THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Improving methods for Genome Wide Analysis of Coeliac Disease

MALIN ÖSTENSSON

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2011

Improving methods for Genome Wide Analysis of Coeliac Disease

Malin Östensson

NO 2011:10
ISSN 1652-9715

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, 2011

Improving methods for Genome Wide Analysis of Coeliac Disease

Malin Östensson

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

Abstract

The topic of this thesis focus on how statistical methods for analyzing Genome Wide Association data can be improved, particularly in the case of Coeliac Disease (gluten intolerance). The key idea is that in complex diseases where several genes are involved, 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 Coeliac Disease. We use this knowledge to illustrate that among healthy individuals, the distribution of other genetic risk variants will depend on if they have the necessary variant or not. In Paper I we propose a method which use the Cochran Armitage test to test for 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 Coeliac Disease case-control material.

Genotype imputation can be used to increase the sample size in Genome Wide Association studies, especially when several studies with different sets of markers are combined. In Paper II we use imputation to increase the sample size at markers which was only genotyped in part of the sample in a Coeliac Disease family material. A version of the TDT in which the imputation probabilities are used, is applied to these data. In addition a likelihood ratio test searching for two-locus interactions by comparing the heterogeneity and epistasis models is applied to the same material.

Keywords: Genome Wide Associtation Studies, Coeliac Disease, Genotype imputation

List of Papers

This Licentiate thesis include the following papers

- **Östensson M**, Naluai A.T, van Heel D., Nilsson S., "Utilizing known risk genes within Coeliac Disease,"
- **Östensson** M., Nilsson S., Naluai A.T, "A sib-pair genome wide association analysis in coeliac disease identifies genetic variants within nutrient sensing and metabolic pathways "

Acknowledgments

I would like to thank my supervisor Staffan Nilsson for his great support and and help during this work. I also would like to thank my co-author Åsa Torinsson Naluai for good collaboration and help in sorting out strand issues.

Finally I would like to thank my friends and my family, thank you for being so encouraging and supportive.

Namaste,

Malin Östensson
Göteborg, April 2011

It runs in the family

Contents

Part I

INTRODUCTION

Chapter 1

Introduction

Genetic association studies aim to identify genetic variants that vary between individuals with different disease states (affected/unaffected). In this chapter we will give the genetic background to the subject.

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 - all 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. Many traits are passed from generation to generation, the units that transmit these traits from parent to offspring are called *genes*, each gene consists of a sequence of DNA which influence some function in the organism. Diploid organisms (like humans) have two copies of each gene

- one on each of the two chromosomes of the same type - which they inherit from their parents. An allele contain genetic information at a certain locus on the parental chromosome. Each gene occupy a certain position (locus) on the chromosome, and the parent randomly pass on one of the two alleles of each gene to its offspring. At each locus in the genome which include population variation, i.e. the alleles are different, 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. A *phenotype* is the physical expression of a genotype, e.g. an individual's eye colour.

1.1.1 Cell Division, genetic maps and linked genes

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 cells, each cell contains one of each chromosome. In **meiosis II** the two *chromatides* of each chromosome are separated in two new haploid cells. A haploid cell contains only one of each chromosome. During meiosis I 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 occurs randomly and sometimes several times 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. 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.

Two genes are said to be *linked together* if it is rare with recombination between these two, these two will then in most situations be passed together from parent to offspring.

1.2 Coeliac Disease

Coeliac disease (CD) is a common, complex and life-long disease. It is autoimmune and unique in the way that the environmental factor is precisely known. The prevalence of CD is about 1-3 % in the caucasian population [1]. Individuals having the disease get an inflammation in the small intestine from a diet containing gluten. Gluten is a protein derived from wheat, barley and rye. CD was previously believed to be a malabsorption syndrome among children, but now it is known that it can be diagnosed at any age. It affects many internal organs and it is associated with several other diseases such as Down's syndrome, Turner's syndrome and Type I Diabetes. The only available therapy is a gluten-free diet, but up to 30 % of the patients still get symptoms [2].

This disorder show a strong association to the *Human Leutocyte Antigen* (HLA) class II region on chromosome 6. About 90% of celiac patients carries alleles that code for the HLA-DQ2 protein, and most of the remaining cases carry the HLA-DQ8 instead. These variants are however present in 25-30 % of the caucasian population, hence these genes are necessary but not sufficient. Twin studies suggest that the contribution of the HLA genes to the total genetic component of CD is less than 50 % [3]. This altogether indicates that there are more genes needed to be affected by CD. In the present search for genes which increase the risk of CD, there are several regions in the human genome which have shown association to CD, but these genes have not yet been verified.

It is quite possible that there is interaction between different genes (epistasis) and rather likely that there is heterogeneity (or phenocopies).

Gluten proteins are digested by enzymes into aminoacids and peptides, gliadin peptides damages the epithelial cells and the increased permeability enables them to get into the lamina propria, where the peptides interact with HLA-DQ2 (or HLA-DQ8), inducing production of *cytokines* which will damage the tissue.

The sympthoms of the disease vary a lot between individuals. Among adults the disease is more common among women than men, similarly to other autoimmune disorders, but this is not the case for individuals older than 65 years. The

heterogeneity in symptoms makes it difficult to diagnose/detect cases and there are many 'silent cases' which sometimes are wrongly diagnosed, but the rate of diagnosed cases is increasing.

The diagnosis includes a *Duodenal Biopsy* and a gluten-free diet. The biopsy often needs to be performed at least 4-6 times, and since it is a unpleasant method of diagnosis, other methods would be preferable for this expensive and socially inconvenient disorder.

Chapter 2

Genome Wide Association Studies

2.1 Genome Wide Association

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.

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.

2.1.1 We inherit our parents traits

Many of our traits are inherited from our parents, and 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.

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, 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 also *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. If there are *phenocopies* $f > 0$ for subjects who do not carry the risk allele (explained by risk variants at other loci or only environmental factors).

A mendelian trait has a clear relation between the causative gene and the phenotype. There are also many *non-mendelian traits* such as polygenic or complex traits and sex-linked traits.

Complex traits which are the subject of this thesis, are traits where typically several genes and environmental factors are involved. These traits or disorders are often common in the population and have various expressions in the affected individuals. Each involved gene commonly has a subtle marginal effect, and it is therefore hard to identify.

2.1.2 Genetic Association

It has been very useful to map disease genes using neutral markers and thereby identify spots where segregation pattern of disease and markers coincide, this

is done using *Linkage Analysis* with samples from families with multiple cases. Many of the mendelian disorders have been successfully analyzed using this method. But Linkage analysis does not perform as well in Complex diseases [4].

Consider two loci M with alleles M_1 and M_2, and D with alleles D_1 and D_2, if the alleles M_1 and D_1 are associated with each other in a non-random manner, that is $P(M_iD_j) \neq P(M_i)P(D_j)$, because of close physical positions on the same chromosome (and recombination between the loci is rare), they are said to be in *Linkage Disequilibrium* (LD). This property can be used to search for genes associated with some disease. If a studied locus is close to the locus of the causal gene, it is likely that the marker allele M_1 is in LD with the disease allele D_1, and M_1 will tend to be inherited together with the disease gene.

With a dense set of markers distributed across parts of or the entire genome it is possible to perform *Association studies*. 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, or if it is a binary trait like CD the proportion of affected will differ between the genotypes. For a binary trait this can also be expressed as different frequencies among affected and healthy individuals. As the genotyping technology has improved it has become possible to obtain data from a much larger sets of markers, improving the precision in the association signals.

Association studies do not require family samples, but can also be performed with 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 have the advantages that it is often easier to recruit cases and controls compared to entire families, and controls can often be re-used in several studies. But families are still useful in association studies, by comparing how frequently each of the alleles from a heterozygous parent is transmitted to the offspring. Family studies has the advantage that they are more robust agains population substructures than case-control studies [5].

The markers which are used to find these associated genes are generally 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.

The genotyping chips which are used in GWAS are designed such that the chosen SNPs are distributed evenly across the genome in a way that they shall be able to detect most of the common genetic risk variants. This design is based on the assumption of *Common Disease, Common Variant* [6], which says that for several common diseases, most of the genetic risk can be explained by variants with allele frequency about 1-5 % with a (marginally) modest effect on the increased risk of the disease.

2.1.3 Missing heritability

In many complex diseases there are many 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. This suggests that the hypothesis of 'Common disease, common variant' is not as valid as was previously believed. [7]

Possible explanations include 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. Other explanations could be some kind of interaction between different genes which could be hard to detect when analyzing one SNP at the time. Such interactions could be *epistasis*, where the effect of one gene is affected by other genes.

2.2 Genetic Interactions

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 [8]. 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 [9], 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* [10] - 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

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

where α_i and β_j are the penetrance factors for the two genetic variants [11]. Locus heterogeneity is similar to a daisy chain, where it is enough for one of the components to break 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*,

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

this model is often considered as epistatic, and we also have the *additive model*

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

both the heterogeneity and the additive model is thought of as non-epistatic. There is an interesting relationship between a heterogeneity disease model and a multiplicative models.

A heterogeneity disease model is equivalent to a multiplicative health model.

In the heterogeneity model you get affected if you have at least one of the risk variants, hence you stay unaffected if you have the 'protective alleles' at both loci.

Further problems appear when considering that both the multiplicative and the heterogeneity models become additive with suitable $\log$ transformations, such as *logistic regression*, which is a popular method among epidemiologists. If the underlying model is heterogeneity logistic regression will interpret it as interaction.

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. If 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, 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 addition 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.

2.3 Imputation of genotypes

During the last few years, collaborations like the **International Hapmap Project** and the **1,000 Genomes Project** have enabled a large catalog of the human genetic variation, which is growing for each month still. When researchers collect several data sets using SNP arrays from platforms with different sets of SNP markers, some markers will only be genotyped in parts of the study material.

Using effective imputation algorithms, we can predict or impute genotypes at these markers and thereby increase the sample size at these loci and the power of the association analysis. The imputation algorithms are based on known genotypes at typed markers and information about LD between markers from a reference sample [12], that have been genotyped on a much more dense set of markers, which is used to predict the genotypes of markers which were not observed in (parts of) the study sample.

The sets of SNPs used in most SNP arrays are chosen in a way that that they should efficiently capture most common variations across the genome, this is well suited for imputation, compared to random selection of markers. But the genotyping platforms often differ in their marker sets, hence imputation is an important tool to merge studies genotyped on different platforms.

HapMap provides references datasets for several human populations, and there are several softwares for imputation, which use varying algorithms, some of the most common are Mach, Beagle, Impute and Plink [12–16].

Most of the algorithms are based on *Hidden Markov Models* and *Markov Chain Monte Carlo* [14] methods and they provide posterior probabilities for

each of the three possible genotypes at each locus, cutoffs can then be applied to impute the most confident genotypes.

One issue that makes this imputation troublesome is that the different providers of SNP arrays use different strands when reading the alleles at the markers. This implies that, when allele A is observed at a specific locus using one platform, the complementary allele T would be observed with some other platform. Hence we need to flip the alleles at loci where the possible alleles in the study material is different from the reference sample. This gets even more difficult at loci where the possible alleles are A and T (or C and G), then it is not possible to directly spot if the alleles are read on the same strands or not, in the two materials. Hence a 'strand translation' code is needed to flip the alleles which need to be flipped in the study sample before the imputation can be performed. If this information is not directly provided from the genotyping platform it also is time consuming to sort this out by comparing the alleles in different data sets.

Association tests for imputed markers should be similar to test signals for other markers on surrounding loci, hence it is important to be cautious with checking if an imputed marker has a very different association signal compared to the surrounding markers.

2.4 Statistical methods in GWAS

If a genetic marker is associated to a particular disease, then the genotype or allele frequencies will be different among cases and controls. A commonly used test for searching for associated SNPs in case-control studies is a Pearson χ_1^2 test applied to a 2-by-2 table of allele counts in the two groups. In complex traits it is commonly assumed that the contribution to the genetic effect from each SNP is roughly additive [17]. This test is powerful for additive models, where the penetrance for heterozygous are somewhere in between the penetrance for the two homozygotes, whereof the popularity of this test in these studies. Other common tests include a Pearson χ^2 test comparing the genotype frequencies instead of alleles, 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, evaluating the transmission of an allele from a heterozygous parent to the offspring. This test was introduced by Spielman et al. [18]. It 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 heteroxygous 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 has a genetic disease, we suspect that the allele is associated to the disease. The test statistic has the following form

$$T = \frac{(b-c)^2}{b+c}, \tag{2.1}$$

where b is the number of heterozygous parents who transmits allele M_1 to their offspring, and c is the number of heterozygous parents who transmits allele M_2. This test, which is sometimes called "McNemar Test", is equivalent to a Pearson χ^2-test,

$$\sum \frac{(O-E)^2}{E},$$

with $E = (b+c)/2$ and follows a χ^2_1-distribution. Let b_k and c_k be the corresponding counts for each *trio* - father, mother and affected offspring - then $b = \sum_{k=1}^{K} b_k$, $c = \sum_{k=1}^{K} c_k$ and K is the number of trios.

Chapter 3

Summary of Papers

3.1 Paper I

A common test in Genome Wide case-control associaton studies is the Pearson χ^2_1-test comparing allele frequencies among the two groups.

In coeliac disease 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, the same allele should 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 [19]. 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 [20] celiac disease case-control study and compare the result on genes which was replicated in further studies.

3.2 Paper II

In this applied paper we run a GWAS on a family material in Coeliac Disease. Since the the material was genotyped using two different platforms we use imputed genotypes to increase the sample size at genetic markers that was not observed in all subjects of the sample.

The first analysis is performed with the Transmission Disequilibrium test (TDT) defined in (2.1). The imputation algorithm do not account for the family structure of the sample, hence we will obtain positive probabilities for impossible combinations of genotypes in a family. For a biallelic marker, there are 15 possible trio combinations,we denote these by

$$\tau_i = \{F_i, M_i, C_i\}, i = 1, \dots 15,$$

where F_i, M_i and C_i denotes the genotypes of the members in the trio τ_i. Let k denote the studied trio with posterior probabilities $P_{F,k}, P_{M,k}$ and $P_{C,k}$ The probability for each of these can be calculated from the imputation probabilities as

$$P_k^T(\tau_i) = P_{F,k}(F_i)P_{M,k}(M_i)P_{C,k}(C_i). \tag{3.2}$$

If the imputation resulted in impossible trios, $\sum_i P_k^T(\tau_I) < 1$, we then normalize this posterior distribution by

$$P_k^*(\tau_i) = \begin{cases} \frac{P_k^T(\tau_i)}{\sum_{i=1}^{15} P_k^T(\tau_i)} & \text{if } \sum_{i=1}^{15} P_k^T(\tau_i) > c, \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

for some threshold c. Based on this posterion distribution P_k^* we calculate the expected counts b_k and c_k in each trio k,

$$\mathbf{E}_{P_k^*}[b_k] = \sum_{i=1}^{15} b(\tau_i) P_k^*(\tau_i), \qquad \mathbf{E}_{P_k^*}[c_k] = \sum_{i=1}^{15} c(\tau_i) P_k^*(\tau_i),$$

The expected counts will then replace the observed counts in (2.1) which then get the following form,

$$T_{imp} = \frac{\left(\sum_k \mathbf{E}_{P_k^*}[b_k - c_k]\right)^2}{\sum_k \mathbf{E}_{P_k^*}[b_k + c_k]} \tag{3.4}$$

and the test statistic follows the χ_1^2 distribution, like the standard TDT.

This method is applied to the the CD material with $c = 0.7$, as well as the standard TDT with a threshold of 0.95 for the imputation probabilities.

A two-locus interaction analysis based on the test in [21] was performed comparing the heterogeneity model with the full model (epistasis) using a likelihood ratio test with 4 degrees of freedom. The likelihood of the penetrances and allele frequencies,

$$L(f, p_A, p_B) = \prod_{i,j} \left(\frac{f_{ij} P(A_i B_j)}{\sum_{k,l} f_{kl} P(A_k B_l)} \right)^{n_{ij}} P(A_i B_j)^{m_{ij}},$$

where $P(A_i B_j) = h_{ij}(p_A, p_B)$. will not have a unique solution to the maximization problem if we do not fix the disease prevalence $\sum_{k,l} f_{kl} P(A_k B_l)$. The likelihood maximization is performed numerically using the software R.

This method was applied to the CD material, with one affected child from each family, on a set of genetic markers which was chosen based on the TDT analysis and previous results.

Bibliography

[1] Catherine Dube, Alaa Rostom, Richmond Sy, Ann Cranney, Navaaz Saloolee, Chantelle Garritty, Margaret Sampson, Li Zhang, Fatemeh Yazdi, Vasil Mamaladze, Irene Pan, Joanne Macneil, David Mack, Dilip Patel, and David Moher, "The Prevalence of Celiac Disease in Average-Risk and At-Risk Western European Populations: A Systematic Review," *GASTROENTEROLOGY*, vol. 128, no. S, pp. 57 – 67, 2005.

[2] Peter H.R. Green and Christohe Cellier, "Medical Progress: Celiac Disease," *N Engl Med*, vol. 357, pp. 1731 – 43, 2007.

[3] L. Greco, R Romino, and et al I. Coto, "The first large poulation based twin study of coeliac diease," *Gut*, vol. 50, pp. 624 – 8, 2002.

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

[5] Nan M. Laird and Christoph Lange, "Family-based designs in the age of large-scale gene-association studies," *Nature Reviews Genetics*, vol. 7, no. 5, pp. 385 – 394, May 2006.

[6] 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*, December 2007, pp. 1–4.

[7] 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*, vol. 461, no. 7265, pp. 747 – 753, October 2009.

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

[9] R.A. Fisher, "The correlation between relatives on the supposition of Mendelian inheritance," *Trans. R. Soc. Edin.*, vol. 52, pp. 399 – 433, 1918.

[10] John P. Rice Rosalind J. Neuman, "Two-locus models of disease," *Genetic Epidemiology*, vol. 9, no. 5, pp. 347–365, 2005.

[11] Neil Risch, "Linkage strategies for genetically complex traits. i. multilocus models.," *Am J Hum Genet.*, vol. 46, no. 2, pp. 222–228, 1990.

[12] 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*, December 2009, pp. 1–5.

[13] 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*, vol. 81, pp. 559 – 575, September 2007.

[14] 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*, vol. 5, no. 6, pp. e1000529, June 2009.

[15] "Mach," `http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html`.

[16] "Beagle," `http://faculty.washington.edu/browning/beagle/beagle.html`.

[17] David J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781 – 791, October 2006.

[18] 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.*, vol. 52, no. 3, pp. 506–516, 1993.

[19] P. Armitage, "Tests for linear trends in proportions and frequencies," *Biometrics*, vol. 11, no. 3, pp. 375–386, September 1955.

[20] Hunt et al., "Newly identified genetic risk variants for celiac disease related to the immune response," *Nature Genetics*, vol. 40, no. 4, pp. 395–402, 2008.

[21] F. Clerget-Darpoux S. Kotti, H. Bickeböller, "Strategy for detecting susceptibility genes with weak or no marginal effect," *Human Heredity*, vol. 63, pp. 85–92, 2007.