

Thesis Introduction: A Survey on Nonparametric Linkage Analysis and Related Issues

Lars Ängquist¹

4th March 2007

¹Department of Mathematical Statistics, Centre for Mathematical Sciences,
Lund University, Lund, Sweden.

Abstract

In linkage analysis or, in a wider sense, gene mapping one searches for disease loci along a genome. This is done by observing so called marker genotypes and phenotypes of a pedigree set, i.e. a set of multigenerational families, in order to locate the loci corresponding to the underlying disease genes or, at least, to narrow down the interesting genome regions.

In this context the key concept is the genetic inheritance of alleles with respect to the phenotype outcomes. A significant deviation from what is expected under random inheritance is taken as statistical evidence of existing genetic components suggested to be located at the loci giving significant results.

In this thesis introduction we begin by outlining the needed genetical foundation of statistical genetics as well as some basic concepts, for instance, the process of allelic inheritance, the genetic model, the pedigree set, the inheritance vector and various types of genetic information. Next, we give an introduction to one-locus nonparametric linkage analysis focusing on significance calculations of nonparametric linkage (NPL) scores and, moreover, make some comments on the generalizations to two-locus procedures and the, related but contrasting, approach of parametric linkage analysis. In the third section we very briefly discuss some competing and complementary sub-fields within the context of statistical genetics and finally we put the papers included in this thesis into context by summarizing their content.

Table of Contents

1	Genetics	2
1.1	Basic Notation and Key Concepts	2
1.2	The Inheritance of Alleles	4
1.3	Mendel, Markers and General Information	8
1.4	The Genetic Model	10
1.5	The Pedigree Set and Allele-Sharing	12
1.6	The Inheritance Vector and Entropy-Based Information Con- tents	14
1.7	How to Collect Information	17
2	Nonparametric Linkage Analysis	18
2.1	Parametric Linkage Analysis	19
2.2	Score Functions	20
2.3	The NPL Score	22
2.4	Calculating the Statistical Significance	24
2.5	Significance Calculations through Theoretical Approximation .	26
2.6	Significance Calculations through Monte Carlo Simulation . .	28
2.7	Two-Locus NPL Analysis	29
3	Other Statistical Genetics Procedures	32
4	Outline of Papers in Thesis	34
4.1	Paper A	34
4.2	Paper B	35
4.3	Paper C	38
4.4	Paper D	39
	References	42
A	Some Notation Used in the Thesis Introduction	52

1 Genetics

In this introductory section we will present some background notation and information, hopefully laying the foundation for subsequent understanding, possibly increasing the utility, of the material to come. More thorough and detailed treatments of included topics are e.g. given in the textbooks Sham (1998), Ott (1999), Almgren et al. (2003), Strachan and Read (2003) and Haines and Pericak-Vance (2006).

1.1 Basic Notation and Key Concepts

The human *genome*, i.e. operating manual, consists of 23 pair of *chromosomes*. In total 22 pairs are so called *autosomes* which are structurally equal with respect to the sexes, whereas 1 pair constitutes the *sex-chromosomes* which are content-wise sex-dependent. Throughout this work we will only consider autosomes. The main chemical structure of the chromosomes is the *deoxyribonucleic acid (DNA)*, which is based on units called *nucleotides* that each consists of a sugar (deoxyribonucleic), a phosphate and a nitrogenous base. There are 4 possible bases available: adenine (A), cytosine (C), guanine (G) and thymine (T).

The physical structure of DNA is actually double-stranded, in the form of a double-helix, but since the two strands are strictly complementary¹ one might look at the chromosomal DNA as a single sequence of nucleotides represented by the underlying bases. An alternative, but essentially similar, view is to consider the whole (or parts of the) sequence as a word written using the $\{A, C, G, T\}$ -alphabet.

The genetic code is organized into nucleotide-triplets, *codons*, which code for specific *amino acids*. Obviously we may define $4^3 = 64$ different codons, but the code is degenerate in the sense that only 20 different amino acids may be produced. This follows since several codons may code for the same amino acid and some codons serve as code punctuation.²

In total the human genome consists of approximately 3×10^9 *base pairs (bp)*.³

¹The genetic codes with respect to the two strands are structurally equivalent since the only admissible bonds between these strands are $A - T$ and $C - G$.

²*Start* and *stop* codons, which tell the code interpreter to start or stop reading, i.e. to begin or end a coding region.

³One base-pair may be seen as one single position in the DNA sequence or, equivalently understood, as one letter in the genomewide genetic word, i.e. as one instance of either A ,

A *gene* is a sequence of DNA, i.e. a unique amino acid sequence, at a specific genome position which specifies the function and structure of a subunit in a *protein*. Some genome regions are in this sense informative coding regions (*exons*) and the space between them are noncoding regions (*introns*). Recent investigations specify the number of distinct genes to be about 20000-25000 (International Human Genome Sequencing Consortium, 2001, 2004). The actual nucleotide-length of genes shows great variation.

A well-defined position⁴ on the genome is called a *locus* (pl. *loci*). Loci located on the same chromosome are *syntenic*. The opposite (complementary) term is *nonsyntenic*. At each locus a human-being hosts two *alleles*.⁵ These alleles constitute the individual's *genotype*. Different nucleotide sequences at the same locus give rise to different *allelic variants* (polymorphisms). If a genotype consists of two copies of the same allelic variant it is called *homozygous*, otherwise it is *heterozygous*.

At a locus we may assume that the genetic variation is summarized by a different allelic variants A_1, A_2, \dots, A_a . For an allele of a randomly chosen individual, these variants occur with probabilities p_1, p_2, \dots, p_a . A common criterion of a *polymorphic locus* is that $a \geq 2$ and $p_i \leq 0.95$ ($\forall i$), but sometimes a slightly less restrictive constraint on the probabilities is used (e.g. $p_i \leq 0.99$). The number of distinct genotypes is $a(a+1)/2$.⁶

A general probabilistic assumption on the formation of distinct genotypes is the *Hardy-Weinberg equilibrium (HWE)*:

Definition 1. *HWE means that the genotype probabilities are directly proportional to the two corresponding allele frequencies, i.e. $P(A_i A_i) = p_i^2$ and $P(A_i A_j) = 2p_i p_j$ ($i \neq j$).*

Considering a randomly chosen individual this reflects a completely random formation of a genotype given the set of allele frequencies.⁷

C, G or T.

⁴Should be understood as a small chromosomal segment.

⁵This makes the human organism a diploid species.

⁶Distinct here means that the unordered genotypes are different, i.e. the genotypes $A_i A_j$ and $A_j A_i$ are not distinct.

⁷The genotype's second allele is not dependent on the first allelic outcome.

1.2 The Inheritance of Alleles

Of the 46 human chromosomes 23 are inherited from the father⁸ (*paternal* chromosomes) and 23 are inherited from the mother (*maternal* chromosomes). This implies that for each individual, at each locus, the genotype consists of one paternal and one maternal allele. A simple inheritance example is given in Figure 1. Knowledge of the parental origins of an individual's alleles implies that one may form an *ordered genotype* where the so called *phase* is known. More generally, knowledge of an individual's ordered genotypes at several syntenic loci implies that one may infer the corresponding *haplotypes*, see Figure 2.

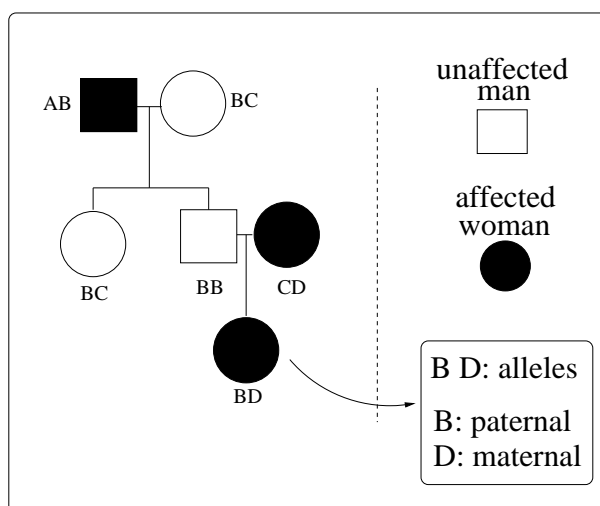


Figure 1: An example of the inheritance of alleles for a single small pedigree.

Each chromosome consists of mixed segments from the two corresponding grandparents. In other words there is a *blockwise chromosomal inheritance* of alleles interchangingly from both the grandparental chromosomes. The positions where one block ends and a new one starts are called *crossovers*. This behaviour is explained by the biological process of the formation of *gametes*, i.e. sperm and ovum cells, which is called *meiosis*.

During meiosis all the chromosomes are duplicated and then the homologous chromosomes pair up, i.e. we have initially an arrangement of four (2×2) DNA strands known as *chromatids*. Now, some physical contact between

⁸One from each pair of *homologous* chromosomes.

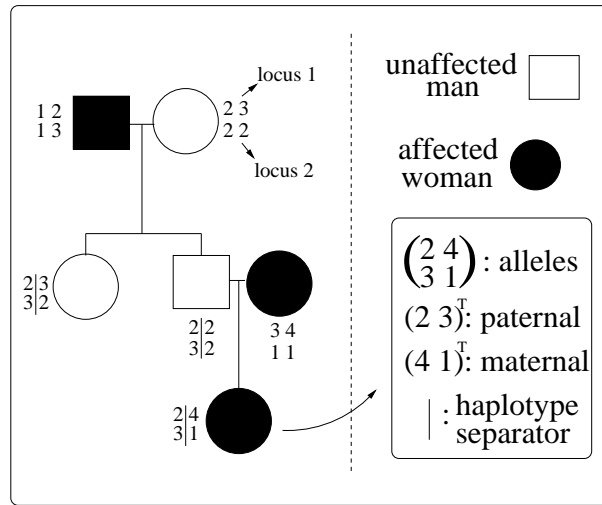


Figure 2: An example of the inheritance of haplotypes for a single small pedigree. We assume that the founder genotypes are phase-unknown. This is typically the case unless information from previous generations is available.

nonidentical chromatids may occur. These positions are known as *chiasmata* and correspond to crossovers. Moreover, there is at least one chiasmata per chromosome pair. Finally, one of these genetically mixed chromatids is chosen, so to speak, for reproduction. This complex process is schematically shown in Figure 3.

Inheritance at different chromosomes is independent. For each chromosome, occurrence of a crossover at a locus usually lowers the probability of having a second crossover in close vicinity of this locus. This phenomenon is referred to as positive *chiasma interference*. Moreover, the human genome hosts both so called *hot spots* and *cold spots*, which are chromosomal regions with high and low intensity of crossovers respectively. Often it is a good approximation to ignore chiasma interference and assume that crossovers occur randomly according to a Poisson process. If variation of spot temperature is ignored as well, the Poisson process has constant intensity.

Equipped with the concept of crossovers one may introduce a new measure of distance. The *genetic distance* g is measured in units of Morgans,⁹ based on the concept of expected number of crossovers between loci. Formally, given two loci, l_1 and l_2 , located 1 Morgan from each other, there is an expected

⁹The alternatively used *physical distance* is simply measured in base-pairs.

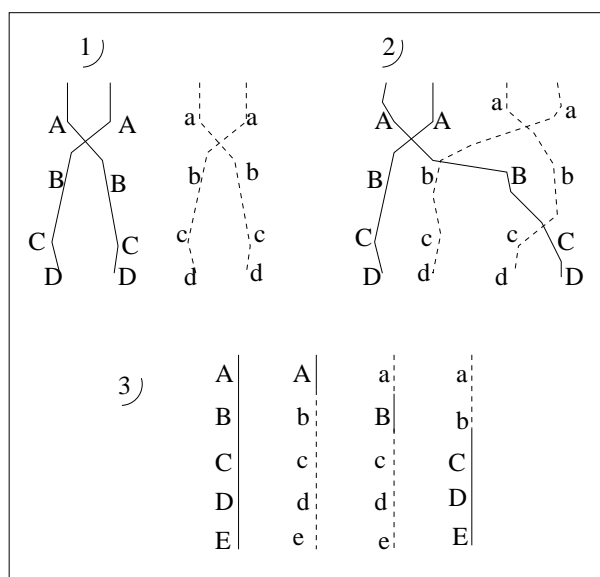


Figure 3: A simple overview of a single meiosis: 1) The replicated homologous chromosomes pair up. 2) Some physical contact at chiasmata occur between the two pairs of chromatids. 3) Four mixed strands of DNA ready for reproduction.

number of 1 crossover for each meiosis, with respect to the actually inherited chromatid, between these loci. Since crossovers occur with higher intensity for females than for males, this distance measure is really sex-dependent, but often one uses sex-averaged numbers. Adopting the latter approach gives us a total genetic length of 35.75 Morgans of the human autosomes (Collins et al., 1996). On average $1cM = 0.01M$ corresponds to a physical distance of 10^6bp although the intensity of crossovers varies along a chromosome.

Alternatively one may define genomic distance using the concept of *recombination*. If the alleles transmitted by a parent at two loci, l_1 and l_2 , are inherited from the same grandparent they are said to be *nonrecombinants*. Otherwise, they are referred to as *recombinants*. If a recombination has occurred between l_1 and l_2 we know, by definition, that there has been an odd number of crossovers between them.

The probability of recombination between two arbitrary loci l_1 and l_2 is called the *recombination fraction* and is denoted by $\theta = \theta(l_1, l_2)$. Obviously, according to the blockwise inheritance, this parameter is an increasing func-

tion of the genetic distance between the loci.

Normally there is a one-to-one function between the recombination fraction and the genetic distance. The link between these concepts is called the *map function*. Several suggested map functions exist in the literature, each choice corresponding to a specific way of modeling crossovers.

The most common one, which we exclusively use in this thesis, is the *Haldane map function* (Haldane, 1919), defined by,

$$g = -\frac{1}{2} \ln(1 - 2\theta), \quad (1)$$

where g is the genetic distance. This function corresponds to lack of interference and, as pointed out above, for each chromosome yields a $\text{Poisson}(g)$ distributed number of crossovers between l_1 and l_2 . From this follows that the distance between crossovers with respect to a single meiosis is exponentially distributed.¹⁰ Other well-known map functions are defined by Morgan (1928) and Kosambi (1944). The *Morgan function*,

$$g = \theta,$$

is valid if excluding multiple crossovers (complete interference) and may therefore be used as an approximation over short distances. Somewhat more involved is the *Kosambi function*,

$$g = \frac{1}{4} \ln \left(\frac{1 + 2\theta}{1 - 2\theta} \right),$$

which models interference as being large at small distances, then decreasing with distance.

To prove (1), we notice that an odd number of crossovers c is needed between two loci for them to be recombinant. If the crossovers are purely random (no interference) they follow a Poisson process with expected value $g = E(c)$. This gives,

$$\begin{aligned} \theta &= \sum_{i=1}^{\infty} P(c = 2i - 1 | g) = \sum_{i=1}^{\infty} \exp(-g) g^{2i-1} / (2i - 1)! \\ &= \left[\frac{1 - \exp(-2g)}{2} \right] \Rightarrow g = -\frac{1}{2} \ln(1 - 2\theta), \end{aligned}$$

¹⁰With intensity and mean value 1 if measured in Morgans. If using centiMorgans these numbers will be 0.01 and 100 respectively.

where we used that $\sum_{i=1}^{\infty} g^{2i-1}/(2i-1)! = \frac{1}{2} [\exp(g) - \exp(-g)]$, which is based on standard Taylor expansions.

Looking at (1), and most of its existing alternatives, one may note that $0 \leq \theta \leq 0.5$ is required.¹¹ This is no coincidence since generally two loci, l_1 and l_2 , are considered to be *unlinked*, i.e. the inheritance at these loci are independent, if $\theta = 0.5$. This is interpreted as the distance between these loci being infinitely large, $g(l_1, l_2) = \infty$, or that there is an infinite expected number of crossover between them, meaning that they are located on different chromosomes. The other extreme case is $\theta = 0$ which corresponds to $l_1 = l_2$, i.e. inheritance at these loci is completely dependent or *linked*. Intuitively this seems surprising, but it makes sense when we compare a hypothetical disease locus with a so called marker locus in close vicinity.

1.3 Mendel, Markers and General Information

In one sense modern research in genetics started with the Austrian monk Johann Gregor Mendel's (1822-1884) publication on the inheritance in the pea plant (Mendel, 1866). He observed the behaviour of *random inheritance (RI)* and *independent assortment (IA)*, which today are known as the first and second *Mendelian laws on inheritance*. Next, we will formally define these concepts and also introduce what is known as the assumption of *random mating (RM)*.

Definition 2. *RI means that the parental alleles are transmitted with equal probabilities to the offspring and this is done in an independent way when considering multiple offspring.*

Definition 3. *IA means that normally the inheritance at distinct loci are performed in an independent way. Deviances from this rule is referred to as genetic (or physical) linkage.*

Definition 4. *RM means that the mating of parents is not dependent on genetic factors, i.e. the probability that a mating couple has genotypes G_P (paternal) and G_M (maternal) is $P(G_P)P(G_M)$.*

Consider $|l|$ different loci. If we denote the number of allelic variants at the i^{th} locus with a_i , then we may form $h = a_1 a_2 \cdots a_{|l|}$ distinct haplotypes. This

¹¹In some cases with interference present, one may find $\theta > 0.5$. Generally this theoretical possibility is not considered to be of practical importance in the context of linkage analysis (Ott, 1999).

gives a total number of $h(h+1)/2$ different multi-loci genotypes and leads to the definition of so called *allelic association (AA)* or *linkage disequilibrium (LD)*.

Definition 5. *AA (or LD) means that the probability for at least one haplotype $A_{i_1}A_{i_2}\cdots A_{i_{|l|}}$ of a randomly chosen individual satisfies,*

$$P\left(A_{i_1}^1A_{i_2}^2\cdots A_{i_{|l|}}^{|l|}\right) \neq p_{i_1}^1p_{i_2}^2\cdots p_{i_{|l|}}^{|l|}.$$

Here $i_j \in \{1, 2, \dots, a_j\}$ is an index number and $A_{i_j}^j$ the corresponding allelic variant at the j^{th} locus which occurs with probability $p_{i_j}^j$.

Assume we want to perform a *genome scan* with respect to a genome region Ω . In order to do so we have to define *genetic markers* throughout Ω , facilitating the investigation of inheritance at these positions. A marker is a locus of known chromosomal position, where it is possible to measure allelic outcomes and where the population shows allelic variation, i.e. each marker locus is polymorphic.

In order to perform *linkage analysis* one needs to define polymorphic markers throughout Ω , estimate allele frequencies corresponding to all possible allelic variants at included markers and use this set of markers to produce a *marker map*.¹² This involves: (i) Ordering the markers with respect to chromosomal position. (ii) Specifying the distances between each consecutive pair of markers. If this is done using genetic distances one has produced a genetic marker map, whereas a physical marker map measures distances in base-pairs.¹³

The degree of polymorphism of a marker at locus $x \in \Omega$, with a allelic variants and corresponding allele frequencies p_1, p_2, \dots, p_a , may be defined in different ways: (i) Through the *heterozygosity (H)* value,

$$H(x) = 1 - \sum_{i=1}^a p_i^2. \quad (2)$$

¹²Strictly speaking, a marker map is only needed when performing multipoint analyses.

¹³Several different techniques for measuring or constructing genotyping markers exist. This includes, for example, *restriction fragment lengths polymorphisms (RFLPs)*, *microsatellite markers* and *single nucleotide polymorphisms (SNPs)*. See Strachan and Read (2003) and Haines and Pericak-Vance (2006).

This is the probability for an arbitrary individual of being heterozygote at locus x . (ii) Through the *polymorphism information content (PIC)* value,

$$PIC(x) = 1 - \sum_{i=1}^a p_i^2 - \sum_{i=1}^{a-1} \sum_{j=i+1}^a 2p_i^2 p_j^2, \quad (3)$$

where the last sum is the probability that a child's genotype is heterozygous with unknown phase.¹⁴ Both (2) and (3) quantify marker informativity on the population level.

1.4 The Genetic Model

Usually in linkage analysis one investigates inheritance of alleles with respect to a given disease. More generally one may divide the set of individuals in the study with respect to their *phenotypes*. This is a non-genetically observable quantity that may be qualitative or quantitative. Throughout this text we will only consider the binary qualitative phenotype of *affection status*. Typical examples of a quantitative phenotypes are body-mass-index (BMI) and body weight.

For an underlying disease to be genetically inheritable, i.e. to include a *genetic component*, some kind of correlation between the phenotype and the disease genotypes must exist. This is described by means of a *genetic model* λ . One may note that λ usually, at least to some extent, is unknown so, if needed, it is estimated prior to analysis using so called segregation analysis. Moreover, the disease may be governed by one or several different possibly interacting genes, *monogenic* and *polygenic* diseases respectively. The latter case is also referred to as a *complex disease*. If several distinct allelic variants at the same locus are susceptible with respect to the disease we speak of *allelic heterogeneity* and if more than one locus independently are susceptible to the disease we phrase this as *locus heterogeneity*.

The complete genetic model may be summarized as,

$$\lambda = (p, f, l), \quad (4)$$

where p is the set of *disease allele frequencies*, f is the set of *penetrance values*, describing the link between phenotypes and disease genotypes, and

¹⁴Each term corresponds to the probability that: (i) An arbitrary pair of parents has the unordered-genotype mating-type $A_i A_j \times A_i A_j$ ($i < j$). (ii) A corresponding offspring inherits the uninformative unordered-genotype $A_i A_j$.

l defines the *disease loci positions*. We will now more formally describe these components for the one-locus (monogenic) case and then make some comments about the two-locus case.

One-Locus Case Generally one assumes a *biallelic* disease locus with disease allele D and normal (wild-type) allele d . The disease allele frequency is denoted by $p = P(D)$ and the normal allele frequency by $q = 1 - p = P(d)$.¹⁵

The probabilistic link between the disease phenotypes and genotypes is given by the *penetrance vector*,

$$f = (f_0, f_1, f_2), \quad (5)$$

where $f_i = P(\text{affected} \mid i \text{ disease alleles})$. This gives the disease structure, whereas the disease allele frequency, with respect to this structure, decides how common the disease will be. Finally, in this case $l = l_1$ gives the actual location of the single disease locus.

Another parameter of interest is the *prevalence* K , i.e. the genetically-unconditional population-wise probability of disease, defined as,

$$K = f_0q^2 + f_12pq + f_2p^2.$$

If there is no genetic component of the disease,

$$f = (f_0, f_1, f_2) = (K, K, K).$$

Two-Locus Case This is a straightforward generalization of the one-locus case. The disease alleles are denoted D_1 and D_2 , where D_i is the disease allele at the i^{th} disease locus, having disease allele frequency parameter $p = (p_1, p_2)$ with $p_1 = P(D_1)$ and $p_2 = P(D_2)$.

The penetrance vector in (5) is generalized to a 3×3 *penetrance matrix*,

$$f = \begin{pmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{pmatrix},$$

where $f_{ij} = P(\text{affected} \mid i \text{ of } D_1, j \text{ of } D_2)$.

Finally, $l = (l_1, l_2)$ denotes the chromosomal positions of the two disease loci. Often l_1 and l_2 are assumed to be nonsyntenic, i.e. $c(l_1) \neq c(l_2)$ where $c(x)$ is the chromosome where x is located.

¹⁵In general D and d may be thought of as collections of several different allelic variants with similar phenotypic effects.

1.5 The Pedigree Set and Allele-Sharing

We assume there is phenotype and genotype information from a given *pedigree set*, which is a set of (possibly) multigenerational families. See Figure 4 for an example.

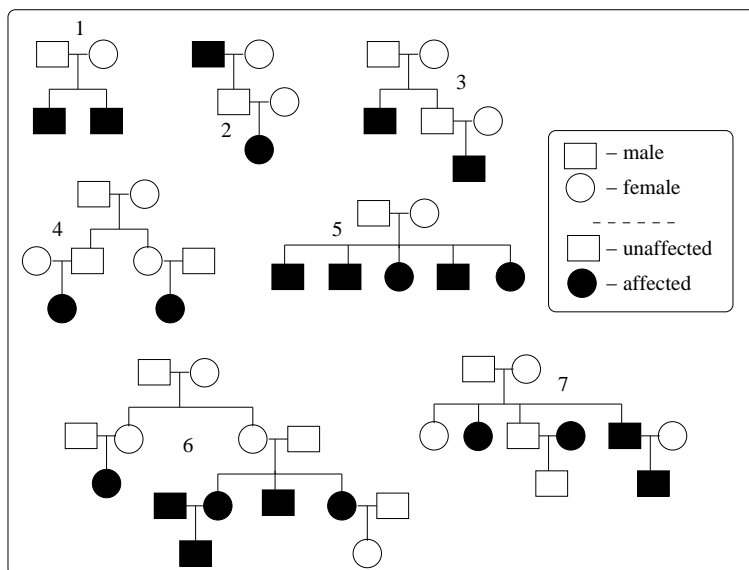


Figure 4: A pedigree set consisting of seven different pedigrees of varying pedigree structure and phenotypic configurations.

Each pedigree may be divided into subsets of *founders* and *nonfounders*, where the parents of the founders are not included in the pedigree, whereas both of the parents of a nonfounder are. The inheritance within a pedigree may be seen as the distribution of alleles from the founders to the corresponding nonfounders (descendants).

Family-based *gene mapping* is, explicitly or implicitly, based on allele-sharing between individuals in a pedigree. In this context, we say that: (i) Two individuals share an allele *identical-by-descent (IBD)* if they have both inherited exactly the same allele, i.e. an identical founder allele, from a common ancestor. (ii) Two individuals share an allele *identical-by-state (IBS)* if they have both inherited a common allelic variant.

Obviously IBD is a stronger sharing property than IBS, since sharing an allele IBD implies sharing IBS as well. Throughout this text we will exclusively be interested in test statistics based on IBD-sharing since they are

more efficient for testing genetic linkage. A simple example of allele-sharing IBD and IBS is given in Figure 5. Moreover, given the pedigree structure

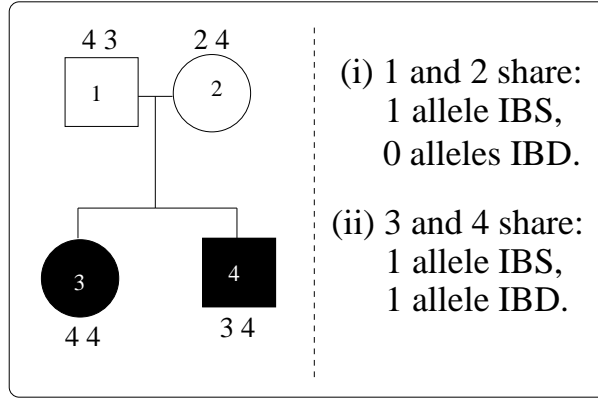


Figure 5: An allele-sharing example of an affected sib-pair (ASP) pedigree.

the IBD-sharing is completely explained by the corresponding inheritance vector, which will be introduced and explained in the next subsection.

Consider an *affected sib-pair (ASP)* pedigree, see Figure 5. In the one-locus case, the sib-pair may share either 0, 1 or 2 alleles IBD. This may be summarized in the *IBD-sharing vector*,

$$z = (z_0, z_1, z_2); \quad \text{where} \quad \sum_{i=0}^2 z_i = 1, \quad (6)$$

and z_i is the probability for the ASP to share i alleles IBD at the disease locus. In the two-locus case one may generalize (6) to the *IBD-sharing matrix*,

$$z = \begin{pmatrix} z_{00} & z_{01} & z_{02} \\ z_{10} & z_{11} & z_{12} \\ z_{20} & z_{21} & z_{22} \end{pmatrix}; \quad \text{where} \quad \sum_{i,j=0}^2 z_{ij} = 1, \quad (7)$$

and z_{ij} is the probability for the ASP of sharing i and j alleles at the 1st and 2nd disease locus respectively. The IBD-sharing vector (matrix) z depends on the genetic model λ and the locus (loci) at which IBD-sharing is (are) evaluated. Discussions on further constraints on z are given, for instance, by Holmans (1993), Dudoit and Speed (1998) and Bengtsson (2001). Such constraints are induced by sets of valid disease models. A classic reference with respect to one-locus IBD-sharing is Suarez (1978).

Example 6. Introduce $z' = (z'_0, z'_1, z'_2)$, where z'_i is the probability for an ASP of sharing i alleles IBS under the null hypothesis of no linkage (random inheritance), conditioned on the founder alleles. Now look at the three ASP pedigrees in Figure 6. From left to right, the IBS-sharing probabilities are $z' = (0.25, 0.5, 0.25)$, $z' = (0.125, 0.5, 0.375)$ and $z' = (0, 0, 1)$ respectively. Contrastingly, the corresponding IBD-sharing vector (6) equals $z = (0.25, 0.5, 0.25)$ for all three pedigrees.

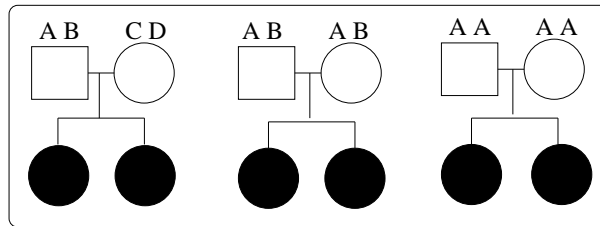


Figure 6: Three ASP pedigrees with different founder-allele configurations.

We end this subsection with a few words on a common abuse of notation. Generally one uses the notion allele for two distinct purposes: (i) For the *type* of allele, i.e. the actual allelic variant. Examples: A , B , 1 or 2. (ii) For the *origin* of the allele, i.e. the actual ancestral founder allele. Examples: 'The paternal allele of the third founder' or 'The fifth founder allele'. In this respect IBD and IBS analysis correspond to allele-sharing with respect to allelic origin and type respectively.

1.6 The Inheritance Vector and Entropy-Based Information Contents

Consider a pedigree consisting of n individuals, including f founders and $n-f$ nonfounders. There are $m = 2(n-f)$ meioses associated with the pedigree, since each nonfounder inherits one paternal and one maternal allele.

One of the core concepts in this thesis is the *inheritance vector* $v(x)$ (Donnelly, 1983). This binary 0-1 vector efficiently summarizes all the inheritance information at locus x for a single pedigree as,

$$v(x) = (p_1, m_1, p_2, m_2, \dots, p_{n-f}, m_{n-f}), \quad (8)$$

where p_i and m_i correspond to the i^{th} nonfounder's paternal and maternal allele respectively, i.e. each value is connected to a specific meiosis.

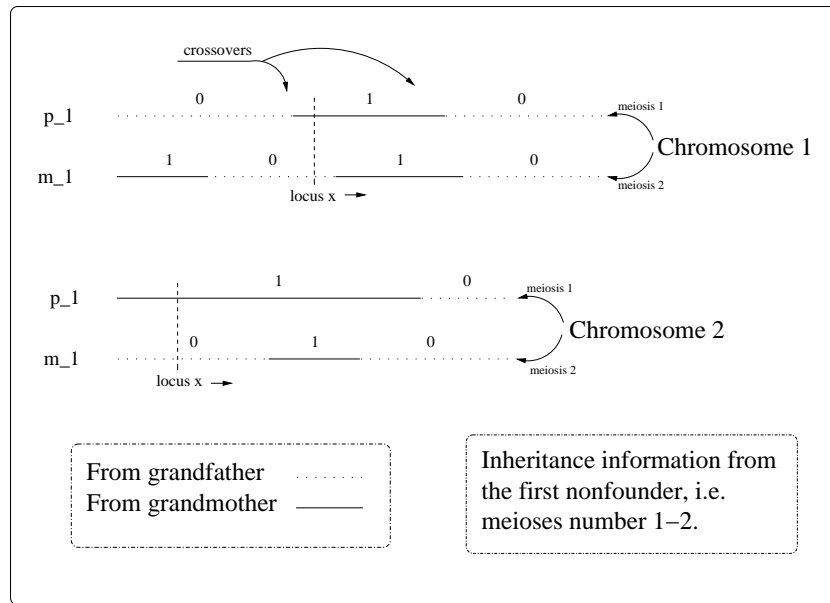


Figure 7: A schematic artificial inheritance vector example of two chromosomes and two meioses.

At a crossover point $v(x)$ will change since the corresponding meiosis, p_i or m_i for some i , switches between 0 and 1. We let 0 and 1 correspond to inheriting the grandpaternal and grandmaternal allele respectively. Given the inheritance vector and pedigree structure, the IBD sharing in the pedigree is unambiguous, i.e. known with probability one. A schematic overview is given in Figure 7 and a small pedigree example in Figure 8.¹⁶

Given data, measures on the locus-specific *information content* or *marker data information* may be based on the certainty of the outcome of $v(x)$. Here we will present the *entropy-based* information measure I_E of Kruglyak et al. (1996). For further discussion and suggestions of other information measures, see Teng and Siegmund (1998), Nicolae et al. (1998), Nicolae (1999) and Nicolae and Kong (2004).

¹⁶Simultaneously considering $|l|$ syntenic loci $\mathbf{x} = (x_1, x_2, \dots, x_{|l|})$, the complete inheritance picture is summarized by the $m \times |l|$ *inheritance matrix*,

$$v(\mathbf{x}) = [v(x_1)^T, v(x_2)^T, \dots, v(x_{|l|})^T],$$

where the i^{th} row and j^{th} column correspond to the equally indexed meiosis and inheritance vector (8) respectively.

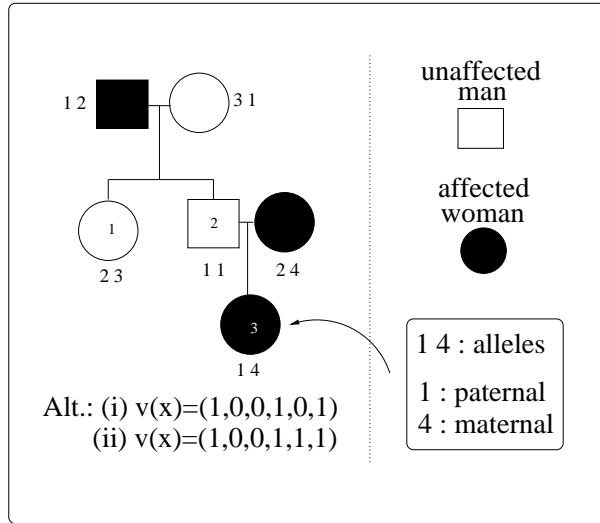


Figure 8: An inheritance vector example of a small three-generational pedigree. In the ideal case when the phase of all founders is known, only two different inheritance vectors are possible given marker genotype data.

Definition 7. Considering a discrete probability distribution based on $|y|$ distinct outcomes $y_1, y_2, \dots, y_{|y|}$ with probabilities $p_1, p_2, \dots, p_{|y|}$, where $p_i = P(y_i)$. The entropy E (Shannon, 1948; Kullback, 1968) with respect to this distribution is,

$$E = - \sum_{i=1}^{|y|} p_i \log_2 p_i. \quad (9)$$

The minimum value 0 of (9) is attained when the p_y -distribution is one-point and the maximum $\log_2(|y|)$ is attained when the distribution is uniform.

In our case, for a single pedigree, we face $m = 2(n - f)$ meioses and therefore $|y| = 2^m$ different valid outcomes of the corresponding inheritance vector, w_1, w_2, \dots, w_{2^m} , with probabilities,

$$p_i = P(v(x) = w_i | \text{MD}); \quad i = 1, 2, \dots, 2^m, \quad (10)$$

where MD is the marker data. The *entropy-based information content* is given by,

$$I_E(x) = \frac{E_0 - E(x)}{E_0 - E_{\min}}, \quad (11)$$

where $E(x)$ is the observed entropy at locus x , E_0 its maximal possible value m and E_{\min} its minimal value. When the phase of all founders is known we

put $E_{\min} = 0$. In general, however, the phase of all founders is unknown and $E_{\min} = f$, since switching of founder alleles results in 2^f equally likely inheritance vectors. In any case, $I_E(x)$ ranges from 0 (no marker information) to 1 (complete marker information).

1.7 How to Collect Information

There are two distinct ways of collecting or extracting the available inheritance information from the complete set of genotypes in the pedigree set: (i) *Single-point analysis*, where for each locus x one uses only marker genotypes at this locus when reconstructing the inheritance distribution (10). (ii) *Multipoint analysis*, where for each locus x one uses all marker genotypes from chromosome $c(x)$ when defining algorithms for computing the inheritance distribution (10).

Given a well-defined genetic marker map and assuming no interference, which leads to the Haldane's map function (1), one may use that, for each pedigree, the *inheritance process* $\{v(x); x \in \Omega\}$ over the genomic region Ω is a time-homogeneous Markov chain with state-space defined by the 2^m inheritance vectors. Given data, one calculates a multipoint-based inheritance distribution (10) using the theory of *hidden Markov models (HMM)* by interpreting marker genotypes and inheritance vectors as observed and hidden variables respectively. The transition matrices between markers may easily be derived assuming a Poisson distributed number of crossovers between consecutive markers.¹⁷

An original HMM-algorithm performing this task was presented by Lander and Green (1987). A detailed review was recently published in the textbook of Ziegler and Koenig (2006). Later, this algorithm has been updated with several speed-ups. Some extensions are described, for instance, by Kruglyak et al. (1995, 1996), Kruglyak and Lander (1998), Gudbjartsson et al. (2000) and Abecasis et al. (2002). General HMM-theory, implementations and applications are discussed, for example, by Rabiner (1989) and Cappé et al. (2005). An alternative algorithmic approach for computing the inheritance distribution (10) was introduced by Elston and Stewart (1971). The complexity of this algorithm increases only linearly with increasing pedigree size

¹⁷The relation between hidden and observed variables may be analyzed by checking the consistency between all possible founder genotype configurations and inheritance vectors, see Sobel and Lange (1996) and Kruglyak et al. (1996) for details.

but exponentially with the number of markers. The opposite is true for the Lander-Green algorithm.

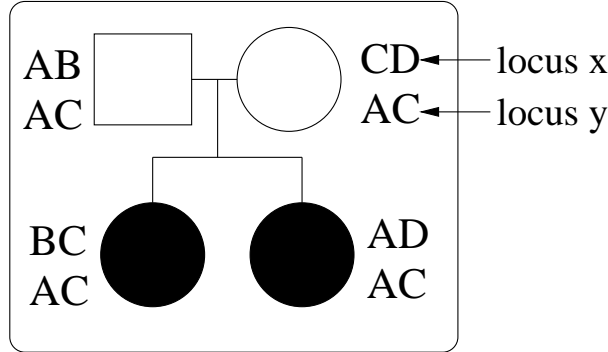


Figure 9: An ASP pedigree, being used in Example 8 for some simple multipoint calculations for loci x and y .

Example 8. Consider the ASP pedigree of Figure 9 and two loci x and y with recombination fraction θ between them. Assume that the phase of the parental genotypes is known, with the convention that the left haplotype is the paternal one. Now, with probability one $v(x) = (1, 0, 0, 1)$ and the possibilities of consistent inheritance vectors for $v(y)$ are,

$$(1, 0, 0, 1), (1, 0, 1, 0), (0, 1, 0, 1), (0, 1, 1, 0).$$

Using multipoint analysis the corresponding probabilities are,

$$(1 - \theta)^4, (1 - \theta)^2\theta^2, \theta^2(1 - \theta)^2, \theta^4,$$

whereas if using single-point analysis at y all four outcomes above are equally likely.

The multipoint approach increases the information content (11) and hence extracts available information more efficiently at the price of increased computational complexity.

2 Nonparametric Linkage Analysis

Linkage analysis aims at using statistical approaches to find locus (loci) involved in the genetic component of a disease under study. Qualitative-phenotype linkage analysis may be performed in two quite different ways,

the *parametric* and the *nonparametric* way, making different assumptions prior to the analysis. In this thesis we adopt the nonparametric approach.

The term nonparametric refers to the fact that no explicit assumptions on the underlying genetic model λ in (4) are made and corresponding nonparametric statistical tests usually depend on the concept of allele-sharing IBD. To perform a test, the actual sharing in the pedigree set is calculated and compared, through a properly chosen *test statistic*, with the expected sharing under the null hypothesis. In nonparametric linkage analysis *genetic linkage* at locus x means that the inheritance of alleles at x is correlated with the phenotype of interest.

This implies, for properly defined disease phenotypes and genetic models, that on average two affected individuals will share more alleles than is expected under the null hypothesis H_0 of no linkage. In the one-locus case one may state the pair of *tested hypotheses* as,

$$\begin{cases} H_0 : \text{Disease locus unlinked to } x. \\ H_1 : \text{A disease locus at } x. \end{cases} \quad (12)$$

when testing a single locus and as,

$$\begin{cases} H_0 : \text{No disease locus linked to } \Omega. \\ H_1 : \text{At least one disease locus along } \Omega. \end{cases} \quad (13)$$

when testing a whole region Ω .

2.1 Parametric Linkage Analysis

Here one assumes a known (or rather estimated) genetic model λ in (4) which must be defined prior to the analysis. The most common test statistic is the likelihood ratio-based *lod-score statistic*, which for the single-point case is defined as,

$$Z(\theta; \lambda) = \log_{10} \left[\frac{P(Y, \text{MD}(x) \mid \theta, \lambda)}{P(Y, \text{MD}(x) \mid 0.5, \lambda)} \right],$$

where Y and $\text{MD}(x)$ refer to phenotypic data and marker data at locus x respectively, θ is the tested recombination fraction between the disease locus and x , and λ is the genetic model. In this case we use $Z(0; \lambda)$ as test statistic for (12) and the maximum lod-score $Z_{\max} = \sup_{0 \leq \theta \leq 0.5} Z(\theta; \lambda)$ as test statistic

for (13) with $\Omega = c(x)$, which can be written as,

$$\begin{cases} H_0 : \theta = 0.5 \\ H_1 : 0 \leq \theta < 0.5 \end{cases}$$

For the multipoint case the lod score statistic is only slightly altered into the form,¹⁸

$$Z(x; \lambda) = \log_{10} \left[\frac{P(Y, \text{MD} \mid x, \lambda)}{P(Y, \text{MD} \mid \infty, \lambda)} \right], \quad (14)$$

where MD is the complete set of marker data from Ω and x is the hypothesized position of the disease locus l . We use $Z(x; \lambda)$ as test statistic for (12) and the maximum lod-score $Z_{\max} = \sup_{x \in \Omega} Z(x; \lambda)$ as test statistic for (13).

The original lod-score reference is Morton (1955) which is based on results and procedures given by Haldane and Smith (1947) and Barnard (1949). Some modern references are Terwilliger and Ott (1994), Kruglyak et al. (1996), Ott (1999), Kurbasic and Hössjer (2004, 2006) and Xing and Elston (2006).

2.2 Score Functions

For the one-locus case, the nonparametric test statistic is based on a score function $S(v)$, which assigns a number to each possible pedigree-wise IBD-sharing structure (or inheritance vector).¹⁹ As noted above, one is normally interested in increased allele-sharing among affecteds since this indicates presence of genetic linkage between the marker and disease loci.

The relative performance of different score functions, in terms of statistical power, depends on the underlying genetic model and the structure of the pedigree set.

Two commonly used score functions were introduced by Whittemore and Halpern (1994). Firstly, S_{pairs} is based on IBD-sharing among all pairs of affected individuals in the pedigree,

$$S_{\text{pairs}}(v) = \sum_{(i,j) \in A} \text{IBD}(i, j),$$

¹⁸An explanation to the ∞ -sign in the denominator of (14) is that, under the null hypothesis, the disease locus is unlinked to all chromosomes constituting Ω .

¹⁹One may note that this notion of a score function may be seen as adopting a data-mining perspective where such functions are used for scoring patterns, in this case inheritance patterns (Hand et al., 2001).

where $i < j$, \mathbb{A} is the set of affecteds and $\text{IBD}(i, j)$ is the number of alleles shared IBD between individuals i and j .

Secondly, S_{all} is based on the simultaneous IBD-sharing among all the affecteds in the pedigree,

$$S_{\text{all}}(v) = \frac{1}{2^{|\mathbb{A}|}} \sum_{h \in \mathbb{H}} \prod_{i=1}^{2f} b_i(h)!,$$

where $|\mathbb{A}|$ is the number of affecteds, \mathbb{H} is a set containing all ways of selecting one allele from each affected, $2f$ is the number of founder alleles in the pedigree and $b_i(h)$ is the number of times the i^{th} founder allele is present in selection h .²⁰

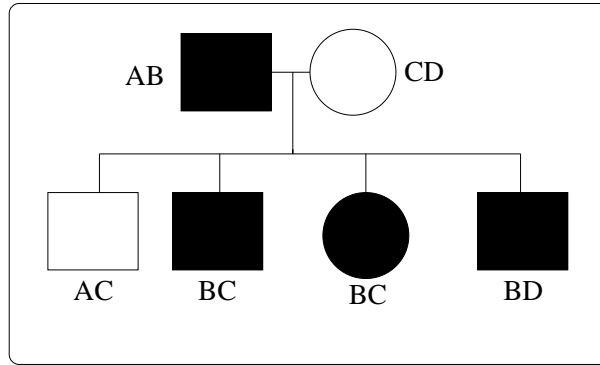


Figure 10: An example pedigree used in Example 9 regarding the comparison of different score functions.

Example 9. Consider the pedigree in Figure 10. Ordering the nonfounders from left to right the inheritance vector $v = (0, 0, 1, 0, 1, 0, 1, 1)$ is fully known if both parents have known phase. Calculating S_{pairs} and S_{all} we get,

$$S_{\text{pairs}}(v) = (1+1+1+2+1+1) = 7,$$

$$S_{\text{all}}(v) = (2+4+6+6+6+6+4+2+2+2+2+2+2+2+2+1+1)/16 = 35/8.$$

For instance, for S_{all} , if $h = \{B, C, C, B\}$ we get $\prod_{i=1}^4 b_i(h) = 0!2!2!0! = 4$.²¹

²⁰The selection h consists of $|\mathbb{A}|$ alleles that may be grouped according to their ancestral history, i.e. each allele is a copy of one of the $2f$ founder alleles. The link to the number of members in the i^{th} group is $b_i(h)$.

²¹Both S_{pairs} and S_{all} are calculated, given v , using the group of affecteds \mathbb{A} only. Each

Given a well-defined genetic model λ in (4) it is possible to derive different kinds of *optimal* score functions, based on different optimality criteria (McPeck, 1999; Hössjer, 2003b, 2005c), which then implicitly leads to the use of both affecteds and unaffecteds through their corresponding definitions. In applications λ is most oftenly not fully known leading to extensive usage of score functions that may be designed to have good performance for a (sufficiently) wide range of genetic models.

2.3 The NPL Score

To ease interpretation and significance calculations we *standardize*²² the score function under H_0 according to,

$$S(v) \leftarrow \frac{S(v) - \mu}{\sigma}, \quad (15)$$

where, for a pedigree with m meioses,

$$\begin{cases} \mu = \sum_i 2^{-m} S(w_i) \\ \sigma^2 = \sum_i 2^{-m} S(w_i)^2 - \mu^2 \end{cases}$$

are the mean and variance of S , prior to standardization, under the null hypothesis H_0 of no linkage.

The standardized score function is used to calculate, at locus x , the *pedigree-specific nonparametric linkage (NPL) score*,

$$Z(x) = \sum_i p(w_i) S(w_i), \quad (16)$$

where $p(w_i) = P(v(x) = w_i | \text{MD})$ is the inheritance distribution at x .

such *traditional* score function S corresponds to an *extended* score function S' ,

$$S'(v) = S'(v | \mathbb{A} \cup \mathbb{UA}) = S(v | \mathbb{A}) + S(v | \mathbb{UA}),$$

where \mathbb{UA} is the set of unaffecteds and $S(\cdot | \mathbb{B})$ is the traditional score function replacing the subgroup of affecteds with the arbitrary subgroup \mathbb{B} . This may increase power since we extract more inheritance information, but the computational complexity may be alarmingly increased and the gain small for many genetic disease models. For more information, and additional extended versions, see Ängquist (2006).

²²Also referred to as *normalization*. Note that we end up with the standardized properties $E(S|H_0) = 0$ and $V(S|H_0) = 1$.

For a pedigree set consisting of N pedigrees we combine the pedigree-specific scores (16) into the (total) *NPL score* as,

$$Z(x) = \sum_{k=1}^N \gamma_k Z_k(x), \quad (17)$$

where Z_k is the NPL score (16) and γ_k the weight assigned to the k^{th} pedigree. The pedigree *weighting scheme* is chosen as $\sum_{k=1}^N \gamma_k^2 = 1$ in order to assure,

$$E(Z(x)|H_0) = 0 \text{ and } V(Z(x)|H_0) \leq 1, \quad (18)$$

with equality for complete marker information.²³ The actual weights may be chosen according to pedigree size, structure, information or inheritance at other loci. We use $Z(x)$ as test statistic for testing a single disease locus through (12). Oftenly, such tests are based on a *perfect-data approximation*, which corresponds to calculating the null distribution assuming that $V(Z(x)|H_0) = 1$ and implies *conservative* p -value estimates when facing imperfect data.²⁴ An alternative definition of total NPL score was proposed by Kong and Cox (1997), which leads to less conservative tests for incomplete marker data.

Calculating the NPL score with respect to a set of loci leads to a stochastic process $\{Z(x); x \in \Omega\}$, the *NPL process*. A simulated example is given in Figure 11. According to the blockwise inheritance of chromosomal segments, NPL scores at linked loci are correlated. The multipoint approach increases the NPL score correlation at closely linked loci, as an effect of smoothing the observed NPL score process.

To test (13) we use the maximum of the NPL process over Ω , referred to as the maximum NPL score $Z_{\max} = \sup_{x \in \Omega} Z(x)$.

²³The inequality in (18) follows for one pedigree from,

$$\begin{aligned} 1 &= V(S[v(x)]) = V[E(S[v(x)]|\text{MD})] + E[V(S[v(x)]|\text{MD})] \\ &\geq V[E(S[v(x)]|\text{MD})] = V[Z(x)], \end{aligned}$$

assuming expectation and variance is taken under H_0 . See also Kruglyak et al. (1996).

²⁴Missing genotypes, homozygosity, usage of single-point analysis or multipoint analysis with a sparse marker map leads to loss of inheritance information, i.e. increased inheritance vector ambiguity, implying a decrease of the NPL score variance. The extreme case of totally uninformative marker data for a pedigree leads to a constant score $Z(x) = 0$ in (16). If the underlying combination of pedigree structure, score function and phenotypic configuration is totally uninformative, the unstandardized scores $S(v)$ are independent of v which, using (15), implies a not even well-defined procedure.

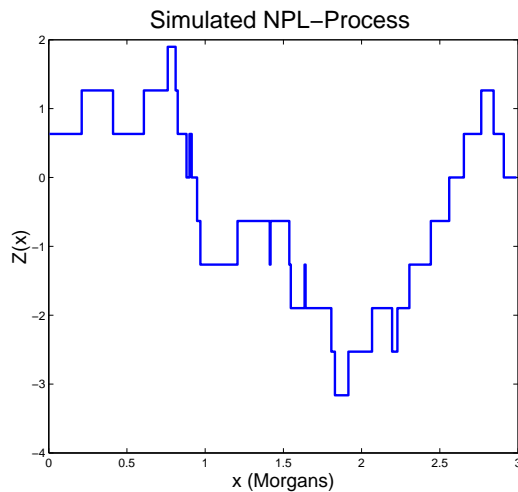


Figure 11: A simulated NPL process along a single chromosome of length 3 Morgans, assuming perfect marker data. The underlying score function is S_{all} , the pedigree set consists of $N = 10$ (homogeneous) pedigrees of the same structure as in Figures 1-2 and 8 with equal weights $\gamma_k = 1/\sqrt{10}$.

2.4 Calculating the Statistical Significance

Consider a test statistic Z , which is either a *pointwise* ($Z = Z(x)$) or *genomewide*²⁵ ($Z = Z_{\text{max}}$) NPL score. The *significance level* of a test which rejects H_0 when $Z \geq T$, where T is a given score threshold, is,

$$\alpha(T) = P(Z \geq T | H_0), \quad (19)$$

when the null hypothesis is simple. The *power function*,

$$\beta(T) = P(Z \geq T | H_1), \quad (20)$$

depends on the genetic model λ . Given a test result $Z = z$, one may calculate the *p-value* $\alpha(z)$.

In the pointwise case, when N is large, one may use the approximation,

$$Z(x) \stackrel{H_0}{\in} N(0, 1), \quad (21)$$

of (17), based on the *Central Limit Theorem*. This leads to the approximation,

$$\alpha(T) \approx 1 - \Phi(T),$$

²⁵That is, if Ω is the whole genome.

where Φ equals the standard normal $N(0,1)$ distribution function. This approximation is usually conservative, due to (18), when marker data is incomplete.

In the genomewide case, the distribution of Z_{\max} is less tractable. It may be computed using either *Monte Carlo simulation* or *analytical approximations*. In the latter case one uses *extreme-value theory* of the stochastic process $\{Z(x); x \in \Omega\}$, see e.g. Leadbetter et al. (1983), Siegmund (1985), Aldous (1989), Lander and Botstein (1989), Feingold (1993), Feingold et al. (1993) and Tu and Siegmund (1999). A small-scale example of the genomewide significance level (19) is given in Figure 12.

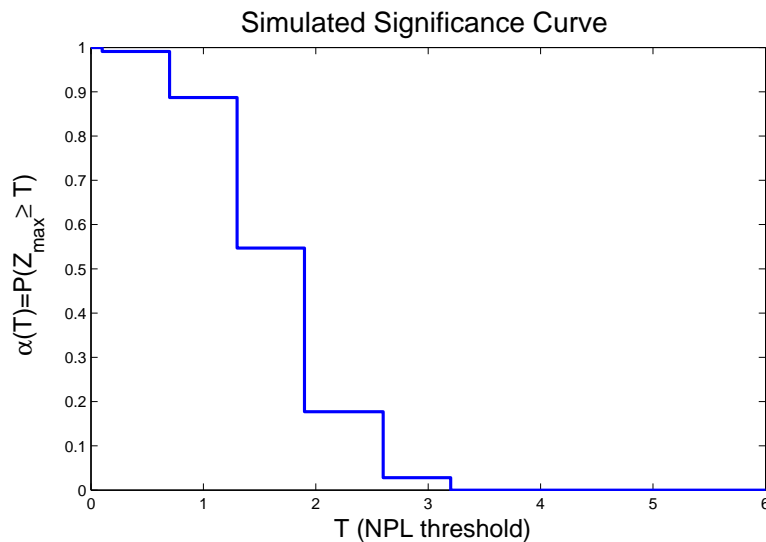


Figure 12: A simulated genomewide significance level curve for a single chromosome of length 3 Morgans, assuming perfect marker data. The score function and pedigree set are as in Figure 11.

The advantage of a closed form analytical approximation is ease of computation, even for large thresholds. On the other hand, the approximation may be more or less accurate according to the assumptions made. It is common to assume:

One That the NPL score (17) is marginally normally distributed, with expected value equal to zero, for all loci under H_0 . Deviations from this assumption will, depending on the type of nonnormality, imply either conservative or anticonservative approximations.

Two That the variance of the NPL score in (17) equals one at all loci, i.e. to use the perfect-data approximation. This implies conservativity according to (18).

Three That Ω is continuous, i.e. an infinitely dense marker map is used. This implies conservativity.

The Monte Carlo approximation of Section 2.6 is more accurate provided the number of replicates is large enough. On the other hand, computationally it can be very slow, especially for large thresholds.

2.5 Significance Calculations through Theoretical Approximation

The most commonly used theoretical approximation (Lander and Botstein, 1989; Feingold et al., 1993; Lander and Kruglyak, 1995) is based on the assumption that the NPL process is a stationary continuous-time Gaussian process with $N(0,1)$ marginals. Moreover, the process is assumed to be *Ornstein-Uhlenbeck*-like (Uhlenbeck and Ornstein, 1930; Hsu and Park, 1985; Blackwell, 2002). By this we mean that the process is non-differentiable and the autocovariance function,

$$r_Z(h) = C [Z(x), Z(x+h)],$$

has a non-zero right-hand side derivative $r'_Z(0)$ at zero.

Following Lander and Kruglyak (1995), the approximation $\hat{\alpha}(T)$ of the genomewide significance level (19) is defined as

$$\hat{\alpha}(T) = 1 - \exp[-\mu(T)], \quad (22)$$

where

$$\mu(T) = [C + 2\rho gT^2]\alpha_{\text{pt}}(T). \quad (23)$$

In (22)-(23) we have that $C = C(\Omega)$ is the number of chromosomes in Ω , $g = g(\Omega)$ is the total genome length of Ω in Morgans, $\rho = -r'_Z(0)/2$ is the crossover rate and $\alpha_{\text{pt}}(T) = 1 - \Phi(T)$ the approximative pointwise significance level with respect to the threshold T .

Extensions of (22) are given by Tang and Siegmund (2001) where *correction* for nonnormality, in the form of *distributional skewness*, of the NPL

score is introduced. Further, based on works of Siegmund (1985) and Feingold et al. (1993) a relaxation of the assumption of a continuous-time process into a finite number of *equidistant markers* is made.

The *crossover rate* ρ in (23) reflects the fluctuation of the NPL process, i.e. how often the score changes and how large these changes are. It may be expressed for arbitrary pedigrees (Hössjer, 2001, 2003a; Ängquist, 2001) as,

$$\rho = \frac{1}{4} 2^{-m} \sum_{w \in \mathbb{V}} \sum_{j=1}^m [S(w) - S(w + e_j)]^2, \quad (24)$$

where S is a standardized score function (15), \mathbb{V} is the set of possible inheritance vectors and e_j is an inheritance vector with one in the j^{th} position and zeros elsewhere, corresponding to a crossover of the j^{th} meiosis. In this sense, $\{w + e_j\}_{j=1}^m$ are *neighbours* of w .²⁶ Hence ρ depends on the score function, pedigree structure and pedigree size.

For a pedigree set consisting of N pedigrees one may combine the pedigree-specific crossover rates (24) into an overall crossover rate,

$$\rho = \sum_{k=1}^N \gamma_k^2 \rho_k,$$

where γ_k and ρ_k are the k^{th} pedigree weight in (17) and crossover rate respectively.

Example 10. Consider an ASP pedigree. If using the standardized version (15) of a symmetric score function,²⁷ $S(w)$ attains the value $-\sqrt{2}$, 0, $\sqrt{2}$ when the ASP shares 0, 1 or 2 alleles IBD.

In this case the number of meioses $m = 4$, the number of possible inheritance vectors $|\mathbb{V}| = 16$. The IBD-status of the ASP along a chromosome changes at points of crossovers according to a Markov chain with transition matrix,

$$P = \{p_{ij}\}_{i,j=0}^2 = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix},$$

²⁶A neighbour w' to w is an inheritance vector that differs from w at only one vector position, i.e. the corresponding *Hamming distance* $H(w', w) = 1$.

²⁷Let s_i be the unstandardized score corresponding to the ASP sharing i alleles IBD. The score function is symmetric if $s_2 - s_1$ equals $s_1 - s_0$. This is true both for S_{pairs} and S_{all} .

where p_{ij} refers to the probability of changing from i to j alleles shared IBD. Hence in all cases, $w \in \mathbb{V}$ and $k \in \{1, 2, \dots, m\}$, $|S(w) - S(w + e_k)| = \sqrt{2}$. Using (24), we obtain $\rho = \rho(\text{ASP}) = 2$.

2.6 Significance Calculations through Monte Carlo Simulation

Using a simulation algorithm is often the easiest way to capture hard-to-get information, by being well-suited to mimic complicated models. The main drawback usually is the computational complexity.

For complete marker data, the pedigree-specific NPL score (16) simplifies to,

$$Z(x) = S[v(x)], \quad (25)$$

where $v(x) = [v_1(x), v_2(x), \dots, v_m(x)]$ is the inheritance vector at locus x . When map distance is measured in Morgans and no chiasma interference is assumed, the components $v_j(\cdot)$ can be simulated under H_0 as independent stationary Markov processes with intensity matrix

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \quad (26)$$

along each chromosome of Ω . From (25) we then get the pedigree-specific NPL score along Ω . Repeating this for all N pedigrees, using (17) and maximizing over Ω , we obtain a maximum NPL score Z_{\max} simulated under H_0 . Further, repeating this for J simulations, with $Z_{\max,j}$ denoting the maximum NPL score for the j^{th} simulation, the crude Monte Carlo estimate of the significance level is,

$$\hat{\alpha}(T) = \frac{1}{J} \sum_{j=1}^J I(Z_{\max,j} \geq T); \quad T \in \mathbb{T}, \quad (27)$$

where $I(A)$ is the indicator function of the event A and \mathbb{T} is a predefined set of score thresholds.

Example 11. Consider a single ASP. Figure 13 illustrates a simulated realization of the inheritance process along a chromosome of length 2 Morgans and Figure 14 shows the corresponding pedigree-specific NPL score process.

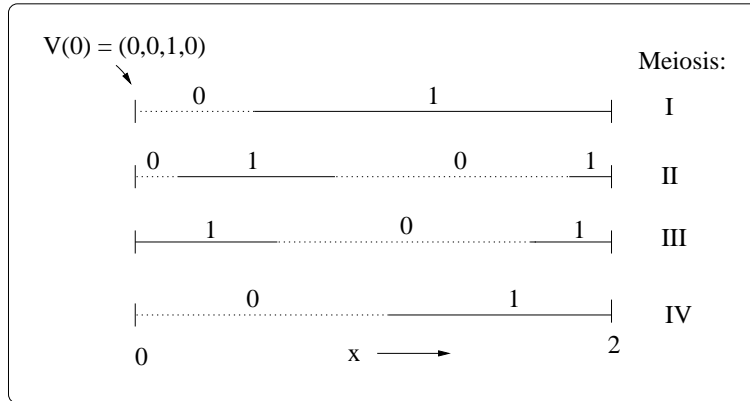


Figure 13: An example illustrating the simulation of the inheritance vector process $v(\cdot)$ of a single ASP along a chromosome of length 2 Morgans.

To use simulation to facilitate calculation of statistical power with respect to the appropriate test statistic, pointwise or genomewide, under a genetic model λ of the alternative hypothesis H_1 and the phenotype vector Y , the H_0 -simulation may be modified as follows for the inheritance process of each pedigree: If l is the disease locus, simulate $v(l)$ from $P(v|\lambda, Y)$. Then generate the components $v_j(\cdot)$ of $v(\cdot)$ independently to the right and left of l , starting at $v_j(l)$, according to the same Markov process (26) as in the H_0 -simulation.

Some references to simulation procedures for incomplete marker data are Boehnke (1986), Ploughman and Boehnke (1989), Ott (1989) and Terwilliger et al. (1993).

2.7 Two-Locus NPL Analysis

One may generalize the NPL procedure above in order to simultaneously, or sequentially, search for two distinct disease loci on Ω . We will now briefly outline, and make some comments on, such procedures.

To be able to perform a so called *unconditional two-locus NPL analysis*, we generalize the pedigree-specific NPL score in (16) to,

$$Z(x_1, x_2) = \sum_{i,j} p(w_i, w_j) S(w_i, w_j), \quad (28)$$

where $p(w_i, w_j) = P(v(x_1) = w_i, v(x_2) = w_j | \text{MD})$ is the joint inheritance

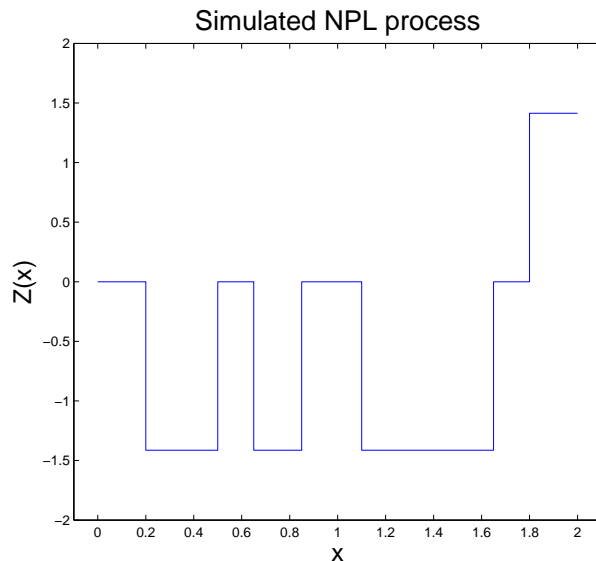


Figure 14: The pedigree-specific NPL score process $Z(x)$ corresponding to the simulated inheritance process of Figure 13.

distribution at loci x_1 and x_2 , see Strauch et al. (2000). We assume that x_1 and x_2 are unlinked²⁸ in (28) but otherwise varied independently. This implies allele-sharing independence at the two loci,

$$p(w_i, w_j) = P(v(x_1) = w_i | \text{MD}) P(v(x_2) = w_j | \text{MD}),$$

and one-locus multipoint algorithms can be used for calculating (28). The (total) NPL score is then calculated as,

$$Z(x_1, x_2) = \sum_{k=1}^N \gamma_k Z_k(x_1, x_2), \quad (29)$$

where $Z_k(x_1, x_2)$ is the pedigree-specific NPL score (28) and γ_k the weight assigned to the k^{th} pedigree.

Equation (28) involves a standardized two-locus score function $S(v_1, v_2)$, depending on the pair of inheritance vectors $(v_1, v_2) \in \mathbb{V} \times \mathbb{V}$. Two simple examples of such score functions are: (i) The *additive* two-locus score function,

$$S(v_1, v_2) = \frac{S(v_1) + S(v_2)}{\sqrt{2}},$$

²⁸Formally, $c(x_1) \neq c(x_2)$.

which essentially is the sum of the corresponding standardized one-locus score functions. (ii) The *multiplicative* two-locus score function,

$$S(v_1, v_2) = S(v_1)S(v_2),$$

which is the product of the corresponding standardized one-locus score functions.

Restricting the two-locus analysis by letting the locus x_2 in (28) be fixed and conditioning on information at this locus yields a *conditional* two-locus NPL analysis. The *conditioning locus* x_2 may be a verified disease locus, $x_2 = l_2$, a suggested or estimated disease locus, $x_2 = \hat{l}_2$ or a locus considered to be interesting for some other reason. The pedigree-specific type of information to condition on may be one-locus inheritance vectors or NPL scores.

The most well-known example of this kind is the conditional multiplicative two-locus pedigree-specific NPL score introduced by Cox et al. (1999). It may be described through the one-locus pedigree-specific NPL scores in (16) as,

$$Z(x_1, x_2) = Z(x_1)f[Z(x_2)], \quad (30)$$

where $f(\cdot)$ is a given function of the pedigree-specific NPL score at the conditioning locus x_2 . When computing the total two-locus NPL score (29) based on pedigree-scores (30), one must replace the unconditional two-locus constraint $\sum_{k=1}^N \gamma_k^2 = 1$ on the pedigree weights with,²⁹

$$\sum_{k=1}^N \gamma_k^2 f[Z_k(x_2)]^2 = 1.$$

Using (30) is basically a way to utilize gene-gene interaction to increase power. Different choices of the weighting function f is suitable for different types of correlation. We refer to *positive* and *negative* interaction (correlation) as *epistasis* and *heterogeneity* respectively, see Cox et al. (1999) and Holmans (2002).

When performing a conditional two-locus NPL analysis one may use a combination of: (i) Several conditioning loci. (ii) Several types of weighting functions. (iii) Several distinct score functions. The use of (i)-(iii) above

²⁹Hence, a conditional two-locus analysis (29)-(30) may be described within the framework of one-locus analysis by interpreting the product $\gamma_k f[Z_k(x_2)]$ as the k^{th} pedigree weight, assigned to $Z_k(x_1)$.

implies the need for *multiple testing correction*. Since the individual tests generally are dependent, the standard *Bonferroni correction* may yield a very conservative upper bound on the *familywise error rate (FWE)*³⁰ under H_0 , i.e. the probability that at least one individual test is declared significant.³¹

Many refinements of the Bonferroni upper bound of the FWE have been proposed, including the sequential procedures of Hochberg (1988) and Benjamini and Hochberg (1995). The latter one controls the *false discovery rate (FDR)*, see Ge et al. (2003) for a recent review.

3 Other Statistical Genetics Procedures

In this section we will briefly comment on some related and complementary analysis procedures in the context of gene mapping.

Prior to linkage analysis, or in its own right, one may perform a *segregation analysis*. This is done in order to determine or estimate genetic model parameters, as well as environmental factors, using phenotype data from pedigrees with many affecteds. For further details see e.g. Khoury et al. (1993) and Haines and Pericak-Vance (2006).

Association analysis may be used for fine-mapping regions pinpointed by an initial linkage analysis or directly for genomewide mapping (Risch and Merikangas, 1996). It aims at finding allelic variants associated with disease. Further, one may split association analysis into *population-based* (Clayton, 2001; Balding, 2006)³² and *family-based* (Terwilliger and Ott, 1992; Spielman et al., 1993; Zhao, 2000) procedures. See also, for instance, Cordell and Clayton (2002, 2005).

Alleles in close physical proximity of the disease locus show association with the disease-causing locus because of linkage disequilibrium. However, association between more distant (and even unlinked) loci may also exist because of *mixture of populations*. The latter source of association is a con-

³⁰This is also referred to as the *global* multiple testing significance level (Hougaard, 2006).

³¹Assume a combined testing-procedure based on n tests with corresponding test statistics T_1, T_2, \dots, T_n and critical regions C_1, C_2, \dots, C_n . Simple manipulations now give $P(\exists i : T_i \in C_i) \leq \sum_{j=1}^n P(T_j \in C_j)$. Taking advantage of this inequality, when testing on a global significance level α , the Bonferroni-procedure now perform each individual test on significance level α/n . If the number of tests is large, especially if many test statistics are positively dependent, this is a severe testing problem with respect to power.

³²Including the biostatistical approach of *case-control studies* which often is considered within the field of *epidemiology* (Clayton and Hills, 1993).

founder in gene mapping of disease loci.

In the population-based procedure, one searches for association between phenotypes and allelic variants of single individuals. The method is powerful but sensitive to population admixture. In the family-based procedure, association between phenotypes and transmission of allelic variants from heterozygous parents is of interest. The method is less powerful but robust towards association due to population admixture. These two procedures can also be combined, see e.g. Clayton (1999) and Shih and Whittemore (2002).

Whereas linkage analysis is based on inheritance of alleles within pedigrees, it does not utilize association between allelic variants and phenotypes on the population level. However, the two approaches can be brought together into joint tests for linkage and association, see for example Fulker et al. (1999), Xiong and Jin (2000), Göring and Terwilliger (2000), Sham et al. (2000) and Hössjer (2005a).³³

Using a quantitative phenotype one may perform a *quantitative trait loci (QTL)* analysis, see Lynch and Walsh (1998). Several different subfields of this approach exist. The traditional *Hase-man-Elston* approach (Hase-man and Elston, 1972) is based on a *regression* of the squared phenotype-difference between sibs with respect to their corresponding allele-sharing and the second, *variance component*, approach is based on splitting up the total phenotype-variation into genetic and environmental factors (Almasy and Blangero, 1998; Cherny et al., 2004). A third approach is based on *IBD-sharing* conditional on observed phenotypes, either by means of a regression model (Sham et al., 2002) or a likelihood score statistic (Tang and Siegmund, 2001; Hössjer, 2005b). For a recent unification of all three types of methods, by means of *generalized estimating equations*, see Chen et al. (2004).

Some general references to linkage analysis, and in a wider sense gene mapping, monographs are Sham (1998), Ott (1999), Almgren et al. (2003), Thomas (2004), Ziegler and Koenig (2006), and Haines and Pericak-Vance (2006). Inference on genetic data is outlined in Thompson (2000).

³³One may note the following: Assume a disease phenotype caused by a *recent* genetic mutation at locus l and, further, a marker locus m in close vicinity of, but not equivalent to, l . Now, with respect to future generations, m will always be linked to l , but eventually the originally corresponding allelic association will *fade away* according to the recombination process.

4 Outline of Papers in Thesis

In this section we briefly outline, through comments and examples, some of the methods and extensions constituting the contents of the four included papers of the thesis.

4.1 Paper A

The approximation of the significance level (22)-(23) is extended to *correct for marginal nonnormality*, under an assumption of fully informative inheritance at all markers, of the NPL score (17). Moreover, the *discreteness correction*, assuming a set of equidistant markers, is incorporated as well.

The nonnormality correction is based on introducing a *link function*, g_{link} , which transforms the NPL process to marginal approximate-normality. Formally,

$$Y(x) = g_{\text{link}}^{-1} [Z(x)]; \quad x \in \Omega, \quad (31)$$

where $g_{\text{link}} = (F^{-1} \circ \Phi)$, $F(z) = P(Z(x) \leq z | H_0)$ is the marginal distribution function of the NPL score process $Z(\cdot)$ under H_0 and $\Phi(z) = \int_{-\infty}^z \phi(z) dz$ is the standard normal distribution function. The transformed process Y in (31) is a stationary process with approximately standard normal marginals, the approximation being due to discreteness of F .

To turn Y into a *continuous-valued* process we use a *linear binning* smoothing procedure to approximate F by a continuous version \hat{F} when defining g_{link} .³⁴ Noting that,

$$\alpha(T) = P(Z_{\max} \geq T | H_0) = P(Y_{\max} \geq g_{\text{link}}^{-1}(T) | H_0),$$

leads to an improved significance approximation of $\alpha(T)$. It is based on (22), but with a refined upcrossing intensity (23) expressed as,

$$\mu(T) = [C + 2\rho_Y g g_{\text{link}}^{-1}(T)^2] \alpha_{\text{pt}} [g_{\text{link}}^{-1}(T)],$$

where ρ_Y is the updated crossover rate with respect to the transformed process Y and g is the genome length. The calculation of ρ_Y is based on the theory of *subordinated* Gaussian processes (Clark, 1973) and a *Hermite polynomial expansion* of g_{link} (Taqqu, 1975). An example of the link function g_{link} is given in Figure 15.

³⁴This transforms g_{link} from being a *nondecreasing step-function* to a *strictly increasing* function.

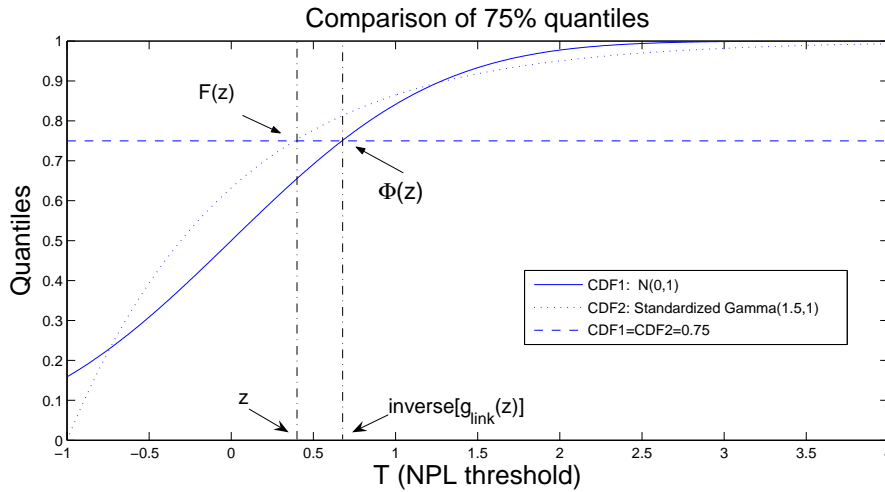


Figure 15: Implicit construction of the link $z \leftrightarrow g_{\text{link}}^{-1}(z)$ in (31) through comparison between the NPL score distribution function F and the standard normal distribution function Φ . Here the unstandardized NPL score is assumed to be distributed as $\Gamma(1.5, 1)$.

Further, extensive simulations and calculations with respect to several pedigree sets are processed and performance comparisons with existing approximation formulas is laid out. In addition, comparison of the original and improved significance level approximation in terms of conservativeness is facilitated using *Edgeworth expansion* approximations of F .

4.2 Paper B

For large thresholds T where $\alpha(T) = P(Z_{\max} \geq T | H_0)$ is very small, we will generally need a very large number of simulations to be able to estimate $\alpha(T)$ using (27) with reasonably small variance. Given a constant time limit the actual number of performed simulations also depends on the computer-time per simulation or, equivalently, the *computational cost*, which is primarily affected by the pedigree sizes. One possible solution to this problem is to use *importance sampling*, or *weighted simulation*, which is a variance reduction technique making it possible to sample from interesting regions of the underlying sample space with higher probability. An early reference is Hammersley and Handscomb (1964) and a modern introduction is given by Ross (2006). See also Hesterberg (1995).

A very brief outline of the general method is as follows. Assume we want to estimate,

$$\alpha = E[f(Z)] = \int f(Z)dP(Z),$$

where E denotes expectation when Z has distribution P . The following reformulation, using the *change of probability measure* from P to \tilde{P} ,

$$\alpha = \int f(Z) \frac{dP(Z)}{d\tilde{P}(Z)} d\tilde{P}(Z) = \tilde{E}[f(Z)L(Z)], \quad (32)$$

is valid if,

$$f(Z)dP(Z) > 0 \Rightarrow d\tilde{P}(Z) > 0.$$

In (32), \tilde{E} denotes expectation when Z has distribution \tilde{P} and the weighting function $L(Z) = dP(Z)/d\tilde{P}(Z)$ is the likelihood ratio with respect to the two measures. Sampling from \tilde{P} , the integral in (32) may be estimated using,

$$\tilde{\alpha} = \frac{1}{J} \sum_{j=1}^J f(Z_j)L(Z_j),$$

where $\{Z_j\}_{j=1}^J$ are independent and identically distributed copies of Z under \tilde{P} .

In our case $\alpha = \alpha(T)$, $Z = \{Z(x); x \in \Omega\}$, $f(Z) = I(Z_{\max} \geq T)$ and P is the H_0 -distribution of Z . We introduce an *exponentially tilted* probability measure \tilde{P} , which for complete marker data along one chromosome $c = [0, l]$ has the form,

$$d\tilde{P}(Z) = \left(\frac{\int_0^l \exp[\delta Z(x)] dx}{lM(\delta)} \right) dP(Z), \quad (33)$$

where $M(\delta) = E(\exp[\delta Z(X)] | H_0)$ is the moment generating function of $Z(x)$ under H_0 and δ is a *design* or *tilting* parameter reflecting the amount of change of measure.

When $\delta = 0$, $\tilde{P} = P$ coincides with the H_0 -distribution. The larger $\delta > 0$ is, the more likely it is for Z_{\max} to attain large values when $Z \sim \tilde{P}$. To actually simulate such a Z , we proceed similarly as when generating NPL scores under H_1 to estimate power in Section 2.6. First select an *artificial disease locus* x_0 according to a uniform distribution on $[0, l]$, then generate $Z(x_0)$ by simulating inheritance vectors at x_0 of all pedigrees under a pointwise version of the exponentially tilted distribution and, finally, generate inheritance vectors to the left and right of x_0 for all pedigrees to compute $Z(\cdot)$ along the whole chromosome.

For incomplete marker data, the formula for \tilde{P} is more complicated than (32), but the procedure is analogous.

Example 12. *To illustrate the distribution of $Z(x_0)$ under \tilde{P} , consider the NPL score (25) of a single pedigree with m meioses. By conditioning on x_0 , it follows from (33) that,*

$$\tilde{P}(v(x_0) = w) = \frac{\exp[\delta S(w)]}{2^m M(\delta)}, \quad (34)$$

where $M(\delta) = 2^{-m} \sum_{w \in \mathbb{V}} \exp[\delta S(w)]$.

In particular, for an ASP with score function as in Example 10, the $2^m = 16$ inheritance vectors may be divided into three groups of sizes 4, 8 and 4, with $S(w) = -\sqrt{2}$, 0 and $\sqrt{2}$ respectively. The three groups correspond to sharing 0, 1 and 2 alleles IBD with allele-sharing probabilities $z_i = P(\text{IBD} = i | H_0)$ under H_0 , where all inheritance vectors $v \in \mathbb{V}$ are equally likely,

$$z = (z_0, z_1, z_2) = (0.25, 0.50, 0.25),$$

as discussed in Example 6. On the other hand, using (34) the allele sharing probabilities $\tilde{z}_i = \tilde{P}(\text{IBD} = i)$ under \tilde{P} are,

$$\tilde{z} = (\tilde{z}_0, \tilde{z}_1, \tilde{z}_2) = c_\delta^{-1} \left[\exp(-\delta\sqrt{2}), 2, \exp(\delta\sqrt{2}) \right],$$

where $c_\delta = [\exp(\delta\sqrt{2}) + 2 + \exp(-\delta\sqrt{2})]$ is a normalizing constant.

In Figure 16 we display the tilted IBD-sharing distribution \tilde{z} , through the $(\tilde{z}_1, \tilde{z}_2)$ -pair, as function of the tilting parameter δ .

As seen from (34), we use $\delta > 0$ at the artificial disease locus x_0 in order to increase the probability $\tilde{P}[v(x_0)] \propto \exp(\delta S[v(x_0)])$ of inheritance vectors $v(x_0)$ corresponding to large positive NPL scores. Since the NPL score at x_0 has an expected value depending on δ , the optimal choice of δ clearly depends on the threshold T of interest, see e.g. Naiman and Priebe (2001).³⁵ The procedure above is also extended to *optimally* form linear combinations of estimates $\tilde{\alpha}_\delta(T)$ using several distinct values of δ . This facilitates simultaneous estimation of $\alpha(T)$ for several thresholds T based on threshold-dependent linear combinations of $\tilde{\alpha}_\delta(T)$ -quantities.

The implementation of the corresponding algorithms and the choice of design parameters are discussed in some detail in the paper. Moreover, the

³⁵Basically one adjusts \tilde{P} in (32) in order to make $\tilde{V}[f(Z)L(Z)]$ small.

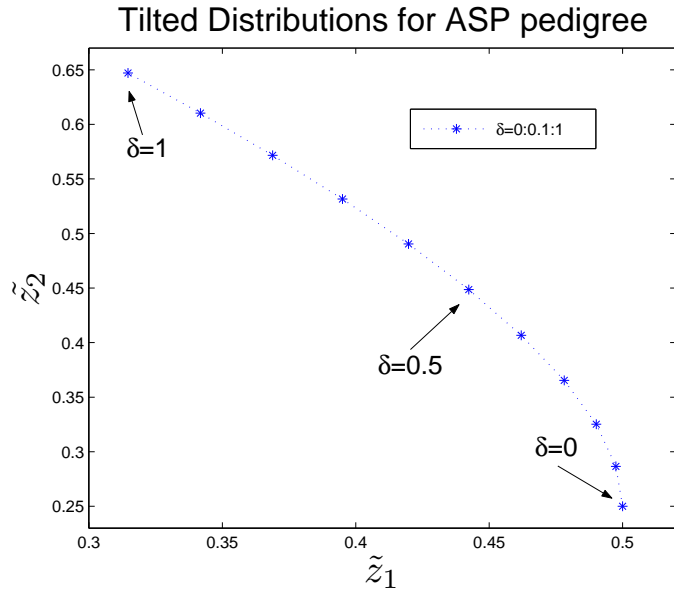


Figure 16: Displaying the IBD-sharing probabilities $(\tilde{z}_1, \tilde{z}_2)$ for an ASP, under \tilde{P} , with respect to the tilting parameter $\delta = (0, 0.1, \dots, 1)$.

suggested approach is evaluated using the concept of *cost-adjusted relative efficiency* with respect to standard Monte Carlo simulation and, in addition, comparison with the importance sampling procedure of Malley et al. (2002) is included.

4.3 Paper C

Here unconditional two-locus NPL analysis with respect to ASP pedigree sets, using (28), is discussed with special emphasis on *gene-gene interaction* (also worded as IBD- or allele-sharing correlation):

Definition 13. Assume two disease loci, l_1 and l_2 , with marginal one-locus IBD-sharing vectors (6), given by $z^1 = (z_0^1, z_1^1, z_2^1)$ and $z^2 = (z_0^2, z_1^2, z_2^2)$ respectively, and a joint IBD-sharing matrix (7). Now if,

$$\exists i, j : z_{ij} \neq z_i^1 z_j^2; \quad i, j \in \{0, 1, 2\},$$

we say there is gene-gene interaction present.

Since $z = \{z_{ij}\}$ is a function of disease allele frequencies and penetrance

values, we may reformulate the genetic model λ in (4) as $\lambda = (z, l)$, where $l = (l_1, l_2)$.

Further, we investigate the differences between using *simple* or *composite* two-locus null hypotheses.³⁶

Example 14. *An example of a composite null hypothesis when considering IBD-sharing with respect to a homogeneous set of ASPs is,*

$$H_0 : z \in \mathbb{Z}_0,$$

where $\mathbb{Z}_0 = \{z; z^1 = (0.25, 0.50, 0.25) \text{ or } z^2 = (0.25, 0.50, 0.25)\}$ corresponds to 'At most one disease locus'. In this case the corresponding alternative hypothesis H_1 is 'Two disease loci'.

Multiple suggestions on how to incorporate a composite null hypothesis into the analysis are made. Significance calculations may, for instance, be based on: (i) The *least-favourable IBD distribution* to derive theoretical significance level approximations (22). This results in a conservative upper bound,

$$\bar{\alpha}(T) = \max_{\lambda \in H_0} \alpha(T|\lambda),$$

of the significance level. (ii) Estimating a one-locus genetic component, for instance with estimates \hat{l}_1 and $\hat{z}^1 = \hat{z}^1(\hat{l}_1)$, and constraining the null hypothesis with respect to this component. Then perform Monte Carlo simulations to estimate p -values.

We also define, discuss and evaluate several classes of score functions. Extensive power calculations are performed based on both simple and composite null hypotheses using a wide range of genetic two-locus models λ , corresponding to varying degree of gene-gene interaction.

4.4 Paper D

This paper deals with two-locus NPL analysis based on (28) in general and conditional two-locus NPL analysis in particular. We adopt a quite general conditional approach which, compared to (30), facilitates *conditioning* on

³⁶Generally, defining a simple hypothesis consists of stating an instance $\theta = \theta_0$ of the unknown parameters $\theta \in \Theta$, whereas a composite hypothesis consists of the union of single-instance, or intervals of, parameter values in Θ . In our case a composite hypothesis refers to allowing for *one* disease locus in the stating of a null hypothesis.

pedigree-specific inheritance vectors rather than NPL scores. We also separate between using previously *known* or *unknown* conditioning loci. The former are defined prior to data analysis, whereas the latter are estimated from an initial one-locus linkage scan. The procedure for unknown loci may be described as follows:

Algorithm 15. (*Unknown conditioning loci*)

1. Perform chromosome-wise one-locus analysis along Ω .
2. Select the (possibly empty) set of conditioning loci \mathbb{X}_2 using some predefined selection-criterion.
3. Perform conditional two-locus analysis over Ω and \mathbb{X}_2 . For $x_2 \in \mathbb{X}_2$, we condition on inheritance information at x_2 and calculate scores when x_1 varies along the remaining chromosomes, i.e. $x_1 \in \Omega \setminus c(x_2)$.

We need to account for the possibly large set of conditioning loci leading to a multiple testing situation. Strategies: (i) We note on a somewhat conservative theoretical approximation. (ii) Our main interest is directed towards Monte Carlo simulation procedures. In this case, when performing n different tests and where n may be stochastic, we base the global significance test on the p -value,

$$\alpha_{\text{global}} = \min_{1 \leq i \leq n} \alpha_i(T_i),$$

where $\alpha_i(T_i)$ is the test-specific p -value corresponding to the i^{th} test, based on test statistic T_i .

Another topic is to derive *optimal score functions* with respect to given disease models, where optimality corresponds to the maximization of *noncentrality parameters (NCPs)* for one-, two- and conditional two-locus cases. In the one-locus case the NCP at disease locus l is defined as,

$$\text{NCP} = \text{NCP}(S, \lambda) = E[Z(l)|\lambda], \quad (35)$$

where S is the underlying score function and $\lambda \in H_1$ the one-locus genetic disease model. It turns out that the score function which maximizes (35) is,

$$S(w) \propto P(v(l) = w|Y, \lambda) - 2^{-m}; \quad \forall w \in \mathbb{V},$$

where Y is the phenotype vector, \mathbb{V} the set of possible inheritance vectors and m the number of meioses in the pedigree. See also Sham et al. (1997),

Nilsson (1999) and Hössjer (2003a). Similar results are obtained in the two-locus and conditional two-locus cases.

The paper also includes a number of explicit NCP and power calculations for several distinct classes of disease models.

Acknowledgements

Many thanks to Professor Ola Hössjer for inspiration, helpful and thoughtful comments, and numerous suggestions on how to improve this review. According to my opinion they surely did!

References

- Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). MERLIN - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, *30*, 97–101.
- Aldous, D. (1989). *Probability approximations via the Poisson clumping heuristic*. New York: Springer-Verlag.
- Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, *62*, 1198–1211.
- Almgren, P., Bendahl, P. O., Bengtsson, H., Hössjer, O. and Perfekt, R. (2003). *Statistics in genetics*. Department of Mathematical Statistics: Lund University.
- Ängquist, L. (2001). *Conditional two-locus NPL-analyses: Theory and applications* (Master's thesis No. 2001:E22). Lund: Department of Mathematical Statistics, Lund University.
- Ängquist, L. (2006, June). *Some notes on the choice of score function in nonparametric linkage analysis*. (Free download from homepage: '<http://www.maths.lth.se/matstat/staff/larsa/>'.)
- Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*, 781–791.
- Barnard, G. A. (1949). Statistical inference [Series B (Methodological)]. *Journal of the Royal Statistical Society*, *11*, 115–149.
- Bengtsson, O. (2001). *Two-locus affected sib-pair identity by descent probabilities: Constraints, parameterisation and estimation* (Licentiate thesis). Göteborg: Department of Mathematical Statistics, Chalmers University of Technology, Göteborg University.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing [Series B (Methodological)]. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

- Blackwell, P. (2002). Ornstein-Uhlenbeck process. In R. C. Elston, J. M. Olson and L. Palmer (Eds.), *Biostatistical genetics and genetic epidemiology* pp. 585–588. New York: John Wiley & Sons.
- Boehnke, M. (1986). Estimating the power of a proposed linkage study: A practical computer simulation approach. *American Journal of Human Genetics*, *39*, 513–527.
- Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models*. New York: Springer.
- Chen, W. M., Broman, K. W. and Liang, K. Y. (2004). Quantitative trait linkage analysis by generalized estimating equations: Unification of variance components and Haseman-Elston regression. *Genetic Epidemiology*, *26*, 265–272.
- Cherny, S. S., Sham, P. C. and Cardon, L. R. (2004). Introduction to the special issue on variance components methods for mapping quantitative trait loci. *Behavior Genetics*, *34*(2), 125–126.
- Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, *41*(1), 135–155.
- Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics*, *65*, 1170–1177.
- Clayton, D. (2001). Population association. In D. J. Balding, M. Bishop and C. Cannings (Eds.), *Handbook of statistical genetics* pp. 519–540. Chichester: John Wiley & Sons.
- Clayton, D. and Hills, M. (1993). *Statistical models in epidemiology*. Oxford: Oxford University Press.
- Collins, A., Frezal, J., Teague, J. and Morton, N. E. (1996). A metric map of humans: 23,500 loci in 850 bands. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 14771–14775.
- Cordell, H. J. and Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Applications to HLA in type 1 diabetes. *American Journal of Human Genetics*, *70*, 124–141.

- Cordell, H. J. and Clayton, D. G. (2005). Genetic association studies. *Lancet*, *366*, 1121–1131.
- Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics*, *21*, 213–215.
- Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, *23*, 34–64.
- Dudoit, S. and Speed, T. P. (1998). *Triangle constraints for sib-pair identity by descent probabilities under a general model for disease susceptibility* (Tech. Rep. No. 527). Department of Statistics, University of California, Berkeley.
- Elston, R. C. and Stewart, J. (1971). A general model for the analysis of pedigree data. *Human Heredity*, *21*, 523–542.
- Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *Journal of Applied Probability*, *30*, 766–779.
- Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics*, *53*, 234–251.
- Fulker, D. W., Cherny, S., Sham, P. C. and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, *64*, 259–267.
- Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Sociedad Española de Estadística e Investigación Operativa Test*, *12*(1), 1–77. (With discussion.)
- Göring, H. H. H. and Terwilliger, J. D. (2000). Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *American Journal of Human Genetics*, *66*, 1310–1327.

- Gudbjartsson, D. F., Jonasson, K., Frigge, M. and Kong, A. (2000). ALLEGRO, a new computer program for multipoint linkage analysis. *Nature Genetics*, *25*, 12–13.
- Haines, J. L. and Pericak-Vance, M. A. (Eds.). (2006). *Genetic analysis of complex disease*. New York: Wiley-Liss.
- Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between loci of linked factors. *Genetics*, *8*, 299–309.
- Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics*, *14*, 10–31.
- Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*. New York: John Wiley & Sons.
- Hand, D. J., Mannila, H. and Smyth, P. (2001). *Principles of data mining*. Cambridge, Massachusetts: The MIT Press.
- Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*, 3–19.
- Hesterberg, R. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, *37*(2), 185–194.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.
- Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, *52*, 362–374.
- Holmans, P. (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity*, *53*, 92–102.
- Hössjer, O. (2001). *Asymptotic estimation theory of multipoint linkage analysis under perfect marker information* (Tech. Rep. No. 2001:16). Lund: Department of Mathematical Statistics, Lund University.

- Hössjer, O. (2003a). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Annals of Statistics*, *31*(4), 1075–1109.
- Hössjer, O. (2003b). Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association*, *98*, 1035–1051.
- Hössjer, O. (2005a). Combined association and linkage analysis for general pedigrees and genetic models. *Statistical Applications in Genetics and Molecular Biology*, *4*(1:11). (Electronic journal, 42 pages)
- Hössjer, O. (2005b). Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics*, *6*(2), 313–332.
- Hössjer, O. (2005c). Information and effective number of meioses in linkage analysis. *Journal of Mathematical Biology*, *50*(2), 208–232.
- Hougaard, P. (2006). *Multiple testing: A clinical trial perspective*. Biostatistics: Lundbeck, Valby, Denmark. (Medicon Valley Academy Course Material: IDEON, Lund, 2006-03-06.)
- Hsu, Y. S. and Park, W. J. (1985). Ornstein-Uhlenbeck process [A Wiley-Interscience Publication]. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia in statistical sciences* Vol. 6, pp. 518–521. New York: John Wiley & Sons.
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*, 931–945.
- Khoury, M. J., Beaty, T. H. and Cohen, B. C. (1993). *Fundamentals of genetic epidemiology*. New York and Oxford: Oxford University Press.
- Khuri, A. I. (2003). *Advanced calculus with applications in statistics* (Second ed.). Hoboken (New Jersey): Wiley-Interscience.
- Kong, A. and Cox, N. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, *61*, 1179–1188.

- Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Annals of Eugenics*, 12, 172–175.
- Kruglyak, L., Daly, M. J. and Lander, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *American Journal of Human Genetics*, 56, 519–527.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics*, 58, 1347–1363.
- Kruglyak, L. and Lander, E. S. (1998). Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, 5(1), 1–7.
- Kullback, S. (1968). *Information theory and statistics* (Second ed.). New York: Dover.
- Kurbasic, A. and Hössjer, O. (2004). On computation of p-values in parametric linkage analysis. *Human Heredity*, 57, 207–219.
- Kurbasic, A. and Hössjer, O. (2006). Relative risks and effective number of meioses: A unified approach for general genetic models and phenotypes. *Annals of Human Genetics*, 70, 907–922.
- Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185–199.
- Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 85, 2363–2367.
- Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11, 241–247.
- Leadbetter, R., Lindgren, G. and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Berlin: Springer-Verlag.
- Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, Massachusetts: Sinauer Associates, Inc.

- Malley, J. D., Naiman, D. and Bailey-Wilson, J. (2002). A comprehensive method for genome scans. *Human Heredity*, 54, 174–185.
- McPeck, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology*, 16, 225–249.
- Mendel, J. G. (1866). Versuche über pflanzen-hybriden. *Verhandlungen Naturforschende Vereinigung Brünn*, 4, 3–47.
- Morgan, T. H. (1928). *The theory of genes*. New Haven: Yale University Press.
- Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7, 277–318.
- Naiman, D. Q. and Priebe, C. (2001). Computing scan statistic p -values using importance sampling, with applications to genetics and medical image analysis. *Journal of Computational and Graphical Statistics*, 10(2), 296–328.
- Nicolae, D. L. (1999). *Allele sharing models in gene mapping: A likelihood approach* (Doctoral thesis). Chicago: Department of Statistics, University of Chicago.
- Nicolae, D. L., Frigge, M. L., Cox, N. J. and Kong, A. (1998). Discussion. *Biometrics*, 54, 1271–1274. (Discussion of article by Teng and Siegmund, 1998)
- Nicolae, D. L. and Kong, A. (2004). Measuring the relative information in allele-sharing linkage studies. *Biometrics*, 60, 368–375.
- Nilsson, S. (1999). *Two contributions to genetic linkage analysis* (Licentiate thesis). Göteborg: Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University.
- Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 86(11), 4175–4178.
- Ott, J. (1999). *Analysis of human genetic linkage* (Third ed.). New York: The John Hopkins University Press.

- Ploughman, L. M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics*, *44*, 543–551.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.
- Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516–1517.
- Ross, S. M. (2006). *Simulation* (Fourth ed.). San Diego: Academic Press.
- Sham, P. (1998). *Statistics in human genetics*. London: Arnold Applications of Statistics.
- Sham, P., Zhao, J. and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Annals of Human Genetics*, *61*, 61–69.
- Sham, P. C., Cherny, S., Purcell, S. and Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components methods, for sibship data. *American Journal of Human Genetics*, *66*, 1616–1630.
- Sham, P. C., Purcell, S., Cherny, S. and Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics*, *71*, 238–253.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 and 623–656.
- Shih, M. C. and Whittemore, A. S. (2002). Tests for genetic association using family data. *Genetic Epidemiology*, *22*, 128–145.
- Siegmund, D. (1985). *Sequential analysis: Tests and confidence intervals*. Berlin: Springer-Verlag.
- Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, *58*, 1323–1337.

- Spielman, R. S., McGinnis, R. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, *52*, 506–516.
- Strachan, T. and Read, A. P. (2003). *Human molecular genetics* (Third ed.). London and New York: Garland Science.
- Strauch, K., Fimmers, R., Kurz, T., Deichmann, K. A., Wienker, T. F. and Baur, M. P. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *American Journal of Human Genetics*, *66*, 1945–1957.
- Suarez, B. K. (1978). The affected sib-pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens*, *12*, 87–93.
- Tang, H. K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, *2*, 147–162.
- Taqqu, M. S. (1975). Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *31*, 287–302.
- Teng, J. and Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, *54*, 1247–1265.
- Terwilliger, J. D. and Ott, J. (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity*, *42*, 337–346.
- Terwilliger, J. D. and Ott, J. (1994). *Handbook of human genetic linkage*. Baltimore and London: The John Hopkins University Press.
- Terwilliger, J. D., Speer, M. and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology*, *10*, 217–224.
- Thomas, D. C. (2004). *Statistical methods in genetic epidemiology*. New York: Oxford University Press.

- Thompson, E. A. (2000). *Statistical inference from genetic data on pedigrees*. Beachwood (Ohio) and Alexandria (Virginia): Institute of Mathematical Statistics and American Statistical Association.
- Tu, I. P. and Siegmund, D. (1999). The maximum of a function of a Markov chain and applications to linkage analysis. *Advances in Applied Probability*, 31, 510–531.
- Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of Brownian motion. *Physical Review*, 36, 823–841.
- Whittemore, A. S. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics*, 50, 118–127.
- Xing, C. and Elston, R. C. (2006). Distribution and magnitude of type 1 error of model-based multipoint lod scores: Implications for multipoint mod scores. *Genetic Epidemiology*, 30(5), 447–458.
- Xiong, M. and Jin, L. (2000). Combined linkage and linkage disequilibrium mapping for genome screens. *Genetic Epidemiology*, 19, 211–234.
- Zhao, H. (2000). Family-based association studies. *Statistical Methods in Medical Research*, 9, 563–587.
- Ziegler, A. and Koenig, I. R. (2006). *A statistical approach to genetic epidemiology: Concepts and applications*. Weinheim: Wiley-WCH.

A Some Notation Used in the Thesis Introduction

Item	Introduced	Comments and Examples
$\{a, b, c\}$	p. 5 p. 23	A set, i.e. a collection of elements. In this case the elements are a , b and c . In the same sense $\{f(x); x \in [a, b]\}$ may denote a continuous process, i.e. an uncountable set, indexed over interval $[a, b]$.
$f(A B)$	p. 16	Conditioning. The function f applied to A conditioned on, i.e. given, B . [Mainly used with respect to probability functions, i.e. $f = P(\cdot)$ for probability measure P .]
A^T	p. 21	Transposing. Interchanging rows and columns in matrix A . [Note that A^T is a transposed vector if A is a $(n \times m)$ -matrix with $n = 1$ or $m = 1$.]
$a!$	p. 28	The factorial function, i.e. $a! = a(a - 1) \cdots 1$, where a is a positive integer.
$[a(b[c(d)])]$	p. 31	Our adopted sequential delimiter system. Applied both to composite functions and separators.
$f(x) \propto x^a$	p. 48	Proportionality symbol. Here $f(x)$ is proportional to the function x^a , i.e. $f(x) = Cx^a$ where C is a constant.

Item	Introduced	Comments and Examples [cont.]
$a : b : c$	p. 49	A vector (v_1, v_2, \dots, v_n) where: (i) The first value is $v_1 = a$. (ii) For all $k : 2 \leq k \leq n$, the k^{th} element is $v_k = v_{k-1} + b$. (iii) Moreover, $c - b < v_n \leq c$, i.e. the final value v_n is at most c , but larger than $c - b$.
$A \setminus B$	p. 51	Complementary set to B with respect to A , i.e. $x \in A \setminus B$ if and only if $x \in A \cap x \notin B$. [Also called the <i>relative complement</i> of B with respect to A (Khuri, 2003).]