UPTEC X 04 040
JUN 2004

ISSN 1401-2138

NICLAS KONGSHOLM

Use of logistic regression to model gene-gene interaction in case-control studies

Master’s degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 04 040	**Date of issue 2004-07**

Author

Niclas Kongsholm

Title (English)

Use of logistic regression to model gene-gene interaction in case-control studies

Title (Swedish)

Abstract

Three candidate genes of Multiple Sclerosis; IL7R, LAG3 and TIM3 have been analyzed for gene-gene interactions using a genotype-based logistic regression model. Our results suggest two interaction effects between the three genes and one interaction effect within TIM3. Hardy-Weinberg Disequilibrium (HWD) could not be ruled out in all markers, and blocks of Linkage Disequlibrium (LD) where observed within IL7R and TIM3, whereas LAG3 lacked a block structured LD pattern.

Keywords

Multiple Sclerosis, SNP, epistasis, logistic regression modelling, HWD, LD

Supervisors

Juni Palmgren
Medical Epidemiology and Biostatistics
KI, Stockholm

Hugh Salter
Genetics and Bioinformatics
AstraZeneca R&D, Södertälje

Scientific reviewer

Elena Jazin
Evolutionary Biology, Uppsala University

Project name	Sponsors
Language **English**	Security **Secret until January 2005**
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **45**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Use of Logistic Regression to Model Gene-Gene Interaction in Case-Control Studies

Niclas Kongsholm

SAMMANFATTNING

Dagens genteknik gör det möjligt att identifiera och jämföra delar av det mänskliga genomet mellan individer. Detta utnyttjas i associationsstudier, där genotypen hos utvalda markörer identifieras och jämförs hos sjuka och friska personer. Tidigare utförda associationsstudier har framgångsrikt lokaliserat enskilda upphovsgener till bland annat cystisk fibros och Parkinsons sjukdom. Flertalet sjukdomar befaras däremot orsakas av ett betydligt mer komplext nätverk av genetiskt arv och miljöfaktorer, vilket försvårar identifieringen av många sjukdomsgener. Ett exempel är multipel skleros (MS), där flertalet, troligtvis interagerande gener ger en förhöjd risk att insjukna i sjukdomen.

Gen-geninteraktioner förknippas i biologisk mening med uttrycket epistasis och definieras generellt som en process där ett genuttryck döljs eller förstärks i närvaro av ett annat genuttryck. Detta skiljer sig något från den matematiska definitionen av en interaktion som syftar på en avvikelse från additivitet.

I detta examensprojekt analyseras interaktionen mellan och inom tre MS-relaterade gener matematiskt genom att utnyttja en logistisk regression modell. Varje identifierad markör (SNP) representeras av en additiv och en dominant variabel och analyseras i två steg. I första steget selekteras de bäst beskrivande markörerna, varefter interaktionstermer hos de utvalda markörerna analyseras i nästa steg. Ytterligare en metod baserad på logistisk regression, främst avsedd för steg ett, presenteras. Dessutom har datamängden analyserats för alternativa förklaringar till påvisade interaktioner genom att utföra tester på Hardy-Weinberg samt kopplingsjämvikt.

Examensarbete 20p i Molekylär bioteknikprogrammet
Uppsala Universitet Juni 2004

CONTENTS

1. INTRODUCTION

Advances in molecular biotechnology and the race to complete the human genome project have started to allow the identification of diseases caused by single gene defects. Most human diseases however, tend to be under influence of multiple, and possibly interacting genes, i.e. they are complex diseases. Localizing such genes is of great importance for pharmaceutical companies because the genes that contribute to these traits might identify drug targets that are difficult to find otherwise.

In a joint project of *AstraZeneca R&D*, *Neurotec* and the department of *Medical Epidemiology and Biostatistics* at Karolinska Institutet in Stockholm, this project covers a detailed study on potential gene-gene interactions involved in the etiology of Multiple Sclerosis.

In a recently performed association study, 123 genotyped SNPs, located in 66 candidate genes were analyzed for significant differences in allele frequencies between 672 Nordic MS patients and 672 healthy Swedish controls [27]. The results suggested that Multiple Sclerosis is influenced by a few polymorphic cites within three genes: IL7R, LAG3 and TIM3. If these cites prove to be genuine susceptibility loci, the etiology of Multiple Sclerosis may be partly explained by gene-gene interactions, often referred to as *epistasis*. Although the definition of epistasis differs slightly between a biological and a mathematical model, we believe that the methods suggested in this thesis are capable of revealing significant gene-gene interactions in complex human diseases.

Analyzing genetically predisposing diseases demands a good understanding of both human and statistical genetics. Chapters 2 and 3 introduces the reader to some important issues and concepts in human and statistical genetics. A reader with modest biological and statistical background should have no problem understanding these initial chapters. Genetic and statistical terms are provided in a glossary, placed at the end of the paper.

Chapter 4 covers gene-gene interaction modelling in case-control studies. A genotype-based and a haplotype-based logistic regression model are proposed and explained in detail. Both methods are useful for evaluating the relative importance of SNPs within a small genetic region, and finding a parsimonious subset of marker loci likely to be closely associated with a disease susceptibility locus. In addition, the genotyped-based method is proposed for modelling epistatic interactions in case-control studies.

Next, Chapter 5 presents the sampled MS data and the initial association analysis performed prior to this project. Such conventional single-locus association methods do not consider gene-gene interactions but may reveal main susceptibility effects, presumably involved in epistasis.

Chapter 6 presents the results of the methods used in this degree project to model gene-gene interaction.

Last, Chapter 7 reviews the performed project, interprets the obtained re-

sults and discusses the continued challenge in localizing susceptibility genes causing complex human diseases.

2. HUMAN GENETICS

The year 2003 marked two major milestones in human genetics. The 50th anniversary of Watson and Cricks discovery of the DNA-helix, and the completion of the sequencing project of the human genome. These breakthroughs in human genetics have boosted interest not only in human genetics, but in interdisciplinary fields such as statistical genetics and bioinformatics. In time, continued research is expected to reveal the complete etiology and genetic variation in complex human diseases.

The following sections describe fundamental concepts in human molecular genetics, genetic variation, and the etiology of complex human disease.

2.1 Basic Concepts in Human Molecular Genetics

In humans, as in other higher organisms, every *cell* contain densely wrapped DNA structures called *chromosomes*. For most of the time, chromosomes are too elongated and tenuous to be seen under a microscope. Only during meiotic or mitotic cell divisions do all chromosomes take on condensed structures.

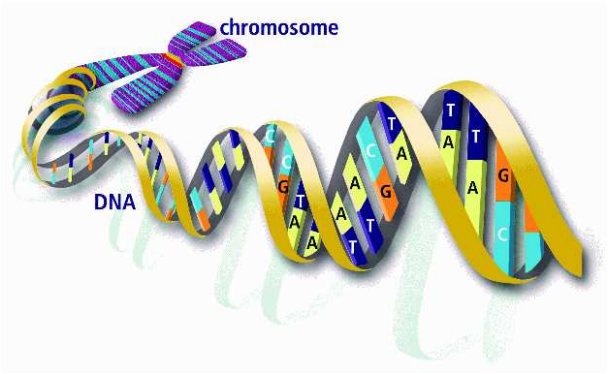

Fig. 2.1: **The molecule of life.** A schematic illustration of a condensed chromosome and the coiled structure of the DNA molecule. The DNA sequence is the particular side-by-side arrangement of bases along the DNA strand (e.g., ATTCCGGA). This order spells out the exact instructions required to create an unique individual.

Image credit: U.S. Department of Energy Human Genome Program, http://www.ornl.gov/hgmis.

Gametes are the only cells in the human body that contain a haploid chromosome set. Other cells, called *somatic cells*, contain a diploid chromosome set. The *haploid* chromosome set consist of 23 single chromosomes, whereas the *diploid* chromosome set consist of 23 pairs of parentally inherited chromosomes. One of the chromosome pairs are sex-linked and is responsible for the gender of an individual. Males carry a X and a Y chromosome, whereas females carry two X chromosomes. This is signified in Figure 2.2, illustrating the chromosomal inheritance caused by meiosis.

In *meiosis*, the chromosomes of a diploid cell are replicated and the cell divided into four gametic daughter cells. During this process, recombination events take place between *homologous* parental chromosomes giving rise to a wide variety of unique gametic cells. Occasionally, mainly during recombination, replication errors occur, increasing the genetic variability even further. If such an error is not repaired, but passed on to the next generation, a meiotic mutation is said to have taken place.

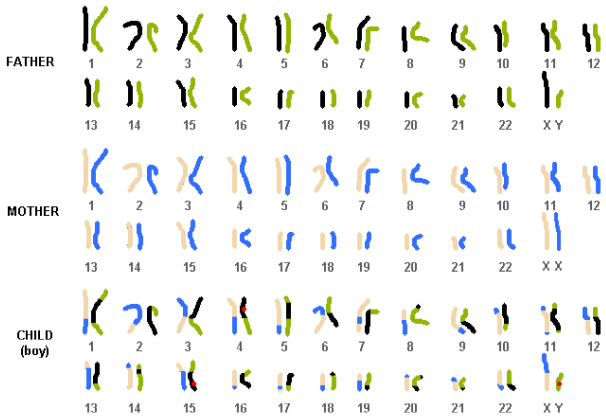

Fig. 2.2: **Chromosomal inheritance.** Illustrates the important meiotic recombination events and the random mutations, responsible for the characteristic genetic setup observed in an offspring. The different colors of the homologous chromosome pairs in the offspring indicates from which paternal chromosome the DNA sequence is inherited. The red dots illustrate random mutations due to irreversible replication errors, either during paternal meiosis or during the offsprings own embryological development. Note how the two sex chromosomes determine the gender of an individual and that no recombination event takes place for the Y chromosome.

In *mitosis*, the genetic material of somatic cells are replicated and the cell split into two daughter cells. This process starts immediately after the haploid genome of the oocyte and the sperm are united. In some cells mitosis continues throughout life, whereas in other cells replication is halted as soon as they are developed.

The DNA molecule, illustrated in Figure 2.1, consists of two strands that wrap around each other to resemble a twisted ladder. Each strand is a linear arrangement of repeating units called nucleotides, all composed of one sugar, one phosphate, and one of four nitrogenous bases: adenine (A), thymine (T),

cytosine (C), or guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence.

A *gene* is a specific stretch of the DNA sequence that carries genetic information, required for constructing proteins. Sometimes, the word gene is confused with two other common used terms in statistical genetics; locus and allele. A *locus* is a position on a chromosome, for example, a gene or a genetic marker. Such genes or genetic markers exist in alternative forms, called *alleles*. If the *phase* of a locus is known, the parental origin of the constituting alleles have been distinguished.

Proteins in turn are long chains of amino acids, that form cell structures and essential biochemical components, such as; enzymes, hormones and antibodies. Each amino acid is coded by a triplet of transcribed and processed RNA bases. In the processing step, the newly transcribed DNA sequence is spliced - that is, non-coding sequences (*introns*) are separated from coding sequences (*exons*).

Even though it is believed that the majority of genes, express both their maternal and paternal gene variant, there are exceptions. This is known as *transcriptional imprinting* and occurs when one parental gene variant is transcribed, whereas the other remains idle. The phenomenon is an imprint in the sense that the paternal gene variant is marked as being either maternal or paternal, such that the chromatin structure of that gene is retained in epigenesis [3].

In human genetics, three important concepts need special attention; phenotype, genotype and haplotype. A *phenotype* refers to the observable traits or characteristics of an individual, such as hair color, weight, or the presence or absence of a disease. Variation in phenotypic traits is not necessarily caused by genes. If you decide to dye your hair or change your diet, the phenotype will no longer be the same. A disease phenotype is no different; many environmental factors can cause or at least trigger a disease. The genetic material influencing an individual's phenotype can be described either by its genotype or by its haplotype. Whereas the *genotype* of an individual refers to pairs of alleles located across the maternally and paternally inherited chromosomes, the *haplotype* of an individual refers to the sequence of alleles located along each parentally inherited chromosome. Genotypes are either phased or unphased, whereas haplotypes are always phased. Figure 2.3 illustrates the difference between unphased/phased genotypes and phased haplotypes. Basically, an unphased genotype assumes that there is no 'haplotype effect' - that is, the parental origin of an allele is irrelevant. For a phased genotype on the other hand, the parental origin of an allele is distinguished such that the genotype abc/ABC is not the same as the genotype AbC/aBc.

Since no human has the same alleles at every loci, all individuals have an unique genome-wide genotype and haplotype. First when geneticists refer to genotypes and haplotypes on a smaller scale, are non-unique variants encountered. Throughout this thesis, the interpretation of a genotype or a haplotype is based on no more than 8 loci.

Sometimes alleles are described as either dominant or recessive for a trait. A *dominant* allele is one that influences the trait even if it is present in just one copy, whereas for a *recessive* allele two copies are needed. For example, consider a single locus with two allelic variants; a dominant allele, denoted A and a recessive allele, denoted a. Whenever an individual has a genotype involving an A - either AA, Aa, or aA - the dominant attribute coded by A is

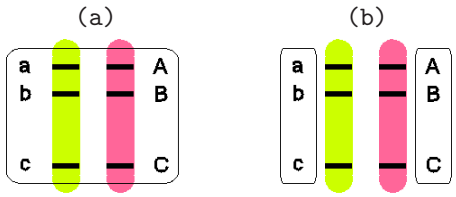

Fig. 2.3: **Difference between genotype and haplotype.** Both figures picture a chromosome segment with three loci, marked with the letters *a/A, b/B* and *c/C*. **(a)** For an unphased genotype it does not matter what chromosome a letter belongs to; *A/a* is equivalent to *a/A* and *aBc/AbC* is equivalent to *ABC/abc*. The left figure illustrates one genotype, namely abc/ABC, out of $2^6 = 64$ phased genotypes or $64/2 + 6/2 = 36$ unphased genotypes. **(b)** As illustrated in the left figure, haplotypes are always phased - that is, the origin of all alleles are distinguished, revealing an allele sequence along each chromosome, namely *abc* and *ABC*.

observed. The only way the recessive attribute, coded by *a* takes effect is when the individual carries an *aa* genotype.

If two identical alleles occupy a locus, an individual is said to be *homozygous* at that locus; if different alleles occupy a locus, an individual is said to be *heterozygous* at that locus. Consequently, *Aa* and *aA* are heterozygous genotypes whereas *aa* and *AA* are homozygous genotypes.

Today, the genetic material, is only partly understood. It is true that we know how amino acids are encoded but we find it hard to distinguish "junk DNA" from coding DNA. In a way, it is like reading a book were the spacings between words are filled and extended with arbitrary letters. If the nonsense words are recognized and removed, the book becomes perfectly readable. Unfortunately, we are not quite there yet. Even though nonsense words have been removed from some pages and put into clear sentences, this is a tedious task and much work remains.

Metaphorically, the human genome project has provided us with a book containing 23 chapters, one for each chromosome, revealing the DNA consensus sequence of the human genome. The finished sequence covers about 99 percent of the gene-containing regions, sequenced to an accuracy of 99.99 percent. Still, only about two percent of the genome is known to make up protein-coding sequences [18].

2.2 Human Genetic Variation

The human genome is estimated to contain 30,000 genes [20]. They vary widely in length, often extending over thousands of bases.

The difference between any two human genomes has been estimated to be less than 0.1 percent overall. Still, this means that there are at least several million nucleotide differences per individual [25].

Today, the most promising way of capturing disease-related genetic variation, is to type the genome for *Single Nucleotide Polymorphisms* (SNPs). Although multi-allelic markers, such as a *microsatellites* are more likely to detect genetic

variation, SNPs are more frequent and mutationally more stable. In fact, most of the genetic variation in the human genome is due to SNPs and occurs, when a single nucleotide in the DNA sequence is altered by a single historical mutation event. Since it is extremely unlikely that two mutations occur twice at the same locus, SNPs mainly exist as diallelic variations. An example of a SNP is the alteration of the DNA segment CCAATGT to CTAATGT, where the second "C" in the first sequence is replaced with a "T".

If two SNPs, caused by historical DNA alterations are located close to each other, they tend to be in *linkage disequilibrium* (LD) - that is, the two loci are inherited together and are strongly correlated. LD is a statistical concept and is explained in detail in section 3.5.

Despite that LD patterns are quite complex, SNPs tend to be structured into haplotype blocks, separated by recombination 'hotspots'. Only a few haplotypes are required to account for the genetic variation within a haplotype block. This has motivated the ongoing *International HapMap Project* in an attempt to develop a map of *haplotype tagging SNPs* (htSNPs), capable of identifying common haplotypes throughout the human genome [12].

2.3 Etiology of Complex Human Disease

Today, it is widely accepted that most common human diseases are caused by a mosaic of genetic and environmental factors. Since the cause of disease is not purely genetic, human geneticists refer to an individual's genetic predisposition or liability to develop disease.

Some liability factors may promote disease, whereas others suppress disease. Each effect is likely to be weak since more than one susceptibility factor is needed to develop disease. Different combinations of alleles and loci may result in a similar or identical disease phenotype. Even though a genetic signal in a multifactorial disease is not more complex than in a single locus disease, it is considerably weaker, making it harder to detect. The situation is likely to be further complicated by complex multi-way interactions among some or all of the contributing loci, loosely defined as epistatis.

The term 'epistatic' was first used in 1909 by Bateson to describe a masking effect whereby an allele at one locus prevents an allele at another locus from manifesting its effect [4]. In a way, this is an extension of the concept of allele dominance within a single locus, where one allele interferes with the effect of another allele, described in section 2.1. For instance, consider two bi-allelic loci, A and B, partly responsible for human eye color[1]. An allele at the A locus is either dominant, A, coding for green eyes; or recessive, a, coding for blue eyes. Similarly, an allele at the B locus is either dominant or recessive, such that, the dominant B allele codes for green eyes, whereas the recessive b allele codes for blue eyes. The possible phenotypes from all possible genotypes are shown in Table 2.1. We see that regardless of genotype at locus A, individuals with one or more copies of the B allele receives brown colored eyes. If no copies of the B allele is present, the genotype of the A locus determines if the eyes receives a blue or green color. Allele B is masking the effect of allele A, i.e.

[1] Human eye color can not be explained by two genes alone. Therefore, the epistatic interaction effect examplified in Table 2.1 only explains the phenotypic interaction effect of a two loci genotype.

allele B is dominant over allele A. Consequently, the effect of allele B at locus B is epistatic to allele A at locus A. This definition of epistasis is similar to how

	Genotype at locus B		
Genotype at locus A	b/b	b/B	B/B
a/a	Blue	Brown	Brown
a/A	Green	Brown	Brown
A/A	Green	Brown	Brown

Tab. 2.1: **Phenotypes obtained from different genotypes at two loci interacting epistatically, under Bateson's (1909) definition of epistasis.** The dominant variant of one locus (B) prevents, 'masks', the dominant variant at another locus (A) from manifesting its effect.

a molecular biologist or biochemist investigate interaction effects in signaling pathways. However, there are some problems with this definition.

In human genetics, the disease phenotype is often *quantitative* and *dichotomous*, indicating presence or absence of disease [4]. Suppose that the two loci from the previous example influence a binary disease trait instead of eye color. If a predisposing allele is required at both loci in order to develop the disease, one or more copies of both allele A and allele B is needed. Then, when the effects of both loci are considered, the penetrance table in Table 2.2 is obtained. In this table, the effect of the two dominant alleles A and B can only be observed jointly. Locus A is equally masked by B, as B is masked by A. Consequently, both locus A and locus B provoke epistatic effects on each other. This corresponds to a more general form of epistasis which implies that both loci have mutual epistatic effects.

	Genotype at locus B		
Genotype at locus A	b/b	b/B	B/B
a/a	0	0	0
a/A	0	1	1
A/A	0	1	1

Tab. 2.2: **Penetrance table for two loci interacting in a general sense.** The dominant variant of one locus (A) prevents, 'masks', the dominant variant at another locus (B) from manifesting its effect. [4]

While there is no unified definition of epistasis, we may broadly define it as *an interaction between genes, where one gene interferes with the effect of another gene.*

2.4 Analysis of Genetic Variation in Complex Human Disease

There is an ongoing debate if genotypes or haplotypes should be used in dissecting the genetic variation predisposing humans to disease. Both approaches have their advantages and disadvantages and their use is largely dependent on the purpose with the genetic study. The two following subsections accentuates their different use.

2.4.1 Genotype Analysis

One reason for genotyping SNPs is that there is a general belief, that SNPs can be used as markers in association studies. The accurate identification and dense distribution of SNPs makes them appropriate for identifying genes that predispose individuals to common, multifactorial diseases.

Moreover, genotypes are believed to be informative for distinguishing disease associated loci that have a direct causal role in the disorder from those only showing association because they are in LD with the primary disease-related polymorphism [5].

Genotype analysis does not infer the phase of the alleles, but this does not necessarily mean that haplotypes are preferable. Theoretically, it is possible to cover the genetic variation in a set of markers using either method. For example, if all main and interaction affects among genotypes are modelled in a logistic regression model, the response is equivalent to performing haplotype analysis on the same marker set. However, such genotype analysis quickly becomes intractable, since the number of interaction components escalates, forcing the choice of a *nested model*. As a result, the information describing the genetic variability is partly lost. Haplotype analysis sustains this variability, but not unless the haplotypes have been correctly inferred. Moreover, if the disease model is simple, the statistical power is likely to be better performing genotype analysis.

2.4.2 Haplotype Analysis

It is now widely accepted that haplotype analysis can be of interest when investigating the role of susceptibility genes in the etiology of complex diseases. Whereas, genotypes are useful for distinguishing a susceptibility locus from neighboring non-susceptibility loci, haplotypes are likely to be more informative when none of the markers have a direct causal role in the disorder, since haplotypes can be "tagged" into parsimonious LD blocks, accounting for the allelic variation of a region, presumably involved in disease (HapMap project).

Eppstein and Satten believe that haplotype-based association methods is inherently more powerful for gene mapping than methods based on single SNPs [8]. The reason for this is that linkage disequilibrium exists, over short genetic distances so traditional association tests have limited power to identify disease-predisposing variants in weak LD.

Another reason for studying haplotypes is that the function of a gene may very well depend on an allele constituting several sites within a gene. Haplotypes provide a greater opportunity to detect such an unknown gene variant than do individual polymorphisms. Morris and Kaplan have published result indicating that the general loss of power observed in association analysis when multiple disease susceptibility loci are present within a gene, is less prominent for multi-allelic markers than for diallelic markers. Suggesting that a haplotype analysis can be advantageous over genotype analysis in the presence of multiple alleles at a disease locus, particularly when SNPs are in weak LD [15].

Unfortunately, establishing haplotypes by genetic assays is extremely expensive. Neither is it possible to deduced the haplotypes from unphased genotypes, unless family data is at hand or if the individuals are heterozygous at no more than one loci. Therefore, haplotypes need to be inferred using statistical meth-

ods. Maximum likelihood algorithms, such as the Newton-Raphson (NR) algorithm and the Expectation-Maximization (EM) algorithm, are often employed in order to infer such haplotypes and their respective frequencies. Both these methods not only assume independent allele frequencies, but also invoke uncertainty in the estimated haplotypes, arguably loosing what one might have gained from studying haplotypes in the first place.

3. STATISTICAL GENETICS

The past decade or so have witnessed an explosion in molecular biotechnology and computer science. With the entire human genome at hand and a continuous growth of marker databases, new opportunities are presented for unravelling the often complex genetic basis of human disease.

Statistical geneticists are currently busy identifying genes influencing multifactorial diseases. So far, the success have been largely restricted to diseases with simple Mendelian inheritance patterns. The main reason for this is likely to be the genetic heterogeneity often observed in complex disease in addition to inappropriately designed population studies.

3.1 Basic Concepts in Statistical Genetics

This section cover some fundamental concepts of probability and inference theory in statistical genetics, and provides a necessary understanding of the methods applied in this thesis.

3.1.1 Probability Theory

A *probability*, is a numeric value that model the likelihood that a specific event occurs and is expressed as the ratio of the number of actual occurrences to the number of possible occurrences. Consider a study population containing 100 individuals, of which 3 are affected by a disease and the remaining 97 are unaffected. The probability of being affected by such a dichotomous trait, often referred as the disease *prevalence*, is

$$P(\omega = \text{Affected}) = \frac{3}{100}, \tag{3.1}$$

where P(ω) is a *probability function* modelling a random event. A probability function must satisfy three intuitively plausible rules, given by Kolmogorov's axioms:

$$\begin{array}{rl} (i) & P(\Omega) = 1 \\ (ii) & \text{If A and B are disjoint, then } P(A \cup B) = P(A) + P(B) \\ (iii) & \text{For any event B, } 0 \leq P(B) \leq 1 \end{array} \tag{3.2}$$

$A \cup B$ refers to the combined event of A and B occurring together. Ω symbolize the complete sample space - that is, the combined occurrence of all possible events, whereas an *event*, symbolizes a disjunct subset of Ω, that an outcome, ω can attain. In the above case, there is only two possible outcomes; affected or unaffected.

If we assume that individuals have the same probability, p to independently develop the disease, the outcome is bernoulli-distributed according to the probability function,

$$P_Y(y|p) = p^y(1-p)^{1-y} \quad , Y = 0, 1 \tag{3.3}$$

where Y is a *random variable* indicating disease status. The sum of k independent bernoulli-distributed variables are binomially distributed,

$$\sum_{i=1}^{k} Y_i = N \sim Bin(k, p) \tag{3.4}$$

and its probability function is given by

$$P_N(n|p,k) = \binom{k}{n} p^n(1-p)^{k-n} \quad , N = n_1, n_2, ..., n_k. \tag{3.5}$$

Note that the bernoulli distribution in equation 3.3 is a special case of the binomial distribution in equation 3.5. It is important to distinguish both these probability functions from their corresponding *likelihood function*,

$$L_Y(p|y) = \prod_{i=1}^{k} p^y(1-p)^{1-y} \tag{3.6}$$

and

$$L_N(p|n,k) = \prod_{i=1}^{k} \binom{k}{n} p^n(1-p)^{k-n}, \tag{3.7}$$

where N is a random variable indicating the number of affected individuals for k sampled individuals. While the probability function returns probabilities of the data, given the parameter p, the likelihood function gives the relative likelihoods for different values of p. Note that the right hand side of the probability functions and likelihood functions are the same when the likelihood function is conditional on an individual, i. The conceptual motivation of using the likelihood function is that the "most likely" parameter, p can be chosen given the data.

3.1.2 Inference Theory

Statistical inference theory uses probability models to describe observed variation in data. Assume we are interested in assessing the association between DNA variants and disease in a case-control study design. The simplest way of doing this is to compare differences in genotype frequencies between affected and unaffected individuals.

A *null hypothesis*, H_0 is formulated, stating that there is no difference in allele frequencies between cases and controls. If the genotype data show a *significant* deviation from the null hypothesis given a fixed threshold, t, the null hypothesis is rejected, and a predefined *alternative hypothesis*, H_a is accepted. In order to specify what a 'significant deviation' implies, it is necessary to define a test statistics[1],

$$\begin{aligned} &T \geq t \Rightarrow \text{reject} H_0 \\ &T < t \Rightarrow \text{do not reject} H_0, \end{aligned} \tag{3.8}$$

[1] One such test statistics is the χ^2-test which approximates a normal distribution for a sufficiently large number of observations.

where T represent the deviation of the genotype data from the null hypothesis. The probability of rejecting the null hypothesis even though it is true is referred to as the *significance level*, α of a the test,

$$\alpha = P(T \geq t|H_0), \tag{3.9}$$

equivalent to the probability of making a *Type I error*. In a similar manner, a *Type II error* is committed if the null hypothesis is incorrectly accepted. The probability of making a Type II error, β is defined as

$$\beta = P(T \leq t|H_a), \tag{3.10}$$

i.e. the probability of rejecting the null hypothesis given that the alternative hypothesis is true. If no Type I or Type II error is committed, the correct decision has been made. The *power* of a statistical test is defined as the complement of β,

$$Power = 1 - \beta = 1 - P(T \leq t|H_a) = P(T \geq t|H_a), \tag{3.11}$$

i.e. the probability of correctly rejecting the null hypothesis when it is truly false. For association studies, the power can be considered as the probability of correctly detecting a genuine association.

Obviously, we can control the significance level and power by our choice of threshold. A lower threshold, that is a larger significance level, generates greater power and consequently less Type II errors, whereas Type I errors increase. The opposite is true for a low significance level. It is important to understand that a statistically significant result does not imply that chance cannot have accounted for the result, only that this is unlikely. Similarly, a non-significant result does not imply that the null hypothesis is true! The study may simply lack power.

If an allele appears to have an effect, it is very important to be able to state with confidence that the effect is due to the allelic variant and not just due to chance. However, as previously discussed, using a larger significance level result in more false positives. So in order to perform an accurate test the significance level needs to remain low (α=0.05).

An alternative to specifying a significance level in advance is to compute a *p-value*. This can be thought of as the significance level achieved by data and is defined as

$$p = P(T(X) \geq T(x)|H_0), \tag{3.12}$$

i.e. the probability, under H_0, of observing a test statistic at least as large as the one we actually observed.

3.2 Statistical Inference in Case-Control Studies

Population based case-control studies are widely used in epidemiological research to identify and characterize genes involved in human disease. The case-control design involves collecting a large number of individuals affected by the disease, the cases; and a large number of individuals not affected by the disease, the controls. If the cases are found to be more frequently exposed to a susceptibility factor than the controls, one can infer that the susceptibility factor is involved in disease pathogenesis. Although extremely simple, care needs to be taken when designing a case-control study. If not, the design might lead to

dubious associations. Such unwanted effects, can often be avoided if a number of design and modelling issues are considered.

In the following sections, unwanted effects, such as *chance*, *confounding* and *heterogeneity* are discussed. Only after these issues have been investigated and approved can an association between a marker and disease locus be considered valid.

3.2.1 Significance, Power and Chance

Recall from section 3.1.2, that the significance level of a statistical test is the probability of falsely rejecting the null hypothesis, whereas the power of a statistical test is the probability of correctly rejecting the null hypothesis when it is truly false.

Significance and power is a tradeoff in any statistical analysis but if a statistical test should be worth wile, the significance level needs to remain low, or else an association is likely to appear by chance. This becomes evident when multiple genes are tested. To see this, consider a box with 20 marbles, 19 black ones and one white. The odds of randomly sampling the white marble by chance is 1 out of 20. Now, assume that you get to sample a single marble 20 times, each time returning the sampled marble to the box. You now have a higher chance to sample the white marble. This is exactly what happens when testing several thousand genes at the same time. The white marble corresponds to a false positive and testing for association multiple times corresponds to repeatedly sampling a marble from the box. If multiple testing is left uncorrected, the significance level for the joint test may be unacceptably high. Moreover, if the marker polymorphism is rare or has only a moderate effect on the disease, very large sample sizes are required to achieve reasonable power in the study.

3.2.2 Confounding

Confounding factors are those that are associated with both the disease and the factor under study [2], illustrated in figure 3.1. In genetic studies one such confounding factor occurs if a locus, believed to be associated with a disease is not directly linked to the susceptibility locus. For example, loci located on different chromosomes may occasionally be in linkage disequilibrium with each other. Therefore, one might incorrectly conclude that a significantly associated marker is located close to the susceptibility locus, when it in reality is located on a completely different chromosome.

A challenge when performing case-control studies is choosing cases and controls from the same study base. If not confounding factors correlated with both the disease and the marker under study, may be present. As an obvious example, consider cases sampled from Sweden and controls from Fiji. Clearly, the cases and the controls are highly likely to have genetic differences at several loci throughout the genome purely because of an inherent genetic distance between populations. Thus, it might prove difficult to know whether any observed difference in allele frequencies between cases and controls reflects the causal impact on the disease or a difference in genetic background.

This is known as *population stratification*. In association studies however, population stratification within cases and controls, is often of larger concern.

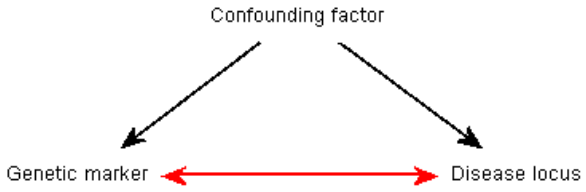

Fig. 3.1: **Spurious association due to confounding.** The confounding factor is associated with both the disease and the marker, denoted by black arrows. This may cause one to conclude that an association between the marker and disease is present when it is not, denoted by the red double headed arrow.

Such stratification can be caused by allele heterogeneity or locus heterogeneity, discussed further in the next section.

3.2.3 Heterogeneity

As previously mentioned, population stratification within cases and controls, might manifest itself as locus or allele heterogeneity.

Locus heterogeneity, implies that different loci combinations affect the trait or disease similarly. This is believed to be the case in several complex diseases and since neither locus is sufficient nor necessary, statistical power is lost in an association test.

Allelic heterogeneity, can create situations where multiple alleles of a gene are associated with the disease, rather than a single specific allele.

Both situations are known to exist in several complex human diseases, so even if cases and controls are sampled wisely, heterogeneity might be present within the two groups. In such cases, not only are large sample sizes needed to find the presumably weak disease association of markers, but it might also be necessary to stratify the cases into subgroups.

The disease status of Multiple Sclerosis is often classified into subtypes and it is therefore possible that these subtypes are caused by heterogeneity in loci or alleles.

3.3 Mendelian Inheritance

In 1865, long before the DNA structure was revealed, Gregor Mendel performed breeding experiments with pea plants enabling him to establish two genetic laws. The first law, the *law of segregation*, states that allelic variants of a trait segregate independently. That is, when a parent passes one of its two copies of a gene to its offspring, either copy is equally likely transmitted. The second law, the *law of independent assortment*, states that alleles at different loci, whether on the same chromosome or not, are distributed independently of one another.

Before Mendelian inheritance was accepted, some scientists believed traits were passed on by *independent blending*. Much like mixing the colors black and white result in grey color, one believed parents passed on an intermediate of their corresponding traits to their offspring. The theory attempted to describe the natural variation in quantitative traits like skin color and height, failing to

realize that the continuity of such traits are not caused by a single gene but by the action of multiple genes.

Today, the general understanding of macro molecular processes within a cell, support what Mendel partly concluded over a century ago. The two laws and the independent transmission of discrete units from parent to offspring, suggested by Mendel, corresponds with the events that occur when haploid gametes are produced during meiosis[2].

However, the concept of linkage disequilibrium made it apparent that exceptions to Mendel's two laws of inheritance exist. Still, few people argue that Mendel's mathematical contribution in segregation analysis, and his conclusion that genetic factors are inherited in discrete units, possibly exhibiting dominance effects, initiated modern science of statistical genetics.

Under the assumption of random mating and diallelic loci, independent blending would cut the population variation of each successive generation in half and eventually, everyone would look the same. Mendelian inheritance however, sustains its allele frequencies in successive generations and the population variation remains unchanged. This is referred to as the *Hardy-Weinberg equilibrium* (HWE).

3.4 Hardy-Weinberg Equilibrium

In its simplest case, a single locus with two alleles, A and a, with allele frequencies, p and q respectively, the principle of Hardy-Weinberg predicts the genotype frequencies for the AA homozygote to be p^2, the Aa heterozygote to be $2pq$ and the other aa homozygote to be q^2, such that

$$p^2 + 2pq + q^2 = 1. \tag{3.13}$$

Provided that no factors of evolution take place, and an infinitely sized, randomly mating population is assumed, one can easily show that genotype frequencies stabilize in the relation given in equation 3.13, after a single round of random mating [23]. It is in this sense that the population frequencies represents an equilibrium. The frequencies of genotype AA, Aa and aa are therefore maintained in a randomly mating population.

Even though deviation from HWE, named *Hardy-Weinberg disequilibrium* (HWD), can potentially be explained by misspecified genotypes, present day techniques for genotyping markers are highly accurate [24]. Therefore, HWD in genotyped case-control studies is more likely to be caused by non-random mating. One form of non-random mating is *population stratification*, where mating of individuals from the same strata is more likely to occur than mating of individuals from different strata. For example, within Sweden, mating tend to be stratified into racial descent, and in the United States even more so. In this situation, individuals within a strata may be in HWD, whereas the whole population might be in HWE. Similar stratification may appear within strata, due to mating between related or similar individuals, e.g. if marriage between cousins is encouraged, or if long women deliberately avoid marrying short men.

[2] Mendelian inheritance implies complete allelic independency, that is, no linkage or genetic imprinting takes place

Testing deviation from HWE is generally performed using Pearson's chi-square test,

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{3.14}$$

i.e. given the genotype i, the observed genotype counts, O_i present in the data are compared to the expected genotype count, E_i under the assumption of HWE. For a diallelic locus the expected genotype frequencies are obtained from

$$\begin{aligned} P(AA) &= p^2 \\ P(Aa) &= 2pq, \\ P(aa) &= q^2 \end{aligned} \tag{3.15}$$

under the assumption of independent allele frequencies, p and q.

When performing association studies, independence between alleles is often assumed. If the population sample is not in HWE, the assumption of allelic independence is violated and the statistical test on association may be invalidated. Testing for Hardy-Weinberg equilibrium is therefore a standard procedure to rule out population stratification.

3.5 Linkage Equilibrium

Whereas Hardy-Weinberg equilibrium refers to independence among alleles on a genotype, linkage equilibrium (LE) refers to independence among alleles on a haplotype. Like the Hardy-Weinberg ratio, LE is an equilibrium in the sense that the allele frequencies stabilize over time and is kept there.

Consider a mutated ancestral haplotype, passed on to successive generations by randomly mating couples. If a neighboring allele segregate independently of the ancestral mutation, the alleles stabilize in the same relation as the Hardy-Weinberg equation, 3.13. Such a scenario is equivalent to testing allele independence between chromosomes. However, because it is highly unlikely that recombination take place between closely spaced loci, neighboring loci tend to co-segregate in haplotype blocks. Such loci are typically in *linkage disequilibrium*, LD with each other (Figure 3.2).

In association studies a marker in LD is assumed to be located close to the predisposing disease locus. However, patterns of LD are generally noisy and unpredictable. For example, pairs of sites that are tens of kilobases apart might be in complete LD, whereas neighboring sites from the same region might be in weak LD [26]. If a mutation influencing disease occurred recently in the population, the mutation will not be present in all affected individuals. Such a locus will only be in weak LD because of the low representation of the allelic variant among the sampled cases. This type of stratification is referred to as allelic heterogeneity and will be discussed further in section 3.2.3.

Several authors have reviewed different types of LD measures and their use in the search for complex disease genes [6, 11, 13, 26]. $|D'|$ and R^2 are two pairwise LD measures often used. Both measures range from 0 (no LD) to 1 (complete LD), but are interpreted slightly different. Whereas, $|D'|$ resembles the recombination rate and is insensitive to allele frequencies, R^2 represents the statistical correlation between two sites and is dependent on allele frequencies.

$|D'|$ may be more appropriate for fine-mapping of disease genes in case-control studies since it is invariant when the disease haplotypes are sampled

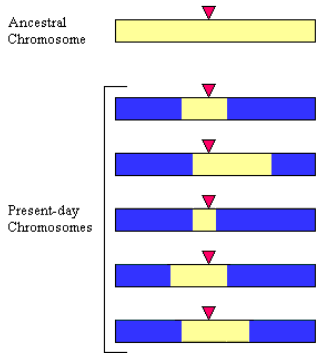

Fig. 3.2: **Linkage Disequilibrium around an ancestral mutation.**The mutation is indicated by the red triangle. The common ancestor (yellow), passes the mutation on to successive generations. Occasionally, homologous recombination causes new haplotypic variants to form, as non-related DNA segments (blue) are exchanged into the chromosome due to homologous recombination. Closely linked loci tend to remain associated (LD) with the mutation in present-day chromosomes.

at a higher rate than their population frequencies [6]. On the other hand, one might argue that R^2 is the most relevant measure since there is a simple inverse relationship between R^2 and the sample size required to detect an association between a susceptibility locus and SNPs [26]. Thus, R^2 may suggest an appropriate sample size in order to achieve reasonable power in an association study.

For a pair of diallelic loci, with allele frequencies p_1, q_1 and p_2, q_2, respectively,

$$D'_{12} = D_{12}/D_{max} \tag{3.16}$$

$$R^2 = \triangle^2 = D^2/(p_1 p_2 q_1 q_2), \tag{3.17}$$

where

$$D_{12} = p_{12} - p_1 p_2,$$

$$D_{max} = min(p_1 q_2, p_2 q_1).$$

The $|D'|$ tends to be upwardly biased for small sample sizes, and intermediate $|D'|$ are generally hard to interpret [26].

4. MODELLING GENE-GENE INTERACTION IN CASE-CONTROL STUDIES

As discussed in section 2.3, virtually all common human diseases are influenced by complex interactions among multiple genes and environmental factors. Such diseases requires sophisticated statistical designs, often consisting of (1) an *additive model*, characterized by non-interacting components, or (2) a *multiplicative model*, characterized by interacting components.

In 1965, Falconer proposed that multifactorial diseases could be mathematically modelled in a threshold liability model [9]. The model assumes that an individual develops disease once the additive effect of genetic and environmental liability factors exceeds a certain threshold value. The liability of developing

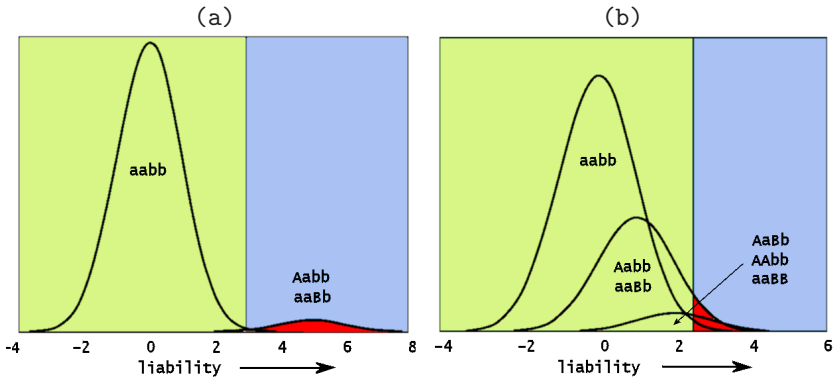

Fig. 4.1: **Mendelian and non-Mendelian two-locus genetic model.** Total liability is defined as the sum of genetic and non-genetic liabilities, and disease occurs when an individual's total liability exceeds a defined threshold. (a) illustrates two dominant Mendelian alleles, A and B. The normal homozygote, $aabb$ has a very low disease risk $\sim 0.13\%$, the heterozygotes, $Aabb$ and $aaBb$ have a very high disease risk $\sim 98\%$, whereas the homozygote, $AABB$ has close to a 100% disease risk (not illustrated). (b) illustrates non-Mendelian alleles. $aabb$ has a very low disease risk, $AaBb$ and $AaBb \sim 0.62\%$ have low disease risk, $AaBb$, $AAbb$ and $aaBB$ have moderate disease risk $\sim 6.7\%$, $AABb$ and $AaBB$ have a fairly high disease risk $\sim 31\%$ (not illustrated), whereas the double homozygote, $AABB$ is has high disease risk $\sim 69\%$ (not illustrated). Although the two loci are additive on the liability scale, the disease risks are non-additive and show both dominance and epistasis effects.

The image has been slightly modified on permission by publicisist and author, [22]. Source: http://www.nature.com.

a disease is assumed to be normally distributed given a certain genotype, an idea partly concluded by Fisher in 1918 [10]. Figure 4.1 illustrates the liability model for a Mendelian and a non-Mendelian two-locus model [22]. A two-locus Mendelian trait is solely dependent on allelic variants at two loci, whereas a two-locus non-Mendelian trait is heterogeneous, meaning that different pairwise combinations of genetic and environmental factors may predispose the disease.

Even though a liability model is intuitively appealing, it does not consider interaction effects. In this chapter, we explain how binary logistic regression can be used to model such interaction effects in human disease. When studying gene-gene interaction it is often necessary to limit the number of interaction variables to a reasonable number.

In an attempt to find a parsimonious subset of disease related markers, two methods are proposed, one based on haplotypes and one based on genotypes. Both these can be extended to model interaction effects.

4.1 Binary Logistic Regression

Logistic regression is a statistical technique useful for binary response problems. As in any other regression modelling the key quantity in logistic regression is the conditional mean - that is, the value of the response variable given the values of the predictor variables.

Suppose we wish to model the presence or absence of disease, denoted

$$Y = \begin{cases} 1, & \text{presence of disease} \\ 0, & \text{absence of disease} \end{cases}, \tag{4.1}$$

as a function of two genetic components, x_1 and x_2. Intuitively, we are interested in modelling the conditional probability of being sick, given the two genetic components, x_1 and x_2. If we apply an ordinary linear regression model,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{4.2}$$

the conditional mean of being affected by the disease, given the predictor variables is,

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{x}) = E(Y|X) = \boldsymbol{\beta}\mathbf{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{4.3}$$

where $\boldsymbol{\beta}$=$(\beta_0, \beta_1, \beta_2)$ and $\mathbf{x} = (1, x_1, x_2)^T$.

However, applying a simple linear regression model to the probability of a binary response variable is inappropriate, since this would imply that $\pi(\mathbf{x})$ take on values outside the probability range, $0 \leq \pi(\mathbf{x}) \leq 1$.

What we would like, is to apply a function, that is zero for very small x values and one for very large x values. As illustrated in Figure 4.2, one such function is the logistic, $a(b) = (1 + e^b)^{-1}$, implying a nonlinear transformation,

$$p(x) = \frac{1}{1 + e^{-\pi(x)}} = \frac{e^{\pi(x)}}{1 + e^{\pi(x)}} \tag{4.4}$$

of equation 4.3. By applying a logistic function, all values are transformed to meet the probability interval constraint. However, the simple linear regression curve is lost, making the parameters hard to interpret. Taking the ratio of

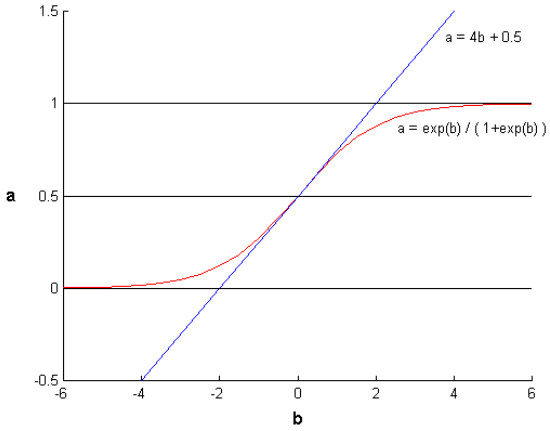

Fig. 4.2: **Linear and logistic function.** The linear function is displayed in blue and the logistic function in red. The 'S-shaped' form of the logistic function (4.4), often called the sigmoid function, maps the interval (-∞,∞) onto (0,1). Note that the logistic function is approximately linear for small $|b|$.

equation 4.4 and its inverse, in this case equivalent to the odds of p and 1-p, followed by yet another transformation, the natural logarithm, the problem is nicely addresses,

$$\ln(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \pi(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{4.5}$$

Consequently, the parameters in the logistic regression model can be interpreted as log odds ratios.

The logistic regression model in (4.5) can easily be extended to handle an arbitrary number predictor variables. For a collection of k predictor variables, denoted by the vector $\mathbf{x} = (1, x_1, x_2, ..., x_k)^T$, the multiple logistic regression model is given by

$$logit(p) = \boldsymbol{\beta}\mathbf{x} = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \tag{4.6}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$.

4.1.1 Parameter Estimation

In logistic regression, the parameters of the model are estimated using the maximum-likelihood method. This optimization method estimates the coefficients of the regression variables as the values that are most likely given the observed state. To be specific, the maximum likelihood estimation (MLE) seeks to maximize the likelihood function, L for the binomial distribution,

$$max_\beta L = \prod_{i=1}^{k} \binom{N}{n} p(\mathbf{x})^n (1-p(\mathbf{x}))^{N-n} \tag{4.7}$$

assuming individuals are independent.

MLE is performed using an iterative algorithm which starts by arbitrarily assigning values to the parameter estimates. Thereafter it determines the direction and size change in the logit coefficients which will increase the L. Based on the estimates, the residuals are tested and a re-estimate is made with an improved function. The process is repeated until the L does not change - that is, the maximum[1] is reached.

4.1.2 Significance Tests

After estimating the coefficients, it is standard practice to assess the significance of the variables in the model. This usually involves comparing two models; one nested within the other. Consider, the nested model

$$logit(p) = \boldsymbol{\beta}\mathbf{x} = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1}, \tag{4.8}$$

containing one variable less than the model depicted in equation 4.6. In order to assess the significance of the variable k, the following hypothesis,

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

is tested in a likelihood ratio test,

$$D = -2 \ln \frac{L_{\text{full model}}}{L_{\text{nested model}}}, \tag{4.9}$$

and corresponds to the deviance in fit of the two models. Under the null hypothesis that β_k is equal to zero,

$$D \sim \chi^2(df_{\text{nested model}} - df_{\text{full model}}), \tag{4.10}$$

i.e. the difference in the number of degrees of freedom for the two models determines the degrees of freedom of the χ^2-test. Since, the nested model is restricted on one variable less than the full model the χ^2-test of the above example has one degree of freedom.

4.2 Coding Genetic Effects

A SNP has three unphased genotypes, two homozygous ones and one heterozygous one, commonly denoted 11, 12 and 22. In order to use these categories as predictors in a logistic regression model, they need to be coded into numerical values. Inevitably, the allele effect will depend on how we decide to code the alleles in reference to one another. For example, an additive allele effect may be coded

$$\{11, 12, 22\} = \{0, 1, 2\} \tag{4.11}$$

whereas a dominant allele may be coded

$$\{11, 12, 22\} = \{0, 1, 1\} \tag{4.12}$$

[1] Logistic regression is a regular model with a convex likelihood surface and a single global maximum.

In the additive coding scheme (4.11), the effect of having two copies of allele 2 is twice that of having a single copy. In the dominant coding scheme (4.12), there is no difference in effect between having two copies of allele 2, from having a single copy, whereas no copy at all is effectless.

Some useful coding schemes for evaluating the relative effects of polymorphisms have been suggested by Cordell and Clayton [5]. If we have reason to believe that the effect of having two copies of a susceptibility allele is twice the effect of having just one, an additive coding scheme,

$$a \sim (11, 12, 22) = (-1, 0, 1), \tag{4.13}$$

models a single locus variation as

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 a. \tag{4.14}$$

The saturated effect of the SNP genotype is modelled by specifying an additional predictor variable,

$$d \sim (11, 12, 22) = (-0.5, 0.5, -0.5) \tag{4.15}$$

resulting in

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 a + \beta_2 d \tag{4.16}$$

The choice of coding for d makes the bases a and d orthogonal, which is an attractive statistical stability property. Note that the dominance effect captures deviations from additivity. The additive model converges to $\beta_2 = 0$ and the dominance model to $\beta_1 = \beta_2$. Interpreted on the logit scale, β_1 represents the additive effect of allele 2, whereas β_2 represents the dominance effect of allele 2 over allele 1. How much the allele effect deviates from additivity is directly dependent on the size of the parameter β_2. A large β_2 value signifies a large deviation from additivity and models a correspondingly large dominance effect; a small β_2 value signifies only a vague dominance effect; whereas a zeroed parameter value models a pure additive effect.

It is important to realize that the dominance component in model 4.16 does not model a dominance effect on its own. Only in addition to the additive component can we talk about a dominance effect[2]. If one insists on interpreting the dominance component alone, the coding scheme resembles a heterozygous advantage/disadvantage effect.

4.3 Stepwise Logistic Regression

When one wishes to reduce the number of predictors in a logistic regression model without markedly reducing the ability to explain variation in the response, two stepwise selection procedures, or a combination of the two, are widely used; forward selection and backward elimination.

Starting from the null model, forward selection iteratively adds and tests components one at a time for an increase in model fit. The algorithm stops

[2] It is possible to model a pure dominant effect using a different coding scheme, such as the one presented in (4.12)

when the full model is reached. A good model will have a large $\Delta\chi$-value and small p-value.

Conversely, the backward elimination procedure starts with the full model, and removes every marker-effect one at a time and the best fitting model is kept. Every successive session reduces the number of parameters by one and not until the null model is encountered does it stop. Be aware of that every nested model is tested for its deviation from the full model, opposed to the procedure in forward selection where a model under test is compared to the previously selected model. A high p-value tells us that the model does not deviate much from the full model.

Stepwise selection is intuitively appealing since it allows one to build models in a sequential manner instead of testing every possible model in a exhaustive and consequently computer intensive way. Unfortunately, it comes with a price. Simple stepwise selection algorithms are irreversible in their selection of variables. Some combinations are never tested and this may result in a suboptimal final set of predictors.

4.4 Parsimonious Selection using Logistic Regression

Once numerous markers within and outside a localized gene have been genotyped, logistic regression can be applied in an attempt to find a parsimonious subset of disease-associated markers.

As discussed along with the concept of linkage disequilibrium in chapter 3, a disease associated polymorphism identified by linkage or association methods, does not necessarily have a causal role in disease etiology. Neighboring sites to a disease susceptibility locus may be associated to disease merely because of LD. In a regression setting, LD corresponds to collinearity between predictor variables making it difficult to distinguish which of two predictors contribute a stronger effect.

Cordell and Clayton suggest that an unconditional logistic regression model based on genotyped cases and controls may be useful for distinguishing between predisposing etiological variants and alleles at neighboring sites, all in LD [5].

Instead of analyzing genotypes, a haplotype based regression model has recently been proposed by Sham et al. [23]. This approach may prove more useful when no disease-related variants are among the genotyped loci.

Both methods can be used to narrow down the number of SNPs within a gene, prior to studying interaction effects. In the following two subsections, the use of the two methods for selecting a parsimonious marker subset are described, and the discussion on genotype and haplotype analysis briefly discussed in section 2.4 is revisited.

4.4.1 Genotype Analysis

Recall that the difference in deviation between the fit of two nested regression models follows a χ^2-distribution. In section 4.3, stepwise regression algorithms applied this idea in order to compare the effect of adding or eliminating variables from the model.

Suppose we have n_1 cases and n_2 controls, each genotyped at G polymorphic loci within a gene. All genotypes are coded into an additive and a dominance

component - that is, the saturated model,

$$logit(p) = \beta_0 + \beta_1 a_1 + \beta_2 d_1 + \beta_3 a_2 + \beta_4 d_2 + ... + \beta_{2G} d_G \tag{4.17}$$

Conditional on the i:th individual, the likelihood of the observed data is

$$L_i = \left(\frac{exp(\pi_i)}{1 + exp(\pi_i)}\right)^{y_i} \left(\frac{1}{1 + exp(\pi_i)}\right)^{1-y_i}, \tag{4.18}$$

where $\pi_i = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$ models the genotype of an individual and y_i indicates wether the individual is a case or a control (equation 4.1). The complete likelihood function for the logistic regression model is

$$L = \prod_{i=1}^{k} L_i. \tag{4.19}$$

To find a parsimonious set of marker loci that are most predictive and therefore most likely to be closely associated with the disease, any of the two stepwise or exhaustive regression models previously described can be used. However, if we expect the gene to possess a large number of functional sites, each coded into an additive and a dominance effect, a backward logistic regression model is more appropriate [5].

4.4.2 Haplotype Analysis

Logistic regression can be used to study 'haplotype effects'. A simple extension of equation 4.6 is to treat each multilocus haplotype in the same way as an allele. Instead of coding genetic effects, each variable represents the number of copies (0, 1, or 2) of a specific haplotype inferred from the multilocus genotype in an individual. A problem with this approach is that many multilocus genotypes do not uniquely specify the underlying constituent haplotypes, making it impossible to define the predictor variables unambiguously as described above [23]. Sham et al. suggest a two-staged approach circumventing this problem[3].

Stage 1. Estimate haplotype frequencies using the genotyped data and use these to obtain the prior probabilities of all possible haplotype combinations. Then use Bayes theorem to obtain the posterior probabilities of the possible haplotype combinations in each individual given their multilocus genotype.

Stage 2. At this stage the coding of the variables needs to be established. Sham et al. use an additive coding scheme. The posterior probabilities calculated in stage 1 is used as weights in a finite mixture regression model. If the posterior probability of the h^{th} possible haplotype is denoted p_h, the likelihood function of an individual observation can be written as a weighted sum

$$L_i = \sum_h p_h L'_h, \tag{4.20}$$

where the summation is over all possible haplotype combinations in the individual and

$$L'_h = \left(\frac{exp(\mu_h)}{1 + exp(\mu_h)}\right)^{y_i} \left(\frac{1}{1 + exp(\mu_h)}\right)^{1-y_i} \tag{4.21}$$

[3] The descriptions of the two stages are partly quoted from Sham et al. [23]

is the likelihood function of the individual conditional on the h^{th} possible haplotype, denoted as $\mathbf{x}_h$ and modelled $\mu_h = \alpha + \boldsymbol{\beta}^T \mathbf{x}_h$. The complete likelihood function for the logistic regression model is given by equation 4.19.

The finite mixture regression model in stage two implies that the likelihood function of the observed parameters in the logit model, is a mixture of likelihood functions each conditional on a specific haplotypic variant. As in genotype analysis, a serie of nested models can be tested for the significance of dropping one or more loci from a set of SNPs by the reduction in minus twice the log-likelihood - that is, the reduction in deviance between two models.

Sham et al. believes that the above suggested haplotype approach is beneficial over the genotype approach suggested by Cordell and Clayton [23]. A problem with the haplotype method is that the haplotypes and its corresponding frequencies need to be estimated, which is likely to introduce uncertainty.

4.5 Modelling Gene-Gene Interaction using Logistic Regression

Assume we are interested in the interaction effect between the two loci, each coded into an additive and a dominance component. Mathematically, the genetic concept of epistasis of the two loci can be linearly model,

$$logit(p) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 a_1 + \beta_2 d_1 + \beta_3 a_2 + \beta_4 d_2$$

$$+\beta_5 a_1 a_2 + \beta_6 a_1 d_2 + \beta_7 d_1 a_2 + \beta_8 d_1 d_2 \tag{4.22}$$

The β_0 coefficient corresponds to the mean effect, whereas β_1, β_2, β_3 and β_4 represent additive or dominance effects of their corresponding variables. The remaining coefficients correspond to the epistatic interaction effects. Lack of epistasis in this model implies that the interaction coefficient, are all zero, resulting in the core model,

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 a_1 + \beta_2 d_1 + \beta_3 a_2 + \beta_4 d_2 \tag{4.23}$$

If we fit model 4.22 and 4.23, and compare their deviances, a χ^2-test with four degrees-of-freedom is obtained. Similarly, individual interaction terms are tested on one df, by removal one of the interaction terms from the full model 4.22.

5. DATA

Prior to this project, a two-staged association analysis was performed on MS cases and controls. A significant difference in allele frequencies between cases and controls were observed in three of the genes, the same genes subject to the interaction study performed in this degree project.

In this chapter, the etiology of Multiple Sclerosis is covered, followed by information on data history. The last section explains the data preprocessing step initializing the work of this degree project.

5.1 Multiple Sclerosis

Multiple Sclerosis (MS) is an autoimmune inflammatory disease of the Central Nervous System which degrades the insulating myelin layer covering nerve cells. Even though the myelin layer is partly rebuilt, hard scar tissue is formed, incidently blocking signal transmission.

Population studies suggest differences in susceptibility to MS, based on geographical distribution and native descent. In northern Europe 0.2-0.3 percent of the population is diagnosed with MS. Within this population, Lapps in Scandinavia appear to be resistant to the disease [14].

The incidence of MS in first degree relatives is 20 times higher than in the general population. Monozygotic twin studies show a concordance rate of 30 percent whereas dizygotic twins show a concordance rate of less than 5 percent. These results suggest that both genetic factors and environmental exposure are important in the disease etiology.

Due to the complex genetic etiology of MS, genetic studies often fail to consistently identify significant linkage or association with genes that module disease. So far, only one locus, the human leukocyte antigen (HLA) locus, has consistently been shown to be associated with increased risk of developing MS.

5.2 Data History

This section is a brief summary of the work performed prior to this degree project [27].

Hoping to uncover non-HLA loci, that may harbor MS susceptibility genes, a two-stage, single-point association study was performed. 672 Nordic MS patients and 672 Swedish healthy controls participated in the study, the same dataset later to be used in this degree project. All MS samples were provided by the Department of Neurology, Karolinska University Hospital, who also provided 288 control samples, collected from a blood bank. The remaining 384 controls were randomly selected from unrelated members of a set of Swedish mono-

and dizygotic twin pairs enrolled in the Swedish Twin Registry. All diagnoses were made according to the established Poser criteria of definite MS [21].

The SNPs used in the study were picked from a variety of public databases available on the Web; the SNP Consortium, the SNP-database of NCBI and proprietary AstraZeneca SNP databases [1, 16].

Standard molecular laboratory techniques and instruments were used in order to extract nuclear DNA from each individual. To amplify marker regions, PCR primers were designed and PCR performed. All SNP genotyping was carried out according to standard manufactural protocols, using the Pyrosequencing technique [17].

The majority of SNPs were selected from 45 candidate genes previously identified as MS susceptibility regions, whereas the remaining ones were choosen based on the biological function of a gene, or due to previous reports. Out of the total 66 genes, 31 were genotyped for at least two SNPs whereas the remaining 35 were only genotyped for a single SNP.

In order to identify genetic variants that increase susceptibility to MS in a cost effective way, a two-stage association analysis was performed. As a first step, all markers; 123 SNPs, located in 66 genes, were tested for association, whereas only a subset of individuals; 149 to 288 cases and 84 to 288 controls, were considered. For those markers that proved significant [$p < 0.08$], a second stage of screening [$p < 0.05$] was performed with additional patients and controls included in the test. Furthermore, in total, 356 to 672 cases and 462 to 672 controls were used for the two stages.

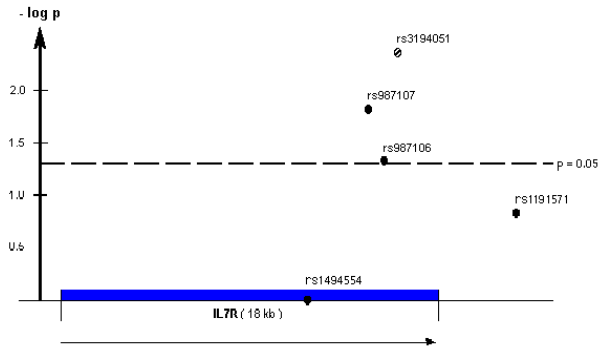

Fig. 5.1: **Results from association analysis and approximate location of genotyped SNPs within and in the vicinity of IL7R**. The significance level is p=0.05. Five SNPs in the IL7R gene were studied. Four of the SNPs are located within the gene; one exonic (rs3194051) and three intronic (rs987106, rs987107, rs1494554), whereas the last of the four SNPs (rs1494571), is flanking the gene. All three intra-genic markers have a significantly different allele distribution among MS cases compared to controls. One of them, the exonic one (rs3194051) is located in the C-terminal domain of the protein and causes a substitution, *Val356Ile*.

Out of the initial 66 genes, 22 genes in the first stage and three genes in the second stage, turned out to be significantly associated with the disease. The fact that only three out of the 22 genes tested in the second stage proved significant emphasizes the need for studying large cohorts as well as testing several markers

within each gene.

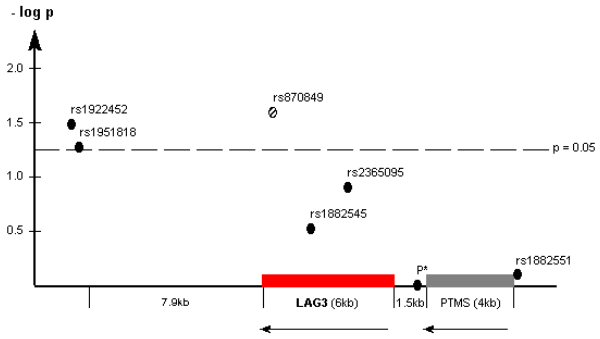

Fig. 5.2: **Results from association analysis and approximate location of genotyped SNPs within and in the vicinity of LAG3.** The significance level is p=0.05. For the LAG3 gene, nine different SNPs were studied. Three of them are located within the gene; one exonic (rs870849) and two intronic (rs1882545, rs2365095). Only the exonic SNP and two additional SNPs, located approximately 7.9kb upstream of the gene showed a positive association to Multiple Sclerosis. Neither of the six remaining SNPs, all located downstream of the gene, indicated any positive association. The striped point represents a nonsynonomous SNP within the coding area of the gene.

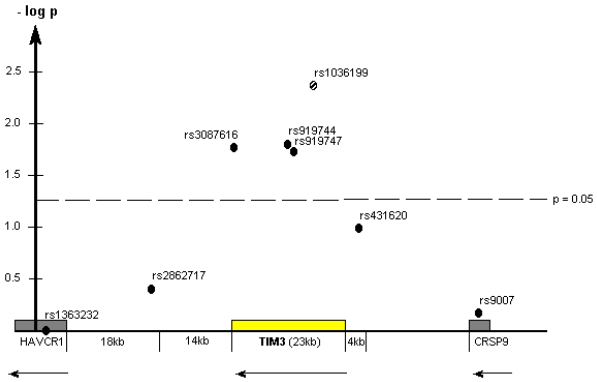

Fig. 5.3: **Results from association analysis and approximate location of genotyped SNPs within and in the vicinity of TIM3.** The significance level is p=0.05. For the TIM3 gene, eight different SNPs were studied. Four of the SNPs are located within the gene; one exonic (rs1036199), two intronic (rs919744, rs919747) and one in the 3'UTR sequence (rs3087616). All the intra-genic SNPs are significantly associated with MS, whereas no association was observed for the four neighboring SNPs up and downstream of the gene. One of the intra-genic markers (rs1036199) causes a substitution, Leu140Arg in the mucin-like domain of the protein. The change from the non-polar hydrophobic amino acid, leucine, for the basic amino acid arginine may cause structural modification of the protein.

The study suggested that three genes; TIM3, LAG3 and IL7R confer susceptibility to Multiple Sclerosis. The first two genes encode molecules that help mediate the inhibition of of activated T cells, while the protein encoded by the

third gene is necessary for a normal maturation of T- and B-cells.

Figure 5.1, 5.2 and 5.3 illustrates the results from the association study for every marker within and in vicinity of the three genes. For exact numbers the reader is referred to the tabulated table in appendix B.

5.3 Data Processing

Often, statistical geneticists and bioinformaticians face a statistical problem referred to as the 'curse of dimensionality'. The reason for this is that genetic studies often have considerably more input variables than observations. In conventional association studies this problem is partly circumvented since no more than one locus is considered at a time. In logistic regression on the other hand, several loci are studied simultaneously and unless the number of observations is larger than the number of variables, standard inference fails. A stepwise logistic regression approach is feasible, but as been discussed earlier, stepwise logistic regression is suboptimal. Therefore, large sample sizes are needed in order to avoid the curse of dimensionality. The more predictor variables we include in our model, the more prominent the problem gets, and the more observations are needed. In order to avoid poor power in the analysis we therefore need to preprocess the genotyped data before studying gene-gene interactions among the three genes.

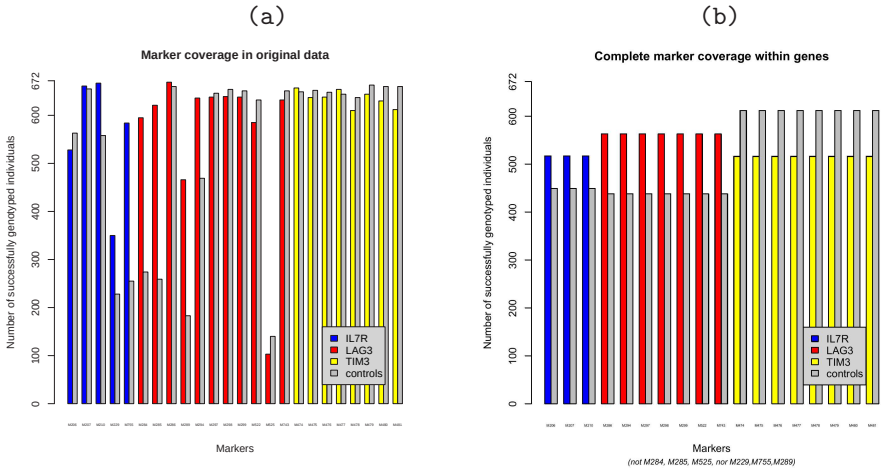

Fig. 5.4: (a) Data provided by AstraZeneca for genotyped markers within and in the vicinity of IL7R, TIM3 and LAG3. In (b) complementary or sparse marker data are removed and individuals with missing data omitted. (Enlarged figures are provided in Appendix B)

Figure 5.4(a) shows a bar plot of the unprocessed data, provided at the start of this degree project. The SNPs belonging to the three genes are colored in yellow (IL7R), red (LAG3) and blue (TIM3), and the height of the bars indicate the number of individuals genotyped at every SNP. As a first preprocessing step, we have excluded SNPs that were typed only in a fraction of the samples. An act that should not be of any concern, since such SNPs have "failed" and has been intentionally aborted for various reasons (e.g. duplicated gene, pseudogenes, no

association). Also, as illustrated in figure 5.4(b), only completely genotyped individuals where kept, the remaining ones were omitted. The reason for this harsh preprocessing is that the generalized linear model used in the implemented programs is best suited for non-missing data. Still, the number of individuals used is reasonably high.

6. METHODS AND RESULTS

6.1 Hardy-Weinberg Test

As discussed in section 3.2.2 population stratification is a legitimate concern of confounding that may generate false positives in an association study. In order to investigate the deviation of genotype frequencies from what is expected on the basis of the Hardy-Weinberg law, a Pearson's chi-square test has been performed and implemented in *R*. The idea and theory behind this test is described in section 3.4.

Three groups where tested; a pooled sample of cases and controls, and cases and controls alone. All test results are tabulated in Appendix C. Conclusions on population stratification is made sole on the control group, whereas the test results from the cases and pooled data are only briefly discussed. The significance level used is α=0.05, unless stated differently.

Neither of the markers within the LAG3 gene deviated significantly from HWE, whereas markers within the IL7R and TIM3 showed mixed test results for the three test groups. Table 6.1 illustrates the test results of the control group for markers located in the vicinity of the LAG3 gene.

	11	12	22	χ^2	p
M286	55(56.64)	205(201.7)	178(179.6)	0.12	0.734
M294	211(213.1)	189(184.8)	38(40.08)	0.22	0.637
M297	53(46.69)	180(192.6)	205(198.7)	1.88	0.170
M298	162(157.3)	201(210.4)	75(70.32)	0.87	0.352
M299	78(71.93)	199(211.1)	161(154.9)	1.45	0.229
M522	41(46.69)	204(192.6)	193(198.7)	1.53	0.217
M743	69(68.33)	208(209.3)	161(160.3)	0.02	0.894

Tab. 6.1: **Hardy-Weinberg test results for markers located in the LAG3 gene of the control group.** At a 5% significance level, neither of the markers show a significant deviation in genotype frequencies from what expected under HWE. Similar resultswhere obtained from the cases and the pooled data.

The test results for markers located in the IL7R gene deviated significantly from HWE for marker M210 within controls [p=0.036] (Figure 6.2). This may indicate population stratification within the controls. Within cases, M206 and M210 showed significant test results [p=0.008, p=0.022], whereas no significant deviation was observed in pooled data. It is possible that such deviations are caused by population stratification but may very well indicate a genuine association between markers and disease.

Figure 6.3 show the test statistics of markers located in the TIM3 gene.

	11	12	22	χ^2	p
M206	254(259)	174(164)	21(26)	1.65	0.198
M207	126(134.8)	240(222.4)	83(91.78)	2.80	0.094
M210	254(262)	178(162)	17(25)	4.41	0.036

Tab. 6.2: **Hardy-Weinberg test results for markers located in the IL7R gene of the control group.** Marker M210 deviates significantly from Hardy-Weinberg equilibrium.

These results suggest a population stratification for markers M474, M475, M476 and M478 in the controls [p=0.029, 0.043, 0.029 and 0.002 respectively]. In addition M479 proved significant at an elevated significance level of 10% [p=0.053]. Interestingly, none of the marker genotypes tested in cases or the pooled data,

	11	12	22	χ^2	p
M474	29(20.86)	168(184.3)	415(406.9)	4.77	0.029
M475	28(20.50)	168(183.0)	416(408.5)	4.12	0.043
M476	29(20.86)	168(184.3)	415(406.9)	4.77	0.029
M477	31(30.22)	210(211.6)	371(370.2)	0.03	0.856
M478	439(428.3)	146(167.3)	27(16.34)	9.94	0.002
M479	407(399.6)	175(189.9)	30(22.56)	3.76	0.053
M480	60(66.67)	284(270.7)	268(274.7)	1.49	0.223
M481	468(471.2)	138(131.6)	6(9.191)	1.44	0.230

Tab. 6.3: **Hardy-Weinberg test results for markers located in the TIM3 gene of the control group.** M474, M475, M476 and M478 deviate significantly from Hardy-Weinberg equilibrium.

deviate significantly from HWE.

6.2 Linkage Disequilibrium

Recall that both the Hardy-Weinberg test and the LD test assumes independent allele frequencies but the HW test is tested on genotype frequencies, whereas the pairwise LD test is tested on allele frequencies between two loci. In order to investigate allelic dependencies within and between IL7R, LAG3 and TIM3, the freeware program EMLD was used to calculate pairwise linkage disequilibrium measures in cases, controls and a pool of the two [7].

The LD test results are tabulated in Appendix D and provide three kind of measures; D, $|D'|$ and R^2. Based on the measures $|D'|$ and R^2, the LD patterns where illustrated in GOLD plots, but only $|D'|$ was critically analyzed. Little difference was observed in LD patterns of cases, controls and pooled data, so only LD patterns from the controls are considered analytically.

Recall that $|D'|$ resembles the recombination rate and that it tends to be upwardly biased for small sample sizes, whereas it is insensitive to allele frequencies. R^2 on the other hand, is a statistical association measure that do depend on allele frequencies and has a simple inverse relationship between its measure and the sample size. Both, $|D'|$ and R^2 takes the value 1 if no more

than two haplotypes are present at the locus and intermediate c-values are hard to interpret.

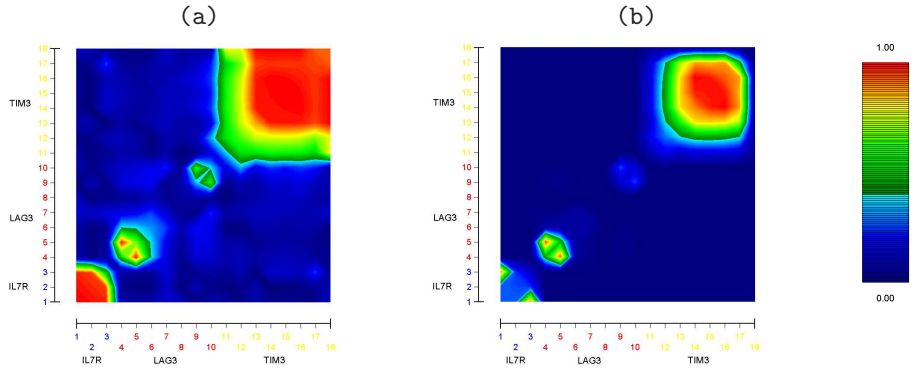

Fig. 6.1: **Pairwise LD of controls within and between IL7R, LAG3 and TIM3.** The genes are colored in different colors along the axis; blue, red and green. A color scale on the right, denotes the extent of LD between two markers. The value 1 (red) illustrates complete LD, whereas the value 0 (blue) illustrates LE. (a) LD plot of $|D'|$ measurement. A clear block structure of LD is observed in IL7R and TIM3, whereas only two markers are in LD within the LAG3 gene (b) LD plot of R^2 measurement. The plot show more moderate values of LD and a slightly different block structure within IL7R compared to the plot generated by the $|D'|$-measure.

The two figures in 6.1 show fairly similar block structures for the $|D'|$ and R^2 measurements, at least in the LAG3 and TIM3 region, whereas the LAG3 region show different LD patterns. It is obvious that $|D'|$ is upwardly biased compared to R^2 and that the choice of LD measure may influence the interpretation of the LD structure. For the purpose of this thesis the plot created by the $|D'|$ measure, is of biggest interest since it is related to the recombination rate between two markers and provides valuable information on the stability and consequently the variation in the region. The results suggest a block structure within the *IL7R* and the *TIM3* region, whereas the LAG3 region is in linkage equilibrium.

In the context of recombination rate, the $|D'|$ results indicate very low recombination rates in IL7R and TIM3, whereas the LAG3 has a fairly high recombination rate. Note, that *rs1922452* and *rs951818* indicate a considerably lower recombination frequency between each other than other surrounding markers.

6.3 Logistic Regression

The goal of a logistic regression analysis is to find the best fitting and most parsimonious, yet biologically reasonable, model to describe the relationship between a response variable and a set of predictor variables. Next, the results from applying forward-, backward- and exhaustive logistic regression in a two-staged process are presented.

6.3.1 Parsimonious Selection

As discussed in section 4.4, the purpose of the first stage is to narrow down the number of markers within each gene to a minimal set, capable of describing the genetic variation between cases and controls. If this pre-processing step is left out, the use of logistic regression to study gene-gene interaction is likely to be a daunting task since the number of possible interaction parameters quickly becomes overwhelming.

All genotypes were coded into an additive and a dominance component in line with the coding scheme suggested by Cordell and Clayton, described in section 4.5. Recall that the dominance component does not result in a dominance effect unless its corresponding additive component is present. If one insists on interpreting the dominance component alone, the coding scheme resembles a heterozygous advantage/disadvantage. Despite the unlikely scenario of heterozygous advantage effects of SNPs, the implemented stepwise logistic regression algorithms performs χ^2-tests where dominant components are treated without its corresponding additive component. The reason for this is that the biological interpretation of the results in stage one is not a serious concern. Instead, the purpose of this stage is to narrow down the number of components within each gene and focus on identifying components that can be left out. The most appropriate and intuitively appealing stepwise approach for leaving terms out is the backward elimination method. However, whenever an exhaustive approach is manageable, it is favored.

In order to depict the most parsimonious subset of markers for the three candidate genes; forward, backward and exhaustive logistic regression have been performed on each gene separately. The complete statistical test results for the stepwise regression models are tabulated and commented in Appendix E, whereas significant results from the exhaustive regression model are presented in the main text.

IL7R

Three SNPs; rs987107, rs987106 and rs3194051, each coded into an additive and a dominance component, were analyzed for the IL7R gene. For convenience the six variables are referred to as components 1-6. Logistic regression modelling of the gene emphasized two tuples; *3 5 6* and *1 3 6*. Both tuples are three component models with 962 degrees of freedom (df), put in relation to the six component full model (959 df) and the zero component null model (965 df). The first tuple, *3 5 6* models two SNPs, the additive effect of rs987106 and a full effect (additive component + dominance component) of rs3194051. The second tuple, *1 3 6* models three SNPs, an additive effect of rs987107 and rs987106 respectively, and a dominance component of rs3194051. The χ^2 difference between the full model and null model was clearly significant [$\Delta\chi^2(\Delta\text{df}=6)=22.86$, $p=0.001$]. Neither of the individual components showed a significant improvement in fit compared to the null model in the forward stepwise procedure. Supported by the both stepwise procedure tests, the results did however indicate that component number *4* can be ruled out as an informative variable. The parsimonious choice of *1 3 6* in the forward selection algorithm and *3 5 6* in the backward elimination algorithm, along with convincing exhaustive modelling results, further clarified this fact. Since the backward regression model is considered more

reliable for the purpose of stage one and the fact that the effect of *3 5 6* is more biologically appealing, involving only two loci; the components *1, 2* and *4* were left out prior to the second stage.

LAG3

The full 7-SNP model consist of 14 components, corresponding to the full effect of the markers; rs870849, rs1882551, rs1882545, rs1922452, rs951818, rs2365095 and rs11227. The components are named as numbers, 1-14. Somewhat surprising, the χ^2 difference between the full model and null model where non-significant [$\Delta\chi^2(\Delta df=14)=10.76$, $p=0.704$], indicating that all markers can be dropped from the full model without significant decline in fit. However, it seems fair to say that a combination of two, possibly three components is a parsimonious subset of the full model. Candidate combinations in the IL7R gene are *5 6 7, 6 7 12, 7 8 12, 7 11 12* and *6 7* [$\Delta\chi^2(11)=5.13$ 2.65, 4.61, 5.53 and $\Delta\chi^2(12)=5.14$; $p=0.925$, 0.995, 0.948, 0.903 and 0.953, respectively]. Only one of these marker showed a significant increase in fit compared to the null model [$\Delta\chi^2(3)=5.63$ 8.11, 6.15, 5.23 and $\Delta\chi^2(2)=5.62$; $p=0.131$, 0.044, 0.948, 0.903 and 0.953, respectively], namely component combination *6 7 12*. However, since component *12* does not seem to add much information, the nested model *6 7* were considered most interesting in an interaction study.

TIM3

For the TIM3 gene, eight SNPs were analyzed; rs919744, rs919747, rs1036199, rs8862717, rs30287616, rs431620, rs1363232 and rs9007. Again, for convenience the sixteen variables are numbered, 1-16. The χ^2 difference between the full model and null model was marginally significant [$\Delta\chi^2(16)=23.73$, $p=0.096$]. The best fitting parsimonious models all included component *10*. The backward stepwise procedure favored *5 10* [$\Delta\chi^2(14)=9.02$, $p=0.830$], the forward stepwise procedure *3 10*[$\Delta\chi^2(1)=6.49$, 8.56 and $p=0.011$, 0.003 successively], and the exhaustive approach *1 10*[$\Delta\chi^2(1)=8.73$, $p=0.848$], as well as *3 10*, and *5 10*. These tuples, all had a strong significant improvement in fit when tested against the null model ($p<0.001$). Presumably, neither of these pairs can be removed from the full model without seriously decrease the fit of the model. The reason is that, a worse fitted pair, *9 10* [$\Delta\chi^2(14)=13.69$, $p=0.473$] significantly worsened the fit when dropped from the full model in the backwards elimination method [$\Delta\chi^2(2)=11.56$, $p=0.003$]. Interestingly, *9* and *10* correspond to the additive component and the dominance component of the same marker, rs30287616. It is worth noting that there is hardly any loss in fit when dropping component *9* from the full model, whereas dropping component *10* affects the fit more [$\Delta\chi^2(1)<0.01$, $p=0.973$ and $\Delta\chi^2(1)=2.04$, $p=0.153$, respectively]. That is, although neither of the two components are significant alone, they are as a pair. In a way this is reassuring, since both components need to be considered when modelling interactions between markers. Since the 3-component model, *3 9 10* models a two locus effect without extensively affecting the significance, it is considered the most parsimonious tuple in TIM3.

No, dominance component where analyzed for interaction effects in absence of its corresponding additive component. In summary, the models investigated for interaction effects were *3 5 6* for the IL7R gene, *5 6 7* for the LAG3 gene

and *3 9 10* for the TIM3 gene.

6.3.2 Modelling Interaction within Genes

Each selected tuple was analyzed for interaction effects between its corresponding components, in order to reveal potential marker interactions within the genes.

TIM3 had a significant improvement in fit when an interaction between the two markers was included [$\Delta\chi^2(1)$=7.09, p=0.008]. The main effect model contains the additive component of marker rs919747 (denoted A475), and the full effect components of marker rs30287616 (denoted F478 or A478+D478). Interaction between the two markers was modelled by the following statistical interaction models

$$logit(p) = A475 + A478 + D478 + A475 * A478 \tag{6.1}$$

and

$$logit(p) = A475 + A478 + D478 + A475 * A478 + A475 * D478 \tag{6.2}$$

The second interaction term in model (6.2) proved redundant. Therefore, model (6.1) and (6.2) are modelling the same genetic effect.

Neither of the other two genes, IL7R and LAG3 had a significant increase in fit when adding interaction components to the main model [$\Delta\chi^2(2)$=0.94, 1.16, p=0.626, 0.560 respectively]. Complete test results and parameter estimations are tabulated in Appendix F.

6.3.3 Modelling Interaction between Genes

To test for statistical interactions between the three genes, the selected components were tested for an increase in fit, when all possible combinations of first order interaction terms were singely added to the main effect model.

Like in stage one, all individuals with incompletely covered genotypes where omitted, leaving 519 controls and 592 cases. Note that this is a larger subset of the original data than used in stage one.

A significant interaction between marker *rs1882545* in the lag3 gene and marker *rs30287616* in the TIM3 gene was observed when testing for deviation in fit between the main effect model and the interaction models [$\Delta\chi^2(2)$=7.88, p=0.019]. Marker *rs1882545* in the lag3 gene also showed a significant statistical interaction with marker, *rs3194051* in the IL7R gene [$\Delta\chi^2(2)$=8.43, p=0.015]. Appendix G tabulates all results and the β-parameters of the models.

Table 6.4 summarizes the significant interaction models in stage two and some additional nested main effect models, all tested against the null model.

ID	Summarized Models
1	Null Model
2	IL7R
3	LAG3
4	TIM3
5	IL7R + LAG3
6	LAG3 + TIM3
7	IL7R + LAG3 + IL7R*LAG3
8	LAG3 + TIM3 + LAG3*TIM3
9	IL7R + LAG3 + TIM3
10	IL7R + LAG3 + TIM3 + IL7R*LAG3 + LAG3*TIM3
11	Full Interaction Model

ID	d.f	$\|D\|$	AIC	Δdf	χ^2	p	ΔAIC
1	1110	1535.37	1537.37	–	–	–	–
2	1107	1512.66	1520.66	3	22.71	$4.6 \cdot 10^{-5}$	−16.71
3	1107	1531.82	1539.82	3	3.55	0.314	2.45
4	1107	1514.91	1522.91	3	20.46	$1.4 \cdot 10^{-4}$	−14.46
5	1104	1509.09	1523.09	6	26.28	$2.0 \cdot 10^{-4}$	−14.28
6	1104	1510.93	1524.93	6	24.44	$4.3 \cdot 10^{-4}$	−12.44
7	1102	1501.22	1519.22	8	34.16	$3.8 \cdot 10^{-5}$	−18.16
8	1102	1502.51	1520.51	8	32.87	$6.5 \cdot 10^{-5}$	−16.87
9	1101	1488.94	1508.94	9	46.43	$5.0 \cdot 10^{-7}$	−28.43
10	1097	1473.27	1501.27	13	62.1	$2.2 \cdot 10^{-8}$	−36.1
11	1070	1445.87	1527.87	40	89.5	$1.2 \cdot 10^{-5}$	−9.5

Tab. 6.4: **Likelihood-Ratio Tests (χ^2) of main and interaction effects of the parsimoniously selected markers of the IL7R, LAG3 and TIM3 gene**. The main effects include all markers selected in stage one for the represented gene, whereas only markers rs3194051, rs1882545 and rs30287616 are included in the interaction terms. The p-values from the likelihood-ratio test provided in the table tells us that all the presented models except the main effect of the LAG3 gene show a significantly better fit compared to the null model.

7. DISCUSSION

Previous association studies have had limited success in revealing the susceptibility loci influencing Multiple Sclerosis [19]. So far, only the HLA II region have shown consistent association to the disease. In a recently performed association study, three additional genes; IL7R, LAG3 and TIM3 showed significant differences in allele frequencies between 672 MS cases and 672 controls, suggesting involvement in disease [27]. All three genes are involved in immune homeostasis and are currently studied in more detail.

In this degree project, IL7R, LAG3 and TIM3 have been analyzed for gene-gene interactions using a genotype-based logistic regression procedure. The analysis was performed in two stages. First, the most parsimonious subset of markers were selected within each gene, using three different methods; forward selection, backward elimination and a best subset selection method. Next, within and between gene interactions were analyzed for the parsimoniously selected markers.

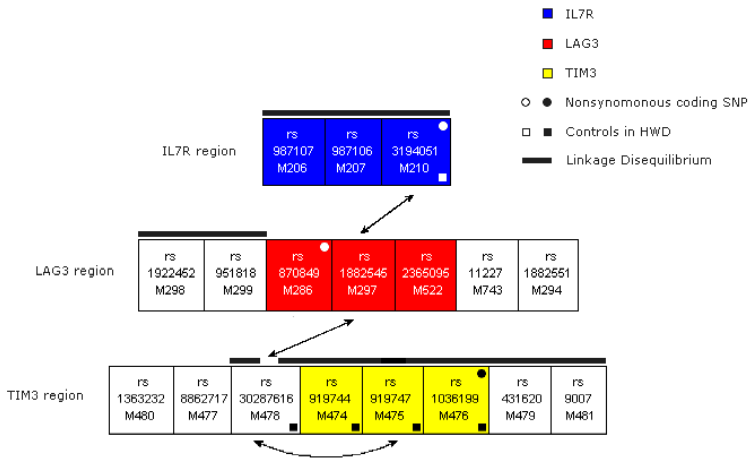

Fig. 7.1: **Gene-Gene Interaction.** The IL7R, LAG3 and TIM3 are displayed in different colors, whereas flanking loci are uncolored. The markers in a genetic region are placed in the sequence they appear in the genome, but no information on interlocus distance is given. Significant interaction results are illustrated with marker joined, double headed arrows. Information on HWD, LD and coding SNPs is also provided.

Our results suggest interaction effects between rs1882545 in the LAG3 gene and two other markers; rs3194051 in the IL7R gene and rs30287616 in the TIM3

gene. An additional interaction was observed between rs919747 and rs30287616 within the TIM3 gene.

Figure 7.1 illustrates statistically significant interaction results between and within the genes, Hardy-Weinberg disequilibrium, linkage disequilibrium and information on non-synonymous coding SNPs. It is alarming that some of the interacting SNPs are in HWD in the controls. This suggests that the controls are genotypically stratified and that the significant interaction effects modelled may be dubious. If a genuine association exists after all, confounding such as population admixture may either weaken or strengthen the association between interacting markers and disease. Consequently, stratification within controls does not necessarily mean that the association is not valid, it might just as well mask a genuine association.

In general however, population stratification is unwanted, so the HWD observed for rs3194051 and the majority of markers within and in the vicinity of the TIM3 gene is a concern. In order to proceed with TIM3, more balanced control data is preferred. In this thesis, population stratification within cases and pooled data have not been considered analytically. Even though such an analysis is of interest, it is not straightforward. There is no obvious way of distinguishing differences in genotype frequencies of a susceptibility locus from those observed in a stratified population.

The results on linkage disequilibrium, provided edged LD maps in IL7R and TIM3, whereas no block structures of LD could be established in the LAG3 region. In addition, if all markers within the LAG3 region where included in a logistic regression model, no significant, or even moderate increase in fit was observed, compared to the null model. This is somewhat surprising since more than one marker in that region showed significant associations in the single-point association study performed prior to this project. However, all those associations where not particularly strong and it is possible that the association was lost in the pre-processing step of the data. Therefore, it would be useful to perform a single-point association study on the processed data to find out if an association still exists.

Still, there are reasons to be optimistic about the obtained results. If the epistatic effects of the unedged LAG3 gene prove consistent in future studies it is supporting evidence for its involvement in disease, since it is not clear from the LD structure, wether the presumable susceptibility locus is located in the LAG3 gene or at a neighboring cite, such as the biologically similar CD4 gene. Another interesting follow up is to identify more coding markers within the three regions. Also, the parsimonious selection was based on the most significant test results, jet restricted to selection of a few representative SNPs within each gene. Additional interaction effects within and between markers located in these genetic regions are expected. For example, it is possible that rs987107, left out in the parsimonious selection of the IL7R gene has similar interaction effects, since it is in strong LD with the two selected loci. Similar effects may be present in the TIM3 region.

Both genotype analysis and haplotype analysis have been discussed in this thesis, and appropriate methods for studying gene-gene interaction suggested. By fitting different models that include main effects and/or statistical interactions, a wide variety of null hypothesis can be tested. Although the statistical methodology is not new, logistic regression procedures of multi-locus genotypes and haplotypes is sparsely documented in human genetics literature. A two-

staged logistic regression procedure is proposed for the identification of main and epistatic effects in complex human disease. Stage one is useful in the evaluation of the relative importance of variants at different sites within a small genetic region and stage two promising in the identification of biological epistatic effects.

Still, there are many conflicting definitions of epistasis and a statistical model does not necessarily imply a biological interaction. Nevertheless, if both a statistical and a biological definition of epistasis is interpreted as a non-additive effect between markers, statistical interactions between non-synonymously coding SNPs, such as the observed interaction of rs3194051 in the LAG3 gene is compelling evidence of a biological interaction.

For a molecular biologist or biochemist, a statistical interaction may provide useful information on inhibitory or enhancing factors in protein signalling pathways and disease etiology.

REFERENCES

[1] Single Nucleotide Polymorphisms for biomedical research. *Internet: http://snp.cshl.org*, 2003.

[2] H. Campbell and I. Rudan. Interpretation of genetic association studies in complex disease. *The Pharmacogenomics journal*, 2:349–360, 2002.

[3] E. Casa-Espern and C. Sapienza. Natural selection and the evolution of genome imprinting. *Annual review of genetics*, 37:349–370, Dec 2003.

[4] H. Cordell. Epistatis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[5] H. Cordell and D. Clayton. A unified stepwise regression procedure for evaluation the relative effects of polymorphisms within a gene using case/control of family data: application to HLA in type 1 diabetes. *American journal of human genetics*, 70:124–141, 2002.

[6] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–322, September 2000.

[7] EMLD. *Internet: http://request.mdacc.tmc.edu/ qhuang/Software/pub.htm*, April 2004. (April 17, 2004).

[8] M.P. Epstein and G.L. Satten. Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet.*, 73(6):1316–1329, Dec 2003.

[9] D.S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annual Human Genetics*, 29:399–433, 1965.

[10] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52:51–76, 1918.

[11] D.B. Goldstein. islands of linkage disequilibrium. *Nature*, 29:109–111, October 2001.

[12] M.R Hoehe. High resolution haplotypes, the HapMap and the analysis of genotype-phenotype relationships. *Poster*.

[13] L.B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome research*, 10:1435–1444, 2000.

[14] J.F. Kurtzke. Epidemiological evidence for multiple sclerosis as an infection. *Clin. Microbiol. Rev.*, 6:382–427, 1993.

[15] R.W. Morris and N.L. Kaplan. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet. Epidemiol.*, 277(48):46328–46337, Nov 2002.

[16] NCBI. Single Nucleotide Polymorphism. *Internet: www.ncbi.nlm.nih.gov/SNP*, 2003.

[17] L. Nordfors, M. Jansson, G. Sandberg, Lavebratt C., S. Sengul, Schalling M., and et al. Large-scale genotyping of single nucleotide polymorphisms by Pyrosequencingtrade mark and validation against the 5'nuclease (Taqman((R))) assay. *Hum Mutat*, 19:395–401, April 2002.

[18] Human Genome Program U.S. Department of Energy. *Genomics and its impact on science and society: A 2003 Primer*. Internet, http://www.ornl.gov/sci/techresources/Human Genome/publicat/primer2001/index.shtml, March 2003. (May 2, 2004).

[19] J.R. Oksenberg, S.E. Baranzini, S.E. Barcellos, and S.L. Hauser. Multiple sclerosis: genomic rewards. *Journal of Neuroimmunology*, 113(2):171–184, 2001.

[20] E. Pennisi. Bioinformatics. Gene counters struggle to get the right answer. *Science*, 301:1040–1041, August 2003.

[21] C.M. Poser, D.W. Paty, Scheinberg L., W.I. McDonald, F.A. Davis, and G.C. et al. Ebers. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Annual Neurology*, 13:227–231, March 1983.

[22] N.J. Risch. Searching for genetic determinants in the new millenium. *Nature*, 405:847–856, June 2000.

[23] P. Sham. *The analysis of segregation and population frequencies*, pages 13–50. Arnold, 1998.

[24] A-C Syvänen. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature*, 2:930–942, December 2001.

[25] J.G. Taylor, E-H Choi, C.B. Foster, and S.J. Chanock. Using genetic variation to study human disease. *Trends in Molecular Medicine*, 7(11):507–512, Nov 2001.

[26] J.D. Wall and J.K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature reviews*, 4:587–597, August 2003.

[27] Z. Zhang, K. Duvefelt, F. Svensson, T. Masterman, G. Jonasdottir, H. Salter, T. Emhaizon, D. Hellgren, G. Falk, T. Olsson, J. Hillert, and M. Anvret. Two genes encoding immune-regulatory molecules (lag3, il7r) confer susceptibility to multiple sclerosis. Unpublished, May 2004.

LITERATURE LIST

D. Clayton. Population association. In *Handbook of statistical genetics*, pages 519–540. John Wiley & Sons, 2001.

R.R. Hudson. Linkage disequilibrium and recombination. In *Handbook of statistical genetics*, pages 309–324. John Wiley & Sons, 2001.

D.C Rao and Micheal A. Province. *Genetic dissection of complex human traits.* Academic press, 2001.

P. Sham. *The analysis of segregation and population frequencies*, pages 13–50. Arnold, 1998.

E.A. Thompson. Linkage analysis. In *Handbook of statistical genetics*, pages 541–561. John Wiley & Sons, 2001.

ACKNOWLEDGEMENTS

I would like to thank Neurotec, AstraZeneca and MEB for enabling and supporting this project.

Special thanks goes to the panel of scientists taking part in the project; my two supervisors, Juni Palmgren and Hugh Salter for excellent guidance, Kristina Duvefelt and Frida Svensson for providing genotype data and knowledge in molecular biology, and lastly, Gudrun Jonasdottir, Keith Humphreys and Mark Iles for their support and knowledge in statistical genetics.

I also wish to express my gratitude to Elena Jazin for scientifically reviewing the report on a tight timeline.

To all biostatisticians at MEB, you are a great mix of people! I can't thank you enough for the relaxed atmosphere and laughs you've provided. Paul, your stories didn't bore me once! Marlies, you're up there, inviting me for dinner and "kubbspel" when thesis writing was taking over.

Appendix A

MARKER NAMES AND GENOTYPE COVERAGE IN DATA

IL7R		
rs	rs	rs
987107	987106	3194051
M206	M207	M210

LAG3						
rs	rs	rs	rs	rs	rs	rs
870849	1882551	1882545	1922452	951818	2365095	11227
M286	M294	M297	M298	M299	M522	M743

TIM3							
rs	rs	rs	rs	rs	rs	rs	rs
919744	919747	1036199	8862717	30287616	431620	1363232	9007
M474	M475	M476	M477	M478	M479	M480	M481

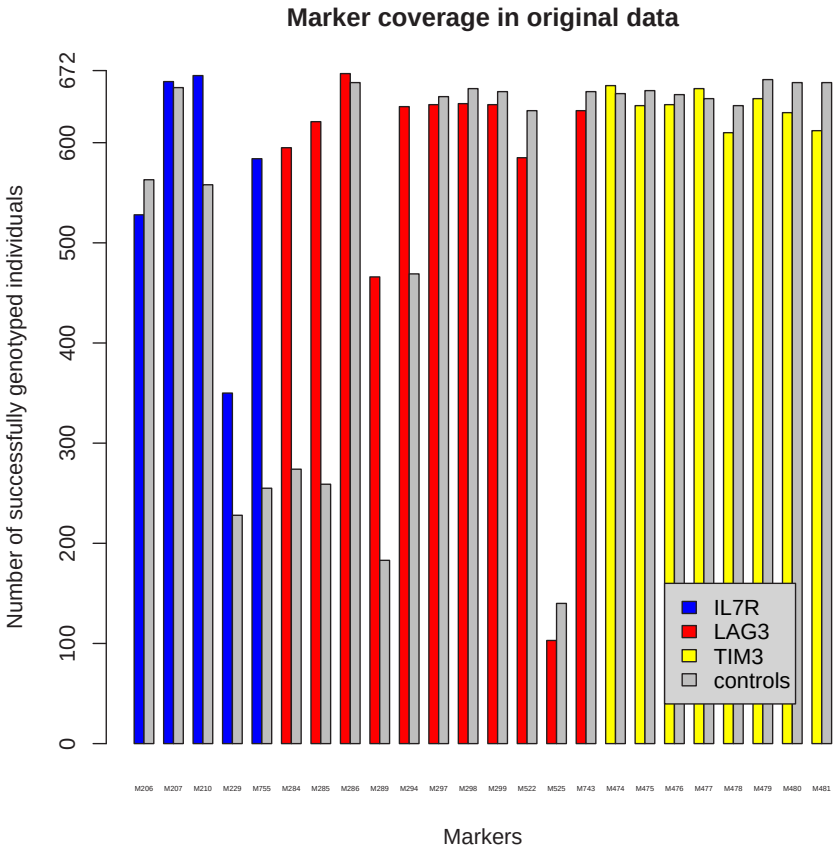

Fig. A.1: Data provided by AstraZeneca for genotyped markers within and in the vicinity of IL7R, TIM3 and LAG3.

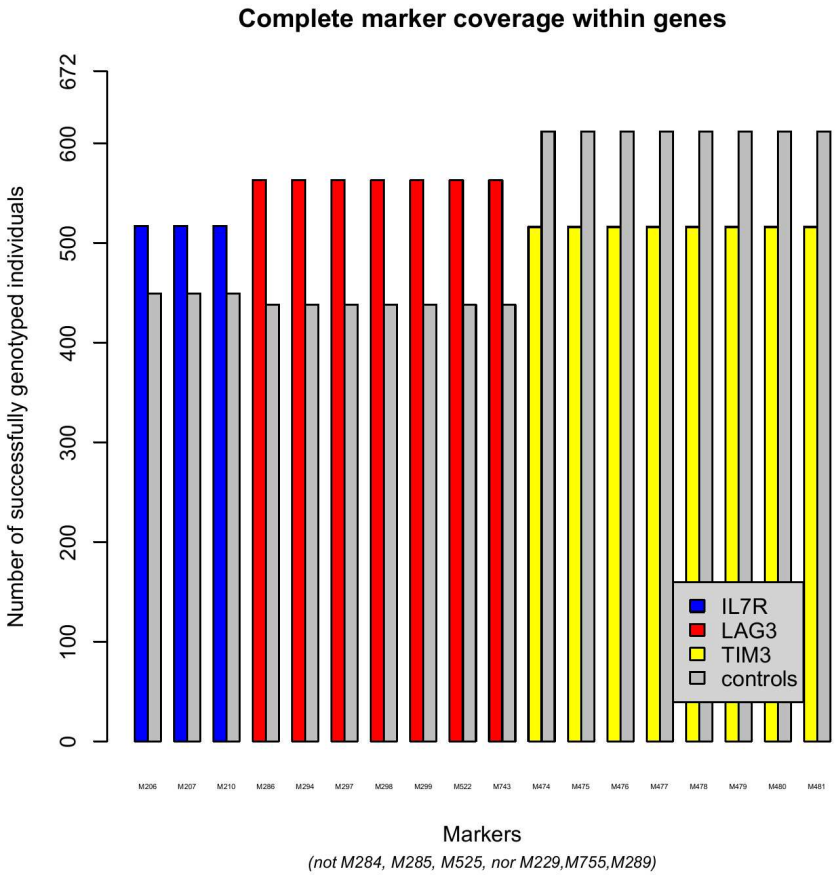

Fig. A.2: Complementary or sparse marker data are removed and individuals with missing data omitted.

Appendix B

DATA HISTORY

	Genotypes of cases				Genotypes of controls									
SNP	11	12	22	Sum	11	12	22	Sum	Base1	Base2	Power	P-value	-log p	OR
IL7R														
rs1494554	195.00	126.00	29.00	350.00	126.00	83.00	19.00	228.00	A	C	0,867	0,915	0,04	0
rs987107	289.00	186.00	53.00	528.00	301.00	228.00	34.00	563.00	C	T	0,992	0,015	1,83	0,58
rs987106	154.00	348.00	157.00	659.00	109.00	283.00	148.00	540.00	T	A	1,000	0,043	1,37	1,28
rs3194051	378.00	235.00	54.00	667.00	305.00	230.00	23.00	558.00	A	G	0,996	0,004	2,36	0,49
rs1494571	57.00	202.00	325.00	584.00	18.00	102.00	135.00	255.00	G	C	0,920	0,207	0,68	0
LAG3														
rs1922452	257.00	296.00	85.00	638.00	231.00	308.00	115.00	654.00	C	T	1,000	0,034	1,47	1,34
rs951818	250.00	294.00	91.00	635.00	225.00	305.00	119.00	649.00	T	G	1,000	0,052	1,28	1,34
rs870849	236.00	248.00	92.00	576.00	276.00	309.00	77.00	662.00	C	T	1,000	0,026	1,58	0,69
rs1882545	280.00	291.00	68.00	639.00	296.00	280.00	82.00	658.00	T	C	1,000	0,305	0,52	0
rs2365095	61.00	247.00	276.00	584.00	68.00	293.00	271.00	632.00	C	T	1,000	0,125	0,90	0
P****	263.00	20.00	0.00	283.00	258.00	19.00	0.00	277.00	C	A	0,453	0,923	0,03	0
rs1882551	50.00	268.00	319.00	637.00	22.00	116.00	133.00	271.00	C	G	0,984	0,783	0,11	0
rs11227	248.00	295.00	90.00	633.00	237.00	306.00	107.00	650.00	T	G	1,000	0,316	0,50	0
rs25557	27.00	181.00	428.00	636.00	11.00	82.00	173.00	266.00	GTTA	–	0,952	0,513	0,29	0
TIM3														
rs1363232	276.00	292.00	63.00	631.00	288.00	307.00	65.00	660.00	G	A	1,000	0,935	0,03	0
rs2862717	379.00	249.00	28.00	656.00	389.00	225.00	34.00	648.00	A	G	0,998	0,406	0,39	0
rs3087616	405.00	192.00	13.00	610.00	431.00	178.00	29.00	638.00	A	G	0,992	0,017	1,78	2,19
rs919744	439.00	203.00	16.00	658.00	441.00	178.00	32.00	651.00	G	C	0,995	0,016	1,80	2,07
rs919747	422.00	200.00	15.00	637.00	442.00	180.00	31.00	653.00	C	T	0,995	0,019	1,72	2,07
rs1036199	419.00	206.00	13.00	638.00	439.00	178.00	32.00	649.00	T	G	0,995	0,004	2,39	2,49
rs431620	413.00	215.00	19.00	647.00	445.00	193.00	31.00	669.00	G	C	0,995	0,106	0,98	0
rs9007	459.00	132.00	11.00	602.00	510.00	141.00	10.00	661.00	T	C	0,966	0,663	0,18	0

Appendix C

HARDY-WEINBERG EQUILIBRIUM

As discussed in section

	11	12	22	χ^2	p
M206	284(272)	182(206)	51(39)	7.02	0.008
M207	129(135.3)	271(258.4)	117(202.3)	1.24	0.266
M210	288(277.8)	182(202.3)	47(36.84)	5.22	0.022

Tab. C.1: IL7Rcases

	11	12	22	χ^2	p
M206	254(259)	174(164)	21(26)	1.65	0.198
M207	126(134.8)	240(222.4)	83(91.78)	2.80	0.094
M210	254(262)	178(162)	17(25)	4.41	0.036

Tab. C.2: IL7Rcontrols

	11	12	22	χ^2	p
M206	538(530.7)	356(370.6)	72(64.7)	1.50	0.221
M207	255(269.8)	511(481.4)	200(214.8)	3.64	0.056
M210	542(539.6)	360(364.7)	64(61.63)	0.16	0.686

Tab. C.3: IL7Rboth

	11	12	22	χ^2	p
M286	80(75.37)	252(261.3)	231(226.3)	0.71	0.401
M294	279(283.5)	241(232.0)	43(47.48)	0.84	0.359
M297	59(63.11)	259(250.8)	245(249.1)	0.61	0.436
M298	226(228.3)	265(260.4)	72(74.28)	0.17	0.678
M299	77(77.22)	263(262.6)	223(223.2)	0.00	0.969
M522	60(56.28)	236(56.28)	267(263.3)	0.53	0.468
M743	81(81.72)	267(265.6)	215(215.7)	0.02	0.897

Tab. C.4: LAG3cases

	11	12	22	χ^2	p
M286	55(56.64)	205(201.7)	178(179.6)	0.12	0.734
M294	211(213.1)	189(184.8)	38(40.08)	0.22	0.637
M297	53(46.69)	180(192.6)	205(198.7)	1.88	0.170
M298	162(157.3)	201(210.4)	75(70.32)	0.87	0.352
M299	78(71.93)	199(211.1)	161(154.9)	1.45	0.229
M522	41(46.69)	204(192.6)	193(198.7)	1.53	0.217
M743	69(68.33)	208(209.3)	161(160.3)	0.02	0.894

Tab. C.5: LAG3controls

	11	12	22	χ^2	p
M286	135(132)	457(463)	409(406)	0.17	0.682
M294	490(496.5)	430(416.9)	81(87.53)	0.98	0.322
M297	112(109.8)	439(443.4)	450(447.8)	0.10	0.752
M298	388(385.3)	466(471.5)	147(144.3)	0.14	0.713
M299	155(148.8)	462(474.3)	384(377.8)	0.67	0.412
M522	101(102.9)	440(436.1)	460(461.9)	0.08	0.779
M743	150(150)	475(475)	376(376)	< 0.01	0.999

Tab. C.6: LAG3both

	11	12	22	χ^2	p
M474	13(16.6)	159(151.8)	344(347.8)	1.15	0.284
M475	12(16.4)	160(151.2)	344(348.4)	1.75	0.186
M476	12(16.76)	162(152.5)	342(346.8)	2.01	0.156
M477	26(29.80)	196(188.4)	294(297.8)	0.84	0.360
M478	347(351.7)	158(148.6)	11(15.70)	2.06	0.151
M479	332(336.2)	169(160.6)	15(19.19)	1.40	0.236
M480	47(54.4)	241(226.3)	228(235.4)	2.19	0.139
M481	393(396.8)	119(111.4)	4(7.814)	2.42	0.120

Tab. C.7: TIM3cases

	11	12	22	χ^2	p
M474	29(20.86)	168(184.3)	415(406.9)	4.77	0.029
M475	28(20.50)	168(183.0)	416(408.5)	4.12	0.043
M476	29(20.86)	168(184.3)	415(406.9)	4.77	0.029
M477	31(30.22)	210(211.6)	371(370.2)	0.03	0.856
M478	439(428.3)	146(167.3)	27(16.34)	9.94	0.002
M479	407(399.6)	175(189.9)	30(22.56)	3.76	0.053
M480	60(66.67)	284(270.7)	268(274.7)	1.49	0.223
M481	468(471.2)	138(131.6)	6(9.191)	1.44	0.230

Tab. C.8: TIM3controls

	11	12	22	χ^2	p
M474	42(37.44)	327(336.1)	759(754.4)	0.83	0.362
M475	40(36.89)	328(334.2)	760(756.9)	0.39	0.532
M476	41(37.62)	330(336.8)	757(753.6)	0.45	0.500
M477	57(59.93)	406(400.1)	665(667.9)	0.24	0.623
M478	786(780)	304(316)	38(32)	1.62	0.202
M479	739(735.7)	344(350.5)	45(41.75)	0.39	0.533
M480	107(121.0)	525(496.9)	496(510.0)	3.60	0.058
M481	861(868)	257(243)	10(17)	3.75	0.053

Tab. C.9: TIM3both

Appendix D

PAIRWISE LINKAGE DISEQUILIBRIUM

A pairwise linkage disequilibrium test have been performed between every possible marker within and among genes, using EMLD. No LD was observed between markers located in different genes, so the below presented tables only illustrate the pairwise LD results within genes.

$$D_{ij} = p_{ij} - p_{i\cdot}p_{\cdot j} \tag{D.1}$$

$$D'_{ij} = D_{ij}/D_{max} \tag{D.2}$$

$$D^2 = \sum_{i,j} D_{ij}^2$$

$$R^2 = \triangle^2 = D^2 / \sum_{i,ki\neq k} \sum_{j,lj\neq l} p_{i\cdot}p_{k\cdot}p_{\cdot j}p_{\cdot l} \tag{D.3}$$

Every marker have been assigned a unique number; IL7R(1-3), LAG3(4-10) and TIM3(11-18).

IL7R

1	2	3
rs	rs	rs
987107	987106	3194051
M206	M207	M210

LAG3

4	5	6	7	8	9	10
rs	rs	rs	rs	rs	rs	rs
1922452	951818	870849	1882545	2365095	11227	1882551
M298	M299	M286	M297	M522	M743	M294

TIM3

11	12	13	14	15	16	17	18
rs	rs	rs	rs	rs	rs	rs	rs
1363232	8862717	30287616	919744	919747	1036199	431620	9007
M480	M477	M478	M474	M475	M476	M479	M481

Pairwise LD in Cases

M1	M2	D	D′	R^2
IL7R				
1	2	−0.130	0.963	0.340
1	3	0.191	0.990	0.943
2	3	−0.127	0.992	0.338
LAG3				
4	5	−0.200	0.875	0.737
4	6	0.063	0.477	0.075
4	7	−0.036	0.171	0.025
4	8	−0.025	0.125	0.013
4	9	−0.047	0.207	0.041
4	10	0.057	0.309	0.067
5	6	−0.061	0.448	0.068
5	7	0.037	0.178	0.026
5	8	0.022	0.110	0.010
5	9	0.035	0.151	0.023
5	10	−0.047	0.261	0.046
6	7	0.054	0.255	0.057
6	8	−0.013	0.118	0.004
6	9	−0.014	0.100	0.003
6	10	0.040	0.380	0.033
7	8	0.007	0.035	0.001
7	9	0.023	0.109	0.010
7	10	0.011	0.117	0.003
8	9	0.013	0.067	0.003
8	10	−0.016	0.082	0.006
9	10	−0.159	0.885	0.528
TIM3				
11	12	0.077	0.485	0.152
11	13	−0.067	0.575	0.140
11	14	0.069	0.583	0.150
11	15	0.071	0.590	0.155
11	16	0.071	0.587	0.155
11	17	−0.072	0.556	0.151
11	18	0.030	0.727	0.037
12	13	−0.113	0.849	0.495
12	14	0.116	0.851	0.511
12	15	0.116	0.853	0.515
12	16	0.118	0.849	0.518
12	17	−0.119	0.795	0.500
12	18	0.027	0.907	0.037
13	14	−0.138	0.971	0.921
13	15	−0.141	0.983	0.949
13	16	−0.140	0.970	0.914
13	17	0.139	0.988	0.857
13	18	−0.021	1.000	0.030
14	15	0.144	0.989	0.973
14	16	0.145	0.989	0.973
14	17	−0.145	1.000	0.906
14	18	0.022	1.000	0.031
15	16	0.145	0.989	0.967
15	17	−0.145	1.000	0.899
15	18	0.022	1.000	0.031
16	17	−0.146	0.994	0.904
16	18	0.023	1.000	0.033
17	18	−0.025	1.000	0.036

Pairwise LD in Controls

M1	M2	D	D′	R^2
IL7R				
1	2	−0.115	0.988	0.277
1	3	0.173	0.969	0.915
2	3	−0.104	0.947	0.238
LAG3				
4	5	−0.237	0.993	0.953
4	6	0.058	0.402	0.060
4	7	−0.032	0.158	0.019
4	8	−0.007	0.033	0.001
4	9	−0.035	0.148	0.021
4	10	0.024	0.128	0.011
5	6	−0.057	0.389	0.058
5	7	0.036	0.179	0.023
5	8	0.011	0.054	0.002
5	9	0.039	0.166	0.025
5	10	−0.028	0.152	0.015
6	7	0.056	0.252	0.061
6	8	−0.010	0.085	0.002
6	9	−0.028	0.204	0.015
6	10	0.019	0.173	0.008
7	8	0.008	0.036	0.001
7	9	−0.002	0.018	0.000
7	10	0.023	0.227	0.011
8	9	−0.010	0.074	0.002
8	10	−0.009	0.046	0.002
9	10	−0.115	0.621	0.259
TIM3				
11	12	0.085	0.562	0.185
11	13	−0.062	0.566	0.127
11	14	0.075	0.600	0.165
11	15	0.074	0.594	0.162
11	16	0.073	0.590	0.160
11	17	−0.069	0.534	0.138
11	18	0.033	0.811	0.046
12	13	−0.097	0.757	0.393
12	14	0.111	0.763	0.459
12	15	0.111	0.769	0.465
12	16	0.109	0.756	0.453
12	17	−0.105	0.688	0.397
12	18	0.021	0.779	0.025
13	14	−0.130	0.976	0.821
13	15	−0.131	0.976	0.831
13	16	−0.129	0.976	0.820
13	17	0.125	0.944	0.729
13	18	−0.020	1.000	0.027
14	15	0.149	0.990	0.975
14	16	0.151	1.000	1.000
14	17	−0.143	0.958	0.867
14	18	0.023	1.000	0.032
15	16	0.149	0.995	0.980
15	17	−0.142	0.952	0.849
15	18	0.022	1.000	0.031
16	17	−0.142	0.957	0.871
16	18	0.022	1.000	0.031
17	18	−0.022	0.925	0.028

Pairwise LD in Pooled Cases and Controls

M1	M2	D	D′	R^2
IL7R				
1	2	−0.122	0.974	0.305
1	3	0.183	0.981	0.931
2	3	−0.117	0.975	0.291
LAG3				
4	5	−0.219	0.937	0.847
4	6	0.061	0.439	0.068
4	7	−0.034	0.165	0.022
4	8	−0.016	0.081	0.005
4	9	−0.041	0.174	0.030
4	10	0.043	0.232	0.037
5	6	−0.059	0.418	0.063
5	7	0.036	0.179	0.025
5	8	0.017	0.084	0.005
5	9	0.037	0.160	0.025
5	10	−0.039	0.214	0.031
6	7	0.055	0.253	0.059
6	8	−0.012	0.103	0.003
6	9	−0.021	0.153	0.008
6	10	0.031	0.287	0.020
7	8	0.008	0.036	0.001
7	9	0.009	0.044	0.002
7	10	0.017	0.171	0.006
8	9	0.001	0.005	0.000
8	10	−0.013	0.067	0.004
9	10	−0.141	0.775	0.405
TIM3				
11	12	0.081	0.523	0.168
11	13	−0.064	0.570	0.133
11	14	0.072	0.592	0.158
11	15	0.072	0.592	0.159
11	16	0.072	0.588	0.158
11	17	−0.071	0.545	0.144
11	18	0.032	0.772	0.042
12	13	−0.105	0.802	0.442
12	14	0.113	0.805	0.484
12	15	0.114	0.809	0.488
12	16	0.113	0.801	0.484
12	17	−0.112	0.741	0.447
12	18	0.024	0.828	0.029
13	14	−0.134	0.973	0.868
13	15	−0.136	0.979	0.887
13	16	−0.134	0.973	0.864
13	17	0.132	0.966	0.790
13	18	−0.021	1.000	0.028
14	15	0.147	0.989	0.974
14	16	0.148	0.995	0.987
14	17	−0.144	0.978	0.886
14	18	0.022	1.000	0.031
15	16	0.147	0.992	0.973
15	17	−0.143	0.975	0.873
15	18	0.022	1.000	0.031
16	17	−0.144	0.975	0.887
16	18	0.023	1.000	0.032
17	18	−0.023	0.955	0.032

Appendix E

PARSIMONIOUS SELECTION

Forward Logistic Regression

Starting from the null model, every successive step picks the best model compared to the previous model and remaining components are added and tested one at a time. The algorithm stops when the full model is reached. A good model will have a large $\Delta\chi$-value and small p-value.

IL7R

rs 987107 M206		rs 987106 M207		rs 3194051 M210							
A	D	A	D	A	D	df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
nullmodel						965	1334.37	1336.37	–	–	–	–
1						964	1331.55	1335.55	1	2.82	0.093	−0.82
	2					964	1333.07	1337.07	1	1.3	0.254	0.7
		3				964	1331.66	1335.66	1	2.71	0.1	−0.71
			4			964	1334.27	1338.27	1	0.1	0.748	1.9
				5		964	1331.97	1335.97	1	2.4	0.121	−0.4
					6	964	1332.34	1336.34	1	2.03	0.155	−0.03
1	2					963	1324.42	1330.42	1	7.14	0.008	−5.14
1		3				963	1323.06	1329.06	1	8.49	0.004	−6.49
1			4			963	1331.47	1337.47	1	0.08	0.772	1.92
1				5		963	1331.52	1337.52	1	0.03	0.867	1.97
1					6	963	1322.22	1328.22	1	9.33	0.002	−7.33
1	2				6	962	1321.79	1329.79	1	0.43	0.51	1.57
1		**3**			**6**	**962**	**1313.18**	**1321.18**	**1**	**9.04**	**0.003**	**−7.04**
1			4		6	962	1321.72	1329.72	1	0.5	0.481	1.5
1				5	6	962	1321.87	1329.87	1	0.35	0.557	1.65
1	2	3			6	961	1312.79	1322.79	1	0.4	0.527	1.6
1		3	4		6	961	1312.98	1322.98	1	0.21	0.651	1.79
1		3		5	6	961	1312.44	1322.44	1	0.74	0.389	1.26
1	2	3		5	6	960	1311.74	1323.74	1	0.7	0.402	1.3
1		3	4	5	6	960	1312.16	1324.16	1	0.28	0.594	1.72
1	2	3	4	5	6	959	1311.51	1325.51	1	0.23	0.632	1.77

LAG3

	rs 870849 M286		rs 1882551 M294		rs 1882545 M297		rs 1922452 M298		rs 951818 M299		rs 2365095 M522		rs 11227 M743						
	A	D	A	D	A	D	A	D	A	D	A	D	A	D	df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
1	nullmodel														1000	1372.03	1374.03	–	–	–	–
2	1														999	1371.95	1375.95	1	0.08	0.772	1.92
3		2													999	1371.62	1375.62	1	0.41	0.52	1.59
4			3												999	1371.67	1375.67	1	0.36	0.55	1.64
5				4											999	1372.02	1376.02	1	0.01	0.913	1.99
6					5										999	1371.88	1375.88	1	0.15	0.696	1.85
7						**6**									**999**	**1369.62**	**1373.62**	**1**	**2.41**	**0.12**	**−0.41**
8							7								999	1369.13	1373.13	1	2.9	0.089	−0.9
9								8							999	1371.89	1375.89	1	0.14	0.711	1.86
10									9						999	1369.56	1373.56	1	2.47	0.116	−0.47
11										10					999	1371.87	1375.87	1	0.16	0.687	1.84
12											11				999	1371.79	1375.79	1	0.24	0.622	1.76
13												12			999	1369.86	1373.86	1	2.17	0.141	−0.17
14													13		999	1371.62	1375.62	1	0.41	0.524	1.59
15														14	999	1372.03	1376.03	1	0	0.984	2
16	1						7								998	1369.1	1375.1	1	0.03	0.861	1.97
17		2					7								998	1368.35	1374.35	1	0.78	0.378	1.22
18			3				7								998	1369.06	1375.06	1	0.07	0.789	1.93
19				4			7								998	1369.11	1375.11	1	0.03	0.873	1.97
20					5		7								998	1368.73	1374.73	1	0.4	0.525	1.6
21						**6**	**7**								**998**	**1366.41**	**1372.41**	**1**	**2.73**	**0.099**	**−0.73**
22							7	8							998	1368.2	1374.2	1	0.93	0.334	1.07
23							7		9						998	1369.12	1375.12	1	0.02	0.902	1.98
24							7			10					998	1368.29	1374.29	1	0.84	0.359	1.16
25							7				11				998	1369.01	1375.01	1	0.13	0.723	1.87
26							7					12			998	1366.97	1372.97	1	2.16	0.141	−0.16
27							7						13		998	1369.03	1375.03	1	0.1	0.747	1.9
28							7							14	998	1369.1	1375.1	1	0.03	0.861	1.97
29	1					6	7								997	1366.26	1374.26	1	0.14	0.707	1.86
30		2				6	7								997	1365.38	1373.38	1	1.03	0.311	0.97
31			3			6	7								997	1366.34	1374.34	1	0.06	0.805	1.94
32				4		6	7								997	1366.38	1374.38	1	0.03	0.874	1.97
33					5	6	7								997	1366.39	1374.39	1	0.01	0.918	1.99
34						6	7	8							997	1365.56	1373.56	1	0.84	0.359	1.16
35						6	7		9						997	1366.37	1374.37	1	0.04	0.847	1.96

36						6	7			10					997	1365.63	1373.63	1	0.77	0.379	1.23
37						6	7				11				997	1366.24	1374.24	1	0.17	0.68	1.83
38						6	7					12			997	1363.92	1371.92	1	2.49	0.115	−0.49
39						6	7						13		997	1366.3	1374.3	1	0.11	0.742	1.89
40						6	7							14	997	1366.37	1374.37	1	0.03	0.855	1.97
41	1					6	7					12			996	1363.74	1373.74	1	0.18	0.668	1.82
42		2				6	7					12			996	1362.91	1372.91	1	1.01	0.314	0.99
43			3			6	7					12			996	1363.87	1373.87	1	0.04	0.832	1.96
44				4		6	7					12			996	1363.88	1373.88	1	0.03	0.853	1.97
45					5	6	7					12			996	1363.91	1373.91	1	0.01	0.916	1.99
46						6	7	8				12			996	1362.92	1372.92	1	0.99	0.319	1.01
47						6	7		9			12			996	1363.92	1373.92	1	0	0.977	2
48						6	7			10		12			996	1363	1373	1	0.92	0.338	1.08
49						6	7				11	12			996	1363.76	1373.76	1	0.16	0.689	1.84
50						6	7					12	13		996	1363.82	1373.82	1	0.1	0.752	1.9
51						6	7					12		14	996	1363.87	1373.87	1	0.05	0.82	1.95
52	1	2				6	7					12			995	1362.9	1374.9	1	0.01	0.931	1.99
53		2	3			6	7					12			995	1362.84	1374.84	1	0.07	0.792	1.93
54		2		4		6	7					12			995	1362.89	1374.89	1	0.02	0.883	1.98
55		2			5	6	7					12			995	1362.9	1374.9	1	0.01	0.937	1.99
56		2				6	7	8				12			995	1361.74	1373.74	1	1.17	0.28	0.83
57		2				6	7		9			12			995	1362.91	1374.91	1	0	0.968	2
58		2				6	7			10		12			995	1361.85	1373.85	1	1.06	0.304	0.94
59		2				6	7				11	12			995	1362.73	1374.73	1	0.17	0.676	1.83
60		2				6	7					12	13		995	1362.79	1374.79	1	0.12	0.727	1.88
61		2				6	7					12		14	995	1362.87	1374.87	1	0.03	0.852	1.97
62	1	2				6	7	8				12			994	1361.73	1375.73	1	0.01	0.92	1.99
63		2	3			6	7	8				12			994	1361.66	1375.66	1	0.08	0.783	1.92
64		2		4		6	7	8				12			994	1361.72	1375.72	1	0.02	0.89	1.98
65		2			5	6	7	8				12			994	1361.74	1375.74	1	0	0.963	2
66		2				6	7	8	9			12			994	1361.74	1375.74	1	0	0.946	2
67		2				6	7	8		10		12			994	1361.72	1375.72	1	0.02	0.882	1.98
68		2				6	7	8			11	12			994	1361.56	1375.56	1	0.18	0.674	1.82
69		2				6	7	8				12	13		994	1361.63	1375.63	1	0.11	0.74	1.89
70		2				6	7	8				12		14	994	1361.72	1375.72	1	0.02	0.882	1.98
71	1	2				6	7	8			11	12			993	1361.56	1377.56	1	0.01	0.93	1.99
72		2	3			6	7	8			11	12			993	1361.48	1377.48	1	0.09	0.77	1.91
73		2		4		6	7	8			11	12			993	1361.55	1377.55	1	0.01	0.905	1.99
74		2			5	6	7	8			11	12			993	1361.56	1377.56	1	0	0.958	2
75		2				6	7	8	9		11	12			993	1361.56	1377.56	1	0	0.958	2
76		2				6	7	8		10	11	12			993	1361.54	1377.54	1	0.03	0.871	1.97
77		2				6	7	8			11	12	13		993	1361.46	1377.46	1	0.11	0.744	1.89

78		2				6	7	8			11	12			14	993	1361.54	1377.54	1	0.02	0.882	1.98
79	1	2				6	7	8			11	12	13			992	1361.45	1379.45	1	0.01	0.917	1.99
80		2	3			6	7	8			11	12	13			992	1361.44	1379.44	1	0.02	0.899	1.98
81		2		4		6	7	8			11	12	13			992	1361.4	1379.4	1	0.06	0.807	1.94
82		2			5	6	7	8			11	12	13			992	1361.45	1379.45	1	0	0.957	2
83		2				6	7	8	9		11	12	13			992	1361.45	1379.45	1	0	0.95	2
84		2				6	7	8		10	11	12	13			992	1361.43	1379.43	1	0.02	0.876	1.98
85		2				6	7	8			11	12	13		14	992	1361.39	1379.39	1	0.07	0.798	1.93
86	1	2				6	7	8			11	12	13		14	991	1361.38	1381.38	1	0.01	0.919	1.99
87		2	3			6	7	8			11	12	13		14	991	1361.37	1381.37	1	0.02	0.897	1.98
88		2		4		6	7	8			11	12	13		14	991	1361.37	1381.37	1	0.02	0.875	1.98
89		2			5	6	7	8			11	12	13		14	991	1361.39	1381.39	1	0	0.946	2
90		2				6	7	8	9		11	12	13		14	991	1361.39	1381.39	1	0	0.948	2
91		2				6	7	8		10	11	12	13		14	991	1361.37	1381.37	1	0.02	0.877	1.98
92	1	2		4		6	7	8			11	12	13		14	990	1361.36	1383.36	1	0.01	0.928	1.99
93		2	3	4		6	7	8			11	12	13		14	990	1361.31	1383.31	1	0.06	0.815	1.94
94		2		4	5	6	7	8			11	12	13		14	990	1361.36	1383.36	1	0	0.959	2
95		2		4		6	7	8	9		11	12	13		14	990	1361.36	1383.36	1	0	0.95	2
96		2		4		6	7	8		10	11	12	13		14	990	1361.34	1383.34	1	0.02	0.875	1.98
97	1	2	3	4		6	7	8			11	12	13		14	989	1361.3	1385.3	1	0.01	0.917	1.99
98		2	3	4	5	6	7	8			11	12	13		14	989	1361.31	1385.31	1	0.01	0.937	1.99
99		2	3	4		6	7	8	9		11	12	13		14	989	1361.31	1385.31	1	0	0.957	2
100		2	3	4		6	7	8		10	11	12	13		14	989	1361.29	1385.29	1	0.02	0.875	1.98
101	1	2	3	4		6	7	8		10	11	12	13		14	988	1361.28	1387.28	1	0.01	0.919	1.99
102		2	3	4	5	6	7	8		10	11	12	13		14	988	1361.28	1387.28	1	0.01	0.942	1.99
103		2	3	4		6	7	8	9	10	11	12	13		14	988	1361.28	1387.28	1	0.01	0.932	1.99
104	1	2	3	4	5	6	7	8		10	11	12	13		14	987	1361.27	1389.27	1	0	0.962	2
105	1	2	3	4		6	7	8	9	10	11	12	13		14	987	1361.27	1389.27	1	0.01	0.927	1.99
106	1	2	3	4	5	6	7	8	9	10	11	12	13		14	986	1361.27	1391.27	1	0	0.967	2

TIM3

	rs 919744 M474 A	D	rs 919747 M475 A	D	rs 1036199 M476 A	D	rs 8862717 M477 A	D	rs 30287616 M478 A	D	rs 431620 M479 A	D	rs 1363232 M480 A	D	rs 9007 M481 A	D	df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
1	nullmodel																1127	1555.56	1557.56	–	–	–	–
2	1																1126	1555.45	1559.45	1	0.11	0.745	1.89
3		2															1126	1554.02	1558.02	1	1.54	0.215	0.46
4			3														1126	1555.48	1559.48	1	0.08	0.774	1.92
5				4													1126	1553.85	1557.85	1	1.71	0.19	0.29
6					5												1126	1555.49	1559.49	1	0.07	0.789	1.93
7						6											1126	1553.46	1557.46	1	2.1	0.147	−0.1
8							7										1126	1554.51	1558.51	1	1.05	0.306	0.95
9								8									1126	1553.92	1557.92	1	1.64	0.201	0.36
10									9								1126	1555.09	1559.09	1	0.47	0.494	1.53
11										10							1126	1549.07	1553.07	1	6.49	0.011	−4.49
12											11						1126	1555.56	1559.56	1	0	0.96	2
13												12					1126	1553.28	1557.28	1	2.28	0.131	−0.28
14													13				1126	1555.48	1559.48	1	0.08	0.777	1.92
15														14			1126	1555.55	1559.55	1	0.01	0.92	1.99
16															15		1126	1555.56	1559.56	1	0	0.97	2
17																16	1126	1555.52	1559.52	1	0.04	0.838	1.96
18	1									10							1125	1540.56	1546.56	1	8.51	0.004	−6.51
19		2								10							1125	1543.09	1549.09	1	5.98	0.014	−3.98
20			**3**							**10**							**1125**	**1540.51**	**1546.51**	**1**	**8.56**	**0.003**	**−6.56**
21				4						10							1125	1542.39	1548.39	1	6.68	0.01	−4.68
22					5					10							1125	1540.85	1546.85	1	8.22	0.004	−6.22
23						6				10							1125	1545.39	1551.39	1	3.68	0.055	−1.68
24							7			10							1125	1548.9	1554.9	1	0.17	0.681	1.83
25								8		10							1125	1549.02	1555.02	1	0.05	0.825	1.95
26									9	10							1125	1545.52	1551.52	1	3.55	0.06	−1.55
27										10	11						1125	1544.24	1550.24	1	4.83	0.028	−2.83
28										10		12					1125	1547.54	1553.54	1	1.53	0.216	0.47
29										10			13				1125	1548.16	1554.16	1	0.91	0.339	1.09
30										10				14			1125	1548.83	1554.83	1	0.24	0.623	1.76
31										10					15		1125	1548.94	1554.94	1	0.13	0.714	1.87
32										10						16	1125	1548.83	1554.83	1	0.24	0.623	1.76
33	1		3							10							1124	1540.41	1548.41	1	0.1	0.758	1.9
34		2	3							10							1124	1538.14	1546.14	1	2.37	0.124	−0.37
35			3	4						10							1124	1538.02	1546.02	1	2.49	0.115	−0.49

36			3		5					10							1124	1540.49	1548.49	1	0.01	0.905	1.99
37			3			6				10							1124	1539.65	1547.65	1	0.85	0.356	1.15
38			3				7			10							1124	1538.76	1546.76	1	1.74	0.187	0.26
39			3					8		10							1124	1540.42	1548.42	1	0.09	0.77	1.91
40			3						9	10							1124	1538.35	1546.35	1	2.15	0.142	−0.15
41			3							10	11						1124	1540.08	1548.08	1	0.42	0.516	1.58
42			3							10		12					1124	1540.21	1548.21	1	0.3	0.584	1.7
43			3							10			13				1124	1540.5	1548.5	1	0.01	0.922	1.99
44			3							10				14			1124	1540.37	1548.37	1	0.14	0.713	1.86
45			3							10					15		1124	1540.5	1548.5	1	0	0.965	2
46			3							10						16	1124	1540.5	1548.5	1	0.01	0.925	1.99
47	1		3	4						10							1123	1537.87	1547.87	1	0.15	0.7	1.85
48		2	3	4						10							1123	1537.9	1547.9	1	0.12	0.73	1.88
49			3	4	5					10							1123	1537.99	1547.99	1	0.03	0.873	1.97
50			3	4		6				10							1123	1537.36	1547.36	1	0.66	0.417	1.34
51			3	4			7			10							1123	1536.17	1546.17	1	1.85	0.174	0.15
52			3	4				8		10							1123	1537.66	1547.66	1	0.36	0.55	1.64
53			3	4					9	10							1123	1537.96	1547.96	1	0.06	0.814	1.94
54			3	4						10	11						1123	1537.56	1547.56	1	0.46	0.497	1.54
55			3	4						10		12					1123	1537.7	1547.7	1	0.31	0.576	1.69
56			3	4						10			13				1123	1538.02	1548.02	1	0	0.959	2
57			3	4						10				14			1123	1537.96	1547.96	1	0.06	0.804	1.94
58			3	4						10					15		1123	1538.01	1548.01	1	0.01	0.942	1.99
59			3	4						10						16	1123	1538.01	1548.01	1	0	0.945	2
60	1		3	4			7			10							1122	1535.96	1547.96	1	0.2	0.652	1.8
61		2	3	4			7			10							1122	1536.05	1548.05	1	0.12	0.731	1.88
62			3	4	5		7			10							1122	1536.11	1548.11	1	0.05	0.815	1.95
63			3	4		6	7			10							1122	1535.56	1547.56	1	0.61	0.436	1.39
64			3	4			7	8		10							1122	1536.13	1548.13	1	0.04	0.838	1.96
65			3	4			7		9	10							1122	1536.12	1548.12	1	0.05	0.824	1.95
66			3	4			7			10	11						1122	1535.74	1547.74	1	0.43	0.512	1.57
67			3	4			7			10		12					1122	1535.88	1547.88	1	0.29	0.593	1.71
68			3	4			7			10			13				1122	1536.06	1548.06	1	0.11	0.741	1.89
69			3	4			7			10				14			1122	1536.07	1548.07	1	0.09	0.759	1.91
70			3	4			7			10					15		1122	1536.16	1548.16	1	0	0.952	2
71			3	4			7			10						16	1122	1536.13	1548.13	1	0.04	0.841	1.96
72	1		3	4		6	7			10							1121	1535.05	1549.05	1	0.51	0.474	1.49
73		2	3	4		6	7			10							1121	1533.64	1547.64	1	1.92	0.166	0.08
74			3	4	5	6	7			10							1121	1535.22	1549.22	1	0.34	0.558	1.66
75			3	4		6	7	8		10							1121	1535.51	1549.51	1	0.05	0.815	1.95
76			3	4		6	7		9	10							1121	1535.52	1549.52	1	0.04	0.835	1.96
77			3	4		6	7			10	11						1121	1535.24	1549.24	1	0.32	0.57	1.68

78			3	4		6	7			10		12					1121	1535.48	1549.48	1	0.08	0.775	1.92
79			3	4		6	7			10			13				1121	1535.44	1549.44	1	0.12	0.73	1.88
80			3	4		6	7			10				14			1121	1535.45	1549.45	1	0.11	0.735	1.89
81			3	4		6	7			10					15		1121	1535.55	1549.55	1	0.01	0.918	1.99
82			3	4		6	7			10						16	1121	1535.5	1549.5	1	0.06	0.807	1.94
83	1	2	3	4		6	7			10							1120	1533.64	1549.64	1	0	0.946	2
84		2	3	4	5	6	7			10							1120	1533.37	1549.37	1	0.28	0.597	1.72
85		2	3	4		6	7	8		10							1120	1533.64	1549.64	1	0.01	0.931	1.99
86		2	3	4		6	7		9	10							1120	1533.61	1549.61	1	0.04	0.851	1.96
87		2	3	4		6	7			10	11						1120	1533.23	1549.23	1	0.41	0.521	1.59
88		2	3	4		6	7			10		12					1120	1533.35	1549.35	1	0.3	0.587	1.7
89		2	3	4		6	7			10			13				1120	1533.51	1549.51	1	0.13	0.718	1.87
90		2	3	4		6	7			10				14			1120	1533.58	1549.58	1	0.06	0.805	1.94
91		2	3	4		6	7			10					15		1120	1533.63	1549.63	1	0.02	0.901	1.98
92		2	3	4		6	7			10						16	1120	1533.57	1549.57	1	0.07	0.788	1.93
93	1	2	3	4		6	7			10	11						1119	1533.15	1551.15	1	0.09	0.768	1.91
94		2	3	4	5	6	7			10	11						1119	1532.67	1550.67	1	0.56	0.455	1.44
95		2	3	4		6	7	8		10	11						1119	1533.23	1551.23	1	0.01	0.933	1.99
96		2	3	4		6	7		9	10	11						1119	1533.23	1551.23	1	0	0.952	2
97		2	3	4		6	7			10	11	12					1119	1533.19	1551.19	1	0.04	0.838	1.96
98		2	3	4		6	7			10	11		13				1119	1533.1	1551.1	1	0.14	0.713	1.86
99		2	3	4		6	7			10	11			14			1119	1533.16	1551.16	1	0.07	0.794	1.93
100		2	3	4		6	7			10	11				15		1119	1533.21	1551.21	1	0.02	0.882	1.98
101		2	3	4		6	7			10	11					16	1119	1533.15	1551.15	1	0.08	0.772	1.92
102	1	2	3	4	5	6	7			10	11						1118	1532.13	1552.13	1	0.54	0.462	1.46
103		2	3	4	5	6	7	8		10	11						1118	1532.66	1552.66	1	0.01	0.918	1.99
104		2	3	4	5	6	7		9	10	11						1118	1532.67	1552.67	1	0	0.993	2
105		2	3	4	5	6	7			10	11	12					1118	1532.67	1552.67	1	0	0.971	2
106		2	3	4	5	6	7			10	11		13				1118	1532.56	1552.56	1	0.12	0.731	1.88
107		2	3	4	5	6	7			10	11			14			1118	1532.61	1552.61	1	0.06	0.807	1.94
108		2	3	4	5	6	7			10	11				15		1118	1532.66	1552.66	1	0.01	0.903	1.99
109		2	3	4	5	6	7			10	11					16	1118	1532.61	1552.61	1	0.07	0.795	1.93
110	1	2	3	4	5	6	7	8		10	11						1117	1532.13	1554.13	1	0.01	0.941	1.99
111	1	2	3	4	5	6	7		9	10	11						1117	1532.13	1554.13	1	0	0.993	2
112	1	2	3	4	5	6	7			10	11	12					1117	1532.11	1554.11	1	0.02	0.875	1.98
113	1	2	3	4	5	6	7			10	11		13				1117	1532	1554	1	0.13	0.716	1.87
114	1	2	3	4	5	6	7			10	11			14			1117	1532.07	1554.07	1	0.06	0.809	1.94
115	1	2	3	4	5	6	7			10	11				15		1117	1532.12	1554.12	1	0.01	0.908	1.99
116	1	2	3	4	5	6	7			10	11					16	1117	1532.07	1554.07	1	0.06	0.799	1.94
117	1	2	3	4	5	6	7	8		10	11		13				1116	1531.99	1555.99	1	0.01	0.937	1.99
118	1	2	3	4	5	6	7		9	10	11		13				1116	1532	1556	1	0	0.987	2
119	1	2	3	4	5	6	7			10	11	12	13				1116	1531.98	1555.98	1	0.02	0.882	1.98

120	1	2	3	4	5	6	7			10	11		13	14			1116	1531.99	1555.99	1	0.01	0.943	1.99
121	1	2	3	4	5	6	7			10	11		13		15		1116	1531.99	1555.99	1	0.01	0.936	1.99
122	1	2	3	4	5	6	7			10	11		13			16	1116	1531.95	1555.95	1	0.05	0.823	1.95
123	1	2	3	4	5	6	7	8		10	11		13			16	1115	1531.94	1557.94	1	0	0.945	2
124	1	2	3	4	5	6	7		9	10	11		13			16	1115	1531.95	1557.95	1	0	0.99	2
125	1	2	3	4	5	6	7			10	11	12	13			16	1115	1531.93	1557.93	1	0.02	0.879	1.98
126	1	2	3	4	5	6	7			10	11		13	14		16	1115	1531.94	1557.94	1	0.01	0.94	1.99
127	1	2	3	4	5	6	7			10	11		13		15	16	1115	1531.87	1557.87	1	0.08	0.774	1.92
128	1	2	3	4	5	6	7	8		10	11		13		15	16	1114	1531.86	1559.86	1	0.01	0.942	1.99
129	1	2	3	4	5	6	7		9	10	11		13		15	16	1114	1531.87	1559.87	1	0	0.989	2
130	1	2	3	4	5	6	7			10	11	12	13		15	16	1114	1531.84	1559.84	1	0.02	0.881	1.98
131	1	2	3	4	5	6	7			10	11		13	14	15	16	1114	1531.86	1559.86	1	0.01	0.936	1.99
132	1	2	3	4	5	6	7	8		10	11	12	13		15	16	1113	1531.84	1561.84	1	0.01	0.94	1.99
133	1	2	3	4	5	6	7		9	10	11	12	13		15	16	1113	1531.84	1561.84	1	0	0.968	2
134	1	2	3	4	5	6	7			10	11	12	13	14	15	16	1113	1531.84	1561.84	1	0.01	0.94	1.99
135	1	2	3	4	5	6	7	8	9	10	11	12	13		15	16	1112	1531.84	1563.84	1	0	0.969	2
136	1	2	3	4	5	6	7	8		10	11	12	13	14	15	16	1112	1531.83	1563.83	1	0	0.945	2
137	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1111	1531.83	1565.83	1	0	0.973	2

Backward Logistic Regression

Starting from the full model, every marker-effect is removed one at a time and the best fitting model kept. Every successive section reduces the number of parameters by one and not until the null model is encountered does it stop. Be aware of that every model is tested for its deviation from the full model and not the from the best model as in the preceding section (forward logistic regression). A high p-value tells us that the model does not deviate much from the full model. Such a model describes the genetic variation just as well as the full model and we can therefore discard the variables not included in the model. It is debatable however if a dominant contribution effect can exist on its own (heterozygous advantage).

IL7R

rs 987107 M206		rs 987106 M207		rs 3194051 M210								
A	D	A	D	A	D	df	$\lvert D\rvert$	AIC	$\Delta\chi^2$	Δdf	p	ΔAIC
1	2	3	4	5	6	959	1311.51	1325.51	–	–	–	–
	2	3	4	5	6	960	1312.11	1324.11	0.6	1	0.44	−1.4
1		3	4	5	6	960	1312.16	1324.16	0.65	1	0.421	−1.35
1	2		4	5	6	960	1320.71	1332.71	9.2	1	0.002	7.2
1	2	3		5	6	960	1311.74	1323.74	0.23	1	0.632	−1.77
1	2	3	4		6	960	1312.62	1324.62	1.11	1	0.292	−0.89
1	2	3	4	5		960	1315	1327	3.49	1	0.062	1.49
	2	3		5	6	961	1312.34	1322.34	0.83	2	0.661	−3.17
1		3		5	6	961	1312.44	1322.44	0.93	2	0.628	−3.07
1	2			5	6	961	1321.22	1331.22	9.71	2	0.008	5.71
1	2	3			6	961	1312.79	1322.79	1.27	2	0.529	−2.73
1	2	3		5		961	1315.16	1325.16	3.65	2	0.162	−0.35
		3		**5**	**6**	**962**	**1313.44**	**1321.44**	**1.93**	**3**	**0.587**	**−4.07**
	2			5	6	962	1321.45	1329.45	9.94	3	0.019	3.94
	2	3			6	962	1329.82	1337.82	18.31	3	0	12.31
	2	3		5		962	1316.67	1324.67	5.16	3	0.161	−0.84
				5	6	963	1322.35	1328.35	10.84	4	0.028	2.84
		3			6	963	1330.65	1336.65	19.14	4	0.001	11.14
		3		5		963	1323.97	1329.97	12.46	4	0.014	4.46
					6	964	1332.34	1336.34	20.83	5	0.001	10.83
				5		964	1331.97	1335.97	20.46	5	0.001	10.46
nullmodel						965	1334.37	1336.37	22.86	6	0.001	10.86

LAG3

	rs 870849 M286		rs 1882551 M294		rs 1882545 M297		rs 1922452 M298		rs 951818 M299		rs 2365095 M522		rs 11227 M743							
	A	D	A	D	A	D	A	D	A	D	A	D	A	D	df	$\lvert D\rvert$	AIC	$\Delta\chi^2$	Δdf	p	ΔAIC
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	986	1361.27	1391.27	–	–	–	–
2		2	3	4	5	6	7	8	9	10	11	12	13	14	987	1361.28	1389.28	0.01	1	0.927	−1.99
3	1		3	4	5	6	7	8	9	10	11	12	13	14	987	1362.26	1390.26	1	1	0.318	−1
4	1	2		4	5	6	7	8	9	10	11	12	13	14	987	1361.32	1389.32	0.06	1	0.81	−1.94
5	1	2	3		5	6	7	8	9	10	11	12	13	14	987	1361.33	1389.33	0.06	1	0.807	−1.94
6	1	2	3	4		6	7	8	9	10	11	12	13	14	987	1361.27	1389.27	0	1	0.967	−2
7	1	2	3	4	5		7	8	9	10	11	12	13	14	987	1363.94	1391.94	2.68	1	0.102	0.68
8	1	2	3	4	5	6		8	9	10	11	12	13	14	987	1362.14	1390.14	0.88	1	0.349	−1.12
9	1	2	3	4	5	6	7		9	10	11	12	13	14	987	1361.38	1389.38	0.11	1	0.738	−1.89
10	1	2	3	4	5	6	7	8		10	11	12	13	14	987	1361.27	1389.27	0.01	1	0.93	−1.99
11	1	2	3	4	5	6	7	8	9		11	12	13	14	987	1361.29	1389.29	0.03	1	0.867	−1.97
12	1	2	3	4	5	6	7	8	9	10		12	13	14	987	1361.44	1389.44	0.17	1	0.679	−1.83
13	1	2	3	4	5	6	7	8	9	10	11		13	14	987	1363.85	1391.85	2.59	1	0.108	0.59
14	1	2	3	4	5	6	7	8	9	10	11	12		14	987	1361.32	1389.32	0.05	1	0.823	−1.95
15	1	2	3	4	5	6	7	8	9	10	11	12	13		987	1361.28	1389.28	0.02	1	0.897	−1.98
16		2	3	4		6	7	8	9	10	11	12	13	14	988	1361.28	1387.28	0.01	2	0.993	−3.99
17	1		3	4		6	7	8	9	10	11	12	13	14	988	1362.26	1388.26	1	2	0.607	−3
18	1	2		4		6	7	8	9	10	11	12	13	14	988	1361.32	1387.32	0.06	2	0.971	−3.94
19	1	2	3			6	7	8	9	10	11	12	13	14	988	1361.33	1387.33	0.06	2	0.969	−3.94
20	1	2	3	4			7	8	9	10	11	12	13	14	988	1364.49	1390.49	3.22	2	0.199	−0.78
21	1	2	3	4		6		8	9	10	11	12	13	14	988	1362.15	1388.15	0.88	2	0.643	−3.12
22	1	2	3	4		6	7		9	10	11	12	13	14	988	1361.38	1387.38	0.11	2	0.945	−3.89
23	1	2	3	4		6	7	8		10	11	12	13	14	988	1361.28	1387.28	0.01	2	0.995	−3.99
24	1	2	3	4		6	7	8	9		11	12	13	14	988	1361.3	1387.3	0.03	2	0.985	−3.97
25	1	2	3	4		6	7	8	9	10		12	13	14	988	1361.44	1387.44	0.17	2	0.917	−3.83
26	1	2	3	4		6	7	8	9	10	11		13	14	988	1363.85	1389.85	2.59	2	0.274	−1.41
27	1	2	3	4		6	7	8	9	10	11	12		14	988	1361.32	1387.32	0.05	2	0.974	−3.95
28	1	2	3	4		6	7	8	9	10	11	12	13		988	1361.28	1387.28	0.02	2	0.991	−3.98
29		2	3	4		6	7	8		10	11	12	13	14	989	1361.29	1385.29	0.02	3	0.999	−5.98
30	1		3	4		6	7	8		10	11	12	13	14	989	1362.27	1386.27	1.01	3	0.799	−4.99
31	1	2		4		6	7	8		10	11	12	13	14	989	1361.33	1385.33	0.07	3	0.995	−5.93
32	1	2	3			6	7	8		10	11	12	13	14	989	1361.34	1385.34	0.07	3	0.995	−5.93
33	1	2	3	4			7	8		10	11	12	13	14	989	1364.49	1388.49	3.22	3	0.358	−2.78
34	1	2	3	4		6		8		10	11	12	13	14	989	1365.57	1389.57	4.31	3	0.23	−1.69
35	1	2	3	4		6	7			10	11	12	13	14	989	1361.4	1385.4	0.13	3	0.988	−5.87

36	1	2	3	4	6	7	8		11	12	13	14	989	1361.3	1385.3	0.03	3	0.998	−5.97
37	1	2	3	4	6	7	8	10		12	13	14	989	1361.45	1385.45	0.18	3	0.98	−5.82
38	1	2	3	4	6	7	8	10	11		13	14	989	1363.92	1387.92	2.65	3	0.448	−3.35
39	1	2	3	4	6	7	8	10	11	12		14	989	1361.33	1385.33	0.06	3	0.996	−5.94
40	1	2	3	4	6	7	8	10	11	12	13		989	1361.29	1385.29	0.03	3	0.999	−5.97
41			3	4	6	7	8	10	11	12	13	14	990	1362.49	1384.49	1.23	4	0.874	−6.77
42		2		4	6	7	8	10	11	12	13	14	990	1361.34	1383.34	0.08	4	0.999	−7.92
43		2	3		6	7	8	10	11	12	13	14	990	1361.35	1383.35	0.08	4	0.999	−7.92
44		2	3	4		7	8	10	11	12	13	14	990	1364.5	1386.5	3.23	4	0.52	−4.77
45		2	3	4	6		8	10	11	12	13	14	990	1365.69	1387.69	4.43	4	0.351	−3.57
46		2	3	4	6	7		10	11	12	13	14	990	1361.41	1383.41	0.14	4	0.998	−7.86
47		2	3	4	6	7	8		11	12	13	14	990	1361.31	1383.31	0.05	4	1	−7.95
48		2	3	4	6	7	8	10		12	13	14	990	1361.46	1383.46	0.2	4	0.996	−7.8
49		2	3	4	6	7	8	10	11		13	14	990	1363.92	1385.92	2.66	4	0.617	−5.34
50		2	3	4	6	7	8	10	11	12		14	990	1361.34	1383.34	0.07	4	0.999	−7.93
51		2	3	4	6	7	8	10	11	12	13		990	1361.3	1383.3	0.04	4	1	−7.96
52			3	4	6	7	8	10	11	12	13		991	1362.52	1382.52	1.25	5	0.94	−8.75
53		2		4	6	7	8	10	11	12	13		991	1361.37	1381.37	0.11	5	1	−9.89
54		2	3		6	7	8	10	11	12	13		991	1361.42	1381.42	0.15	5	1	−9.85
55		2	3	4		7	8	10	11	12	13		991	1364.51	1384.51	3.24	5	0.663	−6.76
56		2	3	4	6		8	10	11	12	13		991	1365.7	1385.7	4.43	5	0.489	−5.57
57		2	3	4	6	7		10	11	12	13		991	1361.43	1381.43	0.16	5	0.999	−9.84
58		2	3	4	6	7	8		11	12	13		991	1361.33	1381.33	0.06	5	1	−9.94
59		2	3	4	6	7	8	10		12	13		991	1361.48	1381.48	0.21	5	0.999	−9.79
60		2	3	4	6	7	8	10	11		13		991	1363.93	1383.93	2.67	5	0.751	−7.33
61		2	3	4	6	7	8	10	11	12			991	1361.34	1381.34	0.08	5	1	−9.92
62			3	4	6	7	8		11	12	13		992	1362.55	1380.55	1.28	6	0.973	−10.72
63		2		4	6	7	8		11	12	13		992	1361.4	1379.4	0.13	6	1	−11.87
64		2	3		6	7	8		11	12	13		992	1361.44	1379.44	0.17	6	1	−11.83
65		2	3	4		7	8		11	12	13		992	1364.53	1382.53	3.26	6	0.775	−8.74
66		2	3	4	6		8		11	12	13		992	1365.73	1383.73	4.47	6	0.614	−7.53
67		2	3	4	6	7			11	12	13		992	1362.49	1380.49	1.22	6	0.976	−10.78
68		2	3	4	6	7	8			12	13		992	1361.5	1379.5	0.23	6	1	−11.77
69		2	3	4	6	7	8		11		13		992	1363.95	1381.95	2.68	6	0.848	−9.32
70		2	3	4	6	7	8		11	12			992	1361.37	1379.37	0.1	6	1	−11.9
71			3	4	6	7	8		11	12			993	1362.59	1378.59	1.32	7	0.988	−12.68
72		2		4	6	7	8		11	12			993	1361.55	1377.55	0.28	7	1	−13.72
73		2	3		6	7	8		11	12			993	1361.48	1377.48	0.21	7	1	−13.79
74		2	3	4		7	8		11	12			993	1364.56	1380.56	3.29	7	0.857	−10.71
75		2	3	4	6		8		11	12			993	1365.89	1381.89	4.63	7	0.705	−9.37
76		2	3	4	6	7			11	12			993	1362.54	1378.54	1.28	7	0.989	−12.72
77		2	3	4	6	7	8			12			993	1361.55	1377.55	0.28	7	1	−13.72

78	2	3	4	6	7	8	11		993	1364	1380	2.73	7	0.909	−11.27
79		3		6	7	8	11	12	994	1362.71	1376.71	1.44	8	0.994	−14.56
80	2			6	7	8	11	12	994	1361.56	1375.56	0.3	8	1	−15.7
81	2	3			7	8	11	12	994	1364.68	1378.68	3.41	8	0.906	−12.59
82	2	3		6		8	11	12	994	1365.95	1379.95	4.68	8	0.791	−11.32
83	2	3		6	7		11	12	994	1362.65	1376.65	1.39	8	0.994	−14.61
84	2	3		6	7	8		12	994	1361.66	1375.66	0.4	8	1	−15.6
85	2	3		6	7	8	11		994	1364.12	1378.12	2.85	8	0.943	−13.15
86				6	7	8	11	12	995	1362.76	1374.76	1.5	9	0.997	−16.5
87	2				7	8	11	12	995	1364.78	1376.78	3.51	9	0.941	−14.49
88	2			6		8	11	12	995	1366.39	1378.39	5.12	9	0.824	−12.88
89	2			6	7		11	12	995	1362.73	1374.73	1.47	9	0.997	−16.53
90	2			6	7	8		12	995	1361.74	1373.74	0.47	9	1	−17.53
91	2			6	7	8	11		995	1364.2	1376.2	2.94	9	0.967	−15.06
92				6	7	8		12	996	1362.92	1372.92	1.66	10	0.998	−18.34
93	2				7	8		12	996	1364.96	1374.96	3.69	10	0.96	−16.31
94	2			6		8		12	996	1366.44	1376.44	5.17	10	0.879	−14.83
95	2			6	7			12	996	1362.91	1372.91	1.64	10	0.998	−18.36
96	2			6	7	8			996	1364.38	1374.38	3.11	10	0.979	−16.89
97				6	7			12	997	1363.92	1371.92	2.65	11	0.995	−19.35
98	2				7			12	997	1366.22	1374.22	4.95	11	0.934	−17.05
99	2			6				12	997	1366.59	1374.59	5.33	11	0.914	−16.67
100	2			6	7				997	1365.38	1373.38	4.11	11	0.967	−17.89
101					7			12	998	1366.97	1372.97	5.7	12	0.93	−18.3
102				6				12	998	1367.15	1373.15	5.88	12	0.922	−18.12
103				**6**	**7**				**998**	**1366.41**	**1372.41**	**5.14**	**12**	**0.953**	**−18.86**
104					7				999	1369.13	1373.13	7.86	13	0.852	−18.14
105				6					999	1369.62	1373.62	8.35	13	0.82	−17.65
106	nullmodel								1000	1372.03	1374.03	10.76	14	0.704	−17.24

TIM3

	rs 919744 M474		rs 919747 M475		rs 1036199 M476		rs 8862717 M477		rs 30287616 M478		rs 431620 M479		rs 1363232 M480		rs 9007 M481							
	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	df	$\lvert D\rvert$	AIC	$\Delta\chi^2$	Δdf	p	ΔAIC
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1111	1531.83	1565.83	–	–	–	–
2		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1112	1532.4	1564.4	0.57	1	0.45	−1.43
3	1		3	4	5	6	7	8	9	10	11	12	13	14	15	16	1112	1533.97	1565.97	2.14	1	0.144	0.14
4	1	2		4	5	6	7	8	9	10	11	12	13	14	15	16	1112	1531.99	1563.99	0.16	1	0.687	−1.84
5	1	2	3		5	6	7	8	9	10	11	12	13	14	15	16	1112	1532.14	1564.14	0.31	1	0.579	−1.69
6	1	2	3	4		6	7	8	9	10	11	12	13	14	15	16	1112	1532.84	1564.84	1.01	1	0.315	−0.99
7	1	2	3	4	5		7	8	9	10	11	12	13	14	15	16	1112	1534.93	1566.93	3.09	1	0.079	1.09
8	1	2	3	4	5	6		8	9	10	11	12	13	14	15	16	1112	1533.13	1565.13	1.3	1	0.254	−0.7
9	1	2	3	4	5	6	7		9	10	11	12	13	14	15	16	1112	1531.84	1563.84	0	1	0.944	−2
10	1	2	3	4	5	6	7	8		10	11	12	13	14	15	16	1112	1531.83	1563.83	0	1	0.973	−2
11	1	2	3	4	5	6	7	8	9		11	12	13	14	15	16	1112	1533.88	1565.88	2.04	1	0.153	0.04
12	1	2	3	4	5	6	7	8	9	10		12	13	14	15	16	1112	1532.03	1564.03	0.2	1	0.658	−1.8
13	1	2	3	4	5	6	7	8	9	10	11		13	14	15	16	1112	1531.86	1563.86	0.02	1	0.879	−1.98
14	1	2	3	4	5	6	7	8	9	10	11	12		14	15	16	1112	1531.91	1563.91	0.08	1	0.783	−1.92
15	1	2	3	4	5	6	7	8	9	10	11	12	13		15	16	1112	1531.84	1563.84	0	1	0.947	−2
16	1	2	3	4	5	6	7	8	9	10	11	12	13	14		16	1112	1531.92	1563.92	0.08	1	0.774	−1.92
17	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		1112	1531.96	1563.96	0.13	1	0.722	−1.87
18		2	3	4	5	6	7	8		10	11	12	13	14	15	16	1113	1532.4	1562.4	0.57	2	0.752	−3.43
19	1		3	4	5	6	7	8		10	11	12	13	14	15	16	1113	1533.97	1563.97	2.14	2	0.344	−1.86
20	1	2		4	5	6	7	8		10	11	12	13	14	15	16	1113	1532.09	1562.09	0.25	2	0.881	−3.75
21	1	2	3		5	6	7	8		10	11	12	13	14	15	16	1113	1532.34	1562.34	0.5	2	0.777	−3.5
22	1	2	3	4		6	7	8		10	11	12	13	14	15	16	1113	1532.84	1562.84	1.01	2	0.603	−2.99
23	1	2	3	4	5		7	8		10	11	12	13	14	15	16	1113	1534.93	1564.93	3.1	2	0.212	−0.9
24	1	2	3	4	5	6		8		10	11	12	13	14	15	16	1113	1533.13	1563.13	1.3	2	0.522	−2.7
25	1	2	3	4	5	6	7			10	11	12	13	14	15	16	1113	1531.84	1561.84	0.01	2	0.997	−3.99
26	1	2	3	4	5	6	7	8			11	12	13	14	15	16	1113	1543.39	1573.39	11.56	2	0.003	7.56
27	1	2	3	4	5	6	7	8		10		12	13	14	15	16	1113	1532.05	1562.05	0.22	2	0.896	−3.78
28	1	2	3	4	5	6	7	8		10	11		13	14	15	16	1113	1531.86	1561.86	0.02	2	0.988	−3.98
29	1	2	3	4	5	6	7	8		10	11	12		14	15	16	1113	1531.91	1561.91	0.08	2	0.963	−3.92
30	1	2	3	4	5	6	7	8		10	11	12	13		15	16	1113	1531.84	1561.84	0.01	2	0.997	−3.99
31	1	2	3	4	5	6	7	8		10	11	12	13	14		16	1113	1531.92	1561.92	0.08	2	0.959	−3.92
32	1	2	3	4	5	6	7	8		10	11	12	13	14	15		1113	1531.96	1561.96	0.13	2	0.938	−3.87
33		2	3	4	5	6	7	8		10	11	12	13		15	16	1114	1532.41	1560.41	0.58	3	0.901	−5.42
34	1		3	4	5	6	7	8		10	11	12	13		15	16	1114	1534	1562	2.17	3	0.539	−3.83
35	1	2		4	5	6	7	8		10	11	12	13		15	16	1114	1532.09	1560.09	0.26	3	0.968	−5.74

36	1	2	3		5	6	7	8	10	11	12	13	15	16	1114	1532.34	1560.34	0.51	3	0.918	−5.49
37	1	2	3	4		6	7	8	10	11	12	13	15	16	1114	1532.85	1560.85	1.02	3	0.796	−4.98
38	1	2	3	4	5		7	8	10	11	12	13	15	16	1114	1534.95	1562.95	3.12	3	0.374	−2.88
39	1	2	3	4	5	6		8	10	11	12	13	15	16	1114	1533.16	1561.16	1.33	3	0.722	−4.67
40	1	2	3	4	5	6	7		10	11	12	13	15	16	1114	1531.84	1559.84	0.01	3	1	−5.99
41	1	2	3	4	5	6	7	8		11	12	13	15	16	1114	1543.39	1571.39	11.56	3	0.009	5.56
42	1	2	3	4	5	6	7	8	10		12	13	15	16	1114	1532.06	1560.06	0.22	3	0.974	−5.78
43	1	2	3	4	5	6	7	8	10	11		13	15	16	1114	1531.86	1559.86	0.03	3	0.999	−5.97
44	1	2	3	4	5	6	7	8	10	11	12		15	16	1114	1531.96	1559.96	0.13	3	0.988	−5.87
45	1	2	3	4	5	6	7	8	10	11	12	13		16	1114	1531.92	1559.92	0.09	3	0.993	−5.91
46	1	2	3	4	5	6	7	8	10	11	12	13	15		1114	1531.96	1559.96	0.13	3	0.988	−5.87
47		2	3	4	5	6	7		10	11	12	13	15	16	1115	1532.42	1558.42	0.59	4	0.964	−7.41
48	1		3	4	5	6	7		10	11	12	13	15	16	1115	1534.04	1560.04	2.21	4	0.697	−5.79
49	1	2		4	5	6	7		10	11	12	13	15	16	1115	1532.1	1558.1	0.26	4	0.992	−7.74
50	1	2	3		5	6	7		10	11	12	13	15	16	1115	1532.34	1558.34	0.51	4	0.972	−7.49
51	1	2	3	4		6	7		10	11	12	13	15	16	1115	1532.86	1558.86	1.03	4	0.905	−6.97
52	1	2	3	4	5		7		10	11	12	13	15	16	1115	1534.99	1560.99	3.15	4	0.532	−4.85
53	1	2	3	4	5	6			10	11	12	13	15	16	1115	1533.64	1559.64	1.81	4	0.77	−6.19
54	1	2	3	4	5	6	7			11	12	13	15	16	1115	1543.4	1569.4	11.56	4	0.021	3.56
55	1	2	3	4	5	6	7		10		12	13	15	16	1115	1532.06	1558.06	0.23	4	0.994	−7.77
56	1	2	3	4	5	6	7		10	11		13	15	16	1115	1531.87	1557.87	0.03	4	1	−7.97
57	1	2	3	4	5	6	7		10	11	12		15	16	1115	1531.97	1557.97	0.14	4	0.998	−7.86
58	1	2	3	4	5	6	7		10	11	12	13		16	1115	1531.93	1557.93	0.09	4	0.999	−7.91
59	1	2	3	4	5	6	7		10	11	12	13	15		1115	1531.97	1557.97	0.14	4	0.998	−7.86
60		2	3	4	5	6	7		10	11		13	15	16	1116	1532.42	1556.42	0.59	5	0.988	−9.41
61	1		3	4	5	6	7		10	11		13	15	16	1116	1534.11	1558.11	2.28	5	0.81	−7.72
62	1	2		4	5	6	7		10	11		13	15	16	1116	1532.11	1556.11	0.28	5	0.998	−9.72
63	1	2	3		5	6	7		10	11		13	15	16	1116	1532.37	1556.37	0.54	5	0.991	−9.46
64	1	2	3	4		6	7		10	11		13	15	16	1116	1532.88	1556.88	1.04	5	0.959	−8.96
65	1	2	3	4	5		7		10	11		13	15	16	1116	1535.02	1559.02	3.19	5	0.671	−6.81
66	1	2	3	4	5	6			10	11		13	15	16	1116	1533.68	1557.68	1.85	5	0.87	−8.15
67	1	2	3	4	5	6	7			11		13	15	16	1116	1543.47	1567.47	11.64	5	0.04	1.64
68	1	2	3	4	5	6	7		10			13	15	16	1116	1532.4	1556.4	0.56	5	0.99	−9.44
69	1	2	3	4	5	6	7		10	11			15	16	1116	1531.99	1555.99	0.16	5	0.999	−9.84
70	1	2	3	4	5	6	7		10	11		13		16	1116	1531.95	1555.95	0.12	5	1	−9.88
71	1	2	3	4	5	6	7		10	11		13	15		1116	1531.99	1555.99	0.16	5	0.999	−9.84
72		2	3	4	5	6	7		10	11		13		16	1117	1532.5	1554.5	0.67	6	0.995	−11.33
73	1		3	4	5	6	7		10	11		13		16	1117	1534.19	1556.19	2.36	6	0.884	−9.64
74	1	2		4	5	6	7		10	11		13		16	1117	1532.2	1554.2	0.36	6	0.999	−11.64
75	1	2	3		5	6	7		10	11		13		16	1117	1532.46	1554.46	0.62	6	0.996	−11.38
76	1	2	3	4		6	7		10	11		13		16	1117	1532.96	1554.96	1.12	6	0.98	−10.88
77	1	2	3	4	5		7		10	11		13		16	1117	1535.1	1557.1	3.27	6	0.774	−8.73

78	1	2	3	4	5	6			10	11	13	16	1117	1533.77	1555.77	1.94	6	0.925	−10.06
79	1	2	3	4	5	6	7			11	13	16	1117	1543.56	1565.56	11.73	6	0.068	−0.27
80	1	2	3	4	5	6	7		10		13	16	1117	1532.49	1554.49	0.65	6	0.995	−11.35
81	1	2	3	4	5	6	7		10	11		16	1117	1532.07	1554.07	0.23	6	1	−11.77
82	1	2	3	4	5	6	7		10	11	13		1117	1532	1554	0.17	6	1	−11.83
83		2	3	4	5	6	7		10	11	13		1118	1532.56	1552.56	0.72	7	0.998	−13.28
84	1		3	4	5	6	7		10	11	13		1118	1534.23	1554.23	2.4	7	0.934	−11.6
85	1	2		4	5	6	7		10	11	13		1118	1532.24	1552.24	0.41	7	1	−13.59
86	1	2	3		5	6	7		10	11	13		1118	1532.5	1552.5	0.67	7	0.999	−13.33
87	1	2	3	4		6	7		10	11	13		1118	1533.02	1553.02	1.19	7	0.991	−12.81
88	1	2	3	4	5		7		10	11	13		1118	1535.14	1555.14	3.3	7	0.856	−10.7
89	1	2	3	4	5	6			10	11	13		1118	1533.79	1553.79	1.96	7	0.962	−12.04
90	1	2	3	4	5	6	7			11	13		1118	1543.66	1563.66	11.82	7	0.106	−2.18
91	1	2	3	4	5	6	7		10		13		1118	1532.53	1552.53	0.7	7	0.998	−13.3
92	1	2	3	4	5	6	7		10	11			1118	1532.13	1552.13	0.3	7	1	−13.7
93		2	3	4	5	6	7		10	11			1119	1532.67	1550.67	0.84	8	0.999	−15.16
94	1		3	4	5	6	7		10	11			1119	1534.34	1552.34	2.51	8	0.961	−13.49
95	1	2		4	5	6	7		10	11			1119	1532.37	1550.37	0.54	8	1	−15.46
96	1	2	3		5	6	7		10	11			1119	1532.64	1550.64	0.81	8	0.999	−15.19
97	1	2	3	4		6	7		10	11			1119	1533.15	1551.15	1.31	8	0.995	−14.69
98	1	2	3	4	5		7		10	11			1119	1535.24	1553.24	3.4	8	0.907	−12.6
99	1	2	3	4	5	6			10	11			1119	1533.8	1551.8	1.97	8	0.982	−14.03
100	1	2	3	4	5	6	7			11			1119	1543.94	1561.94	12.11	8	0.146	−3.89
101	1	2	3	4	5	6	7		10				1119	1532.66	1550.66	0.83	8	0.999	−15.17
102		2		4	5	6	7		10	11			1120	1532.73	1548.73	0.9	9	1	−17.1
103	1			4	5	6	7		10	11			1120	1534.34	1550.34	2.51	9	0.981	−15.49
104	1	2			5	6	7		10	11			1120	1533.47	1549.47	1.63	9	0.996	−16.37
105	1	2		4		6	7		10	11			1120	1533.42	1549.42	1.59	9	0.996	−16.41
106	1	2		4	5		7		10	11			1120	1535.33	1551.33	3.5	9	0.941	−14.5
107	1	2		4	5	6			10	11			1120	1533.99	1549.99	2.15	9	0.989	−15.85
108	1	2		4	5	6	7			11			1120	1544.31	1560.31	12.48	9	0.188	−5.52
109	1	2		4	5	6	7		10				1120	1532.9	1548.9	1.07	9	0.999	−16.93
110				4	5	6	7		10	11			1121	1534.69	1548.69	2.86	10	0.985	−17.14
111		2			5	6	7		10	11			1121	1533.92	1547.92	2.09	10	0.996	−17.91
112		2		4		6	7		10	11			1121	1537.26	1551.26	5.43	10	0.861	−14.57
113		2		4	5		7		10	11			1121	1535.87	1549.87	4.04	10	0.946	−15.96
114		2		4	5	6			10	11			1121	1534.44	1548.44	2.61	10	0.989	−17.39
115		2		4	5	6	7			11			1121	1544.64	1558.64	12.81	10	0.234	−7.19
116		2		4	5	6	7		10				1121	1533.39	1547.39	1.56	10	0.999	−18.44
117				4	5	6	7		10				1122	1535.23	1547.23	3.4	11	0.984	−18.6
118		2			5	6	7		10				1122	1534.54	1546.54	2.7	11	0.994	−19.3
119		2		4		6	7		10				1122	1540.18	1552.18	8.35	11	0.682	−13.65

120	2	4	5		7	10	1122	1536.35	1548.35	4.52	11	0.952	−17.48
121	2	4	5	6		10	1122	1535.12	1547.12	3.28	11	0.986	−18.72
122	2	4	5	6	7		1122	1545.33	1557.33	13.5	11	0.262	−8.5
123			5	6	7	10	1123	1538.44	1548.44	6.61	12	0.883	−17.39
124	2			6	7	10	1123	1541.65	1551.65	9.82	12	0.632	−14.18
125	2		5		7	10	1123	1536.87	1546.87	5.04	12	0.957	−18.96
126	2		5	6		10	1123	1536.15	1546.15	4.31	12	0.977	−19.69
127	2		5	6	7		1123	1545.35	1555.35	13.52	12	0.332	−10.48
128			5	6		10	1124	1540.1	1548.1	8.26	13	0.826	−17.74
129	2			6		10	1124	1541.65	1549.65	9.82	13	0.709	−16.18
130	2		5			10	1124	1538.53	1546.53	6.7	13	0.917	−19.3
131	2		5	6			1124	1547.39	1555.39	15.56	13	0.274	−10.44
132			**5**			**10**	**1125**	**1540.85**	**1546.85**	**9.02**	**14**	**0.83**	**−18.98**
133	2					10	1125	1543.09	1549.09	11.26	14	0.666	−16.74
134	2		5				1125	1550.96	1556.96	19.13	14	0.16	−8.87
135						10	1126	1549.07	1553.07	17.24	15	0.305	−12.76
136			5				1126	1555.49	1559.49	23.66	15	0.071	−6.34
137	nullmodel						1127	1555.56	1557.56	23.73	16	0.096	−8.27

Appendix F

MODELLING INTERACTION WITHIN GENES

IL7R

IL7R. No significant interaction between the components of 3 5 6 was observed.

Model			df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
A207	F210		1107	1512.66	1520.66	–	–	–	–
A207	F210	A207*F210	1105	1511.73	1523.73	2	0.94	0.626	3.06

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.37	0.09	4.09	0.00
A207	0.39	0.10	3.71	0.00
A210	0.55	0.15	3.74	0.00
D210	−0.48	0.17	−2.94	0.00

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.20	81.19	0.04	0.97
A207	3.19	81.19	0.04	0.97
A210	6.26	162.37	0.04	0.97
D210	−6.22	162.37	−0.04	0.97
A207*A210	5.69	162.37	0.04	0.97
A207*D210	−5.84	162.37	−0.04	0.97

LAG3

LAG3. No significant interaction between the components of 5 6 7 was observed.

Model			df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
A298	F297		1107	1531.82	1539.82	–	–	–	–
A298	F297	A298*F297	1105	1530.66	1542.66	2	1.16	0.56	2.84

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.09	0.07	1.24	0.22
A298	−0.15	0.09	−1.71	0.09
A297	0.04	0.10	0.44	0.66
D297	0.10	0.14	0.70	0.48

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.08	0.07	1.20	0.23
A298	−0.12	0.10	−1.24	0.21
A297	0.04	0.10	0.37	0.71
D297	0.08	0.14	0.59	0.56
A298*A297	−0.13	0.14	−0.89	0.38
A298*D297	−0.18	0.19	−0.92	0.36

TIM3

TIM3. If the two-locus interaction effect of 3 9 10 were added to the core model a significant increase in fit was observed.

Model			df	$\|D\|$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
A475	F478		1107	1514.91	1522.91	–	–	–	–
A475	F478	A475*F478	1106	1507.82	1517.82	1	7.09	0.008	−5.09

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.05	0.11	0.49	0.62
A475	1.15	0.37	3.10	0.00
A478	0.79	0.41	1.94	0.05
D478	0.60	0.22	2.76	0.01

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−1.04	0.43	−2.43	0.01
A475	−0.44	0.68	−0.65	0.52
A478	−0.81	0.71	−1.15	0.25
D478	2.83	0.88	3.23	0.00
A475*A478	−2.22	0.84	−2.63	0.01

Appendix G

MODELLING INTERACTION BETWEEN GENES

IL7R-LAG3

	A207 * A297	A207 * D297	A207 * A298	A210 * A297	A210 * D297	A210 * A298	D210 * A297	D210 * D297	D210 * A298	df	$\lvert D\rvert$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
✓										1104	1509.09	1523.09	–	–	–	–
✓			✓							1103	1507.9	1523.9	1	1.19	0.275	0.81
✓	✓									1103	1506.16	1522.16	1	2.93	0.087	−0.93
✓				✓						1103	1503.27	1519.27	1	5.82	0.016	−3.82
✓						✓				1103	1507.8	1523.8	1	1.29	0.256	0.71
✓	✓	✓								1102	1506.16	1524.16	2	2.93	0.231	1.07
✓				✓	✓					1102	1503.18	1521.18	2	5.91	0.052	−1.91
✓				✓			✓			1102	1501.22	1519.22	2	7.88	0.019	−3.88
✓						✓			✓	1102	1507.63	1525.63	2	1.46	0.482	2.54
✓				✓	✓		✓	✓		1100	1501.16	1523.16	4	7.94	0.094	0.06
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1095	1499.41	1531.41	9	9.69	0.377	8.31

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.33	0.10	3.41	0.00
A207	0.42	0.11	3.86	0.00
A210	0.55	0.15	3.76	0.00
D210	−0.48	0.17	−2.93	0.00
A297	0.03	0.10	0.28	0.78
D297	0.10	0.14	0.76	0.45
A298	−0.15	0.09	−1.66	0.10
A207*A298	0.15	0.14	1.09	0.28

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.34	0.10	3.51	0.00
A207	0.47	0.12	4.06	0.00
A210	0.55	0.15	3.75	0.00
D210	−0.49	0.17	−2.94	0.00
A297	0.01	0.11	0.06	0.95
D297	0.09	0.14	0.62	0.53
A298	−0.16	0.09	−1.76	0.08
A207*A297	−0.24	0.14	−1.71	0.09

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.28	0.10	2.80	0.01
A207	0.39	0.10	3.71	0.00
A210	0.44	0.15	2.87	0.00
D210	−0.50	0.17	−2.99	0.00
A297	0.22	0.13	1.65	0.10
D297	0.10	0.14	0.71	0.48
A298	−0.15	0.09	−1.69	0.09
A210*A297	0.37	0.15	2.40	0.02

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.31	0.10	3.14	0.00
A207	0.38	0.10	3.68	0.00
A210	0.51	0.15	3.38	0.00
D210	−0.49	0.17	−2.95	0.00
A297	0.03	0.10	0.25	0.80
D297	0.10	0.14	0.75	0.45
A298	−0.24	0.12	−2.06	0.04
A210*A298	−0.17	0.15	−1.13	0.26

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.34	0.10	3.51	0.00
A207	0.47	0.12	4.00	0.00
A210	0.55	0.15	3.75	0.00
D210	−0.49	0.17	−2.94	0.00
A297	0.01	0.11	0.05	0.96
D297	0.09	0.14	0.61	0.54
A298	−0.16	0.09	−1.76	0.08
A207*A297	−0.24	0.16	−1.49	0.14
A207*D297	−0.01	0.21	−0.04	0.97

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.28	0.10	2.82	0.01
A207	0.39	0.10	3.70	0.00
A210	0.45	0.15	2.89	0.00
D210	−0.50	0.17	−2.99	0.00
A297	0.21	0.14	1.53	0.13
D297	0.07	0.17	0.37	0.71
A298	−0.15	0.09	−1.68	0.09
A210*A297	0.34	0.17	2.06	0.04
A210*D297	−0.07	0.22	−0.30	0.77

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.29	0.10	2.87	0.00
A207	0.39	0.11	3.70	0.00
A210	0.49	0.16	3.08	0.00
D210	−0.60	0.18	−3.30	0.00
A297	0.17	0.14	1.26	0.21
D297	0.10	0.14	0.75	0.45
A298	−0.15	0.09	−1.69	0.09
A210*A297	0.18	0.20	0.91	0.36
D210*A297	0.36	0.25	1.43	0.15

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.29	0.10	2.88	0.00
A207	0.39	0.11	3.69	0.00
A210	0.49	0.16	3.08	0.00
D210	−0.60	0.18	−3.25	0.00
A297	0.16	0.14	1.16	0.25
D297	0.08	0.18	0.44	0.66
A298	−0.15	0.09	−1.68	0.09
A210*A297	0.17	0.22	0.77	0.44
D210*A297	0.35	0.28	1.25	0.21
A210*D297	−0.04	0.29	−0.14	0.89
D210*D297	−0.02	0.37	−0.06	0.95

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.31	0.10	3.12	0.00
A207	0.38	0.10	3.69	0.00
A210	0.50	0.15	3.28	0.00
D210	−0.47	0.17	−2.71	0.01
A297	0.03	0.10	0.24	0.81
D297	0.10	0.14	0.74	0.46
A298	−0.26	0.13	−2.07	0.04
A210*A298	−0.23	0.20	−1.11	0.27
D210*A298	0.10	0.25	0.41	0.68

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.30	0.11	2.86	0.00
A207	0.45	0.13	3.51	0.00
A210	0.50	0.17	2.96	0.00
D210	−0.58	0.19	−3.09	0.00
A297	0.10	0.16	0.64	0.52
D297	0.07	0.21	0.33	0.74
A298	−0.21	0.14	−1.53	0.13
A207*A297	−0.12	0.19	−0.63	0.53
A207*D297	−0.05	0.24	−0.19	0.85
A207*A298	0.08	0.16	0.53	0.60
A210*A297	0.06	0.25	0.24	0.81
A210*D297	−0.06	0.33	−0.18	0.86
A210*A298	−0.16	0.23	−0.68	0.50
D210*A297	0.40	0.29	1.39	0.17
D210*D297	−0.02	0.37	−0.07	0.95
D210*A298	0.15	0.26	0.59	0.56

IL7R-TIM3

core model	A207 * A475	A207 * A478	A207 * D478	A210 * A475	A210 * A478	A210 * D478	D210 * A475	D210 * A478	D210 * D478	df	$\|D\|$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
✓										1104	1492.96	1506.96	–	–	–	–
✓	✓									1103	1492.75	1508.75	1	0.21	0.647	1.79
✓		✓								1103	1491.57	1507.57	1	1.38	0.24	0.62
✓				✓						1103	1492.94	1508.94	1	0.01	0.906	1.99
✓					✓					1103	1492.95	1508.95	1	0.01	0.925	1.99
✓		✓	✓							1102	1491.2	1509.2	2	1.76	0.415	2.24
✓				✓			✓			1102	1490.4	1508.4	2	2.55	0.279	1.45
✓					✓	✓				1102	1492.33	1510.33	2	0.63	0.731	3.37
✓					✓			✓		1102	1489.96	1507.96	2	2.99	0.224	1.01
✓					✓	✓		✓	✓	1101	1487.9	1507.9	3	5.06	0.168	0.94
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1097	1482.28	1510.28	7	10.68	0.153	3.32

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.29	0.13	2.23	0.03
A207	0.44	0.15	2.96	0.00
A210	0.53	0.15	3.59	0.00
D210	−0.45	0.17	−2.72	0.01
A475	1.12	0.37	2.98	0.00
A478	0.76	0.41	1.85	0.07
D478	0.61	0.22	2.76	0.01
A207*A475	−0.08	0.17	−0.46	0.65

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.28	0.13	2.18	0.03
A207	0.53	0.16	3.36	0.00
A210	0.53	0.15	3.58	0.00
D210	−0.45	0.17	−2.69	0.01
A475	1.12	0.37	3.00	0.00
A478	0.75	0.41	1.82	0.07
D478	0.63	0.22	2.81	0.01
A207*A478	0.20	0.17	1.17	0.24

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.30	0.15	1.99	0.05
A207	0.40	0.11	3.75	0.00
A210	0.55	0.21	2.61	0.01
D210	−0.46	0.17	−2.76	0.01
A475	1.11	0.39	2.83	0.01
A478	0.77	0.41	1.88	0.06
D478	0.60	0.22	2.73	0.01
A210*A475	−0.02	0.20	−0.12	0.91

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.28	0.15	1.87	0.06
A207	0.39	0.11	3.75	0.00
A210	0.52	0.21	2.45	0.01
D210	−0.46	0.17	−2.75	0.01
A475	1.12	0.37	3.01	0.00
A478	0.76	0.43	1.79	0.07
D478	0.60	0.22	2.74	0.01
A210*A478	−0.02	0.20	−0.09	0.93

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.29	0.13	2.27	0.02
A207	0.49	0.17	2.91	0.00
A210	0.53	0.15	3.61	0.00
D210	−0.45	0.17	−2.71	0.01
A475	1.12	0.37	2.99	0.00
A478	0.76	0.41	1.86	0.06
D478	0.62	0.22	2.80	0.01
A207*A478	0.07	0.27	0.28	0.78
A207*D478	0.20	0.32	0.62	0.54

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.34	0.16	2.07	0.04
A207	0.40	0.11	3.77	0.00
A210	0.62	0.25	2.52	0.01
D210	−0.47	0.17	−2.81	0.01
A475	1.12	0.37	3.01	0.00
A478	0.89	0.45	1.97	0.05
D478	0.43	0.31	1.40	0.16
A210*A478	0.22	0.37	0.61	0.55
A210*D478	−0.33	0.42	−0.79	0.43

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.58	0.21	2.72	0.01
A207	0.39	0.11	3.74	0.00
A210	1.15	0.38	3.01	0.00
D210	−1.11	0.37	−2.96	0.00
A475	1.15	0.37	3.08	0.00
A478	1.30	0.50	2.62	0.01
D478	0.20	0.33	0.61	0.54
A210*A478	1.01	0.54	1.87	0.06
A210*D478	−0.64	0.44	−1.44	0.15
D210*A478	−0.84	0.42	−2.01	0.04

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.45	0.19	2.31	0.02
A207	0.39	0.11	3.71	0.00
A210	0.88	0.33	2.68	0.01
D210	−0.92	0.36	−2.57	0.01
A475	0.94	0.41	2.27	0.02
A478	0.79	0.41	1.92	0.06
D478	0.56	0.22	2.50	0.01
A210*A475	−0.44	0.35	−1.23	0.22
D210*A475	0.61	0.40	1.52	0.13

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.44	0.19	2.27	0.02
A207	0.39	0.11	3.70	0.00
A210	0.87	0.33	2.65	0.01
D210	−0.96	0.36	−2.66	0.01
A475	1.14	0.37	3.06	0.00
A478	0.98	0.45	2.18	0.03
D478	0.56	0.22	2.50	0.01
A210*A478	0.42	0.36	1.20	0.23
D210*A478	−0.66	0.40	−1.65	0.10

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.67	0.23	2.95	0.00
A207	0.62	0.21	2.97	0.00
A210	1.33	0.42	3.17	0.00
D210	−1.12	0.38	−2.97	0.00
A475	0.90	0.64	1.40	0.16
A478	1.16	0.68	1.70	0.09
D478	0.20	0.36	0.54	0.59
A207*A475	0.93	0.56	1.67	0.09
A207*A478	1.21	0.65	1.85	0.07
A207*D478	0.01	0.42	0.02	0.99
A210*A475	−0.83	0.87	−0.95	0.34
A210*A478	0.44	0.93	0.47	0.64
A210*D478	−0.71	0.57	−1.24	0.21
D210*A475	0.84	0.42	1.99	0.05

LAG3-TIM3

core model	A297 * A475	A297 * A478	A297 * D478	D297 * A475	D297 * A478	D297 * D478	A298 * A475	A298 * A478	A298 * D478	df	$\|D\|$	AIC	Δdf	$\Delta\chi^2$	p	ΔAIC
✓										1104	1510.93	1524.93	–	–	–	–
✓	✓									1103	1509.93	1525.93	1	1	0.317	1
✓		✓								1103	1509.83	1525.83	1	1.11	0.292	0.89
✓							✓			1103	1510.4	1526.4	1	0.54	0.464	1.46
✓								✓		1103	1510.75	1526.75	1	0.18	0.669	1.82
✓								✓	✓	1102	1510.73	1528.73	2	0.2	0.903	3.8
✓	✓			✓						1102	1506.29	1524.29	2	4.64	0.098	−0.64
✓		✓	✓							1102	1509.82	1527.82	2	1.11	0.574	2.89
✓		✓			✓					1102	1502.51	1520.51	2	8.43	0.015	−4.43
✓		✓	✓		✓	✓				1100	1501.74	1523.74	4	9.19	0.056	−1.19
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1095	1493	1525	9	17.93	0.036	0.07

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.06	0.13	0.47	0.64
A297	−0.06	0.16	−0.40	0.69
D297	0.10	0.14	0.71	0.47
A298	−0.16	0.09	−1.81	0.07
A475	1.09	0.38	2.90	0.00
A478	0.81	0.41	1.98	0.05
D478	0.59	0.22	2.70	0.01
A297*A475	0.18	0.18	1.00	0.32

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.07	0.13	0.51	0.61
A297	−0.07	0.16	−0.46	0.64
D297	0.10	0.14	0.71	0.48
A298	−0.16	0.09	−1.81	0.07
A475	1.13	0.37	3.04	0.00
A478	0.86	0.42	2.07	0.04
D478	0.59	0.22	2.68	0.01
A297*A478	−0.19	0.18	−1.05	0.29

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.01	0.12	−0.11	0.91
A297	0.05	0.10	0.52	0.61
D297	0.10	0.14	0.75	0.46
A298	−0.24	0.14	−1.70	0.09
A475	1.16	0.37	3.11	0.00
A478	0.77	0.41	1.86	0.06
D478	0.64	0.22	2.89	0.00
A298*A475	0.12	0.17	0.73	0.47

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.01	0.12	−0.05	0.96
A297	0.05	0.10	0.52	0.61
D297	0.10	0.14	0.73	0.46
A298	−0.21	0.14	−1.44	0.15
A475	1.14	0.37	3.07	0.00
A478	0.76	0.41	1.83	0.07
D478	0.63	0.22	2.87	0.00
A298*A478	−0.07	0.17	−0.43	0.67

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.00	0.12	−0.01	0.99
A297	0.05	0.10	0.51	0.61
D297	0.10	0.14	0.73	0.46
A298	−0.20	0.16	−1.26	0.21
A475	1.14	0.37	3.07	0.00
A478	0.76	0.42	1.83	0.07
D478	0.62	0.23	2.65	0.01
A298*A478	−0.04	0.27	−0.16	0.87
A298*D478	−0.05	0.32	−0.14	0.89

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.06	0.14	0.44	0.66
A297	−0.07	0.19	−0.36	0.72
D297	0.10	0.14	0.71	0.48
A298	−0.16	0.09	−1.81	0.07
A475	1.13	0.37	3.04	0.00
A478	0.86	0.45	1.93	0.05
D478	0.60	0.29	2.09	0.04
A297*A478	−0.18	0.33	−0.55	0.58
A297*D478	−0.01	0.37	−0.03	0.97

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.83	20.30	0.04	0.97
A297	−1.73	40.59	−0.04	0.97
D297	−1.83	40.59	−0.04	0.96
A298	−0.16	0.09	−1.76	0.08
A475	1.13	0.37	3.03	0.00
A478	2.39	40.59	0.06	0.95
D478	−0.90	40.59	−0.02	0.98
A297*A478	−3.41	81.19	−0.04	0.97
A297*D478	2.96	81.19	0.04	0.97
D297*A478	−3.56	81.19	−0.04	0.96
D297*D478	2.81	81.19	0.04	0.97

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.09	0.13	0.69	0.49
A297	−0.20	0.17	−1.15	0.25
D297	−0.25	0.23	−1.08	0.28
A298	−0.16	0.09	−1.76	0.08
A475	1.06	0.38	2.81	0.01
A478	0.84	0.41	2.03	0.04
D478	0.57	0.22	2.60	0.01
A297*A475	0.38	0.21	1.83	0.07
D297*A475	0.52	0.27	1.90	0.06

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.12	0.14	0.85	0.39
A297	−0.30	0.19	−1.56	0.12
D297	−0.43	0.24	−1.75	0.08
A298	−0.16	0.09	−1.77	0.08
A475	1.13	0.37	3.02	0.00
A478	0.94	0.42	2.22	0.03
D478	0.57	0.22	2.53	0.01
A297*A478	−0.50	0.22	−2.25	0.02
D297*A478	−0.76	0.28	−2.66	0.01

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.96	55.17	0.02	0.99
A297	−1.99	110.34	−0.02	0.99
D297	−2.07	110.34	−0.02	0.98
A298	−0.22	0.16	−1.34	0.18
A475	4.53	85.11	0.05	0.96
A478	6.03	139.35	0.04	0.97
D478	−1.15	110.34	−0.01	0.99
A297*A475	−6.08	170.22	−0.04	0.97
A297*A478	−9.99	278.71	−0.04	0.97
A297*D478	3.47	220.69	0.02	0.99
D297*A475	−8.04	170.22	−0.05	0.96
D297*A478	−12.08	278.71	−0.04	0.96
D297*D478	3.39	220.69	0.01	0.99
A298*A475	1.50	0.83	1.80	0.07
A298*A478	1.42	0.87	1.63	0.10
A298*D478	−0.05	0.33	−0.16	0.87

Appendix H

GENETIC GLOSSARY

allele One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type.

chromosome One of the threadlike "packages" of genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers.

diploid cell The number of chromosomes in most cells except the gametes. In humans, the diploid number is 46.

dominant A gene that almost always results in a specific physical characteristic, for example, a disease, even though the patient's genome possesses only one copy.

exon A sequence of DNA that codes information for protein synthesis that is transcribed to messenger RNA

gamete A reproductive cell having the haploid number of chromosomes

gene A gene refers to a segment of DNA found on a chromosome that codes for a particular protein.

genome All the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in mitochondria.

genotype The particular alleles at specified loci present in an organism.

haploid cell The number of chromosomes in a sperm or egg cell, half the diploid number.

haplotype The set, made up of one allele of each gene, comprising the genotype. Also used to refer to the set of alleles on one chromosome or a part of a chromosome.

heterozygosity The presence of different alleles at one or more loci on homologous chromosomes

homologous chromosomes Chromosome containing the same linear gene sequences as another, each derived from one parent.

homozygosity The presence of identical alleles at one or more loci on homologous chromosomes

intron A segment of a gene situated between exons that is removed before translation of messenger RNA and does not function in coding for protein synthesis

liability Susceptibility, vulnerability, predisposition

locus The specific site of a particular gene on its chromosome, alt. the specific site of a SNP on its chromosome.

meiosis Two successive nuclear divisions that produce haploid gametes

microsatellite Short segment of DNA that consist of repeated sequences of usually two to five nucleotides

mitosis The nuclear division producing two daughter nuclei identical to the original nucleus

mutation A permanent structural alteration in DNA

phenotype The observable traits or characteristics of an organism, for example hair color, weight, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic.

recessive A genetic disorder that appears only in patients who have received two copies of a mutant gene, one from each parent.

SNP Single Nucleotide Polymorphism

somatic cell Any of the cells of the body that compose the tissues, organs, and parts of that individual other than the germ cells

trait A genetically determined characteristic or condition of an individual; same as phenotype.

Appendix I

STATISTICAL GLOSSARY

additive A situation in which the best estimate of a dependent variable is obtained by simply adding together the appropriately computed effects of each of the independent variables. Additivity implies the absence of interactions.

concordance rate Probability that both twins will develop a disorder if one twin has the disorder.

confounding When the differences between the treatment and control groups other than the treatment produce differences in response that are not distinguishable from the effect of the treatment, those differences between the groups are said to be confounded with the effect of the treatment (if any).

event An event is a subset of outcome space. An event determined by a random variable is an event of the form A=(X is in A). When the random variable X is observed, that determines whether or not A occurs: if the value of X happens to be in A, A occurs; if not, A does not occur.

dichotomous Divided or dividing into two sharply distinguished parts or classifications.

heteroscedasticity The absence of homogeneity of variance

incidence The rate of occurrence of new cases of a particular disease in a population being studied.

interaction A situation in which the direction and/or magnitude of the relationship between two variables depends on (i.e., differs according to) the value of one or more other variables. When interaction is present, simple additive techniques are inappropriate; hence, interaction is sometimes thought of as the absence of additivity. Synonyms: nonadditivity, conditioning effect, moderating effect, contingency effect.

likelihood The chance that something will happen.

nominal scale A classification of cases which defines their equivalence and nonequivalence, but implies no quantitative relationships or ordering among them

ordinal scale A classification of cases into a set of ordered classes such that each case is considered equal to, greater than, or less than every other case.

parsimonious In statistical genetics, parsimonious refers to the smallest subset of loci that sustains all or most of the information content of a full set. Information refers to variable data in a genetic study.

power The probability of rejecting a false statistical null hypothesis

prevalence The percentage of a population that is affected with a particular disease at a given time. (compare incidence)

probability The probability of an event is a number between zero and 100% (Kolmogorow's axioms)

p-value The smallest significance level p for which a test rejects the null hypothesis.

random variable A variable whose values are random but whose statistical distribution is known.

quantitative 1. A resolution of anything into its constituent or original elements. 2. How much of an element that is present.

significance level The chance that the test erroneously rejects the null hypothesis when the null hypothesis is true.

type I error Occurs when the null hypothesis is rejected erroneously when it is in fact true.

type II error Occurs if the null hypothesis is not rejected when it is in fact false.