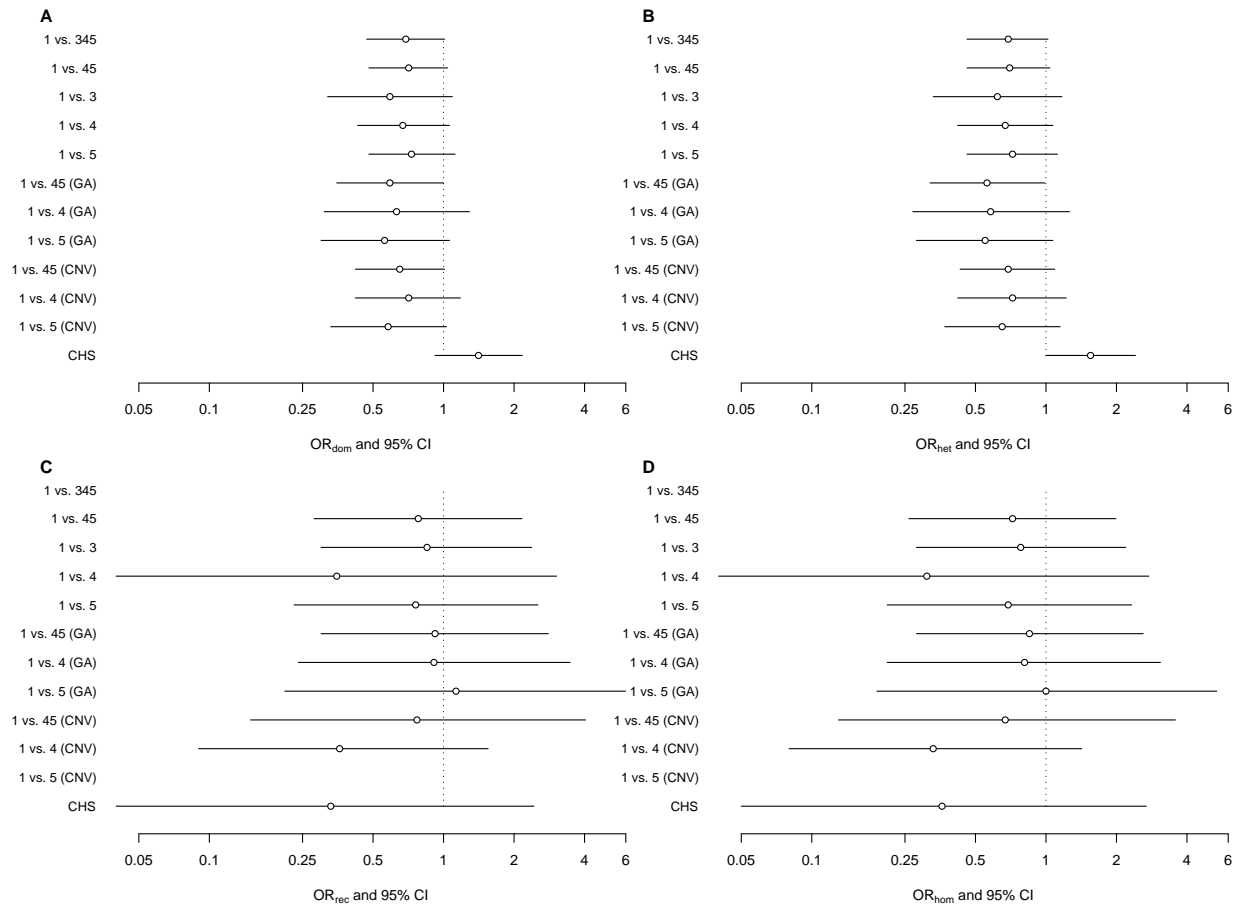


Figure B2: Estimated ORs and 95% CIs for *ELOVL4*. A: OR<sub>dom</sub> for evaluation of dominance effects (AG+GG vs. AA). B: OR<sub>het</sub> for evaluation of the risk for heterozygotes (AG vs. AA). C: OR<sub>rec</sub> for evaluation of recessive effects (GG vs. AG+AA). D: OR<sub>hom</sub> for evaluation of the risk for homozygotes (GG vs. AA). The dotted vertical line marks the null value of OR of 1.



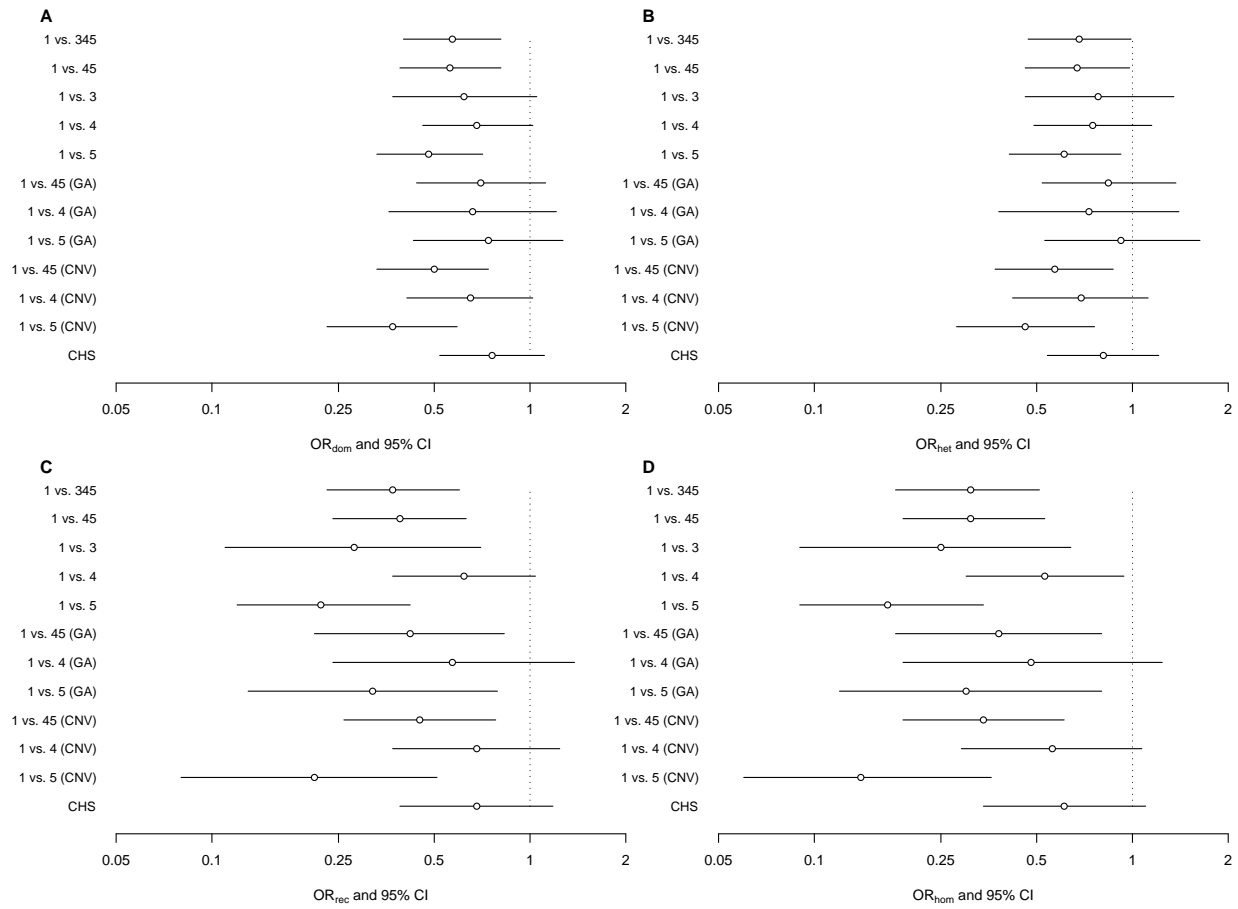


Figure B3: Estimated ORs and 95% CIs for *PLEKHA1*. A: OR<sub>dom</sub> for evaluation of dominance effects (AG+AA vs. GG). B: OR<sub>het</sub> for evaluation of the risk for heterozygotes (AG vs. GG). C: OR<sub>rec</sub> for evaluation of recessive effects (AA vs. AG+GG). D: OR<sub>hom</sub> for evaluation of the risk for homozygotes (AA vs. GG). The dotted vertical line marks the null value of OR of 1.

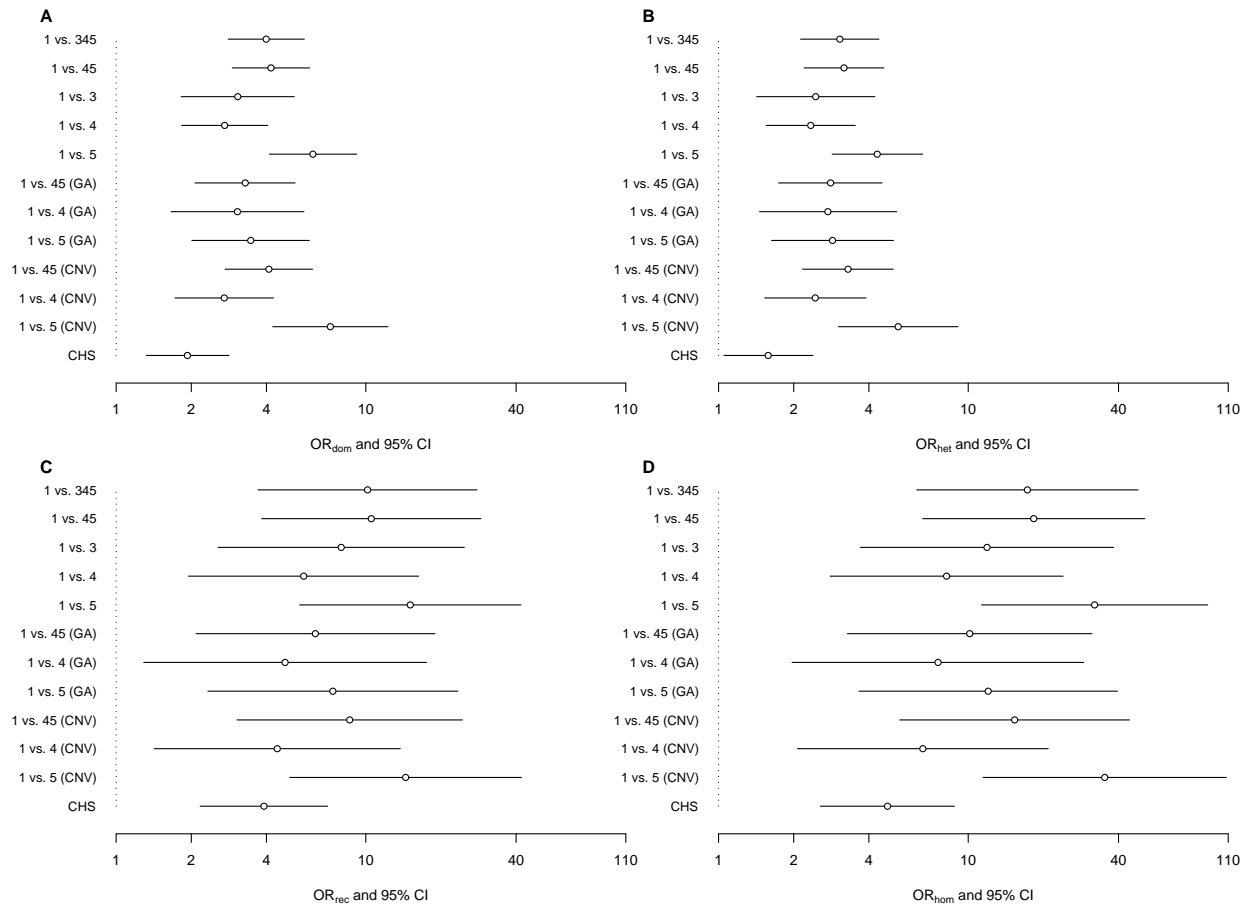


Figure B4: Estimated ORs and 95% CIs for *LOC387715*. A:  $OR_{dom}$  for evaluation of dominance effects (GT+TT vs. GG). B:  $OR_{het}$  for evaluation of the risk for heterozygotes (GT vs. GG). C:  $OR_{rec}$  for evaluation of recessive effects (TT vs. GT+GG). D:  $OR_{hom}$  for evaluation of the risk for homozygotes (TT vs. GG). The dotted vertical line marks the null value of OR of 1.

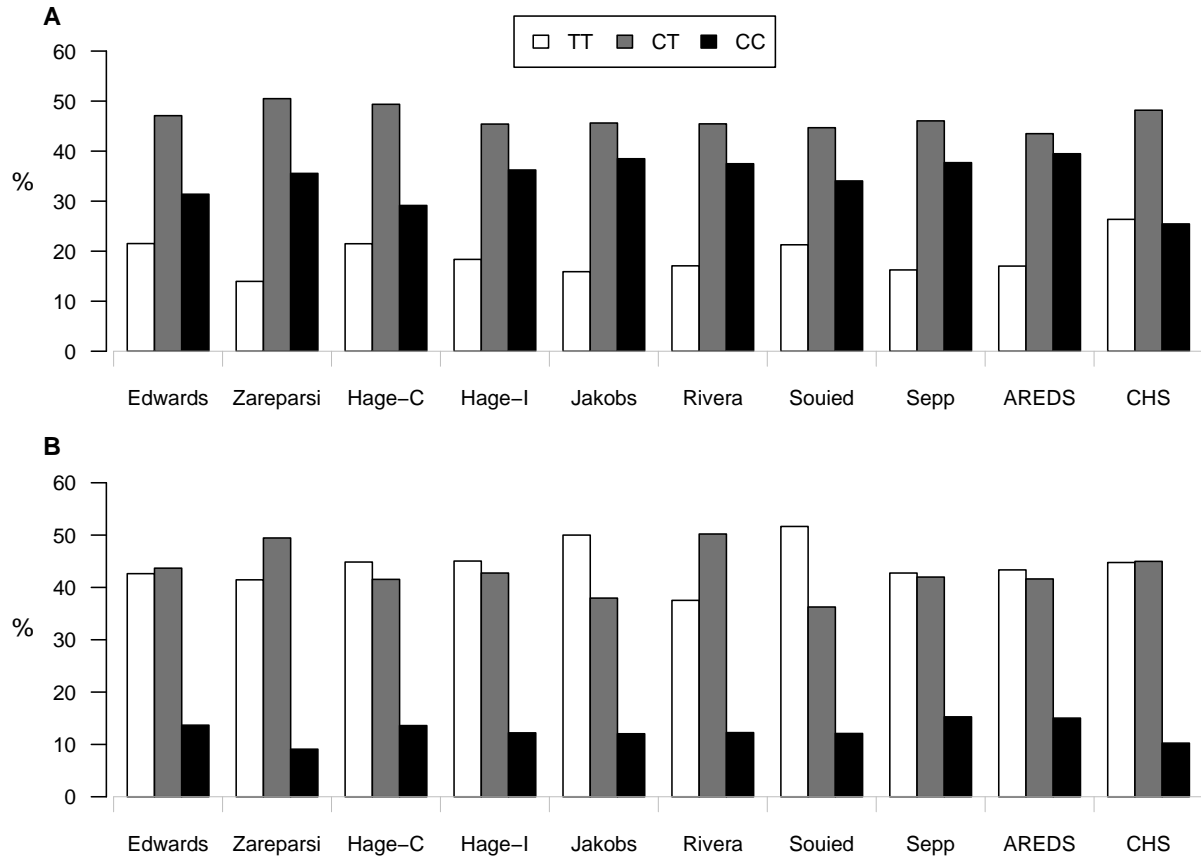


Figure B5: A: Genotype frequencies (%) in unrelated ARM cases, across cohorts included in meta-analysis of Y402H in *CFH*. B: Genotype frequencies (%) in unrelated controls without ARM, across studies included in meta-analysis of Y402H in *CFH*. ‘Hage-C’ and ‘Hage-I’ denote estimates derived from the Columbia and Iowa cohorts of Hageman et al., respectively, and ‘Jakobs’ denotes estimates from the Jakobsdottir et al. paper. Frequencies were not available from Haines et al.

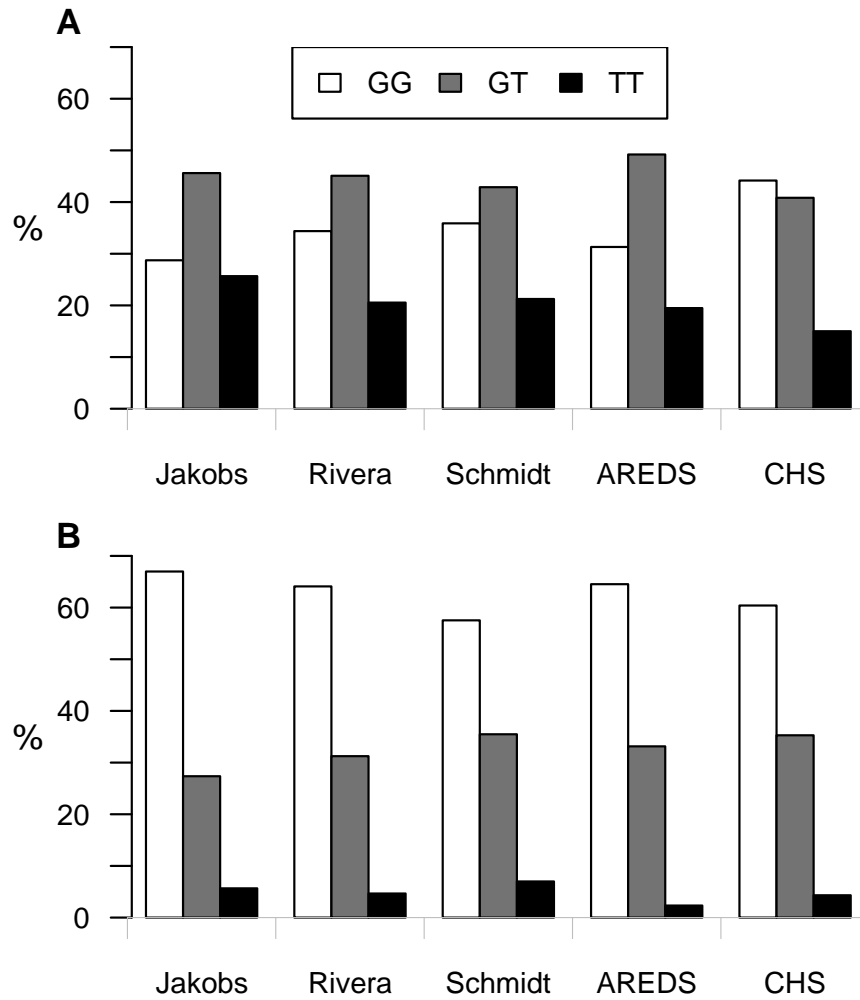


Figure B6: A: Genotype frequencies (%) in unrelated ARM cases, across cohorts included in meta-analysis of S69A in *LOC387715*. B: Genotype frequencies (%) in unrelated controls without ARM, across studies included in meta-analysis of S69A in *LOC387715*. ‘Jakobs’ denotes estimates from the Jakobsdottir et al. paper.

## APPENDIX C

### FOR CHAPTER 5

Here the supplementary material published online as a part of the paper (JAKOBSDOTTIR *et al.* 2008) presented in chapter 5 is given.

#### C.1 LOGISTIC REGRESSION ANALYSES

##### C.1.1 Coding in two-factor models

A series of logistic regression models were fitted to the data in order to find a parsimonious model for the joint effects of each pair of loci. Models allowing for additive effects (ADD1, ADD2, and ADD-BOTH), models incorporating dominance effects (DOM1, DOM2, and DOM-BOTH), and three interaction models (ADD-INT, ADD-DOM, and DOM-INT) were fitted. In the additive models the genotypes,  $RR$ ,  $RN$ , and  $NN$  (or  $PP$ ,  $PN$ , and  $NN$ ), are coded  $-1$ ,  $0$ , and  $1$ , respectively; where  $R$  denotes the assumed risk allele at  $CFH$  or  $LOC387715$ ,  $P$  the assumed protective allele at  $C2$ , and  $N$  the assumed normal allele. The dominance models incorporate a variable to the additive models coded as  $-0.5$  for  $RR$  (or  $PP$ ) and  $NN$  and  $0.5$  for  $RN$  (or  $PN$ ). We let  $x_1$  and  $x_2$  denote the genotype variables in the additive models, and  $z_1$  and  $z_2$  the additional variables incorporated into the dominance models. Then the ADD1, ADD2, and ADD-BOTH models include terms  $(x_1)$ ,  $(x_2)$ , and  $(x_1 \text{ and } x_2)$ , respectively, and the DOM1, DOM2, DOM-BOTH models incorporate terms  $(z_1)$ ,  $(z_2)$ , and  $(z_1 \text{ and } z_2)$  to the ADD1, ADD2, and ADD-BOTH models, respectively. Three further interaction models are fitted: ADD-INT incorporates the product term  $(x_1x_2)$  to

the ADD-BOTH model, ADD-DOM incorporates the product terms ( $x_1x_2$ ,  $x_1z_2$ , and  $z_1x_2$ ), and DOM-INT incorporates the product terms ( $x_1x_2$ ,  $x_1z_2$ ,  $z_1x_2$ , and  $z_1x_2$ ) to the DOM-BOTH model.

### C.1.2 Coding in three-factor models

Since, for each pair of loci, the two-factor analyses implicated additive models as the most parsimonious and to keep the number of parameters as small as possible we only fit three-factor additive models without interaction. The models are ADD1, ADD2, ADD3, ADD12, ADD13, ADD23, and ADD123 and include terms ( $x_1$ ), ( $x_2$ ), ( $x_3$ ), ( $x_1$  and  $x_2$ ), ( $x_1$  and  $x_3$ ), ( $x_2$  and  $x_3$ ), and ( $x_1$ ,  $x_2$ , and  $x_3$ ), respectively, where  $x_1$ ,  $x_2$ , and  $x_3$ , are coded as in the additive two-factor models above.

## C.2 ASSOCIATION ANALYSES—*CFH* AND *LOC387715*

In our prior studies, we tested the associations of *Y402H* in *CFH* (CONLEY *et al.* 2005) and *S69A* in *LOC387715* (JAKOBSDOTTIR *et al.* 2005) in a smaller subset of our data, than we have typed now. In our larger dataset both variants are highly associated with ARM (allelic and genotypic  $P$ -values  $< 0.00001$  in both the case-control and family data, table 5.2) providing further confirmation for the likely involvement of these genes in ARM pathogenesis. The ORs for individuals heterozygous and homozygous for the risk allele at *Y402H* are 4.11 (95% CI 2.28 to 7.40) and 8.96 (95% CI 4.49 to 17.88), and the corresponding PARs are 56% (95% CI 34% to 71%) and 53% (95% CI 31% to 69%), respectively. The ORs for individuals heterozygous and homozygous the *S69A* risk allele are 3.63 (95% CI 2.19 to 6.03) and 8.24 (95% CI 3.81 to 17.81), and the corresponding PARs are 42% (95% CI 26% to 54%) and 32% (95% CI 17% to 44%), respectively.

Table C1: Genotype counts for *C2/CFB* variants, *Y402H* in *CFH*, and *S69A* in *LOC387715*

SNP	Gene	Chr	Bp <sup>a</sup>	Location <sup>a</sup>	Allele		Genotype counts in					
					labeling		Cases			Controls		
					1	2	11	12	22	11	12	22
<i>rs9332739</i>	<i>C2</i>	6	32011783	<i>E318D</i>	<i>C</i>	<i>G</i>	0	10	172	1	9	156
<i>rs547154</i>	<i>C2</i>	6	32018917	<i>IVS10</i>	<i>G</i>	<i>T</i>	170	9	0	130	31	0
<i>rs4151667</i>	<i>CFB</i>	6	32022003	<i>L9H</i>	<i>A</i>	<i>T</i>	0	10	168	1	10	156
<i>rs2072633</i>	<i>CFB</i>	6	32027557	<i>IVS17</i>	<i>A</i>	<i>G</i>	21	74	81	20	88	55
<i>rs1061170</i>	<i>CFH</i>	1	194925860	<i>Y402H</i>	<i>T</i>	<i>C</i>	21	80	60	69	64	22
<i>rs10490924</i>	<i>LOC387715</i>	10	124204438	<i>S69A</i>	<i>G</i>	<i>T</i>	50	74	40	103	42	10

Chr = chromosome

Bp = base pairs

<sup>a</sup> Bp and location within the genes are from NCBI build 127 (human genome build 36.2)

Table C2: Joint and relative genotype frequencies

		2-factor model			3-factor model						
		LOC387715			C2						
		LOC387715			GG			GT			
		GG	GT	TT	LOC387715			LOC387715			
		GG	GT	TT	GG	GT	TT	GG	GT	TT	
<b>Controls</b>	<b>CFH</b>	CC	0.1338	0.0141	0.0070	0.1056	0.0070	0.0070	0.0282	0.0070	0.0000
		CT	0.2465	0.1197	0.0423	0.2113	0.0915	0.0423	0.0352	0.0282	0.0000
		TT	0.2676	0.1479	0.0211	0.2113	0.1127	0.0141	0.0563	0.0352	0.0070
<b>Cases</b>	<b>CFH</b>	CC	0.1188	0.1891	0.0797	0.1109	0.1812	0.0781	0.0078	0.0078	0.0016
		CT	0.1391	0.2219	0.1000	0.1297	0.2047	0.0906	0.0094	0.0172	0.0094
		TT	0.0344	0.0750	0.0422	0.0281	0.0688	0.0328	0.0062	0.0062	0.0094
<b>Control/Case Ratios</b>	<b>CFH</b>	CC	1.13	0.07	0.09	0.95	0.04	0.09	3.62	0.90	0.00
		CT	1.77	0.54	0.42	1.63	0.45	0.47	3.74	1.64	0.00
		TT	7.78	1.97	0.50	7.52	1.64	0.43	9.08	5.68	0.74

The genotype frequencies are for the subset of our data used in the GMDR unadjusted analyses. This includes individuals typed at all three loci (CFH, LOC387715, and C2). The cases include the unrelated cases and those randomly picked from the families (see the main text for details). The Control/Case Ratios are the ratios of the joint allele frequencies in controls versus cases. The gray-highlighted cells correspond to cells with ratio < 1. Note, that those are the same cells that were classified as cases in the GMDR analyses.



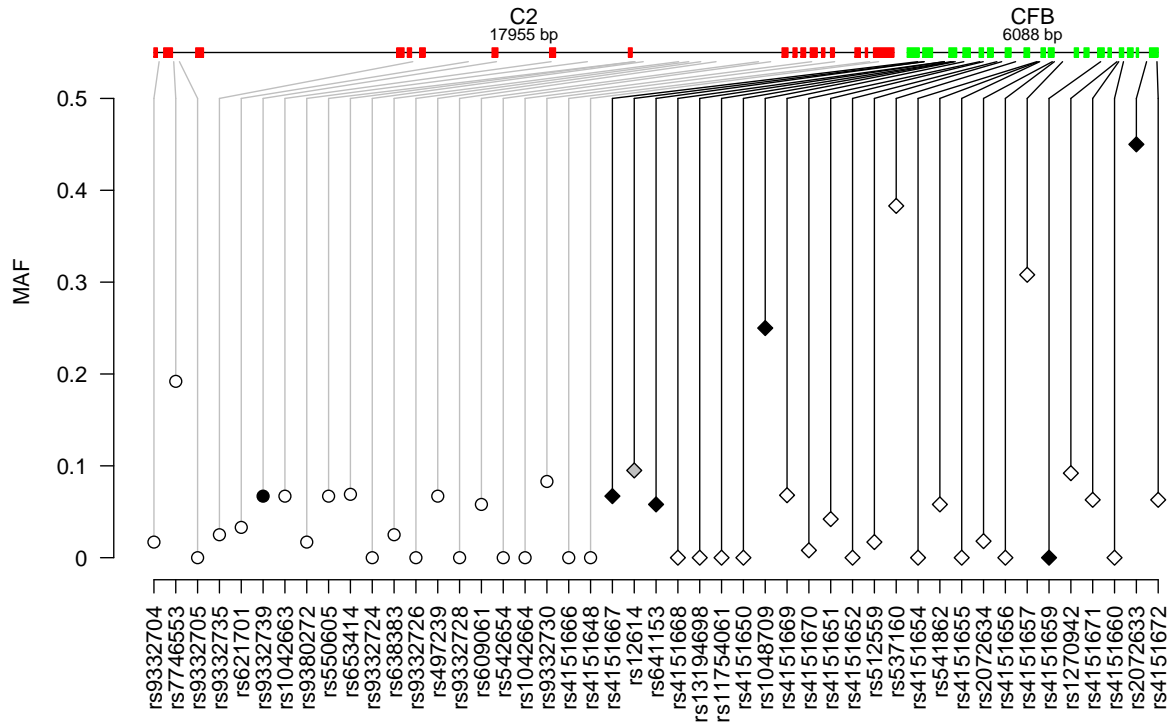


Figure C1: Minor allele frequency (MAF) of SNPs typed for the HapMap CEU population citehapmap in the *C2*/*CFB* region. Locations of the SNPs within the genes are shown. Red lines/boxes show the locations of exons in *C2* and green lines/boxes the locations of exons in *CFB*. White symbols represent SNPs not yet typed in any *C2*/*CFB* study citegold,maller,spencer (including the present study), black filled symbols represent SNPs typed by Gold et al. citegold and grey filled symbols represent SNPs typed in *C2*/*CFB* study citespencer other than the Gold et al. study citegold. Grey lines and circles correspond to SNPs in *C2* and black lines and dimonds correspond to SNPs in *CFB*.

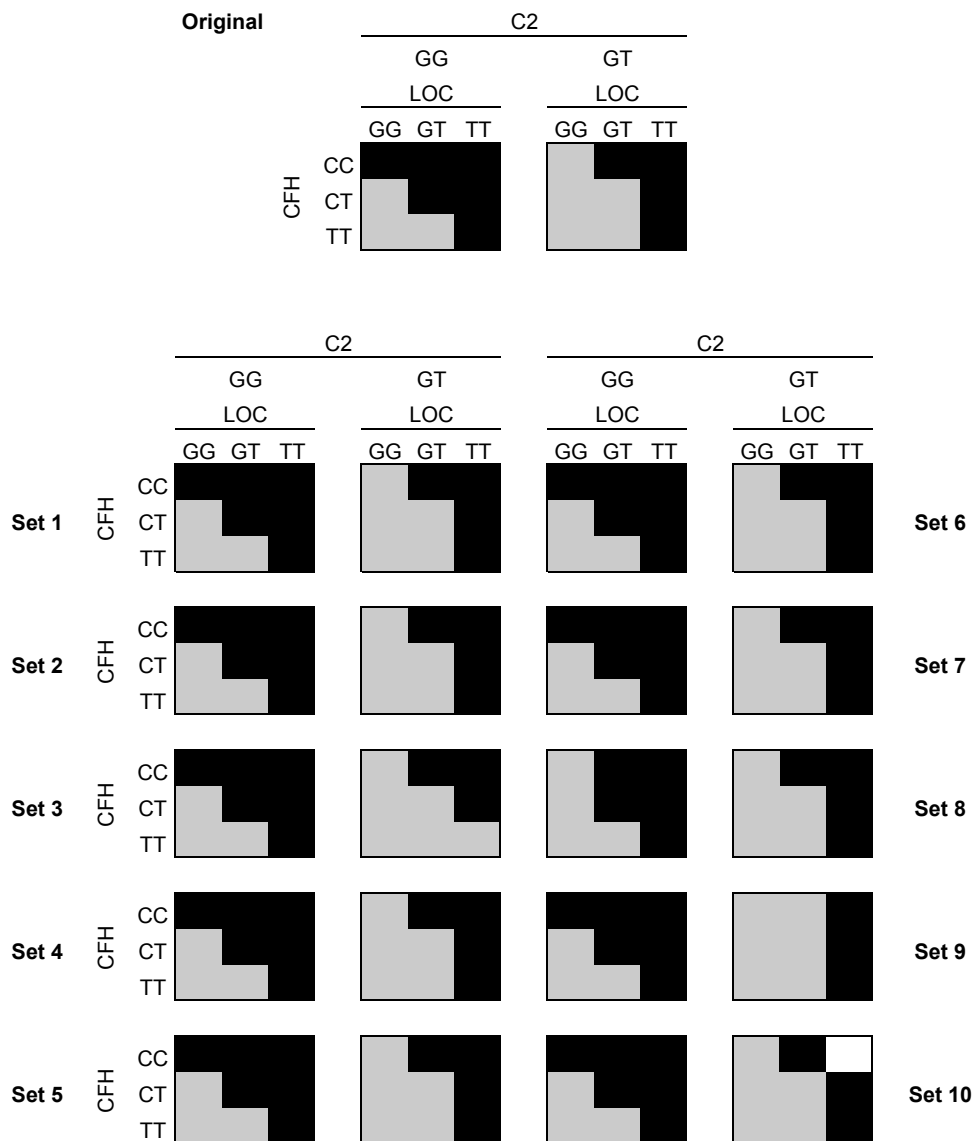


Figure C2: GMDR sensitivity analyses for the three-factor unadjusted model. The classification rules are shown for 10 data-sets, additional to the one used in the main paper (given at top). Each of those 10 data-sets has one case picked at random from each family. Black cells = cases, gray cells = controls, and white cells = empty cells/unknown status.

## APPENDIX D

### FOR CHAPTER 6

Here the supplementary material published online as a part of the paper ([JAKOBSDOTTIR \*et al.\* 2009](#)) presented in chapter 6 is given.

#### D.1 APPLICATION OF CLASSIFICATION-BASED METHODS TO AMD DATA

AMD is a complex, late-onset degenerative disease that is characterized by the disruption of the integrity of the retina, retinal pigment epithelium and choroid that can lead to the loss of central vision and significant visual disability. In recent years, three loci have been found strongly associated with AMD: functional SNPs at *CFH* ([EDWARDS \*et al.\* 2005](#); [HAINES \*et al.\* 2005](#); [KLEIN \*et al.\* 2005](#)) and *LOC387715* (or at the closely linked *PLEKHA1* [MIM 607772] or *HTRA1* [MIM 602194]) ([JAKOBSDOTTIR \*et al.\* 2005](#); [RIVERA \*et al.\* 2005](#)) are thought to increase the risk of AMD while variants at *CFB* (MIM 138470) or *C2* (MIM 217000) ([GOLD \*et al.\* 2006](#)) are thought to decrease risk. These findings appear robust and have been widely replicated ([GORIN 2007](#)). A number of studies have attempted to use variants at these genes to build predictive models for AMD. To our knowledge, none has applied ROC theory to evaluate the classification performance of these variants individually or jointly.

### D.1.1 AMD data

For illustration we use part of our AMD data, which includes 640 cases and 142 controls fully typed at three SNPs at all three loci: *rs1061170* (*Y402H*) in *CFH*, *rs10490924* (*S69A*) in *LOC387715*, and *rs547154* (*IVS10*) in *C2*. For recruitment and phenotyping we refer to our previous publications [8,9] and for genotyping see [JAKOBSDOTTIR \*et al.\* \(2008\)](#).

### D.1.2 Methods

To combine information from the three SNPs for classification using ROC, we use a generalized linear model proposed by Ma and Huang [11]:  $P(Y = 1|X) = G(\beta^T X)$ , where  $Y$  is the disease status ( $Y = 1$  for cases and  $Y = 0$  for controls),  $X$  is the matrix of genotype columns  $X_i = (X_{1,i}, \dots, X_{d,i})^T$  for the  $i$ th subject,  $\beta = (\beta_1, \dots, \beta_d)^T$  is a  $d$ -dimensional vector of unknown regression parameters, and  $G$  is an unknown increasing link function. Since  $G$  is assumed to be increasing, a classification rule can be constructed based on the risk score,  $\beta^T X$ , only. We use the rule: if  $\beta^T X > c$ , we classify this individual as a case, otherwise we classify the individual as a control. This is a sensible approach as decision criteria based on risk are statistically optimal [12]. The overall performance of the classifier is then measured by the AUC of the ROC curve, which is a two-dimensional plot of  $((FPF(c), TPF(c)) : c \in \mathbb{R})$ , where  $FPF(c)$ , and  $TFP(c)$  are the FPF and TPF of the classification rule if  $\beta^T X > c$ . To get all points on the ROC curve the FPF and TPF are estimated for all possible values of  $c$ . The empirical AUC is maximized as a function of  $\beta$ . For each  $\beta$ , the AUC is estimated using a nonparametric trapezoidal estimator [13]. Note that this ROC model is a more general model than the logistic model, as  $G$  needs not to be known. Since many previous studies have found an additive model to be best fitting in both single and multi locus models of *CFH*, *LOC387715*, and *C2* variants, we let  $X_{1,i}$  be the number of risk alleles at *Y402H* at *CFH*,  $X_{2,i}$  be the number of risk alleles at *S69A* at *LOC387715*, and  $X_{3,i}$  be the number of protective alleles at *IVS10* at *C2*. For comparison, we also present the results of logistic regression analyses where the genotypes are coded the same way. In addition to performing ROC and logistic regression analyses, we draw an integrated predictiveness and classification plot [14]. In the integrated plot, there are two aligned plots: In the top plot, ordered individual risks are plotted as function of the risk percentile and, in the bottom plot, the TPF and FPF are plotted as a function of the risk percentile such that at each point the TPF and FPF are calculated for the risk threshold,  $c$ , equal

to the risk associated with the corresponding risk percentile. As we are working with case-control data, we can only calculate individual-level risks from the logistic model (setting the  $G$  function to be the logit function) if the prevalence is known. To be able to draw the plot we therefore need to assume a specific value for the prevalence. Since our data are elderly white individuals (mean age 72.9 and standard deviation [sd] 9.9 in controls) and our cases are all of advanced phenotype, we use a prevalence estimate for advanced AMD in white individuals 65 years and older (approximately 1 sd from the mean) of 5.5%; the US 2000 census data (Table 4: Annual Estimates of the White Alone Population by Age and Sex for the United States: April 1, 2000 to July 1, 2006 [NC-EST2006-04-WA]) were used to project the sex-specific 5-year age interval estimates of Friedman et al. [15] to estimate the AMD prevalence for 65 years and older.

### D.1.3 Accounting for covariates

We also ran the above analysis while adjusted for age, sex, and smoking. The AUC of the genetic model without the covariates was 0.78 and improved to 0.82 when the covariates were added to the model. The AUC of model with only the covariates had an AUC of 0.66. Note that the effective sample size for these new analyses is smaller due to missing covariate information. The AUC of the unadjusted model in the main text (0.79) is therefore, not exactly equal to the AUC of the unadjusted model here (0.78).

## D.2 ESTIMATING THE AUC FROM META-DATA

As science progresses, there is a need for methods to continuously update previous classification models. [LU and ELSTON \(2008\)](#) developed a method to do this when only meta-data and summary statistics are available. This is especially useful if not all markers have been typed in the same samples. Then, if we assume homogeneity across samples, we can combine estimates to form a new classification rule. To compare the AUC of the new classification rule with the old rule, the information we need are 1) allele frequencies in case and control populations or 2) allele frequencies in the general population, risk ratios, and prevalence.

## D.3 DETAILS ON DATA IN OTHER REAL DATA EXAMPLES

### D.3.1 Cardiovascular events

KATHIRESAN *et al.* (2008) investigated whether genetic variants could improve classification accuracy for cardiovascular events beyond standard risk factors. First they tested for single SNP associations of 11 SNPs with low-density lipoprotein (LDL) and high-density lipoprotein (HDL) levels and then identified a set of 9 SNPs that were independently associated with lipid levels. Using these 9 SNPs, they created a simple genotype score based on the total number of unfavorable alleles in all 9 genotypes of the individual, and then evaluated the classification accuracy of the genotype score for the 10-year incidence of cardiovascular events. The  $P$  values for the 9 SNPs ranged from 0.003 to  $10^{-29}$  (table D1) and the adjusted hazard ratio of the genotype score was 1.15 (95% CI 1.07–1.24).

The AUC for prediction of 10-year incidence of cardiovascular events was estimated using model with 14 clinical covariates and no genotype information and found to be 0.80. When the genotype score, which included several highly associated SNPs, was included in the model, the AUC was not improved and also equaled 0.80 even though accounting for the genotype score significantly improved the regression model ( $P$  value 0.0002, Table S3 of KATHIRESAN *et al.* (2008)). The authors additionally looked at whether accounting for the genotype score improved the clinical reclassification and found modest improvement such that the estimated risk correctly increased for individuals who subsequently experienced cardiovascular event and correctly decreased for individuals who remained free of cardiovascular events at 10-year follow-up ( $P$  value 0.01).

### D.3.2 Type 2 diabetes

The 12 SNPs used to generate a classification rule for type 2 diabetes with the Lu and Elston method (LU and ELSTON 2008) come from three studies (table D2) [18-20].

### D.3.3 Prostate cancer

We used the Lu and Elston method (LU and ELSTON 2008) to investigate the classification accuracy of a genetic risk model of two prostate cancer risk SNPs [21] (table D3). We used the information from the combined cohort from the study of YEAGER *et al.* (2007).

Table D1: Association results of 9 SNPs associated with LDL and HDL cholesterol. Information from Table 2 of [KATHIRESAN \*et al.\* \(2008\)](#).

SNP	<i>P</i> value
LDL cholesterol	
<i>rs693</i>	$8 \times 10^{-7}$
<i>rs4420638</i>	$3 \times 10^{-21}$
<i>rs12654264</i>	0.002
<i>rs1529729</i>	0.003
<i>rs11591147</i>	$7 \times 10^{-7}$
HDL cholesterol	
<i>rs3890182</i>	0.003
<i>rs1800775</i>	$2 \times 10^{-29}$
<i>rs1800588</i>	$4 \times 10^{-10}$
<i>rs328</i>	$3 \times 10^{-12}$

Table D2: Association results of 12 type 2 diabetes SNPs.

SNP	Allele frequency in		<i>P</i> value	OR	Study
	Cases	Controls			
<i>rs5219</i>	0.384	0.354	0.0001	1.14	<a href="#">WEEDON <i>et al.</i> (2006)</a>
<i>rs1801282</i>	0.099	0.123	$4 \times 10^{-5}$	1.29	<a href="#">WEEDON <i>et al.</i> (2006)</a>
<i>rs7903146</i>	0.406	0.293	$2 \times 10^{-34}$	Het 1.65, Hom 2.77	<a href="#">SLADEK <i>et al.</i> (2007)</a>
<i>rs13266634</i>	0.254	0.301	$6 \times 10^{-8}$	Het 1.18, Hom 1.53	<a href="#">SLADEK <i>et al.</i> (2007)</a>
<i>rs1111875</i>	0.358	0.402	$3 \times 10^{-6}$	Het 1.19, Hom 1.44	<a href="#">SLADEK <i>et al.</i> (2007)</a>
<i>rs740010</i>	0.336	0.301	$1 \times 10^{-4}$	Het 1.14, Hom 1.40	<a href="#">SLADEK <i>et al.</i> (2007)</a>
<i>rs3740878</i>	0.240	0.272	$1 \times 10^{-4}$	Het 1.26, Hom 1.46	<a href="#">SLADEK <i>et al.</i> (2007)</a>
<i>rs4402960</i>	0.341	0.304	$8 \times 10^{-4}$	1.18	<a href="#">SCOTT <i>et al.</i> (2007)</a>
<i>rs7754840</i>	0.387	0.360	0.0095	1.12	<a href="#">SCOTT <i>et al.</i> (2007)</a>
<i>rs10811661</i>	0.872	0.850	0.0022	1.20	<a href="#">SCOTT <i>et al.</i> (2007)</a>
<i>rs9300039</i>	0.924	0.892	$7 \times 10^{-8}$	1.49	<a href="#">SCOTT <i>et al.</i> (2007)</a>
<i>rs8050136</i>	0.406	0.381	0.017	1.11	<a href="#">SCOTT <i>et al.</i> (2007)</a>

Information from the combined cohort of stage 2 used from the Scott *et al.* study.

Table D3: Association results of two prostate cancer disease SNPs.

SNP	Allele frequency in		$P$ value	OR <sub>het</sub>	OR <sub>hom</sub>
	Cases	Controls			
<i>rs1447295</i>	0.15	0.11	$2 \times 10^{-14}$	1.43	2.23
<i>rs6983267</i>	0.56	0.50	$9 \times 10^{-13}$	1.26	1.58

### D.3.4 Inflammatory bowel diseases

We used the Lu and Elston method (LU and ELSTON 2008) to investigate the classification accuracy of genetic risk model of five SNPs. Two SNPs are in *IL23R* and are thought to be uncorrelated, one in *ATG16CL*, one in *NOD2/CARD15*, and one in *IRGM*; all are associated with Crohns disease (which is a form of inflammatory bowel disease) (table D4).

Table D4: Association results of five Crohns disease SNPs.

SNP	Allele frequency in		$P$ value	OR	Study
	Cases	Controls			
<i>rs11209026</i>	0.019	0.070	$5 \times 10^{-9}$	0.26	DUERR <i>et al.</i> (2006)
<i>rs751784</i>	0.345	0.448	$5 \times 10^{-9}$	0.89	CUMMINGS <i>et al.</i> (2007)
<i>rs2241800</i>	0.61	0.52	$2 \times 10^{-7}$	1.45	CUMMINGS <i>et al.</i> (2007)
<i>rs2076756</i>	0.358	0.244	$7 \times 10^{-14}$	NA	RIOUX <i>et al.</i> (2007)
<i>rs13361189</i>	0.098	0.067	$4 \times 10^{-8}$	1.38	PARKES <i>et al.</i> (2007)



## APPENDIX E

### FOR CHAPTER 7

Here we show how we evaluate  $W$  in our statistic 7.9.2. For simplicity we first look at the case when female dominance is not modeled

$$\begin{aligned} f(\theta) &= \theta^T \mathbf{a} - \frac{1}{2} \theta^T I_0 \theta \\ &= \theta_1 a_1 + \theta_3 a_3 + \sqrt{\theta_1 \theta_3} a_4 - \frac{1}{2} (\theta_1^2 I_{11} + \theta_3^2 I_{33} + \theta_1 \theta_3 I_{44}) \end{aligned}$$

We take the partial first and second derivatives

$$\begin{aligned} \frac{\partial f}{\partial \theta_1} &= a_1 - \theta_1 I_{11} + \frac{1}{2} \sqrt{\frac{\theta_3}{\theta_1}} a_4 - \frac{1}{2} \theta_3 I_{44} \\ \frac{\partial f}{\partial \theta_3} &= a_3 - \theta_3 I_{33} + \frac{1}{2} \sqrt{\frac{\theta_1}{\theta_3}} a_4 - \frac{1}{2} \theta_1 I_{44} \\ \frac{\partial^2 f}{\partial \theta_1^2} &= -I_{11} - \frac{1}{4} \frac{1}{\theta_1} \sqrt{\frac{\theta_3}{\theta_1}} a_4 \\ \frac{\partial^2 f}{\partial \theta_3^2} &= -I_{33} - \frac{1}{4} \frac{1}{\theta_3} \sqrt{\frac{\theta_1}{\theta_3}} a_4 \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_3} &= \frac{1}{4} \sqrt{\frac{1}{\theta_1 \theta_3}} a_4 - \frac{1}{2} I_{44} \end{aligned}$$

Then the Hessian matrix is

$$H = \begin{bmatrix} -I_{11} - \frac{1}{4} \frac{1}{\theta_1} \sqrt{\frac{\theta_3}{\theta_1}} a_4 & \frac{1}{4} \sqrt{\frac{1}{\theta_1 \theta_3}} a_4 - \frac{1}{2} I_{44} \\ \frac{1}{4} \sqrt{\frac{1}{\theta_1 \theta_3}} a_4 - \frac{1}{2} I_{44} & -I_{33} - \frac{1}{4} \frac{1}{\theta_3} \sqrt{\frac{\theta_1}{\theta_3}} a_4 \end{bmatrix}$$

We can solve the set of equations  $\frac{\partial f}{\partial \theta_1} = 0$  and  $\frac{\partial f}{\partial \theta_3} = 0$  explicitly as shown with Mathematic but the solutions are too long to write down. Therefore we solve those equations with numerical methods instead. We then use the Hessian to find which solutions correspond to a local maxima by using the second derivative test. If  $\theta^*$  is the solution that gives the maxima the value of the score statistic when the condition fails becomes  $W = f(\theta^*)$ .

Similarly if the female dominance is modeled the we look at

$$\begin{aligned} \theta^T \mathbf{a} - \frac{1}{2} \theta^T I_0 \theta \\ = \theta_1 a_1 + \theta_2 a_2 + \theta_3 a_3 + \sqrt{\theta_1 \theta_3} a_4 - \frac{1}{2} (\theta_1^2 I_{11} + 2\theta_1 \theta_2 I_{12} + \theta_2^2 I_{22} + \theta_3^2 I_{33} + \theta_1 \theta_3 I_{44}) \end{aligned}$$

and take the partial derivatives as before (which define the Hessian matrix)

$$\begin{aligned} \frac{\partial f}{\partial \theta_1} &= a_1 - \theta_1 I_{11} + \frac{1}{2} \sqrt{\frac{\theta_3}{\theta_1}} a_4 - \theta_2 I_{12} - \frac{1}{2} \theta_3 I_{44} \\ \frac{\partial f}{\partial \theta_2} &= a_2 - \theta_2 I_{22} - \theta_1 I_{12} \\ \frac{\partial f}{\partial \theta_3} &= a_3 - \theta_3 I_{33} + \frac{1}{2} \sqrt{\frac{\theta_1}{\theta_3}} a_4 - \frac{1}{2} \theta_1 I_{44} \\ \frac{\partial^2 f}{\partial \theta_1^2} &= -I_{11} - \frac{1}{4} \frac{1}{\theta_1} \sqrt{\frac{\theta_3}{\theta_1}} a_4 \\ \frac{\partial^2 f}{\partial \theta_2^2} &= -I_{22} \\ \frac{\partial^2 f}{\partial \theta_3^2} &= -I_{33} - \frac{1}{4} \frac{1}{\theta_3} \sqrt{\frac{\theta_1}{\theta_3}} a_4 \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} &= -I_{12} \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_3} &= \frac{1}{4} \sqrt{\frac{1}{\theta_1 \theta_3}} a_4 - \frac{1}{2} I_{44} \\ \frac{\partial^2 f}{\partial \theta_2 \partial \theta_3} &= 0 \end{aligned}$$

## BIBLIOGRAPHY

- ABECASIS, G. R., W. O. COOKSON, and L. R. CARDON, 2002 Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* *30*(1): 97–101.
- ABECASIS, G. R., B. M. YASHAR, Y. ZHAO, N. M. GHIASVAND, S. ZAREPARSI, K. E. BRANHAM, A. C. REDDICK, E. H. TRAGER, S. YOSHIDA, J. BAHLING, E. FILIPPOVA, S. ELNER, M. W. JOHNSON, A. K. VINE, P. A. SIEVING, S. G. JACOBSON, J. E. RICHARDS, and A. SWAROOP, 2004 Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease. *Am J Hum Genet* *74*(3): 482–494.
- AGE RELATED EYE DISEASE STUDY GROUP, 1999 The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clin Trials* *20*(6): 573–600.
- AGE RELATED EYE DISEASE STUDY GROUP, 2000 Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* *107*(12): 2224–32.
- AMOS, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* *54*(3): 535–43.
- AYYAGARI, R., K. ZHANG, A. HUTCHINSON, Z. YU, A. SWAROOP, L. E. KAKUK, J. M. SEDDON, P. S. BERNSTEIN, R. A. LEWIS, J. TAMMUR, Z. YANG, Y. LI, H. ZHANG, B. M. YASHAR, J. LIU, K. PETRUKHIN, P. A. SIEVING, and R. ALLIKMETS, 2001 Evaluation of the ELOVL4 gene in patients with age-related macular degeneration. *Ophthalmic Genet* *22*(4): 233–9.
- BAIRD, P. N., E. GUIDA, D. T. CHU, H. T. VU, and R. H. GUYMER, 2004 The epsilon2 and epsilon4 alleles of the apolipoprotein gene are associated with age-related macular degeneration. *Invest Ophthalmol Vis Sci* *45*(5): 1311–5.
- BAIRD, P. N., F. M. ISLAM, A. J. RICHARDSON, M. CAIN, N. HUNT, and R. GUYMER, 2006 Analysis of the Y402H variant of the complement factor H gene in age-related macular degeneration. *Invest Ophthalmol Vis Sci* *47*(10): 4194–8.
- BARRETT, J. C., B. FRY, J. MALLER, and M. J. DALY, 2005 Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* *21*(2): 263–5.
- BHATTACHARJEE, S., 2008 Variance component score statistics for QTL mapping. Ph. D. thesis, University of Pittsburgh.

- BHATTACHARJEE, S., C. L. KUO, N. MUKHOPADHYAY, G. N. BROCK, D. E. WEEKS, and E. FEINGOLD, 2008 Robust score statistics for QTL linkage analysis. *Am J Hum Genet* *82*(3): 567–82.
- BLANGERO, J. and L. ALMASY, 1997 Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiol* *14*(6): 959–64.
- BOURGAIN, C., E. GÉNIN, N. COX, and F. CLERGET-DARPOUX, 2007 Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? *Eur J Hum Genet* *15*(3): 260–3.
- BROWN, H., 2006 *Applied Mixed Models in Medicine*. West Sussex, UK: John Wiley & Sons, Inc.
- BROWNING, S. R., J. D. BRILEY, L. P. BRILEY, G. CHANDRA, J. H. CHARNECKI, M. G. EHM, K. A. JOHANSSON, B. J. JONES, A. J. KARTER, D. P. YARNALL, and M. J. WAGNER, 2005 Case-control single-marker and haplotypic association analysis of pedigree data. *Genet Epidemiol* *28*(2): 110–22.
- BULMER, M., 1985 *The mathematical theory of quantitative genetics* (2 ed.). Oxford University Press.
- CALEFATO, J. M., I. NIPPERT, H. J. HARRIS, U. KRISTOFFERSSON, J. SCHMIDTKE, L. P. TEN KATE, E. ANIONWU, C. BENJAMIN, K. CHALLEN, A. M. PLASS, R. HARRIS, and C. JULIAN-REYNIER, 2008 Assessing educational priorities in genetics for general practitioners and specialists in five countries: factor structure of the Genetic-Educational Priorities (Gen-EP) scale. *Genet Med* *10*(2): 99–106.
- CAMERON, D. J., Z. YANG, D. GIBBS, H. CHEN, Y. KAMINOH, A. JORGENSEN, J. ZENG, L. LUO, E. BRINTON, G. BRINTON, J. M. BRAND, P. S. BERNSTEIN, N. A. ZABRISKIE, S. TANG, R. CONSTANTINE, Z. TONG, and K. ZHANG, 2007 HTRA1 variant confers similar risks to geographic atrophy and neovascular age-related macular degeneration. *Cell Cycle* *6*(9): 1122–5.
- CARREL, L. and H. F. WILLARD, 2005 X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* *434*(7031): 400–4.
- CHEN, L. J., D. T. LIU, P. O. TAM, W. M. CHAN, K. LIU, K. K. CHONG, D. S. LAM, and C. P. PANG, 2006 Association of complement factor H polymorphisms with exudative age-related macular degeneration. *Mol Vis* **12**: 1536–42.
- CHERNOFF, H., 1954 On the distribution of the likelihood ratio. *Ann Math Stat* *25*(3): 573–578.
- CLAYTON, D. G. SNPHAP: A program for estimating frequencies of large haplotypes of SNPs. Version 1.3.1.
- CLERGET-DARPOUX, F., C. BONAITI-PELLIE, and J. HOCHÉZ, 1986 Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* *42*(2): 393–399.
- CLERGET-DARPOUX, F. and R. ELSTON, 2007 Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum Hered* *64*(2): 91–6.

- CONLEY, Y. P., J. JAKOBSDOTTIR, T. MAH, D. E. WEEKS, R. KLEIN, L. KULLER, R. E. FERRELL, and M. B. GORIN, 2006 CFH, ELOVL4, PLEKHA1 and LOC387715 genes and susceptibility to age-related maculopathy: AREDS and CHS cohorts and meta-analyses. *Hum Mol Genet* *15*(21): 3206–18.
- CONLEY, Y. P., A. THALAMUTHU, J. JAKOBSDOTTIR, D. E. WEEKS, T. MAH, R. E. FERRELL, and M. B. GORIN, 2005 Candidate gene analysis suggests a role for fatty acid biosynthesis and regulation of the complement system in the etiology of age-related maculopathy. *Hum Mol Genet* *14*(14): 1991–2002.
- COOK, N. R., 2007 Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* *115*(7): 928–35.
- CORDELL, H. J. and D. G. CLAYTON, 2005 Genetic association studies. *Lancet* *366*(9491): 1121–31.
- CUMMINGS, J., T. AHMAD, A. GEREMIA, J. BECKLY, R. COONEY, L. HANCOCK, S. PATHAN, C. GUO, L. CARDON, and D. JEWELL, 2007 Contribution of the novel inflammatory bowel disease gene IL23R to disease susceptibility and phenotype. *Inflamm Bowel Dis* *13*(9): 1063–1068.
- CUMMINGS, J., R. COONEY, S. PATHAN, C. ANDERSON, J. BARRETT, J. BECKLY, A. GEREMIA, L. HANCOCK, C. GUO, T. AHMAD, L. CARDON, and D. JEWELL, 2007 Confirmation of the role of ATG16L1 as a Crohn’s disease susceptibility gene. *Inflamm Bowel Dis* *13*(8): 941–946.
- DAVIES, H. T., I. K. CROMBIE, and M. TAVAKOLI, 1998 When can odds ratios mislead? *Bmj* *316*(7136): 989–91.
- DE LUCA, A., M. DE FALCO, L. DE LUCA, R. PENTA, V. SHRIDHAR, F. BALDI, M. CAMPIONI, M. G. PAGGI, and A. BALDI, 2004 Pattern of expression of HtrA1 during mouse development. *J Histochem Cytochem* *52*(12): 1609–1617.
- DEANGELIS, M. M., F. JI, I. K. KIM, S. ADAMS, J. CAPONE, A., J. OTT, J. W. MILLER, and T. P. DRYJA, 2007 Cigarette smoking, CFH, APOE, ELOVL4, and risk of neovascular age-related macular degeneration. *Arch Ophthalmol* *125*(1): 49–54.
- DEEKS, J., 1998 When can odds ratios mislead? Odds ratios should be used only in case-control studies and logistic regression analyses. *Bmj* *317*(7166): 1155–6; author reply 1156–7.
- DERSIMONIAN, R. and N. LAIRD, 1986 Meta-analysis in clinical trials. *Control Clin Trials* *7*(3): 177–88.
- DESPRIET, D. D., C. C. KLAVER, J. C. WITTEMAN, A. A. BERGEN, I. KARDYS, M. P. DE MAAT, S. S. BOEKHOORN, J. R. VINGERLING, A. HOFMAN, B. A. OOSTRA, A. G. UITTERLINDEN, T. STIJNEN, C. M. VAN DUIJN, and P. T. DE JONG, 2006 Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. *Jama* *296*(3): 301–9.
- DEWAN, A., M. BRACKEN, and J. HOH, 2007 Two genetic pathways for age-related macular degeneration. *Curr Opin Genet Dev* *17*(3): 228–33.

- DEWAN, A., M. LIU, S. HARTMAN, S. S. ZHANG, D. T. LIU, C. ZHAO, P. O. TAM, W. M. CHAN, D. S. LAM, M. SNYDER, C. BARNSTABLE, C. P. PANG, and J. HOH, 2006 HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* *314*(5801): 989–92.
- DINU, V., P. L. MILLER, and H. ZHAO, 2007 Evidence for association between multiple complement pathway genes and AMD. *Genet Epidemiol* *31*(3): 224–37.
- DOWLER, S., R. A. CURRIE, D. G. CAMPBELL, M. DEAK, G. KULAR, C. P. DOWNES, and D. R. ALESSI, 2000 Identification of pleckstrin-homology-domain-containing proteins with novel phosphoinositide-binding specificities. *Biochem J* *351*(1): 19–31.
- DRIGALENKO, E., 1998 How sib pairs reveal linkage. *Am J Hum Genet* *63*(4): 1242–5.
- DUERR, R., K. TAYLOR, S. BRANT, J. RIOUX, M. SILVERBERG, M. DALY, A. STEINHART, C. ABRAHAM, M. REGUEIRO, A. GRIFFITHS, T. DASSOPOULOS, A. BITTON, H. YANG, S. TARGAN, L. DATTA, E. KISTNER, L. SCHUMM, A. LEE, P. GREGERSEN, M. BARMADA, J. ROTTER, D. NICOLAE, and J. CHO, 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* *314*(5804): 1461–1463.
- DUNAI, G., B. VASARHELYI, M. SZABO, J. HAJDU, G. MESZAROS, T. TULASSAY, and A. TRESZL, 2008 Published genetic variants in retinopathy of prematurity: random forest analysis suggests a negligible contribution to risk and severity. *Curr Eye Res* *33*(5): 501–5.
- EDITORIAL, 2007 Risky business. *Nat Genet* *39*(12): 1415.
- EDWARDS, A. O., R. RITTER, R., K. J. ABEL, A. MANNING, C. PANHUYSEN, and L. A. FARRER, 2005 Complement factor H polymorphism and age-related macular degeneration. *Science* *308*(5720): 421–4.
- EKSTRØM, C. T., 2004 Multipoint linkage analysis of quantitative traits on sex-chromosomes. *Genet Epidemiol* *26*(3): 218–30.
- ELSTON, R. C., S. BUXBAUM, K. B. JACOBS, and J. M. OLSON, 2000 Haseman and Elston revisited. *Genet Epidemiol* *19*(1): 1–17.
- ENNIS, S., S. GOVERDHAN, A. CREE, J. HOH, A. COLLINS, and A. LOTERY, 2007 Fine-scale linkage disequilibrium mapping of age-related macular degeneration in the complement factor H gene region. *Br J Ophthalmol* *91*(7): 966–70.
- ESPARZA-GORDILLO, J., J. SORIA, A. BUIL, L. ALMASY, J. BLANGERO, J. FONTCUBERTA, and S. RODRIGUEZ DE CORDOBA, 2004 Genetic and environmental factors influencing the human factor H plasma levels. *Immunogenetics* *56*(2): 77–82.
- FAN, J. and A. MALIK, 2003 Toll-like receptor-4 (TLR4) signaling augments chemokine-induced neutrophil migration by modulating cell surface expression of chemokine receptors. *Nat Med* *9*(3): 315–321.
- FEERO, W. G., 2008 Genetics of common disease: a primary care priority aligned with a teachable moment? *Genet Med* *10*(2): 81–2.

- FISHER, S., G. ABECASIS, B. YASHAR, S. ZAREPARSI, A. SWAROOP, S. IYENGAR, B. KLEIN, R. KLEIN, K. LEE, J. MAJEWSKI, D. SCHULTZ, M. KLEIN, J. SEDDON, S. SANTANGELO, D. WEEKS, Y. CONLEY, T. MAH, S. SCHMIDT, J. HAINES, M. PERICAK-VANCE, M. GORIN, H. SCHULZ, F. PARDI, C. LEWIS, and B. WEBER, 2005 Meta-analysis of genome scans of age-related macular degeneration. *Hum Mol Genet* *14*(15): 2257–2264.
- FISHER, S. A., A. RIVERA, L. G. FRITSCH, G. BABADJANOVA, S. PETROV, and B. H. WEBER, 2007 Assessment of the contribution of CFH and chromosome 10q26 AMD susceptibility loci in a Russian population isolate. *Br J Ophthalmol* *91*(5): 576–8.
- FLAQUER, A., G. A. RAPPOLD, T. F. WIENKER, and C. FISCHER, 2008 The human pseudoautosomal regions: a review for genetic epidemiologists. *Eur J Hum Genet* *16*(7): 771–779.
- FRANCIS, P., B. APPUKUTTAN, E. SIMMONS, N. LANDAUER, J. STODDARD, S. HAMON, J. OTT, B. FERGUSON, M. KLEIN, J. STOUT, and N. M., 2008 Rhesus monkeys and humans share common susceptibility genes for age-related macular disease. *Hum Mol Genet* *17*(17): 2673–80.
- FRANCIS, P. J., S. GEORGE, D. W. SCHULTZ, B. ROSNER, S. HAMON, J. OTT, R. G. WELEBER, M. L. KLEIN, and J. M. SEDDON, 2007 The LOC387715 gene, smoking, body mass index, environmental associations with advanced age-related macular degeneration. *Hum Hered* *63*(3–4): 212–8.
- FRIED, L. P., N. O. BORHANI, P. ENRIGHT, C. D. FURBERG, J. M. GARDIN, R. A. KRONMAL, L. H. KULLER, T. A. MANOLIO, M. B. MITTELMARK, A. NEWMAN, and ET AL., 1991 The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* *1*(3): 263–76.
- FRIEDMAN, D. S., B. J. O’COLMAIN, B. MUNOZ, S. C. TOMANY, C. MCCARTY, P. T. DE JONG, B. NEMESURE, P. MITCHELL, and J. KEMPEN, 2004 Prevalence of age-related macular degeneration in the United States. *Arch Ophthalmol* *122*(4): 564–72.
- FUSE, N., A. MIYAZAWA, M. MENGKEGALE, M. YOSHIDA, R. WAKUSAWA, T. ABE, and M. TAMAI, 2006 Polymorphisms in Complement Factor H and Hemicentin-1 genes in a Japanese population with dry-type age-related macular degeneration. *Am J Ophthalmol* *142*(6): 1074–6.
- GOETZ, T., 2007 23AndMe Will Decode Your DNA for \$1000. Welcome to the Age of Genomics. *Wired Magazine* **15.12**: 256–265, 283.
- GOLD, B., J. E. MERRIAM, J. ZERNANT, L. S. HANCOX, A. J. TAIBER, K. GEHRS, K. CRAMER, J. NEEL, J. BERGERON, G. R. BARILE, R. T. SMITH, G. S. HAGEMAN, M. DEAN, and R. ALLIKMETS, 2006 Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat Genet* *38*(4): 458–62.
- GOLDGAR, D. E., 1990 Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* *47*(6): 957–67.
- GORIN, M. B., 2007 A clinician’s view of the molecular genetics of age-related maculopathy. *Arch Ophthalmol* *125*(1): 21–9.
- GOTOH, N., R. YAMADA, H. HIRATANI, V. RENAULT, S. KUROIWA, M. MONET, S. TOYODA, S. CHIDA, M. MANDAI, A. OTANI, N. YOSHIMURA, and F. MATSUDA, 2006 No association be-

- tween complement factor H gene polymorphism and exudative age-related macular degeneration in Japanese. *Hum Genet* *120*(1): 139–43.
- GUDBJARTSSON, D. F., K. JONASSON, M. L. FRIGGE, and A. KONG, 2000 Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* *25*(1): 12–13.
- HADDAD, S., C. A. CHEN, S. L. SANTANGELO, and J. M. SEDDON, 2006 The genetics of age-related macular degeneration: a review of progress to date. *Surv Ophthalmol* *51*(4): 316–63.
- HAGEMAN, G. S., D. H. ANDERSON, L. V. JOHNSON, L. S. HANCOX, A. J. TAIBER, L. I. HARDISTY, J. L. HAGEMAN, H. A. STOCKMAN, J. D. BORCHARDT, K. M. GEHRS, R. J. SMITH, G. SILVESTRI, S. R. RUSSELL, C. C. KLAVER, I. BARBAZETTO, S. CHANG, L. A. YANNUZZI, G. R. BARILE, J. C. MERRIAM, R. T. SMITH, A. K. OLSH, J. BERGERON, J. ZERNANT, J. E. MERRIAM, B. GOLD, M. DEAN, and R. ALLIKMETS, 2005 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* *102*(20): 7227–32.
- HAGEMAN, G. S., P. J. LUTHERT, N. H. VICTOR CHONG, L. V. JOHNSON, D. H. ANDERSON, and R. F. MULLINS, 2001 An integrated hypothesis that considers drusen as biomarkers of immune-mediated processes at the RPE-Bruch’s membrane interface in aging and age-related macular degeneration. *Prog Retin Eye Res* *20*(6): 705–32.
- HAGEMAN, G. S. and R. F. MULLINS, 1999 Molecular composition of drusen as related to sub-structural phenotype. *Mol Vis* **5**: 28.
- HAINES, J. L., M. A. HAUSER, S. SCHMIDT, W. K. SCOTT, L. M. OLSON, P. GALLINS, K. L. SPENCER, S. Y. KWAN, M. NOUREDDINE, J. R. GILBERT, N. SCHNETZ-BOUDAUD, A. AGARWAL, E. A. POSTEL, and M. A. PERICAK-VANCE, 2005 Complement factor H variant increases the risk of age-related macular degeneration. *Science* *308*(5720): 419–21.
- HALDANE, J., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of genetics* *8*(4): 299–309.
- HARIBABU, B. and R. SNYDERMAN, 1993 Identification of additional members of human G-protein-coupled receptor kinase multigene family. *Proc Natl Acad Sci USA* *90*(20): 9398–9402.
- HASEMAN, J. K. and R. C. ELSTON, 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* *2*(1): 3–19.
- HOLLBORN, M., A. REICHENBACH, P. WIEDEMANN, and L. KOHEN, 2004 Contrary effects of cytokines on mRNAs of cell cycle- and ECM-related proteins in hRPE cells in vitro. *Curr Eye Res* *28*(3): 215–223.
- HOWSON, J. M., B. J. BARRATT, J. A. TODD, and H. J. CORDELL, 2005 Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. *Genet Epidemiol* *29*(1): 51–67.
- HUANG, Q., S. SHETE, and C. CI AMOS, 2004 Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* *75*(6): 1106–1112.



- HUGHES, A. E., N. ORR, H. ESFANDIARY, M. DIAZ-TORRES, T. GOODSHIP, and U. CHAKRAVARTHY, 2006 A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat Genet* 38(10): 1173–7.
- IYENGAR, S. K., D. SONG, B. E. KLEIN, R. KLEIN, J. H. SCHICK, J. HUMPHREY, C. MILLARD, R. LIPTAK, K. RUSSO, G. JUN, K. E. LEE, B. FIJAL, and R. C. ELSTON, 2004 Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. *Am J Hum Genet* 74(1): 20–39.
- JACQUARD, A., 1966 Logique du calcul des coefficients d’identiti entre deux individuals. *Population* (French Edition) 21e(4): 751–776.
- JACQUARD, A., 1974 *The genetic structure of populations*. New York: Springer-Verlag.
- JAKOBSDOTTIR, J., Y. P. CONLEY, D. E. WEEKS, R. E. FERRELL, and M. B. GORIN, 2008 C2 and CFB genes in age-related maculopathy and joint action with CFH and LOC387715 genes. *PLoS ONE* 3(5): e2199.
- JAKOBSDOTTIR, J., Y. P. CONLEY, D. E. WEEKS, T. S. MAH, R. E. FERRELL, and M. B. GORIN, 2005 Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet* 77(3): 389–407.
- JAKOBSDOTTIR, J., M. B. GORIN, Y. P. CONLEY, R. E. FERRELL, and D. E. WEEKS, 2009 Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers. *PLoS Genetics* 5(2): e1000337.
- JANES, H. and M. PEPE, 2006 The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. *Biostatistics* 7(3): 456–68.
- JANSSENS, A. C., Y. S. AULCHENKO, S. ELEFANTE, G. J. BORSBOOM, E. W. STEYERBERG, and C. M. VAN DUIJN, 2006 Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 8(7): 395–400.
- JANSSENS, A. C., R. MOONESINGHE, Q. YANG, E. W. STEYERBERG, C. M. VAN DUIJN, and M. J. KHOURY, 2007 The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet Med* 9(8): 528–35.
- JOHNSON, L. V., W. P. LEITNER, M. K. STAPLES, and D. H. ANDERSON, 2001 Complement activation and inflammatory processes in Drusen formation and age related macular degeneration. *Exp Eye Res* 73(6): 887–96.
- JOHNSON, L. V., S. OZAKI, M. K. STAPLES, P. A. ERICKSON, and D. H. ANDERSON, 2000 A potential role for immune complex pathogenesis in drusen formation. *Exp Eye Res* 70(4): 441–9.
- JULIAN-REYNIER, C., I. NIPPERT, J. M. CALEFATO, H. HARRIS, U. KRISTOFFERSSON, J. SCHMIDTKE, L. TEN KATE, E. ANIONWU, C. BENJAMIN, K. CHALLEN, A. M. PLASS, and R. HARRIS, 2008 Genetics in clinical practice: general practitioners’ educational priorities in European countries. *Genet Med* 10(2): 107–13.
- KANDA, A., W. CHEN, M. OTHMAN, K. E. BRANHAM, M. BROOKS, R. KHANNA, S. HE, R. LYONS, G. R. ABECASIS, and A. SWAROOP, 2007 A variant of mitochondrial protein

- LOC387715/ ARMS2, not HTRA1, is strongly associated with age-related macular degeneration. *Proc Natl Acad Sci U S A* *104*(41): 16227–32.
- KATHIRESAN, S., O. MELANDER, D. ANEVSKI, C. GUIDUCCI, N. P. BURTT, C. ROOS, J. N. HIRSCHHORN, G. BERGLUND, B. HEDBLAD, L. GROOP, D. M. ALTSHULER, C. NEWTON-CHEH, and M. ORHO-MELANDER, 2008 Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* *358*(12): 1240–9.
- KAUR, I., A. HUSSAIN, N. HUSSAIN, T. DAS, A. PATHANGAY, A. MATHAI, A. HUSSAIN, R. NUTTHETI, P. K. NIRMALAN, and S. CHAKRABARTI, 2006 Analysis of CFH, TLR4, and APOE polymorphism in India suggests the Tyr402His variant of CFH to be a global marker for age-related macular degeneration. *Invest Ophthalmol Vis Sci* *47*(9): 3729–35.
- KENEALY, S. J., S. SCHMIDT, A. AGARWAL, E. A. POSTEL, M. A. DE LA PAZ, M. A. PERICAK-VANCE, and J. L. HAINES, 2004 Linkage analysis for age-related macular degeneration supports a gene on chromosome 10q26. *Mol Vis* *10*: 57–61.
- KENT, J. W., J., T. D. DYER, and J. BLANGERO, 2005 Estimating the additive genetic effect of the X chromosome. *Genet Epidemiol* *29*(4): 377–88.
- KLAVER, C. C., M. KLIFFEN, C. M. VAN DUIJN, A. HOFMAN, M. CRUTS, D. E. GROBBEE, C. VAN BROECKHOVEN, and P. T. DE JONG, 1998 Genetic association of apolipoprotein E with age-related macular degeneration. *Am J Hum Genet* *63*(1): 200–6.
- KLEIN, M. L., D. W. SCHULTZ, A. EDWARDS, T. C. MATISE, K. RUST, C. B. BERSELLI, K. TRZUPEK, R. G. WELEBER, J. OTT, M. K. WIRTZ, and T. S. ACOTT, 1998 Age-related macular degeneration. Clinical features in a large family and linkage to chromosome 1q. *Arch Ophthalmol* *116*(8): 1082–8.
- KLEIN, R., B. E. KLEIN, E. K. MARINO, L. H. KULLER, C. FURBERG, G. L. BURKE, and L. D. HUBBARD, 2003 Early age-related maculopathy in the cardiovascular health study. *Ophthalmology* *110*(1): 25–33.
- KLEIN, R. J., C. ZEISS, E. Y. CHEW, J. Y. TSAI, R. S. SACKLER, C. HAYNES, A. K. HENNING, J. P. SANGIOVANNI, S. M. MANE, S. T. MAYNE, M. B. BRACKEN, F. L. FERRIS, J. OTT, C. BARNSTABLE, and J. HOH, 2005 Complement factor H polymorphism in age-related macular degeneration. *Science* *308*(5720): 385–9.
- KONG, A. and N. COX, 1997 Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* *61*(5): 1179–88.
- KONG, X., K. MURPHY, T. RAJ, C. HE, P. S. WHITE, and T. C. MATISE, 2004 A combined linkage-physical map of the human genome. *Am J Hum Genet* *75*(6): 1143–1148.
- KOSAMBI, D., 1944 The estimation of map distances from recombination values. *Annals of Eugenics* *12*: 172–175.
- KRUGLYAK, L., M. DALY, M. REEVE-DALY, and E. LANDER, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Gene* *58*(6): 1347–1363.

- LANGE, K., 2003 *Mathematical and statistical methods for genetic analysis* (2 ed.). Statistics for biology and health. New York: Springer-Verlag.
- LANGE, K., R. M. RM CANTOR, S. HORVATH, M. PEROLA, C. SABATTI, J. SINSHEIMER, and E. SOBEL, 2001 MENDEL version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. , , Am J Hum Genet Suppl **69**: A1886.
- LANGE, K. and E. SOBEL, 2006 Variance component models for X-linked QTLs. Genet Epidemiol *30*(5): 380–3.
- LAU, L. I., S. J. CHEN, C. Y. CHENG, M. Y. YEN, F. L. LEE, M. W. LIN, W. M. HSU, and Y. H. WEI, 2006 Association of the Y402H polymorphism in complement factor H gene and neovascular age-related macular degeneration in Chinese patients. Invest Ophthalmol Vis Sci *47*(8): 3242–6.
- LEBREC, J., H. PUTTER, and J. C. HOUWELINGEN, 2004 Score test for detecting linkage to complex traits in selected samples. Genet Epidemiol *27*(2): 97–108.
- LI, C., L. J. SCOTT, and M. M BOEHNKE, 2004 Assessing whether an allele can account in part for a linkage signal: the Genotype-IBD Sharing Test (GIST). Am J Hum Genet *74*(3): 418–431.
- LI, H., 2001 A permutation procedure for the haplotype method for identification of disease-predisposing variants. Ann Hum Genet *65*(2): 189–196.
- LI, M., P. ATMACA-SONMEZ, M. OTHMAN, K. E. BRANHAM, R. KHANNA, M. S. WADE, Y. LI, L. LIANG, S. ZAREPARSI, A. SWAROOP, and G. R. ABECASIS, 2006 CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. Nat Genet *38*(9): 1049–54.
- LOHMUELLER, K. E., C. L. PEARCE, M. PIKE, E. S. LANDER, and J. N. HIRSCHHORN, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet *33*(2): 177–82.
- LOU, X. Y., G. B. CHEN, L. YAN, J. Z. MA, J. ZHU, R. C. ELSTON, and M. D. LI, 2007 A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet *80*(6): 1125–37.
- LU, Q. and R. C. ELSTON, 2008 Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet *82*(3): 641–51.
- MAGNUSSON, K. P., S. DUAN, H. SIGURDSSON, H. PETURSSON, Z. YANG, Y. ZHAO, P. S. BERNSTEIN, J. GE, F. JONASSON, E. STEFANSSON, G. HELGADOTTIR, N. A. ZABRISKIE, T. JONSSON, A. BJORNSSON, T. THORLACIUS, P. V. JONSSON, G. THORLEIFSSON, A. KONG, H. STEFANSSON, K. ZHANG, K. STEFANSSON, and J. R. GULCHER, 2006 CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD. PLoS Med *3*(1): e5.
- MAJEWSKI, J., D. W. SCHULTZ, R. G. WELEBER, M. B. SCHAIN, A. O. EDWARDS, T. C. MATISE, T. S. ACOTT, J. OTT, and M. L. KLEIN, 2003 Age-related macular degeneration—a genome scan in extended families. Am J Hum Genet *73*(3): 540–50.

- MALLER, J., S. GEORGE, S. PURCELL, J. FAGERNESS, D. ALTSHULER, M. J. DALY, and J. M. SEDDON, 2006 Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet* *38*(9): 1055–9.
- MALLER, J. B., J. A. FAGERNESS, R. C. REYNOLDS, B. M. NEALE, M. J. DALY, and J. M. SEDDON, 2007 Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet* *39*(10): 1200–1201.
- MALCOT, G., 1948 *Les mathématiques de l'hérédité*. Paris: Masson et Cie.
- MITKA, M., 1998 Genetics research already touching your practice. *American Medical News* **April 6, 1998; News section 3**.
- MORATZ, C., K. HARRISON, and J. H. KEHRL, 2004 Regulation of chemokine-induced lymphocyte migration by RGS proteins. *Methods Enzymol* **389**: 15–32.
- MORI, K., K. HORIE-INOUE, M. KOHDA, I. KAWASAKI, P. L. GEHLBACH, T. AWATA, S. YONEYA, Y. OKAZAKI, and S. INOUE, 2007 Association of the HTRA1 gene variant with age-related macular degeneration in the Japanese population. *J Hum Genet* *52*(7): 636–41.
- MORTON, N., 1955 Sequential tests for the detection of linkage. *Am J Hum Genet* *7*(3): 277–318.
- MOSKOWITZ, C. S. and M. S. PEPE, 2004 Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Stat Med* *23*(10): 1555–70.
- MUKHOPADHYAY, N., L. ALMASY, M. SCHROEDER, W. P. MULVIHILL, and D. E. WEEKS, 2005 Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* *21*(10): 2556–2557.
- MUKHOPADHYAY, N., S. G. BUXBAUM, and D. E. WEEKS, 2004 Comparative study of multipoint methods for genotype error detection. *Hum Hered* *58*(3-4): 175–89.
- MULLINS, R. F., S. R. RUSSELL, D. H. ANDERSON, and G. S. HAGEMAN, 2000 Drusen associated with aging and age-related macular degeneration contain proteins common to extracellular deposits associated with atherosclerosis, elastosis, amyloidosis, and dense deposit disease. *Faseb J* *14*(7): 835–46.
- MURWANTOKO, M. M YANO, Y. UETA, A. MURASAKI, H. KANDA, C. OKA, and M. M KAWAICHI, 2004 Binding of proteins to the PDZ domain regulates proteolytic activity of HtrA1 serine protease. *Biochem J* *381*(3): 895–904.
- NARAYANAN, R., V. BUTANI, D. S. BOYER, S. R. ATILANO, G. P. RESENDE, D. S. KIM, S. CHAKRABARTI, B. D. KUPPERMANN, N. KHATIBI, M. CHWA, A. B. NESBURN, and M. C. KENNEY, 2007 Complement factor H polymorphism in age-related macular degeneration. *Ophthalmology* *114*(7): 1327–31.
- NORMAND, S. L., 1999 Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med* *18*(3): 321–59.
- NORTH, B. V., D. CURTIS, and P. C. SHAM, 2005 Application of logistic regression to case-control association studies involving two causative loci. *Hum Hered* *59*(2): 79–87.

- O'CONNELL, J. R. and D. E. WEEKS, 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* *63*(1): 259–66.
- OKA, C., R. TSUJIMOTO, M. KAJIKAWA, K. KOSHIBA-TAKEUCHI, J. INA, M. YANO, A. TSUCHIYA, Y. UETA, A. SOMA, H. KANDA, M. MATSUMOTO, and M. KAWAICHI, 2004 HtrA1 serine protease inhibits signaling mediated by Tgf $\beta$  family proteins. *Development* *131*(5): 1041–1053.
- OKAMOTO, H., S. UMEDA, M. OBAYAZAWA, M. MINAMI, T. NODA, A. MIZOTA, M. HONDA, M. TANAKA, R. KOYAMA, I. TAKAGI, Y. SAKAMOTO, Y. SAITO, Y. MIYAKE, and T. IWATA, 2006 Complement factor H polymorphisms in Japanese population with age-related macular degeneration. *Mol Vis* *12*: 156–8.
- OTT, J., 1983 Linkage analysis and family classification under heterogeneity. *Ann Hum Genet* *47*(Pt 4): 311–320.
- OTT, J., 1999 *Analysis of human genetic linkage*. Baltimore: Johns Hopkins University Press.
- PAN, L., C. OBER, and M. ABNEY, 2007 Heritability estimation of sex-specific effects on human quantitative traits. *Genet Epidemiol* *31*(4): 338–47.
- PARKES, M., J. BARRETT, N. PRESCOTT, M. TREMELLING, C. ANDERSON, S. FISHER, R. ROBERTS, E. NIMMO, F. CUMMINGS, D. SOARS, H. DRUMMOND, C. LEES, S. KHAWAJA, R. BAGNALL, D. BURKE, C. TODHUNTER, T. AHMAD, C. ONNIE, W. MCARDLE, D. STRACHAN, G. BETHEL, C. BRYAN, C. LEWIS, P. DELOUKAS, A. FORBES, J. SANDERSON, D. JEWELL, J. SATSANGI, W. T. C. C. C. MANSFIELD, J.C, L. CARDON, and C. MATHEW, 2007 Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* *39*(7): 830–832.
- PAWITAN, Y., 2001 *In all likelihood: Statistical modelling and inference using likelihood* (1 ed.). Oxford University Press.
- PENCINA, M. J., S. D'AGOSTINO, R. B., J. D'AGOSTINO, R. B., and R. S. VASAN, 2008 Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* *27*(2): 157–72; discussion 207–12.
- PEPE, M. S., H. JANES, G. LONGTON, W. LEISENRING, and P. NEWCOMB, 2004 Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* *159*(9): 882–90.
- PEPE, M. S., Y. ZHENG, Y. JIN, Y. HUANG, C. R. PARIKH, and W. C. LEVY, 2008 Evaluating the ROC performance of markers for future events. *Lifetime Data Anal* *14*(1): 86–113.
- PURCELL, S. and P. SHAM, 2002 Variance components models for gene-environment interaction in quantitative trait locus linkage analysis. *Twin Res* *5*(6): 572–6.
- PUTTER, H., L. A. SANDKUIJL, and J. C. VAN HOUWELINGEN, 2002 Score test for detecting linkage to quantitative traits. *Genet Epidemiol* *22*(4): 345–55.
- R DEVELOPMENT CORE TEAM, 2005 R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.

- RAKOCZY, P., M. YU, S. NUSINOWITZ, B. CHANG, and J. HECKENLIVELY, 2006 Mouse models of age-related macular degeneration. *Exp Eye Res* 82(5): 741–52.
- RINALDO, A., S. A. BACANU, B. DEVLIN, V. SONPAR, L. WASSERMAN, and K. ROEDER, 2005 Characterization of multilocus linkage disequilibrium. *Genet Epidemiol* 28(3): 193–206.
- RIOUX, J., R. XAVIER, K. TAYLOR, M. SILVERBERG, P. GOYETTE, A. HUETT, T. GREEN, P. KUBALLA, M. BARMADA, L. DATTA, Y. SHUGART, A. GRIFFITHS, S. TARGAN, I. A.F., E. BERNARD, L. MEI, D. NICOLAE, M. REGUEIRO, L. SCHUMM, A. STEINHART, J. ROTTER, R. DUERR, J. CHO, M. DALY, and S. BRANT, 2007 Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 39(5): 596–604.
- RISCH, N., 2001 Implications of multilocus inheritance for gene-disease association studies. *Theor Popul Biol* 60(3): 215–20.
- ITCHIE, M. D., L. W. HAHN, N. ROODI, L. R. BAILEY, W. D. DUPONT, F. F. PARL, and J. H. MOORE, 2001 Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69(1): 138–47.
- RIVERA, A., S. A. FISHER, L. G. FRITSCH, C. N. KEILHAUER, P. LICHTNER, T. MEITINGER, and B. H. WEBER, 2005 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* 14(21): 3227–36.
- ROSNER, B., 2000 *Fundamentals of biostatistics*. Pacific Grove, CA: Duxbury.
- ROSS, R. J., C. M. BOJANOWSKI, J. J. WANG, E. Y. CHEW, E. ROCHTCHINA, R. FERRIS, F. L., P. MITCHELL, C. C. CHAN, and J. TUO, 2007 The LOC387715 polymorphism and age-related macular degeneration: replication in three case-control samples. *Invest Ophthalmol Vis Sci* 48(3): 1128–32.
- ROSS, R. J., V. VERMA, K. I. ROSENBERG, C. C. CHAN, and J. TUO, 2007 Genetic markers and biomarkers for age-related macular degeneration. *Expert Rev Ophthalmol* 2(3): 443–457.
- SANTANGELO, S. L., C.-H. YEN, S. HADDAD, J. FAGERNESS, C. HUANG, and J. M. SEDDON, 2005 A discordant sib-pair linkage analysis of age-related macular degeneration. *Ophthalmic Genet* 26(2): 61–67.
- SCHAID, D. J., C. M. ROWLAND, D. E. TINES, R. M. JACOBSON, and G. A. POLAND, 2002 Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70(2): 425–34.
- SCHAUMBERG, D. A., W. G. CHRISTEN, P. KOZLOWSKI, D. T. MILLER, P. M. RIDKER, and R. Y. ZEE, 2006 A prospective assessment of the Y402H variant in complement factor H, genetic variants in C-reactive protein, and risk of age-related macular degeneration. *Invest Ophthalmol Vis Sci* 47(6): 2336–40.
- SCHAUMBERG, D. A., S. E. HANKINSON, Q. GUO, E. RIMM, and D. J. HUNTER, 2007 A prospective study of 2 major age-related macular degeneration susceptibility alleles and interactions with modifiable risk factors. *Arch Ophthalmol* 125(1): 55–62.

- SCHICK, J. H., S. K. IYENGAR, B. E. KLEIN, R. KLEIN, K. READING, R. LIPTAK, C. MILLARD, K. E. LEE, S. C. TOMANY, E. L. MOORE, B. A. FIJAL, and R. C. ELSTON, 2003 A whole-genome screen of a quantitative trait of age-related maculopathy in sibships from the Beaver Dam Eye Study. *Am J Hum Genet* 72(6): 1412–1424.
- SCHMIDT, S., M. A. HAUSER, W. K. SCOTT, E. A. POSTEL, A. AGARWAL, P. GALLINS, F. WONG, Y. S. CHEN, K. SPENCER, N. SCHNETZ-BOUTAUD, J. L. HAINES, and M. A. PERICAK-VANCE, 2006 Cigarette smoking strongly modifies the association of LOC387715 and age-related macular degeneration. *Am J Hum Genet* 78(5): 852–64.
- SCHMIDT, S., C. KLAVER, A. SAUNDERS, E. POSTEL, M. DE LA PAZ, A. AGARWAL, K. SMALL, N. UDAR, J. ONG, M. CHALUKYA, A. NESBURN, C. KENNEY, R. DOMURATH, M. HOGAN, T. MAH, Y. CONLEY, R. FERRELL, D. WEEKS, P. T. DE JONG, C. VAN DUIJN, J. HAINES, M. PERICAK-VANCE, and M. GORIN, 2002 A pooled case-control study of the apolipoprotein E (APOE) gene in age-related maculopathy. *Ophthalmic Genet* 23(4): 209–23.
- SCHMIDT, S., A. M. SAUNDERS, M. A. DE LA PAZ, E. A. POSTEL, R. M. HEINIS, A. AGARWAL, W. K. SCOTT, J. R. GILBERT, J. G. MCDOWELL, A. BAZYK, J. D. GASS, J. L. HAINES, and M. A. PERICAK-VANCE, 2000 Association of the apolipoprotein E gene with age-related macular degeneration: possible effect modification by family history, age, and gender. *Mol Vis* 6: 287–93.
- SCHMIDT, S., W. K. SCOTT, E. A. POSTEL, A. AGARWAL, E. R. HAUSER, M. A. DE LA PAZ, J. R. GILBERT, D. E. WEEKS, M. B. GORIN, J. L. HAINES, and M. A. PERICAK-VANCE, 2004 Ordered subset linkage analysis supports a susceptibility locus for age-related macular degeneration on chromosome 16p12. *BMC Genet* 5: 18.
- SCHORK, N. J., 1993 Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet* 53(6): 1306–19.
- SCOTT, L. J., K. L. MOHLKE, L. L. BONNYCASTLE, C. J. WILLER, Y. LI, W. L. DUREN, M. R. ERDOS, H. M. STRINGHAM, P. S. CHINES, A. U. JACKSON, L. PROKUNINA-OLSSON, C. J. DING, A. J. SWIFT, N. NARISU, T. HU, R. PRUIM, R. XIAO, X. Y. LI, K. N. CONNEELY, N. L. RIEBOW, A. G. SPRAU, M. TONG, P. P. WHITE, K. N. HETRICK, M. W. BARNHART, C. W. BARK, J. L. GOLDSTEIN, L. WATKINS, F. XIANG, J. SARAMIES, T. A. BUCHANAN, R. M. WATANABE, T. T. VALLE, L. KINNUNEN, G. R. ABECASIS, E. W. PUGH, K. F. DOHENY, R. N. BERGMAN, J. TUOMILEHTO, F. S. COLLINS, and M. BOEHNKE, 2007 A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316(5829): 1341–5.
- SEDDON, J. M. and C. A. CHEN, 2004 The epidemiology of age-related macular degeneration. *Int Ophthalmol Clin* 44(4): 17–39.
- SEDDON, J. M., J. COTE, W. F. PAGE, S. H. AGGEN, and M. C. NEALE, 2005 The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Arch Ophthalmol* 123(3): 321–7.
- SEDDON, J. M., P. J. FRANCIS, S. GEORGE, D. W. SCHULTZ, B. ROSNER, and M. L. KLEIN, 2007 Association of CFH Y402H and LOC387715 A69S with progression of age-related macular degeneration. *Jama* 297(16): 1793–800.

- SEDDON, J. M., S. L. SANTANGELO, K. BOOK, S. CHONG, and J. COTE, 2003 A genomewide scan for age-related macular degeneration provides evidence for linkage to several chromosomal regions. *Am J Hum Genet* **73**(4): 780–90.
- SEITSONEN, S., S. LEMMELA, J. HOLOPAINEN, P. TOMMILA, P. RANTA, A. KOTAMIES, J. MOILANEN, T. PALOSAARI, K. KAARNIRANTA, S. MERI, I. IMMONEN, and I. JARVELA, 2006 Analysis of variants in the complement factor H, the elongation of very long chain fatty acids-like 4 and the hemicentin 1 genes of age-related macular degeneration in the Finnish population. *Mol Vis* **12**: 796–801.
- SELF, S. and K. LIANG, 2007 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc* **82**(398): 605–610.
- SEPP, T., J. C. KHAN, D. A. THURLBY, H. SHAHID, D. G. CLAYTON, A. T. MOORE, A. C. BIRD, and J. R. YATES, 2006 Complement factor H variant Y402H is a major risk determinant for geographic atrophy and choroidal neovascularization in smokers and nonsmokers. *Invest Ophthalmol Vis Sci* **47**(2): 536–40.
- SHI, G. X., K. HARRISON, S. B. HAN, C. MORATZ, and J. H. JH KEHRL, 2004 Toll-like receptor signaling alters the expression of regulator of G protein signaling proteins in dendritic cells: implications for G protein-coupled receptor signaling. *J Immunol* **172**(9): 5175–5184.
- SHIH, M. and A. WHITTEMORE, 2001 Allele-sharing among affected relatives: non-parametric methods for identifying genes. *Stat Methods Med Res* **10**(1): 27–55.
- SIEGMUND, D. and B. YAKIR, 2007 *The statistics of gene mapping*. New York: Springer.
- SIMONELLI, F., G. FRISSE, F. TESTA, R. DI FIORE, D. F. VITALE, M. P. MANITTO, R. BRANCATO, E. RINALDI, and L. SACCHETTI, 2006 Polymorphism p.402Y<sub>L</sub>H in the complement factor H protein is a risk factor for age related macular degeneration in an Italian population. *Br J Ophthalmol* **90**(9): 1142–5.
- SIVAPRASAD, S., T. ADEWOYIN, T. A. BAILEY, S. S. DANDEKAR, S. JENKINS, A. R. WEBSTER, and N. V. CHONG, 2007 Estimation of systemic complement C3 activity in age-related macular degeneration. *Arch Ophthalmol* **125**(4): 515–9.
- SLADEK, R., G. ROCHELEAU, J. RUNG, C. DINA, L. SHEN, D. SERRE, P. BOUTIN, D. VINCENT, A. BELISLE, S. HADJADJ, B. BALKAU, B. HEUDE, G. CHARPENTIER, T. J. HUDSON, A. MONTPETIT, A. V. PSHEZHETSKY, M. PRENTKI, B. I. POSNER, D. J. BALDING, D. MEYRE, C. POLYCHRONAKOS, and P. FROGUEL, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**(7130): 881–5.
- SOUIED, E. H., N. LEVEZIEL, F. RICHARD, M. A. DRAGON-DUREY, G. COSCAS, G. SOUBRANE, P. BENLIAN, and V. FREMEAUX-BACCHI, 2005 Y402H complement factor H polymorphism associated with exudative age-related macular degeneration in the French population. *Mol Vis* **11**: 1135–40.
- SPENCER, K. L., M. A. HAUSER, L. M. OLSON, S. SCHMIDT, W. K. SCOTT, P. GALLINS, A. AGARWAL, E. A. POSTEL, M. A. PERICAK-VANCE, and J. L. HAINES, 2007 Protective



- Effect of Complement Factor B and Complement Component 2 Variants in Age-related Macular Degeneration. *Hum Mol Genet* *16*(16): 1986–92.
- STRAUCH, K., R. FIMMERS, M. BAUR, and T. WIENKER, 2003 How to model a complex trait. 1. General considerations and suggestions. *Hum Hered* *55*(4): 202–210.
- SWAROOP, A., K. E. BRANHAM, W. CHEN, and G. ABECASIS, 2007 Genetic susceptibility to age-related macular degeneration: a paradigm for dissecting complex disease traits. *Hum Mol Genet* **16 Spec No. 2**: R174–82.
- SZATKIEWICZ, J., 2004 Mapping genes for quantitative traits using selected samples of sibling pairs. Ph. D. thesis, University of Pittsburgh.
- TANG, H. K. and D. SIEGMUND, 2001 Mapping quantitative trait loci in oligogenic models. *Biostatistics* *2*(2): 147–62.
- TANIMOTO, S., H. TAMURA, T. UE, K. YAMANE, H. MARUYAMA, H. KAWAKAMI, and Y. KIUCHI, 2007 A polymorphism of LOC387715 gene is associated with age-related macular degeneration in the Japanese population. *Neurosci Lett* *414*(1): 71–4.
- TEDESCHI-BLOK, N., J. BUCKLEY, R. VARMA, T. J. TRICHE, and D. R. HINTON, 2007 Population-based study of early age-related macular degeneration: role of the complement factor H Y402H polymorphism in bilateral but not unilateral disease. *Ophthalmology* *114*(1): 99–103.
- THE INTERNATIONAL HAPMAP CONSORTIUM, 2003 The International HapMap Project. *Nature* *426*(6968): 789–796.
- THE INTERNATIONAL HAPMAP CONSORTIUM, 2005 A haplotype map of the human genome. *Nature* *437*(7063): 1299–1320.
- THOMAS, D., 2004 *Statistical methods in genetic epidemiology*. New York: Oxford University Press.
- THOMPSON, C. L., B. E. KLEIN, R. KLEIN, Z. XU, J. CAPRIOTTI, T. JOSHI, D. LEONTIEV, K. E. LEE, R. C. ELSTON, and S. K. IYENGAR, 2007 Complement Factor H and Hemicentin-1 in Age-Related Macular Degeneration and Renal Phenotypes. *Hum Mol Genet* *16*(17): 2135–2148.
- THORNTON, J., R. EDWARDS, P. MITCHELL, R. A. HARRISON, I. BUCHAN, and S. P. KELLY, 2005 Smoking and age-related macular degeneration: a review of association. *Eye* *19*(9): 935–44.
- THORNTON, T. and M. MCPEEK, 2007 Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet* *81*(2): 321–37.
- TUO, J., B. NING, C. M. BOJANOWSKI, Z. N. LIN, R. J. ROSS, G. F. REED, D. SHEN, X. JIAO, M. ZHOU, E. Y. CHEW, F. F. KADLUBAR, and C. C. CHAN, 2006 Synergic effect of polymorphisms in ERCC6 5' flanking region and complement factor H on age-related macular degeneration predisposition. *Proc Natl Acad Sci U S A* *103*(24): 9256–61.
- UKA, J., H. TAMURA, T. KOBAYASHI, K. YAMANE, H. KAWAKAMI, A. MINAMOTO, and H. K. MISHIMA, 2006 No association of complement factor H gene polymorphism and age-related macular degeneration in the Japanese population. *Retina* *26*(9): 985–7.

- VALDES, A. M. and G. THOMSON, 1997 Detecting disease-predisposing variants: the haplotype method. *Am J Hum Genet* *60*(3): 703–16.
- VAN DER VAAR, A., 1998 *Asymptotic statistics* (1 ed.). Caimbridge University Press.
- VAN HOUWELINGEN, H. C., L. R. ARENDS, and T. STIJNEN, 2002 Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* *21*(4): 589–624.
- VAN LEEUWEN, R., C. C. KLAVER, J. R. VINGERLING, A. HOFMAN, and P. T. DE JONG, 2003 Epidemiology of age-related maculopathy: a review. *Eur J Epidemiol* *18*(9): 845–54.
- WALTER, S. D., 1975 The distribution of Levin’s measure of attributable risk. *Biometrika* *62*(2): 371–372.
- WANG, J. J., R. J. ROSS, J. TUO, G. BURLUTSKY, A. G. TAN, C. C. CHAN, E. J. FAVALORO, A. WILLIAMS, and P. MITCHELL, 2007 The LOC387715 Polymorphism, Inflammatory Markers, Smoking, and Age-Related Macular Degeneration A Population-Based Case-Control Study. *Ophthalmology* *115*(4): 693–699.
- WANG, K., 2002 Efficient score statistics for mapping quantitative trait loci with extended pedigrees. *Hum Hered* *54*(2): 57–68.
- WANG, K. and J. HUANG, 2002a A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *Am J Hum Genet* *70*(2): 412–24.
- WANG, K. and J. HUANG, 2002b Score test for mapping quantitative-trait loci with sibships of arbitrary size when the dominance effect is not negligible. *Genet Epidemiol* *23*(4): 398–412.
- WARNES, G. and L. FRIEDRICH, 2006 *Population Genetics*.
- WEEDON, M. N., M. I. MCCARTHY, G. HITMAN, M. WALKER, C. J. GROVES, E. ZEGGINI, N. W. RAYNER, B. SHIELDS, K. R. OWEN, A. T. HATTERSLEY, and T. M. FRAYLING, 2006 Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med* *3*(10): e374.
- WEEKS, D. E., Y. P. CONLEY, T. S. MAH, T. O. PAUL, L. MORSE, J. NGO-CHANG, J. P. DAILEY, R. E. FERRELL, and M. B. GORIN, 2000 A full genome scan for age-related maculopathy. *Hum Mol Genet* *9*(9): 1329–49.
- WEEKS, D. E., Y. P. CONLEY, H. J. TSAI, T. S. MAH, P. J. ROSENFELD, T. O. PAUL, A. W. ELLER, L. S. MORSE, J. P. DAILEY, R. E. FERRELL, and M. B. GORIN, 2001 Age-related maculopathy: an expanded genome-wide scan with evidence of susceptibility loci within the 1q31 and 17q25 regions. *Am J Ophthalmol* *132*(5): 682–92.
- WEEKS, D. E., Y. P. CONLEY, H. J. TSAI, T. S. MAH, S. SCHMIDT, E. A. POSTEL, A. AGARWAL, J. L. HAINES, M. A. PERICAK-VANCE, P. J. ROSENFELD, T. O. PAUL, A. W. ELLER, L. S. MORSE, J. P. DAILEY, R. E. FERRELL, and M. B. GORIN, 2004 Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. *Am J Hum Genet* *75*(2): 174–89.

- WEGSCHEIDER, B. J., M. WEGER, W. RENNER, I. STEINBRUGGER, W. MARZ, G. MOSSBOCK, W. TEMMEL, Y. EL-SHABRAWI, O. SCHMUT, R. JAHRBACHER, and A. HAAS, 2007 Association of complement factor H Y402H gene polymorphism with different subtypes of exudative age-related macular degeneration. *Ophthalmology* *114*(4): 738–42.
- WHITTEMORE, A. and J. HALPERN, 1994 A class of tests for linkage using affected pedigree members. *Biometrics* *50*(1): 118–127.
- WIENER, H., R. ELSTON, and H. TIWARI, 2003 X-linked extension of the revised Haseman-Elston algorithm for linkage analysis in sib pairs. *Hum Hered* *55*(2-3): 97–107.
- WIGGINTON, J. E. and G. R. ABECASIS, 2005 PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* *21*(16): 3445–7.
- WOLOSHIN, S., L. M. SCHWARTZ, W. C. BLACK, and H. G. WELCH, 1999 Women’s perceptions of breast cancer risk: how you ask matters. *Med Decis Making* *19*(3): 221–9.
- YANG, Z., N. J. CAMP, H. SUN, Z. TONG, D. GIBBS, D. J. CAMERON, H. CHEN, Y. ZHAO, E. PEARSON, X. LI, J. CHIEN, A. DEWAN, J. HARMON, P. S. BERNSTEIN, V. SHRIDHAR, N. A. ZABRISKIE, J. HOH, K. HOWES, and K. ZHANG, 2006 A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science* *314*(5801): 992–3.
- YATES, J. R., T. SEPP, B. K. MATHARU, J. C. KHAN, D. A. THURLBY, H. SHAHID, D. G. CLAYTON, C. HAYWARD, J. MORGAN, A. F. WRIGHT, A. M. ARMBRECHT, B. DHILLON, I. J. DEARY, E. REDMOND, A. C. BIRD, and A. T. MOORE, 2007 Complement C3 Variant and the Risk of Age-Related Macular Degeneration. *N Engl J Med* *357*(6): 553–561.
- YEAGER, M., N. ORR, R. B. HAYES, K. B. JACOBS, P. KRAFT, S. WACHOLDER, M. J. MINICHELLO, P. FEARNHEAD, K. YU, N. CHATTERJEE, Z. WANG, R. WELCH, B. J. STAATS, E. E. CALLE, H. S. FEIGELSON, M. J. THUN, C. RODRIGUEZ, D. ALBANES, J. VIRTAMO, S. WEINSTEIN, F. R. SCHUMACHER, E. GIOVANNUCCI, W. C. WILLETT, G. CANCEL-TASSIN, O. CUSSENOT, A. VALERI, G. L. ANDRIOLE, E. P. GELMANN, M. TUCKER, D. S. GERHARD, J. FRAUMENI, J. F., R. HOOVER, D. J. HUNTER, S. J. CHANOCK, and G. THOMAS, 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* *39*(5): 645–9.
- YOSHIDA, T., A. DEWAN, H. ZHANG, R. SAKAMOTO, H. OKAMOTO, M. MINAMI, M. OBAZAWA, A. MIZOTA, M. TANAKA, Y. SAITO, I. TAKAGI, J. HOH, and T. IWATA, 2007 HTRA1 promoter polymorphism predisposes Japanese to age-related macular degeneration. *Mol Vis* **13**: 545–8.
- YOUNG, I., 2007 *Introduction to risk calculation in genetic counseling* (3 ed.). Oxford University Press.
- ZAREPARSI, S., K. E. BRANHAM, M. LI, S. SHAH, R. J. KLEIN, J. OTT, J. HOH, G. R. ABECASIS, and A. SWAROOP, 2005 Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration. *Am J Hum Genet* *77*(1): 149–53.
- ZAREPARSI, S., M. M BURACZYNSKA, K. E. BRANHAM, S. SHAH, D. ENG, M. LI, H. H PAWAR, B. M. YASHAR, S. E. MOROI, P. R. LICHTER, H. R. PETTY, J. E. RICHARDS, G. R. ABECASIS, V. M. ELNER, and A. A SWAROOP, 2005 Toll-like receptor 4 variant D299G is

associated with susceptibility to age-related macular degeneration. *Hum Mol Genet* 14(11): 1449–1455.

ZAREPARSI, S., A. C. REDDICK, K. E. BRANHAM, K. B. MOORE, L. JESSUP, S. THOMS, M. SMITH-WHEELOCK, B. M. YASHAR, and A. SWAROOP, 2004 Association of apolipoprotein E alleles with susceptibility to age-related macular degeneration in a large cohort from a single center. *Invest Ophthalmol Vis Sci* 45(5): 1306–10.

ZEEGERS, M., F. RIJSDIJK, and P. SHAM, 2004 Adjusting for covariates in variance components QTL linkage analysis. *Behav Genet* 34(2): 127–33.

ZHOU, X., N. OBUCHOWSKI, and D. MCCLISH, 2002 *Statistical Methods in Diagnostic Medicine* (1 ed.). Wiley series in probability and statistics. New York: John Wiley & Sons, Inc.