Thesis for the Degree of Doctor of Philosophy

Which genes are involved?
- Statistical planning and analysis of human genetic samples

Staffan Nilsson

CHALMERS | GÖTEBORGS UNIVERSITET

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, August 2001

Which genes are involved?
- Statistical planning and analysis of human genetic disease samples.
Staffan Nilsson
ISBN 91-7291-061-5

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 1744
ISSN 0346-718X

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46-(0)31-772 1000

Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2001

Abstract

This thesis consists of six papers related to the statistical planning and analysis of human genetic samples. The first four mainly concern statistical methodology, while the last two are applied papers in genetics and medicin. In paper A an alternative method of estimating the locus specific sibling relative risk $\lambda_{s(HLA)}$ is derived, assuming a multiplicative model in coeliac disease. The method is compared to the commonly used $\lambda_{s(HLA)} = 0.25/p_0$. In paper B the nonparametric mean test is modified by using weighted IBD scores based on model assumptions. Selective sampling is also considered. In paper C some TDT-like statistics for detecting linkage disequilibrium with nuclear families of one of more affected children are introduced and compared. In paper D methods for multiple inference in association analysis of many loci using a case-control design are investigated. The keys to the methods are closed tests and permutation of case-control labels. In paper E the candidate gene CTLA4 is showed to be associated to coeliac disease in a Scandinavian population. In paper F the incidence of CSF abnormalities in siblings of multiple sclerosis (MS) patients and unrelated controls are examined in order to identify a subphenotype that can be utilised in the search for disease genes.

Keywords: Statistical genetics, Linkage analysis, Association analysis, Linkage disequilibrium, Affected sibpair

2000 Mathematical subject classification: 92D10,62P10

Förord

Först vill jag rikta ett varmt kollektivt tack till min familj, vänner, medarbetare på Matematisk statistik och Klinisk genetik samt medförfattare till artiklarna i avhandlingen. Jag vill också tacka Institutet för tillämpad matematik (ITM) som bidragit med finansiering.

Sen vill jag individuellt tacka en handfull personer, som jag har haft särskilt stor hjälp och glädje av. Först och främst essen i mitt liv: Suzanne(♡), Kalle, Petter och Karin. Några kungar: Olle Nerman - handledare och trumfkung, Jan Wahlström, Tommy Martinsson och Mikael Knutsson. Några damer: Lena Samuelsson, Marita Olsson, Kajsa Larsson och Åsa Naluai. Till sist en knekt: Per-Olof Flodin, som byggde altan och cykelstall.

När jag för många år sen gick på gymnasiet, skulle man göra ett specialarbete i årskurs 3. En smula beräknande valde jag en lättvindig lösning; en diktsamling. Här återges nu titeldikten.

Ett ansikte förnöjsamt ler
en mor som på sin gosse ser
han blir nog någonting förmer
-Vad vill du bli för något? Läkare? Kamrer?

Ett ansikte av smuts och snor
en liten hand i ögat gnor
han muttrar buttert till sin mor
-Jag vill inte bli nånting när jag blir stor.

Föga anade jag då att mitt slutbetyg skulle komma att innehålla den pinsamt avslöjande raden:

Specialarbete: `Svenska, Jag vill inte bli nånting när jag blir stor.`

Genom åren har detta skapat en viss munterhet vid anställningsintervjuer. Denna munterhet vill jag gärna bjuda på.

Staffan Nilsson
Göteborg, augusti 2001

List of papers

The thesis consists of the following papers.

A: Nilsson S, Naluai Å, Gudjonsdottir AH, Ascher H, A new method for estimation of $\lambda_{\mathbf{S(nonHLA)}}$ and its relevance in celiac disease. (submitted)

B: Nilsson S, Model based sampling and weights in affected sib pair methods. (submitted)

C: Nilsson S, Association analysis with TDT in locus heterogeneity models.

D: Nilsson S, Permutation based multiple inference in association studies with case-control design.

E: Haghighi S, Andersen O, Rosengren L, Bergström T, Wahlström J, Nilsson S, Incidence of CSF abnormalities in siblings of multiple sclerosis patients and unrelated controls. *J Neurol*, 247(8):616-622.

F: Naluai Å, Nilsson S, Samuelsson L, Gudjonsdottir AH, Ascher H, Ek J, Hallberg B, Kristiansson B, Martinsson T, Nerman O, Sollid LM, Wahlström J, The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders. *J Invest Dermatol*, 116(5):750-754

Summary

Introduction to genetic concepts

The key to the development of all living species is the *DNA*. For humans it consists of 23 pairs of *chromosomes* that are long sequences of *nucleotides* of four different kinds. These are adenine (A), guanine (G), cytosine (C) and thymine (T), a four letter alphabet of life. The nucleotides are also called *bases* and all together the 23 chromosomes contains approximately 3 billions of bases.

Subsequences of a chromosome consist of functional units called *genes*. After the completion of the human genome project (Lander *et al.*, 2001),(Venter *et al.*, 2001) the number of genes are estimated to be around 30 000. In the DNA sequence of a gene the nucleotides are grouped into a more complex alphabet of consecutive triplets called *codons* that will through intermediate stages eventually translate to sequences of *amino acids* building up *proteins*.

During evolution genes are targets for *mutations* that are changes of bases in the sequence. This gives rise to variants of genes. The variants are called *alleles*. The variation of the genes can cause variation in functions. The variation in function can be the cause of a disease. A well defined physical position on a chromosome is called a *locus*. A pair of alleles from a chromosome pair at a specific locus is called a *genotype*. A segment of alleles on the same chromosome and from the same parental origin is called a *haplotype*.

Between the genes is junk DNA that do not code for proteins, but are still carried around from generation to generation. These non coding regions on a chromosome also have variations. Loci with known position and several alleles are called *markers*.

In each pair of chromosomes one originates from the mother, the *maternal*, and one from the father, the *paternal*. The maternal chromosome is an alternating sequence of the mothers maternal and paternal chromosome. The positions where the sequences switch are called crossovers. This mixing is done during the *meiosis* when the egg is formed. Similarly, in the male meiosis, when the sperm is formed, the paternal chromosome becomes an alternating sequence of the fathers maternal and paternal chromosomes.

The locations of the crossovers are seemingly random. A common simplified but useful way to model this randomness is to assume that the crossovers follow a Poisson process and that the starting state is maternal or paternal with equal probability. The crossover process of sperms and eggs have in re-

ality different probability laws. This is sometimes taken into account, but in this thesis we will use the common simplifying assumption of a single process for both sexes.

The main focus in this thesis is on gene localisation and two general approaches in finding them. These are *linkage analysis* and *association analysis*.

Linkage analysis

Two loci are said to be *linked* if they are physically linked in the sense that they are positioned on the same chromosome. Two alleles at different loci from the same parent are said to be recombined if one comes from grandfather and the other from grandmother. If the loci are on different chromosomes the recombination probability, called *recombination fraction* θ, is $1/2$ due to the independence between chromosomes in the meiosis. If they are located on the same chromosome, the recombination fraction is equal to the probability of an odd number of crossovers between them. In this case $\theta < 1/2$. The closer the loci are the smaller is θ.

The distance between two loci can be measured in the number of bases. Another measure used in linkage analysis is the *genetic distance* with the unit Morgan (M). It is defined as a distance between two points where the expected number of crossovers is one. Since usually small genetic distances are of interest, the most used unit is cM. By the additivity of expectations for sums of random variables the Morgan scale is additive. On each chromosome there exist genetic maps of loci with empirically known genetic distances between them. In the adopted Poisson crossover, essentially due to (Haldane, 1919), the recombination fraction θ between two loci at distance m Morgan is

$$\theta = \frac{1}{2}(1 - e^{-2m}) \tag{1}$$

In linkage analysis the aim is to map a possible disease locus by studying how markers have segregated in a pedigree in relation to the *phenotypes* (affection status) of the pedigree members. In a candidate approach markers are selected in the vicinity of some candidate loci or regions, while in a full genome screen entire sets of markers that are distributed on approximately equal distances all over the genome are used. In either case, the collected data consist of genotypes and phenotypes of family members of one or more pedigrees.

One way of formulating a hypothesis for testing linkage to a single marker locus is as

$$H_0 : \theta = \frac{1}{2} \text{ (no linkage) vs } H_1 : \theta < \frac{1}{2} \text{ (linkage)}$$

As a test statistic the likelihood ratio can be used

$$\Lambda = \frac{L_{\theta=\hat{\theta}}}{L_{\theta=\frac{1}{2}}} \tag{2}$$

For historical reasons a unit called LOD score is defined as $\log_{10} \Lambda$. In *parametric linkage analysis* (or LOD score analysis) an assumption of a genetic model is explicitly made. For a biallelic monogene disease the parameters (p, f_0, f_1, f_2) are specified, where $p =$ disease allele frequency, while $f_i = P(\text{affected}|i \text{ disease alleles})$ are the genotype *penetrances*. For example a dominant and fully penetrant disease has the parameters $(p, 0, 1, 1)$. This statistic is typically applied on large pedigrees where a dominant highly penetrant monogene disorder like Huntingtons chorea is segregating or on nuclear families where a recessive disease like cystic fibrosis is present in 25% of the children. There are variants called *multipoint lodscores* where the position of a potential disease locus somewhere on a chromosome are tested for using individual genotype data from many marker loci simultaneously.

Another approach to linkage testing is known as *affected sib pair analysis*. As the term indicates only parents and their affected children are used in the analysis. When the disease has low penetrance there can be lots of siblings that are not (yet) affected, although they share the risk genotype with their affected sib. The reason for this reduced penetrance is a lack of other components that are necessary for disease development such as age, environment or other genetic components.

When the lacking component is additional genes, large pedigrees are usually rare, but affected sib pairs and their parents can still be sufficiently common. These diseases are called *complex*, *polygene* or *multifactorial*, with no clear distinctions between the terms. By simply looking at the segregation pattern of the disease phenotype (*segregation analysis*) it is practically impossible to reliably guess the genetic model when several genes are involved.

In the absence of good guesses for the model parameters *nonparametric* methods have been developed. The basic idea is to count the number of alleles at a marker locus that are *identical by descent* (IBD) in a sib pair.

The outcomes are $0, 1, 2$ with probabilities $\boldsymbol{p} = (p_0, p_1, p_2) = (1/4, 1/2, 1/4)$ and expected value $\mu = 1$ when there is no linkage.

The hypothesis is formulated as

$$H_0 : \boldsymbol{p} = (1/4, 1/2, 1/4) \text{ (no linkage) vs } H_1 : \boldsymbol{p} \neq (1/4, 1/2, 1/4) \text{ (linkage)}$$

which is tested with a likelihood ratio $L_{\hat{\boldsymbol{p}}}/L_{(1/4,1/2,1/4)}$ or as

$$H_0 : \mu = 1 \text{ (no linkage) vs } H_1 : \mu > 1 \text{ (linkage)}$$

which is tested with a sum of ibd counts.

Most nonparametric methods are more or less sophisticated variants of these two basic test statistics.

The resolution for mapping disease genes into genomic regions by linkage analysis is limited to something like 10-30 cM, since for the typical sample sizes used in linkage analysis there are no big difference between the statistical linkage signal of a genetic marker in the disease gene compared to that of a marker located at a distance of 10 cM (on the average $1/10$ crossover per meios).

Association analysis

A phenomena related to linkage is called *linkage disequilibrium* (LD). The first disease mutation in a given population existed on a background haplotype in that populations ancient disease founder, which either migrated into the population or is the person where the mutation occurred. The alleles on closely linked loci ($<$ 1cM), will continue to follow the disease mutation on the same haplotype through many generations of meiosis. As a result the haplotype frequencies will differ from the expected frequencies under the assumption of independence between the alleles at the disease and marker loci. The discrepancy is measured by the linkage disequilibrium coefficient.

$$D_{ij} = h_{ij} - a_i b_j \tag{3}$$

where a_i is the frequency of allele i at the disease locus A, b_j is the frequency of allele j at the marker locus B and h_{ij} is the frequency of haplotypes with allele i in haplotype with allele j.

The disequilibrium will decay as time goes by, due to recombinations between the loci. The more distant the loci are, the faster is the decay. Thus

the candidate region that was achieved by the linkage analysis is narrowed down to something like a 1-3 cM region.

The amount of LD is unfortunately randomly varying and it depends not only on the genetic distance, the marker allele frequencies and the ancestral pedigree of the common disease allele, but also on single recombination events and which marker allele that happened to be on the ancestral mutant haplotype.

In a case control study unrelated patients and controls are compared with respect to genotype or allele frequencies to test if there is LD between a marker allele and a disease allele. Care must be taken to ensure that cases and controls are from the same population so the marker frequencies does not differ for any other reason than LD.

In family based association studies the cases genotypes are compared with the genotypes of the (hypothetical) complementary children's genotypes, which are formed by combining the mothers and fathers nontransmitted alleles. The currently most used test is the transmission disequilibrium test (TDT) (Spielman *et al.*, 1993), where only the transmissions from heterozygous parents are recorded and used in a test. Under the null hypothesis a heterozygous parent with genotype 12 will transmit allele 1 or 2 with equal probability $\frac{1}{2}$ to an affected child. Families with just one affected child, called trios, can be collected in addition to the pedigrees and/or sib pair families that are used for linkage.

Further reading

Some textbooks in human genetics (Strachan and Read, 1996) and (Haines and Pericak-Vance, 1998) are recommended and for more detailed statistical aspects (Sham, 1998), (Ott, 1999) and (Balding *et al.*, 2000). An excellent and amusing review paper, still with deep insight is (Terwilliger and Göring, 2000).

Summary of paper A

A disease is hereditary when an observed familiar clustering is not only due to the affected family members common social and physical environment. One general measure of the familiar clustering is the relative risk for siblings $\lambda_S = K_S/K_P$, where K_P is the population prevalence and K_S is the sibling

prevalence. Similar measures exist for other kind of relatives. A necessary criteria for a hereditary disease is thus that $\lambda_S > 1$

For a monogene disorder the sibling relative risk λ_S can be estimated by $0.25/p_0$, where p_0 is the probability of sharing no haplotypes identical by descent (IBD) at the disease locus. In a multiplicative genetic model with two unlinked, independently penetrating genetic components A and B that are both necessary for disease development $\lambda_S = \lambda_{S(A)} * \lambda_{S(B)}$. When A is a single locus $\lambda_{S(A)} = 0.25/p_0$. Here p_0 is the probability of no allele sharing at locus A. All these facts are shown in (Risch, 1987).

In celiac disease (gluten intolerance) a multiplicative model is generally assumed and A corresponds to two known close consecutive loci in the HLA-region on the human chromosome 6. Several evidences are available that indicate the existence of a *non-HLA* component. In planning a genome scan with an affected sib pair design a suitable size of the study depends largely on the size of $\lambda_{S(nonHLA)}$. From the results above it follows that it could be estimated by $\lambda_S/\lambda_{S(HLA)}$, but that requires that at least a fair number of sib pairs are already collected and of course the multiplicative assumption.

In contrast we derive in the paper a method to estimate $\lambda_{S(HLA)}$ from more easily collected samples.

Let G denote the event that an individual has a certain genotype or class of genotypes with equal penetrances. The derived formula is

$$\lambda_{S(HLA)} = \frac{P(G|\text{aff sib})P(G|\text{affected})}{P(G|\text{affected, aff sib})P(G)} \tag{4}$$

and it is compared with Rischs

$$\lambda_{S(HLA)} = \frac{0.25}{p_0}. \tag{5}$$

We can estimate $P(G)$ from controls and $P(G|\text{affected})$ from cases. These are easily collected in large numbers. $P(G|\text{aff sib})$ can be estimated by use of *all* siblings of cases. $P(G|\text{affected, aff sib})$ is the only quantity that requires samples of sib pairs, but it is a *large* number as opposed p_0. Furthermore we need to collect and genotype parents as well for estimation of p_0.

Finally the power of a planned full genome screen of 100 affected sib pairs is calculated under some different complex models, based on the estimation of $\lambda_{S(nonHLA)}$.

Summary of paper B

When the power is calculated for a genome scan with a non parametric statistic based on IBD counts, we show that although only the IBD counts for the affected sibs are included, the affection status of the other sibs and parents have a substantial impact on the power. The reason is that the expected deviation from the null hypothesis mean of 1 IBD per sib pair varies a lot conditional on family size and affection structure.

We utilize these differences by two strategies: weighting and selective sampling. The considered statistic is

$$T = \sum_i w_i Z_i.$$

where w_i is the weight for family i and

$$Z_i = \frac{X_i - \binom{a_i}{2}}{\sqrt{\binom{a_i}{2}\frac{1}{2}}},$$

is a normalisation of X_i, which is the sum of IBD count of all ASP in family i where a_i children are affected.

Some different weighting schemes are designed that gives T maximum power (these weights depend on significance level), maximum asymptotic power or maximum expectation. The power of these test statistics are compared with the power of tests with equal weights on the normalised family scores or equal weights on each sib pair in the total sample.

We also investigate the strategy of selecting more informative families and indicate that this strategy, although good on its own, should be combined with weighting.

The maximum expectation weights are found to be easy to implement, powerful and robust against misspecification of genetic model. They can be made even more robust by averaging them over some different models.

Every now and then there is a debate in the literature about the choice between parametric and nonparametric methods for linkage analysis. This paper advocates a compromise.

Summary of paper C

To be able to use ASP families for TDT analysis when the null hypothesis is "no linkage disequilibrium" a valid solution is to randomly select one af-

fected child from each family. To avoid discarding information randomly we derive some TDT inspired tests that stratifies the families based on the IBD configuration. The underlying idea is that only transmissions from those heterozygous parents that transmitted the same allele to both affected children are of interest.

The ASP families are classified according to IBD sharing. The families with 0 IBD are discarded for the reason given above. Then the IBD transmissions from heterozygous parents in 1 resp 2 IBD families form two different classes and the transmissions from parents in trios, i.e. a single affected child and two parents, form a third class.

The transmission distortion in these different categories are generally rather different implying that a simple merging of the data is not recommended. The transmissions in trios are less informative in models where a substantial amount of the cases are phenocopies or when locus heterogeneity is present.

We investigate how the three classes of data can be combined by considering the model space of the alternative hypothesis. The resulting statistics are either fix weighted mixtures of the three classes or of order regression type.

From the simulations of some phenocopy models and locus heterogeneity models it turns out that the fix weights that are implicitly used in the statistic T_{sp} (Martin *et al.*, 1997) are competitive in the models covered in the study.

Summary of paper D

The paper describes how various methods for multiple inference can be used in genetic association analysis with a case-control design. The data are genotype information on many SNP markers in cases and controls where each individual locus is tested for equality of genotype or allele frequency in the two groups.

We show how substantial improvement over the simple Bonferroni method can be achieved. The most important components in the test design is

1. The closed test procedure (Marcus *et al.*, 1976).

2. Using minP, the minimum of p-values, as test statistic for the the hypothesis intersections in the closed test.

3. Evaluating the p-value of minP by permutation of case-control labels.

When only the first two components are used and the critical values for the minP statistic are set by Booles inequality, the beautiful Holm-Bonferroni method (Holm, 1979) follows.

Holm-Bonferroni is an excellent method when the individual tests are almost independent, but when they are correlated, as they typically are when close markers are used, it is outperformed by the permutation strategy. The use of a permutation method for p-value estimation is described in general in (Westfall and Young, 1993).

Another type of dependency between tests comes from using several slightly different statistics on the same sample. In genetic case - control studies the counts of the three genotypes that are possible with biallelic markers can be collapsed by pooling the heterozygotes with either of the two homozygotes or by just counting the alleles. Homogeneity tests are then performed on the three different collapsed 2 by 2 contingency tables. These tests are obviously dependent, which can be accounted for by the permutation strategy.

Since the hypotheses corresponding to the three different tests are all equal, there is a very simple logical constraint that can be utilized in the multiple inference framwork (Shaffer, 1986). Whenever one null hypothesis is false, both the other are false as well. This reduces the test multiplicity and we get a modest - but still - improvement over the unrestricted method.

When loci are tight it might or might not be better to study haplotypes of various lengths. This does of course immediately increase the number of tests, but once again, the dependencies between the tests are accounted for by permutation and also some logical constraints between the test statistics can be utilized.

Finally we indicate the possibility to distribute the familywise error rate unevenly by weighting the individual p-values. The basis for the weighting should be prior knowledge and that prior knowledge can even be achieved by studying the pooled sample.

Summary of paper E

The cytotoxic T-lymphocyte associated (CTLA4) gene has been investigated in several autoimmune diseases as a candidate gene due to its involvement in T-cell activation and profileration.

We have performed a candidate approach for celiac disease using five microsatellite markers surrounding CTLA4 and one SNP (+49 A/G) in exon

1 on a material of 104 Swedish and Norwegian families, where at least two of the children have celiac disease (CD). In a French population the A allele has been found to associate with CD (Djilali-Saiah *et al.*, 1998). We confirm this result with a TDT test. There were 114 transmitted versus 79 non-transmitted A-alleles among the heterozygous parents. Using a one-sided binomial test this gives a p-value of 0.007 for preferential transmission of the A-allele.

Interestingly the other autoimmune diseases that have been studied report association with the G-allele. Most striking is the French IDDM study (Marron *et al.*, 1997), since it is the same poulation that shows association to the A-allele in celiac disease. Thus either CTLA4 plays a different role in celiac disease as opposed to other autoimmune diseases or, if the association is due to LD, the CD susceptibility gene must differ from the other whether it is an allelic difference at another loci or two different genes.

Summary of paper F

MS patients are known to have increased levels of measles antibodies in serum and cerebral spinal fluid (CSF) as well as encriched oligoclonal bands (OCB) in CSF. This partially hyperimmune phenotype was termed "MS-trait" in (Poser, 1993) It has been found that healthy siblings of MS patients share these conditions to various extent.

We investigate these three conditions in 47 pairs of MS patients/healthy sibling and in 50 unrelated controls.

To measure the measles antibodies the samples were incubated in twofold dilutions starting from 1/10 for CSF and 1/100 for serum. Titers were given as the highest dilution giving positive reaction.

One sided tests were performed comparing patients vs sibs and sibs vs controls. Nonparametric methods were used since the data is highly skewed and downwards truncated. Previous studies have reported higher measles titers in women, so the two sexes are compared separately and then combined into a single test.

Some technical details, that are not given in the paper, of the statistics used in comparing the titers are given below. For comparing sibs with controls a Mann-Whitney tests is performed for each sex. The two test statistics

W_1 (men) and W_2 (women) are combined by

$$Z = a_1 \frac{W_1 - E[W_1]}{\sqrt{VAR[W_1]}} + a_2 \frac{W_2 - E[W_2]}{\sqrt{VAR[W_2]}} \tag{6}$$

which is asymptotically normal for all choice of weights a_1, a_2 such that $a_1^2 + a_2^2 = 1$. The weights have been choosen to maximize Z when W_1 and W_2 reach there maximum.

For comparing patients with their sibs we divide the pairs in three different groups:

1. pairs of the same sex
2. MS patient male, sibling female
3. MS patient female, sibling male

In the first group the Wilcoxon signed rank test is used on the differences between the siblings. In the second and third groups the differences between the female data and the male data is formed and then a Mann-Whitney test is used on the resulting two groups of differences.

The two different test statistics are finally combined as above.

A general trend of more measles antibodies in both serum and CSF as well as more OCB in the direction controls<sibs<patients is found. The sibs with two or more OCB also have more antibodies than the sibs with one or less.

In the search for the genetic causes of MS the most important aspect of the paper is that it defines a sub phenotype that is more common than MS itself and that can be assumed to have a more simple genetic model, perhaps an autosomal dominant gene. This will give a possibility to use parametric linkage analysis to locate that gene.

References

Balding, D., Bishop, M., and Cannings, C., editors (2000). *Satistical Genetics*. John Wiley & sons.

Djilali-Saiah, I., Schmitz, J., Harfouch-Hammoud, E., Mougenot, J., Bach, J., and Caillat-Zucman, S. (1998). Ctla-4 gene polymorphism is associated with predisposition tocoeliac disease. *Gut*, **43**.

Haines, J. and Pericak-Vance, M., editors (1998). *Approaches to gene mapping in complex human diseases*. John Wiley & sons.

Haldane, J. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *J Genet*, **8**, 299–309.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Statist*, **6**, 65–70.

Lander, E. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Marron, M., Raffel, L., HJ, G., *et al.* (1997). Insulin-dependent diabetes mellitus (iddm) is associated withctla4 polymorphisms in multiple ethnic groups. *Hum Mol Genet*, **6**, 1275–1282.

Martin, E., NL, K., and Weir, B. (1997). Tests for linkage and association in nuclear families. *Am J Hum Genet*, **61**, 439–448.

Ott, J. (1999). *Analysis of Human Genetic Linkage*. John Hopkins.

Poser, C. (1993). The pathogenesis of multiple sclerosis. additional considerations. *J Neurol Sci*, **115**, 3–15.

Risch, N. (1987). Assessing the role of hla-linked and unlinked determinants of disease. *Am J Hum Genet*, **40**, 1–14.

Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**, 826–831.

Sham, P. (1998). *Statistics in Human Genetics*. Arnold.

Spielman, R., McGinnis, R., and Ewens, W. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, **52**, 506–516.

Strachan, T. and Read, A. (1996). *Human molecular geneticsStatistical Inference under Order Restrictions*. BIOS Scientific Publishers Ltd.

Terwilliger, J. and Göring, H. (2000). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology*, **72**, 63–132.

Venter, J. *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.

Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing: examples and methods for p-value adjustment*. John Wiley and Sons.

A new method for estimation of $\lambda_{\mathbf{S(nonHLA)}}$ and its relevance in celiac disease

Staffan Nilsson[1], Åsa Naluai[2], Audur Gudjonsdottir[3] and Henry Ascher[3]
[1]Chalmers University of Technology, [2]Department of Clinical Genetics, Göteborg University, Sahlgrenska University Hospital/Östra and [3]The Queen Silvia Children's Hospital, Göteborg, Sweden

30 January 2001

Abstract

We address the question if it is possible to find other genes than the known coeliac disease HLA-component with a linkage analysis of 100 affected sib pair (ASP) families. The siblings relative risk, λ_S, is a strong determinant for the power in linkage studies. Assuming a multiplicative model, Risch has shown that the relative risk at HLA for siblings can be estimated by $0.25/p_0$, where p_0 is the probability of two affected sibs sharing no haplotypes identical by descent (IBD) at HLA. From this the relative risk for a nonHLA component can be determined. We derive an alternative way of estimating the sibling relative risk for the nonHLA component, when not only linkage to HLA, but also the disease alleles are known. Applied on coeliac disease, we get $\lambda_{S(nonHLA)} = 9$ with our method and $\lambda_{S(nonHLA)} = 4$ with Risch's method. This difference has a substantial impact in the power calculations.

Introduction

Celiac disease is defined as a permanent intolerance to dietary gluten, i.e. wheat, rye, barley and possibly oats. Ingestion of these products causes an

HLA class II driven autoimmune reaction leading to intestinal lesions characterised by crypt hyperplasia, villous atrophy and infiltration of intraepithelial lymphocytes (Mäki and Collin, 1997) The disease is unique since the trigger factor is known and a complete remission is seen if this factor is withdrawn. An autoantigen, tissue transglutaminase (tTG), has been found (Schuppan *et al.*, 1998). Gluten is preferentially bound to tTG. Further, tTG is modifying gluten peptides by deamidation increasing its affinity to the HLA class II molecules on intestinal antigen presenting cells (Sollid, 2000). HLA class II genes seems to be a prerequisite for disease development.

The HLA alleles DQA1*0501 and DQB1*02 are present in at least 90% of the patients, either in DR3-DQ2 positive individuals where the alleles are located in cis-position or, less often, in DR5-DQ7/DR7-DQ2 heterozygous individuals where the alleles are located in trans-position. These alleles are encoding the DQ(a1*0501, b1*02) heterodimer. Almost all of the remaining patients carry the DQ(a1*0301, b1*0302) heterodimer in DR4-DQ8 positive individuals. Interestingly, the heterodimers in DQ2 and DQ8 positive individual have a strong structural similarity. However these genes are common in Caucasian populations and less than 1% of the carriers develop celiac disease. In monozygotic twins there is a concordance of at least 80% compared with app. 10% in HLA identical siblings (Sollid and Thorsby, 1993). All these facts indicate that other genes are involved as well.

In the planning of a full genome scan of affected sib pair families we have calculated the power to detect a nonHLA component under a number of different models. To make the models realistic they were all constrained to fit the same relative risk, $\lambda_s = K_S/K_P$, of the disease, where K_P is the population prevalence and K_S is the sibling prevalence. The size of λ_S will influence the power to detect disease loci in a genome scan and specifically the component relative risk, $\lambda_{s(nonHLA)}$ (Risch, 1990), will influence the power to detect nonHLA disease genes.

Material

For the genome scan we have access to 107 nuclear families, where at least two children are affected by celiac disease. The affected children are diagnosed according to the ESPHGHAN criteria (Walker-Smith, 1990). The families were collected from Sweden and south Norway.

Figures on sibling prevalence is based on a family study including all 164

siblings in 96 families with 97 probands with celiac disease fulfilling strict diagnostic criteria (Meeuwisse, 1970),(Ascher *et al.*, 1997). A diagnostic intestinal biopsy was performed in all siblings with HLA DR3-DQ2/DR3-DQ2, DR3-DQ2/x (where x are configurations other than DR3-DQ2), DR5/DR7-DQ2 or who had positive serological markers for celiac disease.

Derivation

In order to estimate $\lambda_{S(nonHLA)}$ we need to model the interaction between the HLA component and the nonHLA component. Since we assume that all patients carry risk alleles in HLA, we adopt a multiplicative model (Risch, 1990).

A multiplicative model has the property that if we have two loci, A and B, with genotypes G_{Ai} and G_{Bj}, where f_{ij} is the penetrance for the composite genotype, then for all (i,j), the penetrances have a structure $f_{ij} = r_i s_j$, r_i and s_j being penetrance factors. We define $K_{P(A)} = \sum r_i P(G_i)$ which we can think of as the population prevalence for an unobserved "A-disease" caused by locus A. With this interpretation in mind we consequently define $K_{S(A)} = P(A \text{ affected}|A \text{ affected sib})$ and similar for $K_{P(B)}, K_{S(B)}$.

With these assumptions and definitions, the model can thus be viewed as if celiac disease is the result of having two independent diseases caused by an HLA component and a nonHLA component.

If we further assume that the nonHLA component is not linked to HLA, then

$$\lambda_S = \frac{K_S}{K_P} = \frac{K_{S(HLA)} K_{S(nonHLA)}}{K_{P(HLA)} K_{P(nonHLA)}} = \lambda_{S(HLA)} \lambda_{S(nonHLA)}.$$

With a multiplicative model it was derived in (Risch, 1987)

$$\lambda_{S(HLA)} = \frac{0.25}{p_0}, \tag{1}$$

where

$$p_0 = P(\text{sharing 0 HLA-haplotypes}|\text{ASP})$$

can be estimated by genotyping markers at HLA in the nuclear families.

This gives finally

$$\lambda_{S(nonHLA)} = \frac{p_0 K_S}{0.25 K_P}, \tag{2}$$

where K_S and K_P are estimated from population genetic data and p_0 can be estimated by observing the IBD-sharing at HLA in the family material.

We can roughly think of the estimate of p_0 as a simple relative frequency. In practice problem arises with partially observable meioses and even with fully informative meioses, a maximum likelihood estimation in Holmans triangle (Holmans, 1993) would be an alternative to the relative frequency.

Estimating p_0 with accuracy will require large samples of sib pair families and genotyping at HLA. The phenotype "HLA affected" is unobservable and merely an abstract concept, so we can not directly estimate $K_{S(HLA)}$ and $K_{P(HLA)}$ from population data. When we know not only the location of the disease locus, but also the alleles involved (in our case DQA1*0501,DQB1*02) (Sollid *et al.*, 1989) we could perhaps calculate $K_{S(HLA)}$ and $K_{P(HLA)}$ directly from definitions. For instance

$$K_{P(HLA)} = \sum_i r_i P(G_i) \tag{3}$$

where r_i is the penetrance of the disease genotype G_i. However only one genotype is known and its penetrance unknown. Still, by using the knowledge of one of the risk genotypes, it is possible to derive an alternative formula.

Let G be a class of genotypes, in our case at least one DQA1*0501 and one DQB1*02, that are all sharing the same penetrance $f = P(\text{HLA affected}|G)$, Then

$$\begin{aligned}
\lambda_{S(nonHLA)} &= \frac{fK_{S(nonHLA)}}{fK_{P(nonHLA)}} \\
&= \frac{P(\text{HLA aff}|G)P(\text{nonHLA aff}|\text{nonHLA aff sib})}{P(\text{HLA aff}|G)P(\text{nonHLA aff})} \\
&= \frac{P(\text{aff}|G, \text{nonHLA aff sib})}{P(\text{aff}|G)} \qquad (4) \\
&= \frac{P(\text{aff}|G, \text{aff sib})P(G)}{P(G|\text{aff})K_p}. \qquad (5)
\end{aligned}$$

In equality (4) the independence of "nonHLA aff" and the HLA-genotype class G together with the multiplicative model is used in both the numerator and the denominator.

In equality (5) Bayes formula is applied on the denominator. For the numerator the common assumption that the affection status depends on a siblings affection status only through the genotype is used.

We estimated the different factors in (5) from results in a previous study (Ploski *et al.*, 1996) to get an estimate of $\lambda_{S(nonHLA)}$.

To compare the two methods we can from the expression (5) derive

$$\lambda_{S(nonHLA)} = \frac{P(G|\text{affected, aff sib})P(G)K_S}{P(G|\text{aff sib})P(G|\text{affected})K_P}. \tag{6}$$

The expression (6) has K_S and K_P in common with (2). Note that we have implicitly shown

$$\lambda_{S(HLA)} = \frac{P(G|\text{aff sib})P(G|\text{affected})}{P(G|\text{affected, aff sib})P(G)}. \tag{7}$$

The fact that the known genotype G is fairly common in the population and very common in affecteds implies that all quantities $P(G)$, $P(G|\text{affected})$, $P(G|\text{aff sib})$ and $P(G|\text{affected, aff sib})$ are large compared to p_0 resulting in less variance of the estimate.

There will be no need to genotype the parents for this estimation, but the population frequency of G is used. This is normally already estimated if G is really verified to be disease causing.

It would be possible though to avoid estimating $P(G)$. Formula (7) is true for any G as long as there is a risk associated with it. If all genotype classes G_i are at risk we could calculate a separate estimate $\lambda^{(i)}_{S(HLA)}$ for each of them and then as a total estimate take a weighted sum of the separate estimates, with $P(G_i)$ used as weights,

$$\begin{aligned}\lambda_{S(HLA)} &= \sum_i \lambda^{(i)}_{S(HLA)} P(G_i) \\ &= \sum_i \frac{P(G_i|\text{aff sib})P(G_i|\text{affected})}{P(G_i|\text{affected, aff sib})}.\end{aligned}$$

If there is a class G_0 with no risk, we can use the weights $\frac{P(G_i)}{1-P(G_0)}$. In this way we only need to estimate $P(G_0)$.

Power calculation

Using the sib pair family material and microsatellite markers for HLA, D6S273, D6S1019 and TNFa gives an estimate of $p_0 = 0.05$ calculated by GeneHunter 2.0 (Kruklyak *et al.*, 1996). From data in (Ascher *et al.*, 1997) we reestimated

$K_S = 0.08$ and from a Swedish population study (Cavell *et al.*, 1992) we got $K_P = 0.004$ resulting in $\lambda_{s(nonHLA)} = 4$ by Risch method.

Other studies that have estimated p_0 are (Greco *et al.*, 1998) that found $p_0 = 0.09$ and (Petronzelli *et al.*, 1997) with $p_0 = 0.07$.

Using the data in (Ploski *et al.*, 1996) we were able to estimate $P(G) = 0.22$, $P(G|\text{affected}) = 0.92$ and $P(\text{affected}|G, \text{affected sib}) = 0.16$ resulting in $\lambda_{S(nonHLA)} = 9$.

To get some idea of the power to find the nonHLA components we chose to study a medium complex model. Heterogeneity was assumed between two dominant loci, i.e. each of the two disease genotypes are, together with the necessary HLA component, sufficient to cause the disease. Even with this rather specific model there are infinitely many sub models that fits with the same $\lambda_{S(nonHLA)}$. We varied the frequencies of the disease allele at the two loci from equal to more skewed (table 1). The penetrances were arbitrarily set to 1, but the impact on the power is negligible as long as the penetrances are equal at the two loci. The significance levels were varied between suggestive (0.00074) and significant (0.000022) in accordance with (Lander and Kruglyak, 1995).

Table 1: *Power to detect linkage for various models*

$\lambda_{S(nonHLA)}$	disease allele freq		α	power		
	locus 1	locus 2		locus 1	locus 2	any locus
4	0,061	0,012	0,000740	0,94	0,01	0,94
4	0,061	0,012	0,000022	0,69	0,00	0,69
4	0,049	0,025	0,000740	0,69	0,08	0,72
4	0,049	0,025	0,000022	0,32	0,01	0,33
4	0,037	0,037	0,000740	0,32	0,32	0,55
4	0,037	0,037	0,000022	0,08	0,08	0,15
9	0,025	0,005	0,000740	0,99	0,02	0,99
9	0,025	0,005	0,000022	0,92	0,00	0,92
9	0,020	0,010	0,000740	0,88	0,14	0,90
9	0,020	0,010	0,000022	0,57	0,02	0,58
9	0,015	0,015	0,000740	0,50	0,50	0,77
9	0,015	0,015	0,000022	0,17	0,17	0,32

Discussion

We have compared two methods of estimating $\lambda_{S(nonHLA)}$. With the method from (Risch, 1987) no knowledge of the disease susceptibility alleles at HLA are necessary. Thus it can be used directly in the same genome scan that discovers linkage, although the argument for a multiplicative model must be much weaker when the gene is not yet found. In celiac disease this knowledge is however available and we use it to derive an alternative estimator. Since the methods require different kind of samples, it is not straightforward to compare them, but in general it is a good principle to use all available knowledge.

The main purpose of estimating $\lambda_{S(nonHLA)}$ in the study was to be able to perform power calculations for a genome scan. Risch's method requires that we collect a quite large preliminary sample in order to be able to estimate how many that are needed in the final sample, with the risk of having collected the preliminary sample in vain, when the amount of affected sib pairs needed to achieve reasonable power is not possible to reach. In contrast, with our method (5) we can estimate $\lambda_{S(nonHLA)}$ before we collect a large amount of ASP. The only ASP we need are those we find when we estimate the entity P(affected|G,aff sib).

When applying the two methods we got the estimate $\lambda_{S(nonHLA)} = 4$ using Risch's (2) and $\lambda_{S(nonHLA)} = 9$ using our new (5). The difference can be explained by chance, but in a recent study (Lie *et al.*, 1999) an indication of additional susceptibility modifiers in the HLA-region was found. Risch's method still gives an unbiased estimate, but our method would typically underestimate $\lambda_{S(HLA)}$ and thereby overestimate $\lambda_{S(nonHLA)}$, which fits with the direction of the observed difference.

Although the power calculations are made on several idealistic assumptions, but also on rather conservative significance values, the difference in power between the different estimates of $\lambda_{S(nonHLA)}$ illustrates the impact of its size and thus the importance of an accurate estimate.

References

Ascher, H., Krantz, I., Rydberg, L., Nordin, P., and Kristiansson, B. (1997). Influence of infant feeding and gluten intake in coeliac disease. *Arch Dis Child*, **76**, 113–117.

Cavell, B., Stenhammar, L., *et al.* (1992). Increasing incidence of childhood coeliac disease in sweden. *Acta Paediatr*, **81**, 589–592.

Greco, L. *et al.* (1998). Genome search in celiac disease. *Am J Hum Genet*, **62**, 669–675.

Holmans, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *Am J Hum Genet*, **52**, 362–374.

Kruklyak, L., Daly, M., Reeve-Daly, M., and Lander, E. (1996). Parametric and nonparametrice linkage analysis: A unified multipoint approach. *Am J Hum Genet*, **58**, 1347–1363.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting results. *Nature Genetics*, **11**, 241–247.

Lie, B., Sollid, L., Ascher, H., Ek, J., Akselsen, H., Ronningen, K., Thorsby, E., and Undlien, D. (1999). A gene telomeric of the hla class i region is involved in predisposition to both type 1 diabetes and coeliac disease. *Tissue Antigens*, **54**, 162–168.

Meeuwisse, G. (1970). Diagnostic criteria in coeliac disease. *Acta Paediatr Scand*, **59**, 461–463.

Mäki, M. and Collin, P. (1997). Coeliac disease [see comments]. *Lancet*, **349**, 1755–1759.

Petronzelli *et al.* (1997). Genetic contribution of the hla region to familial clustering of coeliac disease. *Ann hum Genet*, **61**, 307–317.

Ploski, R., Ascher, H., and Sollid, L. (1996). Hla genotypes and the increased incidence of coeliac disease in sweden. *Scand J Gastroentero*, **31**, 1092–1097.

Risch, N. (1987). Assessing the role of hla-linked and unlinked determinants of disease. *Am J Hum Genet*, **40**, 1–14.

Risch, N. (1990). Linkage strategies for genetically complex traits. i. multilocus models. *Am J Hum Genet*, **46**, 222–228.

Schuppan, D., Dieterich, W., Ehnis, T., Bauer, M., Donner, P., Volta, U., *et al.* (1998). Identification of the autoantigen of celiac disease. *Ann N Y Acad Sci*, **859**, 121–126.

Sollid, L. (2000). Molecular basis of celiac disease. *Annu Rev Immunol*, **18**, 53–81.

Sollid, L. and Thorsby, E. (1993). Hla susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology*, **105**, 910–922.

Sollid, L., Markussen, G., Ek, J., Gjerde, H., Vartdal, F., and Thorsby, E. (1989). Evidence for a primary association of celiac disease to a particular hla-dq α/β heterodimer. *J Exp Med*, **169**, 345–350.

Walker-Smith, J. (1990). Report of working group of european society of paediatric gastroenterology and nutrition. *Arch Dis Child*, **65**, 909–911.

Model based sampling and weights in affected sib pair methods

Staffan Nilsson
Chalmers University of Technology

30 January 2001

Abstract

When running a genome scan with affected sib pairs and a non-parametric statistic in the form of sums of IBD counts, it turns out that, depending on genetic model, some families are more useful than others if we consider the phenotypes of all family members. This can be utilised by performing selective sampling and/or putting weights on the IBD counts. Some different approaches are introduced and compared on simulated examples.

Introduction

In the planning of a genome scan with nuclear families we can get a rough idea about a suitable sample size for an Affected Sib Pair (ASP) statistic by performing power calculations. Later on, when the sample has been collected, the power calculation can be made more precise by taking into account the fact that the expected contribution to the over all sib pair statistic varies between sib pairs from different families due to differences in family size and affection status of family members.

The fact that the contribution is different implies that some sib pairs will be better to use than others. This observation can be utilised in two ways to achieve higher power. We can take a selective sample (McCarthy *et al.*, 1998) by preferring 'good' sib pairs and we can use different weights (Sham *et al.*, 1997) for different sib pairs.

The two approaches can of course also be combined. It is perhaps easiest to illustrate the idea by giving examples of 'bad' families.

- **Phenocopies**

  If the phenocopy rate is high, as it is in e.g. breast cancer, an affected pair from a nuclear family, where all the parents and other siblings are healthy, will often be a pair of phenocopies. In this case the segregation at the risk locus will be random.

- **Gene overdose**

  Think of a simple monogene dominant disease. A family with many children, where all have the disease, is most likely the result of one of the parents being homozygous for the disease allele. The segregation at the disease locus is random from this homozygous parent. Since it requires both a large family and a homozygous parent, it will be a problem of practical importance only for common disease alleles in populations where the family sizes are large.

We can identify three separate issues that influence the usefulness of a particular family.

- **Number of affected siblings**

  With a affected siblings in a sibship we have $\binom{a}{2}$ affected sib pairs in a family. The corresponding $\binom{a}{2}$ IBD counts will not be independent. This single issue has been considered in (Suarez and Hodge, 1979; Hodge, 1984) among others.

- **Number of unaffected siblings**

  Although the unaffected siblings are not genotyped, their very existence influence the power and can be considered as input in designing weights or as input to a decision to select or reject the family. In practice we have to consider the cost of phenotyping the siblings with unknown affection status as a part of the whole strategy.

- **Affection status of parents**

  The parents affection status will effect the power and can thus be taken into account as well. Alternatively it can be used more explicitly as is done in (Whittemore and Halpern, 1994).

In (Sham *et al.*, 1997) the first two issues are used to assign weights that gives optimal power under a certain model and this weighting scheme is compared with some other proposed weighting schemes.

In (McCarthy *et al.*, 1998) all three issues are used to compare the power of different pedigree structures, with the purpose of selectively collect the more powerful pedigrees. The study does however not consider selecting more than one sib-pair from a multi affected pedigree.

This study extends (Sham *et al.*, 1997) by considering the affection status of the parents and by including the selective sampling strategy as well. We also introduce some other weighting schemes besides the optimal power scheme.

We will only compare different weights and sampling strategies for IBD sharing statistics and not cover the question whether an ordinary likelihood based statistic would be better.

Derivations of IBD expectation and variance

If we intend to collect n sib pairs and use a statistic of the form $T = \sum X_i$ where X_i is the number of alleles shared IBD in the i'th sib pair, we will need $p_k = P(k \text{ IBD at disease locus}|2 \text{ affected})$ for $k = 0, 1, 2$ to calculate the power.

These quantities will, when the disease model is a biallelic monogene disorder, depend on the disease allele frequency p and the genotype specific penetrances f_0, f_1, f_2 (Suarez, 1978) through

$$p_0 = \frac{1}{4} - \frac{\frac{1}{2}V_A + \frac{1}{4}V_D}{4(K_P^2 + \frac{1}{2}V_A + \frac{1}{4}V_D)}$$

$$p_1 = \frac{1}{2} - \frac{\frac{1}{2}V_D}{4(K_P^2 + \frac{1}{2}V_A + \frac{1}{4}V_D)}$$

$$p_2 = \frac{1}{4} + \frac{\frac{1}{2}V_A + \frac{3}{4}V_D}{4(K_P^2 + \frac{1}{2}V_A + \frac{1}{4}V_D)},$$

where

$$\begin{aligned} K_P &= p^2 f_2 + 2pq f_1 + q^2 f_0 \text{ (population prevalence)} \\ V_A &= 2pq(p(f_2 - f_1) + q(f_1 - f_0))^2 \text{ (additive variance)} \\ V_D &= p^2 q^2 (f_2 - 2f_1 + f_0)^2 \text{ (dominance variance).} \end{aligned}$$

Now we easily get $\mu = E(X_i)$ and $\sigma^2 = Var(X_i)$. The power to detect linkage at significance level α with n sib pairs can with a normal approximation be calculated as

$$\Phi\left(\frac{z_\alpha - n\mu}{\sigma\sqrt{n}}\right) \text{ where } z_\alpha = \Phi^{-1}(1-\alpha).$$

This formula assumes that all sib pairs come from different families and it doesn't take the affection status of any other family members into account.

In reality however we will know the affection status of the family members and we will also occasionally have $a > 2$ affected siblings in which case we have $\binom{a}{2}$ sib pairs in the family.

Conditioning on the affection status of the entire family will give different IBD probabilities for different family affection structures. By doing these more detailed calculations on a collected sample of nuclear families we can get a more accurate power estimate.

We will generalise $P(k \text{ IBD}|2 \text{ affected sibs})$ to $P(I|A)$, where I is an IBD configuration for a set of affected sibs and A is the affection status of all family members.

Applying Bayes formula

$$P(I|A) = \frac{P(A|I)P(I)}{P(A)}. \tag{1}$$

$P(I)$ is easy to derive by Mendels laws. Conditioning on the parents genotypes G

$$\begin{aligned} P(A) &= \sum_G P(A|G)P(G) \\ P(A|I) &= \sum_G P(A|G,I)P(G|I) \\ &= \sum_G P(A|G,I)P(G). \end{aligned}$$

If we let $I = I_{ij}$ in (1) be the IBD count between the sibs i and j, we can derive the family specific formulas for p_0, p_1 and p_2. This is sufficient for calculating the total expectation of all $\binom{a}{2}$ sib pair IBD counts in one family, but for the total variance in a family we need the covariances between all pairs of sib pair IBD counts in the same sibship. There are three different

kind of pairs of sib pairs to consider. If i, j, k, l all refer to different sibs we need

$$c_2 = Cov(I_{ij}, I_{ij})$$
$$c_1 = Cov(I_{ij}, I_{jl})$$
$$c_0 = Cov(I_{ij}, I_{kl}),$$

where the index of c refers to the number of siblings in common, so c_2 is just the variance. To calculate c_1 we set $I = (I_{ij}, I_{jl})$ in (1) and similar for c_0. This is partly done in (Sham *et al.*, 1997) but with the mistake of assuming $c_0 = 0$. Under the null hypothesis we will obviously have $c_0 = 0$, but the power calculations are performed for a disease locus. The results of all calculations above are shown in appendix Tables VIII,IX,X and XI.

If the genome scan is going to be performed with a dense set of markers we need not worry about the recombination distance. But with a more sparse set of markers, say 10 cM between adjacent markers, we might prefer to optimise the weights on the maximum possible distance 5 cM, corresponding to the recombination fraction $\theta \approx 0.05$. So rather than observing I_{ij} at the disease locus we will observe the IBD-count Y_{ij} at a marker locus a fraction θ off the spot. We have

$$P(Y_{ij}|A) = \sum_{I_{ij}} P(Y_{ij}|A, I_{ij})P(I_{ij}|A) = \sum_{I_{ij}} P(Y_{ij}|I_{ij})P(I_{ij}|A). \qquad (2)$$

It is straightforward to calculate $P(Y_{ij}|I_{ij})$. The results are taken from (Sham *et al.*, 1997) and shown in appendix Table XII. This is all we need to calculate expectations, but for the variance we also need $P(Y_{ij}, Y_{kl}|A)$ and $P(Y_{ij}, Y_{jk}|A)$.

With i, j, k, l all different

$$\begin{aligned} & P(Y_{ij}, Y_{kl}|A) \\ &= \sum P(Y_{ij}, Y_{kl}|I_{ij}, I_{kl}, A)P(I_{ij}, I_{kl}|A) \\ &= \sum P(Y_{ij}|I_{ij}, I_{kl}, A)P(Y_{kl}|I_{ij}, I_{kl}, A)P(I_{ij}, I_{kl}|A) \qquad (3) \\ &= \sum P(Y_{ij}|I_{ij})P(Y_{kl}|I_{kl})P(I_{ij}, I_{kl}|A), \end{aligned}$$

but if $j = k$ the equality (3) does not hold as is claimed in (Sham *et al.*, 1997) since Y_{ij} and Y_{jl} are not independent conditional on (I_{ij}, I_{jl}, A) due to the cross over process in the common sibling j.

If $(I_{ij}, I_{jl}) \neq (1,1)$ we can proceed by conditioning on the number of recombinations R_j in the common sibling j

$$\begin{aligned} & P(Y_{ij}, Y_{jl}|I_{ij}, I_{jl}, A) \\ &= \sum_{R_j} P(Y_{ij}, Y_{jl}|I_{ij}, I_{jl}, A, R_j)P(R_j) \\ &= \sum_{R_j} P(Y_{ij}|I_{ij}, I_{jl}, A, R_j)P(Y_{jl}|I_{ij}, I_{jl}, A, R_j)P(R_j) \qquad (4) \\ &= \sum_{R_j} P(Y_{ij}|I_{ij}, R_j)P(Y_{jl}|I_{jl}, R_j)P(R_j). \end{aligned}$$

The conditional independence used in equality (4) is obvious in the case $R_j = 0$ or $R_j = 2$. For $R_j = 1$ the fact that one of I_{ij} and I_{jl} must be 0 or 2 is used. Assume it is I_{ij}. Then Y_{ij} depends entirely on the recombinations in i.

The case $(I_{ij}, I_{jl}) = (1,1)$ needs a special treatment

$$\begin{aligned} & P(Y_{ij}, Y_{jl}|(I_{ij}, I_{jl}) = (1,1), A) \\ &= \sum_{R_j} P(Y_{ij}, Y_{jl}|(I_{ij}, I_{jl}) = (1,1), A, R_j)P(R_j) \\ &= P(Y_{ij}, Y_{jl}|(I_{ij}, I_{jl}) = (1,1), A, R_j = 1)P(R_j = 1) \\ &\quad + \sum_{R_j=0,2} P(Y_{ij}|I_{ij} = 1, R_j)P(Y_{jl}|I_{jl} = 1, R_j)P(R_j) \end{aligned}$$

Splitting $(I_{ij}, I_{jl}) = (1,1)$ in the two events $(1,1)^S$ = "The same IBD allele in all three siblings" and $(1,1)^D$ = "Different IBD alleles in i,j pair compared to j,l pair" and setting

$$\gamma = P((1,1)^S|(I_{ij}, I_{jl}) = (1,1), A) = \frac{P(A|(1,1)^S)}{2P(A|(I_{ij}, I_{jl}) = (1,1))}, \qquad (5)$$

we get

$$\begin{aligned} & P(Y_{ij}, Y_{jl}|(I_{ij}, I_{jl}) = (1,1), A, R_j = 1) \\ &= P(Y_{ij}, Y_{jl}|(1,1)^S, R_j = 1)\gamma \\ &\quad + P(Y_{ij}, Y_{jl}|(1,1)^D, R_j = 1)(1-\gamma). \end{aligned}$$

The details of the components that are necessary to finally calculate $P(Y_{ij}, Y_{jl}|A)$ are given in Appendix Tables XIII,XIV and XV.

We illustrate the different contributions from different families with an example where $(p, f_0, f_1, f_2) = (0.01, 0.005, 0.005, 1) \equiv M_1$. Only two percent of the cases have a genetic cause, while an affected sib pair has a genetic cause in about half of the cases and with more than two affected sibs the cause is most likely genetic.

For each family i with a_i affected children we create a score $X_i = \sum_{j<k} I_{jk}$ where j, k are affected siblings in i. We then normalise this score

$$Z_i = \frac{X_i - \binom{a_i}{2}}{\sqrt{\binom{a_i}{2}\frac{1}{2}}}, \tag{6}$$

so that under the null hypothesis of no linkage $E_{H_0}(Z_i) = 0$ and $Var_{H_0}(Z_i) = 1$. The ratio between expectation and standard deviation (E/SD) for Z_i in various family structures is shown in Table I. Calculated for an affected sib pair, disregarding any knowledge of other family members, we get E/SD=1.76. As we would expect, we see reduced values in the column with only two affected siblings due to the phenocopy effect. Even more reduced values are seen in the main diagonal due to the gene overdose effect when two parents are affected.

Weighting IBD scores

The statistic we will first consider is a sum of normalised scores

$$T = \sum_i Z_i, \tag{7}$$

where Z_i is defined in (6) and based on a sum of IBD counts for all possible affected sib pairs in the family.

In a family with two affected sibs the segregation of the alleles has 16 different outcomes. Many ASP statistics are based on mapping these outcomes into three categories, defined by 0,1,2 alleles identical by descent. The corresponding score X has the admittingly natural values 0,1,2 with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ under the null alternative of no linkage, but there are also statistics that in principle map the three categories onto non equidistant scores $0, 1, a$ (where $a \neq 2$) (Knapp *et al.*, 1995).

These simple mappings disregard pieces of information that are essential in the alternative. One allele IBD from an affected parent tells us something different than one allele IBD from an unaffected parent and with many

Table I: *E/SD for model (0.01,0.005,0.005,1)*

affected parents	total children	affected children						
		2	3	4	5	6	7	8
	2	*0.55*						
	3	*0.46*	2.02					
	4	*0.39*	1.98	2.50				
0	5	*0.32*	1.93	2.50	2.82			
	6	*0.26*	1.87	2.50	2.82	3.10		
	7	*0.21*	1.80	2.50	2.82	3.10	3.35	
	8	*0.16*	1.71	2.49	2.82	3.10	3.35	3.59
	2	*0.62*						
	3	*0.56*	1.11					
	4	*0.48*	1.14	1.37				
1	5	*0.39*	1.17	1.38	1.62			
	6	*0.31*	1.21	1.39	1.62	1.86		
	7	*0.24*	1.26	1.42	1.61	1.85	2.08	
	8	*0.18*	1.30	1.45	1.60	1.83	2.07	2.29
	2	*0.22*						
	3	*0.59*	*0.20*					
	4	*0.52*	1.10	*0.15*				
2	5	*0.43*	1.12	1.37	*0.10*			
	6	*0.34*	1.14	1.38	1.62	*0.06*		
	7	*0.26*	1.17	1.39	1.62	1.86	*0.04*	
	8	*0.20*	1.21	1.40	1.61	1.86	2.09	*0.02*

affected siblings the total sum of IBD counts for each pair hides relevant information. This difference is taken into account in (Whittemore and Halpern, 1994) and is implemented in GENEHUNTER (Kruklyak *et al.*, 1996).

Other statistics are focused on estimation of the unconditional (p_0, p_1, p_2) with (Holmans, 1993) or without (Risch, 1990) constraints.

All of these statistics have in common that they are considered non-parametric. Their relative performance depends however on the underlying model. So, already when we choose a statistic among them, we are implicitly favouring some models before others, although we might not be aware of it.

We will adopt a semiparametric approach in modifying (7) by putting model dependent weights, w_i, on each family score

$$T = \sum_i w_i Z_i. \tag{8}$$

If we set the additional constraint $\|\boldsymbol{w}\| = \sum w_i^2 = 1$ on the weights, T will be approximately standard normal under the null hypothesis for all set of weights. A similar approach is taken in (Sham *et al.*, 1997) but with the sum directly on sib pair level rather than family level. The resulting statistic is equivalent, but the weights will be different.

We will assume that all meioses are informative. This is of course an unrealistic assumption, but with highly polymorphic markers we will be close to this assumption and in any case the purpose of this study is to compare strategies. In a real implementation of the method, the families with only partially observed IBD counts need to have special underlying scores and the weights could be based on the parental marker genotypes.

Our goal is of course to come up with some good weights, but one can easily think of a number of goodness criteria. Some weighting schemes have been proposed in (Suarez and Hodge, 1979) and (Hodge, 1984). These are compared in (Sham *et al.*, 1997) with weights based on maximum power and equal weights on each sib pair. We will consider the following weighting schemes.

- **Maximum power**

  Choose your favourite significance level α and then optimise the normal approximation power

$$\Phi\left(\frac{\sum w_i \mu_i - z_\alpha}{\sqrt{\sum w_i^2 \sigma_i^2}}\right). \tag{9}$$

Since Φ is monotone the optimal weights are achieved by maximising its argument and it turns out to be the solutions of a set of polynomial equations.

In general we need to solve the optimisation problem numerically and the solution is only optimal for a particular significance level α. So in order to choose weights we have to face the question of what is most important, optimal power for significant linkage, suggestive linkage (Lander and Kruglyak, 1995) or some other level.

As an extreme illustration of the dependence of α we can study what happens to the weights as $\alpha \to 0$. Then $z_\alpha \to \infty$ and the term $\sum w_i\mu_i$ becomes negligible, so (9) will be maximised when $\sum w_i^2\sigma_i^2$ is maximum, which happens when $w_i = 1$ for the Z_i with the largest variance σ_i^2 and all other weights equal to zero.

- **Maximum expectation**

  By this approach we hope that the power will be fair for all kind of small significance levels. The simple solution to maximising $\sum w_i\mu_i$ when $\boldsymbol{w}$ is of constant length is to make $\boldsymbol{w}$ parallel to $\boldsymbol{\mu}$. With the constraint $\|\boldsymbol{w}\| = 1$ this gives $w_i = \mu_i/\|\boldsymbol{\mu}\|$.

  This corresponds to the dual problem of fixing the power to 50% (thereby setting $\sum w_i\mu_i = z_\alpha$) and then design the weights in order to get the smallest possible corresponding significance level, which occurs when $\sum w_i\mu_i$ is max.

- **Maximum asymptotic power**

  If we fix α and let the sample size $n \to \infty$ as the power is maximised the term $-z_\alpha$ in (9) becomes negligible and we will end up with a solution that does not depend on α.

  Solving the polynomial equations with $z_\alpha = 0$ gives a solution where w_i are proportional to μ_i/σ_i^2. This solution can be achieved easier by

$$\frac{\sum w_i\mu_i}{\sqrt{\sum w_i^2\sigma_i^2}} = \frac{\sum (w_i\sigma_i)(\frac{\mu_i}{\sigma_i})}{\sqrt{\sum (w_i\sigma_i)^2}} \leq \sqrt{\sum (\frac{\mu_i}{\sigma_i})^2},$$

  where the inequality is Cauchy-Schwartz, with equality iff $w_i\sigma_i \propto \frac{\mu_i}{\sigma_i}$ i.e. $w_i \propto \frac{\mu_i}{\sigma_i^2}$. If n is very large though, we are not too worried about the power.

Since we achieved the weights by setting $z_\alpha = 0$ these weights will obviously give maximum power for the modest $\alpha = 0.5$ without the asymptotic argument, although the sample size need to be large enough to use normal approximation in the power calculation.

- **Equal weights on every family**

  The constraint $\|\boldsymbol{w}\| = 1$ gives weights $1/\sqrt{n}$ to each family. This is the solution adopted by GENEHUNTER (Kruklyak *et al.*, 1996).

- **Equal weights on every sib pair**

  If we do not consider the concept family score we can sum all sib pair IBD scores from all families and then normalise it. This corresponds to having family weights $w_i \propto \sqrt{\binom{a_i}{2}}$.

- **Weights inverse proportional to coefficient of variation**

  One way to compare the utility of two families is to compare the power to detect linkage at significance level α between two samples with n families of each type. This will reveal that family type 1 is better than type 2 if

  $$\phi^{-1}(1-\alpha)(\sigma_2 - \sigma_1) < (\mu_1\sigma_2 - \mu_2\sigma_1)n. \tag{10}$$

  The RHS of inequality (10) is positive if the coefficient of variation is smaller (E/SD is larger) in type 1, which means that, for sufficiently large n, type 1 should be preferred.

  Thus E/SD is one reasonable way of ordering the families and the weights reflect this order.

A simulation was done of a sample from the model in Table I with the family sizes following a Poisson distribution with intensity 3 truncated to max 8 children. This size structure fits well with a large sample of 11,000 Swedish psoriasis families [personal observation]. The simulated sample of 40 families were distributed as shown in Table II and the power was calculated for optimum weights at significance levels 0.01, 0.00074 (suggestive linkage) and 0.000022 (significant linkage) (Lander and Kruglyak, 1995), equal family weights (EqF), equal sib pair weights (EqSP) and weights proportional to expectation (Exp), expectation/variance (E/V) and expectation/standard deviation(E/SD). Results are shown in Table III. The maximal power is of course obtained, due to the construction, at the corresponding optimal

weights. For practical purposes the power for all weights but the two equal weighting approaches are the same. The difference is substantial although we admit that the example was deliberately chosen for illustration purposes. Other models can show less difference, but what we have seen so far the choice between the good weights are not important. This indicates that the simple weights 'proportional to expectation' which is easiest to implement is a good candidate.

Table II: *Sample of 40 families from model (0.01,0.005,0.005,1)*

affected parents	total children	affected children						
		2	3	4	5	6	7	8
	2	2						
	3	6	0					
	4	13	1	0				
0	5	7	1	0	0			
	6	6	1	0	0	0		
	7	1	0	0	0	0	0	
	8	0	1	0	0	0	0	0
	2	0						
1	3	0	0					
	4	1	0	0				

Deviations from the correct model

Our assumptions of the genetic model will be more or less founded. Studying population data like the relative risk λ_R for different kind of relatives R will only give unique solutions when we put restrictions on (p, f_0, f_1, f_2) (James, 1971). The usefulness of weighting will therefore, not surprisingly, depend on the accuracy of the assumed model. If for example the known estimates of population parameters are population prevalence (K_P) and the siblings relative risk (λ_S), both the models $M_1 \equiv (0.01, 0.005, 0.005, 1)$ illustrated above and $M_2 \equiv (0.06, 0.0045, 0.0045, 0.165)$ fits the same population parameters $K_P = 0.005$ and $\lambda_S = 2$.

Table III: *Power at 3 different levels and different weights*

	significance level		
weights	0.01	sugg	sign
0.01	*0.9788*	0.8779	0.5967
sugg	0.9787	*0.8780*	0.5973
sign	0.9787	0.8779	*0.5976*
EqF	0.8338	0.5550	0.2278
EqSP	0.9264	0.7297	0.3898
Exp	0.9786	0.8778	0.5976
E/V	0.9786	0.8766	0.5930
E/SD	0.9788	0.8779	0.5966

Table IV: *Effects of model deviations*

		assumed M_1			assumed M_2		
	weights	0.01	sugg	sign	0.01	sugg	sign
	0.01	*0.9911*	0.9437	0.7747	0.9665	0.8446	0.5569
	sugg	0.9910	*0.9441*	0.7778	0.9757	0.8774	0.6176
	sign	0.9907	0.9437	*0.7790*	*0.9825*	*0.9048*	*0.6768*
correct	EqF	0.8359	0.5612	0.2349	0.8359	0.5612	0.2349
M_1	EqSP	0.9682	0.8542	0.5806	0.9682	0.8542	0.5806
	Exp	0.9903	0.9425	0.7780	0.9820	0.9025	0.6714
	E/V	0.9905	0.9388	0.7560	0.9321	0.7470	0.4151
	E/SD	0.9911	0.9438	0.7754	0.9646	0.8386	0.5470
	0.01	0.7487	0.4287	0.1391	*0.7821*	0.4727	0.1659
	sugg	0.7500	0.4304	0.1401	0.7821	*0.4727*	0.1659
	sign	0.7513	0.4320	0.1411	0.7820	0.4727	*0.1660*
correct	EqF	*0.7817*	*0.4722*	*0.1656*	0.7817	0.4722	0.1656
M_2	EqSP	*0.7817*	*0.4722*	*0.1656*	0.7817	0.4722	0.1656
	Exp	0.7488	0.4288	0.1392	0.7821	0.4727	0.1659
	E/V	0.7446	0.4233	0.1359	0.7819	0.4722	0.1655
	E/SD	0.7468	0.4262	0.1376	0.7821	0.4726	0.1658

We see in Table IV that when M_1 is true the optimal weights strategy outperforms equal weights, under both the correct and wrong assumption. When M_2 is true however the equal weights happen to be somewhat better. Although the weights seem to be fairly robust against model deviations in this case and many other that we have studied we can not claim that this is always the case.

It is sometimes argued that equal weights can be a kind of assurance to get the best power for the worst case. However in our case the gain of equal weights is small when M_2 is true compared to the loss when M_1 is true. In case we are indifferent regarding the models, in the sense that our prior probability is $\frac{1}{2}$ for each of them, this assurance strategy is a very pessimistic one. A better alternative for assurance is to design weights for a mixed assumption by looking at the expected value over both models

$$E[Z_i] = \frac{1}{2}E[Z_i|M_1] + \frac{1}{2}E[Z_i|M_2]. \tag{11}$$

The strategy is applied to the previous example and the results are shown in Table V.

Table V: *Power comparisons between mixed strategy and equal weights*

model	weights	0.01	sugg	sign
M_1	mix	*0.9898*	*0.9384*	*0.7618*
	EqF	0.8359	0.5612	0.2349
	EqSP	0.9682	0.8542	0.5806
M_2	mix	0.7762	0.4645	0.1605
	EqF	*0.7817*	*0.4722*	*0.1656*
	EqSP	*0.7817*	*0.4722*	*0.1656*

In principle this idea could be generalised to cases where we have a more complex prior probability distribution of the model space and we could calculate $E[Z_i] = E[E[Z_i|M]]$ where M is a random model. But to be honest, who would have any clear idea of this distribution?

Sampling strategies

The idea is that we shall of course preferably collect good families before bad. In (McCarthy *et al.*, 1998) a number of general strategies for collecting sib pairs are compared with a random collection at different models, both one and two locus models. However they only collect *one* sib pair from any family. This is too much when gene overdose is likely, but a waste of useful information when gene overdose is not likely, which it rarely is.

The concept good is as we saw before somewhat unclear. One way is to order the families according to coefficient of variation of Z_i as was indicated in (10).

The sampling strategy consists of two steps. In the first we examine the phenotypes of an available population A of families with at least two affected sibs and in the second we select the N best families in A, i.e. the families with smallest coefficient of variation.

We do not claim that these N families necessarily is the best theoretical choice among all subsamples of size N, on the contrary, it is not hard to construct counterexamples, but we trust it is good enough.

In Table VI we have simulated an available population of 100 families, with M_1 as well as M_2. For each of these two populations we select the 40 best families according to an assumption of M_1 or M_2

Table VI: *Power to detect significant linkage with selective sampling*

Correct Assumed	M_2 M_2	M_2 M_1	M_1 M_1	M_1 M_2
0.01	0.466	0.427	0.977	0.916
sugg	0.469	0.431	0.978	0.939
sign	*0.470*	0.435	*0.979*	0.959
EqF	0.351	0.351	0.694	0.694
EqSP	0.440	*0.440*	0.955	0.955
Exp	0.470	0.437	0.976	*0.964*
E/V	0.452	0.417	0.972	0.844
E/SD	0.465	0.427	0.979	0.921

We could apply a mixed strategy as well, with weights based on (11) and with a selection criteria where for each family $E/SD_i^{min} = min(E/SD_i^{M_1}, E/SD_i^{M_2})$

and the 40 families are selected according to the largest E/SD_i^{min}.

If we do not weight the families, a peculiar situation might occur with selective sampling of fix sizes. The power can be larger with some of the families removed. Return to the example in Table III and define $T_k = \sum_1^k Z_i$ for $k = 1, ..40$ where the Z_i are ordered according to E/SD such that Z_i is the smallest.

The four families with 3 affected children had the largest E/SD. Using only them gives a power as high as 0.32. When we increase the sample by adding the remaining 36 families with only two affected children, the power of the optimal weights statistic slowly increases up to 0.60, while the power for the equal family weights statistic rapidly decreases, oscillates and remains low ending in 0.23.

Table VII: *Power to detect significant linkage at various sample sizes*

	weights	
size	sign	EqF
4	0.325	0.325
10	0.402	0.200
15	0.453	0.196
20	0.494	0.202
25	0.533	0.219
30	0.561	0.227
35	0.583	0.231
40	0.598	0.228

Discussion

We have studied two methods of increasing the power in a genome scan with affected sib pairs; selective sampling and weighting. The strategies can also be combined. If our model assumption is correct the power can be considerably increased.

To use weights with maximum power we have to decide what significance level that is most important, but for the cases we have presented, and for other cases we have seen during several trials, it is for practical purposes

sufficient to maximise the expected value of the statistic. This has a simple solution.

The selective sampling rejects the 'bad' families and selects the 'good' families. It is possible to further increase the power by putting the highest weights to the very best families amongst those selected. If we are so sure of our model assumptions, that we deliberately bias our sample, we should definitely go for a weighted statistic as well in order to further increase the power.

In case we hesitate between models, we propose weights that are proportional to the expected values taken over the model candidates, rather than a unreflected arbitrary choice of equal weights.

We will always be more or less wrong in our model assumption. With selective sampling the most severe risk is that we select bad families rather that good. There will be no magic way to repair this damage, while if we recruit families randomly, we can apply a weighted statistic and if we fail to find linkage, we could try another model with other weights. This would of course be fishing, with all of its danger, but at least we would have a sea to fish in.

The study was limited to comparing weights and sampling strategies with a pure sib pair IBD count based statistic as the basic statistic. In a future work we plan to compare these approach with an ordinary LOD score analysis with respect to power and robustness.

A Appendix

Table VIII: *ASP probabilities given IBD count and mating types*

G	p_{0G}	p_{1G}	p_{2G}
DDxDD	f_2^2	f_2^2	f_2^2
DDxDd	$f_1 f_2$	$(f_1+f_2)^2/4$	$(f_1^2+f_2^2)/2$
DDxdd	$(f_0 f_2+f_1^2)/2$	$f_1(f_0+f_2)/2$	$(f_0^2+2f_1^2+f_2^2)/4$
DdxDd	f_1^2	f_1^2	f_1^2
Ddxdd	$f_0 f_1$	$(f_0+f_1)^2/4$	$(f_0^2+f_1^2)/2$
ddxdd	f_0^2	f_0^2	f_0^2

G = mating type, $p_{iG} = P(\text{ASP}|i \text{ IBD alleles,mating type } G)$

Table IX: *Disease probabilities of family members given mating types*

G	f_G	h_{0G}	h_{1G}	h_{2G}
DDxDD	f_2	$(1-f_2)^2$	$2f_2(1-f_2)$	f_2^2
DDxDd	$(f_1+f_2)/2$	$(1-f_1)(1-f_2)$	$f_1+f_2-2f_1f_2$	f_1f_2
DDxdd	f_1	$(1-f_0)(1-f_2)$	$f_0+f_2-2f_0f_2$	f_0f_2
DdxDd	$(f_0+2f_1+f_2)/4$	$(1-f_1)^2$	$2f_1(1-f_1)$	f_1^2
Ddxdd	$(f_0+f_1)/2$	$(1-f_0)(1-f_1)$	$f_0+f_1+2f_0f_1$	f_0f_1
ddxdd	f_0	$(1-f_0)^2$	$2f_0(1-f_0)$	f_0^2

G = mating type, $f_G = P(\text{affected child}|G)$,
$h_{mG} = P(m \text{ affected parents}|G)$

Table X: *Two sib pairs with one sib in common*

G	IBD	$P(A\|I_{ij}, I_{jk}, G)$
DDxDD	any	$h_{mG}(1-f_G)^{n-a}f_G^a$
DDxDd	0, 0	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^2f_2+f_1f_2^2)/2$
	0, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^2f_2+f_1f_2^2)/2$
	0, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^2f_2+f_1f_2^2)/2$
	1, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1+f_2)^3/8$
	1, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^3+f_1^2f_2+f_1f_2^2+f_2^3)/4$
	2, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^3+f_2^3)/2$
DDxdd	any	$h_{mG}(1-f_G)^{n-a}f_G^a$
DdxDd	0, 0	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^2f_2+f_0f_2^2+2f_1^3)/4$
	0, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}f_1(2f_0f_2+f_0f_1+f_1f_2)/4$
	0, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0f_2^2+2f_1^3+f_0^2f_2)/4$
	1, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}f_1(2f_1f_2+2f_0f_1+2f_0f_2+f_2^2+f_0^2)/8$
	1, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}f_1(f_0^2+f_0f_1+f_1f_2+f_2^2)/4$
	2, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^3+2f_1^3+f_2^3)$
Ddxdd	0, 0	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^2f_1+f_0f_1^2)/2$
	0, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^2f_1+f_0f_1^2)/2$
	0, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^2f_1+f_0f_1^2)/2$
	1, 1	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0+f_1)^3/8$
	1, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^3+f_0^2f_1+f_0f_1^2+f_1^3)/4$
	2, 2	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^3+f_1^3)/2$
ddxdd	any	$h_{mG}(1-f_G)^{n-a}f_G^a$

n = number of children, a = number of affected children,

$p_{iG} = P(\text{ASP}|i \text{ IBD alleles,mating type } G)$

$f_G = P(\text{affected child}|G), h_{mG} = P(m \text{ affected parents}|G)$

Table XI: *Two sib pairs with no sib in common*

IBD	$P(A\|I_{ij}, I_{kl}, G)$
$0,0$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{0G}^2$
$0,1$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{0G}p_{1G}$
$0,2$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{0G}p_{2G}$
$1,1$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{1G}^2$
$1,2$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{1G}p_{2G}$
$2,2$	$h_{mG}(1-f_G)^{n-a}f_G^{a-4}p_{2G}^2$

$n =$ number of children, $a =$ number of affected children,
$p_{iG} = P(\text{ASP}|i \text{ IBD alleles,mating type } G)$
$f_G = P(\text{affected}|G), h_{mG} = P(m\text{affected parents}|G)$

Table XII: *IBD probabilities at marker given IBD count at disease locus*

	Disease Locus IBD		
Marker Locus IBD	0	1	2
0	Ψ^2	$\Psi(1-\Psi)$	$(1-\Psi)^2$
1	$2\Psi(1-\Psi)$	$1-2\Psi(1-\Psi)$	$2\Psi(1-\Psi)$
2	$(1-\Psi)^2$	$\Psi(1-\Psi)$	Ψ^2

$\Psi = \theta^2 + (1-\theta)^2$, $\theta =$ recombination fraction

Table XIII: *IBD probabilities at marker given IBD count at disease locus. Three siblings.*

Marker IBD	Disease Locus IBD 0	1	2	
0	$(1-\theta)^2$	$\theta(1-\theta)$	θ^2	$R_j = 0$
1	$2\theta(1-\theta)$	$\theta^2+(1-\theta)^2$	$2\theta(1-\theta)$	
2	θ^2	$\theta(1-\theta)$	$(1-\theta)^2$	
0	$\theta(1-\theta)$	$\theta^2+(1-\theta)^2/2$	$\theta(1-\theta)$	$R_j = 1, (I_{ij}, I_{jl}) \neq (1,1)$
1	$\theta^2+(1-\theta)^2$	$2\theta(1-\theta)$	$\theta^2+(1-\theta)^2$	
2	$\theta(1-\theta)$	$\theta^2+(1-\theta)^2/2$	$\theta(1-\theta)$	
0	θ^2	$\theta(1-\theta)$	$(1-\theta)^2$	$R_j = 2$
1	$2\theta(1-\theta)$	$\theta^2+(1-\theta)^2$	$2\theta(1-\theta)$	
2	$(1-\theta)^2$	$\theta(1-\theta)$	θ^2	

θ = recombination fraction, R_j = number of recombinations in common sib

Table XIV: *Joint IBD probabilities at marker in special case with three siblings.*

Y_{ij}, Y_{jl}	$P(Y_{ij}, Y_{jl}\|(I_{ij}, I_{jl}) = (1,1), R_j = 1, A)$
$0,0$	$\frac{1}{2}(\theta^4 + (1-\theta)^4)\gamma + \theta^2(1-\theta)^2(1-\gamma)$
$0,1;1,0$	$\theta^3(1-\theta) + \theta(1-\theta)^3$
$0,2;2,0$	$\frac{1}{2}(\theta^4 + (1-\theta)^4)(1-\gamma) + \theta^2(1-\theta)^2\gamma$
$1,1$	$4\theta^2(1-\theta)^2$
$1,2;2,1$	$\theta^3(1-\theta) + \theta(1-\theta)^3$
$2,2$	$\frac{1}{2}(\theta^4 + (1-\theta)^4)\gamma + \theta^2(1-\theta)^2(1-\gamma)$

θ = recombination fraction, $\gamma = P((1,1)^S|(I_{ij}, I_{jl}) = (1,1), A)$,
R_j = number of recombinations in common sib

Table XV: *Probabilities needed for calculations of* γ

G	$P(A\|(1,1)^S, R_j = 1, G)$
`DDxDD`	$h_{mG}(1-f_G)^{n-a}f_G^a$
`DDxDd`	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^3 + f_1^2 f_2 + f_1 f_2^2 + f_2^3)/4$
`DDxdd`	$h_{mG}(1-f_G)^{n-a}f_G^a$
`DdxDd`	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}f_1(f_0^2 + f_0 f_1 + f_1 f_2 + f_2^2)/4$
`Ddxdd`	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^3 + f_0^2 f_1 + f_0 f_1^2 + f_1^3)/4$
`ddxdd`	$h_{mG}(1-f_G)^{n-a}f_G^a$

n = number of children, a = number of affected children,
$p_{iG} = P(\text{ASP}|i$ IBD alleles,mating type $G)$
$f_G = P(\text{affected}|G), h_{mG} = P(m$ affected parents$|G)$

References

Hodge, S. (1984). The information contained in multiple sibling pairs. *Genetic Epidemiology*, **1**, 109–122.

Holmans, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *Am J Hum Genet*, **52**, 362–374.

James, J. (1971). Frequency in relatives for an all-or-none trait. *Ann Hum Genet*, **35**, 47–49.

Knapp, M., Seuchter, S., and Bauer, M. (1995). Linkage analysis in nuclear families, i. optimality criteria for affected sib-pair tests. *Human Heredity*, **44**, 37–43.

Kruklyak, L., Daly, M., Reeve-Daly, M., and Lander, E. (1996). Parametric and nonparametrice linkage analysis: A unified multipoint approach. *Am J Hum Genet*, **58**, 1347–1363.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting results. *Nature Genetics*, **11**, 241–247.

McCarthy, M., Kruglyak, L., and Lander, E. (1998). Sib-pair collection strategies for complex diseases. *Genetic Epidemiology*, **15**, 317–340.

Risch, N. (1990). Linkage strategies for genetically complex traits. iii. the effect of marker polymorphisms on analysis of affected relative pairs. *Am J Hum Genet*, **46**, 242–253.

Sham, P., Zhao, J., and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann Hum Genet*, **61**, 61–69.

Suarez, B. (1978). The affected sib pair ibd distribution for hla-linked disease susceptibility genesd. *Tissue Antigens*, **12**, 87–93.

Suarez, B. and Hodge, S. (1979). A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clinical Genetics*, **18**, 126–136.

Whittemore, A. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

Association analysis with TDT in locus heterogeneity models

Staffan Nilsson

26 June 2001

Abstract

When the TDT statistic is used for linkage analysis, nuclear families with at least one affected child are used and each transmission from a heterozygous parent to an affected child is counted. However, when the TDT statistic is used for association analysis in the presence of linkage, the transmissions from heterozygous parents to different affected children are not independent. In many studies this problem has been solved by the poor solution of randomly selecting one child from each sibship. Other methods have been developed that use more of the information from families with multiple affected sibships, ASP-families, such as the T_{sp} statistic of Martin et al. We develop this method further to increase the power. The increase is most pronounced in genetic models with locus heterogeneity or when the phenocopy rate is large.

Introduction

In fine mapping of disease susceptibility genes, when linkage analysis has identified a candidate region, association analysis is the most used tool. It relies on the possible linkage disequilibrium (LD) between a marker allele and a disease allele, when the marker and disease loci are tightly linked. There are several methods available for association analysis. From a sampling point of view an ordinary case-control study is preferred, but in mixed populations spurious association can occur (Ewens and Spielman, 1995).

The transmission disequilibrium test (TDT), which was introduced in (Spielman *et al.*, 1993), is a test that eliminates the risk above. The smallest unit in the test statistic is the transmission from a heterozygous parent to

an affected child. Families with just one affected child, called trios, can be collected in addition to the pedigrees and/or sib pair families that are used for linkage. This is a clear advantage, since in general trios are more common than ASP-families, although their frequency ratio depends on the genetic model and the family size structure in the studied population. The basic idea is that, under the null hypothesis, the heterozygous parents two different marker alleles are transmitted with equal probability 1/2. If then all units are statistically independent a simple binomial test can be performed.

The TDT is used in two different ways depending on the choice of null hypothesis being tested. When the null hypothesis is no linkage, all such transmissions are independent and thus possible to include in the statistic, while a null hypothesis of no linkage disequilibrium requires unrelated children to get the independence needed for a binomial test.

To still be able to use ASP families for TDT analysis a valid solution is to randomly select one child, but this is intuitively an unattractive solution since part of the information is lost. To use more information, but still have the independence, the construction of an alternative test statistic, T_{sp} (Martin *et al.*, 1997), was done.

In this article we will, focusing on genetic models with locus heterogeneity or phenocopies, introduce statistics that extract information from an ASP family in a way equivalent to T_{sp}. The extracted information will be classified according to the marker allele sharing in the sib pair. It will be shown that TDT statistics from different kind of samples can have rather different transmission distortion and that this phenomena can be utilised to combine the samples in a way that will increase the power in many natural genetic models.

The basic idea is that, when a disease allele at the studied candidate locus is not necessary for the development of the disease, families where the disease allele is not segregated will show the normal populations marker allele frequencies and 50/50 transmission patterns, leading to large statistical noise. By using ASP families the phenocopy rate is reduced compared to that of trios and by selecting only transmissions that caused an identity by descent (IBD) allele in the sib pair, the rate of transmissions of a disease allele at the studied locus can be increased in locus heterogeneity models.

Methods

Extracting family data

Assume a genetic model with locus heterogeneity and/or phenocopies and a sample of ASP families with two affected children and their parents. Consider a specific disease locus and a tightly linked marker. In each family there are four transmissions of marker alleles to the affected children. Suppose that the IBD-counts (0,1 or 2) are precisely known in each family. If the IBD-count is 0, there is a strong reason to believe that no disease allele is segregated in that family at the considered disease locus. In such families the transmissions from heterozygous parents will be 50/50 transmissions and will just increase the noise. For the same reason, if the IBD count is 1, only the parent that transmitted the IBD-allele is of interest, the transmissions from the other parent is likely to be noise.

Table I: TDT table of IBD transmissions for a marker with r alleles

	Non transmitted			
Transmitted	1	2	...	r
1	N_{11}	N_{12}		N_{1r}
2	N_{21}	N_{22}		N_{2r}
...				
r	N_{r1}	N_{r2}		N_{rr}

From the transmissions of IBD-alleles an rxr contingency table of transmitted IBD-alleles vs non transmitted can be constructed (Table I). In cell (i, j) is the count N_{ij} of parents with marker genotype ij that transmit i and keep j. We simplify matters by assuming that the IBD-status always can be determined. For a single marker this is not true, but it can usually be established by use of other nearby markers. Since, conditionally of the IBD fact, each transmission now would be duplicated, only one of them should be counted.

It can easily be shown (Martin *et al.*, 1997) that the contingency table is made up by observations that are independent under the null hypothesis of no LD and with the probability to belong to cell (i, j) being $m_i m_j$, where m_i is the frequency of marker allele i.

Once this has been established several statistics, not only the TDT, can be used on this data structure. Specifically it should be technically possible to also use the statistics HRR (Falk and Rubinstein, 1987) or AFBAC (Thomson, 1995). The limitations of these statistics are the same here as when the samples are trios (Ewens and Spielman, 1995) and the subject will not be pursued further here, only variants of the TDT statistic will be studied. Most of the ideas can be carried over to the use of HRR or AFBAC.

In the most basic form of TDT there is a diallelic marker, such as an SNP. In an ASP family the two counts of interest are the IBD transmissions from heterozygous (genotype 12) parents. If these counts are N_{12} and N_{21}, then the statistic $T = N_{12}$ is binomially distributed with parameters $(N_{12}+N_{21}, \frac{1}{2})$ under the null hypothesis. It is often formulated as a McNemar test

$$\mathrm{TDT}_{IBD} = \frac{(N_{12} - N_{21})^2}{N_{12} + N_{21}} \tag{1}$$

which under the null hypothesis is asymptotically $\chi^2_{(1)}$-distributed. This turns out to be equivalent to the statistic T_{sp} in (Martin *et al.*, 1997), where the derivation is motivated by the ambition to design a TDT-like statistic, which is a valid test for LD in the presence of linkage. T_{sp} is derived as a sum of scores based on observed transmissions from heterozygous ij parents. Their score is

$$X_{ij} = \begin{cases} 2 & \text{if the parent transmits i to both children} \\ 1 & \text{if the parent transmits i to one and j to the other} \\ 0 & \text{if the parent transmits j to both children.} \end{cases}$$

From the viewpoint of our derivation the basic score is

$$X'_{ij} = \begin{cases} 1 & \text{if the parent transmits i to both children} \\ -1 & \text{if the parent transmits j to both children} \end{cases}$$

but the final statistics are the same.

Combining TDT statistics

For a locus with two marker alleles a TDT-statistic is based on an underlying binomial experiment with n transmissions from heterozygous parents such that

$$\mathrm{TDT} = \frac{T^2}{n} \text{ where } T = \sum_{k=1}^{n} X_k \text{ and } X_k = \{-1, 1\} \tag{2}$$

This statistic can be formed on different samples and a sample can be subdivided into different classes as long as the null distribution hold.

Assume there are m different classes with n_i observations $X_{ik}, k = 1, \ldots, n_i$ in class i and that $n = \sum n_i$. The partition that will be used here is

1. Transmissions in 1 IBD families
2. Transmissions in 2 IBD families
3. Transmission in family trios

To each class a separate statistic $T_i = \sum_{k=1}^{n_i} X_{ik}$ is formed. These t statistics can then be directly aggregated as

$$T_{eq} = \frac{T^2}{n}, \text{ where } T = \sum_{i=1}^{m} T_i. \tag{3}$$

T_{eq} is asymptotically $\chi^2_{(1)}$-distributed under H_0. As a result the classes have been merged into one and the usual TDT statistic has been constructed.

However, some of the classes can be considered more important than others by assigning different weights in the sum

$$T_w = \frac{T^2}{\sum n_i w_i^2} \text{ where } T = \sum_{i=1}^{t} w_i T_i \tag{4}$$

T_w is still asymptotically $\chi^2_{(1)}$ under the null hypothesis. So under the alternative hypothesis when there is association, if $\mu_i = E[X_{ik}]$ and the μ_i are very different in size the weights should try to reflect these differences. With this argument T_{eq} implicitly assumes that all μ_i are *equal*.

The T_{sp} statistic is an example where different weights are used. First of all, the classes (1) and (2) are never considered as different in the derivation so $w_1 = w_2$ automatically. Secondly, when the third class, i.e. the family trios, is included the weights become $w_1 = w_2 = 2w_3$. The last equality comes from studying heterozygous parents in the trios and introducing the scores

$$Y_{ij} = \begin{cases} 1 & \text{if the parent transmits i to the child} \\ 0 & \text{if the parent transmits j to the child.} \end{cases}$$

When the two classes are combined it is done through

$$\sum_{ASP} X_{ij} + \sum_{trio} Y_{ij}$$

The difference between the opposite scores is twice as high in the sib pairs, which explains the weights. Any reason for this is not given in (Martin *et al.*, 1997). It turns out though to be reasonable weights for the genetic models that are studied here.

We will study some different statistics that have the structure of T_w. Which to prefer is a general question in the design of statistics. The principals are outlined below.

Let Z_{ij}, $1 \leq i \leq k$, $1 \leq j \leq n_i$, be normally distributed $\mathrm{N}(\mu_i, 1)$ and let T^R be the LR test of

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \text{ vs } H_1^R : \boldsymbol{\mu} \in \Omega^R$$

where R is some restriction of the full parameter space $\Omega = \Re^k$. A classical statistical problem in this situation is the choice of the restriction R. Assume that $\Omega^R \subset \Omega^{R'}$ and that $\boldsymbol{\mu} \in \Omega^{R'}$. If also $\boldsymbol{\mu} \in \Omega^R$ then it is natural to prefer T^R to $T^{R'}$, but even if $\boldsymbol{\mu} \notin \Omega^R$ it can still be in a neighbourhood of Ω^R where T^R is more powerful than $T^{R'}$.

In our application the X'_{ij} are independent, have mean 0 and variance 1 under the null hypothesis, but they are not normally distributed. Furthermore, in the alternative the variance is not 1, nor are all the X'_{ij} independent, since the two parents transmissions are generally dependent. In spite of these discrepancies, we will use LR tests for the canonical normal case. The null distributions are asymptotically the same. T_{eq} and T_{sp} are both LR tests of hypothesis in the family above with the one-dimensional restriction spaces $\boldsymbol{\mu} = (1, 1, 1)\mu$ resp $\boldsymbol{\mu} = (2, 2, 1)\mu$.

Cell probabilities

To guide the choice of a suitable restriction it is natural to study how different genetic models are related to the mean vector $\boldsymbol{\mu}$. The dependence is through the cell probabilities in the TDT tables for the different classes. The details of the derivations are shown in Appendix. For TDT class c and a multiallelic marker with allele frequencies m_i the probability, when there is tight linkage ($\theta = 0$), that a parent transmits allele i and keeps allele j is

$$p_{ij}^{(c)} = m_i m_j + m_j \Delta_i^{(c)} \tag{5}$$

where

$$\Delta_i^{(c)} = \sum_k d_{ik} b_k^{(c)} \tag{6}$$

and d_{ik} is the disequilibrium coefficient for marker allele i and the disease locus allele k, i.e. if a_k is an allele frequency at the disease locus and h_{ik} is a haplotype frequency then $d_{ik} = h_{ik} - m_i a_k$. Further $b_k^{(c)}$ is a quantity that only depends on the disease model and the class c.

If there is allelic heterogeneity at the disease locus a marker allele i can be positively associated with one disease allele k, $d_{ik} >> 0$, while for another disease allele l, $d_{il} << 0$, with the possible result that $\Delta \approx 0$. This annoying feature with association analysis is well known (Terwilliger and Göring, 2000) and clear from (6). From here on only disease loci with two alleles are considered. In these models $\Delta_i^{(c)} = d_{i1}(b_1^{(c)} - b_2^{(c)})$, since $d_{i1} + d_{i2} = 1$.

When the marker has $r > 2$ alleles there are a number of approaches to construct test statistics on the r x r contingency table of transmissions such as ETDT (Sham and Curtis, 1995), maxTDT (Schaid, 1996), T_{mhet}, (Spielman and Ewens, 1996).

Only markers with two alleles will be considered here. The availability of such markers is rapidly increasing as more and more single nucleotide polymorphisms (SNP's) on the human chromosomes are mapped. With $\delta = d_{11}$ and $\beta_c = b_1^{(c)} - b_2^{(c)}$ the cell probabilities $p_{ij}^{(c)}$ are shown in Table II.

Table II: Transmission probabilities

	Non transmitted	
Transmitted	1	2
1	$m_1^2 + m_1\delta\beta_c$	$m_1 m_2 + m_2\delta\beta_c$
2	$m_2 m_1 - m_1\delta\beta_c$	$m_2^2 - m_2\delta\beta_c$

The quantity of interest in a TDT test is $\mu_i = 2p_i - 1$ where

$$p_i = \frac{p_{12}^{(i)}}{p_{12}^{(i)} + p_{21}^{(i)}}.$$

If the disease allele is named 1 as well as its associated marker allele, then $p_i > 1/2$.

The values for the three classes in some specific models are shown in Table III. It should be noted that these models represent a rather wide class

of other models as far as the sizes of μ_i or equivalently p_i are concerned. From the formulas in Appendix it is clear that the μ_i depend on the disease allele frequency p and penetrances f_0, f_1, f_2 for the genotypes at the disease locus, but only through the component specific population prevalence K_{2p} (Risch, 1990) and the corresponding sib prevalence K_{2s} of the alternative genetic mechanism. This means that the other unlinked component can be ever so complicated as long as it penetrates independently of the current locus. Furthermore the total genetic model can be multiplicative with two genetic components A and B, where the studied locus is part of component A. As before the structure of B is not important as long as it penetrates independently of A and is unlinked to A since then

$$P(M_{ij}|\text{Aff}) = P(M_{ij}|\text{Aff}_A, \text{Aff}_B) = P(M_{ij}|\text{Aff}_A).$$

For the study of the current locus the only impact of the additional B component is then a reduction of the disease prevalence and a change in the ratio ASP/Trios.

Choosing weights

When an allele has been found to be overrepresented in one study, a confirmation study can use the estimates $\hat{\mu}_i$ from the previous study as weights in a one sided test with

$$Z = \frac{\sum \hat{\mu}_i T_i}{\sqrt{\sum \hat{\mu}_i^2 n_i}}$$

where Z is asymptotically N(0,1) in the null hypothesis and where rejection is done for large Z.

When no such prior information is available regarding the direction of the LD (the sign of δ) or the size of the μ_i, we can try to guess the genetic model. That will determine the β_c and thereby

$$\frac{\mu_i}{\mu_j} = \frac{\beta_i}{\beta_j} \frac{2m_1m_2 + (m_2 - m_1)\delta\beta_j}{2m_1m_2 + (m_2 - m_1)\delta\beta_i}, \tag{7}$$

which reduces to

$$\frac{\mu_i}{\mu_j} = \frac{\beta_i}{\beta_j} \tag{8}$$

if $m_1 = m_2$, but also in the limit as $\delta \to 0$. This gives that the β_i are reasonable weights and they will be used in the simulations as a reference statistic called T_{ref}.

Table III: The transmission distortion in some models

model	p	f_0	f_1	f_2	D'	K_{2p}	K_{2s}	μ_1	μ_2	μ_3
DP50	0.001	0.002	1	1	0.5	0	0	0.496	0.249	0.125
DP20	0.001	0.008	1	1	0.5	0	0	0.463	0.242	0.050
DH50	0.001	0	1	1	0.5	0.002	0.5	0.249	0.166	0.125
DH20	0.001	0	1	1	0.5	0.008	0.5	0.099	0.083	0.050
RP50	0.03	0.0009	0.0009	1	0.3	0	0	0.292	0.300	0.150
RP20	0.03	0.0036	0.0036	1	0.3	0	0	0.216	0.296	0.060
RH50	0.03	0	0	1	0.3	0.0009	0.27	0.032	0.237	0.150
RH20	0.03	0	0	1	0.3	0.0036	0.28	0.009	0.141	0.060

The models are designated $XYnn$, where X indicates that the studied locus is D - dominant or R - recessive and Y indicates that there are P - phenocopies ($f_0 > 0$) or H - heterogeneity (genes at other loci with similar effect). Finally nn indicates that the loci is roughly responsible for nn percent of the affected population. The marker is an SNP with two equally frequent alleles ($m_1 = m_2 = 0.5$) and the disequilibrium is expressed as

$$D' = \frac{h_{11} - a_1 m_1}{a_1 m_2}$$

(Lewontin, 1964).

If the mle estimates $\hat{\mu}_i$ from the same investigation are adaptively used as weights in (4) it turns out (Robertson *et al.*, 1988) that

$$T_{\hat{\boldsymbol{\mu}}} = \frac{(\sum \hat{\mu}_i T_i)^2}{\sum \hat{\mu}_i^2 n_i} = \sum \hat{\mu}_i^2 n_i. \tag{9}$$

This is actually the LR test in the generic normal model. It is $\chi^2_{(1)}$-distributed when the alternative is restricted to a line $\mu \boldsymbol{w}$ as in T_{sp}, but has other distributions for other kind of restrictions.

When there are no restrictions on $\boldsymbol{\mu}$ the mle $\hat{\mu}_i = \frac{T_i}{n_i}$ and the LR test, here called $T_u r$ for this unrestricted case assuming m classes is

$$\begin{aligned} T_{ur} &= T_{\boldsymbol{\mu}} & (10) \\ &= \sum (\frac{T_i}{n_i})^2 n_i & (11) \end{aligned}$$

$$= \sum_{i=1}^{m} \frac{T_i^2}{n_i} \tag{12}$$

which is $\chi^2_{(m)}$-distributed under H_0. In a meta analysis of association studies in *different* populations this statistic could be justified, since the associated marker allele could differ between populations, giving rise to large T_i, with different signs, that would otherwise cancel if T_{eq} was used.

While the unrestricted T_{ur} has the χ^2 disadvantage of sensitivity to deviations in all directions, the true alternative $(\mu_1, \mu_2, \ldots, \mu_m)$ could be too distant from the space defined by the somewhat arbitrary restrictions for T_{sp} resp T_{eq}.

We will continue by introducing alternatives that are more restricted than T_{ur} but less than T_{eq}.

The T_{neq} statistic

When the study classes are from the *same* population, we expect the same direction of the associations in all classes. This puts a minimum requirement on $\boldsymbol{\mu}$, namely the restriction $\mu_i \geq 0$ for all i or $\mu_i \leq 0$ for all i. The mle becomes either $\hat{\mu}_i^- = \max(0, \hat{\mu}_i)$ for all i or $\hat{\mu}_i^+ = \min(0, \hat{\mu}_i)$ for all i.

This gives

$$T_{neq} = \max\left(\sum_{T_i>0} \frac{T_i^2}{n_i}, \sum_{T_i<0} \frac{T_i^2}{n_i}\right) \tag{13}$$

The null distribution is found by conditioning on N^+, the number of non-negative T_i. Under H_0 N^+ is binomial$(m, \frac{1}{2})$ distributed. The two sums in (13) are both χ^2 distributed, further they are independent. By defining $\chi^2_{(0)}$ as the constant 0

$$\begin{aligned} F_{T_{neq}(x)} &= \sum_{i=0}^{m} P(T_{neq} \leq c | N^+ = i) P(N^+ = i) \\ &= \frac{1}{2^m} \sum_{i=0}^{m} \binom{m}{i} F_{\chi^2_{(i)}}(x) F_{\chi^2_{(m-i)}}(x). \end{aligned} \tag{14}$$

The T_{so} statistic

If the model assumptions are weak, but sufficiently precise to give an order of the magnitudes of μ_i, the parameter space can be restricted accordingly. When the order reflects reality this additional restriction will typically result

in an improvement compared with T_{neq}. The transmissions in the sibs are in general more skewed. The estimates are achieved by isotonic regression (Barlow *et al.*, 1972) and the usual notation for the isotonic mle is μ_i^*.

In the case of a *simple order*, $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_m$, the estimates are calculated by using PAVA, Pool Adjacent Violators Algorithm (**?**). This setup is different in two ways. There is the additional restriction $0 \leq \mu_1$ and the reverse order is also possible, thus

$$R = (0 \leq \mu_1 \leq \mu_2 \ldots \leq \mu_m) \cup (0 \geq \mu_1 \geq \mu_2 \ldots \geq \mu_m)$$

It is straightforward to adjust PAVA accordingly.

1. Get the preliminary estimates $\mu_i^{*<}$ and $\mu_i^{*>}$ by applying the ordinary PAVA on the simple order $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$ ($\mu_i^{*<}$) and its reverse ($\mu_i^{*>}$).

2. Adjust for the restriction that the estimates must have equal signs by $\mu_i^{*+} = \max(0, \mu_i^{*<})$ and $\mu_i^{*-} = \min(0, \mu_i^{*>})$

3. Compare the two likelihoods and let

$$\hat{\boldsymbol{\mu}}^{**} = \begin{cases} \boldsymbol{\mu}^{*+} & \text{if } L(\boldsymbol{\mu}^{*+}) \geq L(\boldsymbol{\mu}^{*-}) \\ \boldsymbol{\mu}^{*-} & \text{otherwise} \end{cases}$$

Then

$$T_{so} = \sum n_i (\hat{\mu}_i^{**})^2 \tag{15}$$

is the LR statistic. In a case with only two classes

$$T_{so} = \begin{cases} \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} & \text{if } \hat{\boldsymbol{\mu}} \in R \\ \max(\frac{(T_1+T_2)^2}{n_1+n_2}, \frac{T_2^2}{n_2}) & \text{otherwise} \end{cases} \tag{16}$$

In the standard isotonic regression, where $H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$ is tested against $H_1 : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_m$ with some strict inequality, the distribution of the LR test statistic denoted $\bar{\chi}^2$ is a mixture of χ^2 distributions, but this test has no such simple structure. Although the p-values for an observed value t_{so} can be found by integrating the normal distribution function over the m-dimensional rejection area defined by $P_{H_0}(T_{so} \leq t_{so})$, simulating the null distribution is easier and it is also more accurate when the sample sizes

are too small for the asymptotic distribution to be good in the actual binomial case.

The more important question is how to guess the order. The relationships relies entirely on the relations between the different β_i. By using expressions derived in Appendix and some algebra these relationships can be investigated

In a monogene disease $\beta_1 \geq \beta_3$ is equivalent to

$$(f_1 q + f_2 p)(f_0 q + f_1 p)((f_1 - f_0)q + (f_2 - f_1)p) \geq 0. \tag{17}$$

The first factor is > 0. The second factor is ≥ 0 with equality in a recessive model with no phenocopies $(0, 0, f_2)$. A sufficient, but not necessary, criteria for the third factor to be nonnegative is that the penetrances have the "natural" order $f_0 \leq f_1 \leq f_2$, i.e. no heterozygous advantage or disadvantage.

Furthermore $\beta_2 \geq \beta_3$ is equivalent to

$$f_1 f_2 (f_2 - f_1)p^2 + f_0 f_1 (f_1 - f_0)q^2 + f_0 f_2 (f_2 - f_0)pq \geq 0 \tag{18}$$

which is always true when $f_0 \leq f_1 \leq f_2$.

Finally $\beta_1 \geq \beta_2$ is equivalent to

$$(f_1^2 - f_0 f_2)(f_1(f_1 - f_0) + f_0(f_2 - f_1) + 2p(f_1 - f_0)(f_2 - f_1)) \geq 0. \tag{19}$$

The second factor is > 0 in natural penetrance orders and the first factor is 0 in the genetic models with a constant genotype relative risk

$$\gamma = \frac{f_1}{f_0} = \frac{f_2}{f_1}.$$

So this specific model, which is so mathematically convenient in power calculations (Risch and Merikangas, 1996), since parents transmissions to children become independent, is thus the border of choosing whether to prefer 1 IBD families or 2 IBD families. By insertion of parameters for dominant resp. recessive model into (19) it follows, what was of course intuitively immediately clear, that 1 IBD families are best for dominant models, while 2 IBD families are best for recessive models.

Further subclasses

In a sense the philosophy behind only using IBD transmissions from heterozygous parents is to remove data that are less informative. Is a similar

strategy possible with the trios? The triple heterozygous families contribute to the same increase of both N_{12} and N_{21}. If they are removed the numerator in the statistic will be the same, but the denominator decreases. The statistic will however no longer be $\chi^2_{(1)}$, but that can be corrected for.

Single out the double heterozygous matings and construct a TDT statistic from these families. Assume there are c_{11} families with a 11 child, c_{12} with 12 and c_{22} with 22. The TDT statistic of this subclass becomes

$$T' = \frac{(2c_{22} - 2c_{11})^2)}{2c_{22} + 2c_{12} + 2c_{11}} = \frac{2(c_{22} - c_{11})^2}{c_{22} + c_{12} + c_{11}} \tag{20}$$

If now the triple heterozygous families are discarded from the subsample, we can compare c_{11} with c_{22} by

$$T'' = \frac{(c_{22} - c_{11})^2}{c_{22} + c_{11}}. \tag{21}$$

Both statistics are asymptotically $\chi^2_{(1)}$ under H_0 and it is easy to see that $T'' > T'$ iff $c_{12} > c_{11} + c_{22}$. However from the triangle restriction (Holmans, 1993) $E[c_{12}] \leq E[c_{11} + c_{22}]$, so the strategy is no good.

It could still be possible though to split the class into two subclasses 3.1 and 3.2 with this principle and a similar approach can be applied for the 2 IBD class, but this is not done in the paper.

Comparisons

It is not so clear how the different statistics should be compared. The power can be calculated before sample collection, when the collection has been done, when the IBD status of the sib pairs have been found and finally when the number of heterozygous transmissions have been counted in each class. The latter has been chosen here since it is the point, when as much information as possible is available for guessing the model. In Table (IV) the results of simulated powers for some different statistics is shown. The significance level is everywhere 0.01 and the critical value for this level comes from simulations under the null hypothesis for every model. The proportions of sib pairs to trios have been set to 1:3. The number of families has been chosen so that the power for $T_{eq} \approx 0.50$. The numbers of heterozygous transmissions in each category have been set to the expected numbers rounded up to an integer. The statistic T_{ref} is a statistic with weights $\boldsymbol{w} = (\beta_1, \beta_2, \beta_3)$, which should

Table IV: Power simulations for some different models

model	#ASP	T_{eq}	T_{sp}	T_{neq}	T_{so1}	T_{so2}	T_{so3}	T_{ref}	T_{ASP}
DP50	50	0.48	0.63	0.49	0.56	0.57	0.57	0.63	0.36
DP20	150	0.52	0.74	0.83	0.86	0.84	0.77	0.91	0.88
DH50	80	0.53	0.55	0.41	0.47	0.47	0.49	0.53	0.15
DH20	500	0.50	0.52	0.42	0.44	0.46	0.45	0.53	0.18
RP50	40	0.56	0.62	0.48	0.54	0.56	0.53	0.59	0.39
RP20	90	0.48	0.70	0.63	0.70	0.72	0.43	0.77	0.74
RH50	50	0.50	0.47	0.36	0.37	0.42	0.41	0.48	0.16
RH20	340	0.50	0.53	0.45	0.40	0.46	0.36	0.57	0.16

$T_{so1} : \mu_3 \leq \mu_2 \leq \mu_1, T_{so2} : \mu_3 \leq \mu_2 = \mu_1, T_{so3} : \mu_3 = \mu_2 \leq \mu_1$ The models are described in Table III.

be close to optimal and T_{ASP} is a statistic that does not use the trios at all ($\boldsymbol{w} = (1, 1, 0)$). Three statistics based on isotonic regression are investigated. They all assume that $0 \leq \mu_3 \leq \mu_1$ or the reverse and differ by the position of μ_2. In T_{so1} it is between the other two, in T_{so2} $\mu_2 = \mu_1$ and in T_{so3} $\mu_2 = \mu_3$. From the simulations some observations can be made

- T_{sp} is to be preferred before T_{eq}. In fact T_{sp} competes well with T_{ref}.
- T_{neq} is generally not as good as T_{eq} except in the models DP20 and RP20 where most of the trios are phenocopies. In these models the trios do more harm then good, which is most evident when looking at the power for T_{ASP}.
- The statistics that are based on isotonic regression are in general not as good as T_{sp}, partly because the restriction is not correct in all models, but also because of "more degrees of freedom". The models DP20 and RP20 are the exceptions here as well. In DP20 the order is correct for T_{so1} and the differences in μ_i sufficiently large. In RP20 μ_1 and μ_2 are rather close and much larger then μ_3, which explains why T_{so1} and T_{so2} are significantly better than T_{so3}.

An example of a collection of families with psoriasis is shown in table V. The families are from south west of Sweden and the marker is IAP80 located at 3q21. The different statistics are $T_{eq} = 5.7$ (p=0.021), $T_{sp} = 8.4$ (p=0.004),

Table V: TDT tables: Psoriasis example

1 IBD

	1	2
1		6
2	3	

2 IBD

	1	2
1		11
2	1	

Trios

	1	2
1		50
2	38	

$T_{neq} = 11.0$ (p=0.004), $T_{so1} = 9.7$ (p=0.004), $T_{so2} = 9.7$ (p=0.004), $T_{so3} = 5.8$ (p=0.028), $T_{ASP} = 8.0$ (p=0.007). There is a remarkable difference in the transmission distortion. It is interesting to note that if we take T_{ASP} as a base for comparisons, the addition of all the trios reduce the evidence if we use T_eq or T_{so3}, while it increases slightly when the other are used. T_eq does not take into account that the sib pairs are generally better and T_{so3} is designed for a dominant alternative, while inspection of data reveals that it looks very much recessive.

Discussion

Tests statistics similar to T_{sp} (Martin *et al.*, 1997) have been derived here from the perspective of genetic models with phenocopies and/or locus heterogeneity. The transmissions from heterozygous parents have been classified according to their origin from trios, ASP families with 1 IBD and ASP families with 2 IBD.

The different classes show clearly different transmission distortion. In phenocopy models or locus heterogeneity models the selected transmissions in the sib pairs are more valuable than the transmissions in the trios. For this reason T_{sp}, which puts twice as much weight on the sibpairs is generally more powerful than T_{eq} which uses equal weights.

Although the examples shown in the paper are limited to monogenic disorders with phenocopies and two-locus heterogeneity models, the usefulness of the strategies is much wider. If the susceptibility gene at the locus that is tested for association is not mandatory, there will be different skewness in different TDT classes.

We investigated some statistics based on isotonic regression, where the restrictions were variants of simple order. Unless there is a large difference in the skewness between the different TDT classes and the order is correctly

guessed, these statistics are less powerful than a predefined set of reasonable weights. Other restrictions like the partial order $0 \leq \mu_3 \leq [\mu_1, \mu_2]$ (or reverse order) which means that the trios are not as skewed as any of the other could also be used. Another possible restriction is $0 \leq \mu_3 \leq \max(\mu_2, \mu_1)$ (or reverse), indicating that the trios can not be most skewed.

It should be noted that the linkage results, in particular the observed IBD-distribution, can be used as a guidance for guessing the inheritance model at a linked loci and then design a statistic based on a suitable restriction for that model.

There are other ways of partitioning the observed transmissions into subclasses. One argument is to take the affection status of the parents into account. In the case of exactly one affected parent it could make sense to include an IBD-transmission only if it comes from the affected parent and exclude it otherwise.

With more than two affected sibs a direct generalisation is to consider only transmission that are IBD in *all* siblings. The transmission probabilities will in general not be the same as with a single ASP, but to avoid too many classes it seems reasonable two merge the observations from all families.

When the problem with the trios is due to a high rate of phenocopies, it is possible to decrease the phenocopy rate by having a family history as an inclusion criteria. When such strategies are used difference in transmission probabilities will decrease.

In conclusion the T_{sp} statistic with its ad hoc weights (2,2,1) behaves well in the models that have been simulated, but so would a more speculative set like $(\pi, e, 1)$ do. The investigator must, as always, make a choice based on whatever knowledge is available.

Appendix

Derivation of TDT cell probabilities

Assume that there is a disease locus with allele frequencies a_k and a marker locus with allele frequencies m_i. The marker alleles and disease locus alleles form haplotypes with frequencies h_{ik}. The marker and the disease locus might be in disequilibrium with the coefficient $d_{ik} = h_{ik} - m_i a_k$. Consider an IBD transmission in an affected sib pair with one or two IBD-alleles and let M_{ij} be the event that the parent transmitted marker allele i and kept marker allele j. Let A_{kl} be defined similarly for the disease locus. Also let X be the IBD count at the marker locus in the sib pair and $b_k^{(n)} = P(\text{ASP}|A_{kl}, X = 1)/P(\text{ASP}|X = 1)$. Assume that there are no recombinations between the marker and the disease locus. Then

$$
\begin{aligned}
p_{ij}^{(n)} &= P(M_{ij}|X = 1, ASP) \qquad (22)\\
&= \sum_{k,l} P(M_{ij}|A_{kl}, X = 1, ASP) P(A_{kl}|X = 1, ASP)\\
&= \sum_{k,l} P(M_{ij}|A_{kl}) \frac{P(ASP, A_{kl}, X = 1)}{P(ASP, X = 1)}\\
&= \sum_{k,l} P(M_{ij}|A_{kl}) \frac{P(ASP|A_{kl}, X = 1) P(A_{kl}|X = 1)}{P(ASP|X = 1)}\\
&= \sum_{k,l} \frac{h_{ik} h_{jl}}{a_k a_l} \frac{P(ASP|A_{kl}, X = 1) a_k a_l}{P(ASP|X = 1)}\\
&= \sum_{k,l} (m_i a_k + d_{ik})(m_j a_l + d_{jl}) b_k^{(n)}\\
&= m_i m_j + \sum_{k,l} (m_i a_k d_{jl} + m_j a_l d_{ik} + d_{ik} d_{jl}) b_k^{(n)}\\
&= m_i m_j + m_j \sum_k d_{ik} b_k^{(n)}.
\end{aligned}
$$

The derivation of the probabilities for singletons is similar. If the TDT class for trios is denoted 3, then $b_k^{(3)} = P(\text{Aff}|A_{kl})/P(\text{Aff})$.

The bivariate distribution of the two parents contribution to the TDT table can be derived in a way analogous to (23).

$$
p_{ij,i'j'}^{(2)} = P(M_{ij}, M_{i'j'}|X = 2, \text{ASP}) \qquad (23)
$$

$$
\begin{aligned}
&= \sum_{k,l,k',l'} P(M_{ij}, M_{i'j'}|A_{kl}, A_{k'l'}, X=2, \text{ASP}) P(A_{kl}, A_{k'l'}|X=2, \text{ASP}) \\
&= m_i m_j m_{i'} m_{j'} + m_j m_{j'} \sum_{k,k'} (m_i a_k d_{i'k'} + m_{i'} a_{k'} d_{ik} + d_{ik} d_{i'k'}) b^{(2)}_{kk'}
\end{aligned}
$$

with

$$
b^{(2)}_{kk'} = \frac{P(ASP|IBD=2, A_{kl}, A_{k'l'})}{P(ASP|IBD=2)}
$$

In a locus heterogeneity model b_k and $b_{kk'}$ are found by conditioning on the number of sibs, S, in the pair that are affected through the other mechanism. Let $\gamma_i = P(S=i)$ and $p_{ki} = P(\text{ASP}|IBD=1, A_{kl}, S=i)$. Then

$$
b^{(1)}_k = \frac{\sum_{i=0}^{2} p_{ki}\gamma_i}{\sum_k \sum_{i=0}^{2} p_{ki}\gamma_i a_k} \tag{24}
$$

where $\gamma_2 = K_{2p}K_{2s}, \gamma_1 = 2K_{2p}(1-K_{2s}), \gamma_0 = 1 - K_{2p}(2-K_{2s})$.

The necessary entities for trios and for ASP families with 2 IBD are calculated in a similar way using the probabilities in Table VI

Table VI: *Conditional probabilities*

k	k‘	i	p_{ki}	$q_{kk'i}$	$r_{kk'i}$
1	1	0	$(f_1q + f_2p)^2$	f_2^2	f_2
1	1	1	$f_1q + f_2p$	f_2	1
1	2	0		f_1^2	f_1
1	2	1		f_1	1
2	1	0	$(f_0q + f_1p)^2$	f_1^2	f_1
2	1	1	$f_0q + f_1p$	f_1	1
2	2	0		f_0^2	f_0
2	2	1		f_0	1
.	.	2	1	1	NA

$p_{ki} = P(\text{ASP}|IBD=1, A_{kl}, S=i)$, $q_{kk'i} = P(\text{ASP}|IBD=2, A_{kl}, A_{k'l'}, S=i)$,
$r_{kk'i} = P(\text{Aff}|A_{kl}, A_{k'l'}, S=i)$

When the disease locus has only two alleles define $\beta_c = b^{(c)}_1 - b^{(c)}_2$

$$
\beta_1 = \frac{\gamma_1(f_1q + f_2p - f_0q - f_1p) + \gamma_0((f_1q + f_2p)^2 - (f_0q + f_1p)^2)}{\gamma_2 + \gamma_1 K_{1p} + \gamma_0(p(f_1q + f_2p)^2 + q(f_0q + f_1p)^2)}
$$

$$\beta_2 = \frac{\gamma_1(f_2p + f_1q - f_0q - f_1p) + \gamma_0(f_2^2p + f_1^2q - f_0^2q - f_1^2p)}{\gamma_2 + \gamma_1 K_{1p} + \gamma_0(f_2^2p^2 + 2f_1^2pq + f_0^2q^2)}$$

$$\beta_3 = \frac{(1 - K_{2p})(f_1q + f_2p - f_0q - f_1p))}{K_{1p} + K_{2p} - K_{1p}K_{2p}}$$

The cell probabilities $p_{ij}^{(c)}$ are calculated under the assumption of no recombination between the marker and the disease locus, which is relevant when $\theta = 0$. For a general recombination fraction θ, condition on the recombination events and find

$$p_{\theta ij}^{(3)} = (1 - \theta)p_{ij}^{(3)} + \theta p_{ji}^{(3)} \tag{25}$$

$$p_{\theta ij}^{(c)} = (1 - \theta)^2 p_{ij}^{(c)} + \theta^2 p_{ji}^{(c)} + 2\theta(1 - \theta)r_{ij}^{(c)}, c = 1, 2 \tag{26}$$

where $r_{ij}^{(c)} = P(M_{ij}|\text{recombination in one sib}, X = c, ASP)$. From (25) the important feature of the original TDT to only detect association when there is linkage follows by noting that $p_{\frac{1}{2}ij}^{(3)} = p_{\frac{1}{2}ji}^{(3)}$. For symmetry reason $r_{ij}^{(c)} = r_{ji}^{(c)}$, so property is present in the ASP classes as well.

References

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference under Order Restrictions*. John Wiley & Sons, London.

Ewens, W. and Spielman, R. (1995). The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet*, **57**, 455–464.

Falk, C. and Rubinstein, P. (1987). Haplotype relative risk: an easy reliable way to construct a proper control sample for rsik calculations. *Ann Hum Genet*, **51**, 227–233.

Holmans, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *Am J Hum Genet*, **52**, 362–374.

Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, **49**, 49–67.

Martin, E., NL, K., and Weir, B. (1997). Tests for linkage and association in nuclear families. *Am J Hum Genet*, **61**, 439–448.

Risch, N. (1990). Linkage strategies for genetically complex traits. i. multilocus models. *Am J Hum Genet*, **46**, 222–228.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley and Sons Ltd.

Schaid, D. (1996). General score tests for association of genetic markers with disease using cases and parents. *Genet Epidemiol*, **13**, 423–450.

Sham, P. and Curtis, D. (1995). An extended transmission/disequilibrium test (tdt) for multi-allele marker loci. *Ann Hum Genet*, **59**, 323–336.

Spielman, R. and Ewens, W. (1996). The tdt and other family based tests for linkage disequilibrium and association. *Am J Hum Genet*, **59**, 983–989.

Spielman, R., McGinnis, R., and Ewens, W. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, **52**, 506–516.

Terwilliger, J. and Göring, H. (2000). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology*, **72**, 63–132.

Thomson, G. (1995). Mapping disease genes: family-based association studies. *Am J Hum Genet*, **57**, 487–498.

Permutation based multiple inference in association studies with case-control design

Staffan Nilsson

October 18, 2001

Abstract

In testing many hypothesis, which is typically done in genetic association studies with a large number of investigated loci, the individual p-values have to be interpreted in a multiple inference framework. A simple Bonferroni correction is usually too crude. We propose the use of step-down tests based on the permutation distribution, which can be achieved in a case control study by randomly assigning the labels case and control.

Introduction

Testing many hypothesis is always a problem when it comes down to sound interpretation of the achieved nominal p-values. In such genetic studies where the aim is to localise susceptibility genes a large number of tests is a rule rather than an exception. The fast development of technology to produce massive amount of data on many items will continually make this problem even more pronounced.

A typical genetic screening study consists of genotyping a large number of loci and then performing many statistical tests for specific questions adapted to the chosen study design. The results must then be interpreted in a multiple inference framework.

When family data are collected and linkage analysis is performed there exist guidelines, based on theoretical modelling of the recombination process, that permits an infinite number of markers to be tested (Lander and Kruglyak, 1995).

For association analysis too little is known about the linkage disequilibrium and its variation over the genome and between populations to have a similar theoretical approach. In (Risch and Merikangas, 1996) the crude Bonferroni method was used to correct for the huge amount of SNPs investigated.

In this paper a method is presented for dealing with multiple inference when searching for association in a classical case-control design. It uses a permutation step down strategy (Westfall and Young, 1993) to incorporate the dependencies between test statistics on tightly linked markers as well as the dependencies between different test statistics on the same marker. In (Westfall *et al.*, 2001) a closely related approach to the problem is presented.

Methods

The setup that will be considered is a case-control study with m cases and n controls that are genotyped at k loci. The k null hypothesis $H_1, ..., H_k$, where H_i is "no genetic effect of loci i", are tested and achieve corresponding nominal p-values $p_1, ..., p_k$. The effect can be directly functional or caused by linkage disequilibrium to a functional loci.

The key concept when multiple tests are performed is the multiple significance level also known as the family wise error rate (FWE). The interpretation of FWE=α for a set of tests is that the a priori risk that any of the true null hypothesis are rejected is at most α.

For any general set of hypotheses the crude and simple method for multiple inference is the Bonferroni method where, within FWE=α, all hypotheses H_i corresponding to nominal p-values $p_i \leq \alpha/k$ can be rejected. An improved version of the Bonferroni method, which is also valid for any set of hypotheses and almost as easy to apply is the Holm(-Bonferroni) (Holm, 1979) method. In this test the p-values are first sorted in increasing order $p_{(1)} \leq p_{(2)} ... \leq p_{(k)}$. The algorithm then sequentially compares $p_{(i)}$ to α/i for $i = 1, 2, \ldots, k$ and rejects the corresponding hypothesis until an i is reached where $p_{(i)} > \alpha/i$. The procedure stops and as a result $H_{(1)}, ..., H_{(i-1)}$ are rejected, while $H_{(i)}, ..., H_{(k)}$ are accepted.

The Holm method can be explained by considering the principle of closed tests (Marcus *et al.*, 1976).

In a closed test the k base hypotheses and all their $2^k - k - 1$ possible nonempty intersections are tested preliminary with a level α-test. A base

hypothesis H_i is then finally rejected at FWE=α if all intersections of hypotheses that include H_i, and H_i itself, are rejected by the preliminary tests.

If, for all index sets τ, the corresponding hypotheses intersection $H_\tau = \cap_{i\in\tau} H_i$ is tested with $\min_{i\in\tau} p_i$, where p_i is the p-value of the preliminary test of H_i, then the Holm method follows by applying Boole's inequality in all intersection tests, i.e. by using Bonferroni preliminary tests.

Although the Holm method is an improvement over the original Bonferroni, it is clear that the power to reject at least one hypothesis is not improved and while it works well with nearly independent test statistics, its performance can be poor when the test statistics are highly correlated.

Consider a multiple hypotheses situation with a case-control design. Rather than setting the threshold through the Boole's inequality in the Holm test, it can be taken from the permutation distribution (Westfall and Young, 1993), i.e. by calculating the statistics for all $\binom{m+n}{m}$ possible ways of assigning m labels marked "case" on $m+n$ individuals. For smaller sample sizes all these different permutations can be generated and for larger sample sizes the permutation distribution can be simulated.

When the hypothesis intersection H_τ is tested with $T_\tau = \min_{i\in\tau} p_i$ with a permutation procedure, the distribution of the corresponding genotype vector $\mathbf{G}_\tau$ must be identical in cases and controls under H_τ to justify the permutation of labels and the calculation of p-values for T_τ from the permutation distribution. Thus it does not suffice that the distribution of G_i are equal in cases and controls for each single $i \in \tau$. In theory it is possible to construct joint genotype distributions where the marginal distributions of each of the genotypes at the individual loci are equal in cases and controls, but the joint distributions differs, although this would correspond to both a strange genetic disease model and a coincidence of frequency parameters for which there is no natural evolutionary force.

However a natural interpretation of "no genetic effect" for all hypotheses H_i where $i \in \tau$ is that the joint distribution of $\mathbf{G}_\tau$ is the same in both cases and controls. The definition of

$$H_i : \text{No genetic effect of loci } i$$

has to be consistent with this interpretation and requirement.

Definition 1 (No genetic effect) *A loci i has no genetic effect if the following statements are true*

1. $G_i^{case} \stackrel{d}{=} G_i^{control}$

2. $(G_i, \mathbf{G}_\tau)^{case} \stackrel{d}{=} (G_i, \mathbf{G}_\tau)^{control}$ *for all index sets* τ *where* $\mathbf{G}_\tau^{case} \stackrel{d}{=} \mathbf{G}_\tau^{control}$

otherwise loci i has a genetic effect.

The property of equal distribution in cases and control for the joint distribution of G_τ in any set τ of true null hypothesis follows by induction from this definition.

The distribution of G_τ under the null hypotheses intersection H_τ can now be estimated by the permutation distribution. The permutation step down procedure (PSTD) (Westfall and Young, 1993) becomes

1. Calculate the k p-values corresponding to the k ordinary tests of the k hypotheses and order them such that $p_{(1)} \leq p_{(2)} ... \leq p_{(k)}$. Name the corresponding null hypotheses $H_{(1)}, H_{(2)}, ..., H_{(k)}$.

2. Permute the case-control labels randomly and independently some large number n times and calculate p-values $p_{\pi i}$ for each hypothesis $H_{(i)}$ and permutation π.

3. Calculate $p'_{\pi i} = \min_{i \leq j \leq k} p_{\pi j}$ for $i = 1$ to k and for each π.

4. Let $p'_i = \frac{1}{n} \sum_\pi I(p'_{\pi i} \leq p_{(i)})$ where I is the indicator function.

5. Let $p_{i,c} = \max_{1 \leq j \leq i} p'_j$.

Now $p_{i,c}$ is the p-value corresponding to $H_{(i)}$ and corrected for multiple inference, i.e. rejection of $H_{(i)}$ for all i with $p_{i,c}$ has a FWE=α.

The p-values that are used in the first two steps can be viewed as test statistics. They can be approximate or taken from conditional tests. The validity of the step down procedure is motivated by the closed test procedure using a monotone structure among the H_τ tests. For this to hold it is not essential that the original p_i are p-values.

One of the advantages with Holm compared to ordinary Bonferroni is that a very false hypothesis, which surely will give a p-value less than $\alpha/(k+1)$

if it is tested, can be added to a set of k hypotheses at no cost, since it will not affect the rejection of them. PSTD shares that property.

When a hypothesis with an individual test of very low power is added to the other hypotheses in a multiple inference framework it will in general be harder to reject the false hypotheses. In Holm it will typically move all thresholds one step, but in the permutation procedure a completely meaningless test (individual power = 0 due to the discrete nature of the possible p-values) does not affect the other tests at all since it will never affect the distribution of the minimum p-value statistic in its lower α tail.

In Table 1 the result from an investigation of a number of SNPs in a candidate gene HCR (α-helix coiled-coil rod homolog gene) for psoriasis is shown. In addition to HCR the presence of the CW6-allele in the HLA-C locus, which is known to associate with psoriasis, was examined. A total of 42 cases and 38 controls were tested (O'Brien *et al.*, 2001). We calculated the nominal p-values by Fishers exact test on a 2x2 contingency table with counts of the presence of the rare nucleotide in cases and controls.

The obvious software implementation in this situation is to precalculate all possible (not many) nominal p-values for each polymorphism based on the marginal counts and then use a simple table lookup in each permutation. In (O'Brien *et al.*, 2001) we implemented the procedure in Splus and more involved test statistics than Fisher p-values were also used. For the simple Fisher calculation though a very fast permutation step-down procedure is already implemented by the PROC MULTTEST in SAS (Westfall *et al.*, 1999), which was used to create the data in Table 1 from 10 million simulations.

In the columns marked Holm(26) and PSTD(26) all 26 polymorphic loci are included in the multiple inference. The difference in p-values is something like a factor 3 in favour of PSTD for the relevant SNPs. The careful investigator only using Holm could improve the situation by noting that there are 9 polymorhisms where 5 or less individuals (pooling cases and controls) have the rare allele. These polymorphisms can be removed without embarrassment since they can not even reach a nominal p-value of 0.05 with a 5-0 split between cases and controls. In the columns Holm(17) and PSTD(17) the results of this operation can be seen. The corrected p-values in the Holm procedure are obviously improved with the reduction of the number of tests, but PSTD is unaffected since it has no problem with these polymorphisms when they are included in PSTD(26).

Table 1: HCR SNPs

	frequency		p-values				
	case	ctrl	nominal	Holm(26)	PSTD(26)	Holm(17)	PSTD(17)
CW6	29	7	0.00000	0.00014	0.00004	0.00009	0.00004
B384	9	2	0.05106	0.76601	0.25726	0.30640	0.21221
B386	29	14	0.00670	0.14084	0.04651	0.08048	0.04651
B404	29	14	0.00670	0.14084	0.04651	0.08048	0.04651
B556	4	11	0.04282	0.68521	0.18552	0.29978	0.16921
B571	27	25	1.00000	1.00000	1.00000	1.00000	1.00000
B614	5	0	0.05626	0.78770	0.25726		
B850	17	11	0.35021	1.00000	0.91339	1.00000	0.77922
B904	13	24	0.00670	0.14084	0.04651	0.08048	0.04651
B1178	2	1	1.00000	1.00000	1.00000		
B1328	41	23	0.00003	0.00074	0.00025	0.00046	0.00025
B1329	3	1	0.61729	1.00000	0.99758		
B1354	11	11	0.80702	1.00000	0.99887	1.00000	0.97940
B1364	36	14	0.00000	0.00019	0.00005	0.00012	0.00005
B1426	0	2	0.22246	1.00000	0.75472		
B1471	9	0	0.00262	0.05776	0.01843	0.03413	0.01843
B1579	1	1	1.00000	1.00000	1.00000		
B1601	1	1	1.00000	1.00000	1.00000		
B1802	31	17	0.01177	0.21193	0.06785	0.10596	0.06785
B1959	4	3	1.00000	1.00000	1.00000	1.00000	1.00000
B1990	21	11	0.06936	0.90178	0.26725	0.34684	0.23226
B2254	1	0	1.00000	1.00000	1.00000		
B2257	3	11	0.01669	0.28381	0.06977	0.13356	0.06977
B2277	1	0	1.00000	1.00000	1.00000		
B2278	1	0	1.00000	1.00000	1.00000		
B2406	36	19	0.00072	0.01657	0.00494	0.01008	0.00494

Choice of test statistics

The data from the experiment at one locus can be summarised in a 2xk contingency table, where k is the number of possible genotypes at the locus, i.e. $k = \frac{a(a+1)}{2}$ for a locus with a alleles. So typically, for an SNP, $k = 3$

(Table 2). Statistics are usually best applied on collapsed versions (Table 3).

Table 2: Genotype counts for cases and controls

	11	12	22
Case	n_{11}	n_{12}	n_{22}
Control	m_{11}	m_{12}	m_{22}

Table 3: Collapsed tables

	DOM		REC		ALLELE	
	11	12/22	11/12	22	1	2
Case	n_{11}	$n_{12}+n_{22}$	$n_{11}+n_{12}$	n_{22}	$2n_{11}+n_{12}$	$2n_{22}+n_{12}$
Control	m_{11}	$m_{12}+m_{22}$	$m_{11}+m_{12}$	m_{22}	$2m_{11}+n_{12}$	$2m_{22}+m_{12}$

From these 3 tables we can perform 3 different tests of homogeneity corresponding to 3 null hypothesis

$$H_{iD} = H_{iR} = H_{iA} : \text{No genetic effect of loci } i$$

where the letters D, R, A in the indices denote that the tests are adapted to a dominant, recessive resp allele count alternative. In the example in Table 1 the permutation strategy accounted for the dependencies between statistics on different loci. If there is doubt about the best (most powerful) way to collapse Table 2 then H_{iD}, H_{iR} and H_{iA} can be used simultaneously. Whatever test statistic, Fisher or χ^2, that is used on the corresponding $2x2$ tables, there will be a fair amount of dependencies between them. However, if the closed set principle is applied to the set of basic hypotheses H_{iD}, H_{iR}, H_{iA} for the loci $i = 1, \ldots, k$, this dependency will be accounted for by the permutation strategy. In Table 4 the effect of using all three tests simultaneously is shown. The loci used for illustration is a subset of the HCR polymorphisms in Table 1.

The Fishers exact test has been used to get the p-values corresponding to contingency tables for dominant and recessive models. This test fits well with

the permutation strategy since the p-values are exactly those that would be found empirically if each permutation of labels was used exactly once. For that contingency table, where the alleles are counted, the usual Fisher exact test will not correspond to such empirical permutation p-value. Therefore we have used a similar permutation test that uses one of the cell counts (e.g. $2n_{11} + n_{12}$) as a test statistic, conditional on the pooled genotype counts rather than the allele counts.

Compared to the situation where only one model is used the corrected p-values are increased by a factor 1.5. This price has to be paid to avoid guessing the wrong model, but the price is at least not a factor 3, which would be the case with Bonferroni corrections.

It should be remembered that Table 2 was collapsed to get more powerful tests on the 2x2 tables. If *all three* tests are used simultaneously this gain might be lost. It is merely an illustration of how the permutation strategy can handle the dependencies. However the multiple test can immediately be

Table 4: Three different ways of collapsing genotype counts

	Nominal p-values			Corr. one by one			Corr. simultaneously		
SNP	Dom	Rec	Allele	Dom	Rec	Allele	Dom	Rec	Allele
B386	0.0067	0.0563	0.0014	0.0329	0.1176	0.0068	0.0556	0.3031	0.0103
B556	0.0428	0.6018	0.0718	0.14	0.9712	0.2168	0.2566	0.9950	0.3179
B571	1.0000	0.0303	0.1813	1	0.1113	0.4611	1	0.2033	0.6745
B850	0.3502	0.4159	0.7267	0.6721	0.7531	0.8859	0.8947	0.9104	0.9950
B904	0.0067	0.0007	0.0003	0.0329	0.0014	0.0018	0.0556	0.0063	0.0026
B1471	0.0026	1.0000	0.0026	0.0129	1	0.0129	0.0208	1	0.0208
B1802	0.0118	0.2687	0.0088	0.0451	0.7531	0.0352	0.08	0.8947	0.0580
B1959	1.0000	1.0000	1	1	1	1	1	1	1.0000
B1990	0.0694	0.6018	0.1728	0.194	0.9712	0.4611	0.3179	0.995	0.6745
B2257	0.0167	1.0000	0.0482	0.0547	1	0.1863	0.0994	1	0.2735
B2406	0.0007	0.0594	0.0003	0.0043	0.1363	0.0023	0.0063	0.3179	0.0032

improved by noting the obvious constraint that whenever one in the group $\{H_{iD}, H_{iR}, H_{iA}\}$ is false, the other two must be false as well. Use of logical constraints with step down tests was introduced by (Shaffer, 1986) and its use with a permutation strategy is described in (Westfall and Young, 1993).

With these most simple constraints the step-down procedure is easily revised. When the first hypothesis in a group is rejected the other two are also rejected regardless of their individual p-values. All three are removed

from the set of hypothesis to be tested in the next step. This is a uniformly more powerful statistic than the procedure where the logical constraint was not used, but the probability to reject at least one hypothesis is however the same. In the section *Haplotype testing* below is another example of more involved logical constraints.

Another approach that will give identical result is to formulate only one single hypothesis

$$H_i : \text{No genetic effect of loci } i$$

and use $T_i = \min(p_{iD}, p_{iR}, p_{iA})$ as test statistic.

In Table 5 the effect of using only one hypothesis and the combined test statistic T_i is illustrated on five polymorhisms from the HCR example. The corrected p-values $p_{i,c}$ when 5 hypotheses are tested simultaneously with $T_i = \min(p_{iD}, p_{iR}, p_{iA})$ are compared to the smallest of the corrected p-values $p_{iD,c}, p_{iR,c}, p_{iA,c}$ when the $5*3$ hypotheses H_{ij} are tested without utilising the logical constraints.

Table 5: Min-of-models strategy

SNP	$\min(p_{D,c}, p_{R,c}, p_{A,c})$	p_c
B556	0.0943	0.0636
B614	0.1230	0.0636
B1471	0.0091	0.0091
B1802	0.0273	0.0273
B2257	0.0337	0.0273

Weighting the original p-values

In a simple Bonferroni it is free to choose any set of thresholds $\alpha_1, \alpha_2, \ldots, \alpha_k$ corresponding to some tests of $H_1, H_2, \ldots, H_k$ as long as $\sum \alpha_i = \alpha$, hereby increasing the power to reject some of the hypotheses, while decreasing it for others.

In a randomisation procedure this goal can be achieved by using weighted p-values rather than the original p-values as test statistics, by substituting each nominal p_i with $p'_i = p_i/w_i$ for some positive weights $w_1, \ldots, w_k$. There can be a variety of reasons for employing such a procedure. Some of the loci

could be more interesting prior to the study, either for functional reasons or when some of the loci have been indicated by earlier linkage or association studies.

Before performing the tests it is possible to legally "cheat" by looking at the pooled sample, hiding the labels for the investigator. In this inspection some variables can turn out to be more interesting than others. An SNP where only a few individuals carry the mutant allele is in general less interesting, since even if these individuals would all turn out to be cases it can not explain other genetic parameters seen from population data or linkage studies.

The marker loci of the cases will typically not be in Hardy-Weinberg equilibrium (HWE) in the neighbourhood of a disease locus. If the number of cases are sufficiently large compared to the controls there will not be HWE in the pooled sample either. The pooled sample can be tested for HWE in advance to get hints on which loci are most interesting. The weights can then be set as a function of some measure of the HWE. Such procedure would not be meaningful if there are a lot more controls than cases; there must be a reasonable power to detect deviation from HWE in the pooled sample.

In Table 6 an illustration of the weighting technique is shown. For each of the three collapsed contingency tables in Table 3, the weights are set to the minimum of the two marginal counts in the pooled sample. In the first two these counts are genotype counts, while the third table has marginal allele counts. The latter are actually divided by 2 to get the same range of weights.

Using Haplotypes

When several marker loci are tightly linked it can be useful to study haplotypes of alleles rather than the single alleles. This is done in EH (Xie and Ott, 1993) and EHPLUS (Zhao *et al.*, 2000), where the haplotype frequencies are estimated for the cases, the controls and for the pooled sample and the test statistic is the likelihood ratio

$$\Lambda = \frac{L_{case} L_{control}}{L_{pool}}. \tag{1}$$

If there are h different haplotypes then $2 \ln \Lambda$ is approximately $\chi^2_{(h-1)}$-distributed.

It is not easy to know in advance though how many loci to include in a haplotype test. If there are n such tightly linked loci and n is large, then

Table 6: Using weights on p-values

SNP	p_D	w_D	p_R	w_R	p_A	w_A	p_c	$p_{c,w}$
B386	0.0067	37	0.0563	5	0.0014	24	0.0107	0.0087
B556	0.0428	15	0.6018	3	0.0718	9	0.1687	0.1198
B571	1.0000	28	0.0303	18	0.1813	35	0.1510	0.1695
B850	0.3502	28	0.4159	6	0.7267	17	0.7213	0.4012
B904	0.0067	37	0.0007	9	0.0003	23	0.0025	0.0015
B1471	0.0026	9	1.0000	1	0.0026	5	0.0191	0.0406
B1802	0.0118	32	0.2686	8	0.0088	28	0.0555	0.0431
B1959	1.0000	7	1.0000	2	1.0000	4.5	1.0000	1.0000
B1990	0.0694	32	0.6018	3	0.1728	17.5	0.1855	0.1478
B2257	0.0167	14	1.0000	2	0.0482	8	0.0912	0.1149
B2406	0.0007	25	0.0594	8	0.0003	31.5	0.0020	0.0020

experience shows that the strongest statistical evidence of association is usually when some smaller number k of consecutive loci are analysed regarding haplotypes.

For a study with n such loci the investigator could be interested in testing n individual loci, $n-1$ consecutive pairs of loci, $n-2$ consecutive triplets etc, giving a total of $\frac{n(n+1)}{2}$ possible tests, with a great deal of dependency.

These tests can be dealt with in the same framework although computation time can become critical when, due to the phase unknown situation in case-control genotypes, the EM-algorithm is used for likelihood estimation in the calculation of the test statistics as is the case in EH.

The hypothesis used for extended genotypes spanning loci $i_1, i_2, \ldots, i_k$ must be analogue to the definition for single loci, i.e. the extended genotype $G_{i_1 \ldots i_k}$ shall replace G_i in a definition of "no effect of the extended locus $i_1, \ldots, i_k$". This implies that

$$H_{i_1 i_2 \ldots i_k} = H_{i_1} \cap \ldots \cap H_{i_k}.$$

There are logical constraints among the hypotheses that can be utilised to decrease the corrected p-values in the spirit of (Shaffer, 1986) applied on resampling-based testing in (Westfall and Young, 1993).

Let s and S be some index sets and let H_s and H_S be corresponding hypotheses intersections. Then

- If H_s is false then all H_S where $s \subseteq S$ are false.
- If H_S is false then there exist some index i such that $i \in S$ and H_i is false.

When all hypotheses can be true or false in any combination they are said to have *free combinations* otherwise they are *restricted* (Holm, 1979). As an example, assume there are 4 tightly linked loci and 10 hypotheses H_1, H_2, H_3, H_4, H_{12}, H_{23}, H_{34}, H_{123}, H_{234} and H_{1234} that are tested and obtain nominal p-values $p_{123} < p_1 < \ldots$. In the first step the corrected p-value of p_{123} is calculated from the permutation distribution

$$p_{123,c} = P(\min(P_1, P_2, P_3, P_4, P_{12}, P_{23}, P_{34}, P_{123}, P_{234}, P_{1234}) \leq p_{123}) \qquad (2)$$

just as in the case of free combinations, but then H_{1234} must be false as well. Furthermore at least one of H_1, H_2 and H_3 must be false, so in the next step

$$p'_1 = \max(P(\min(P_1, P_3, P_4, P_{34}) \leq p_1), P(\min(P_1, P_2, P_4, P_{12}) \leq p_1) \qquad (3)$$

rather than

$$p'_1 = P(\min(P_1, P_2, P_3, P_4, P_{12}, P_{23}, P_{34}, P_{234}, P_{1234}) \leq p_1) \qquad (4)$$

In Table 7 and 8 are the results of a simulated example of 100 cases and 100 controls genotyped at 3 loci. The three loci are in moderate linkage disequilibrium and 30 of the cases carry a dominant disease allele on the haplotype 1-1-1. For simplicity the haplotypes are assumed to be known. A simple Pearson chisquare test is performed on the allele counts for the three loci and on the haplotypes spanning loci (1,2),(2,3) and (1,2,3).

When these particular logical constraints and the closed test principle are used together, it will not necessarily result in a sequential procedure. The sequential procedure described above is still valid, but it can be improved by directly calculating the p-values for all the $\min p_i$ statistics that are prescribed by the closed test principle.

Table 7: Haplotype testing

allele	control	case	haplotype	control	case	haplotype	control	case
1xx	98	119	11x	54	84	111	40	65
2xx	102	81	12x	44	35	112	14	19
x1x	103	120	21x	49	36	121	6	9
x2x	97	80	22x	53	45	122	38	26
xx1	111	114	x11	78	91	211	38	26
xx2	89	86	x12	25	29	212	11	10
			x21	33	23	221	27	14
			x22	64	57	222	26	31

Table 8: Haplotype testing

locus	nominal	free	restricted
1	0.0154	0.0564	0.0435
2	0.0454	0.0971	0.0971
3	0.2259	0.2644	0.2235
1,2	0.0011	0.0039	0.0039
2,3	0.1996	0.2644	0.0971
1,2,3	0.0191	0.0564	0.0039

Testing several candidate regions

When several regions are tested with a fair number of linked loci in each region, it is primarily most important to identify which regions that are associated, rather than having the full picture of which loci in a particular region that are associated.

A single hypothesis H_i can then be formulated for each region i with n_i examined loci

$$H_i = \cap_{j=1}^{n_i} H_{ij} \text{ (no effect of region } i\text{)}, \tag{5}$$

where H_{ij} are the single loci null hypotheses. If H_i is tested with $T_i = \min_j p_{ij}$ we get a test capable of rejecting more regions than those that would be implied by the rejected loci.

This improvement is easy to understand if the null hypotheses H_{ij} are

reformulated as "no effect of any loci in region i", but still the single loci test statistics with p-values p_{ij} are used. It is then just an application of utilising the logical constraint that if H_{ij} is false than H_{ik} is false for all k.

In Table 9 are the results of a simulated example of 100 cases and 100 controls genotyped in two regions with 3 tightly linked loci in each region. Fifty of the cases carry a disease allele which is in complete LD with haplotype 1-1-1 in region 1, while the other 50 cases carry a disease allele in complete LD with haplotype 2-1-2 in region 2. The smallest corrected p-value (0.0365 for region 2 and loci 2,2) coincide with each other as they must do, but the corrected p-value for region 1 in the region strategy is 0.0383, while the best loci (1,2) for that region in the locus strategy has a corrected p-value of 0.0609.

Table 9: Testing two regions

region, locus	nominal p-value	corrected p-value 6 loci	corrected p-value 2 regions
1,1	0.0620	0.1100	0.0382
1,2	0.0139	0.0609	0.0382
1,3	0.0257	0.0689	0.0382
2,1	0.0597	0.1100	0.0365
2,2	0.0072	0.0365	0.0365
2,3	0.0195	0.0689	0.0365

Discussion

This paper has described multiple inference in genetic association studies with a case-control design. A permutation based method has been used in conjunction with the closed set principle as described in (Westfall and Young, 1993). This results in a multiple inference method which is fairly easy to implement as software; for some classes of primary test statistics it is even available in SAS.

The great advantage with a permutation strategy is that it handles the correlation between tests on loci in linkage disequilibrium as well as correlation between different tests on the data.

The null hypothesis for a single loci can be tested by different test statistics that are adapted to different model alternatives. There is naturally a loss of power compared to using the best test for the model, but that loss is an insurance that can be considered to avoid speculating about the model.

For tightly linked loci test statistics can be used on haplotypes as well as on a individual loci. The expected dependencies between these statistics is incorporated by the permutation strategy, so it is possible to perform test statistics on different sizes of haplotypes without paying the usual Bonferroni price for the multiplicity.

There are logical constraints when single loci hypotheses are combined with hypotheses for larger sequences of loci. These can be utilised as indicated in (Shaffer, 1986) to improve the corrected p-values.

The total FWE=α can be divided between the tests in a non uniform way, if some loci are considered more important or if we believe more in some of them before the tests are performed. It shall be noted that the beliefs can be based on an inspection of the data in the pooled sample with the case-control label hidden for the investigator.

A number of techniques have been used to improve the situation when many loci are screened and tested simultaneously. While we have used a small sample of real genetic data from a candidate gene for psoriasis, more research is needed on larger sets of real genetic data as well as on different designs such as genome wide screens or a screening with many candidate genes.

For some common simple statistics a very efficient implementation is available in SAS, but most of the material herein was not possible to implement in SAS. The current implementation was done in Splus, which can be used to run some 10.000 permutations on problems of this size, but we intend to implement a more efficient software package in the near future.

References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Statist*, **6**, 65–70.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting results. *Nature Genetics*, **11**, 241–247.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

O'Brien, K., Holm, S., Nilsson, S., Carlén, L., Rosenmüller, T., Enerbäck, C., Inerot, A., and Ståhle-Bäckdah, M. (2001). Despite disease-associated polymorphisms, the hcr gene on 6p21 is not a likely psoriasis susceptibility gene. *J Invest Dermatol*, **116**(5), 750–754.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**, 826–831.

Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing: examples and methods for p-value adjustment*. John Wiley and Sons.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Publishing.

Westfall, P., Zaykin, D., and Young, S. (2001). Multiple tests for genetic effects in association studies. In S. Looney, editor, *Statistical Methods in Molecular Biology*.

Xie, X. and Ott, J. (1993). Estimating haplotype-frequencies. *Am J Hum Genet*, **suppl 53**, 1107.

Zhao, Curtis, and Sham (2000). Model-free analysis and permutation tests for allelic associations. *Human Heredity*, **66**, 1661–1668.

Å. Torinsson Naluai
S. Nilsson
L. Samuelsson
A.H. Gudjónsdóttir
H. Ascher
J. Ek
B. Hallberg
B. Kristiansson
T. Martinsson
O. Nerman
L.M. Sollid
J. Wahlström

The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders

Key words:
CTLA4; celiac disease; genetic study; IDDM; linkage analysis; transmission disequilibrium test

Acknowledgments:
We would like to thank Britt-Marie Käck for great help in collecting families and blood samples, as well as The Celiac Society in Sweden. This work was supported by grants from the Foundation for Strategic Research, the Swedish Medical Research Council-Project no. K98-13X-12568-01A and K2000-27X-12568-03A, the Gothenburg Medical Society, the Research Council of Norway, Aktieselskabet Freia Chocolade Fabriks Medicinske Fond, The Swedish Society of Medicine, Frimurare-Barnhusdirektionen and the Wilhelm and Martina Lundgren Research Foundation.

Abstract: The effect of the gene region on chromosome 2q33 containing the CD28 and the cytotoxic T-lymphocyte associated (CTLA4) genes has been investigated in several diseases with chronic inflammatory nature. In addition to celiac disease (CD), type I diabetes, Grave's disease, rheumatoid arthritis and multiple sclerosis have all demonstrated associations to the A/G single nucleotide polymorphism (SNP) in exon 1, position +49 of the CTLA4 gene. The purpose of this study was to investigate this gene region in a genetically homogeneous population consisting of 107 Swedish and Norwegian families with CD using genetic association and linkage methods. We found a significant association with preferential transmission of the A-allele of the exon 1 +49 polymorphism by using the transmission disequilibrium test (TDT). Suggestive linkage of this region to CD was moreover demonstrated by non-parametric linkage (NPL) analysis giving a NPL-score of 2.1. These data strongly indicates that the CTLA4 region is a susceptibility region in CD. Interestingly, of the several chronic inflammatory diseases that exhibit associations to the CTLA4 +49 A/G dimorphism, CD appears to be the only disease associated to the A allele. This suggests that the +49 alleles of the CTLA4 gene are in linkage disequilibrium with two distinct disease predisposing alleles with separate effects. The peculiar association found in the gut disorder CD may possibly relate to the fact that the gastrointestinal immune system, in contrast to the rest of the immune system, aims to establish tolerance to foreign proteins.

Authors' affiliations:

Å. Torinsson Naluai[1],
S. Nilsson[3],
L. Samuelsson[1],
A.H. Gudjónsdóttir[2],
H. Ascher[2],
J. Ek[5],
B. Hallberg[1],
B. Kristiansson[2],
T. Martinsson[1],
O. Nerman[3],
L.M. Sollid[4],
J. Wahlström[1]

[1]Department of Clinical Genetics, The Queen Silvia Children's Hospital, Sahlgrenska University Hospital/Östra, Gothenburg, Sweden,

[2]Department of Pediatrics, The Queen Silvia Children's Hospital, Sahlgrenska University, Hospital/Östra, Gothenburg, Sweden,

[3]Chalmers University of Technology, Gothenburg, Sweden,

[4]Institute of Immunology, Rikshospitalet, University of Oslo, Oslo, Norway,

[5]Department of Pediatrics, Buskerud Central Hospital, Drammen, Norway

Correspondence to:
Åsa T. Naluai
Department of Clinical Genetics
Sahlgrenska University Hospital/Östra
416 85 Gothenburg
Sweden
Tel: +46 31 343 54 64
Fax: +46 31 84 21 60
e-mail: asa.torinsson@sahlgrenska.se

Received 21 April,
accepted for publication 19 June 2000

Copyright © Munksgaard 2000
Tissue Antigens . ISSN 0001-2815

Tissue Antigens 2000: **56**: 350–355
Printed in Denmark . All rights reserved

Celiac disease (CD) or gluten sensitive enteropathy is a chronic inflammatory disorder of the small intestine with multifactorial etiology (1). The prevalence of the disease was long believed to be in the range of 1 per 2,000 to 10,000. Recent population-based screening studies have, however, shown that the majority of cases are undiagnosed and that the prevalence rather ranges around 1 per 100 to 200 (2). In Sweden, childhood CD was increasingly diagnosed in the 1980s and early 1990s (3–5) and the disease is today the second most common chronic disease among Swedish children, surpassed only by asthma and allergies. HLA genes and non-HLA genes, together with gluten and possibly other environmental fac-

tors, are involved in disease development. More than 90% of CD patients carry the HLA-DQA1*0501 and -DQB1*02 alleles (coding for a variant of DQ2) compared to 20–30% of controls (6). Accumulating evidence suggests that the DQ2 molecule confers susceptibility to the disease by presenting gluten epitopes to CD4+ T cells in the small intestinal mucosa (1). Studies of celiac families indicate that non-HLA genes may contribute even more to the genetic susceptibility than the HLA genes (7, 8), but so far knowledge about non-HLA genes that predispose to CD is scarce.

Two genome-wide scans have been published on CD, one in families from western Ireland (9) and one performed in Italian families (10). The candidate regions retained this way have been followed up by independent studies from the UK (11, 12) and Finland (13). Two regions on chromosome 5q and chromosome 15q have been implicated in at least two of the studies (9–11).

Three independent association studies have been published on CD and the chromosome region 2q33 which contains the two genes CTLA4 and CD28 (13–15). This chromosome region has also been implicated in a variety of chronic inflammatory and autoimmune diseases including type 1 diabetes (where it is denoted IDDM12), autoimmune thyroid disease, rheumatoid arthritis and multiple sclerosis (16–34).

CTLA4 and CD28 are both involved in T-cell activation and proliferation. T cells require two signals in order to become fully activated. The first signal is antigen specific and occurs when the antigenic peptide bound to a major histocompatibility complex (MHC) molecule is recognized by the T-cell receptor (TCR). The second signal is a so called co-stimulatory signal where CD28 and CTLA4 interact with their ligands B7–1 (CD80) and B7–2 (CD86). CTLA4 plays an important role in maintaining tolerance to self-antigens, both as a negative regulator of this second co-stimulatory signal and as an inducer of clonal anergy. Severe autoimmune disease develop in mice in which CTLA4 is deleted by targeted gene disruption (35, 36). In CD, it has been shown using an organ culture model that CTLA4Ig controls many immunological features including apoptosis of gliadin-specific T cells (37). These findings make CTLA4 an interesting candidate for CD and other chronic inflammatory disorders.

In this study, a large and homogeneous family material of Swedish and Norwegian families is investigated with the aim to study the involvement of the 2q33 chromosome region in CD.

A) Norwegian family structures

Number of families	Affected parents	Affected children	Non-affected children
12	0	2	0

B) Swedish family structures

Number of families	Affected parent	Affected children	Non-affected children
24	0	2	0
27	0	2	1
9	0	2	2
2	0	2	3
0	0	2	4
2	0	2	5
1	0	2	6
1	0	3	0
2	0	3	1
1	0	3	2
0	0	3	3
1	0	3	4
1	0	3	5
2	1	2	0
9	1	2	1
3	1	2	2
2	1	2	3
1	1	3	0
2	1	3	1
2	1	3	2
1	3	7	3

Table 1

Material and methods

Subjects

The study included 104 nuclear families with a minimum of two affected siblings and one extended pedigree with three generations, resulting in a total of 107 nuclear families. The families were contacted via advertisement in the journal of the Swedish Celiac Society, inquiries to all Swedish pediatric departments and via our own patient registries. Families were included in the study only if all affected members strictly fulfilled the ESPGHAN diagnostic criteria (38). Family structures are presented in Table 1.

Genotyping

A total of six markers were analyzed, one SNP within the CTLA4 gene (an A/G substitution, position +49 in exon 1) and five microsatellites in the region outside the gene (Fig. 1).

PCR reactions for the +49 (A/G) SNP in exon 1 was performed

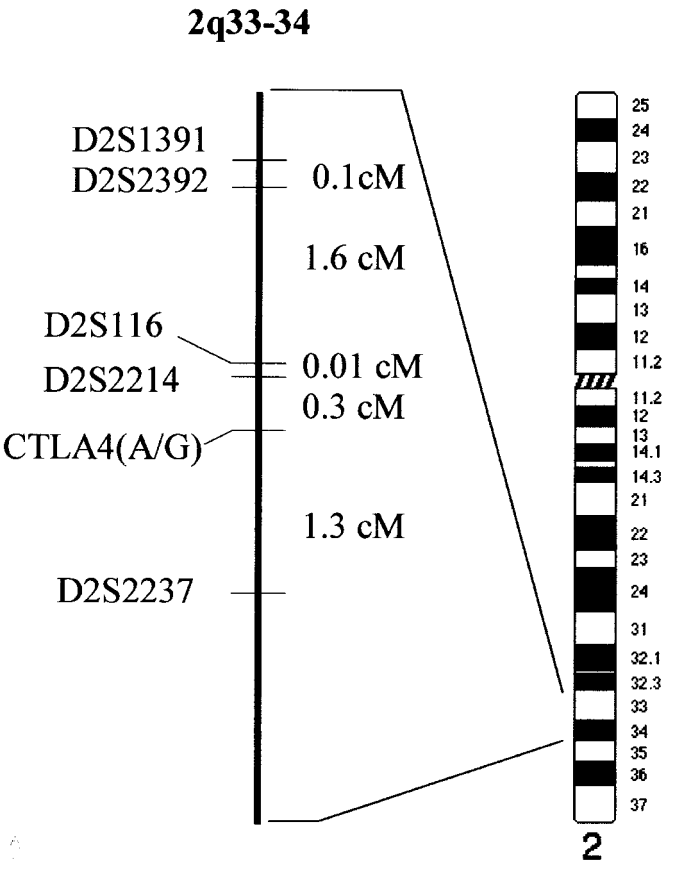

Fig. 1. **The chromosome region 2q33–34 including the CTLA4 A/G +49 exon 1 SNP and microsatellite markers used in the study.** The order of markers were taken from the work of Hadano et al. (44) and is slightly different from the order presented by the CEPH/Genethón genetic map (45).

using specific primers previously described by Marron et al. (16). The polymerase chain reaction (PCR) reactions were performed in a total volume of 20 µl with cycling conditions of 94°C for 45 s, 62°C for 45 s, 72°C for 60 s, for 30 cycles, followed by 72°C for 7 min.

This resulted in a 152-bp fragment which was cleaved overnight at 60°C using 5 units of BstEII per reaction. The digested A allele yields a fragment of 130 bp and the G allele yields an intact 152-bp fragment. Digested products were separated by electrophoresis on a 3.5% agarose gel.

The microsatellite markers D2S1391, D2S2392, D2S116, D2S2214 and D2S2237 were amplified using an ABI 877 PCR robot under standard conditions, annealing temp 54°C, 30 cycles. Products were analyzed with an ABI 377 sequencer using the GENESCAN ANALYSIS and GENEOTYPER software (PE Biosystems, Foster City, CA, USA; http://www.pecorporation.com).

Statistical analysis

For statistical analysis we used the transmission disequilibrium test (TDT) (40) as well as non parametric linkage analysis (NPL). NPL was calculated using the software GENEHUNTER version 2.0 (http://www.genome.wi.mit.edu/ftp/distribution/software/genehunter). For the TDT we used a binomial test for the exon 1 A/G SNP and the Extended TDT (ETDT) statistic program (allele-wise mode) for multiple allele markers (41).

We included all siblings in our TDT study, as we were interested in showing linkage in this region as well as confirming association. Haplotypes over the region were constructed using GENEHUNTER version 2.0 (http://www.genome.wi.mit.edu/ftp/distribution/software/genehunter).

Results

The CTLA4 gene contains three known polymorphisms: a (C/T) single nucleotide polymorphism (SNP) at position −318 in the promoter region (39); a non-synonymous (A/G) SNP at position +49 in exon 1, which corresponds to a Thr/Ala amino acid change in the leader peptide of the expressed protein and an (AT)n microsatellite in exon 3 (Fig. 2). In this study the subjects have been typed for the SNP at position +49. In addition they were typed for 5 microsatellite markers in the region around the CTLA-4 and CD28 genes (Fig. 1).

Analysis of the CTLA4 position +49 (A/G) SNP in Swedish and Norwegian CD patients

Results from the TDT test showed 114 transmitted vs. 79 non-transmitted A-alleles at the +49 position of the CTLA4 gene. This gives a *P*-value of 0.007 for a one-sided test of preferential transmission of the A-allele, i.e. an association to the A-allele. Since the test was an attempt to confirm association presented in a previous study, we consider a *P*-value of 0.007, as obtained in this study, significant.

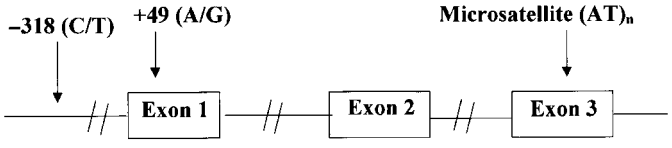

Fig. 2. **Polymorphisms in the CTLA4 gene.**

Results from the Extended transmission disequilibrium test (ETDT)

Marker:	*P*-value, ETDT	NPL score
D2S1391	NS	2.1
D2S2392	0.037	2.0
D2S116	NS	1.7
D2S2214	0.044	1.8
CTLA4(A/G) +49 SNP	0.012, 0.007*	1.2
D2S2237	NS	0.8

* One-sided binomial test
NS=non-significant

Table 2

Analysis of five microsatellite markers in the CTLA4/CD28 chromosome region in Swedish and Norwegian CD patients

Five microsatellite markers in the region around the CTLA4 and CD28 genes were typed (Fig. 1).

Two of the microsatellite markers, D2S2392 and D2S2214, showed signs of association with *P*-values of <0.05 as calculated with the ETDT program (Table 2). The non-parametric linkage analysis had a maximum at marker D2S1391, with an NPL score of 2.1, which corresponds to a *P*-value of 0.018. NPL-scores for all the markers are presented in Table 2.

No particular haplotype was significantly over represented, neither in affected nor in healthy individuals.

Discussion

T cells are considered to be important for controlling the development of HLA-associated diseases that have a chronic inflammatory nature. CTLA4 and CD28 and their interacting ligands are involved in T-cell regulation and are therefore potential candidates conferring susceptibility to such diseases. The association studies of CD published to date on the CTLA4/CD28 gene region have reported deviant results for different populations. In a study on French Caucasians (14), association was found to the A-allele of the CTLA4 exon 1 +49 polymorphism, while a Finnish study found that the marker D2S116, located immediately adjacent to the gene, was associated to the disease (13). A study on Italian and Tunisian patients failed to confirm association to the exon 1 +49 polymorphism of CTLA4 (15). However, a polymorphism which is not associated with the disease in a certain population can still be linked to the disease in the very same population. This is the result of the complex interaction between marker allele frequencies, number of founder mutations, disease heterogeneity and recombinations over time. Even if the very marker analyzed also happens to be the etiological disease causing polymorphism or mutation, these disturbing factors could still be present. These factors could all play a part in explaining why an association to a certain marker is not replicated in different populations. Other reasons for discrepant results could be that the sample size is too small to detect linkage disequilibrium, even if the factors mentioned above are minimal, or the sample size is too small to detect a disease allele that has a minor effect in the disease population.

As mentioned earlier, the CTLA4/CD28 region has been implicated in a number of other autoimmune diseases. Moreover, many autoimmune disorders including type 1 diabetes and autoimmune thyroid diseases are over-represented in CD and vice-versa (42, 43). *A priori*, one would expect that within the same ethnical population, CD and the other CTLA4-associated diseases should demonstrate associations to the same CTLA4 allele. This is, however, not the case. For instance, studies of French Caucasians clearly have revealed that type 1 diabetes is associated with the G allele and CD with the A allele of the CTLA4 +49 exon 1 dimorphism (14, 16). In fact, CD appears to be the sole disease where association is with the A allele. The gastrointestinal immune system has an important role in promoting tolerance to many of the new antigens encountered after birth, while the rest of the immune system mainly focuses on attacking newly encountered antigens. It is an intriguing thought that this locus could represent part of the specific functional genetic differences between CD as a gastrointestinal immune-mediated disease and other immune-mediated disorders.

We conclude that the chromosome region surrounding CTLA4 is likely to be a susceptibility region in CD. Further studies will be required to determine whether the CTLA4 polymorphisms are functionally important in these diseases or if not then to identify the true disease-promoting gene(s), or regulatory element(s) in this region, and the variants conferring susceptibility to CD as well as other chronic inflammatory and autoimmune diseases.

References

1. Sollid LM. Molecular basis of celiac disease. *Ann Rev Immunol* 2000: **18**: 53–81.
2. Ascher H. The role of quantity and quality of gluten-containing cereals in the epidemiology of coeliac disease. In: Mäki M, Collin P, Visakorpi JK, eds. *Coeliac Disease. Proceedings of the Seventh International Symposium on Coeliac Disease*. Tampere: Coeliac Disease Study Group, 1997: 15–22.
3. Ascher H, Krantz I, Kristiansson B. Increasing incidence of coeliac disease in Sweden. *Arch Dis Child* 1991: **66**: 608–11.
4. Cavell B, Stenhammar L, Ascher H et al. Increasing incidence of childhood coeliac disease in Sweden. Results of a national study. *Acta Paediatr* 1992: **81**: 589–92.
5. Ivarsson A, Persson LÅ, Nyström L et al. Epidemic of coeliac disease in Swedish children. *Acta Paediatr* 2000: **89**: 165–71.
6. Sollid LM, Thorsby E. HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis [published erratum appears in *Gastroenterology* 1994: **106**: 1133]. *Gastroenterology* 1993: **105**: 910–22.
7. Risch N. Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet* 1987: **40**: 1–14.
8. Petronzelli F, Bonamico M, Ferrante P et al. Genetic contribution of the HLA region to the familial clustering of coeliac disease. *Ann Hum Genet* 1997: **61**: 307–17.
9. Zhong F, McCombs CC, Olson JM et al. An autosomal screen for genes that predispose to celiac disease in the western counties of Ireland. *Nat Genet* 1996: **14**: 329–33.
10. Greco L, Corazza G, Babron MC et al. Genome search in celiac disease. *Am J Hum Genet* 1998: **62**: 669–75.
11. Houlston RS, Tomlinson IP, Ford D et al. Linkage analysis of candidate regions for coeliac disease genes. *Hum Mol Genet* 1997: **6**: 1335–9.
12. Brett PM, Yiannakou JY, Morris MA et al. A pedigree-based linkage study of coeliac disease: failure to replicate previous positive findings. *Ann Hum Genet* 1998: **62**: 25–32.
13. Holopainen P, Arvas M, Sistonen P et al. CD28/CTLA4 gene region on chromosome 2q33 confers genetic susceptibility to celiac disease. A linkage and family-based association study. *Tissue Antigens* 1999: **53**: 470–5.
14. Djilali-Saiah I, Schmitz J, Harfouch-Hammoud E, Mougenot JF, Bach JF, Caillat-Zucman S. CTLA-4 gene polymorphism is associated with predisposition to coeliac disease. *Gut* 1998: **43**: 187–9.
15. Clot F, Fulchignoni-Lataud MC, Renoux C et al. Linkage and association study of the CTLA-4 region in coeliac disease for Italian and Tunisian populations. *Tissue Antigens* 1999: **54**: 527–30.
16. Marron MP, Raffel LJ, Garchon HJ et al. Insulin-dependent diabetes mellitus (IDDM) is associated with CTLA4 polymorphisms in multiple ethnic groups. *Hum Mol Genet* 1997: **6**: 1275–82.
17. Nistico L, Buzzetti R, Pritchard LE et al. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Hum Mol Genet* 1996: **5**: 1075–80.
18. Owerbach D, Naya FJ, Tsai MJ, Allander SV, Powell DR, Gabbay KH. Analysis of candidate genes for susceptibility to type I diabetes: a case-control and family-association study of genes on chromosome 2q31-35. *Diabetes* 1997: **46**: 1069–74.
19. Pozzilli P, Buzzetti R, Nistico L, Romiti A, Giovannini C. Non-HLA genes and the susceptibility to insulin dependent diabetes: the role of the CTLA-4 gene. *Acta Diabetol* 1996: **33**: 250–2.
20. Awata T, Kurihara S, Iitaka M et al. Association of CTLA-4 gene A-G polymorphism (IDDM12 locus) with acute-onset and insulin-depleted IDDM as well as autoimmune thyroid disease (Graves' disease and Hashimoto's thyroiditis) in the Japanese population. *Diabetes* 1998: **47**: 128–9.
21. Donner H, Seidl C, Braun J et al. CTLA4 gene haplotypes cannot protect from IDDM in the presence of high- risk HLA DQ8 or DQ2 alleles in German families. *Diabetes* 1998: **47**: 1158–60.
22. Djilali-Saiah I, Larger E, Harfouch-Hammoud E et al. No major role for the CTLA-4 gene in the association of autoimmune thyroid disease with IDDM. *Diabetes* 1998: **47**: 125–7.
23. Heward JM, Allahabadia A, Carr-Smith J et al. No evidence for allelic association of a human CTLA-4 promoter polymorphism with autoimmune thyroid disease in either population-based case-control or family-based studies. *Clin Endocrinol (Oxf)* 1998: **49**: 331–4.
24. Sale MM, Akamizu T, Howard TD et al. Association of autoimmune thyroid disease with a microsatellite marker for the thyrotropin receptor gene and CTLA-4 in a Japanese population. *Proc Assoc Am Physicians* 1997: **109**: 453–61.
25. Braun J, Donner H, Siegmund T, Walfish PG, Usadel KH, Badenhoop K. CTLA-4 promoter variants in patients with Graves' disease and Hashimoto's thyroiditis. *Tissue Antigens* 1998: **51**: 563–6.
26. Donner H, Rau H, Walfish PG et al. CTLA4 alanine-17 confers genetic susceptibility to Graves' disease and to type 1 diabetes mellitus. *J Clin Endocrinol Metab* 1997: **82**: 143–6.
27. Heward JM, Allahabadia A, Armitage M et al. The development of Graves' disease and the CTLA-4 gene on chromosome 2q33. *J Clin Endocrinol Metab* 1999: **84**: 2398–401.
28. Yanagawa T, Hidaka Y, Guimaraes V, Soliman M, DeGroot LJ. CTLA-4 gene polymorphism associated with Graves' disease in a Caucasian population. *J Clin Endocrinol Metab* 1995: **80**: 41–5.
29. Yanagawa T, Taniyama M, Enomoto S et al. CTLA4 gene polymorphism confers susceptibility to Graves' disease in Japanese. *Thyroid* 1997: **7**: 843–6.
30. Kotsa K, Watson PF, Weetman AP. A CTLA-4 gene polymorphism is associated with both Graves disease and autoimmune hypothyroidism. *Clin Endocrinol (Oxf)* 1997: **46**: 551–4.
31. Gonzalez-Escribano MF, Rodriguez R, Valenzuela A, Garcia A, Garcia-Lozano JR, Nunez-Roldan A. CTLA4 polymorphisms in Spanish patients with rheumatoid arthritis. *Tissue Antigens* 1999: **53**: 296–300.
32. Seidl C, Donner H, Fischer B et al. CTLA4 codon 17 dimorphism in patients with rheumatoid arthritis. *Tissue Antigens* 1998: **51**: 62–6.
33. Harbo HF, Celius EG, Vartdal F, Spurkland A. CTLA4 promoter and exon 1 dimorphisms in multiple sclerosis. *Tissue Antigens* 1999: **53**: 106–10.

34. Ligers A, Xu C, Saarinen S, Hillert J, Olerup O. The CTLA-4 gene is associated with multiple sclerosis. *J Neuroimmunol* 1999: **97**: 182–90.
35. Waterhouse P, Penninger JM, Timms E et al. Lymphoproliferative disorders with early lethality in mice deficient in Ctla-4. *Science* 1995: **270**: 985–8.
36. Tivol EA, Borriello F, Schweitzer AN, Lynch WP, Bluestone JA, Sharpe AH. Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity* 1995: **3**: 541–7.
37. Maiuri L, Auricchio S, Coletta S et al. Blockage of T-cell costimulation inhibits T-cell action in celiac disease. *Gastroenterology* 1998: **115**: 564–72.
38. Revised criteria for diagnosis of coeliac disease. Report of working group of European Society of Paediatric Gastroenterology and Nutrition. *Arch Dis Child* 1990: **65**: 909–11.
39. Deichmann K, Heinzmann A, Bruggenolte E, Forster J, Kuehr J. An Mse I RFLP in the human CTLA4 promotor. *Biochem Biophys Res Commun* 1996: **225**: 817–8.
40. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993: **52**: 506–16.
41. Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995: **59**: 323–36.
42. Ventura A, Magazzu G, Greco L. Duration of exposure to gluten and risk for autoimmune disorders in patients with celiac disease. SIGEP Study Group for Autoimmune Disorders in Celiac Disease. *Gastroenterology* 1999: **117**: 297–303.
43. Kaukinen K, Collin P, Mykkanen AH, Partanen J, Maki M, Salmi J. Celiac disease and autoimmune endocrinologic disorders. *Dig Dis Sci* 1999: **44**: 1428–33.
44. Hadano S, Nichol K, Brinkman RR et al. A yeast artificial chromosome-based physical map of the juvenile amyotrophic lateral sclerosis (ALS2) critical region on human chromosome 2q33-q34. *Genomics* 1999: **55**: 106–12.
45. Dib C, Faure S, Fizames C et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 1996: **380**: 152–4.

J Neurol (2000) 247 : 616–622
© Steinkopff Verlag 2000

ORIGINAL COMMUNICATION

Sara Haghighi
Oluf Andersen
Lars Rosengren
Tomas Bergström
Jan Wahlström
Staffan Nilsson

Incidence of CSF abnormalities in siblings of multiple sclerosis patients and unrelated controls

Received: 29 June 1999
Received in revised form: 2 February 2000
Accepted: 15 March 2000

S. Haghighi (✉) · O. Andersen · L. Rosengren
Institution of Clinical Neuroscience, Sahlgrenska University Hospital, 41345 Gothenburg, Sweden
Tel.: 46–31–342 15 69

T. Bergström
Institution of Virology, Sahlgrenska University Hospital, Gothenburg, Sweden

Jan Wahlström
Institution of Clinical Genetics, Sahlgrenska University Hospital, Gothenburg, Sweden

S. Nilsson
Institution of Mathematical Statistics, Chalmers University of Technology Gothenburg, Sweden

Abstract We found that 19 % (9/47) of healthy siblings of patients with clinically definite multiple sclerosis had an intrathecal immunological reaction with two or more 2 CSF-enriched oligoclonal bands (OCBs), in contrast to (4 %) (2/50) unrelated healthy controls. Furthermore, in this group of nine healthy sibs the measles CSF IgG antibody titers were higher than that of the other sibs and that of controls. There were also differences in the serum titers for measles IgG antibody, which were higher in the group of all healthy sibs than in healthy volunteers, and (as with CSF titers) higher in the subgroup of healthy sibs with two or more 2 CSF-enriched OCBs than the other sibs. Thus a significant proportion of healthy siblings to MS patients have a partially hyperimmune condition similar to that occurring in MS, which in 19 % manifested itself as an OCB reaction, in 9 % as increased CSF measles IgG antibody titers, and in 21 % as increased serum measles IgG antibody titers, these phenomena tending to occur in the same individuals. This condition is characterized by CSF-enriched OCBs with undefined specificity, although some increased antiviral reactivity is found both in the serum and CSF. While it needs further characterization, a genetic trait interacting with common infections is suggested. The recurrence risk of this condition is approximately five times higher than the 3–4 % recurrence risk for manifest MS reported for sibs.

Key words Multiple sclerosis · Siblings · Genetics · Oligoclonal bands · Measles

Introduction

Elevated measles antibodies in serum and CSF in multiple sclerosis (MS) patients were first demonstrated by Adams and Imagawa [1]. Subsequent studies have also shown elevated titers of measles serum and CSF antibodies in healthy siblings of MS patients, but data from healthy unrelated controls were provided only for serum [6, 13, 24, 34]. Furthermore, an intrathecal immunological reaction with oligoclonal bands (OCBs) has been found in healthy dizygotic twins to MS patients [35]. Poser has proposed the term "MS trait" for this partially hyperimmune phenotype [25, 26]. To determine the frequency of this condition we examined the CSF-enriched OCBs with IgG immunoblot and the CSF and serum viral antibody titers in 47 pairs of MS patients and their healthy sibs, and in 50 unrelated healthy volunteers.

Materials and methods

Ascertainment of patients, sibs, and volunteers

The dimension of the study was planned to be 50 patients with clinically definite MS (CDMS), their 50 healthy siblings, and 50 healthy unrelated volunteers, since this would allow detection of the one-fourth of phenotype-expressing individuals expected in recessive heredity. The study was approved by the Medical Ethics Committee

at the University of Gothenburg, and informed consent was obtained after written information. The MS patients were recruited from the Gothenburg MS register [28] and from local MS societies, yielding a sample of approximately 400 accessible MS patients. A letter was sent explaining the study, and from 200 answers obtained we initially included 50 consecutive eligible pairs of Scandinavian descent. When a CDMS patient had more than one sibling consenting to participate, we included the sibling whose age was closest to the patient's age. Three pairs were excluded, one of half-siblings, one of monozygotic twins, and one with symptomatic neuroborreliosis revealed during the present study. The results were therefore based upon 47 patients with clinically definite MS and their 47 healthy siblings.

A request for healthy volunteers free from current symptomatic disease was made at the Sahlgrenska University Hospital Blood Bank. Hospital staff was not invited to participate. We obtained 76 answers, and from these we included fifty consecutive volunteers in the study. Except for one volunteer who was of southern European origin, all were of Scandinavian descent.

A complete history was taken and a neurological examination was performed and scored with the Regional Functional Scoring System (RFSS), utilizing Kurtzke's Functional Systems in both patients and sibs [19]. Healthy volunteers were examined with ophthalmoscopy and blood pressure measurement. CSF and blood samples were obtained from all groups as an outpatient procedure for analysis of virus IgG antibody titers, total IgG and albumin assays, isoelectric focusing (IEF) with immunoblot, and CSF cell counts. In the subgroup with at least two OCBs in the healthy sibling we also analyzed measles IgM antibodies in the blood.

Basic and demographic parameters

Sex and age were equally distributed between the 47 patients and sibs, while the 50 healthy volunteers were younger (the possible effect of age was examined in subsequent analyses). Sex was unequally distributed between the sib pairs and the unrelated volunteers; the proportion of men was higher among the healthy controls (Table 1). Of the 47 patients with CDMS 3 were primary progressive while 22 were in the relapsing-remitting and 22 in the secondary progressive phase. Of the 47 sibs three suffered from migraine, three had had a hysterectomy, two were on antidepressive treatment, two were on antihypertensive treatment, and six had each one of the following: previous transient cerebrovascular disease, hip operation, ileus, orally treated diabetes, and well-controlled hypothyroidism. There was no age difference between the sibling group with fewer than two CSF-enriched OCBs (mean 45.4 years) and those with two or more (mean 45.7 years). In the group of healthy volunteers two persons were on antidepressive treatment, one had been operated on for lumbar disc hernia 1988, and one had a previous syncopal episode.

Table 1 Sex distribution and median age (range) in the analyzable material, 47 sibling pairs, and 50 unrelated healthy volunteers

	CDMS patients	Healthy siblings	Unrelated healthy volunteers
Women	31	29	15
Men	16	18	35
Total	47	47	50
Median age (range; years)	45 (21–64)	45 (18–66)	33 (18–57)

CSF

Quantitative determinations of albumin and IgG in serum and CSF were performed using the Behring nephelometer analyzer (Behringwerke). The CSF/serum albumin ratio [30] was calculated as: CSF albumin (mg/l)/serum albumin ×1000, and was used as a measure of blood-brain barrier (BBB) function. IEF of samples was performed in polyacrylamide gel, and the gels were silver-stained as previously described [33]. All samples were also transblotted onto Immobilon-P membranes (Mollipore) after IEF and immunoreacted using an antiserum against the gamma chain of human IgG (goat antihuman IgG; BioRad). One of the authors (L.R.), who was blinded to the clinical data, was responsible for the analytical procedures and the interpretation of the IEFs and immunoblots. The presence of two or more CSF-enriched OCBs on the IEF that positively were identified as IgG bands by the immunoblots was considered a sign of local IgG synthesis, with the precondition that there were no corresponding OCBs in the blood. Using the cutoff level of two or more 2 positively identified CSF-enriched oligoclonal IgG bands, the frequency of positive and negative IEF tests were determined in the MS and control study populations.

Viral serology

Serum and CSF samples from patients, sibs, and controls were analyzed simultaneously by enzyme-linked immunosorbent assay (ELISA) for IgG measles antibodies, together with positive and negative control sera. Antigen was prepared as described previously [12]. Briefly, cell cultures of human fibroblasts infected with strain Edmonston E139 were allowed to develop until a complete cytopathogenic effect was observed. Cells were scraped down, freeze-thawed, ultrasonicated, and centrifuged, and the supernatant was saved as antigen, which was passively coated on to microtiter plates (Nunc Immuno Maxisorp, Nunc, Gothenburg). Patient samples were incubated in twofold dilutions starting from 1/10 for CSF or 1/100 for serum. For determination of IgG antibodies, peroxidase-conjugated affinity-purified goat anti-human IgG antibodies (Jackson Immunoresearch, Gothenburg) were used, and the color reaction after addition of *o*-phenylenediamine and H_2O_2 was determined by a spectrophotometer (Molecular Devices VMax kinetic microplate reader). One of the authors (T.B.), who was blinded to the clinical data, was responsible for the analytical procedures and the interpretation of the ELISA. Titers were given as the highest dilution giving positive reaction as compared to the negative controls. The individual cutoff value for pathological titers of the CSF and serum measles antibody, corresponding to the 95% percentile of the titers of the unrelated healthy controls, was a CSF titer higher than 20 and a serum titer higher than 6400.

IgM antibodies to measles were determined by immunofluorescence, using measles-infected GMK-AH1 cells as antigen. Serum samples were incubated in twofold dilutions, and fluorescein-labeled goat anti-human IgM liquid globulin (bioMeriux) was used as a conjugate.

Herpes simplex virus (HSV) serology was performed by the use of a HSV-type common antigen as described previously [4]. In brief, a DOC-solubilized membrane antigen of HSV–1-infected cells were used in ELISA, and the procedure was carried out as described above for measles serology as regards sample dilution, conjugate, substrate, and absorbance determinations. HSV–2 specific antibodies were determined using purified glycoprotein G2 as antigen [4]. IgG antibodies to varicella zoster virus were determined by ELISA, and to Epstein-Barr virus by immunofluorescence, using in-house routine tests. Antibodies to human immunodeficiency virus (HIV) and rubella were determined by commercially available IMX system assays (Abbot Diagnostics).

Statistical methods

Differences in proportions of persons with two or more OCBs between patients, siblings, and healthy volunteers were tested by Fisher's exact test ("CSF-enriched OCBs without unequivocal specificity in the CSF"). Measles and HSV–1 CSF and serum titers and antibody index were compared using the Mann-Whitney test and the Wilcoxon signed rank test, with stratification for sex ("Measles antibody titers in the CSF and serum," and "Relationship between the occurrence of at least two CSF-enriched oligoclonal IgG bands and measles antibody titers and indices"). The effect of sex was examined by the Mann-Whitney test. The effect of age was examined by regression analysis for the measles antibody variables, and by a *t* test for the presence of at least two 2 CSF-enriched OCBs. Comparison of RFSS scores in the patients whose sibs had at least two 2 CSF-enriched OCBs with patients whose sibs had not was performed with the Mann-Whitney test ("Quantitative neurological score").

Results

Borrelia, syphilis, and HIV antibodies in the serum and CSF

Serum and CSF specimens were negative for *Borrelia* antibodies, apart from four individuals. One sib had a history of neuroborreliosis with headache and slight cognitive symptoms which subsided after treatment, remaining low *Borrelia* antibody titers in serum and negative *Borrelia* titers in the CSF, and at least two CSF-enriched OCBs. This sib pair was excluded from the study. One MS patient (who had at least two CSF-enriched OCBs) and one healthy volunteer (with fewer than two OCBs) had slightly elevated *Borrelia* IgG titers in serum, and one sib (who had at least two OCBs) had slightly elevated IgG and IgM titers in serum. The anamneses of these three individuals revealed no previous clinical episodes suggestive of neuroborreliosis. All CSF samples from patients, sibs, and healthy volunteers were negative for *Borrelia* antibodies. Syphilis and HIV serology was negative in the CSF and serum in the three groups.

CSF-enriched OCBs without unequivocal specificity in the CSF

The presence of at least two CSF-enriched OCBs was observed in 45 of 47 (96%) of CDMS patients, 9 of 47 (19%, 95% confidence intervals 8–30%) of their sibs, and 2 of 50 (4%) healthy unrelated volunteers. There was a significantly higher proportion of sibs with at least two CSF-enriched OCBs than of healthy volunteers (P=0.0196). Generally the healthy sibs with at least two CSF-enriched OCBs had fewer bands than the patients.

Measles antibody titers in the CSF and serum

In the subsequent sections, individuals with BBB lesion as defined by a CSF/serum albumin ratio above 9.8 were excluded from all results presented as measles CSF titers. In comparisons between patients and sibs the pair was excluded if either patient or sib had a BBB lesion [5].

Serum measles IgG antibody titers tended to be higher in the patients than in their sibs (P=0.096). There were significantly higher serum measles IgG antibody titers in sibs than in controls (P=0.017). However, the significance was entirely due to the difference among the men. The CSF measles titers were significantly higher in the patients than in their sibs (n=40 pairs, after exclusion of individuals with BBB lesion, P< 0.001), and tended to be higher in the sibs (n=42) than in the healthy volunteers (n=47, P=0.094). Serum IgM measles antibody titers, when analyzed in nine pairs in whom the healthy sib had at least two CSF-enriched OCBs, were all negative. Previous studies [6, 13] reported higher measles titers in women. This was confirmed in the present study for healthy unrelated volunteers only (P=0.05), while there was no significant sex difference of measles titers in the sibs or the patients. There was no significant effect of age on CSF or serum measles antibody titers.

A viral antibody index [11] was also used, but a substantial number of individuals with a CSF titer below 10 (not detectable) then had to be excluded because these individuals probably have very different true antibody ratios. The index provided the same information as the CSF antibody titers in this study.

Relationship between the occurrence of at least two CSF-enriched oligoclonal IgG bands and measles antibody titers

The relationship between the occurrence of at least two CSF-enriched OCBs and the measles antibody titers in the serum and CSF measles titer in the healthy siblings is displayed in Fig. 1. Measles titers were higher in the sibs with at least two CSF-enriched OCBs than in the sib group with fewer than two OCBs. The difference was significant both in serum (P=0.040, Fig. 2) and in the CSF (after exclusion of individuals with BBB lesion, n=9 and 33 in the two groups, P=0.0067; Fig. 2). The measles antibody titers were also higher in the siblings with at least two CSF-enriched OCBs than in the healthy unrelated volunteers (for serum samples, n=9 sibs and 50 volunteers; for CSF samples n=9 sibs and 47 volunteers, P=0.0039 for serum and P=0.0092 for CSF). Four healthy siblings (9%) had pathological measles IgG antibodies titers in the CSF in the absence of a BBB lesion, and two of these had fewer than two CSF-enriched OCBs (Fig. 1). If these two were nevertheless considered to have an intrathecal reaction similar to MS, the proportion of healthy siblings with this condition would be (9+2)/47, or 23%.

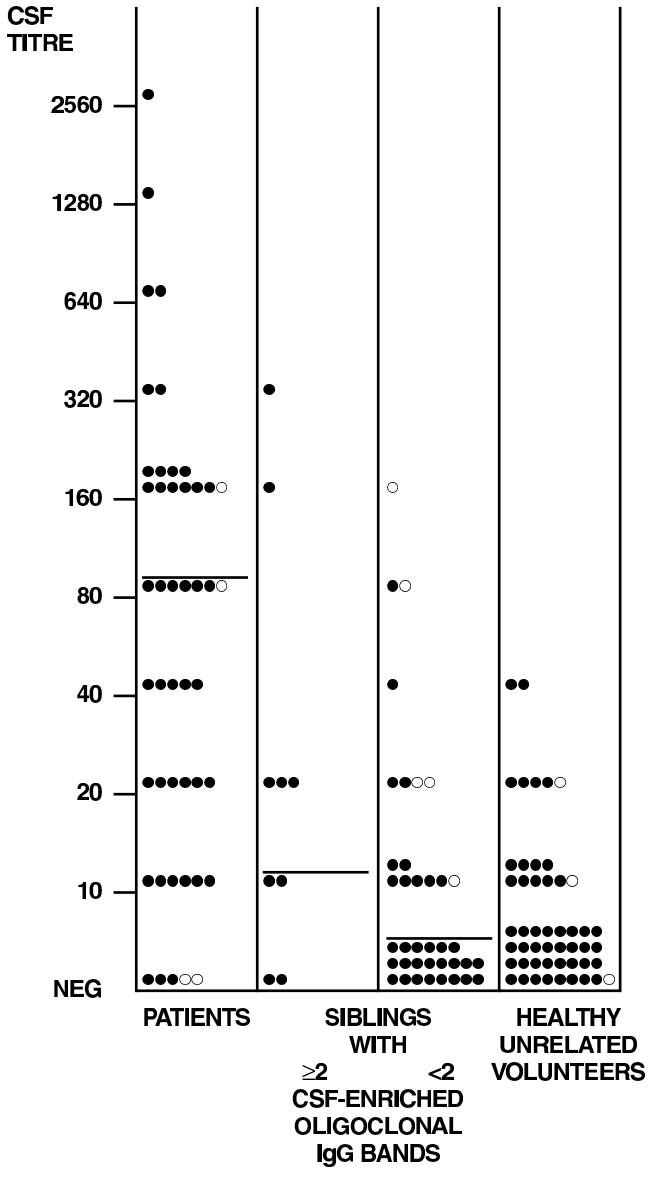

Fig. 1 Measles virus ELISA titers in CSF, in 47 CDMS patients, 9 siblings with at least two CSF-enriched OCBs, 38 siblings with fewer than two CSF-enriched OCBs, and 50 unrelated healthy volunteers. *Dots* Individuals; *circles* individuals with a BBB damage defined as an albumin ratio > 9.8; *horizontal bars* median values. The figure suggests a tendency to increasing titers through sibs with fewer than two CSF-enriched OCBs, sibs with at least two CSF-enriched OCBs, and particularly in CDMS patients

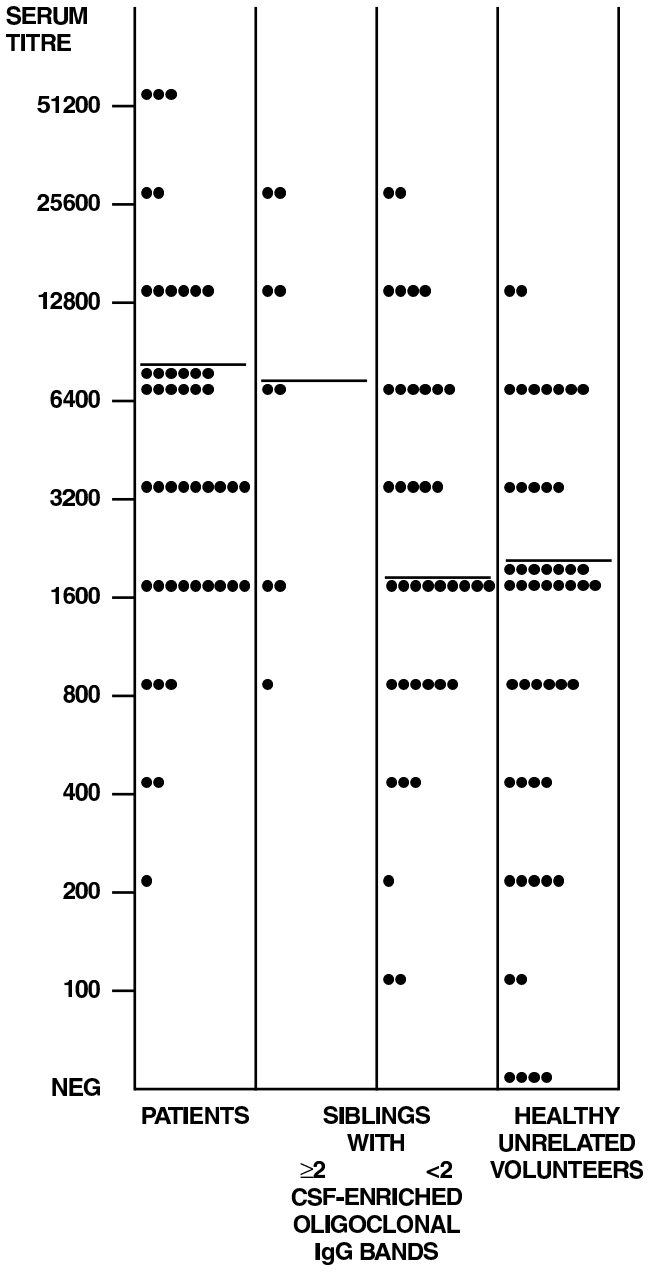

Fig. 2 Measles virus ELISA titers in serum, in 47 CDMS patients, 9 siblings with at least two CSF-enriched OCBs, 38 siblings with fewer than two CSF-enriched OCBs, and 50 unrelated healthy volunteers. *Dots* Individuals; *horizontal bars* median values. The figure suggests a tendency to increasing titers in sibs with at least two CSF-enriched OCBs and in CDMS patients

HSV–1 IgG antibody titer

The median CSF HSV–1 antibody titer in CDMS patients and healthy sibs after exclusion of cases with BBB lesion was 10, while it was negative for the healthy unrelated controls. The proportion of negatives was 19 of 43 (44%) CDMS, 18 of 42 (43%) healthy sibs, and 26 of 45 (58%) controls; the range was negative to 160 in patients, and negative to 80 in both healthy siblings and controls. There were no significant group differences between the HSV–1 titers of healthy sibs and healthy unrelated volunteers in serum (P=0.52) or CSF (P=0.33). There was also no difference between healthy sibs with at least two CSF-enriched OCBs in this respect and the sibs with fewer than two CSF-enriched OCBs (P=0.52 for serum, P=0.38 for CSF) or from the healthy unrelated controls (P=0.49 for serum, P=0.32 for CSF).

Other viral antibody titers

Screening was also performed for HSV–2 CSF antibody in the group of nine healthy siblings with at least two CSF-enriched OCBs. These titers were all negative. The varicella zoster virus IgG titers were negative in six, and low in two individuals. The Epstein-Barr virus IgG titers were negative in three, and low in five individuals (ND=1). The rubella IgG titer was negative in 3 (ND=6).

Measles history and vaccination status

Measles was previously endemic in Sweden. A history of measles was confirmed in 40 of 43 CDMS patients, 38 of 45 healthy siblings, and 20 of 26 healthy controls who gave a precise statement on their measles history. Of 16 individuals who denied previous measles infection 9 had a history of measles vaccination. Measles vaccination was first introduced in Sweden in 1969 and was fully implemented by 1980 when all individuals were vaccinated both at age 2 and age 12 years. A precise statement on measles vaccination status was provided by 33 patients, 33 healthy sibs, and 24 healthy controls. Two patients, both with at least two CSF-enriched OCBs, 6 healthy sibs, and 6 healthy controls, all 12 with fewer than two CSF-enriched OCBs, had been vaccinated against measles. Thus the majority of healthy siblings with at least two CSF-enriched OCBs was measles unvaccinated.

Quantitative neurological score

CDMS patients having a sib with at least two CSF-enriched OCBs (n=9) and those without (n=38) had essentially the same median RFSS [18], of 11.77 and 10.53 (P=0.4).

Discussion

We found that 9 of 47 (19%) of the healthy sibs to CDMS patients had an intrathecal immunological reaction with at least two CSF-enriched OCBs, distinct from 2 of 50 (4%) unrelated healthy controls. Furthermore, in these 9 sibs the measles CSF IgG antibody titers were higher than in the other sibs and in controls. There were also differences in serum titers for measles IgG antibody, which were higher in the group of all healthy sibs than in healthy volunteers, and (as with the CSF titers) higher in the subgroup of healthy sibs with at least two CSF-enriched OCBs than the other sibs. Thus a significant proportion of healthy siblings to MS patients have a partially hyperimmune condition, in 19% manifesting itself as an OCB reaction, in 9% as increased CSF measles IgG antibody titers, in 21% as increased serum measles IgG antibody titers. These phenomena tended to occur in the same individuals. Earlier studies were smaller, provided no serological data to exclude *Borrelia* and other CNS infections, and lacked normal CSF. In one study 5 of 28 (17%) of healthy siblings of MS patients had abnormal CSF immunoglobulins [8]. We found a 4% frequency of at least two CSF-enriched OCBs in our healthy volunteers. A frequency of 4–7% intrathecal immunological reactions were found in healthy students and tension headache patients using IEF [16, 32]. However, specific serology was not reported in these studies either. Some asymptomatic persons with CSF-enriched OCBs were reported to develop later manifest MS [35]. One healthy sib who has been reported to have an oligoclonal CSF reaction [2] (not in the present series) now has CDMS. The proportion of sibs to MS patients presenting with CSF-enriched OCBs in this study (19%) is much higher than the proportion of siblings to MS reported to develop manifest MS (3–4%) [27]. With a median age of 44 years in the present study, our individuals carrying at least two CSF-enriched OCBs are by and large beyond the risk age of MS. Thus only a minority of this group could be presymptomatic MS.

Our patients tended to have detectable CSF IgG antibodies to HSV–1 more often than siblings and controls. However, only measles antibody titers were significantly increased in the present study. MS patients harbor immunological reactivity against a number of viral and myelin antigens, while viral, predominantly measles, antigens have been reported in healthy siblings [25]. Imprint electroimmunofixation may reveal the antibody specificity of individual IgG bands [22]. The virus-specific antibodies constitute only a minor fraction of the IgG. This method may be used to study the specificity in the group of healthy siblings with CSF-enriched OCBs.

This intrathecal reaction was apparently not the sequela of an intercurrent CNS infection (HIV, *Borrelia*, lues, or acute meningitis). We did not include individuals with a history or signs of previous CNS infections. Usually some specificity remains in the intrathecal reaction after neu-

roborreliosis, but sometimes only an unspecific oligoclonal reaction in the CSF remains [14, 15]. However, these infections are usually severe enough to be revealed by clinical history, leading to the exclusion of one sibling pair from the present study. A positive *Borrelia* serology in an MS patient with no suggestive features of the infection is unlikely to indicate neurological Lyme disease [7]. Notably, the healthy volunteer with low serum *Borrelia* titers in the present series was not one of the individuals with at least two CSF-enriched OCBs. One individual with low positive *Borrelia* titers in the serum was found in each of our three groups, approaching the number expected from the 1.3–1.4% prevalence of serum *Borrelia* antibody-positive asymptomatic individuals in Gothenburg (Leif Dotevall, personal communication).

Could the MS-related partially hyperimmune state be identified by a noninvasive method, at least in the CNS compartment? Magnetic resonance imaging (MRI) detects subclinical abnormalities in the CNS. However, cerebrovascular disease is a differential problem in elderly persons. In previous studies two healthy siblings of 40 healthy relatives under the age of 50 years had MRI abnormalities, and 3 of 27 asymptomatic siblings had demyelinating MRI lesions, using strict criteria [20, 31]. In the British Isles survey only 3 of 33 discordant dizygotic twins showed MRI abnormalities fulfilling the Fazekas criteria [29].

The origin of this immunological reaction in MS patients and in their healthy sibs is unknown, but infectious and genetic factors and interaction between these two have been reviewed [3]. Twin studies [9, 21] suggest that both genetic and nongenetic factors are of importance. However, adoptee and half-sibling studies [10, 27] strongly indicate that the increased recurrence risk in siblings is predominantly genetic. We hypothesize that the analogously increased recurrence risk of CSF-enriched OCBs in siblings is genetically determined. Natural antibodies occur normally in low titers, which are genetically determined, and they may be important for control of early viral and bacterial distribution [23]. One way of differentiating antibodies in respect to their maturity concerning target recognition is analysis of the antibody avidity (which increases after the antibody recognized its antigen) using increasing salt concentrations in the antigen-antibody reaction [17]. It was reported that antibodies against measles in MS patients are predominantly of low affinity, while antibodies in patients with a primary viral infection are predominantly of high affinity [18]. Further characterization of the partially hyperimmune MS-related condition in larger materials should include viral and anti-myelin antibody specificity and avidity studies in the CSF and blood, cytokine patterns, and CSF markers for subclinical CNS damage, magnetic resonance spectroscopy, and HLA association studies.

In summary, we report that a MS-related partial hyperimmune state, which is predominantly but not totally confined to the CNS compartment, occurs in healthy siblings to MS patients with a considerable frequency, approximately five times higher than that of manifest MS in sibs. Pending further characterization of this condition, an analogy points to a genetic trait, or a genetic effect expressed in joint action with infections. However, data from the present study do not allow us to draw conclusions on the possible mode of inheritance of this condition. Both monogenetic inheritance with high gene frequency and polygenetic inheritance could explain our findings. Determining whether and how this phenotype is inherited requires family studies on its segregation. Further virological and genetic studies should compare the clinically asymptomatic and symptomatic subgroups of this MS-related hyperimmune state.

Acknowledgements This study was supported by the NHR (National Society of Neurologically Disabled), by the Gothenburg division of the NHR, and by the Faculty of Medicine, University of Gothenburg (L. U. A.).

References

1. Adams JM, Imagawa DT (1962) Measles antibodies in multiple sclerosis. Proc Soc Exp Biol Med 111:562–566
2. Andersen O, Lindholm A (1976) Tissue typing of CSF lymphocytes in MS. Acta Neurol Scand 54:464–466
3. Andersen O (1999) The rationale for anti-viral therapies in multiple sclerosis. Fredrikson S, Link H (eds) Advances in multiple sclerosis. Dunitz, London
4. Bergström T, Vahlne A, Alestig K, Forsgren M, Lycke E (1990) Primary and recurrent herpes simplex virus type 2-induced meningitis. J Infect Dis 162:322–330
5. Blennow K, Fredman P, Wallin A, Gottfries CG, Frey H, Pirttilä T, Skoog I, Wikkelsö C, Svennerholm L (1994) Formulas for the quantification of intrathecal IgG production. J Neurol Sci 121:90–96
6. Brody JA, Sever JL, Henson TE (1971) Virus antibody titers in multiple sclerosis patients, siblings and controls. JAMA 216:1441–1446
7. Coyle P, Krupp L, Doscher C (1993) Significance of reactive Lyme serology in multiple sclerosis. Ann Neurol 34:745–747
8. Duquette P. Charest L (1986) Cerebrospinal fluid findings in healthy siblings of multiple sclerosis patients. Neurology 36:727–729
9. Ebers GC, Bulman DE, Sadovnick AD, Paty DW, Warren S, Hader W, Murray TJ, Seland P, Duquette P, Grey T, Nelson R, Nicolle M, Brunet D (1986) A population-based study of multiple sclerosis in twins. N Engl J Med 315:1638–1642
10. Ebers GC, Sadovnick AD, Risch NJ (1995) A genetic basis for familial aggregation in multiple sclerosis. Nature 377:150–151

11. Felgenhauer K, Reiber H (1992) The diagnostic significance of antibody specificity indices in multiple sclerosis and herpes virus induced diseases of the nervous system. Clin Investig 70:28–37
12. Forghani S, Schmidt N (1979) Antigen requirements, sensitivity and specificity of enzyme immunoassay for measles and rubella virus antibodies. J Clin Microbiol 9:657–664
13. Henson TE, Brody JA, Sever JL, Dyken ML, Cannon J (1970) Measles antibody titers in multiple sclerosis patients, siblings and controls. JAMA 211:1985–1988
14. Kaiser R (1994) Variable CSF findings in early and late Lyme neuroborreliosis: a follow-up study in 47 patients. J Neurol 242:26–36
15. Kaiser R (1995) Intrathecal immune response in patients with neuroborreliosis: specificity of antibodies for neuronal proteins. J Neurol 242:319–325
16. Kostulas V, Link H, Lefvert AK (1987) Oligoclonal IgG bands in cerebrospinal fluid. Principles for demonstration and interpretation based on findings in 1114 neurological patients. Arch Neurol 44:1041–1044
17. Luxton RW, Thompson EJ (1990) Affinity distributions of antigen-specific IgG in patients with multiple sclerosis and in patients with viral encephalitis. J Immunol Methods 131:277–282
18. Luxton RW, Zeman A, Holzel H, Harvey P, Wilson J, Kocen R, Morgan-Hughes J, Miller DH, Compston A, Thompson E (1995) Affinity of antigen-specific IgG distiguishes multiple sclerosis from encephalitis. J Neurol Sci 132:11–19
19. Lycke J, Svennerholm B, Hjelmquist E, Frisén L, Badr G, Andersson M, Vahlne A, Andersen O (1996) Acyclovir treatment of relapsing-remitting multiple sclerosis. A randomized, placebo-controlled, double-blind study. J Neurol 243:214–224
20. Lynch SG, Rose JW, Smoker W, Petajan JH (1990) MRI in familial multiple sclerosis. Neurology 40:900–903
21. Mumford CJ, Wood NW, Keller-Wood H, Thorpe JW, Miller DH, Compston DAS (1994) The British Isles survey of MS in twins. Neurology 44:11–15
22. Nordal HJ, Vandvik B, Norrby E (1978) Multiple sclerosis Local synthesis of electrophoretically restricted measles, rubella, mumps and herpes simplex virus antibodies in the central nervous system. Scand J Immunol 7:473–479
23. Ochsenbein AF, Fehr T, Lutz C, Suter M, Brombacher F, Hengartner H, Zinkernagel RM (1999) Control of early viral and bacterial distribution and disease by natural antibodies. Science 286:2156–2159
24. Panelius M, Salmi A, Halonen PE, Kivalo E, Rinne UK, Penttinen K (1973) Virus antibodies in serum specimens from patients with multiple sclerosis, from siblings, and matched controls. A final report. Acta Neurol Scand 49:85–107
25. Poser CM (1986) Pathogenesis of multiple sclerosis. A critical reappraisal. Acta Neuropathol (Berl) 71:1–10
26. Poser CM (1993) The pathogenesis of multiple sclerosis. Additional considerations. J Neurol Sci 115:S3–S15
27. Sadovnick AD, Ebers GC, Dyment DA, Risch NJ (1996) Evidence for genetic basis of multiple sclerosis. Lancet 347:1728–1730
28. Svenningsson A, Runmarker B, Lycke J, Andersen O (1996) Incidence of multiple sclerosis during two fifteen-year periods in the Göteborg region of Sweden. Acta Neurol Scand 82:161–168
29. Thorpe JW, Mumford CJ, Compston DAS, Kendall BE, MacManus DG (1994) British Isles survey of multiple sclerosis in twins: MRI. J Neurol Neurosurg Psychiatry 57:491–496
30. Tibbling G, Link H (1977) Principle of albumin and IgG analyses in neurological disorders. Establishment of reference values. Scand J Clin Lab Invest 37:385–390
31. Tienari PJ, Salonen O, Wikström J, Valanne L, Palo J (1992) Familial multiple sclerosis: MRI findings in clinically affected and unaffected siblings. JNNP 55:883–886
32. Tourtellotte WW, Straugaitis SM, Walsh MJ, Shapshak P, Baumhefner RW, Potvin AR, Syndulko K (1985) The basis of intra-blood-brain-barrier IgG synthesis. Ann Neurol 17:21–27
33. Wikkelsö C, Andersson M, Andersson R, Blomstrand C (1984) Isoelectric focusing followed by silver staining. A suitable method for routine investigation of cerebrospinal fluid proteins. Eur Neurol 23:306–312
34. Woyciechowska JL, Dambrozia J, Leinikki P, Shekarchi C, Wallen W, Sever J, McFarland H, McFarlin D (1985) Viral antibodies in twins with multiple sclerosis. Neurology 35:1176–1180
35. Xu X, McFarlin DE (1984) Oligoclonal bands in CSF: twins with MS. Neurology 34:769–774