THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Statistical Methods for Genome Wide Association Studies

MALIN ÖSTENSSON

CHALMERS | GÖTEBORGS UNIVERSITET

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2012

Statistical Methods for Genome Wide Association Studies
Malin Östensson
ISBN 978-91-7385-742-0

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny Serie Nr 2012:3423
ISSN 0346-718X

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Typeset with LaTeX.
Printed in Göteborg, Sweden, 2012

Statistical Methods for Genome Wide Association Studies

Malin Östensson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis focus on various statistical methods for analyzing Genome Wide Association data. The thesis include four papers, three of them considers the analysis of complex traits, and the last one a method for analyzing mendelian traits.

Although GWAS have identified many associated regions in the genome for many complex diseases, there is still much of the genetic heritability that remains unexplained. The power of detecting new genetic risk variants can be improved by considering several genes in the same model.

A genetic variant in the HLA region on chromosome 6 is necessary but not sufficient to develop Celiac Disease. In the first two papers we utilize this information to discover additional genetic variants. In Paper I this is done by a method which use the 'Cochran Armitage trend test', to find a trend in allele frequencies. Simulations are used to evaluate the power of this test compared with the commonly used Pearson 1 df chisquare test and the test is then applied to a previously published Celiac Disease case-control material.

In paper II the HLA information is utilized by a stratified TDT, conditioning on the HLA variants. In addition, an imputation-based version of the TDT is presented, as well as a likelihood ratio test searching for two-locus interactions by comparing the heterogeneity and epistasis models. Here the candidates for interaction analysis are chosen by a two-step approach, combining the results from the TDT and prior information from previous studies.

In contrast to the approach used in paper II for identifying interactions between genes, in paper 3 we instead consider the method of performing a full Genome Wide Interaction Analysis. By examining how commonly we will find interactions without marginal effects in a GWIA we discuss what conclusions can be drawn from such findings.

In the final paper we develop a program locating a region containing a causal gene for rare monogenic traits. This program can be used in large pedigrees with multiple affected cases, and discerns the causal region by coloring them according to how common they are in the population.

Keywords: Genome Wide Association Studies, gene-gene interactions, Genotype imputation, allele sharing, haplotype sharing, Single Nucleotide Polymorphism, Celiac Disease

List of Papers

This thesis include the following papers

- ▷ **Östensson M**, Naluai A.T, van Heel D., Nilsson S.,
  "Utilizing known risk genes within Celiac Disease"
- ▷ **Östensson M**, Montén C., Bacelis J., Gudjonsdottir A., Adamovic S., Ek J., Ascher H., Pollak E., Fischler B., Arnell H., Browaldh L., Agardh D., Wahlström J., Nilsson S., Naluai AT,
  "A genome-wide linkage and association analysis in celiac disease families identifies genetic variants within DUSP10 and implicates genes within protein and energy homeostasis"
- ▷ **Östensson M**,
  "Are two-locus interactions without marginal effects in Genome Wide Association Studies really that interesting?"
- ▷ **Östensson M**, Martinsson, T.,
  "The Color Method – a simplified tool for locating risk regions with GWA data in mendelian disorders"

”Att livet överhuvudtaget uppstod på den här planeten
är statistiskt sett så orimligt att vi sannolikt inte finns.”

ur ’Vips så blev det liv’, Bob Hansson

Acknowledgments

I would like to express my greatest appreciation to my supervisor Staffan Nilsson for his great guidance and support during these years and my work with this thesis. I am also grateful to my co-supervisor Marina Axelson-Fisk for your support.

To my co-authors, particularly Åsa Torinsson Naluai, I would like to thank you for good collaboration. It has been a true pleasure working, discussing research and enjoying these years with all of you.

I would also like to thank all my colleagues at the department of Mathematical Science and elsewhere. Thank you for contributing to my understanding of mathematics and its applications, and thank you for your friendship!

To all my friends outside the department, thank you for your great friendship, love and encouragement! You have given me the support and pauses I have needed in order to keep this going.

To the Sadhana Crew and other yogi/yogini friends - thank you for supporting me in substituting the the C in locus ;) This has been amazing and helped me enormously. Namaste.

Finally, to my dear family. I love you! Thank you for your never ending love and support!

Malin Östensson
Göteborg, September 2012

Contents

Chapter 1

Introduction

Genetic association studies aim to identify genetic variants that vary between individuals with different disease status (affected/unaffected). In this chapter the genetic background to the subject is presented, explaining concepts and properties which are important for making inference from such studies.

1.1 Background

The DNA is built up by different arrangements of the four nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T). The DNA molecule has the shape of a double helix where each nucleotide pairs up with its complementary nucleotide - A binds to T and C binds to G, and the DNA is tightly packed into chromosomes. The human genome consist of 23 pairs of chromosomes, 22 pairs of autosomes - chromosomes which are present in two copies in both males and females - and one pair of sex chromosomes. Females have two X chromosomes and males have one X and one Y chromosome. In each pair of chromosomes, one of the chromosomes is inherited from the mother and the other from the father. A *gene* is a segment of DNA that provides coded instructions for synthesis of RNA, which when it is transcribed into protein contributes to the expression of a hereditary character. Diploid organisms (like humans) have two copies of each gene - one on each of the two *homologous* chromosomes (of the same type) - which they inherit from their parents. Each gene occupy a certain position (*locus* plur. loci) on the chromosome, and the parent randomly pass on one of the two *alleles* (defined below) of each gene to its offspring with probability $1/2$. The distance between two loci can be measured in base pairs (bp, also kb=1000 bp and Mb=10^6 bp), which corresponds to the number of nucleotides there are between these loci, this distance measure is referred to as *physical distance*.

The allele is the unit containing genetic information at a certain locus on the parental chromosome. Mutation of an allele will change its form and create a new mutated variant of the gene, causing genetic variation between individuals at that locus. *Allele frequencies* p_A and p_a are used to denote the relative population frequency of the alleles A and a at the locus. At those loci in the genome which include population variation, there are several possible *genotypes* - combinations of alleles on the same locus of two homologous chromosomes. A genotype is *heterozygous* if the two alleles are different, and *homozygous* if they are equal. Under the assumption of random mating, absence of disturbances like migration, selection and mutation at the gene in question, the population is said to be in *Hardy Weinberg Equilibrium* (HWE) meaning that the genotype frequencies only depend of the allele frequencies. This implies that the frequency of a homozygous genotype AA is p_A^2, and for the heterozygous Aa the frequency is $2p_Ap_a$. A *phenotype* is the physical expression of a genotype, e.g. an individual's eye colour.

1.1.1 Genetic models

Genetic models describe the relation between an individual's genotype(s) and some specific trait. A parameter which is often used to describe genetic models for (binary) traits is *penetrance*, the probability of a particular phenotype F for a given genotype G_i,

$$f_i = P(F|G_i).$$

There are several genetic models. A *mendelian trait* is determined by one gene, where a mutation in the gene cause the trait. There are two types of purely mendelian traits; for a *dominant* trait it is enough for one of the two alleles at the loci to be of the susceptible type for the trait to be expressed in the organism, and for a *recessive* trait both of the alleles need to be of this type. For completely dominant and recessive traits penetrances are either 0 or 1. There are many traits which follows *incomplete penetrance models*, where some of the penetrance parameters are below 1, hence the trait is expressed in some, but not all, of the individuals with that genotype. Other models include *phenocopies* where some individuals have a trait induced by environmental factors, resembling the phenotype which is usually caused by a specific genotype.

There are also many *non-mendelian traits* such as polygenic or complex traits and sex-linked traits. *Complex traits* which are the subject of some of the papers in this thesis, are traits that do not follow a classic mendelian inheritance pattern, where typically several genes and environmental factors are involved. Here a positive penetrance ($f > 0$) for subjects who do not carry the risk allele at one risk locus can be explained by environmental factors, risk variants at other loci, possibly *heterogeneity* (different genotypes cause the same phenotype) and/or *interactions* between genes. Complex disorders are often common in the population, but it is hard to identify the risk variants. This is partially because the disease has various expressions among the cases, but also because each involved gene has a subtle marginal effect on disease risk.

One example of a common complex disorder is *Celiac Disease*, this disease show a strong association to the *Human Leutocyte Antigen* (HLA) class II region on chromosome 6. In addition to this necessary genetic risk factor there are also more genetic variants and the environmental factor of gluten which contributes to the development of the disease.

1.1.2 Inheritance

Many of our traits are inherited from our parents. By studying and comparing our genotypes and traits with the genotypes and traits of other related and unrelated indiviuals we can identify which genotypes give rise to different phenotypes.

In 1865 Gregor Mendel discovered what is today known as *Mendels laws* [1], which was later rediscovered and reformulated in the early 20th century as the Chromosomal theory of inheritance.

The section below contains descriptions of how genetic variation is created during reproduction and how dependence between loci can be measured.

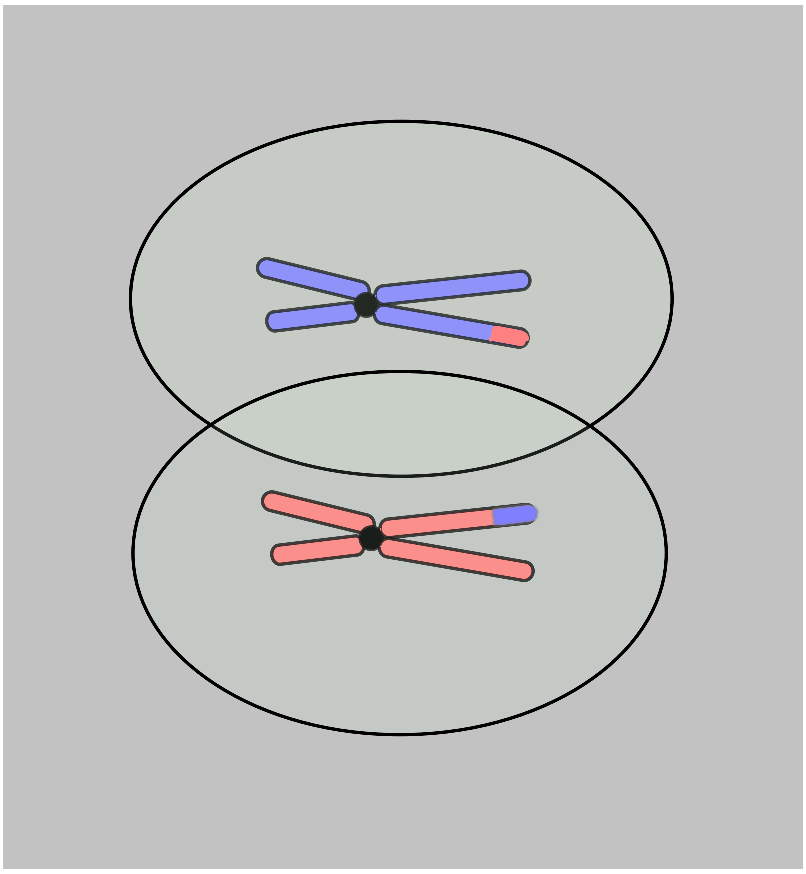
Cell Division, linked genes and genetic maps

During reproduction, the cell divides in new cells through *meiosis* in two different stages. During **meiosis I** the homologous chromosomes are separated in two new *haploid* cells, each cell contains one of each chromosome. In **meiosis II** the two *chromatides* of each chromosome are separated in two new haploid cells.

Before the formation of haploid cells during meiosis I homologous chromatides will cross over each other, both chromatides will break at the same positions and the broken piece will join the other chromatide. This event, which is illustrated in Figure 1.1, occurs randomly and sometimes at multiple positions on each chromosome. The result of this will be an alternating sequence with pieces from both of these chromatides, which creates genetic variation. The probability of a cross-over will increase with increased distance between the loci. In some regions of the genome the intensity for crossovers are higher than in other regions. The frequency of crossovers is measured with *recombination rate* θ, the probability of observing a single crossover between the two loci during meiosis. In many regions of the genome the recombination rate is very low, and in such regions there will be association between pairs of loci.

Consider two loci situated at the same chromosome, with alleles A, a and B, b respectively. If two loci are situated close to each other then it is less likely for cross-overs to occur between these loci, and the alleles tend to be inherited together during meiosis. Two loci are said to be *linked* if θ is less than 0.5, i.e. it is rare with crossovers between these loci. If the two loci are linked the possible alleles will be correlated, e.g. it holds that $P(AB) \neq P(A)P(B)$,

Figure 1.1: *Crossovers during Meiosis I*

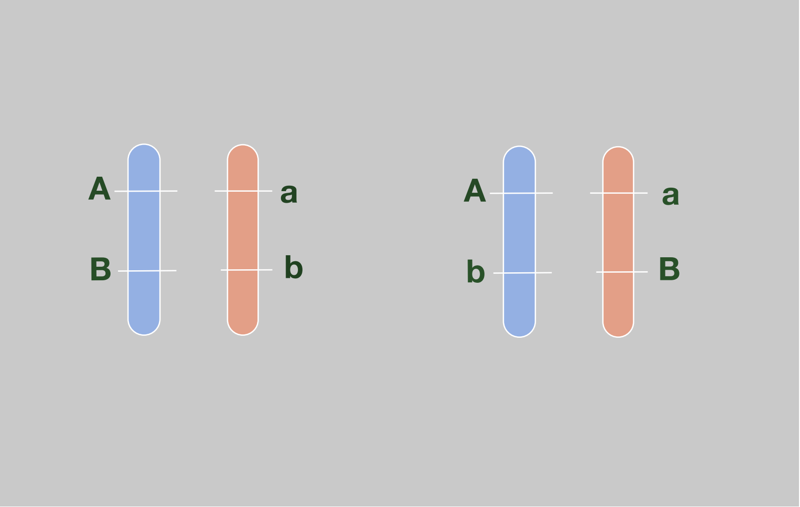


and the loci are then said to be in *Linkage Disequilibrium* (LD). LD is created by evolutionary forces like mutation, drift and selection, and is diminished by recombinations [2, 3].

The combination of one allele from each of these two loci which an individual receives from one parent is called a *haplotype*. The concept of haplotypes can also be generalized to involving more than two loci.

If an individual has genotypes Aa and Bb, then there are two possibilities for how the alleles at these two loci were inherited from the parents, either as AB and ab or as Ab and aB. These are called the *phases* of the genotype combination, and are illustrated in Figure 1.2. It is often possible to determine the phase of the haplotype by studying the parents' genotypes.

Figure 1.2: *The two possible phases for a doubly heterozygous individual.*



For a *doubly heterozygous* parent Aa/Bb the child can receive any of the four possible haplotypes AB, Ab, aB, ab. If the loci would be independent each one is inherited with probability $1/4$ each, but when there is LD between the loci then

$$P(AB|\text{parent is } AB/ab) > 1/4.$$

That is, it is more likely that the child inherits one of the two haplotypes with the same phase as the parent's haplotypes. For two such loci it also holds that the population frequency of some haplotype(s) A_iB_j is

$$P(A_iB_j) \neq p_{A_i}p_{B_j}$$

A genetic map contains information about the frequency of cross-overs across the entire genome. To measure the intensity of cross-overs between two loci in the genome we use *genetic distance* Morgan (M), where 1 M corresponds to an expected number of 1 cross-over between two loci. With a genetic map we can translate the physical distance between two loci into genetic distance. With this distance measure the occurence of crossovers is uniform over the chromosome.

1.2 Mapping the Human Genome

"One can systematically discover the genes causing inherited diseases without any prior biological clue as to how they function."

Eric S. Lander and Nicholas J. Schork [4]

The genome contains approximately 3 billions of base pairs. Most of the genome is identical for all humans, but about 0.1 % varies between different individuals. It is these variations that influence many of our variable traits such as height and eye colour. With this genetic knowledge comes also an urge to explain the biological mechanisms behind diseases and other traits which seem to be inherited from parent to offspring.

The task of making a thorough catalog of the human genome involves scientists from fields like Molecular Biology, Computer Science and Mathematical Statistics. One aim is to measure the genetic variations and identify their function in genetic diseases. By identifying the genetic variants which affect the risk of a certain disease it might be possible to diagnose cases at an earlier stage of the disease, and patients can start treatment before the disease is severe. Since not all patients are helped by the same kind of treatment, it would be desirable to choose treatment based on genetic tests. In this way patients could start the appropriate treatment earlier, without having to try out treatments which are inefficient for them.

1.2.1 From Linkage to Association

In this historical review several key concepts are introduced, more detailed descriptions of these are given throughout this chapter.

Already in the early 20th century, mapping of genes to positions in genomes was performed with experimental organisms using controlled crosses [4, 5]. But for ethical reasons these kind of experiments was not possible with humans. Therefore it was not until 1980 that it became possible to perform genetic mapping in humans. Following the discovery that highly polymorphic genetic markers could be used to trace inheritance in human pedigrees, researchers started constructing *Linkage Maps* of the human genome [6]. This enabled searching any genomic region for *genetic linkage*, positions where chromosomal segments are co-inherited with the trait in families. With hundreds of neutral genetic markers distributed across (parts of) the genome, linkage analysis successfully 'mapped' hundreds of, mostly mendelian, traits [4]. But Linkage analysis does not perform as well in Complex diseases [7].

Recently there has been a shift of focus from family based linkage studies towards population based *case-control* and *cohort studies* with thousands of unrelated subjects. The first step towards *association analysis* was taken in the later part of the 1980's when the idea arouse that LD patterns across the genome could be used for mapping disease genes [8]. Association analysis incorporates the concept of indirect association between genetic markers and disease status described in Section 1.2.3. Now it was possible to also use unrelated individuals to locate disease genes. In complex diseases each genetic risk variant often have a small effect on the disease risk. According to Risch and Merikangas [9] association analysis has a greater power to detect these small effects compared to linkage analysis, also when the markers are chosen without prior knowledge of the genetics of the disease.

In 1996 Lander [10] proposed the hypothesis of *Common Disease Common Variant*, which was adopted as a strategy for the GWAS. In order to capture the risk loci for these diseases a sufficient number of genetic markers (SNPs) where needed. To explore how dense set of markers where needed to capture most of the common genetic variation, the HapMap project was initiated [11]. And today biotechnology companies are developing high-throughput genotyping technologies based on that 500 000-1 000 000 SNPs can be used for 'tagging' about 80 % of the common SNPs, if chosen suitably [12].

Both Linkage and Association analysis rely on the property of linkage disequilibrium, linkage exploits the LD within a pedigree, and with assocition we incorporate the LD on population level.

1.2.2 Genetic Linkage

With Linkage Analysis disease genes can be mapped using neutral markers and thereby identify spots where segregation pattern of disease and markers coincide.

The basic idea of linkage analysis is that by studying the pedigree of a family with some affected individuals, it is possible to picture where different crossovers have occured during each meiosis, and thereby locate a narrow interval which includes the disease locus.

Some properties of linkage analysis:

1. Can only be performed with data from related individuals with known pedigree,

2. It is not always possible to determine the phase of a haplotype, and it can therefore be difficult to distinguish where there have been cross-overs and where there have not.

3. For many regions of the genome the recombination fraction varies depending on the gender of the parent transmitting the haplotype.

Parametric Linkage Analysis assumes a mendelian trait (e.g. recessive) model M_1 including a position for the causal gene. The model M_1 is much more likely to have produced the observed data than the model M_0 where there is no linkage to the disease. These models are compared using a *Likelihood Ratio test*, measured by the *lod score*,

$$\mathcal{Z} = \log_{10} \frac{P(data|M_1)}{P(data|M_0)}.$$

The models M_1 and M_0 contains (apart from the loci of the disease gene) parameters for penetrance, recombination fraction and allele frequency of the disease and marker loci. For complex traits complete multipoint linkage analysis becomes a large computational challenge for general pedigrees, even for a handfull of loci [13].

Since parametric linkage is sensitive to misspecification of the linkage model [14], *Non-parametric Linkage* has been useful for the more complicated models.

Many of the mendelian disorders, and also a few non-mendelian disorders (with locus heterogeneity and/or interactions) have been successfully analyzed using this method. The poor results from Linkage analysis in Complex diseases can partially be explained by that the effect sizes usually are too small to be detected by cosegregation within pedigrees [3].

1.2.3 Genetic Association

In contrast to linkage studies, where we examine which haplotypes are *inherited* from parents to affected offspring, when performing association analysis we instead search for loci where the allele or genotype frequencies vary between healthy and affected individuals.

A genetic locus is *associated* with a trait if different genotypes at the locus have different distributions for the trait. E.g. if individuals with one genotype tend to be taller than other individuals, then this locus could be associated with human height. If it is a binary trait (like many diseases) the proportion of cases will differ between the genotypes. This is equivalent to that (some of) the genotype frequencies differ between cases and controls.

Assume that we have a disease locus D with alleles D_1 and D_2, where D_1 is the allele that gives an increased risk for the studied disorder. Consider one marker locus M with alleles M_1 and M_2. Let the studied locus M be close to the locus D of the causal gene, and assume there is LD between these loci, such that the alleles D_1 and M_1 are positively correlated. Then the haplotype D_1M_1 will be more common than expected under the assumption of independence. Because of strong correlation between the alleles D_1 and M_1, M_1 will often be inherited together with the disease gene. This property can be used to search for genes associated with some disease.

One of the main advantages of Association studies compared with Linkage analysis is that they do not require family samples. Instead we can use samples consisting of 'unrelated' cases and controls. When performed in case-control studies associated regions are identified by comparing allele or genotype frequencies among the cases and controls. Case-control studies has the advantage that it is often easier to recruit cases and controls compared to entire families, especially for diseases with late onset. Also, the control samples can often be re-used in several studies. For this reason case-control studies are the most common type of association studies performed.

But families are still useful in association studies. Using *allele sharing* methods, risk genes are identified by searching for loci where heterozygous parents overtransmit one of the two alleles. Family studies has the advantage that they are more robust against population substructures than case-control studies [15, 16], both in the sense of *population stratification* (cases and controls may have differing ancestral backgrounds) and *cryptic relatedness* (the affected individuals tend to be more closely related than the controls).

Chapter 2

Genome Wide Association Studies

In the early 1990's researchers started conducting *candidate gene* studies. After more than 10 years of these studies, few of the associated disease genes had been replicated [17]. This problem can partially be explained by several different issues regarding the study design and the nature of the disorders.

Following this Genome Wide Association (GWA) studies have identified more than 2000 common variants which influence the genetic susceptibility to over 200 complex diseases [3, 11, 18, 19]. The main breakthrough of GWAS was when the *Wellcome Trust Case Control Consortium* published their study in Nature 5 years ago [19]. Many of the detected variants have been previously unsuspected candidates, leading to a better understanding of the biological mechanisms of each trait as well as a general knowledge of the allelic architecture of complex traits. In this section follows a description of the background and different aspects of GWA analysis.

2.1 Data Collection and Methods

The most common study design i GWAS is the *case-control design*, also known as a *retrospective* study design, where 'unrelated' affected and healthy individuals are collected for genotyping. When using a *family based study design*, the samples are collected from families where at least one of the members are affected by the disease.

The case-control design is sensitive to population stratification between case and control samples, which can cause false positives. It is therefore important to consider the optimal selection of samples to minimize or correct for these effects. Family studies are less sensitive to these population substructures, but has a reduced power compared to case-control studies. In case-control studies phenotypic and genetic heterogeneity will often occur in the samples, and family designs are robust against this type of heterogeneity [20]. In addition, case-control design has the advantage that it is easier to collect unrelated subjects, compared to families where complete families are not always available [15].

Following the *Wellcome Trust Case Control Consortium* study [19], it has also become possible to use common controls samples in several studies. One potential problem with such common control samples is unidentified cases among the controls, which might reduce the power if the trait is common. Another possible problem is that some studies use public control data from other countries, not quite matching the case sample.

Starting the era of GWAS, 'Population Stratification' was believed to be a major threat to the success of the case-control approach, suggesting family-based controls [4]. However, it has turned out not to be a large problem if matching or adjusting for reported ethnicity is applied [21]. It also turns out that the GWA data itself can be used to identify the substructures [22]. In order to have enough power to detect effects with *genome wide significance* (p-value $< 5 \cdot 10^{-8}$) it has been necessary to build consortia for large GWAS. With the possibility of collecting such large samples, it is quite easy to detect and correct for population substructures. However, in many studies it is still either hard to find enough cases to collect, or for financial reasons not enough individuals can be genotyped.

Lately the case-control design has also been extended to *population based cohort studies*, usually designed to investigate various traits from the same data [22]. These studies are more useful for continuous traits, and still has quite limited power for dichotomous phenotypes. *Meta analysis* is another approach to overcome the sample size issue, but unfortunately there are difficulties of standardizing studies performed with varying sampling strategies, genotyping arrays etc.

Implementation

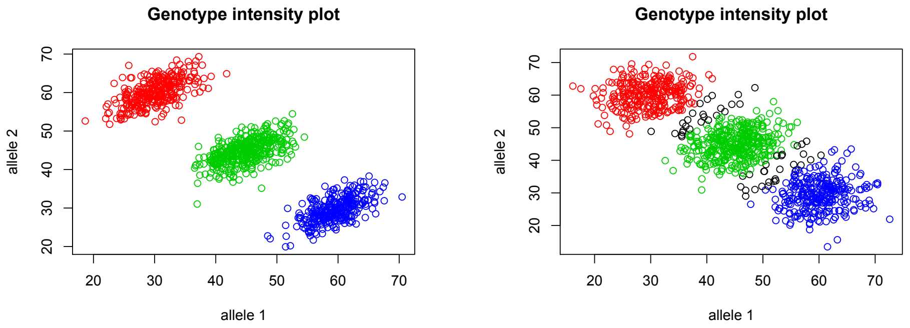
The markers which are used to find these associated genes are generally at positions which vary between individuals, but where the genetic variation is not associated with any traits. The markers used in *Genome Wide Association Studies* (GWAS) are Single Nucleotide Polymorphisms (SNPs). SNPs are variations in the genome where one single nucleotide has been substituted to another, without affecting the neighbouring nucleotides. E.g. if a C nucleotide have been substituted with a T in some individuals, then that locus is a SNP with alleles C and T.

In order to make powerful GWA analyses the SNPs need to be chosen wisely, distributed in a way that reflects the genetic variation. When GWAS was introduced, there was a debate regarding the selection of markers [23, 24]. This resulted in a set of genome-wide chips to choose from. These chips are designed such that they should be able to identify most of the genetic variations. Progress in the technology of these chips have enabled an increased coverage of markers, improving the precision in the association signals. But this does not necessarily imply increased power of detecting associated loci, since it also increases the challenge of adjusting for multiple testing.

This design is based on the assumption of *Common Disease, Common Variant* [9, 10, 25]. This hypothesis is commonly expressed as: 'for several common diseases, most of the genetic risk can be explained by variants with allele frequency about 1-5 % and with a (marginally) modest effect on the increased risk of the disease’. Reich and Lander [26] formulated the hypothesis as ’if the number of loci contributing to disease risk is moderate, then a few disease alleles should account for a large proportion of the genetic risk’.

The raw experimental data obtained from a genome wide experiment does not consist of discrete genotypes, but rather bivariate intensity signals for each of the two alleles. For each marker locus these can be viewed in a two-dimensional plot in order to define three clusters corresponding to each of the genotypes AA, Aa and aa, examples of such plots are given in Figure 2.1. The left panel of the figure illustrates the preferable situation, where we can separate the clusters and assign genotypes to each signal. In comparison the right panel of Figure 2.1 illustrates the case where some of the signals are in between two clusters. The method of assigning discrete genotypes is being replaced by algorithms that assign posterior probabilities to each genotype [16].

Figure 2.1: *Illustration of Raw genotype signals*



In the left panel the clusters are well defined for all three genotypes, in the right panel there is overlap between the clusters which will result in no call for some of the genotypes.

2.2 Missing heritability

In many complex diseases there are numerous genetic variants which have been identified. But for many of the recent studies these common variants only explain a small fraction of the increased risk. Most of those that have been identified have no established biological relevance to the disease and often they are not located inside 'active' genes [3]. From the last years of GWAS it is clear that the common variants fail to explain the majority of the genetic heritability of most human diseases [27].

This suggests that the hypothesis of 'Common disease, common variant' is not as valid as was previously believed. The problem is that the biological reality does not correspond to the study design and assumption of GWAS, and the solution is not to increase the sample size even further but to improve the study design and statistical methods.

One possible explanation to the missing heritability could be some kind of interaction between different genes(*epistasis*). These interactions could be hard to detect when analyzing one SNP at the time, as the marginal effect of a single SNP will be small. Another explanation is that part of the increased risk can be explained by many rare variants, which are present among less than 1 % of the population. This suggests that there could be *heterogenetiy*, where different genetic profiles can cause diseases that are diagnostically the same.

Genetic Interactions

A general definition of genetic interaction (epistasis) is that the effect (penetrance) of one locus varies according to the genotype present at another locus. To detect interactions we need to define how a 'natural' combined effect of two risk loci would be expressed in the organism. The concept of gene-gene interactions is not new, but still it is confusing since the term is used in various ways. *Biological interaction* or epistasis was defined first by Bateson in 1909 [1]. In that example one of the alleles at one locus G is preventing the alleles at locus B from being expressed in the organism. This relation does not necessarily have to be symmetric. This definition is similar to the definition biologists use to examine a biological interaction between proteins, where proteins interact to regulate several cellular processes.

In statistics the definition of *interaction* is usually a deviation from a linear model. In 1918 Fisher made a statistical definition of epistasis [28], as deviation from additivity in effects of the alleles at different loci on a quantitative trait. This definition is more similar to the classical statistical definition of interaction and do not quite correspond to the biological definition of epistasis.

These definitions get troublesome when the trait is binary, in these cases the mathematical modelling often focus on the penetrances. Hence the definitions of epistasis need to be modified. For binary traits an example could be that both allele A and allele B at two different loci are needed to develop the trait. In this case A is epistatic to B, and B is epistatic to A, hence the epistasis is symmetric - in contrast to the definition by Bateson.

A classic way to represent lack of epistasis has been the *heterogeneity model* [29] - a person gets the trait by possessing (at least) one of the predisposing genotypes. This definition actually falls under Bateson's definition of epistasis, for example if a person has both risk variants (situated at different loci) the effect of allele A will be masked by allele B - another confusing issue about these genetic interactions.

There are two types of genetic heterogeneity, *allelic heterogeneity* is when several mutations on the same allele cause the same disease. *Locus heterogeneity* means that mutations in several unrelated loci can cause the same disorder. The above example of locus heterogenetity could be generalized to a situation without full penetrance, that is $0 < f_{i,j} < 1$ for some of the penetrances. Mathematically, locus heterogeneity can be expressed as

$$f_{ij} = \alpha_i + \beta_j - \alpha_i\beta_j, \tag{2.1}$$

where α_i and β_j are the penetrance factors for the two genetic variants [30]. Locus heterogeneity is similar to a daisy chain, where it is enough for one of the components to break (caused by having at least one of the risk variants) for the entire system to malfunction, i.e. to obtain the disease.

There are two other common two-locus models for binary traits, the *multiplicative model* and the *additive model*. The multiplicative model can be expressed as

$$f_{ij} = \alpha_i\beta_j, \tag{2.2}$$

this model is often considered as epistatic. Both the additive model

$$f_{ij} = \alpha_i + \beta_j,$$

and the heterogeneity model are thought of as non-epistatic by most authors. However, some authors [14] considers epistasis as departure from the multiplicative model.

Further problems appear when considering that both the multiplicative and the heterogeneity models become additive with suitable $\log$ transformations.

It will be difficult to really model the true epistatic interactions in complex diseases, and discovered epistatic effects may have limited input to the understanding of the disease. Still, models that allow for interactions can improve the statistical power of detecting the genetic risk variants [31].

The main issue in finding interactions, independent of how you define epistasis, is how you should detect it in complex diseases when analyzing millions of genetic markers. Assume that the disease is caused by different mutations on different loci in various families, and these genes have a strong effect in each of the subpopulations. Then the heterogenetic risk genes will probably show a very weak marginal effect when the markers are analyzed one at the time.

For epistatic interactions it will be very computationally demanding to examine all possible gene-gene interactions, in addititition to the issue of correcting for testing multiple hypotheses. One way to handle this is to first test for marginal main effects for each marker in the sample, and hope that the genes involved in interactions will also show at least a modest marginal effect. Then the results from this analysis is combined with biological knowledge to suggest a number of candidates for interaction analysis.

Imputation of genotypes

The different genotyping platforms often differ in their marker sets, this can cause problems when researchers want to combine several data sets, since some markers will only be genotyped in parts of the study material. During the last few years, collaborations like the *International Hapmap Project* [11] and the

1,000 Genomes Project [32] have enabled a large catalog of the human genetic variation, which is growing for each month still. These reference haplotypes, which are assayed over a dense set of SNPs, are useful for predicting unobserved genotypes through Genotype Imputation. The way SNP arrays are designed make them well suited for imputation, since they efficiently capture most common variations across the genome.

Using effective imputation algorithms, we can predict or impute genotypes at (partially) unobserved markers and thereby increase the sample size at these loci and thus improve the power and accuracy of the association analysis. The algorithms are based on known genotypes at typed markers and information about LD patterns in a reference sample [33], which is used to predict the genotypes of markers which were not observed in (parts of) the study sample.

Most of the algorithms are based on *Hidden Markov Models* and *Markov Chain Monte Carlo* [34] methods and they provide posterior probabilities for each of the three possible genotypes at each locus. It is then possible to apply cutoffs to these probabilities in order to impute the most confident genotypes, or perform *imputation-based association analysis* [35]. Association tests for imputed markers should be similar to test signals for other markers on surrounding loci. Therefore it is important to be cautious with checking if an imputed marker has a very different association signal compared to the surrounding markers.

One important issue of genotype imputation is that the different providers of SNP arrays present the alleles relative to either the ’+’ or the ’-’ strand of the human genome reference. This implies that, when alleles A and C are observed at a specific locus using one platform, the complementary alleles T and G could be observed with some other platform. If annotation files are available it is simple to flip the alleles in the study material that are different from the expressed alleles in the reference sample before the imputation is performed [36].

HapMap provides references datasets for several human populations, enabling to choose a reference with an ancestry matching the studied sample. There are several softwares for imputation, which use varying algorithms, some of the most common are Mach, Beagle, Impute and Plink [33, 34, 37–41]. The increased availability of Next Generation Sequencing (NGS) data, such as the

1000 Genomes Project, will influence how imputation is used. This data will notably increase the available number of SNPs, haplotypes and populations, compared to the HapMap 2 and 3. This might also enable identification of rarer variants [36].

2.3 Statistical methods in GWAS

If a genetic marker is associated to a particular disease, then the genotype or allele frequencies will be different among affected and healthy individuals. A commonly used test for searching for associated SNPs in case-control studies is a Pearson χ^2_1 test applied to a 2-by-2 table of allele counts in the two groups. For complex traits it is commonly assumed that the contribution to the genetic effect from each SNP is roughly additive [42], i.e. the penetrance for heterozygous are somewhere in between the penetrance for the two homozygotes. This test is powerful for additive models, whereof the popularity of this test in these studies. Other common tests include a Pearson χ^2_2 test comparing the genotype frequencies instead of allele frequencies, Cochran Armitage test for trend in penetrances, and logistic regression.

The *Transmission Disequilibrium Test* (TDT) is an association test using data from families with at least one affected child. This test was introduced by Spielman et al. [43], and the test evaluates the transmission of an allele from a heterozygous parent to the offspring.

The TDT is based on the assumption that each of the two alleles M_1 and M_2 at a locus is transmitted with equal probability to the offspring, hence for a sample of heterozygous parents we expect approximately half of them to transmit the allele M_1. If one of the alleles is transmitted more often among families where the children have a genetic disease, we suspect that the allele is associated to the disease.

Let b denote the number of heterozygous parents who transmits allele M_1 to their offspring, and c the number of heterozygous parents who transmits allele M_2. Conditioned on $b + c$, b is is binomially distributed, but usually the test statistic has the following form

$$T = \frac{(b-c)^2}{b+c}, \tag{2.3}$$

This test asymptotically follows a χ^2_1-distribution and is equivalent to a Pearson χ^2-test.

Logistic Regression

Generalized Linear Models (GLMs) [44] extend the ordinary regression models to other response variables than the Normal distributed. GLMs are applicable if the response variable has a distribution which belongs to the natural exponential family. One of those distributions is the Binomial distribution, and with *Logistic Regression* we model the binomial probability $p(\mathbf{x}) = P(Y = 1|\mathbf{x})$ as

$$\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \alpha + \sum_j \beta_j x_j \tag{2.4}$$

Here x_j denotes the value of the jth element in the predictor $\mathbf{x}$. In the simple logistic regression with one binary predictor x, β is equal to the log odds ratio

$$\beta = \frac{p(x=1)/(1-p(x=1))}{p(x=0)/(1-p(x=0))}$$

In *retrospective* (individuals are sampled based on their affection status) studies the effect parameter β will be the same as in the *prospective* (sampling based on the predictors) design, if we assume that the sampling probability is independent of x. This is one of the main reasons for using this method in biomedical studies [45]. Another advantage with the logistic regression is that it is easy to include several predictor in the analysis and make inference for interactions between genes and environment, as well as gene-gene interactions.

Schaid [46] described a univariate method for case-parent data, modelling genotype relative risks with *conditional logistic regression* using three pseudo-controls based on the parents' untransmitted alleles. This method can be generalized to two loci. For case-control data logistic regression can be used to analyse interactions by comparing the saturated model to an additive model, specified on the form of (2.4) [31].

The additive logistic model is roughly equivalent to the heterogeneity model if the relative risk (RR) or odds ratio (OR) is of moderate size. However, North et. al [47] show examples of heterogeneity models which are marginally recessive (marginal RR$\approx$ 150), in this case the logistic regression yields non-zero interaction estimates.

Hence, to really examine deviations from the heterogeneity model (and not the multiplicative or logistic model) more advanced methods need to be applied.

Chapter 3

Summary of Papers

There are four papers included in this thesis, the first two papers are two GWA studies of Celiac Disease. The second and the third paper covers inference of two-locus interactions in GWA studies. The final paper describes a method for locating a causal variant for mendelian diseases by haplotype sharing.

Paper I:
Utilizing known risk genes within Celiac Disease

A common test in Genome Wide case-control associaton studies is the Pearson χ_1^2-test comparing allele frequencies among the two groups.

In celiac disease (CD) a genetic variant in the HLA-region on chromosome 6 in the human genome is necessary but not sufficient, for developing the disease. As this variant also is present in healthy individuals, other risk variants should be less common among the controls who possess the necessary gene, compared to the controls who lacks this variant. Similarly, these additional risk variants should also be more common among the cases. Hence we have refined the alternative hypothesis to

$$H_1 : p_A^{ctrl+} < p_A^{ctrl-} < p_A^{case},$$

where p_A^* denotes the frequency of the risk increasing allele A in each of the three subpopulations. $ctrl+$ denotes the population of individuals who has the necessary genetic component, denoted by H, but is not affected by the studied disease. Similarly, $ctrl-$ consist of all individuals who do not have the gene H, and finally the cases. In the paper we derive a test that can examine this kind of genetic model.

A test for trends in proportions is the Cochran-Armitage test [48]. This test needs a parameter ρ describing the relative differences between the proportions, that is

$$\rho = \frac{p_A^{ctrl-} - p_A^{ctrl+}}{p_A^{case} - p_A^{ctrl+}}. \tag{3.1}$$

We show that $\rho = P(\text{aff}|H)$, hence we estimate ρ by the disease prevalence among the individuals who has the necessary gene H. This entity is thus independent of the marginal model for any other gene that we are searching for.

We use simulations from various genetic models of this type to estimate the power of this test compared with the Pearson 1 df chi-square test. We also apply this method to a previously published [49] celiac disease case-control study and compare the result on genes which were replicated in further studies.

Paper II:
Genome-wide linkage and association analysis in celiac disease families identifies genetic variants within DUSP10 and implicates genes involved in metabolism and energy homeostasis

The aim of this applied paper is to uncover additional genetic risk factors in Celiac Disease. To accomplish this we perform a Genome Wide Linkage and Association analysis on a scandinavian family material, where at least two members of each family are affected by CD. In addition, we also perform pathway analysis and two-locus interaction analysis of the 383 top-scoring regions, as well as comparing gene expression levels between cases and controls. When combining association analyses with pathway and expression analysis we verified several previous findings and identified new variants involved in growth and energy homeostasis.

Since the the material was genotyped using two different arrays we impute unobserved genotypes using Impute2 with the HapMap 2 as a reference to increase the sample size and improve accuracy for the association analysis.

Univariate Association analysis

In the first analysis we perform an *imputation based* Transmission Disequilibrium test (TDT). Out of the genotypes included in imputation, 88 % have a posterior probability above 0.95. Therefore we present a test based on 'posterior expected transmission counts'.

We compared this test with the standard TDT defined in (2.3), with a threshold of 0.95 for the imputation probabilities. The expected counts TDT was able to boost the association signals and incorporates a check for mendelian errors which were created when imputation was performed without considering the relationships.

In addition, based on the prior knowledge of the necessary HLA risk variants, we performed a stratified TDT analysis that identified a genome-wide significant association to the DUSP10 gene for the low-risk group.

Two-locus Interaction analysis

Based on the top results from the TDT analysis we chose 383 genomic regions for two-locus interaction analysis. For this analysis one affected child from each family was chosen. For imputed SNPs, we imputed genotypes if the joint likelihood of the three subjects was above 0.95 for some of the possible genotype vectors (according to mendelian inheritance).

A Likelihood ratio test was applied, comparing four models of no association, heterogeneity, multiplicative and epistasis. The maximum likelihood estimates of the penetrances and allele frequencies for each of the models were obtained numerically. We identified 15 pairs which deviated from the heterogeneity model, 5 of these were interactions with the HLA region. In addition we identified 7 pairs of loci which had a joint heterogenic effect on disease risk.

Paper III:
Are two-locus interactions without marginal effects in Genome Wide Association Studies really that interesting?

One suggested explanation to the missing heritability in complex genetic diseases has been interactions between genes (epistasis). However it has previously not been very common to perform a complete genome wide interaction analysis even for two-locus interactions. Instead the interaction analysis has been done by a two-step approach.

In this paper we examine how common interactions without marginal effects will be in a GWIA, assuming the null hypothesis that none of the predictors has any effects. With the use of small sample examples we illustrate the phenomenon of significant interactions without marginal effects. We consider two different study designs, the retrospective (case-control) and the prospective study design.

We found that the possible outcomes with the most significant interactions will not have any marginal effects at all. But for large samples these events will hardly occur.

Paper IV:
The Color Method – a simplified tool for locating risk regions with GWA data in mendelian disorders

With Homozygosity mapping, a recessive trait can be mapped using cases from large inbred families. A region containing a risk variant is located by searching for regions where affected individuals are homozygous for the same allele at each of multiple consecutive markers.

In this paper we develop a Simplified Linkage Program called the *Color Method* that constructs illustrations of candidate regions for the causal locus of rare mendelian traits, i.e. also for dominant traits. The method assumes that the obligate haplotype is the only one which is shared IBD by all cases. To discern the obligate region from other candidates, the method estimates the frequencies of the haplotypes using a public reference sample and assign a score for each region based on these frequencies.

The assumption of only one shared haplotype is crucial. To validate this assumption crossovers and IBD sharing was modelled in simple pedigrees. By theory and simulations we model the crossovers using the *Ehrenfest Urn Model* [50] for a random walk on the hypercube $\{0, 1\}^k$.

To assertain that the strongest signal is from the causal variant we model and simulate meiotic crossovers using HapMap reference haplotypes and measure the distribution of both IBD and IBS sharing for a given pedigree. We find that in order to discern the causal haplotype we need not only a sufficiently large number of cases, but also they need to be rather distantly related.

The method is applied to three different datasets, one recessive and two dominant traits. For all three datasets a unique region is successfully identified.

Bibliography

[1] W. Bateson. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, 1909.

[2] D. L. Hartl and A. G. Clark. *Principles of Population Genetics, Second Edition*. Sunderland: Sinauer Associates, 1997.

[3] Peter M. Wisscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five years of gwas discovery. *American Journal of Human Genetics*, 90:7–24, 2012.

[4] ES Lander and NJ Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, 1994.

[5] A. H. Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59, 1913.

[6] David Botstein, Raymond L. White, Mark Skolnick, and Ronald W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331, 1980.

[7] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33:228 – 237, March 2003.

[8] ES Lander and D. Botstein. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Cold Spring Harbor Symposia Quantitative Biology*, 51:49–62, 1986.

[9] N Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.

[10] Eric S. Lander. The new genomics: Global views of biology. *Science*, 274(5287):536–539, 1996.

[11] Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins. A hapmap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605, 5 2008.

[12] Leonid Kruglyak. The road to genome-wide association studies. *Nature Reviews Genetics*, 9(4):314 – 318, April 2008.

[13] Leonid Kruglyak, Mark J. Daly, Mary Pat Reeve-Daly, and Eric S. Lander. Parametric and nonparametric linkage analysis:a unified multipoint approach. *Am. J.Hum. Genet.*, 58:1347–1363, 1996.

[14] F Clerget-Darpoux, C Bonaiti-Pellie, and J Hochez. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, 42:393–399, 1986.

[15] Nan M. Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385 – 394, May 2006.

[16] Mark I. McCarthy, Goncalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356 – 369, May 2008.

[17] Joel N. Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genetics in medicine*, 4(2):45–61, 2002.

[18] LA Hindorff, J MacArthur (European Bioinformatics Institute), A Wise, HA Junkins, PN Hall, AK Klemm, and TA Manolio. A catalog of published genome-wide association studies., 2012. [accessed on 3 May 2012].

[19] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661 – 678, June 2007.

[20] Nan M. Laird and Christoph Lange. The role of family-based designs in genome-wide association studies. *Statistical Science*, 24(4):388 – 397, 2009.

[21] Sholom Wacholder, Nathaniel Rothman, and Neil Caporaso. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *Journal of the National Cancer Institute*, 92(14):1151–1158, 2000.

[22] David J. Hunter. Lessons from genome-wide association studies for epidemiology. *Epidemiology*, 23(3):363 – 367, May 2012.

[23] Jeffrey C Barrett and Lon R Cardon. Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6):659 – 662, June 2006.

[24] Itsik Pe'er, Paul I W de Bakker, Julian Maller, Roman Yelensky, David Altshuler, and Mark J Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, 38(6):663 – 667, June 2006.

[25] Gary K Chen, Eric Jorgenson, and John S Witte. An empirical evaluation of the common disease-common variant hypothesis. In *BMC Proceedings, Genetic Analysis Workshop 15*, pages 1–4, December 2007.

[26] David E Reich and Eric S Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502 – 510, 2001.

[27] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747 – 753, October 2009.

[28] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edin.*, 52:399 – 433, 1918.

[29] John P. Rice Rosalind J. Neuman. Two-locus models of disease. *Genetic Epidemiology*, 9(5):347–365, 2005.

[30] Neil Risch. Linkage strategies for genetically complex traits. i. multilocus models. *Am J Hum Genet.*, 46(2):222–228, 1990.

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

[32] Richard M. Durbin, David Altshuler, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, and Francis S. Collins et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[33] Joanna M Biernacka, Rui Tang, Jia Li, Shannon K MvDonnel, Kari G Rabe, Jason P Sinnwell, David N Rider, Mariza de Andrade, Ellen L Goode, and Brooke L Fridley. Assessment of genotype imputation methods. In *BMC Proceedings, Genetic Analysis Workshop 16*, pages 1–5, December 2009.

[34] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, June 2009.

[35] Yu-Fang Pei, Lei Zhang, Jian Li, and Hong-Wen Deng. Analyses and comparison of imputation-based association methods. *PLoS ONE*, 5(5):e10827, 05 2010.

[36] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11:499– 511, 2010.

[37] Mach. `http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html`. [accessed on 7 July 2012].

[38] Y Li, CJ Willer, J Ding, P Scheet, and GR Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34:816–834, 2010.

[39] Brian L. Browning. Beagle. `http://faculty.washington.edu/browning/beagle/beagle.html`. [accessed on 7 July 2012].

[40] Brian L. Browning and Sharon R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210 – 223, 2009.

[41] Shaun Purcell, Benjamin Neale, Kathe Todd-Brownand Lori Thomas, Manuel A. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. de Bakker, Mark J.

Daly, and Pak C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81:559 – 575, September 2007.

[42] David J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781 – 791, October 2006.

[43] Warren J. Ewens Richard S. Spielman, Ralph E. McGinnis. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet.*, 52(3):506–516, 1993.

[44] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972.

[45] A. Agresti. *Categorical Data Analysis, Second Edition*. Wiley Series in Probability and Statistics, 2002.

[46] Daniel J. Schaid. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, 13(5):423–449, 1996.

[47] BV North, D Curtis, and PC Sham. Application of logistic regression to case-control association studies involving two causative loci. *Human Heredity*, 59:79–87, 2005.

[48] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, September 1955.

[49] Hunt et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics*, 40(4):395–402, 2008.

[50] P Ehrenfest and P Ehrenfest. *The Conceptual Foundations of the Statistical Approach in Mechanics*. Cornell University Press, Ithaca, NY, 1959.