Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 3

Global tests for linkage

R. el Galta, J.C. van Houwelingen and J.J. Houwing-Duistermaat

Abstract

To test for global linkage along a genome or in a chromosomal region, the maximum over the marker locations of mean alleles shared identical by descent of affected relative pairs, Z_{max}, can be used. Feingold et al. (1993) derived a Gaussian approximation to the distribution of the Z_{max}. As an alternative we propose to sum over the observed marker locations along the chromosomal region of interest. Two test statistics can be derived. (1) The likelihood ratio statistic (LR) and (2) the corresponding score statistic. The score statistic appears to be the average mean IBD over all available marker locations. The null distribution of the LR and score tests are asymptotically a 50:50 mixture of chi-square distributions of null and one degree of freedom and a normal distribution, respectively.

We compared empirically the type I error and power of these two new test statistics and Z_{max} along a chromosome and in a candidate region. Two models were considered, namely (1) one disease locus and (2) two disease loci. The new test statistics appeared to have the right type I error. Along the chromosome, for both models we concluded that for very small effect sizes, the score test was more powerful than the other test statistics. For large effect sizes, the likelihood ratio statistic and Z_{max} were comparable and performed much better than the score test. For candidate regions of about 30 cM, all test statistics were comparable.

To be submitted for publication

3.1 Introduction

In complex genetic diseases, multiple genes are assumed to cause a predisposition to disease. Each single susceptibility gene might contribute little to disease, and therefore the statistical power to detect such a gene is low, especially if the gene is common and has low penetrance. Furthermore, some of these genes might lie on the same chromosomal region of interest.

Regions harboring genes responsible for a trait are often identified by means of genome wide linkage analysis, which studies co-segregation of an unobserved disease locus and a marker locus. Two approaches can be considered, namely parametric and nonparametric linkage analysis. Parametric methods require allele frequencies at the disease locus and the penetrances to be known. Parametric linkage analysis is the most powerful when the genetic parameters are correctly specified. For many complex traits the mode of inheritance is unknown. For such traits non-parametric methods are suitable, as they do not make any assumption about the mode of inheritance. Nonparametric methods rely only on the information of sharing alleles identical by descent (IBD) between relatives at a locus to study whether it is genetically linked to the unobserved disease locus. Linkage between a disease locus and marker genotypes can be studied by comparing the observed IBD sharing of affected relative pairs to the expected IBD sharing under random segregation. An increase in IBD sharing indicates the presence of a susceptibility gene in the region. Conventionally testing for linkage is carried out at each observed locus throughout the human genome. To adjust for multiple testing, only p-values smaller than 2.2×10^{-5} are considered to be significant (Lander and Kruglyak, 1995). In this paper we propose two global tests for linkage, which use IBD information from observed markers all together. All tests considered in this paper, assume that each of unlinked chromosomal regions carries one disease locus at most.

A global test for linkage, which tests all observed markers simultaneously, can be obtained by summing the likelihood of the data over all marker locations along the region of interest. Throughout this paper, we will refer to this approach as the averaging approach. Siegmund (2001) discussed briefly the averaging approach for complete IBD information. However, we think that this approach merits more consideration. For a candidate region, Liang et al. (2001) proposed a generalized estimating equations approach to primarily estimate the location of a disease gene. The authors used IBD information from all markers jointly. Liang's method can also be regarded as an averaging

approach.

For various types of affected relative pairs, and dense and fully informative markers about IBD status, Feingold et al. (1993) proposed a global test for linkage, which appeared to be the maximum of mean IBD over all markers (Z_{max}). The authors derived a Gaussian approximation to the significance level of Z_{max} for large sample sizes. Teng and Siegmund (1998) extended this approach to relative pairs with partial information about the IBD status.

In this paper, we considered the averaging approach for both complete and partial information about the IBD status. For simplicity, we restricted to affected sibling pairs (ASP). Two test statistics were derived namely the likelihood ratio statistic and the corresponding score statistic. The score statistic appeared to be the average mean IBD over all available marker locations. The null distribution of the likelihood ratio and score statistic are asymptotically a 50:50 mixture of chi-square distributions of null and one degree of freedom and a normal distribution, respectively.

For complete IBD information, we compared empirically the type I error and power of these two test statistics and Z_{max}. To generate data two models were considered, namely (1) single-locus disease model and (2) two-locus disease model. The new test statistics appeared to have the right type I error. For both models we concluded that for a sample of 200 ASPs and a small effect sizes ($\lambda_s < 1.17$, with λ_s the siblings relative recurrence risk (Risch, 1990a)), the score test had slightly more power than the other test statistics. For large effect sizes or large sample sizes, the likelihood ratio statistic and Z_{max} were comparable and perform better than the score test. Further, we studied the effect of the information loss on the performance of the test statistics when only partial information is available.

3.2 Methods

Complete IBD information

First we describe briefly the approach proposed by Feingold et al. (1993). Let T be the length (in cM) of a chromosome of interest. Suppose that we have N affected sib pairs. For a marker locus at the position t let $X_{t,i}^k$ be the event that the members of the i^{th} sib-pair share k alleles identical by descent (IBD), for $k = 0, 1, 2$ and let $X_{t,i} = \sum_{k=0}^{2} k X_{t,i}^k$ the number of marker alleles shared IBD. Assume that all markers are fully informative about the IBD status. Hence $X_{t,i}$ has expectation $E_0[X_{t,i}] = 1$ and variance $Var(X_{t,i}) = 1/2$ under the null hypothesis of no linkage. Let $X_t = \sum_{i=1}^{N} X_{t,i}$

and $Z_t = \frac{X_t - N}{\sqrt{N}}$. Assuming Haldane's mapping function, $\{Z_t, 0 \leq t \leq T\}$ is approximately a Gaussian Markov process which has zero mean and covariance $R(t,s) = 1/2 \exp(-0.04|t-s|)$ under the null hypothesis. Under the alternative hypothesis of the presence of one susceptibility gene at the location τ on the chromosome, this process $\{Z_t, 0 \leq t \leq T\}$ is superimposed by $0.5\sqrt{N}\alpha \exp(-0.04|t-\tau|)$ with α representing the excess IBD sharing. The parameter α varies between 0 and 1. For an additive model, $\alpha = \frac{\lambda_s - 1}{\lambda_s}$ with λ_s, the sibling risk ratio (Risch, 1990a). For small α an approximation of the log likelihood of this process is

$$L(\tau, \alpha | Z_t, 0 \leq t \leq T) = L(\alpha | Z_\tau) \propto exp(\sqrt{N}\alpha Z_\tau - N\alpha^2/4). \qquad (3.1)$$

It is not known which locus is the disease locus. To test the null hypothesis $H_0 : \alpha = 0$ i.e. none of the markers is linked with the disease, Feingold et al. (1993) proposed to use the maximum of the corresponding likelihood ratio over the parameters α and τ, which appeared to be

$$Z_{max} = \sqrt{\max_\tau \max_\alpha 2 log L(\tau, \alpha)} = \max_\tau \sqrt{2} Z_\tau.$$

Feingold et al. (1993) derived the following approximation to calculate the significance level of Z_{max}

$$P_0(Z_{max} > b) \approx 1 - \Phi(b) + 0.04 T \varphi(b)$$

with φ and Φ are the standard normal density and distribution functions, respectively.

Averaging approach

As an alternative to maximizing the conditional likelihood over τ one can take the average of conditional likelihoods given the location of the disease locus over all marker loci assuming that they are all equally likely to be in complete linkage with the disease locus. Hence the average likelihood is approximately

$$L(\alpha | Z_t, 0 \leq t \leq T) \propto \sum_t L(\tau = t, \alpha) = \sum_t exp(\sqrt{N}\alpha Z_t - N\alpha^2/4). \qquad (3.2)$$

Here, we assume also the presence of a single disease locus in the region of interest. As test statistic we propose to use either the corresponding likelihood ratio test

$$\Lambda = \max_{0 \leq \alpha \leq 1} 2 \log(\frac{L(\alpha)}{L(\alpha = 0)}) = \max_{0 \leq \alpha \leq 1} 2 \log(\sum_t exp(\sqrt{N}\alpha Z_t - N\alpha^2/4)) \qquad (3.3)$$

or the score test

$$U = \frac{\partial \log(L(\alpha = 0))/\partial\alpha}{\sqrt{-E[\partial^2 \log(L(\alpha = 0))/\partial\alpha^2]}} = \frac{\sum_t Z_t}{\sqrt{\sum_{t,s} Cov_0(Z_t, Z_s)}}, \tag{3.4}$$

with $L(\alpha)$ given in formula (3.2) and $\sum_{t,s} Cov_0(Z_t, Z_s) = \sum_{t,s} R(t,s)/2$, the variance under the null hypothesis. Since the parameter α is positive, the null distribution of Λ is approximately a 50:50 mixture of χ^2_1 distribution and a point mass at zero ($\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0$). The score test U follows asymptotically the normal distribution under the null hypothesis. It is a one-sided test and rejects the null hypothesis only for positive values of U. Let $U^2_+ = U^2$ if $U > 0$ otherwise $U^2_+ = 0$. Then U^2_+ is asymptotically distributed as $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0$. Note that U^2_+ approximates Λ for α small. Hence Λ and U are locally asymptotically equivalent. The theoretical power of the score test U is given by

$$\beta = 1 - \Phi(\frac{b\sqrt{Var_0(Z)} - E_\alpha[Z]}{\sqrt{Var_\alpha(Z)}}), \tag{3.5}$$

with $Z = \sum_t Z_t$, and $E_\alpha[Z]$ and $Var_\alpha[Z]$, the expectation and the variance of Z under the alternative hypothesis. The critical value b is the normal percentile corresponding to a prespecified significance level. For instance $b = 1.64$ corresponds to a significance level of 0.05. The covariance matrix of the process $\{Z_t, 0 \leq t \leq T\}$ is given in the appendix for two models, namely (1) one disease locus on the chromosome and (2) two disease loci on the same chromosome. For the likelihood ratio test Λ, the power can be estimated by means of a simulation study. The value of α that maximize the average likelihood can be used as an estimate of the effect size of the susceptibility gene. The posterior probabilities are of interest; for instance the largest posterior probability indicates the most likely location of a disease gene.

Incomplete IBD information

Since genotyped markers are typically not fully informative about the IBD status, the IBD sharing $X_{t,i}$ can be replaced by $\hat{X}_{t,i} = E_0[X_{t,i}|G_i]$ the mean IBD under the null hypothesis of no linkage given the genotypes, G_i, of the ith sib-pair for all observed markers. Given the disease locus at the location τ Teng and Siegmund (1998) used the following likelihood to derive a corresponding test statistic to Z_{max} for incomplete IBD information

$$\widetilde{L}(\tau, \alpha|G) \propto \prod_i^N (1 + \alpha(\hat{X}_{\tau,i} - 1)). \tag{3.6}$$

Similarly to the case of complete IBD information the average likelihood can be obtained by summing over all available marker positions. Hence the average likelihood is

$$\widetilde{L}(\alpha) \propto \sum_{t} \prod_{i}^{N} (1 + \alpha(\hat{X}_{t,i} - 1)). \tag{3.7}$$

The likelihood ratio corresponding to $\widetilde{L}(\alpha)$

$$\hat{\Lambda} = \max_{0 \leq \alpha \leq 1} 2log(\widetilde{L}(\alpha)/\widetilde{L}(\alpha = 0))$$

Let $\hat{Z}_t = \sum_{i=1}^{N} (\hat{X}_{t,i} - 1)/\sqrt{N}$. An estimate of the score test is

$$\hat{U} = \frac{\sum_t \hat{Z}_t}{\sqrt{\sum_{t,s} \hat{Cov}(\hat{Z}_t, \hat{Z}_s)}}$$

where the covariances are the sample covariances. The likelihood ratio $\hat{\Lambda}$ follows approximately a 50:50 mixture of χ^2_1 distribution and a point mass at zero. The score test $\hat{U}$ asymptotically follows the normal distribution.

3.3 Simulation

Complete IBD information

Single-locus disease model

Whole chromosome

For practical reasons we simulated data from a multivariate normal distribution instead of generating the process of number of IBD sharing for affected sib pairs. We generated a vector of length 100, which corresponds to IBD data on 100 equidistant markers along a chromosome of 300 cM, from a multivariate normal distribution with the mean and covariance being evaluated under the alternative hypothesis (Liang et al., 2001). Data at each position were considered as an average IBD sharing of 200 affected sib-pairs. We repeated the procedure 10,000 times. We positioned one disease gene at 75 cM on the chromosome, with varying values of α. The results are summarized in Figure 7.1. Type I error and power were calculated at significance level of 0.05. All test statistics provided reasonable type I error rates. For small effect size ($\alpha \leq 0.15$), the score statistic U appeared to have slightly more power than

Z_{max} and Λ. For larger effect size ($\alpha > 0.15$), the test statistics Λ and Z_{max} had similar power and performed better than U. From the simulation results we observed that Λ_s and Z_{max} statistics attained the power of 80 % at $\alpha \approx 0.37$, which corresponds to $\lambda_s = 1.59$. Hence, the sample size N required to achieve the power of 80 % for a given α, can be approximated by using the following formula

$$N = 200(0.37/\alpha)^2.$$

For example the sample size required to achieve a power of 80 % when $\alpha(\lambda_s) = 0.05$ (1.05), 0.1 (1.11), 0.15 (1.17) and 0.23 (1.3) are 10765, 2692, 1196 and 500 ASPs, respectively. The sample size required to attain a power of 80 % can be approximated by the using the corresponding formula to the formula (3.5). The score test U achieved a power of 80 % at $\alpha(\lambda_s) = 0.6(2.5)$. The sample size required to achieve a power of 80 % when $\alpha(\lambda_s) = 0.05$ (1.05), 0.1 (1.11), 0.15 (1.17) and 0.23 (1.3) are 28800, 7200, 3200 and 800 ASPs, respectively. Similar results were also obtained by using

$$N = 200(0.6/\alpha)^2.$$

Note that these results are only for one chromosome, and thus the sample sizes required for genome scan are much higher, see also Cordell (2001).

Candidate region

To study the performance of the test statistics in a candidate region, we generated 10000 data sets. Each data set consists of fully IBD information for 500 ASPs, on 10 equidistant markers spanning a chromosomal region of 30 cM. The disease gene was positioned at the middle of the chromosomal region, with varying values of α. The results are depicted in Figure 3.2. A nominal significance level of 0.05 is considered. All test statistics were comparable in terms of the power.

Two-locus disease model

In order to study the robustness of these test statistics we considered the presence of two disease loci on the same chromosome. Data were generated similar to the case of one disease locus from a multivariate normal distribution with the mean and covariance matrix as given in the appendix (3.5) for various values of α_1 and α_2, the parameters of increased IBD at the first and the second disease locus, respectively. Unlike for the single-locus disease model, the IBD

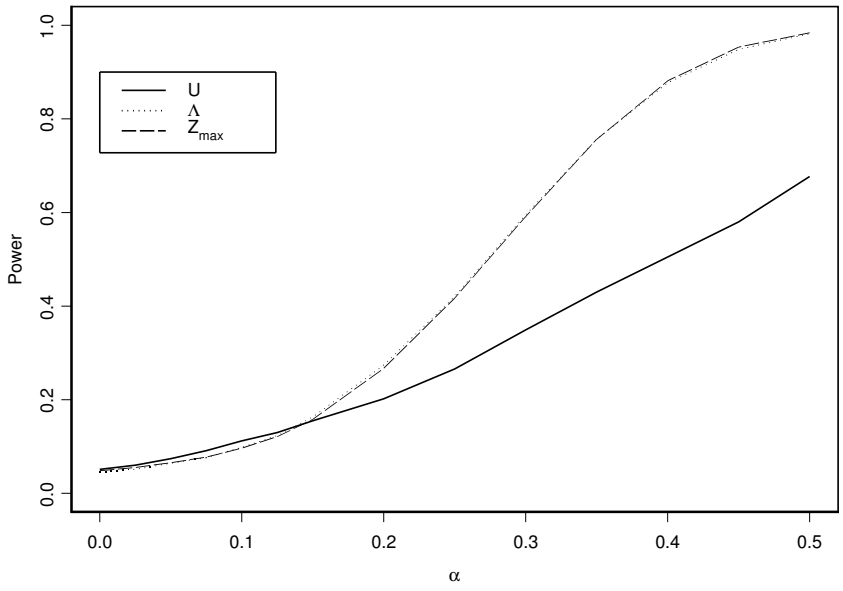

FIGURE 3.1: *Power to detect linkage due to a single disease locus at $\tau = 75$ cM when IBD data are available on 100 fully equidistant markers along a chromosome of length 300 cM for 200 ASPs.*

process is now controlled by α_1, α_2, and the covariance matrix of IBD at the disease loci through the penetrance matrix and the allele frequencies at the disease loci. The disease loci were located at $(\tau_1, \tau_2) = (90, 150)$. The results are shown in Table 3.1. A nominal significance level of 0.05 is used. Compared to the single-locus disease model, all test statistics appeared to gain power when two disease loci exist. The power of the score statistic U was especially improved. For effect sizes $\alpha_1 \leq 0.10$ and $\alpha_2 \leq 0.10$ the score statistic yielded the highest power. For larger effect sizes the corresponding likelihood ratio statistic Λ often performed the best. The Z_{max} statistic has good power relative to U.

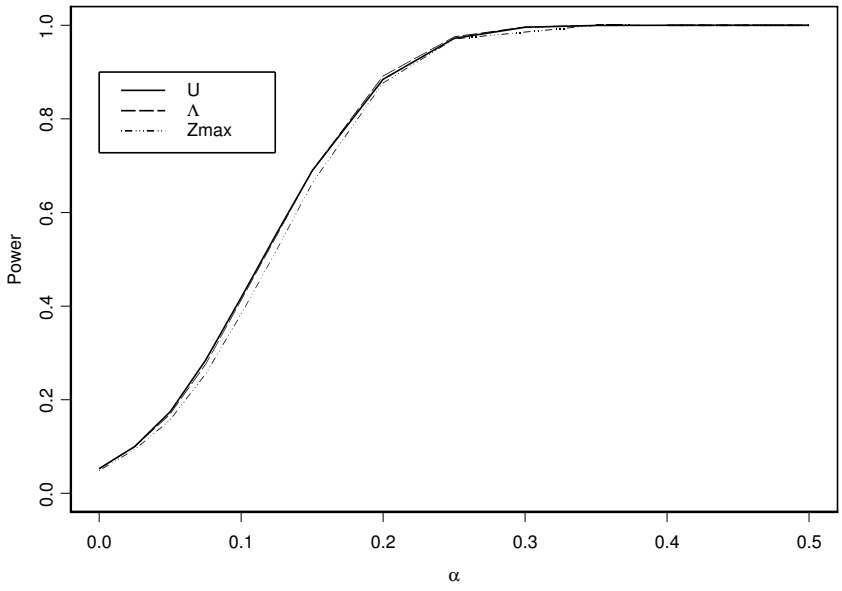

FIGURE 3.2: *Power to detect linkage due to a single disease locus at $\tau = 15$ cM when IBD data are available on 10 fully equidistant markers along a chromosome of length 30 cM for 500 ASPs.*

Partial IBD information

We generated 10000 data sets under the null model and 1000 data sets under the alternative model of one disease locus, using the ALLEGRO program (Gudbjartsson et al., 2000). Each data set consisted of 200 affected sib-pairs and their parents. We considered a chromosome with a length of 315 cM. Genotypes were simulated for 41 markers spaced about 7.8 cM on average. The disease locus was located at $\tau = 75$ cM. The two adjacent observed markers to τ were located at $t_9 = 73$ cM and $t_{10} = 81$ cM. We varied α from 0 to 0.3. Multipoint IBD's were calculated using the Merlin program (Abecasis et al., 2002). The simulation results are depicted in Figure 3.3 for affected sib-pairs, with and without information on parental genotypes at the left and right panel, respectively. A nominal significance level of 0.05 is used. Here

TABLE 3.1: Power to detect linkage due to two disease loci located on the same chromosome at $\tau_1 = 90$ and $\tau_2 = 150$, when IBD data are available on 100 fully informative markers in 200 ASPs.

p_1	p_2	α_1	α_2	λ_s	U	Λ	Z_{max}
0.033	0.033	0.05	0.05	1.051	0.10	0.07	0.07
0.05	0.033	0.10	0.05	1.082	0.14	0.12	0.12
0.05	0.05	0.10	0.10	1.113	0.18	0.16	0.15
0.063	0.031	0.15	0.05	1.113	0.19	0.21	0.21
0.064	0.05	0.15	0.10	1.146	0.23	0.24	0.22
0.065	0.065	0.15	0.15	1.183	0.28	0.30	0.28
0.08	0.066	0.20	0.15	1.226	0.35	0.44	0.42
0.081	0.081	0.20	0.20	1.267	0.44	0.54	0.51
0.092	0.068	0.25	0.15	1.267	0.43	0.59	0.59
0.096	0.084	0.25	0.20	1.317	0.48	0.65	0.63
0.1	0.1	0.25	0.25	1.370	0.55	0.70	0.67

p_1 and p_2 are the frequencies of disease alleles A and B at loci τ_1 and τ_2 respectively. The data were generated under genetic models with the following penetrance matrix

	Genotypes at τ_1		
Genotypes at τ_2	AA	Aa	aa
BB	0.95	0.95	0.95
Bb	0.95	0.09	0.09
bb	0.95	0.09	0.09

we compared the score test $\hat{U}$, the likelihood ratio $\hat{\Lambda}$ and the maximum of the score statistics $\hat{Z}_{max}$ proposed by Teng and Siegmund (1998). When parental genotypes were available, all test statistics showed the same pattern as for the perfect IBD information. The score test $\hat{U}$ had the highest power for effect sizes $\alpha < 0.15$, and for $\alpha \geq 0.15$ the test statistics $\hat{\Lambda}$ and $\hat{Z}_{max}$ performed similarly and had the highest power. When parental genotypes were not available, all test statistics appeared to be conservative and the power decreased. The type I error rates of $\hat{U}$, $\hat{\Lambda}$ and $\hat{Z}_{max}$ were about 0.048, 0.045 and 0.052 when parental genotypes were available respectively, and they dropped to 0.04, 0.42 and 0.039 when parental genotypes were not available. In terms of the power the $\hat{Z}_{max}$ statistic suffered the most from the loss of IBD information, whereas the score statistic was the least affected. For small $\alpha < 0.15$ the score statistic

$\hat{U}$ had the highest power, and for $\alpha \geq 0.15$ the likelihood ratio test Λ had the highest power.

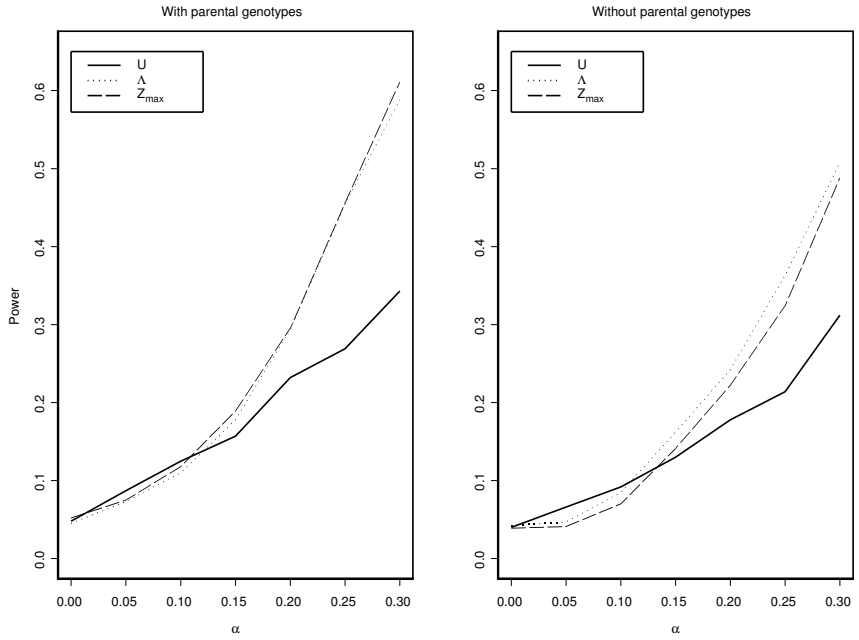

FIGURE 3.3: *Power to detect linkage due to a single disease locus when IBD information is partially available on 41 markers along a chromosome of length 315 cM in 200 ASPs.*

3.4 Discussion

In this paper, we proposed an averaging approach to test for linkage of an unobserved disease locus and a marker locus when IBD data are available on multiple markers. We first considered marker data to be fully informative about the IBD status. We assumed that all observed marker loci are equally likely to be the disease locus. As a basic model to our approach we used the model proposed by Feingold et al. (1993). The model assumes the presence of one disease locus at most on a chromosome. The likelihood of the data was computed by summing the conditional likelihood given a marker locus being the disease locus, over all observed marker loci. The corresponding likelihood

ratio test or the score test can be used. Both test statistics are easy to compute and have known asymptotic distributions under the null hypothesis. Further, we adapted the method to the case of partial information on the IBD status.

We performed a simulation study to compare the performance of the averaging approach in comparison to the maximising approach (Feingold et al., 1993; Teng and Siegmund, 1998). For complete IBD information we considered a single-locus disease model as well as a two-locus disease model. The score test of the averaging approach appeared to perform slightly better when each single susceptibility gene contributed little to the disease and the sample size was small. However, this difference in power may disappear if the sample size is large. For large effect sizes ($\alpha > 0.15$), the likelihood ratio test of the averaging approach performed best when two disease loci existed on the same chromosome or when the IBD information was partial, and it had similar power to the maximising approach when IBD information was complete. In a candidate region the averaging approach has slightly higher power to detect linkage relative to the maximising approach. The score test U and the likelihood ratio Λ performed equally.

When information about IBD status is not fully known, the amount of IBD sharing is often estimated by its expectation given the available genotypes data from all observed markers assuming linkage equilibrium. However this assumption may not be valid if the marker map is dense. The less the IBD information the less accurate are the IBD estimates. Further, the variance of the estimate is unknown and it should in turn be estimated by its sample variance (Teng and Siegmund, 1998). Unlike the $\hat{Z}_{max}$ and $\hat{U}$ statistics , the likelihood ratio $\hat{\Lambda}$ does not need to estimate the variance, and therefore it may be less affected by the loss of IBD information when sample size is large. To reduce the bias one should include all available parental genotypes in analysis (Risch, 1990c). Recently, Bacanu (2005) proposed an approach to eliminate the bias due to the presence of linkage disequilibrium between adjacent markers. The author partitioned the markers into interlaced and non-overlapping subsets, and then analyzed each set separately. A final test statistic for linkage is the standardized average of subset-specific statistics.

Liang et al. (2001) proposed a GEE method to estimate the location of a single susceptibility gene in a candidate region. The authors proposed also a GEE-based test statistic of linkage. The test statistic appeared to be a weighted sum of the number of IBD sharing over the markers positions along the region of interest. This statistic assigns more weight to markers at the ends of the chromosomal region of interest. Power calculations using the formula (3.5)

and the corresponding formula in Liang et al. (2001) (data not shown) show that the score test proposed in this paper has more power than the GEE-based test statistic as long as the susceptibility gene does not lie at the end of the chromosomal region of interest. Moreover the GEE-based test statistic becomes very conservative relative to the test statistic $\hat{U}$ when parental genotypes are missing (data not shown), which is in agreement with the findings of Lebrec et al. (2005).

Biernacka et al. (2005) extended the work of Liang et al. (2001) to address the presence of two disease loci on the same chromosome. The averaging approach can also be extended to test for the presence and estimate the effect sizes of two disease loci on the same chromosome. The likelihood can be obtained by taking the sum of the conditional likelihoods given two disease loci over all marker pairs. the corresponding likelihood ratio asymptotically follows a 0.25, 0.5 and 0.25 mixture of chi squares with zero, one and two degrees of freedom, respectively (Self and Liang, 1987) Moreover, the application of the approach for other relative pairs, i.e. half sibling, grandparent-grandchild, cousin pairs, etc, is straightforward.

We conclude that the averaging approach improves the power to detect linkage relative to the maximising approach when a single disease-locus of small effect size exists on a chromosome or two disease loci lie on the same chromosome.

3.5 Appendix

The expectation and the covariance of the process $\{Z_t, 0 \leq t \leq T\}$ when two disease loci exist

Suppose that two loci disease loci are located at τ_1 and τ_2. Let $\alpha_1 = E[X_{\tau_1,i} - 1]/2$ and $\alpha_2 = E[X_{\tau_2,i} - 1]/2$ be the deviation of the expectation of the number of IBD sharing at τ_1 and τ_2 from the mean under the null hypothesis. Using the fact that given the IBD at the disease loci the IBD process on the same chromosome does not involve α_1 and α_2 (Teng and Siegmund, 1998), the expectation and covariance matrix of the Gaussian process $\{Z_t, 0 \leq t \leq T\}$ can be calculated using the following formulae

$$\begin{aligned} E[Z_t] &= E[E_0[Z_t|Z_{\tau_1}, Z_{\tau_2}]] \\ Cov(Z_t, Z_s) &= E[Cov_0(Z_t, Z_s|Z_{\tau_1}, Z_{\tau_2})] + Cov(E_0[Z_t|Z_{\tau_1}, Z_{\tau_2}], E_0[Z_s|Z_{\tau_1}, Z_{\tau_2})]) \end{aligned}$$

Using the theory of the conditional multivariate normal distributions (Anderson, 1984, p. 37) and the fact that the Gaussian process is Markovian, the

expectation and the covariance are

$$E[Z_t] = \begin{cases} 0.5\sqrt{N}\alpha_1 e^{-0.04|\tau_1 - t|} & \text{for } t < \tau_1 \le \tau_2 \\ 0.5\sqrt{N}\alpha_2 e^{-0.04|\tau_2 - t|} & \text{for } \tau_1 \le \tau_2 < t \\ 0.5\sqrt{N}(\alpha_1 c_1(t) + \alpha_2 c_2(t)) & \text{for } \tau_1 \le t \le \tau_2 \end{cases}$$

$$Cov(Z_t, Z_s) = \begin{cases} e^{-0.04|s-t|}/2 + (Var(Z_{\tau_1}) - 1/2)e^{-0.04(|s-\tau_1|+|t-\tau_1|)} & \text{for } s \le t \le \tau_1 \le \tau_2 \\ e^{-0.04|s-t|}/2 + (Var(Z_{\tau_2}) - 1/2)e^{-0.04(|t-\tau_2|+|s-\tau_2|)} & \text{for } \tau_1 \le \tau_2 \le s \le t \\ Cov(Z_{\tau_1}, Z_{\tau_2})e^{-0.04(|s-\tau_1|+|t-\tau_2|)} & \text{for } s \le \tau_1 \le \tau_2 \le t \\ Var(Z_{\tau_1})c_3(s,t) + Cov(Z_{\tau_1}, Z_{\tau_2})c_4(s,t) & \text{for } s \le \tau_1 \le t \le \tau_2 \\ Var(Z_{\tau_2})c_5(s,t) + Cov(Z_{\tau_1}, Z_{\tau_2})c_6(s,t) & \text{for } \tau_1 \le s \le \tau_2 \le t \\ c(s,t)/2 + Var(Z_{\tau_1})c_1(s)c_1(t) + Var(Z_{\tau_2})c_2(s)c_2(t) & \\ \quad + Cov(Z_{\tau_1}, Z_{\tau_2})(c_1(s)c_2(t) + c_2(s)c_1(t)) & \text{for } \tau_1 \le s \le t \le \tau_2 \end{cases}$$

with

$$\begin{aligned} c_1(s) &= \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-\tau_1|} \\ c_2(s) &= \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-\tau_2|} \\ c_3(s,t) &= \frac{1 - e^{-0.08|t-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-t|} \\ c_4(s,t) &= \frac{1 - e^{-0.08|t-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_2|)} \\ c_5(s,t) &= \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04|s-t|} \\ c_6(s,t) &= \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_2|} \\ c(s,t) &= e^{-0.04|s-t|} - \frac{1 - e^{-0.08|s-\tau_2|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_1|+|t-\tau_1|)} \\ &\quad - \frac{1 - e^{-0.08|s-\tau_1|}}{1 - e^{-0.08|\tau_2-\tau_1|}} e^{-0.04(|s-\tau_2|+|t-\tau_2|)} \end{aligned}$$

Similar formula of the expectation was also given by Biernacka et al. (2005). The variance and covariance of Z_{τ_1} and Z_{τ_2} can be directly calculated by

$$\begin{aligned} Cov(Z_{\tau_1}, Z_{\tau_2}) &= Cov(X_{\tau_1,1}, X_{\tau_2,1}) \\ &= 4g_{22} + 2g_{12} + 2g_{21} + g_{11}, \end{aligned}$$

with g_{ij} the probability that a pair share i and j alleles IBD at disease locus τ_1 and τ_2 respectively:

$$\begin{aligned} g_{ij} &= P(X_{\tau_1,1} = i, X_{\tau_2,1} = j|\phi) \\ &= \frac{\sum_G f_{G_{11}\times G_{12}} f_{G_{21}\times G_{22}} P(G_{11}, G_{21}|X_{\tau_1,1} = i) P(G_{12}, G_{22}|X_{\tau_2,1} = j) P(X_{\tau_1,1} = i, X_{\tau_2,1} = j)}{\sum_{i,j}\sum_G f_{G_{11}\times G_{12}} f_{G_{21}\times G_{22}} P(G_{11}, G_{21}|X_{\tau_1,1} = i) P(G_{12}, G_{22}|X_{\tau_2,1} = j) P(X_{\tau_1,1} = i, X_{\tau_2,1} = j)} \end{aligned}$$

where the sum is taken over all possible genotypes, $G = (G_{11} \times G_{12}, G_{21} \times G_{22})$, at both disease loci τ_1 and τ_2 of the first and second relative, respectively (Biernacka, 2004). The function $f_{G_{r1}\times G_{r2}}$ is the penetrance given the genotypes G_{r1} and G_{r2} at both disease loci of the pair member r with $r = 1, 2$. The joint probabilities of sharing i and j IBD at τ_1 and τ_2 $P(X_{\tau_1,1} = i, X_{\tau_2,1} = j|\phi)$ were given by Haseman and Elston (1972). The probabilities that a pair has genotypes G_1 and G_2 given that they share j IBDs, $P(G|X_\tau = j)$, for $j = 0, 1, 2$, are summarized in Table 3.2 according to Thompson (1975).

TABLE 3.2: Probability $P(G_1, G_2|X_\tau = j)$, for $j = 0, 1, 2$

G_1	G_2	$X_\tau = 2$	$X_\tau = 1$	$X_\tau = 0$
A/A	A/A	p^2	p^3	p^4
A/A	A/a	0	$2p^2q$	p^3q
A/A	a/a	0	0	p^2q^2
A/a	A/A	0	$2p^2q$	p^3q
A/a	A/a	pq	$pq(p+q)$	p^2q^2
A/a	a/a	0	$2pq^2$	pq^3
a/a	A/A	0	0	p^2q^2
a/a	A/a	0	$2pq^2$	pq^3
a/a	a/a	q^2	q^3	q^4

The locus has 2 alleles A and a with frequencies p and q.