Two Contributions to Genetic Linkage Analysis

Staffan Nilsson
Chalmers University of Technology

November 22, 1999

Abstract

The thesis consists of two papers related to a method in statistical genetics called linkage analysis.

Estimation of $\lambda_{S(non-HLA)}$ in coeliac disease.

We focus on the question if it is possible to find other genes than the known coeliac disease HLA-component with a linkage analysis of 100 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 $p_0/0.25$, where p_0 is the probability of two affected sibs sharing no haplotypes ibd at HLA. When not only linkage to HLA, but also the disease alleles are known we derive an alternative way of estimating the sibling relative risk for the nonHLA component.

Model based sampling and weights in affected sib pair analysis.

When running a genome scan with affected sib pairs and a nonparametric 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.

Acknowledgements

Professor Olle Nerman has been my supervisor. He is a man with a warm heart.

I have had the opportunity to cooperate with geneticists, doctors and other professionals at Sahlgrenska University Hospital in the past few years. It has been an interesting time and I've met a lot of gentle people there. Thank you everybody!

Kajsa Fröjd has been my room mate and "little sister". We've had long discussions and laughs over the most various subjects. No subject too big or too small. The longest laugh over the shortest word "a".

Finally of course, Suzanne, Kalle, Petter and Karin have kindly accepted the financial suicide that I have committed, although I can't complain about my current salary. After all Kalle, when having a quick glance in my books, summarised what all descent people think:
- This is no real job!

1 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 each chromosome is many millions of bases long. Since the chromosomes come in pairs a commonly used term for a unit is *base pair (bp)*. Subsequences of a chromosome consists of functional units called *genes*. They are the templates for building proteins. The position of a gene on a chromosome is called a *locus*. During evolution genes are targets for *mutations* that are changes of letters in the sequence. This gives rise to variants of genes. The variants are called *alleles*. The variation of the genes cause variation in functions. A pair of alleles from a chromosome pair at a specific locus is called a *genotype*.

Between the genes are monkey typed messages that don't 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 comes from the mother, *maternal*, and one comes from the father, *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 *meios* when the egg is formed. Similarly, in the male meios, 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 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.

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

Our focus will be on finding *disease loci* by *linkage analysis*. Two loci are said to be *linked* if they are physically linked in the sense that they reside

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 reside on different chromosomes the recombination probability, called *recombination fraction* θ is $1/2$ due to the independence between chromosomes in the meios. If they reside on the same chromosome the recombination fraction is equal to the probability of an odd number of crossovers between them. In this case $\theta \leq 1/2$.

In linkage analysis we test whether a possible disease locus resides on a certain chromosome by studying how markers have segregated in a pedigree in relation to the *phenotypes* (affection status) of the pedigree members. So the collected data consists of genotypes and phenotypes of members of the pedigree.

One way of formulating hypothesises for testing linkage is as

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

As a test statistic we can use the likelihood ratio

$$\frac{L_{\hat{\theta}}}{L_{1/2}}$$

In *parametric linkage 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. Their meaning are $p = P(\text{disease allele frequency})$ while $f_i = P(\text{affected}|i \text{ disease alleles})$ are the genotype *penetrances*. For example a monogene, fully penetrant diease has the parameters $(p, 0, 1, 1)$

Another approach to linkage testing is known as *affected sib pair analysis*. The idea is to count the numbers of alleles at a marker loci 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 hypothesises is then 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

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

which is tested with a sum of ibd counts.

This was a rather condensed description of the concepts used in the papers. For more details refer to any textbook in genetics, e.g. Human Molecular genetics, Strachan and Read, Wiley, 1997 and for the statistical aspects of linkage consult, Analysis of Human Genetic Linkage, Jurg Ott, John Hopkins University Press, 1999.

Estimation of $\lambda_{S(nonHLA)}$ in coeliac disease

Staffan Nilsson
Chalmers University of Technology

Abstract

We focus on the question if it is possible to find other genes than the known coeliac disease HLA-component with a linkage analysis of 100 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 $p_0/0.25$, where p_0 is the probability of two affected sibs sharing no haplotypes ibd at HLA. When not only linkage to HLA, but also the disease alleles are known we derive an alternative way of estimating the sibling relative risk for the nonHLA component.

1 Introduction

Coeliac disease is defined as a permanent intolerance to dietary gluten, i.e. the main storage protein in wheat, rye oats and corn. It has a genetic component, which is assumed to be multifactorial. Association to the HLA-region on chromosome 6 is well established [Petronzelli:1997]. However this does not explain all the relative risk seen in siblings and other relatives, which indicates that there should be a nonHLA component.

In an ongoing study we have access to 108 nuclear families, where at least two children are affected by coeliac disease. We use the software GeneHunter [Kruglyak:1996] to perform non-parametric linkage (NPL) analysis.

In an early phase of the project we were concerned whether linkage analysis would have enough power to detect an assumed nonHLA component. The doubts were based on [Risch:1996] and [Camp:1997]. Thus, admittingly too late, we decided to perform power calculations for some reasonable alternatives. To this end we needed to examine the strength of the nonHLA-component.

One general measure of the genetic strength of a disease is the relative risk $\lambda_s = K_S/K_P$, where K_P is the population prevalence and K_S is the sibling prevalence. The size of λ_S will obviously influence the power of a genome scan.

A similar measure can be defined for each of the various genetic subcomponent of the disease. For a subcomponent A we have $\lambda_{s(A)} = K_{S(A)}/K_{P(A)}$ [Risch:1990a]

2 Estimating $\lambda_{S(nonHLA)}$ through two methods

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

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 the assumptions and definitions in the last paragraph, the model can thus be viewed as if coeliac disease was 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

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

With a multiplicative model it was derived in [Risch:1987]

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

where

$$p_0 = P(\text{sharing 0 HLA-haplotypes} | \text{2 affected siblings})$$

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 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, an MLE 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. If there was a phenotype "HLA affected" we could directly estimate $K_{S(HLA)}$ and $K_{P(HLA)}$ from population data. For us "HLA affected" is merely an abstract concept. When we know, not only the location of the disease locus, but also the alleles involved (as in our case DQA1*0501,DQB1*02) [Sollid:1989] we could perhaps calculate $K_{S(HLA)}$ and $K_{P(HLA)}$ directly. 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. But we don't know all the genotypes G_i and we don't know the penetrance for the one we know. Still, we will use our knowledge of one of the genotypes to derive an alternative formula.

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

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

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 but perhaps questionable 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: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 at this stage, but it will be necessary to estimate the population frequency of G. The latter is of course something that is normally already done if G is really known 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 $P(G_0)$.

3 Power calculations

Using the sib pair family material gives an estimate of $p_0 = 0.05$ from GeneHunter 2.0. Together with the estimates $K_S = 0.08$ and $K_P = 0.004$ from

a previous study [Ascher:1997] this gives an estimate of $\lambda_{s(nonHLA)} = 4$ by Risch method.

Other studies that estimated p_0 are [Greco:1998] that found $p_0 = 0.09$ and [Petronzelli:1997] with $p_0 = 0.07$.

Using the data in [Ploski: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. We also speculated that there could be a lot more loci, but if the two most common causes are much more frequent than the rest this model will be sufficiently close for our purpose.

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:1995]. The power was calculated for an ASP method with 100 sib pairs and a dense set of completely informative markers. In reality there will of course not be any completely informative markers, but on the other hand we run multi-locus linkage and among the 108 coeliac disease families there are several with more than two affected sibs.

4 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 coeliac disease this knowledge is however available and we use it to derive an alternative estimator. (In the current situation this gave a better estimate in the sense of having less variance.) Since the methods require different kind of samples it is not straightforward to compare them, but in general it is a good principle

Table 1: Power to detect linkage for various models

$\lambda_{S(HLA)}$	allele freq		α	power		
	1	2		1	2	any
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

to use all available knowledge.

The main purpose of estimating $\lambda_{S(nonHLA)}$ was to be able to perform power calculations for a genome scan and an important use of power calculations is a base for choosing suitable sample sizes, although in the current application the sample was already collected. By using (5) we can estimate $\lambda_{S(nonHLA)}$ before we collect a large amount of sib pairs. The only sib pairs we need are for the estimation of the entity P(affected|G,aff sib). In contrast (2) requires that we collect a large amount of the final sample in order to be able to estimate how many we finally need, with the risk of collecting nuclear families in vain since the amount needed to achieve reasonable power is not possible to reach.

When applying the two methods we got the estimates $\lambda_{S(nonHLA)} = 4$ with (2) and $\lambda_{S(nonHLA)} = 9$ with (6).

Although we have made lots of more or less well founded model assumptions, the results of the power calculations illustrates the impact of the size of $\lambda_{S(nonHLA)}$ and thus the importance of an accurate estimate.

With $\lambda_{s(HLA)} = 9$ the worst case (power 32%) under the assumption of nonHLA heterogeneity is, not surprisingly, dominant inheritance with equally frequent disease alleles and a conservative significance level.

With such power one would generally not start an expensive investigation. However there are several reasons to not be that pessimistic.

- There is no reason why the frequencies should be equal, since they come from different mutation.
- The power is calculated for a pure ASP method, i.e. the parents affection status is not taken into account. The affection set method of Genehunter is generally more powerful in dominant cases.
- We have DNA samples of healthy siblings and could include them to look for decreased allele sharing between pairs of affected - unaffected that both have disease haplotypes in HLA.
- The point wise significance level 0.000022 corresponds to a probability of 0.05 of any false positives in a whole genome scan. If we use suggestive linkage we expect 1 false positive. The time that will be wasted on deeper investigations of a few false positive could be acceptable as long as a true positive is among the candidate regions.

Based on all this, and perhaps a flavour of hope, we decided to continue with a full genome scan.

References

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

[Camp:1997] Camp NJ. Genomewide transmission/disequilibrium testing - consideration of the genotypic relative risks at disease loci. *Am J Hum Genet.* **61**:1424-1430.

[Greco:1998] Greco et al. Genome search in celiac disease. *Am J Hum Genet.* **62**:669-675.

[Holmans:1993] Holmans P. Asymptotic properties of affected sib-pair linkage analysis. *Am J Hum Genet.* **52**:362-374.

[Kruklyak:1996] Kruklyak L, Daly MJ, Reeve-Daly MP, and Lander ES. Parametric and nonparametrice linkage analysis: A unified multipoint approach. *Am J Hum Genet.* **58**:1347-1363.

[Lander:1995] Lander E and Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting results. *Nature Genetics.* **11**:241-247.

[Petronzelli:1997] Petronzelli et al. Genetic contribution of the HLA region to familial clustering of coeliac disease. *Ann Hum Genet.* **61**:307-317.

[Ploski:1996] Ploski R, Ascher H, and Sollid LM. HLA genotypes and the increased incidence of coeliac disease in sweden. *Scand J Gastroentero.* **31**:1092-1097.

[Risch:1987] Risch N. Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet.* **40**:1-14.

[Risch:1990a] Risch N. Linkage strategies for genetically complex traits. I. multilocus models. *Am J Hum Genet.* **46**:222-228.

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

[Sollid:1989] Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, and Thorsby E. Evidence for a primary association of celiac disease to a particular HLA-DQ α/β heterodimer. *J Exp Med.* **169**:345-350.

Model based sampling and weights in affected sib pair methods

Staffan Nilsson
Chalmers University of Technology

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.

1 Introduction

In the planning of a genome scan with nuclear families we can get a rough idea about suitable sample sizes for an Affected Sib Pair (ASP) statistic by performing power calculations. Later on, when the samples have 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 difference in size and affection statuses 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:1998] by preferring good sib pairs and we can use different weights [Sham: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.

2 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 $p_k = P_k($ ibd at disease locus$|2$ affected$)$ for $k = 0, 1, 2$ to calculate the power.

These quantities will, when the alternative hypothesis is a biallelic monogenic 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

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

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 statuses 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 this 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

$$\begin{aligned} c_2 &= Cov(I_{ij}, I_{ij}) \\ c_1 &= Cov(I_{ij}, I_{jl}) \\ c_0 &= Cov(I_{ij}, I_{kl}), \end{aligned}$$

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:1997] but with the mistake of assuming $c_0 = 0$. The results of the calculations are shown in appendix A.

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:1997] and shown in appendix table 13. 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:1997] since Y_{ij} and Y_{jl} are not independent conditional on (I_{ij}, I_{jl}, A).

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

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

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 14,15 and 16.

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 coefficient of variation (CV) for Z_i in various family structures is

Table 1: CV 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*

shown in table 2. Calculated for an affected sib pair, disregarding any knowledge of other family members, we get CV=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.

2.1 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: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. This difference is taken into account in [Whittermore:1994b] and is implemented in Genehunter [Kruglyak:1996].

Other statistics focus on the estimation of the unconditional (p_0, p_1, p_2) with [Holmans:1993] or without [Risch:1990c] 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: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 would need to have special scores and weights as well.

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:1979] and [Hodge:1984]. These are compared in [Sham: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

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

  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: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

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

  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 [Kruglyak:1996].

- **Equal weights on every sib pair**

  If we don't 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 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

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

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

Thus CV 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 2 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 3 and the power was calculated for optimum weights at significance levels 0.01, 0.00074 (suggestive linkage) and 0.000022 (significant linkage) [Lander:1995], equal family weights (EqF), equal sib pair weights (EqSP) and weights proportional to expectation (Exp), expectation/variance (E/V) and coefficient of variation (CV). Results are shown in table 4. 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.

2.2 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

Table 2: 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				

Table 3: 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
CV	0.9788	0.8779	0.5966

parameters $K_P = 0.005$ and $\lambda_S = 2$. We see 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. 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

Table 4: 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
	CV	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
	CV	0.7468	0.4262	0.1376	0.7821	0.4726	0.1658

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

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

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?

2.3 Sampling strategies

The idea is that we shall of course preferably collect good families before bad. In [McCarthy:1998] a number of general strategies for collecting sib pairs are

Table 5: 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*

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).

We can describe the sampling strategy as a two step strategy, where we first examine the phenotypes of an available population A of families with at least two affected sibs and then select the N best families in A, i.e. the families with largest 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 7 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 We could apply a mixed strategy as well, with weights based on (11) and with a selection criteria where for each family $CV_i^{min} = min(CV_i^{M_1}, CV_i^{M_2})$ and the 40 families are selected according to the largest CV_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 4 and define $T_k = \sum_1^k Z_i$ for $k = 1, ..40$ where the Z_i are ordered according to their coefficient of variation such that Z_i has the largest.

The four families with 3 affected children had the largest CV. Using only

Table 6: Power to detect significant linkage with selective sampling

Correct	M_2	M_2	M_1	M_1
Assumed	M_2	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
CV	0.465	0.427	0.979	0.921

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 7: 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

2.4 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.

Every now and then we will of course be 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.

A Appendix

Table 8: IBD count probabilities given mating types

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

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

Table 9: 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, $f_G = P(\text{affected}|G, h_{mG} = P(m \text{ affected parents}|G)$

Table 10: 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,
$p_{iG} = P(\text{i IBD alleles}|\text{mating type } G)$, G = mating type,
$f_G = P(\text{affected}|G)$, $h_{mG} = P(m \text{ affected parents}|G)$

Table 11: 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,
$p_{iG} = P(\text{i IBD alleles}|\text{mating type } G), G =$ mating type,
$f_G = P(\text{affected}|G), h_{mG} = P(m\text{affected parents}|G)$

Table 12: 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

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

Table 13: 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	

Table 14: Joint IBD probabilities at marker in special case with three siblings.

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

$$\gamma = P((1,1)^S \mid (I_{ij}, I_{jl}) = (1,1), A)$$

Table 15: 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	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_1^3+f_1^2f_2+f_1f_2^2+f_2^3)/4$
DDxdd	$h_{mG}(1-f_G)^{n-a}f_G^a$
DdxDd	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}f_1(f_0^2+f_0f_1+f_1f_2+f_2^2)/4$
Ddxdd	$h_{mG}(1-f_G)^{n-a}f_G^{a-3}(f_0^3+f_0^2f_1+f_0f_1^2+f_1^3)/4$
ddxdd	$h_{mG}(1-f_G)^{n-a}f_G^a$

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

References

[Hodge:1984] Hodge SE. The information contained in multiple sibling pairs. *Genetic Epidemiology.* **1**:109-122.

[Holmans:1993] Holmans P. Asymptotic properties of affected sib-pair linkage analysis. *Am J Hum Genet.* **52**:362-374.

[James:1971] James J. Frequency in relatives for an all-or-none trait. *Ann Hum Genet.* **35**:47-49.

[Knapp:1995] Knapp M, Seuchter SA, and Bauer MP. Linkage analysis in nuclear families, i. optimality criteria for affected sib-pair tests. *Human Heredity.* **44**:37-43.

[Kruglyak:1996] Kruklyak L, Daly MJ, Reeve-Daly MP, and Lander ES. Parametric and nonparametrice linkage analysis: A unified multipoint approach. *Am J Hum Genet.* **58**:1347-1363.

[Lander:1995] Lander E and Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting results. *Nature Genetics.* **11**:241-247.

[McCarthy:1998] McCarthy MI, Kruglyak L, and Lander ES. Sib-pair collection strategies for complex diseases. *Genetic Epidemiology.* **15**:317-340.

[Risch:1990c] Risch N. 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:1997] Sham PC, Zhao JH, and Curtis D. Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann Hum Genet.* **61**:61-69.

[Suarez:1978] Suarez BK. The affected sib pair ibd distribution for hla-linked disease susceptibility genesd. *Tissue Antigens.* **12**:1-14.

[Suarez:1979] Suarez BK and Hodge SE. A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clinical Genetics*. **18**:126-136.

[Whittermore:1994b] Whittemore AS and Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics*. **50**:118-127.