

Conditional Two-Locus  
NPL-Analysis: Theory and  
Applications

Lars Ängquist

2001-08-22

## Abstract

*Type 2 diabetes* is a serious, genetically influenced disease for which no fully effective treatments are available (Gelder Ehm et al.(2000)) [15]. The molecular basis of type 2 diabetes is unknown, to a great extent because of the substantial locus heterogeneity that is associated with diabetes risk (Gelder Ehm et al.(2000) [15] and Parker et al.(2001) [17]). However, studies indicate that a genetic component exists. Type 2 diabetes is a *complex disorder* and therefore it is assumed to depend on the actions and interactions of multiple genetical and environmental factors (see for example Cox et al.(1999) [9]). Simultaneous consideration of susceptibility from *multiple regions* may improve the possibility to find genes that are involved in the mechanism behind a complex disorder [9].

In this work conditional two-locus NPL-analyses will be performed. That means that one uses one-locus family scores from interesting regions (markers) to weight the results from other regions and finally, as a result, get the conditional two-locus NPL-score. This aims to find two correlated regions in linkage with the disease. Theory that describes how to calculate p-values under the null hypothesis of no linkage, which will make it possible to draw conclusions about any possible significance of the results, will be described and applied to the present data set. Aspects of robustness will be discussed.

The data set consists of 2606 individuals belonging to 337 families originating from Sweden and Finland.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Contents: Chapter by Chapter . . . . .	4
1.2	Thank You! . . . . .	5
<b>2</b>	<b>A Brief Introduction to Genetics</b>	<b>6</b>
2.1	Chromosomes and Genes . . . . .	6
2.2	Recombinations and Crossovers . . . . .	7
<b>3</b>	<b>Introduction to Non-Parametric Linkage Analysis</b>	<b>9</b>
3.1	Basic Definitions . . . . .	9
3.1.1	The Inheritance Vector . . . . .	9
3.1.2	Markers and Multipoint Analysis . . . . .	11
3.2	Score Functions . . . . .	12
3.2.1	One-Locus Score Functions . . . . .	12
3.2.2	Two-Locus Score Functions . . . . .	14
3.3	Parametric vs. Non-Parametric Linkage Analysis . . . . .	15
<b>4</b>	<b>NPL-Analysis: Further Theoretical Notes</b>	<b>17</b>
4.1	Conditional NPL-Analysis . . . . .	17
4.2	Calculations: $p$ – values . . . . .	19
4.3	Calculations: $\rho$ – values . . . . .	22
4.3.1	Theoretical Notes . . . . .	22
4.3.2	Examples: Different Pedigree Structures . . . . .	24
<b>5</b>	<b>Results</b>	<b>25</b>
5.1	Analyzing with Allegro . . . . .	25
5.2	Single-Locus Analyses . . . . .	28
5.3	Conditional Two-Locus Analyses . . . . .	29
5.4	Finding the $\rho$ – value/ $p$ – value . . . . .	30
5.4.1	Calculations . . . . .	30
5.4.2	$\rho$ : Monte Carlo Simulations and Further Properties . . . . .	31
5.5	Robustness . . . . .	33
5.5.1	Subject: Different Weights and Outliers . . . . .	34
5.5.2	Subject: Score Functions . . . . .	36
5.6	Conditioning on Regions Suggested by Meta-Analysis . . . . .	37
<b>6</b>	<b>Summary</b>	<b>40</b>
<b>A</b>	<b>Appendix: Graphs and Results</b>	<b>41</b>

## List of Figures

1	Chromatids crossing over each other during meioses. . . . .	9
2	Representation of inheritance at one locus with one family. . .	10
3	$\rho$ : pedigrees 1-5. . . . .	25
4	$\rho$ : pedigrees 6-7. . . . .	26
5	Plotting $\rho$ against $m$ and $n$ . . . . .	33
6	Investigating interesting markers on chromosomes 7 and 19 respectively. . . . .	34
7	Comparisons $S_{all}$ vs. $S_{pairs}$ . . . . .	37
8	Plotting family scores to compare robustness. . . . .	38
9	Results: single-locus analyses. . . . .	41
10	Results: conditioning on chr.4 <sub>max</sub> . . . . .	42
11	Results: conditioning on chr.7 <sub>max</sub> . . . . .	43
12	Results: conditioning on chr.12 <sub>max</sub> . . . . .	44
13	Results: conditioning on chr.16 <sub>max</sub> . . . . .	45
14	Results: conditioning on chr.18 <sub>max</sub> . . . . .	46

# 1 Introduction

*Type 2 diabetes* is a serious, genetically influenced disease for which no fully effective treatments are available (Gelder Ehm et al.(2000)) [15]. The molecular basis of type 2 diabetes is unknown, to a great extent because of the substantial locus heterogeneity that is associated with diabetes risk (Gelder Ehm et al.(2000) [15] and Parker et al.(2001) [17]). But, however, studies indicate that a genetic component exist, with a recurrence risk to first degree relatives of about 3.5 (Gelder Ehm et al.(2000) [15] and Rich(1990) [16]).

Type 2 diabetes is a *complex disorder*. According to Cox et al.(1999) [9] the transmission of such disorders is complex and dependent on the actions and interactions of multiple genetical and environmental factors. They also point out that simultaneous consideration of susceptibility from *multiple regions* may improve the possibility to find genes that are involved in the mechanism behind a complex disorder. That kind of considerations will be used in this work.

Further articles about the genomewide search for type 2 diabetes susceptibility genes may be found in, for example, Cox et al.(1999) [9], Gelder Ehm et al.(2000) [15], Parker et al.(2001) [17] and Mahtani et al.(1996) [18].

## 1.1 Contents: Chapter by Chapter

In *chapter 2* the biological definitions and notations that are needed to understand the text and what it is all about is introduced. *Chapter 3* is the start for a more mathematical description of this biological issue. Basic theory regarding non-parametric linkage analysis, including the inheritance vector and the concept of score functions, is explained and set into the appropriate context. In the following section, *chapter 4* the more specific theory that is needed to perform the analyses in this work and later on be able to fully interpret the results is introduced. The  $\rho$ -parameter is defined. How to calculate an overall significance value(p-value) that is corrected for multiple testing, using asymptotic approximations, is explained. In *chapter 5* these theoretical ideas are applied to a real data set and analyses according to different weighting schemes and to conditioning on different markers, positioned on distinct chromosomes, are performed. P-values are calculated. Some properties of the fluctuation parameter,  $\rho$ , are discussed. Monte Carlo simulations of inheritance vectors, to be able to calculate the  $\rho$ -value for large pedigrees, are briefly mentioned. Robustness related to usage of different score functions, the influence of questioned outliers and to different weighting schemes are discussed. The results conditioning on some prede-

find interesting regions are presented. In the last part, *chapter 6* a short summary and discussions about the results of this work are given. In the *appendix* a total presentation of the results will be given in the form of graphs and as plain and simple numerical values.

## 1.2 Thank You!

I would like to thank my supervisor Ola Hössjer<sup>1</sup> who always has an idea at hand when needed and who also was extremely helpful throughout the process behind this project. I would also like to thank the people at the Wallenberg Laboratory<sup>2</sup> in Malmö, especially Leif C. Groop and Peter Almgren for their help and for giving me the opportunity to work with some of their good, well prepared, data.

---

<sup>1</sup>Department of Mathematical Statistics, Lund University, Lund.

<sup>2</sup>Wallenberg Laboratory, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö.

## 2 A Brief Introduction to Genetics

This will be a short summary of biological definitions that are needed in the following chapters. For a deeper description of this area see for example [1], [2] or [3]. If you really want to understand these complex processes you might consider to read a textbook in genetics. The books mentioned above have been the source of information for the following sections.

### 2.1 Chromosomes and Genes

In the human individual there is a set of 23 pairs of *chromosomes*, located to the cell nucleus, in every cell of the body. Each chromosome contains two long *strands of DNA*<sup>3</sup> that are twisted around each other and normally bound to each other by hydrogen bonds. Each strand of DNA is made up by a sequence of subunits called *nucleotides*. These subunits is made up by a *sugar*, a *nitrogenous base* and *phospates*. There are four different kinds of nitrogenous bases<sup>4</sup> that may be seen as letters in a four-letter-alphabet and, according to the specific order of the bases, build different words. This long word is translated into *proteins* that is made up by so called *amino acids*. In a simplified manner you may define a *gene* as a segment of DNA, within a chromosome, that specifies the amino acid sequence of a single subunit of a protein and therefore are responsible for the expressions of specific characteristics. Interaction among genes that leads to specific expressions is called *epistasis*.

All the cells in the human being are derived from a single cell called the *zygote*. This cell is formed by the union of two *gametes*<sup>5</sup>. The special form of cell division that produces gametes is called *meiotes* (daughter cells contains 23 chromosomes) in contrast to normal cell division that is called *mitosis* (daughter cells contains 46 chromosomes). The two chromosomes of a chromosomal pair originates from different gametes. One gamete normally contains 22 non-sex chromosomes (autosomes) and one sex-chromosome. Two chromosomes of the same sort (for example chromosome 1) are called *homologous*.

The combination of all the DNA characteristics in a human being is called the human *genome*. A representation of the location of genes on the chromosome are called a *genetic map*. A *locus* may be defined as a specific position in the genome and will be an important concept later on. An *allele* may be seen as a specific DNA sequence at a locus. Different kinds of sequences

---

<sup>3</sup>DNA is shorthand for deoxyribonucleic acid

<sup>4</sup>Adenine(A), guanine(G), cytosine(C) and thymine(T)

<sup>5</sup>The two kinds of gametes are the ovum and the sperm

corresponds to different alleles. The terminology in this area is discussed and may be somewhat different.

## 2.2 Recombinations and Crossovers

The two different alleles at a locus constitutes the so called *genotype* of the present individual. The expression of a given genotype is called a *phenotype*.

For example, if one consider the *ABO* blood group locus there are three different kinds of alleles  $\{A, B, O\}$  that with all possible combinations forms six distinct genotypes  $\{AA, AB, AO, BB, BO, OO\}$  and these combinations corresponds to the four possible phenotypes (bloodgroups)

$\{A_{AA/AO}, B_{BB/BO}, AB_{AB}, O_{OO}\}$ , where the lowercase indexes corresponds to the genotypes that constitutes that specific phenotype. A genotype consisting of two alleles that are of the same sort  $\{AA, BB, OO\}$  are called *homozygous* and genotypes with distinct alleles  $\{AB, AO, BO\}$  are called *heterozygous*.

At a specific locus one allele will be inherited from the mother and the other one will be inherited from the father. The mother and father themselves have inherited their alleles from the grandfathers and grandmothers. If one considers two different loci A and B, then the maternal(paternal) gamete is defined to be *non-recombinant* with respect to these two loci if the maternal(paternal) alleles in the offspring (at these loci) both originates from the same grandparent. Otherwise the gamete are defined to be *recombinant*. The probability that two loci are recombinant is called the *recombination fraction*,  $\theta$ , (with respect to these specific loci) with the restriction that  $\theta$  is defined to be  $1/2$  if the probability is greater than  $1/2$ . Loci on the same chromosome are said to be *syntenic* and loci on different chromosomes are said to be *non-syntenic*. In the latter case  $\theta$  is defined to be  $1/2$  because the inheritance of gametes from different chromosomes are supposed to be independent. In the former case the gametes inherited are the same, but a phenomenon called *crossover* slightly complicates the picture.

During meioses, cell division leading to the formation of gametes, homologous chromosomes pair up. A chromosome pair consists of four strands that here will be called *chromatids*. The non-sister chromosomes is in contact with each other at zones called *chiasmata*. At such positions a so called crossover takes place. See figure 1 to, hopefully, get a clearer view.

Now there are four different chromatids where some of them may be built by alternating fragments originating from both of the grandparents. Each gamete then receives one of the four chromatids to construct 23 chromosomes. The combined gametes from the maternal/paternal meioses then form the

offsprings human genome. If one once again looks at figure 1 and focuses on loci A/D then inheritance of the, starting from left, first or third chromatids corresponds to a gamete non-recombination and therefore, somewhat self-explained, the second or fourth chromatid will naturally correspond to gamete recombinations.

Crossovers appear quite random with the exception that the probability of a crossover closely following another one is much lower than otherwise. This is called *chiasma interference*. The recombination fraction are related to the concept of *map distance*,  $m(\theta)$ , in such a way that map distance, between two loci, is defined to be the expected number of crossovers taking place between them on a single chromatid during meiosis. The distance is measured in Morgans(M). The human sex-averaged autosomal map length usually is measured to be about 33 Morgans but newer investigations show that it may be slightly longer, about 36 Morgans. The average chromosome is then about 1.5 Morgans long, which according to the definition above means that the expected value of crossovers on the average chromosome equals 1.5. Exact correspondence between map distance and the recombination fraction may be described with so called *map functions*<sup>6</sup>. Descriptions of the most important map functions may be found in, for example, [1] or [2].

---

<sup>6</sup>For example the Haldane map function is described as:  $m(\theta) = -\frac{1}{2} \ln(1 - 2\theta)$

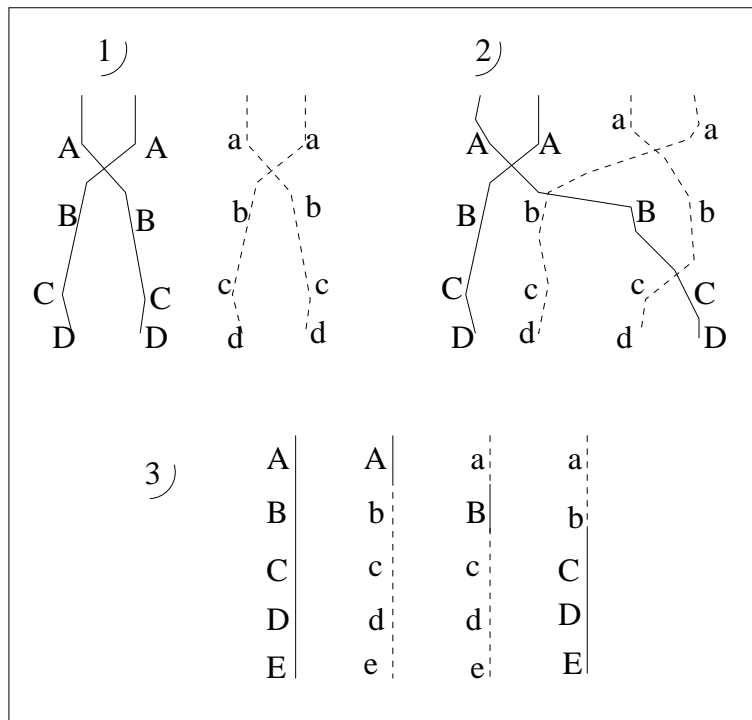


Figure 1: Chromatids crossing over each other during meiosis.

### 3 Introduction to Non-Parametric Linkage Analysis

Here, in this chapter, basic definitions and methods considering *NPL-analysis* will be described. The alternative approach, *parametric linkage analysis*, and some comparisons between these different ways of performing linkage analysis will be discussed in the last section below.

#### 3.1 Basic Definitions

In the following subsections important concepts as the inheritance vector and the difference between singlepoint and multipoint linkage analysis will be explained.

##### 3.1.1 The Inheritance Vector

A *pedigree* is a set of relatives. The structure of a pedigree may be represented as a mathematical tree. The members of a pedigree may be divided into two subgroups: *founders* and *non-founders*. Founders are individuals

whose parents are not in the pedigree and non-founders are individuals whose parents belong to the present pedigree.

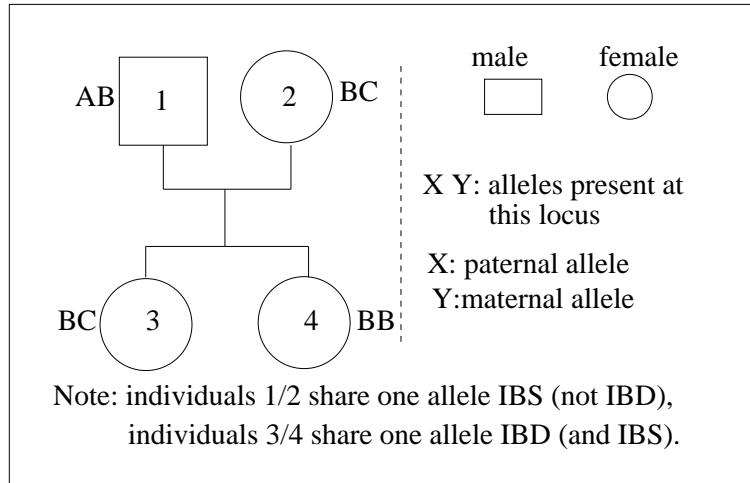


Figure 2: Representation of inheritance at one locus with one family.

Two individuals are said to share an allele that is *identical-by-descent* (*IBD*) if they have inherited the same specific allele originating from a specific founder. If two individuals carry an allele of the same sort, with the same DNA-sequence, at a locus they are said to share the allele *identical-by-state* (*IBS*). Obviously sharing an allele IBD implies sharing the allele IBS, but the reverse implication is certainly not true. For an example, take a look at figure 2.

Every individual have two alleles present at each locus. One allele is inherited from the mother and may be called the *maternal allele*, the other allele is inherited from the father and may be called the *paternal allele*. At a specific locus,  $x$ , the whole inheritance process, for a given pedigree, may be described using a so called *inheritance vector*,  $v(x)$ . Some notation:  $n$  is the number of individuals in the pedigree,  $f$  is the number of founders,  $n - f$  is the number of non-founders and  $m = 2(n - f)$  is the number of meioses present during the inheritance of alleles for this pedigree. Then  $v(x)$  may be defined as:

$$v(x) = (p_1, m_1, p_2, m_2, \dots, p_{n-f}, m_{n-f}) \quad (1)$$

In (1)  $p_i$  and  $m_i$  are defined to be equal to 0 if the  $i$ :th nonfounders paternal and maternal allele respectively origin from a grandfather and to be equal to 1 if they origin from a grandmother. To get an example look at figure 2 where the inheritance vector at that locus,  $x$ , equals  $v(x) =$

(1, 1, 1, 0).

### 3.1.2 Markers and Multipoint Analysis

A genetic *polymorphism* refers to the situation where several different alleles (DNA sequences), at a locus, exists in a population. If a polymorphism is reliably detectable and highly variable (c.f. Sham [1]) it may be used as a genetic *marker*. A marker is a locus, at a known or estimated position in the genome, where it has been possible to measure the alleles present for the actual pedigrees. There exist several different methods to measure the alleles (see for example [1] or [2]).

In NPL-analysis one defines genetic *linkage*, to a locus, to be deviations from random inheritance of the founder alleles among people with a certain genetical expression (in our case affected individuals) at that locus. If one wants to find a possible linkage within a region, then one has to test for linkage against markers that are covering the whole region of interest. The included markers gives us a *marker map*. If one finds linkage to a specific marker, then one can draw the conclusions that there is linkage to a locus in that region, which could be the marker locus or some non-marker locus positioned in the neighbourhood. That conclusion is based on the relationship between map distance and recombinations. A situation with a short map distance between loci implicates a small probability of recombination between loci which itself implicates that linkage to a region will be detectable at neighbouring marker loci. To find a more precise position one has to include more markers in that region, which will give a *denser* map. It may be important to stress the fact, as Terrwilliger/Ott do [3], that the term linkage refers to loci not to specific alleles at these loci.

Sometimes there will be incomplete *information* in the actual data set that one works with. Data may be lost, unable to measure, non-polymorphic etc. For example if an individual is homozygous at a locus it may be impossible to find out if it was the maternal allele or the paternal allele that was inherited. Several measures of the information contained in the data set at a locus exists. In my investigations I used the computer program Allegro (see [5]) and that program uses measures given by equations (2.2) (combined for all families) and (2.4) (per family) in Nicolae (1999) [6]. The information measure is usually ranging from 0 (no information) to 1 (full information).

If one just uses the information contained in the data set at the present specific locus of investigation one performs a *singlepoint* linkage analysis, but if one is using the information contained in the data set at the surrounding marker loci as well one performs a *multipoint* linkage analysis. In the lat-

ter case the extracted information when performing linkage analysis will be increased. In other words one uses the information contained in the inheritance vectors from the surrounding markers to try to, in a probabilistic way, fill the non-informative gaps at the investigated locus. So, in the multipoint case one can perform the analysis using the multipoint inheritance vector  $v(x)$  (see 1), where  $x$  belongs to the set of markers and where the elements changing between 0 and 1 when a recombination has taken place between the two marker positions at that specific meioses. In the investigations performed in this work multipoint analysis has been performed and Allegro (see [5]) uses the theory of Hidden Markov Models (HMM), and known recombination fractions between markers, to gain information from the neighbouring markers in that case. For further reading about related topics it might be interesting to consider, for example, Kruglyak et al. (1996) [4] and Lander/Green (1987) [20].

### 3.2 Score Functions

So how does one measure linkage in the non-parametric case? Usually one uses mathematical functions called non-parametric *score functions*.

#### 3.2.1 One-Locus Score Functions

In the *one-locus* linkage analysis case the non-parametric score function may be written as  $S_1(v) = S_1(v; Y)$  where  $v$  is the inheritance vector at the particular locus of interest (found by singlepoint or multipoint analysis) and  $Y$  is the phenotype vector (in our case: diseased/not diseased).

One simple and common score function is the so called  $S_{pairs}$  function which is defined as follows:

$$S_{pairs}(v) = \sum_{(p,q)} S_{pq}(v), \quad (2)$$

where  $(p, q)$  in the summation refers to pairs of affected people in the pedigree,  $S_{pq}(v) = \sum_{i=0}^1 \sum_{j=0}^1 \phi_{ij}(p, q)$  and  $\phi_{ij}(p, q)$  is 1 if allele  $i$  of  $p$  and allele  $j$  of  $q$  are IBD and 0 otherwise.

The function that will be used later on in this work is called  $S_{all}$ . It is defined as follows:

$$S_{all}(v) = 2^{-a} \sum_h \prod_{i=1}^{2f} b_i(h)! \quad (3)$$

where  $a$  is the number of affected individuals in the pedigree,  $h$  is a selection that picks one of the two present alleles for each affected individual,  $b_i(h)$  is the number of times founder allele number  $i$  ( $i = 1 \dots 2f$ ) appears in

$h$  (given  $v$ ) and the sum is taken over all the possible ways of choosing  $h$ . The score may be seen as the average number of permutations that preserve a collection obtained by choosing one allele from each of the affected individuals in the pedigree (c.f. [4]). The two functions described above have both been explained by for example Whittemore and Halpern (1994) [7]. In fact they even introduced the latter one of these two formulas.

The reason to favour  $S_{all}$  is that one can often gain statistical power when considering larger sets of affected individuals than just pairs (c.f. [4] and [7]). It is in some fashion, as Kruglyak et al(1996) [4] points out, more impressive to find that, for example, six individuals share the same allele IBD than that each pair of them share some allele IBD. Kruglyak et al. (1996) have compared these two methods and found that the latter score function performed the best in almost all their investigations. On the negative side, as we shall see later on, that score may in some situations not be that robust because of the sharply increased weight when the number of affected individuals sharing a specific allele increases.

The computer program Allegro [5], that I used in my investigations, is also able to perform NPL-analysis using three other score functions called  $S_{homoz}$ ,  $S_{robdom}$  and  $S_{mnallele}$ . None of these three score functions will be described any further in this work.

The rest of this subsection will closely follow the presentation by Kämpe (2001) [8] included in his Master's thesis. Let us get back to the formal description of the one-locus score function  $S_1(v)$ . To get a proper test statistic one usually normalizes the score and as a result one gets the *normalized one-locus score function*  $Z_1(v)$ :

$$Z_1(v) = \frac{S_1(v) - \mu_1}{\sigma_1} \quad (4)$$

$$\text{where } \mu_1 = \sum_w S_1(w)p_v(w)$$

$$\text{and } \sigma_1^2 = \sum_w S_1(w)^2 p_v(w) - \mu_1^2$$

In the equations shown above  $p_v(w)$  refers to the probability distribution of  $w$  under the null hypothesis  $H_0$ : *no linkage*. It equals  $p_v(w) = P(v = w) = 2^{-m}$ . It is easy to show that under  $H_0$ ;  $E(Z_1(v)) = 0$  and  $V(Z_1(v)) = 1$ .

A more common way to present this normalized score is as a function of the marker position instead of as a function of the inheritance vector at a specific marker.

$$\bar{Z}_1(x) = \sum_w P(v(x) = w)Z_1(w), \quad x \in \Omega, \quad (5)$$

where  $\Omega$  is the union of all marker positions and  $P(v(x) = w)$  is the probability function for the inheritance vector  $v(x)$  given the marker data. With perfect information contained in the marker data at position  $x$  the equation,  $\bar{Z}_1(x) = Z_1(v(x))$ , holds because  $v(x)$  is determined to be a single specific value with probability one.

### 3.2.2 Two-Locus Score Functions

*Two-locus* score functions are mainly just a generalization of the one-locus case. In this situation one wants to be able to capture problems where two different loci in some fashion cooperates according to genetical expressions as in, for example, diseases. The two-locus score function may formally be written as  $S_2(v, v') = S_2(v, v'; Y)$  where  $v, v'$  are the inheritance vectors at the two different loci involved and  $Y$ , as before, is the phenotype information from the pedigree. The normalized score function will then be possible to express as:

$$Z_2(v, v') = \frac{S_2(v, v') - \mu_2}{\sigma_2} \quad (6)$$

$$\text{where } \mu_2 = \sum_{w, w'} S_2(w, w') p_{v, v'}(w, w')$$

$$\text{and } \sigma_2^2 = \sum_{w, w'} S_2(w, w')^2 p_{v, v'}(w, w') - \mu_2^2.$$

As a generalized version of the one-locus case above  $p_{v, v'}(w, w')$  refers to the joint probability distribution of  $v, v'$  under the null hypothesis  $H_0$  : *no linkage*. If the two loci are situated on different chromosomes it equals  $p_{v, v'}(w, w') = P(v = w, v' = w') = 2^{-2m}$ .

As above this normalized score usually is presented in a slightly different form.

$$\bar{Z}_2(x, x') = \sum_{w, w'} P(v(x) = w, v(x') = w') Z_2(w, w'), \quad x, x' \in \Omega, \quad (7)$$

where  $P(v(x) = w, v(x') = w')$  is the joint probability function for the inheritance vectors  $v(x)$  and  $v(x')$  given the marker data. If these two positions,  $x$  and  $x'$ , are situated on different chromosomes, and therefor  $v(x)$  and  $v(x')$  are mutually independent, one may simplify this expression a bit further:

$$\bar{Z}_2(x, x') = \sum_{w, w'} P(v(x) = w) P(v(x') = w') Z_2(w, w'), \quad x, x' \in \Omega, \quad (8)$$

where  $c(x) \neq c(x')$  and  $c(x)$  denotes the chromosome where marker position  $x$  is situated.

According to Ott (1999) [2] two-locus inheritance appears quite frequently in nature, so it may be of interest, in some cases, to perform analysis with aid of that kind of methods. Of course, there is nothing that says that the structure behind certain genetical expressions, for example complex diseases, would not be more complex so generalizations of higher degree may certainly be interesting in some situations. The problem will be, as usual, that more complex models will increase the number of possible analyses to perform and maybe with even higher degree the complexity of the computations.

One may perform different kinds of two-locus linkage analysis. One approach is to look simultaneously for the two marker loci involved  $(x, x')$ , but the approach in this work will be to perform a so called *conditional two-locus NPL-analysis*. In this case one does not need to consider jointly two-locus score functions, but rather two-locus score functions that are made up by combinations of one-locus score functions. The conditional approach has been described, for example, by Cox et al.(1999) [9].

### 3.3 Parametric vs. Non-Parametric Linkage Analysis

In *parametric linkage analysis* one assumes that the genetical model of the disease is known. One then usually uses the so called *lod score method* where one performs some sort of *likelihood ratio test* with the recombination fraction as an unknown parameter and under the null hypothesis set as  $H_0 : \theta = \frac{1}{2}$  corresponding to no linkage. Using this method one maximizes the lod score  $Z(\theta)$  with respect to  $\theta$  where:

$$Z(\theta) = \log_{10} \left( \frac{L(\theta)}{L(\frac{1}{2})} \right), \quad (9)$$

then the found value will be the best estimate of  $\theta$  called  $\hat{\theta}$  and a large positive value of  $Z(\hat{\theta})$  will favour the alternative hypothesis of linkage.

Of course, there are both positive and negative sides of choosing either of these two approaches. As Kruglyak et al. (1996) [4] point out NPL-analysis is much more robust considering uncertainty about the mode of inheritance according to the disease. They also notify that even if the parametric approach performs better under the true model the NPL-analysis loses relatively little power relative to that approach in that situation. They suggest that NPL-analysis should be performed when there is not much known about the mode of inheritance, which is the case when considering complex diseases. This is the situation one has to confront in this work when type 2 diabetes is considered. The same opinion seems to be shared by Terrwilliger/Ott (1994) [3].

On the negative side, according to Sham (1998) [1], is that when performing NPL-analysis one has to choose between a lot of different models when deciding which weighting scheme to use for different kind of pedigrees/individual pairs in a pedigree/individuals in a pedigree etc. Sham points out that the optimal weighting system depends on the true model of inheritance so in that aspect the non-parametric analysis is in fact parametric!

## 4 NPL-Analysis: Further Theoretical Notes

In this chapter the theory that is needed to understand the results described in the next chapter will be introduced. First the important concept of conditional NPL-analysis will be described (c.f. for example Cox et al.(1999) [9]). It is a method well suited for the two-locus non-parametric linkage analysis case.

### 4.1 Conditional NPL-Analysis

This subsection will quite closely follow the presentation made by Kämpe (2001) [8]. The following *expected value calculations* are performed under the null hypothesis of no linkage and assuming that the two loci involved are situated on different chromosomes. First we create the multiplicative two-locus score function.

$$S_2(v, v') = \tilde{S}_1(v) \frac{S_1(v') - \mu_1}{\sigma_1} = \tilde{S}_1(v) Z_1(v') \quad (10)$$

Using that  $\mu_2 = E(S_2(v, v')) = 0$  and  $\sigma_2^2 = V(S_2(v, v')) = \tilde{\mu}_1^2 + \tilde{\sigma}_1^2$  we can then, as in the one-locus case (see equations 4 and 6) form the normalized two-locus score function.

$$Z_2(v, v') = \frac{S_2(v, v') - \mu_2}{\sigma_2} = \frac{\tilde{S}_1(v)}{\sqrt{\tilde{\mu}_1^2 + \tilde{\sigma}_1^2}} Z_1(v'), \quad (11)$$

where  $\tilde{\mu}_1$  and  $\tilde{\sigma}_1$  are defined in the same manner as in equation (4).

Now we face a situation where we have to, in some way, use the results from the one-locus case,  $Z_1(v)$ , to be able to get an explicit value out of the equation in the conditional two-locus case. This has been discussed by for example Cox et al. (1999) [9]. In their paper they have presented three different functions which are all of the form  $\tilde{S}_1(v) = \gamma(Z_1(v))$ . These functions are exactly the same functions that has been used during the analyses performed in this work. These three different functions are given here:

$$\begin{aligned} \gamma_{prop}(Z) &= Z \cdot \mathbf{1}_{\{Z>0\}} \\ \gamma_{01}(Z) &= \mathbf{1}_{\{Z>0\}} \\ \gamma_{het}(Z) &= \mathbf{1}_{\{Z<0\}} \end{aligned} \quad (12)$$

The first two functions are suited for situations when there are *positive interactions* (epistasis) between genes while the last function is appropriate to use in situations when *genetic heterogeneity* is present. Of course, when considering complex diseases there may be a complex explanation to the disease that

will consist of both a system of genetic interactions (epistasis/heterogeneity) and certain environmental factors. According to Cox et al.(1999) [9] animal studies have suggested that epistatic interactions between genes at least partially are responsible for type 2 diabetes which is the disease that will be studied, in a statistically applied way, in the next chapter.

To be able to perform real-data analyses one needs to be able to weight the NPL-scores from different pedigrees together to get an overall NPL-score. In this work the theory above has been used to perform a conditional weighting. Then the  $\gamma$ -functions, that were defined above, may be seen as family-specific weights. The two-locus score function for pedigree number  $k$  may be described as (c.f. equations 5 and 7):

$$\begin{aligned}\bar{S}_{2k}(x, x') &= \sum_{w, w'} P(v_k(x) = w)P(v_k(x') = w')S_2(w, w') \\ &= \left( \sum_w P(v_k(x) = w)\tilde{S}_1(w) \right) \bar{Z}_{1k}(x') \\ &= \bar{\gamma}_k(x)\bar{Z}_{1k}(x')\end{aligned}\quad (13)$$

As noted before  $\tilde{S}_1 = \gamma(Z_1(w))$  and then  $\bar{\gamma}_k(x)$  may be expressed as:

$$\bar{\gamma}_k(x) = E(\gamma(Z_1(v_k(x)))) = \sum_w P(v_k(x) = w)\gamma(Z_1(w)) \quad (14)$$

It is computationally preferable to move the expectation in (14) within  $\gamma(\cdot)$ , since then the one-locus score function (5) can be utilized. This will give us an expression that will look like...

$$\bar{\gamma}_k(x) = \gamma(E(Z_1(v_k(x)))) = \gamma\left(\sum_w P(v_k(x) = w)Z_1(w)\right) = \gamma(\bar{Z}_{1k}(x)) \quad (15)$$

Now the equality between (15) and the equations (13) and (14) does not hold but in return we get an expression where all we need to know to calculate the overall NPL-score are the individual NPL-scores for the different pedigrees included in the study. Conditioning on the one-locus NPL-scores at a specific marker  $x$  the weights  $\bar{\gamma}_k$  in (15) will be constants and it is an easy task to calculate the two-locus normalized weighted score function:

$$\begin{aligned}\bar{Z}_2(x, x') &= \frac{\sum_{k=1}^P \bar{S}_{2k}(x, x') - \sum_{k=1}^P E(\bar{S}_{2k}(x, x')|\bar{Z}_{1k}(x))}{\sqrt{(\sum_{k=1}^P V(\bar{S}_{2k}(x, x')|\bar{Z}_{1k}(x)))}} \\ &= \frac{\sum_{k=1}^P \bar{\gamma}_k(x)\bar{Z}_{1k}(x')}{\sqrt{(\sum_{k=1}^P \bar{\gamma}_k(x)^2)}}\end{aligned}\quad (16)$$

$$= \frac{\sum_{k=1}^P \gamma(\bar{Z}_{1k}(x)) \bar{Z}_{1k}(x')}{\sqrt{(\sum_{k=1}^P \gamma(\bar{Z}_{1k}(x))^2)},$$

where  $P$  is the number of pedigrees. This is how the overall NPL-score is calculated by the computer program Allegro [5] which has been used to perform the investigations discussed in the next chapter.

## 4.2 Calculations: $p$ – values

Now we are facing a situation where one may perform genome-wide NPL-analyses and find a maximum two-locus NPL-score. To put more meaning into that single value we need to calculate some sort of probability, a so called  $p$ -value, of how unusual it is to find a value like that we have found under some predefined *null hypothesis*. This question will be dealt with in the following paragraphs of this section.

One way to calculate a  $p$ -value is by using the technique of *permutation tests* and then, for example, pick the family weights, with/without replacement, from the pooled sample of all the family weights given by the results found at the conditioning marker. Several alternative methods to calculate related but different kind of  $p$ -values may be used. We will use one of them, which follows a quite different approach compared to the one mentioned above.

First we need some further notations. The indexing etc will, in as many ways as possible, be consistent with the notations in the presentation by Kämpe (2001) [8].

The interesting one-point regions, marker positions, where the family-specific NPL-scores which will be used to conditionally perform two-locus analyses have been calculated, will be defined as follows:

$$D = \{\hat{x}_1, \hat{x}_2, \dots\}, \bar{Z}_1(\hat{x}_i) > T, \quad (17)$$

where  $T$  is a NPL-score threshold,  $\bar{Z}_1(\hat{x}_i)$  is the maximum NPL-score among the markers located on chromosome  $c_i$ ,  $c_i \in C$ ,  $C$  is consisting of the 22 different autosomes (non-sex chromosomes) and all the chromosomes  $c_i(\hat{x}_i)$  present are different.

When one conditions on the results at several different markers and/or use more than one different weighting scheme (see equation 12) one needs to correct the found  $p$ -value for *multiple testing*. This issue will be discussed in more detail in the end of this section. For now we assume that we perform a conditionally two-locus NPL-analysis conditioning on the results at one marker position,  $\hat{x}$ , and that we will only use one single weighting scheme.

The statistic for the maximum NPL-score may then be written as:

$$\bar{Z}_{max} = \sup\{\bar{Z}_2(\hat{x}, x'), x' \in \Omega, c(x') \neq c(\hat{x})\}, \quad (18)$$

which means that we found the maximum value of the present two-locus NPL-score function after it has been calculated at all the marker positions available at all the 21 autosomes that are different than the chromosome where one finds the marker where the one-locus weights, that we conditioned on, were calculated. Moreover  $\bar{Z}_2(x, x')$  is given by (16).

Now one is ready to formulate the null hypothesis  $\bar{H}_0$ :

$$\begin{aligned} \bar{H}_0 : \quad & \{v_k(x), 0 \leq x \leq l\} \text{ and } \{v_k(x'), 0 \leq x' \leq l'\} \\ & \text{are independent given phenotypes } Y_k, k = 1, \dots, P, \quad (19) \\ & \text{and the marginal distribution for } \{v_k(x'), 0 \leq x' \leq l'\} \\ & \text{is uniformly distributed,} \end{aligned}$$

where  $P$  is the number of pedigrees included in the study and  $l, l'$  are the lengths of the chromosomes corresponding to  $x$  and  $x'$ .

The probability that one is interested to find, a sub-destination of this mathematical walk,  $\bar{p}(\bar{z}_{max})$  will now be expressed:

$$\bar{p}(\bar{z}_{max}) = P(\bar{Z}_{max} \geq \bar{z}_{max} | \bar{H}_0, \{v_k(x), 0 \leq x \leq l\}_{k=1}^P) \quad (20)$$

If one studies (16) and the derivation of the expressions included in that equation (see chapter 3) it is easy to show that under the assumption of perfect marker information and under the null hypothesis the following statements holds:

$$E(\bar{Z}_2(\hat{x}, \cdot) | \bar{H}_0, \text{ and } \{v_k(x), 0 \leq x \leq l\}_{k=1}^P) = 0 \quad (21)$$

$$V(\bar{Z}_2(\hat{x}, \cdot) | \bar{H}_0, \text{ and } \{v_k(x), 0 \leq x \leq l\}_{k=1}^P) = 1 \quad (22)$$

When there is an imperfect-data situation present, the variance in the expression above is less than 1 under the null hypothesis (see for example Kruglyak et al. (1996) [4] or the technical report that belongs to Allegro (2000) [5]), and this will make the p-value conservative when one calculates it using the theory/equations described above. An analysis performed with data where imperfect-data information are present is therefor said to be performed under *perfect-data approximations*. Kruglyak(1996) [4] also points out that

though the p-value is conservative it sacrifices relatively little power except for situations where the information content is very low. To try to find the probability that we are looking for we will use an asymptotic approximation to calculate the p-value in (20).

According to Kämpe(2001) [8] the following expression will asymptotically be true:

$$(\bar{Z}_2(\hat{x}, \cdot) | \bar{H}_0, \{v_k(x), 0 \leq x \leq l\}_{k \geq 1}) \xrightarrow{L} V(\cdot), \quad (23)$$

as  $P \rightarrow \infty$  where  $V(\cdot)$  is a Gaussian process (Ohrnstein-Uhlenbeck). According to Lander/Kruglyak(1995) [11] the following approximation is increasingly accurate as  $\bar{z}_{max}$  grows:

$$P(\sup\{V(x')\} > \bar{z}_{max}) \approx 1 - e^{-\mu(\bar{z}_{max})}, \quad (24)$$

where  $x'$  ranges over the same set as in (18). Now (24) may be used to formulate an asymptotic formula regarding the probability  $\bar{p}(\bar{z}_{max})$ :

$$\bar{p}(\bar{z}_{max}) \approx 1 - e^{-\mu(\bar{z}_{max})}, \quad (25)$$

where (c.f. Lander/Kruglyak(1995) [11])  $\mu(z)$  is an approximation for the mean number of regions where the normal process  $V$  exceeds the threshold. This parameter may be written in a more explicit form:

$$\mu(z) = [C + 2\rho Gz^2]\alpha(z), \quad (26)$$

where  $C$  is the number of chromosomes included in the scan,  $G$  is the genetic length of these chromosomes (According to the lengths presented in Ott(1999) [2] the result will be: 35.73 Morgans if all the 22 autosomes are included in the scan),  $z$  is the already mentioned threshold (for example  $\bar{z}_{max}$ ),  $\alpha(z)$  is the pointwise significance level of exceeding  $z$  ( $\alpha(z) = 1 - \Phi(z)$ , where  $\Phi$  is the cumulative distribution function of a standard normal random variable) and where  $\rho$  measures the fluctuation of the score-statistic  $S(x)$  and therefore depends on the family structure and the information regarding the present individuals disease status (see the next section).

As was earlier remarked one must correct for multiple testing when using more than one marker and/or one weighting scheme when performing analyses of the kind described above. Let us consider the possible situation where we have performed  $N$  different genome scans and found an overall top NPL-score  $z_{top}$ , then we may use a so called *Bonferroni correction* in the following way:

$$P_{overall} = P(\bar{Z}_{top} \geq \bar{z}_{top}) \leq \sum_{i=1}^N P(\bar{Z}_{max,i} \geq \bar{z}_{top}) \quad (27)$$

For further reading about this/related topics consider, for example, Lander/Botstein (1989) [13] or Feingold (1993) [12], where one for example may find discussions about the extreme value theory forming the basis of (24)-(25).

### 4.3 Calculations: $\rho$ – values

In the subsection that follows right after this note a description of how to calculate the family-specific  $\rho$  – value, using a general formula that will be independent of the pedigree structure and therefore applicable for all kinds of pedigrees, will be described. These values will then be weighted to produce an overall  $\rho$  – value that will make it possible to calculate  $p$  – values in the manner described in the last section. The theory presented in the following subsection depends on the work produced by Hössjer (2001) [10].

The second subsection will give a few examples of  $\rho$  – calculations for pedigrees with different structures (disease-schedules and sizes).

#### 4.3.1 Theoretical Notes

According to Lander/Kruglyak(1995) [11]  $\rho$  is related to the autocorrelation function  $C$  under  $H_0$  of the stationary process  $S(v(x))$  and may be defined as

$$\rho = -\frac{C'(0)}{2}, \quad (28)$$

where the derivative is taken from above. They also point out that in human linkage analysis, depending on the one-sided nature of the test,  $X$  is asymptotically distributed as a 1/2:1/2 mixture of a chi-squared distribution and a single-point distribution at 0.

Assume that

$$\sum_w S(w) = 0, \sum_w S(w)^2 = 1 \quad (29)$$

and define

$$\Delta(h) = S(t+h) - S(t), \quad h > 0 \quad (30)$$

and then decompose the expression of the variance:

$$Var(\Delta(h)) = E(Var(\Delta(h)|S(t))) + Var(E(\Delta(h)|S(t))), \quad (31)$$

and now it is possible to simplify this formula further after working with the two inner expressions:

$$E(\Delta(h)|S(t) = w) = \lambda h \sum_{j=1}^{m_i} (S(w + e_j) - S(w)) + o(h), \quad (32)$$

$$\text{Var}(\Delta(h)|S(t) = w) = \lambda h \sum_{j=1}^{m_i} (S(w + e_j) - S(w))^2 + o(h), \quad (33)$$

inserting this into equation (31), noticing that  $\text{Var}(\Delta(-h)) = \text{Var}(\Delta(h))$ , letting  $h \rightarrow 0$  and using equation (28) and the following discussion one may then finally get:

$$\begin{aligned} 4 \cdot \rho &= \lim_{h \rightarrow 0} \frac{\text{Var}(\Delta(h))}{h} \\ &= \lambda \sum_{w \in Z_2^m} P(w) \sum_{j=1}^m (S(w + e_j) - S(w))^2 \\ &= \lambda \sum_{w' \in Z_2^m} \sum_{w \in Z_2^m} S(w') B(w', w) S(w), \end{aligned} \quad (34)$$

where  $\lambda$  in our case equals 1 because the genetic length is given in Morgans(0.01 if given in centiMorgans), the inheritance vector  $(w + e_j)$  is equal to the inheritance vector  $w$  in all but the  $j$ :th position where an addition with  $1 \pmod{2}$  has been performed,  $P(w) = P(v(t) = w|Y) = 2^{-m}$  under the null hypothesis of no linkage and under the same condition  $B(w', w)$  equals:

$$B(w', w) = \begin{cases} 2m \cdot 2^{-m} & , w' = w \\ -2 \cdot 2^{-m} & , |w' - w| = 1 \\ 0 & , |w' - w| > 1 \end{cases}, \quad (35)$$

where  $|w' - w|$  is the Hamming distance, which to use more explicit words means the number of positions where the values of the two inheritance vectors,  $w'$  and  $w$ , are distinct from each other.

To further complicate the picture we may consider the case when the data set consists of information about several different pedigrees and then we may proceed as follows:

Assume data from  $P$  different pedigrees. According to equation (16) we may rewrite the normalized two-locus conditional score function  $Z(\cdot)$  in the following way:

$$\begin{aligned} Z(x') &= \frac{\sum_{i=1}^P \gamma_i Z_i(x')}{\sqrt{\sum_{i=1}^P \gamma_i^2}} \\ &= \sum_{i=1}^P \epsilon_i Z_i(x'), \end{aligned} \quad (36)$$

where  $\gamma_i = \gamma(\bar{Z}_{1i}(\hat{x}))$  and  $\sum_{i=1}^P \epsilon_i^2 = 1$ . Then one may form the overall  $\rho$  - score using the given values of the weights  $\epsilon_i$ :

$$\rho_{\text{overall}} = \sum_{i=1}^P \epsilon_i^2 \rho_i, \quad (37)$$

where  $\rho_i$  is the  $\rho$  – score for the  $i$ :th pedigree.

### 4.3.2 Examples: Different Pedigree Structures

Now some examples of  $\rho$  – calculations, considering pedigrees of different structure and size, will be given. To be able to perform these calculations I have written a computer program using the m-file system of *MATLAB*. This program has also been used to perform these operations when dealing with the real-data analyses. The score function that has been used to complete the necessary calculations, within the computer program, is  $S_{all}$ . More details will be given in the next chapter.

Seven different cases have been considered. The results of the calculations are shown in the table found below. Examples number 1-4 replicates the results found by Lander/Kruglyak(1995) [11]. Pedigrees number 6-7 are examples of the pedigrees that were the source for the real-data analyses described in the result chapter below.

Ex.	Pedigree description	$\rho$ – value ( $S_{all}$ )
1	sib-pairs	2.0000
2	grandparent/grandchild	1.0000
3	uncle/nephew	2.5000
4	first cousins	2.6667
5	five affected siblings	2.0847
6	real pedigree example	2.5053
7	real pedigree example	2.1880

As pointed out above the  $\rho$  – value measures the fluctuation rate of the score function  $S(x)$ . This means, loosely speaking, that a high value of  $\rho$  will lead to a situation where it is easier for the approximated gaussian process to exceed a given value (the threshold  $T$ ) and therefore the  $p$  – value will be higher than in the case of a low value of  $\rho$ .

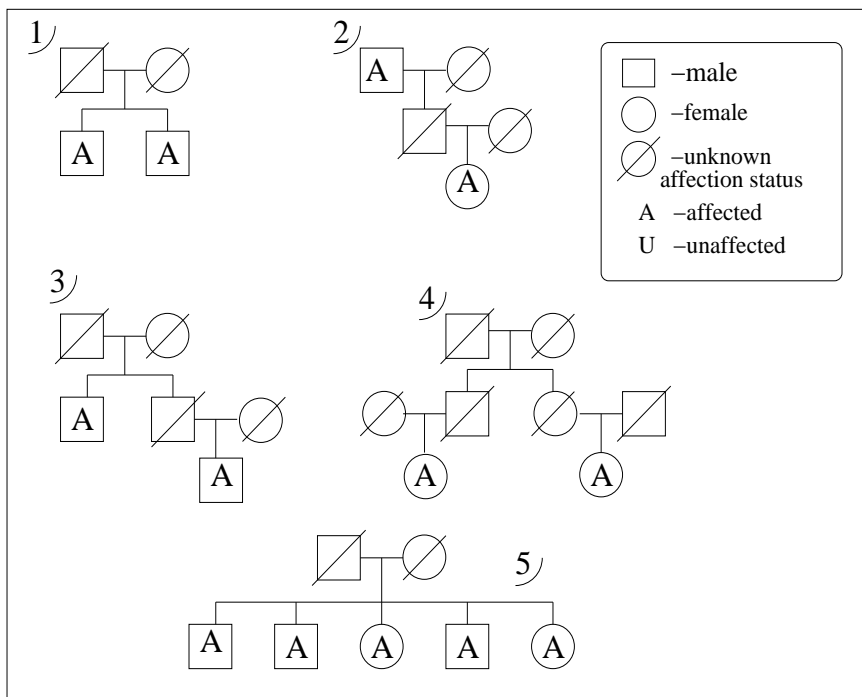


Figure 3: Pedigrees corresponding to the first five  $\rho$  – values given in the table above.

## 5 Results

### 5.1 Analyzing with Allegro

In the investigations that will be described below the computer program Allegro has been used (see [5]). To be able to perform analyses with that program three different files have to be created for each chromosome that one wants to investigate. The files may be written in different formats and the one that is used when describing the data set in this work is called the *Linkage* format (see [3]). The first two of these files have been created at the Institute of Endocrinology in Malmö, who kindly made them available for me to use in this work.

The *pedigree file* defines all the information needed about the pedigree structure and related issues. For each individual this involves information about *family number*, *individuals identification number*, *father identification number*, *mothers identification number*, *sex number*, *disease-status number* and *coded numbers for the alleles* present at the included markers. All this information except the family number and the individuals identification number may be coded as *unknown*.

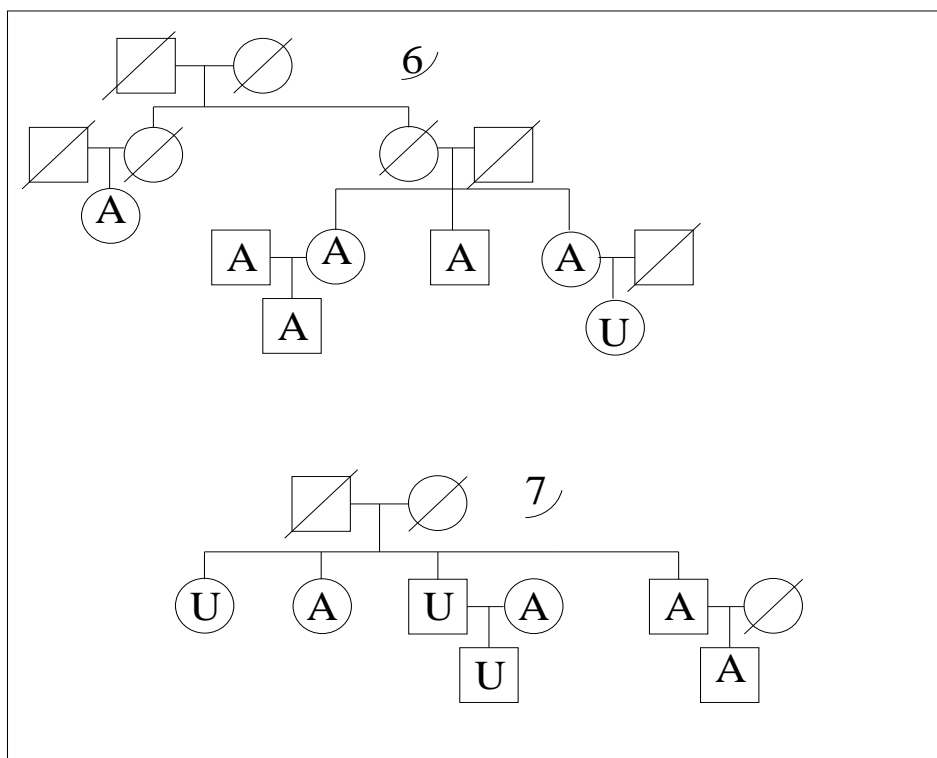


Figure 4: Pedigrees corresponding to the last two  $\rho$  – values given in the table above.

Specific information about the loci is supposed to be presented in the *data file*. Such information is for example *allele frequencies* at each marker, *recombination fractions* between markers etc.

The third file, the *options file*, defines the exact performance of the wanted investigation. This includes for example to define the *names of the two files* mentioned above, *the chosen score function*, *single-point/multipoint analysis*, *weighting schemes* (defined as *equal* to get the single-locus analysis). Several analyses may be performed, considering one specific chromosome, at the same time. The output will for each distinct investigation of a single chromosome produce two different data files. One of these files consists of, for example, the overall normalized NPL-scores at each present marker at the interesting chromosome and in the other file one may find the specific family scores at each of these markers. To get a look of an option file consider the following one that is a real example from the file that was written to perform the analyses, in the two-locus case, for chromosome 9. PREFILE and DATFILE are the pedigree and datafile that was mentioned above,

```

% read input(Linkage format)

PREFILE pedtest.dat
DATFILE chr9dat.dat

% analyses

MODEL mpt exp all propvikt4
MODEL mpt exp all 01vikt4
MODEL mpt exp all hetvikt4
MODEL mpt exp all propvikt7
MODEL mpt exp all 01vikt7
MODEL mpt exp all hetvikt7
MODEL mpt exp all propvikt12
MODEL mpt exp all 01vikt12
MODEL mpt exp all hetvikt12
MODEL mpt exp all propvikt16
MODEL mpt exp all 01vikt16
MODEL mpt exp all hetvikt16
MODEL mpt exp all propvikt18
MODEL mpt exp all 01vikt18
MODEL mpt exp all hetvikt18

% mpt: multipoint linkage analysis
% exp: allele sharing model, only
%       interesting when considering
%       non-parametric lod-scores.
% all: score function S(all)
% ex/ hetvikt18 : weighting scheme

MAXMEMORY 200

% memory that the program is set to be
% allowed to occupy.

```

The disease that we wanted to try to search for linkage to specific genome regions was *type 2 diabetes*. The present data set consisted of *2606 individuals* who belonged to *337 different families*. The families originated from Sweden and Finland.

Some more information that will be needed to make this presentation

complete is given in the table below. The lengths stated in the table are taken from a table in Ott (1999) [2] who himself got the information from Collins et al. (1996) [14]. The lengths were originally given in a sex-sensitive form but here they are presented as sex-averages of the male and female lengths.

Chr.	map length(cM)	Nbr.of markers
1	298.5	42
2	245.0	30
3	237.5	28
4	215.5	25
5	208.0	31
6	182.0	29
7	194.0	23
8	180.5	25
9	153.0	21
10	168.0	23
11	157.0	19
12	184.0	33
13	132.0	16
14	128.5	20
15	116.5	14
16	131.0	18
17	130.0	25
18	130.0	62
19	117.0	12
20	112.0	23
21	71.5	9
22	83.5	12

## 5.2 Single-Locus Analyses

First of all to be able to perform conditional two-locus NPL-analyses one needs to calculate the single locus NPL-scores for each interesting chromosome and according to some criteria decide which marker/markers family-score information to use when conditioning. Somewhat ad hoc we chose

to condition on the marker positions with the maximum NPL-score for each chromosome where that value exceeded  $1.75$ . Using that criteria it was found that the family scores at the appropriate markers at chromosomes  $4$ ,  $7$ ,  $12$ ,  $16$  and  $18$  should be used as weights in the two-locus analyses that would follow.

Chr.	$Z_1 - max$	Marker
1	1.3750	D1S2766
2	1.3376	D2S1776
3	1.2186	D3S3038
4	2.4329	D4S3248
5	0.6205	D5S1456
6	1.5123	D6S261
7	2.4789	D7S1808
8	1.5874	D8S1752
9	0.4578	D9S1830
10	0.6689	D10S1432
11	0.6120	D11S910
12	1.7627	D12S349
13	0.6722	D13S779
14	1.2762	D14S1060
15	0.1733	ATC3C11
16	2.2862	D16S419
17	1.5189	D17S1294
18	2.1319	D18S873
19	1.1649	D19S587
20	1.3323	D20S887
21	-0.2591	D21S258
22	1.0713	D22S274

### 5.3 Conditional Two-Locus Analyses

To perform these conditional two-locus NPL-analyses we conditioned on the information contained at the five specific markers mentioned above. Three different weighting schemes were used (see equation (12)) called *prop* (proportional epistatic weight), *01* (epistatic weight) and *het* (heterogeneity weight). The

ten marker combinations with the highest NPL-score are presented in the table below.

A complete list of all the results found when performing these analyses are included in the *appendix*. The first part of the appendix consists of all the results presented in the form of graphical figures and the second part consists of the same results presented in a mainly numerical form.

Pos.	Chr.	Weight	Cond.on	$Z_2 - max$
1	7	prop	16	3.8658
2	19	prop	16	3.0901
3	12	01	18	2.8694
4	7	prop	12	2.8286
5	17	prop	16	2.8199
6	4	het	16	2.7189
7	7	het	16	2.6632
8	4	01	7	2.6230
9	17	het	7	2.5665
10	2	prop	12	2.5270

#### 5.4 Finding the $\rho$ - value/ $p$ - value

In the first subsection of this chapter the theory that was described in the last chapter will be applied to our data set and therefore make it possible for us to calculate approximations of the p-values. The second subsection will describe the procedure of using so called Monte Carlo simulations to be able to calculate the  $\rho$ -value in case of very large pedigrees. Some further properties of the  $\rho$ -parameter will be discussed as well.

##### 5.4.1 Calculations

The highest NPL-score that we found was, as shown above, 3.8658. When removing a single family (called 1065.0) that possibly shouldn't be included in the study (see the next subsection about *robustness*) the highest score decreased to 2.8694.

To be able to find some sort of information about any possible significance value of the possibility of finding such high scores under the null hypothesis of no linkage one may calculate the p-value according to the theory presented in the preceding chapter. Using equations (34) and (37) one may calculate

the appropriate value of the  $\rho$  – *parameter* for this specific data set. This value will then combined with equations (25) and (26) make it possible to calculate the p-values for each specific test. The results of these calculations are shown in the table below.

Nbr.	Cond.on	Weight	G (cM)	$\rho$ – <i>value</i>	$p : T = 3.8658$	$p : T = 2.8694$
1	4	prop	3357.5	2.3030	0.1211	0.9311
2	4	01	3357.5	2.1213	0.1122	0.9141
3	4	het	3357.5	2.0175	0.1071	0.9034
4	7	prop	3379.0	2.1107	0.1124	0.9144
5	7	01	3379.0	2.1043	0.1120	0.9138
6	7	het	3379.0	2.0335	0.1085	0.9065
7	12	prop	3389.0	2.0895	0.1116	0.9129
8	12	01	3389.0	2.0916	0.1117	0.9131
9	12	het	3389.0	2.0472	0.1095	0.9086
10	16	prop	3442.0	2.1410	0.1158	0.9210
11	16	01	3442.0	2.0889	0.1132	0.9161
12	16	het	3442.0	2.0460	0.1110	0.9118
13	18	prop	3443.0	2.0603	0.1118	0.9133
14	18	01	3443.0	2.0841	0.1130	0.9157
15	18	het	3443.0	2.0509	0.1113	0.9123

Using these results and equation 27 to correct for multiple testing one finally finds that in this situation and using this theory it is not possible to draw any significant conclusions about linkage to any combination of regions, because ...

$$P_{overall} = P(\bar{Z}_{top} \geq \bar{z}_{top}) \leq \min(\sum_{i=1}^{15} p_i, 1) = 1 \quad (38)$$

This holds for both of the distinct  $T$  – *values* but there is still a big difference in the results if one decides to include or exclude the extreme family in the study. The large p-values in the rightmost column of the table above reflect the fact that the asymptotic approximation (25) becomes bad for small values of  $\bar{z}_{max}$ .

#### 5.4.2 $\rho$ : Monte Carlo Simulations and Further Properties

For pedigrees where the total number of meioses,  $m$ , exceed 20 the computational complexity gets so high that the calculations, at least with my

computer program, can not be performed within a reasonable time limit. This forces us to use Monte Carlo simulations for these pedigrees (13 families).

Notice first that under the null hypothesis,  $H_0$ , of no linkage (34)-(35) may be rewritten as:

$$4 \cdot \rho = \lambda \cdot 2^{-m} \sum_{w \in Z_2^m} \sum_{j=1}^m (S(w + e_j) - S(w))^2 \quad (39)$$

$$= \lambda \cdot E(f(v)), \quad (40)$$

where  $f(w) = \sum_{j=1}^m (S(w + e_j) - S(w))^2$  and  $v$  is random with  $P(v = w) = 2^{-m}$  for each  $w \in Z_2^m$ . A proper Monte Carlo approximation of  $\rho$  is thus:

$$\hat{\rho} = \frac{\lambda}{4} \frac{1}{U} \sum_{i=1}^U f(v_i), \quad (41)$$

where  $\{v_i\}_{i=1}^U$  are i.i.d. with the same distribution as  $v$  and  $U$  is the number of simulated inheritance vectors.

We picked  $U=1000$  inheritance vectors at random. The table below describes the fluctuations of the outcomes of the  $\rho$ -calculations for the two subpopulations of pedigrees for which exact calculations and Monte Carlo simulations was used respectively.

Calc.	Nr of ped.	$\rho$ -max	$\rho$ -min(>0)	$\rho$ -min	mean	std
Exact	324	3.1069	1.0000	0.0000	2.0246	0.4031
Monte Carlo	13	3.5620	2.0348	2.0348	2.6208	0.4868
Total	337	3.5620	1.0000	0.0000	2.0476	0.4218

To get a better view of the deviations of the outcomes of the  $\rho$ -calculations (total) consider figure 5. The figure shows plots of  $\rho$  against both the variables  $m$  and  $n$ . It seems to be some positive correlation between the variables in both of these cases.

The exact values of the estimated correlations between  $\rho$  and the number of meioses and the number of individuals in a pedigree respectively are shown in the small table below. Because of the obvious high correlation between  $m$  and  $n$  it might be no surprise that these two correlations almost coincide.

The reason why the  $\rho$ -values for some pedigrees equals 0 is that in these specific cases the value of the score function is independent of the inheritance

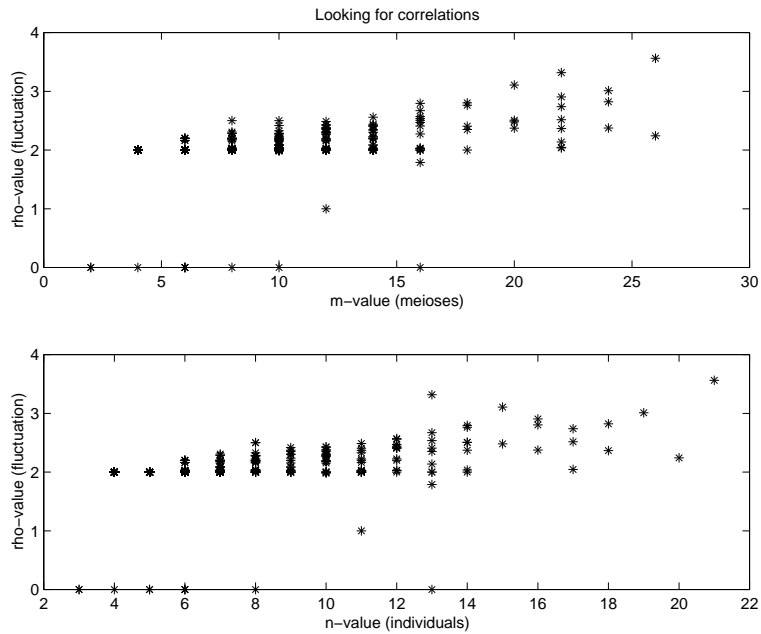


Figure 5: Plotting  $\rho$  against  $m$  and  $n$ .

vectors. This means that the score function equals the same value for all the inheritance vectors and may therefore, in this case, be seen as a constant. Using (34) one may easily draw the conclusion that  $\rho$  in this case will be equal to zero. To get a simple example of a pedigree structure that leads to this situation one may for example consider a family only consisting of two unaffected parents and one affected child.

Var.1	Var.2	Correlation
$\rho$	m	0.4115
$\rho$	n	0.4140
m	n	0.9738

## 5.5 Robustness

In this subsection discussions about the issue of *robustness* will be given. This will involve discussions about using different score functions, family outliers with extreme influence on the overall NPL-score and comparisons between using different weighting schemes.

### 5.5.1 Subject: Different Weights and Outliers

In figure 6 it is possible to see that, when one plots the family scores from the marker combinations that produces the two highest NPL-scores, one family is an extreme outlier at both marker positions in both of the combinations. In fact, that outlier is the same family in both cases. It is called family number 1065.0. According to the opinions of the people at the *Department of Endocrinology* in Malmö there may be some biological reason/ genotype error that makes this family inappropriate to include and use in a linkage study.

To see how much that single family influenced the overall NPL-score all the conditional two-locus NPL-investigations were performed once more, this time with the family 1065.0 removed from the data set. The top-ten results are presented in the table below.

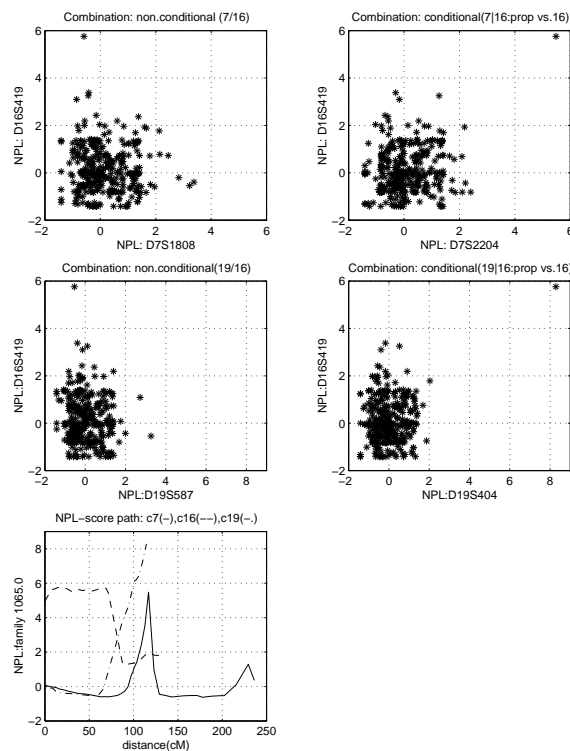


Figure 6: Plotting the family scores at the two markers that formed the combination for the maximum two-locus NPL-score at chromosomes 7/19 (conditional) and the combination of the family scores at the markers where one found the maximum single-locus NPL-scores at these chromosomes (non-conditional).

Pos.	Chr.	Weight	Cond.on	$Z_2 - max$	Marker
1	12	01	18	2.8694	D12S84
2	7	prop	12	2.8286	D7S1808
3	4	het	16	2.7189	D4S2367
4	7	prop	16	2.6966	D7S1830
5	18	prop	16	2.6793	D18S976
6	7	het	16	2.6632	D7S503
7	4	01	7	2.6230	D4S2361
8	18	het	7	2.5275	D18S976
9	2	prop	12	2.5270	D2S1400
10	18	prop	4	2.5262	D18S63

In this table we can see that the overall highest score decreased from 3.8658 to 2.8694 and that the combination that produced the highest score in the first round in fact fell to 2.6966 and the correlation between the family scores in that case changed from 0,202 to 0.097. Considering the combination that produced the second highest score in the first genomewide linkage investigation, 3.0901, the NPL-score decreased all the way down to 0.3424.

The possibility to get combinations of family scores with extreme outliers with such remarkable influence on the total NPL-score is depending on, for example, the choice of score function, the choice of weighting scheme and possible typing/measuring errors.

The *proportional weight* has both advantages and disadvantages. On the positive side it will in a situation of real linkage with positive epistasis be a better tool than the other present weights in extracting information from the data set and finding the genome positions of linkage. This depends on the fact that in this situation one do not only divide the family scores from the single locus case into one interesting group that is used in the further two-locus investigations and one uninteresting group that is not included in that investigations, but rather also take into account the exact individual value of each family score in the interesting group. This might also be a great backlash to the proportional weight because it makes it possible for single family NPL-scores to get an inappropriate influence of the overall NPL-score, which in fact means that the weight is lacking in robustness.

A kind of informal criteria if a high NPL-score with the proportional weight is trustworthy is the performance of the corresponding *01-weight*. If

that weight shows a quite high NPL-score that might be a sign of the proportional weight indicating a possible real linkage and if there is a big discrepancy between the high proportional weight NPL-score and a low corresponding 01-weight NPL-score that might indicate that the high NPL-score in the former case depends on the extreme influence of single family scores. Of course one always, even in the case of a relatively high 01-weight NPL-score, also has to, for example, produce scatter plots and make sure that there are not one or two individual families that makes the whole difference between the 01-weight situation and the proportional-weight situation. In that case the extreme families perhaps should be removed or the proportional-weight situation may be considered uninteresting.

### 5.5.2 Subject: Score Functions

One might also consider the possibility of using different kind of score functions when one discusses the issue of robustness. For the same two combinations of family scores as in the last subsection, the scores that produced the two highest overall NPL-scores in the original case, we performed the investigation once more and this time the score function  $S_{pairs}$  (see equation 2) was used. The results are presented in the table below and in figure 7.

Analysis	Chr.	Weight	Cond.on	NPL-pairs	NPL-all
Single-locus	16	-	-	2.2365(D16S419)	2.2862(D16S419)
Two-locus	7	prop	16	2.8689(D7S1830)	3.8658(D7S2204)
Two-locus	7	01	16	2.3928(D7S1830)	2.1491(D7S1830)
Two-locus	7	het	16	2.2948(D7S503)	2.6632(D7S503)
Two-locus	19	prop	16	1.4160(D19S418)	3.0901(D19S404)
Two-locus	19	01	16	1.3760(D19S178)	1.1679(D19S418)
Two-locus	19	het	16	2.0933(D19S1034)	1.7642(D19S1034)

Kruglyak et al. (1996) [4] point out that  $S_{all}$  may extract more information from the data set than  $S_{pairs}$  and therefore produces better results in most situations. This is probably true but in some situations, with quite large families of certain pedigree structure, the  $S_{all}$  function seems to, see equation (3), be very non-robust and therefore it rather easily, by chance, produces extremely influential outlier family scores. In figure 8 one can see that the extreme family scores are being held back when considering the

$S_{pairs}$  weight and therefore they will have less extreme influence on the overall NPL-score. This can be seen in the table above, where the NPL-scores in the proportional-weights situation have decreased.

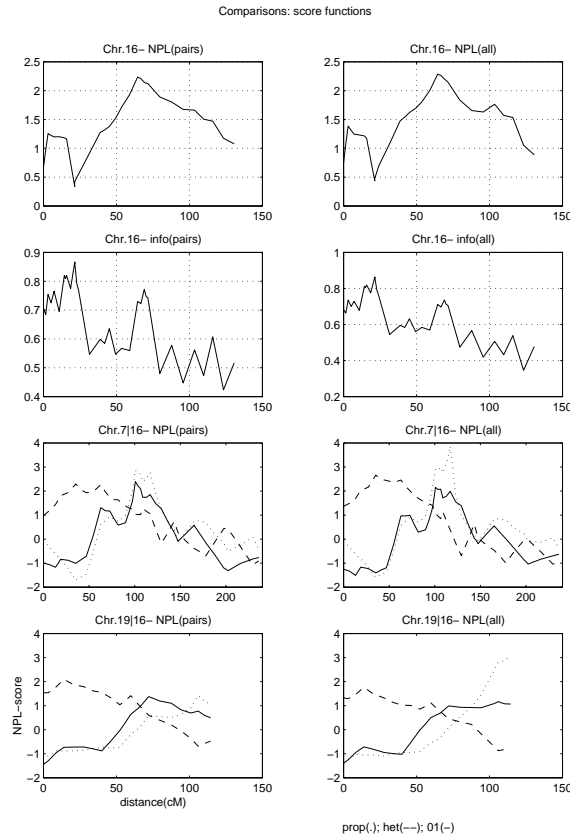


Figure 7: Comparisons  $S_{all}$  vs.  $S_{pairs}$ .

## 5.6 Conditioning on Regions Suggested by Meta-Analysis

Finally, the Institute of Endocrinology gave us the information about five different regions that had been pinpointed by a *meta analysis*<sup>7</sup> and asked us to perform conditional two-locus NPL-analyses conditioning on markers related to these regions. The present regions may be seen in the table below. The specific markers that were used in the investigations were chosen to be the ones closest to the middle of the defined regions.

<sup>7</sup>A meta analysis is an analysis that uses the merged information and results from several different previous investigations. See Wise et al. [21].

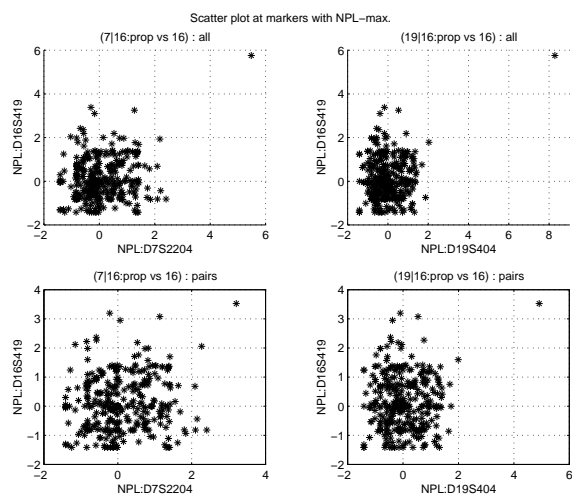


Figure 8: Plotting family scores to compare robustness.

Chr.	Region (cM)	Marker	NPL(single locus)
2	60-90	D2S441	-0.30
6	120-150	D6S261	1.51
12	120-150	D12S332	0.45
16	30-60	D16S769	1.62
17	30-60	D17S799	0.82

We performed these conditional two-locus NPL-analyses, ranging over all the 22 autosomes, and the highest score found was 3.1849. The top-ten list of the highest scores found during these investigations is presented in the table below.

Pos.	Chr.	Weight	Cond.on chr.	$Z_2 - max$	Marker
1	7	het	17	3.1849	D7S1808
2	18	prop	12	2.8075	D18S976
3	4	het	16	2.7848	D4S2361
4	7	het	6	2.7184	D7S1818
5	7	prop	16	2.6037	D7S1830
6	16	prop	17	2.4841	D16S3253
7	12	prop	2	2.3882	D12S395
8	8	het	6	2.3819	D8S1752
9	6	01	17	2.3787	D6S1581
10	4	het	6	2.3772	D4S1643

The  $\rho$ -values and the p-values were then calculated. As before, no significant results were found. The information that was needed to calculate these values, using the theory described in chapter 4, is presented in the following table...

Nbr.	Cond.on	Weight	G (cM)	$\rho - value$	$p : T = 3.1849$
1	2	prop	3328	2.0999	0.6471
2	2	01	3328	2.1127	0.6493
3	2	het	3328	2.0374	0.6362
4	6	prop	3391	2.1171	0.6569
5	6	01	3391	2.1147	0.6565
6	6	het	3391	2.0219	0.6402
7	12	prop	3389	2.1076	0.6550
8	12	01	3389	2.1304	0.6589
9	12	het	3389	2.0251	0.6406
10	16	prop	3442	2.0743	0.6549
11	16	01	3442	2.0944	0.6584
12	16	het	3442	2.0418	0.6492
13	17	prop	3443	2.1161	0.6622
14	17	01	3443	2.1156	0.6621
15	17	het	3443	2.0016	0.6421

## 6 Summary

In this work the ideas described in the Master's thesis written by Markus Kämpe (2001) [8] have been applied to a real data set and some further theory have been described and applied to that same set of pedigrees.

The data set consisted of 337 families which in total contained 2606 individuals from Sweden and Finland. The disease of interest was type 2 diabetes.

Conditional two-locus non-parametric linkage analysis was performed throughout all the 22 autosomes. Significance values, p-values, were calculated according to theory derived by Ola Hössjer (2001) [10] and partially described in this work, but no significant results were found.

Further, we discussed robustness according to different score functions, using different kind of weights, dealing with extreme family outliers etc. Appropriate comparisons were graphically presented and discussed. In some situations the proportional-weight and the score function  $S_{all}$  may be very non-robust and this will, somewhat self-explained, especially occur when these two situations coincide. It seems to be a quite trustworthy positive correlation between the crossover rates and the number of meioses related to the pedigrees,  $m$ , and the number of individuals in the pedigrees,  $n$ , respectively.

Conditional two-locus NPL-analyses were performed conditioning on five different markers related to five predefined interesting regions, but no significant results were found.

Some notes for interesting further research will finally be proposed:

- New theoretical methods that make it possible to search for more regions of susceptibility at the same time <sup>8</sup>, three-locus analysis, four-locus analysis and so on.
- Perhaps new methods that condition on, for example, haplotypes instead of family scores [19].
- To complement the asymptotic p-values based on normal approximations with *Monte Carlo* approximations of the exact genomewide p-values. The normal approximation is likely to perform worse when some outlying family scores have large impact on the overall NPL-score.

---

<sup>8</sup>A big problem is of course, as usual, the issue of having to correct the p-values according to multiple testing which makes it hard to find significant results

## A Appendix: Graphs and Results

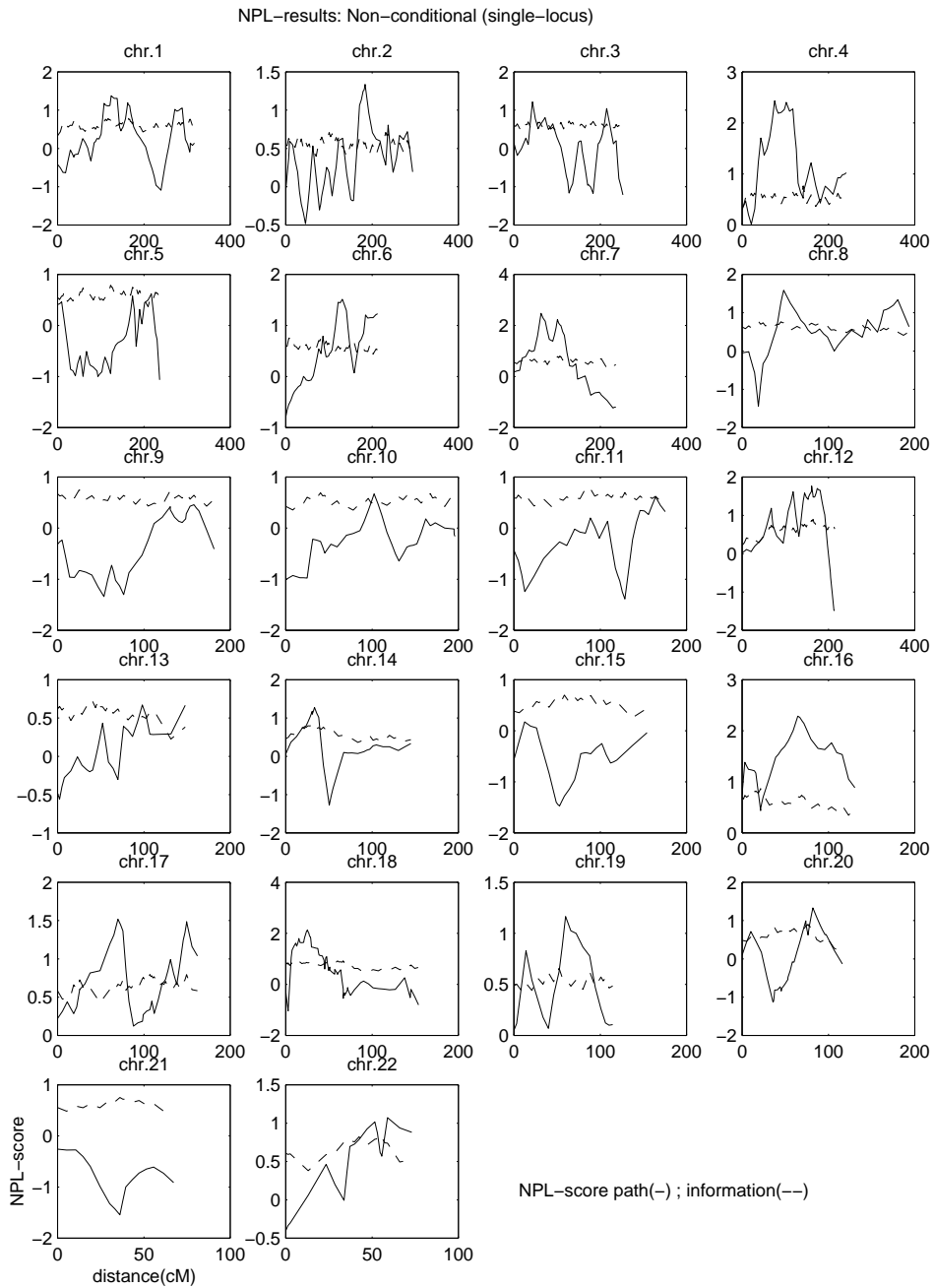


Figure 9: Results: single-locus analyses.

NPL-results: Conditional chr.4

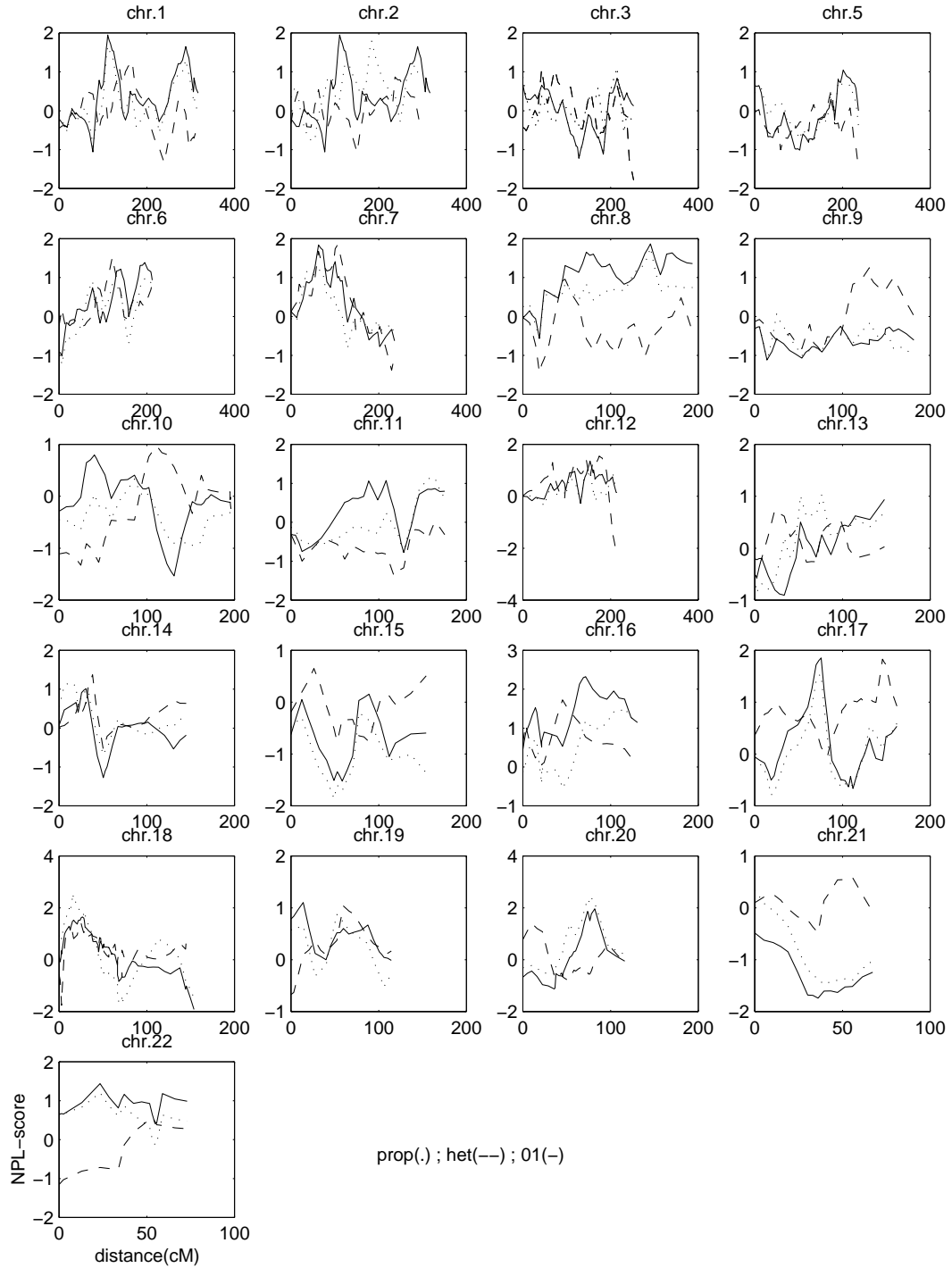


Figure 10: Results: conditioning on chr.4<sub>max</sub>.

NPL-results: Conditional chr.7

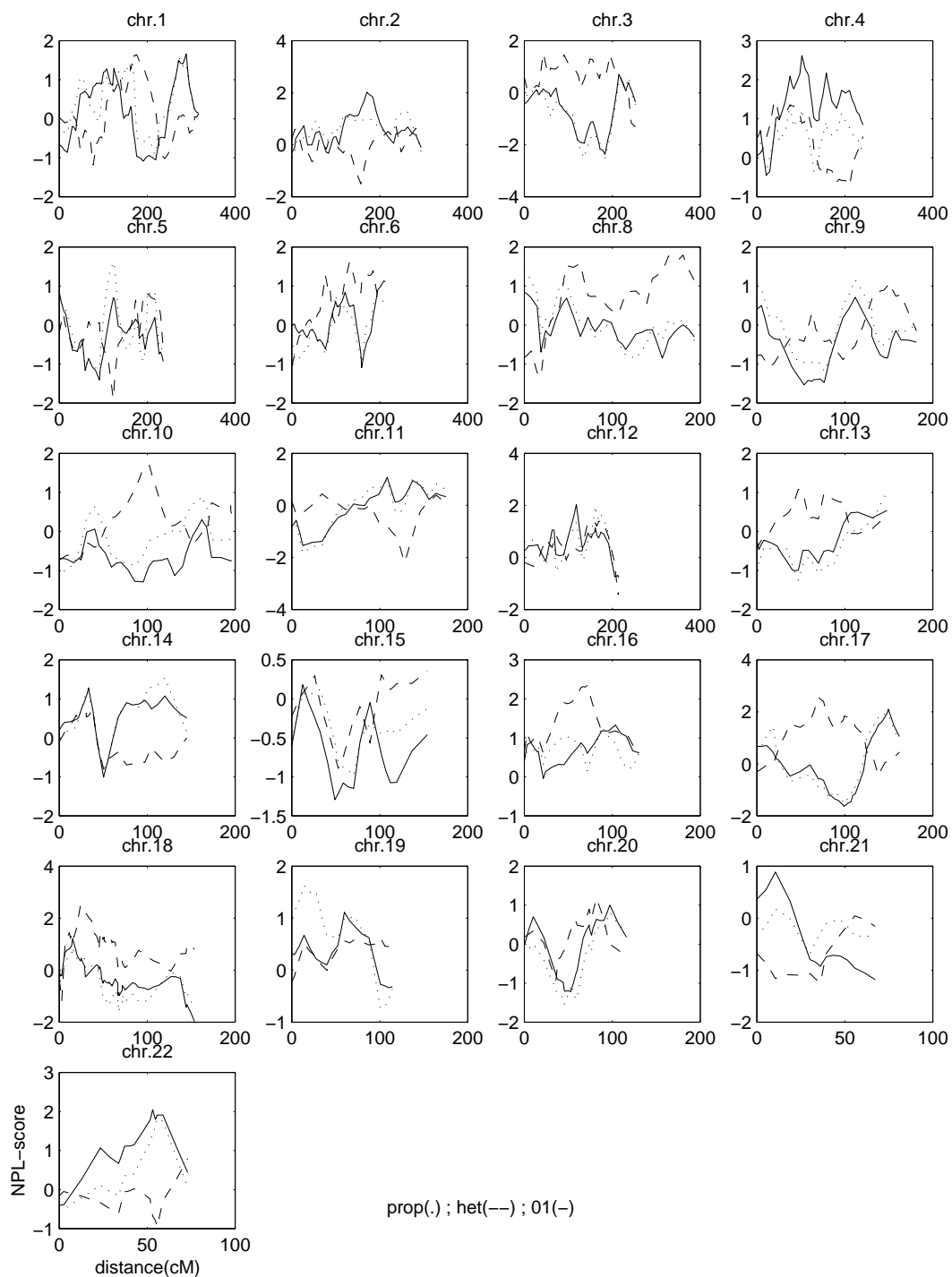


Figure 11: Results: conditioning on  $chr.7_{max}$ .

NPL-results: Conditional chr.12

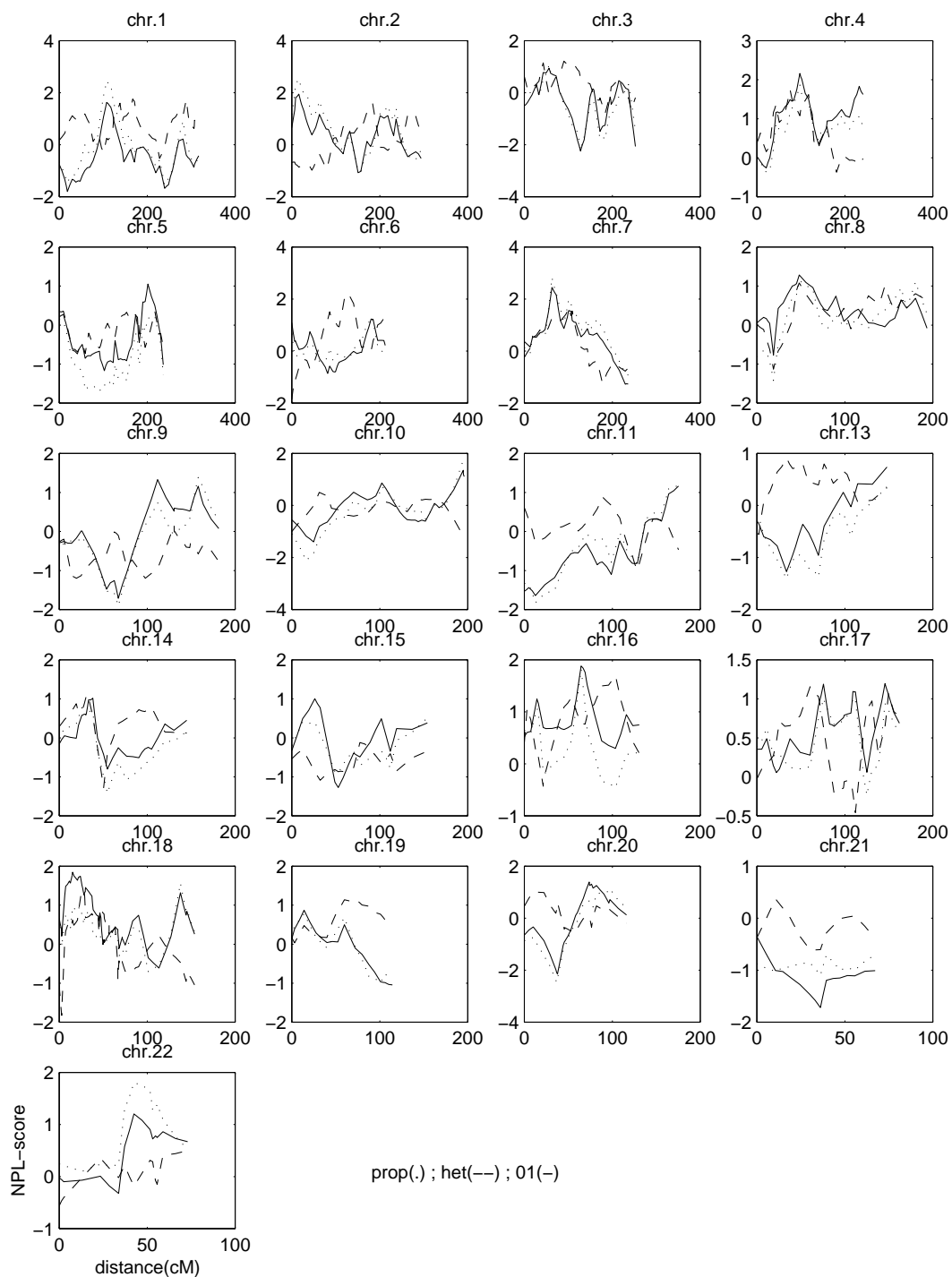


Figure 12: Results: conditioning on chr.12<sub>max</sub>.

NPL-results: Conditional chr.16

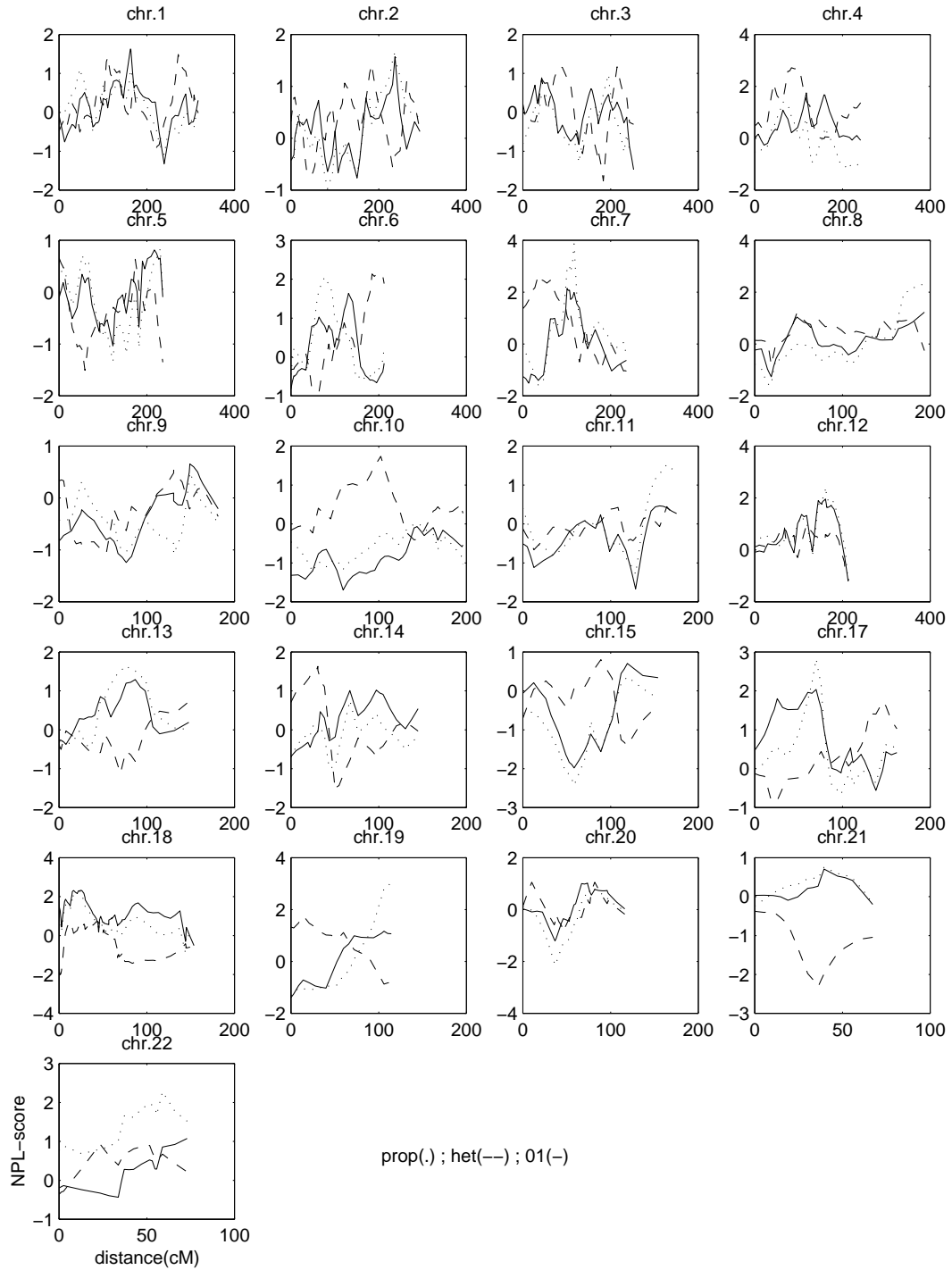


Figure 13: Results: conditioning on chr.16<sub>max</sub>.

NPL-results: Conditional chr.18

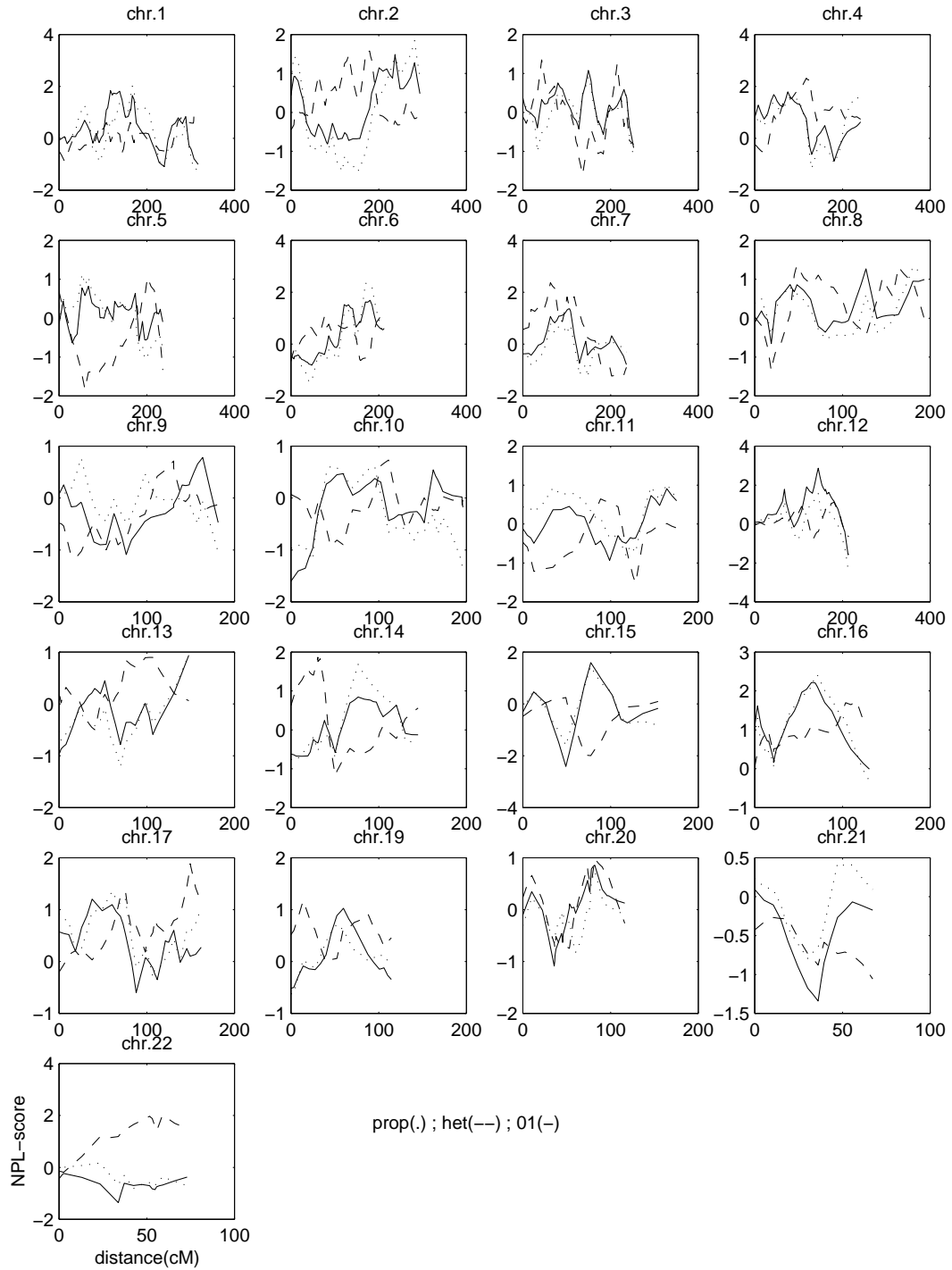


Figure 14: Results: conditioning on chr.18<sub>max</sub>.

Complete list of all conditional two-locus NPL-results:

Chromosome	Cond.chr.	Weight	NPL-max.	Marker pos.
1	4	prop	1.6573	D1S464
	4	01	1.9495	D1S464
	4	het	1.2618	D1S498
	7	prop	1.6020	D1S549
	7	01	1.6608	D1S3462
	7	het	1.6555	D1S498
	12	prop	2.4529	D1S1665
	12	01	1.6305	D1S1665
	12	het	1.7556	D1S498
	16	prop	1.1170	D1S199
	16	01	1.6258	D1S453
	16	het	1.4903	D1S549
	18	prop	2.0098	D1S498
	18	01	1.8399	D1S1728
	18	het	1.1375	D1S1609
2	4	prop	1.7915	D2S1776
	4	01	2.0457	D2S1776
	4	het	0.8675	D2S434
	7	prop	1.2558	D2S427
	7	01	2.0224	D2S1353
	7	het	0.6904	D2S434
	12	prop	2.5270	D2S1400
	12	01	1.9475	D2S1400
	12	het	1.5779	D2S1776
	16	prop	1.6681	D2S434
	16	01	1.5709	D2S434
	16	het	1.4264	D2S1776
	18	prop	1.8409	D2S338
	18	01	1.4771	D2S434
	18	het	1.5772	D2S1395
3	4	prop	1.1058	D3S2427
	4	01	0.8370	D3S2427
	4	het	1.0785	D3S3038

	7	prop	0.5758	D3S2398
	7	01	0.7075	D3S2427
	7	het	1.5259	D3S3038
	12	prop	1.0622	D3S2432
	12	01	0.9445	D3S2432
	12	het	1.2174	D3S1287
	16	prop	0.9754	D3S1275
	16	01	0.8576	D3S3038
	16	het	1.1689	D3S2427
	18	prop	0.9087	D3S2460
	18	01	1.0782	D3S2460
	18	het	1.3506	D3S3038
4	7	prop	1.3511	D4S3248
	7	01	2.6230	D4S2361
	7	het	1.4611	D4S1643
	12	prop	1.8654	D4S3243
	12	01	2.1611	D4S3243
	12	het	1.7526	D4S3248
	16	prop	1.5307	D4S1627
	16	01	1.7507	D4S1647
	16	het	2.7189	D4S2367
	18	prop	1.5952	D4S1540
	18	01	1.7907	D4S3248
	18	het	2.3209	D4S1647
5	4	prop	0.7342	D5S636
	4	01	1.0458	D5S1465
	4	het	0.5114	D5S816
	7	prop	1.6127	D5S428
	7	01	0.8460	D5S2488
	7	het	0.8048	D5S1465
	12	prop	0.2319	D5S1456
	12	01	1.0514	D5S1465
	12	het	0.3579	D5S1456
	16	prop	0.8335	D5S2111
	16	01	0.8130	D5S1456
	16	het	0.6601	D5S2488
	18	prop	1.1326	D5S819
	18	01	0.8194	D5S395

	18	het	1.0575	D5S1465
6	4	prop	0.9714	D6S1277
	4	01	1.3849	D6S1277
	4	het	1.5248	D6S1021
	7	prop	0.7663	D6S1270
	7	01	1.1366	D6S503
	7	het	1.6101	D6S261
	12	prop	1.2946	D6S1007
	12	01	1.2305	D6S1007
	12	het	2.1868	D6S1021
	16	prop	2.0531	GATA11E02
	16	01	1.6360	D6S261
	16	het	2.1337	D6S1581
	18	prop	2.3676	D6S2436
	18	01	1.6852	D6S1007
	18	het	1.2351	D6S503
7	4	prop	1.1938	D7S1808
	4	01	1.8378	D7S1808
	4	het	1.8484	D7S502
	12	prop	2.8286	D7S1808
	12	01	2.4464	D7S1808
	12	het	1.6899	D7S1830
	16	prop	3.8658	D7S2204
	16	01	2.1491	D7S1830
	16	het	2.6632	D7S503
	18	prop	0.7918	D7S1808
	18	01	1.3611	D7S502
	18	het	2.3673	D7S1808
8	4	prop	1.7655	D8S592
	4	01	1.8668	D8S592
	4	het	0.9581	D8S1752
	7	prop	1.4047	D8S264
	7	01	0.8440	D8S264
	7	het	1.7943	D8S1100
	12	prop	1.1460	D8S1752
	12	01	1.2833	D8S1752
	12	het	1.0799	D8S1752

	16	prop	2.3613	D8S373
	16	01	1.2276	D8S373
	16	het	1.2085	D8S1752
	18	prop	1.3202	D8S1100
	18	01	1.2679	GAAT1A4
	18	het	1.3613	D8S1752
9	4	prop	0.0606	D9S157
	4	01	-0.2540	D9S257
	4	het	1.2510	D9S1675
	7	prop	1.1568	D9S910
	7	01	0.7146	D9S910
	7	het	1.0933	D9S1825
	12	prop	1.3840	D9S1830
	12	01	1.3277	D9S910
	12	het	0.1671	D9S1675
	16	prop	0.4899	D9S1825
	16	01	0.6548	D9S1825
	16	het	0.5340	D9S930
	18	prop	0.7678	D9S157
	18	01	0.7830	D10S1818
	18	het	0.7156	D9S1675
10	4	prop	0.3303	D10S1225
	4	01	0.7980	D10S674
	4	het	0.9693	D10S2327
	7	prop	0.9041	D10S1230
	7	01	0.3064	D10S1230
	7	het	1.6977	D10S1432
	12	prop	1.7469	D10S169
	12	01	1.3501	D10S169
	12	het	0.5014	D10S2325
	16	prop	0.0118	D10S1795
	16	01	-0.0655	D10S1795
	16	het	1.7358	D10S1432
	18	prop	0.6149	D10S1423
	18	01	0.5426	D10S1230
	18	het	0.7305	D10S2327
11	4	prop	1.1356	D11S910

	4	01	1.0709	D11S2002
	4	het	0.0188	D11S910
	7	prop	1.0541	D11S1998
	7	01	1.0851	D11S2002
	7	het	0.4450	D11S1981
	12	prop	1.1965	D11S968
	12	01	1.1613	D11S968
	12	het	0.8868	D11S987
	16	prop	1.5256	D11S910
	16	01	0.4635	D11S912
	16	het	0.4423	D11S910
	18	prop	1.0290	D11S910
	18	01	0.9066	D11S910
	18	het	0.6456	D11S987
12	4	prop	0.9864	D12S2070
	4	01	1.3560	D12S2070
	4	het	1.5539	D12S342
	7	prop	1.9696	D12S349
	7	01	2.0370	D12S1064
	7	het	1.4124	D12S342
	16	prop	2.4142	D12S349
	16	01	1.9627	D12S349
	16	het	0.9802	D12S1064
	18	prop	1.5653	D12S84
	18	01	2.8694	D12S84
	18	het	1.1257	D12S340
13	4	prop	1.0315	D13S767
	4	01	0.9392	D13S285
	4	het	0.7476	GGAA29H03
	7	prop	1.0094	D13S285
	7	01	0.5402	D13S285
	7	het	1.0793	D13S788
	12	prop	0.3592	D13S285
	12	01	0.7375	D13S285
	12	het	0.9158	D13S894
	16	prop	1.6490	D13S767
	16	01	1.2913	D13S786
	16	het	0.7236	D13S285

	18	prop	0.9514	D13S285
	18	01	0.9385	D13S285
	18	het	0.8980	D13S779/D13S797
14	4	prop	1.1747	D14S50/D14S264
	4	01	1.0146	D14S1071/D14S1040
	4	het	1.3697	D14S599
	7	prop	1.5402	D14S749
	7	01	1.2852	D14S1060
	7	het	0.7318	D14S599
	12	prop	0.7730	D14S597
	12	01	1.0154	D14S599
	12	het	1.0688	D14S1071
	16	prop	0.7541	D14S592
	16	01	1.0221	D14S606
	16	het	1.6319	D14S1071
	18	prop	1.6862	D14S588
	18	01	0.8432	D14S588
	18	het	1.8651	D14S1071
15	4	prop	-0.2050	D15S131
	4	01	0.1531	D15S205
	4	het	0.6502	D15S144
	7	prop	0.1220	D15S118
	7	01	0.1811	ATC3C11
	7	het	0.3610	D15S642
	12	prop	0.5122	D15S642
	12	01	1.0061	D15S144
	12	het	-0.1197	D15S131
	16	prop	0.3628	D15S657
	16	01	0.7051	D15S657
	16	het	0.8020	D15S205
	18	prop	1.4491	D15S131
	18	01	1.5939	D15S131
	18	het	0.2459	D15S659
16	4	prop	1.5162	D16S422
	4	01	2.3218	D16S3253
	4	het	1.7271	D16S769
	7	prop	1.2827	D16S2624

	7	01	1.3274	D16S516
	7	het	2.3468	D16S3253
	12	prop	1.8192	D16S419
	12	01	1.8810	D16S419
	12	het	1.6985	D16S516
	18	prop	2.4075	D16S3253
	18	01	2.2165	D16S419/D16S771
	18	het	1.7435	D16S422
17	4	prop	1.5284	D17S1293
	4	01	1.8522	D17S1293
	4	het	1.8289	D17S836
	7	prop	1.9290	D17S836
	7	01	2.1028	D17S1830
	7	het	2.5665	D17S1294
	12	prop	1.0137	D17S1835
	12	01	1.1958	D17S836
	12	het	1.1565	D17S122
	16	prop	2.8199	D17S1294
	16	01	2.0382	D17S1294
	16	het	1.7266	D17S836
	18	prop	1.3533	D17S122
	18	01	1.2034	D17S799
	18	het	1.8847	D17S1830
18	4	prop	2.4890	D18S63
	4	01	1.6536	FB25F12
	4	het	1.4180	D18S976
	7	prop	0.8057	D18S459
	7	01	1.2117	D18S459
	7	het	2.4814	D18S976
	12	prop	1.5620	D18S485
	12	01	1.8475	D18S459
	12	het	1.3735	FB25F12
	16	prop	2.2657	D18S976
	16	01	2.2394	D18S873
	16	het	0.7696	D18S873
19	4	prop	0.8727	D19S433
	4	01	1.1045	D19S1034

	4	het	1.0349	D19S587
	7	prop	1.6214	D19S1034
	7	01	1.1147	D19S587
	7	het	0.6344	D19S589
	12	prop	0.7783	D19S1034
	12	01	0.8711	D19S1034
	12	het	1.1717	D19S178
	16	prop	3.0901	D19S404
	16	01	1.1679	D19S418
	16	het	1.7642	D19S1034
	18	prop	0.7205	D19S433
	18	01	1.0252	D19S587
	18	het	1.1803	D19S1034
20	4	prop	2.3914	D20S891
	4	01	1.9632	D20S887
	4	het	1.4025	D20S116
	7	prop	0.8104	D20S469
	7	01	1.0013	D20S469
	7	het	1.1823	D20S887
	12	prop	1.1261	D20S469
	12	01	1.3918	D20S119
	12	het	1.0239	D20S116
	16	prop	0.7847	D20S887
	16	01	1.0213	D20S119
	16	het	1.0579	D20S887
	18	prop	0.7345	D20S887
	18	01	0.8589	D20S887
	18	het	0.9831	D20S887
21	4	prop	0.3442	D21S258
	4	01	-0.4838	D21S258
	4	het	0.5908	D21S1260
	7	prop	0.1760	D21S1441
	7	01	0.8865	D21S1441
	7	het	0.0452	D21S1260
	12	prop	-0.7021	D21S1446
	12	01	-0.3539	D21S258
	12	het	0.3703	D21S1441
	16	prop	0.7575	D21S1440

	16	01	0.7108	D21S1440
	16	het	-0.3777	D21S258
	18	prop	0.4276	D21S156
	18	01	0.0922	D21S258
	18	het	-0.2670	D21S1441
22	4	prop	1.2016	D22S268
	4	01	1.4400	D22S268
	4	het	0.5848	D22S423
	7	prop	1.8171	D22S1140
	7	01	2.0472	D22S423
	7	het	0.7924	D22S1169
	12	prop	1.8012	D22S683
	12	01	1.2057	D22S683
	12	het	0.4997	D22S1169
	16	prop	2.2754	D22S274
	16	01	1.0747	D22S1169
	16	het	0.9302	D22S268
	18	prop	0.1703	D22S268
	18	01	-0.2339	D22S420
	18	het	2.0223	D22S274

## References

- [1] Sham, P. *Statistics in Human Genetics*. Arnold Applications of Statistics, 1998.
- [2] Ott, J. *Analysis of Human Genetic Linkage*, 3:rd edition. The John Hopkins University Press, 1999.
- [3] Terrwilliger, J.D. and Ott, J. *Handbook of Human Genetic Linkage*. The John Hopkins University Press, 1994.
- [4] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach, *American Journal of Human Genetics* 55: 1347-1363, 1996.
- [5] Gudbjartsson, D.F., Jonasson, K., Frigge, M. and Kong, A. Allegro, a new computer program for multipoint linkage analysis, *Nature Genetics* 25: 12-13, 2000.
- [6] Nicolae, D.L. *Allele sharing models in gene mapping: a likelihood approach.*, PhD thesis, Department of Statistics, University of Chicago, 1999.
- [7] Whittemore, A.S. and Halpern, J. A class of tests for linkage using affected pedigree members, *Biometrics* 50: 118-127, 1994.
- [8] Kämpe, M. *Two-Locus Nonparametric Linkage Analysis for Complex Diseases*, Master's thesis, Lund Institute of Technology, Lund University, 2001.
- [9] Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nature Genetics* 21: 213-215, 1999.
- [10] Hössjer, O. *Asymptotic Estimation Theory of Multipoint Linkage Analysis Under Perfect Marker Information*, Lund University, 2001.
- [11] Lander, E. and Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* 11: 241-247, 1995.
- [12] Feingold, E. Markov Processes for modeling and analyzing a new genetic mapping method, *Journal of Applied Probability* 30: 766-779, 1993.

- [13] Lander, E.S. and Botstein, D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps, *Genetics* 121:185-199, 1989.
- [14] Collins, A., Teague, J. and Morton, N.E. A metric map of humans: 23.500 loci in 850 bands, *Proc. Natl. Acad. Sci. USA* 93: 14771-14775, 1996.
- [15] Gelder Ehm, M., Karnoub, M.C., Sakul, H., Gottschalk, K., Holt, D.C., Weber, J.L. et al. Genomewide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations, *American Journal of Human Genetics* 66: 1871-1881, 2000.
- [16] Rich, S. Mapping Genes in Diabetes: Genetic Epidemiological Perspective, *Diabetes* 39: 1315-1319, 1990.
- [17] Parker, A., Meyer, J., Lewitzky, S., Rennich, J.S., Chan, G., Thomas, J.D. et al. A Gene Conferring Susceptibility to Type 2 Diabetes in Conjunction With Obesity Is Located on Chromosome 18p11, *Diabetes* 50: 675-680, 2001.
- [18] Mahtani, M.M., Widén, E., Lehto, M., Thomas, J., McCarthy, M., Brayer, J. et al. Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families, *Nature Genetics* 14: 90-94, 1996.
- [19] Daly, M. Seminar given at the Wallenberg Institute in Malmö, May 23rd 2001.
- [20] Lander, E.S. and Green, P. Construction of multilocus genetic linkage maps in humans, *Proc. Natl. Acad. Sci. USA* 85: 2363-2367, 1987.
- [21] Wise, L.H., Lanchbury, J.S. and Lewis, C.W. Meta-analysis of genome searches, *American Journal of Human Genetics* 63: 263-272, 1999.