Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 2

Score statistic to test for genetic correlation for proband-family design

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

Abstract

In genetic epidemiological studies, informative families are often oversampled to increase the power of a study. For a proband-family design, where relatives of probands are sampled, we derived the score statistic to test for clustering of binary and quantitative traits within families due to genetic factors. The derived score statistic is robust to ascertainment scheme. We considered correlation due to unspecified genetic effects and/or due to sharing alleles identical by descent (IBD) at observed marker locations in a candidate region. A simulation study was carried out to study the distribution of the statistic under the null hypothesis in small data-sets. To illustrate the score statistic, data on 33 families with type 2 diabetes mellitus (DM2) were analyzed. In addition to the binary outcome DM2, we also analyzed the quantitative outcome, body mass index (BMI). For both traits, familial aggregation was highly significant. For DM2, including also IBD sharing at marker D3S3681 as a cause of correlation gave an even more significant result, which suggests the presence of a trait gene linked to this marker. We conclude that for the proband-family design the score statistic is a powerful and robust tool for detecting clustering of outcomes.

Published in *Ann Hum Genet*. 2005; **69**:373-81

2.1 Introduction

Complex genetic traits are often determined by multiple genetic and environmental factors with small effects (Lander and Schork, 1994) necessitating large-scale studies to obtain enough power to identify genetic factors. The power can also be increased by enrichment of the sample for the presence of genetic factors via selection of families with an unusual distribution of the trait (Amos and de Andrade, 2001; Carey and Williamson, 1991; Elston, 2000; Liang and Beaty, 2000; Risch and Zhang, 1995). For binary and quantitative traits, the selected families often have a high proportion of affected individuals and of subjects with phenotypes exceeding some threshold T, respectively. In this paper, we consider a proband-family design where relatives are sampled through probands. We define a set of probands to be a set of family members chosen in such a way that the remaining relatives are not related to ascertainment (Ewens and Shute, 1986). A score statistic is derived to test for clustering of a trait due to genetic factors within these families.

In family studies, a first question to be answered is whether genetic factors play a role in the observed trait. Closely related individuals should tend to have similar outcomes compared to non or distantly related individuals. Another question of interest may be whether correlation exists due to the presence of a genetic factor located in a genomic region. If a genetic factor is linked to a marker, relatives with excess sharing of marker alleles identical by descent (IBD) should tend to have similar outcomes compared to relatives with less sharing of alleles IBD.

Since the families selected through probands are non random, the statistical methods used should take into account the ascertainment procedure (Elston and Sobel, 1979; Morton, 1959). For single ascertainment, Cannings and Thompson (1977) proposed to condition on the phenotypes of the probands. However if single ascertainment cannot be assumed, conditioning on affected probands may not be sufficient. Ewens and Shute (1986) showed that if remaining family members are related to ascertainment, the estimates of genetic parameters may be biased. They proposed to split the family in a set of probands related to ascertainment and a set of relatives not related to ascertainment. For example if nuclear families with at least one affected offspring are ascertained, the set of probands should consist of all observed offspring regardless of their affection status and the only family members not related to ascertainment are the parents. Like Ewens and Shute (1986) we also assume that family members who do not belong to the set of probands are not related

to ascertainment.

For randomly selected families, Houwing-Duistermaat et al. (2003, 1995) derived the score statistic to test for correlation due to genetic factors. The test can be applied before complex genetic models are fitted. The score statistic does not assume any distribution of random effects. Adjustments for covariates can be made. In this paper we derive the score statistic for clustering of binary and quantitative traits in the proband family design by using the conditional likelihood given the phenotypes of the probands. By conditioning on the phenotypes of the individuals relevant for ascertainment, the method is robust to the ascertainment scheme (Ewens and Shute, 1986). To take into account the relationship between covariates and outcome, we propose to use data from other sources, such as large-scale epidemiological studies.

Commenges et al. (1995) proposed also a score statistic for testing familial aggregation of binary traits in families ascertained via probands. However, they only considered a random intercept model, which yields equal correlation between family members and they assumed single ascertainment. Furthermore the effects of covariates on the trait are estimated under the null hypothesis using the data on the relatives. However for larger sets of probands, it can be impossible to estimate the parameters. For example for age dependent traits, the effect of age cannot be estimated from the data if the set of probands consists of all offspring and the remaining relatives are only the parents. Moreover, Rao et al. (1988) and de Andrade and Amos (2000) showed that conditioning on the trait value of the proband is not efficient for parameter estimation. Therefore we suggest to obtain parameters from an available population based study. Commenges et al. (1995) used the normal distribution for the distribution of the statistic under the null hypothesis. As alternative, we propose the scaled χ^2 distribution. By means of simulation we study the performance of the scaled χ^2 and the normal distribution in small data-sets.

As an illustration we applied the score statistics to a sample of first-degree relatives of probands with type 2 diabetes mellitus (DM2). Probands were patients with DM2 living in the GRIP population (Genetic Research in Isolated Populations), and known to be affected by the physicians participating in GRIP. To study correlation due to sharing alleles IBD at marker positions, we used genotypes at five makers located in a region earlier identified as possibly harboring a genetic factor that plays a role in the distribution of DM2 in this set of families (Aulchenko et al., 2003). We also studied body mass index (BMI) to illustrate the use of the score statistic on quantitative traits. Age and sex specific distributions were obtained from the Rotterdam Study (Hofman

et al., 1991), a large population-based follow up study.

2.2 Methods

The generalized linear mixed model (GLMM)

Let $Y_k = (Y_{k,1}, ..., Y_{k,n_k})'$ be the response vector of a family with n_k relatives. We assume that each component of Y_k has a distribution f in the exponential family with a dispersion parameter ϕ. Let $\mu_{k,i}$ be the expected value of $Y_{k,i}$ which may depend on a vector of covariates $x_{k,i}$. To model genetic correlation among family members, the following generalized linear mixed model (GLMM) is proposed

$$\begin{aligned} Y_{k,i} &\sim f(\mu_{ki}, \phi) \\ \mu_{k,i} &= E(Y_{k,i}|x_{k,i}, u_{k,i}) = h^{-1}(x_{k,i}\beta + u_{k,i}), \end{aligned}$$

with h a link function (McCullagh and Nelder, 1989). For a quantitative outcome h may be the identity function and for a qualitative outcome h may be the logit function. The vector β is a vector of regression coefficients and $u_k = (u_{k,1}, ..., u_{k,n_k})'$ is a random vector with mean 0 and covariance matrix $\tau^2 R_k$. If $\tau^2 = 0$ the variables $Y_{k,i}$ are independent and the model $\mu_{k,i} = h^{-1}(x_{k,i}\beta)$ is simply a generalized linear model (McCullagh and Nelder, 1989).

For u_k equal to an additive genetic effect, the correlation structure R_k has elements $R_{k,ij}$ equal to the coefficients of relationships (Sham, 1998). The coefficient of relationships can be written as follows

$$R_{k,ij} = \pi^2_{ij} + \frac{1}{2}\pi^1_{ij}, \tag{2.1}$$

with $\pi^l_{ij}, l = 0, 1, \text{or } 2$ the probability of sharing l alleles identical by descent (IBD) between individual i and j. When genetic markers are typed, correlation due to sharing alleles IBD at marker positions may be of interest. For a certain marker, we propose to extend the correlation structure (2.1) to the following set of correlation structures:

$$R_{k,ij} = \rho(\tilde{\pi}^2_{ij} + \frac{1}{2}\tilde{\pi}^1_{ij}) + (1-\rho)(\pi^2_{ij} + \frac{1}{2}\pi^1_{ij}), \text{ with } 0 \leq \rho \leq 1, \tag{2.2}$$

with $\tilde{\pi}^l_{ij}, l = 0, 1, \text{or } 2$ the conditional probability of sharing l alleles IBD between relative i and j at the marker locus, given the marker data and the family structure. Now the random effect u_k represents correlation due to tight linkage between a marker and a gene involved in the etiology of the trait and due

to residual genetic factors. The proportion of the additive genetic variance explained by a locus is modelled by the parameter ρ. For $\rho = 0$, R_k equals correlation structure (2.1) and no contribution of the locus to the genetic correlation is modelled. For $\rho = 1$, R_k is the mean IBD sharing matrix at a locus and hence this model assumes that the total genetic variance is explained by this locus.

The choice of ρ depends on the population and trait studied. For a genetically homogeneous population where only one gene is expected to be involved in the etiology of the studied trait, ρ may be set to 1. For complex genetic traits studied in the general population, small values of ρ may be of interest. Alternatively plots of ρ versus the p-value may be made to study if the p-value decreases by adding the correlation due to IBD sharing to the correlation structure (1).

For randomly chosen families, Houwing-Duistermaat et al. (1995) used the GLMM above and derived the score statistic to test the null hypothesis H_0 : $\tau^2 = 0$ of no correlations between relatives of randomly chosen families. The statistic is given by

$$\sum_{1 \le k \le m} Q_k = \sum_{1 \le k \le m} (Y_k - \mu_k)' R_k (Y_k - \mu_k),$$

with m the number of families.

Score statistic for proband family design

Since the score statistic adds over independent families, it suffices to give the statistic and its distribution only for a single family. Therefore we drop the family index k. Suppose a family has n members with n_p probands related to ascertainment and $n_r = n - n_p$ remaining relatives not related to ascertainment. Let $Z_i = Y_i - \mu_i$ with μ_i the known mean. To let μ_i depend on covariates a marginal model may be used (Diggle, 1994). Now let the first n_p observations belong to the set of probands then we can write $Z' = (Z^{p\prime}, Z^{r\prime})$ with $Z^p = (Z_1, ..., Z_{n_p})'$ and $Z^r = (Z_{n_p+1}, ..., Z_n)'$. Analogously we decompose the correlation matrix R into four blocks, $R = \begin{pmatrix} R^{pp} & R^{pr} \\ R^{pr\prime} & R^{rr} \end{pmatrix}$. The logarithm of the conditional likelihood $l(Z|Z^p, \tau)$ of Z given the outcomes of the probands Z^p is

$$\begin{aligned} l(Z|Z^p, \tau) &= l(Z|\tau) - l(Z^p|\tau) \\ &= \log(E_u[f(Z|u, \tau)]) - \log(E_u[f(Z^p|u, \tau)]). \end{aligned}$$

The corresponding score statistic to test $H_0 : \tau^2 = 0$ is obtained by taking the first derivative of $l(Z|Z^p, \tau)$ with respect to τ^2 at $\tau^2 = 0$

$$\frac{\partial}{\partial \tau^2} l(Z|Z^p, \tau^2)_{|\tau^2=0} = Q - E_{|\tau^2=0}(Q),$$

with Q a quadratic form equal to the stochastic part of the first derivative plus a constant. The statistic Q can be written as follows

$$Q = Z'RZ = Z^{r\prime}R^{rr}Z^r + 2Z^{p\prime}R^{pr}Z^r + Z^{p\prime}R^{pp}Z^p.$$

Note that the stochastic part of Q can be written as a sum of two statistics $L = 2Z^{p\prime}R^{pr}Z^r$ and $Q^r = Z^{r\prime}R^{rr}Z^r$. The linear term L measures correlation between probands and relatives. The quadratic form Q^r is the score statistic applied to the relatives and ignoring the probands. If the outcome of the proband is unknown, no information is available about the correlation between the proband and the family members. Furthermore from the formula for Q it is clear that information is lost when more relatives are allocated to the set of probands (see also Ewens & Shute, 1986).

Under the null hypothesis of no correlation, the conditional expectation $E(Q|Z_p)$ and variance $Var(Q|Z_p)$ of the statistic given the outcomes of the probands are

$$E(Q|Z^p) = E(Q^r) + Z^{p\prime}R^{pp}Z^p = \sum_{i=n_p+1}^{n} R_{ii}Var(Z_i) + Z^{p\prime}R^{pp}Z^p,$$

and

$$\begin{aligned} Var(Q|Z^p) &= Var(L|Z^p) + 2Cov(L; Q^r|Z^p) + Var(Q^r) \\ &= 4\sum_{j=n_p+1}^{n}(\sum_{i=1}^{n_p} Z_i R_{ij})^2 Var(Z_j) + 4\sum_{j=n_p+1}^{n}\sum_{i=1}^{n_p} Z_i R_{ij} E(Z_j^3) \\ &\quad + \sum_{i=n_p+1}^{n} R_{ii}^2(E(Z_i^4) - 3Var(Z_i)^2) + 2\sum_{i,j=n_p+1}^{n} R_{ij}^2 Var(Z_i)Var(Z_j). \end{aligned}$$

For binomially and normally distributed outcomes, formulae for the expectation and the variance of Q are given in the appendix. Asymptotically, the statistic $\frac{Q-E(Q|Z^p)}{\sqrt{Var(Q|Z^p)}}$ follows a standard normal distribution $N(0,1)$. Alternatively, the distribution of Q under $H_0 : \tau^2 = 0$ can be approximated by a scaled chi-square distribution $c\chi_v^2$ with the scale parameter c given by $c = \frac{Var(Q|Z^p)}{2E(Q|Z^p)}$ and the degrees of freedom v given by $v = \frac{2E(Q|Z^p)^2}{Var(Q|Z^p)}$ (le Cessie and van Houwelingen, 1995).

2.3 Simulation study

In order to study the performance of the $c\chi^2$ and normal distributions as approximations of the null distribution of the score statistic, we performed a simulation study. For sake of simplicity we used the data structure of our example of 33 families (see below). We generated 100,000 data sets of independently binomially distributed outcomes and 100,000 data sets of independently normally distributed outcomes. The score statistics were calculated using correlation structure (2.1) based on the coefficients of relationship. We also studied the performance of the distributions in a very small set of nine families.

In table 2.1, the actual p-values corresponding to a nominal p-value of 0.05, 0.01, 0.001 and 0.0001 are given. The results were in favour of the $c\chi^2$ distribution for both binomially and normally distributed outcomes. Even for the set of nine families, the $c\chi^2$ distribution performed very well.

TABLE 2.1: Type I error rate when using $c\chi^2$ distribution and normal distribution as approximation for the distribution of Q under the null hypothesis. The estimates are based on 100,000 simulations.

		33 families		9 families	
	nominal	$c\chi^2$	normal	$c\chi^2$	normal
Binomial (DM2)					
	0.05	0.0547	0.0606	0.0550	0.0649
	0.01	0.0143	0.0194	0.0137	0.0239
	0.001	0.0020	0.0041	0.0017	0.0070
	0.0001	0.0004	0.0011	0.0002	0.0019
Normal (BMI)					
	0.05	0.0538*	0.0615*	0.0566	0.0651
	0.01	0.0125*	0.0196*	0.0151	0.0233
	0.001	0.0016*	0.0047*	0.0027	0.0069
	0.0001	0.0002*	0.0011*	0.0004	0.0023

* based on 27 families

A data example

Description of families

To illustrate the score statistic, we used data from 79 patients with type 2 diabetes mellitus (DM2), their first-degree relatives and spouses (Aulchenko et al., 2003). These families were derived from the GRIP population (Genetic Research in Isolated Populations), an isolated village in the Southwest of the Netherlands. The GRIP population is described in detail elsewhere (Aulchenko et al., 2003; Vaessen et al., 2002; van Duijn et al., 2001). Probands are patients with DM2 treated by physicians participating in GRIP. Among the relatives are patients not related to ascertainment namely patients of other physicians and subjects who did not know that they have DM2. In a combined linkage and association study, a genome scan was carried out on these data and Aulchenko et al. (2003) found a borderline association between marker D3S3681 and DM2 (LOD score of 1.20, P=0.01).

For DM2 we analysed 33 families informative for linkage. One of these families was a combination of two nuclear families. Three families had probands with unknown disease status. In total 31 probands and 65 relatives were observed. The percentage of women was 60%. The mean age in years was 62 (range 45-94). We did not use subjects younger than 45 years, because we do not have information on the prevalence of DM2 for these age groups. In table 2.2 the number of families for combinations of number of affected relatives and number of observed relatives in the family are given. The mean size of the families was 2.9 (range 2 to 5) and the number of affected relatives per proband was 0.72 (range 0 to 4). The quantitative outcome body mass index (BMI) was known for a subset of 27 families with 46 relatives. BMI was only known for 7 probands. In this subset, the distributions of age and sex agreed with those of larger set of 33 probands. Also for this outcome we assumed that the family members are not related to ascertainment if they are not probands.

The age and gender specific distributions of DM2 and BMI were obtained using data from the Rotterdam Study (Hofman et al., 1991). The Rotterdam Study is a population based follow up study of the elderly with about 8000 subjects aged 55 and over. For both sexes, we fitted logistic and linear regression models to estimate the relationschip between age and DM2 and BMI respectively. The following marginal models were obtained

TABLE 2.2: Counts of families per combinations of number of affected relatives and number of observed relatives in the family.

		Number of relatives				
		1	2	3	4	total
Number of affected relatives	0	11	3	3	0	17
	1	3	4	3	0	10
	2	-	1	3	1	5
	3	-	-	0	0	0
	4	-	-	-	1	1
	total	14	8	9	2	33

$$logit(\mu(DM2)) = \begin{cases} -4.379 + 0.035 * age & \text{for women} \\ -3.529 + 0.025 * age & \text{for men} \end{cases} \quad (2.3)$$

$$\text{and } \mu(BMI) = \begin{cases} 25.48 + 0.018 * age & \text{for women} \\ 28.10 - 0.036 * age & \text{for men} \end{cases} \quad (2.4)$$

and $\sigma^2(BMI) = 13.62$. We used these models also for the distributions of DM2 and BMI for subjects aged between 45 and 55 years. In table 2.3, the observed and expected prevalence of DM2 and the observed and expected mean of BMI in the relatives are given. The expected values were computed using model (2.3) and (2.4) respectively. The prevalence of DM2 and the mean of BMI was higher than expected.

The conditional probabilities of sharing zero, one or two alleles IBD at marker D3S3681 and four informative proximal markers namely D3S1276, D3S3634, D3S1603, and D3S1271 were computed using the multipoint option in GENEHUNTER (Kruglyak et al., 1996) and using all available family members regardless of their age. Unfortunately no informative distal marker was available. The genetic distances between adjacent markers are 2.67, 2.67, 0.53, and 2.67 cM successively. The markers appeared to be highly informative (> 0.89), using the entropy as a measure of the informativeness (Kruglyak et al., 1996), hence also the Spearman's rank correlations between the estimated proportion of alleles shared IBD at each marker locus and the coefficient of relationship are rather small (< 0.53). The Spearman's rank correlations between the estimated proportion of alleles shared IBD at pairs of

TABLE 2.3: Observed and expected prevalence of DM2 and mean of BMI

	Observed	Expected*
Prevalence of DM2		
women (n=40)	0.35	0.10
men (n=25)	0.40	0.11
Mean of BMI (standard error)		
women (n=28)	29.27 (0.91)	26.40
men (n=18)	28.63 (0.93)	25.98

* expected values are based on the Rotterdam Study

markers varied from 0.74 (D3S3681 and D3S1271 at distance 8.54 cM) to 0.90 (D3S3634 and D3S1603 at distance 0.53 cM). Due to recombination between two physically close located markers D3S3634 and D3S1603, the estimated proportion of alleles shared IBD differed between these marker loci.

Results

We applied the score statistics to test for clustering of DM2 and BMI due to genetic factors. Correlation structure (2.1) based on familial relationship appeared to be highly significant for both traits ($P < 0.00001$). For testing correlation structure (2.2) for the five markers, plots of ρ versus minus log_{10} of p-value are given in figure 2.1 for DM2 and in figure 2.2 for BMI. All p-values were highly significant ($P < 0.00001$). Especially for DM2, adding correlation due to sharing allele IBD at marker D3S3681 to the familial correlation decreased the p-value for clustering.

2.4 Discussion

In this paper we proposed a score statistic for the proband family design to test for the presence of a prespecified correlation structure for binary and quantitative outcomes. The score statistic allows for adjusting of covariates. No assumption about the distribution of the random effects is made. Furthermore by conditioning on the trait value of all individuals related to ascertainment the method is robust to the ascertainment scheme. By means of a simulation study we showed that the $c\chi^2$ distribution performs well as an approximation of the distribution of the score statistic under the null hypothesis even in very small data-sets.

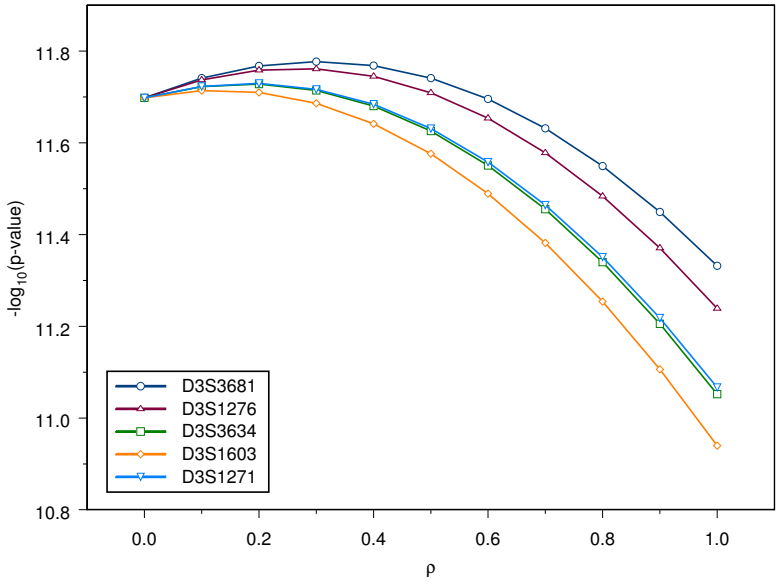

FIGURE 2.1: *For DM2 the p-values for testing correlation due to sharing alleles IBD at the five marker positions and residual genetic correlation for* $\rho = 0, 0.1, ..., 1$. *The parameter* ρ *models the proportion of genetic variances explained by IBD sharing.*

We analysed the clustering of DM2 and BMI in families of DM2 cases. Age and sex specific distributions of DM2 and BMI were obtained from the Rotterdam Study. The number of DM2 cases was higher than expected taking into account the age and sex distributions. Also the mean BMI was higher than expected in these families. This may indicate that genetic factors play a role in these families. Application of the score statistics indeed showed significant familial clustering of DM2 and BMI. Furthermore for DM2 adding IBD sharing at the five marker locations decreased the p-value. This decrease was most pronounced for marker D3S3681, which also showed some association with DM2 (Aulchenko et al., 2003). The next step in analysing these data will be estimation of the parameters. However the methodology has to be developed (see below).

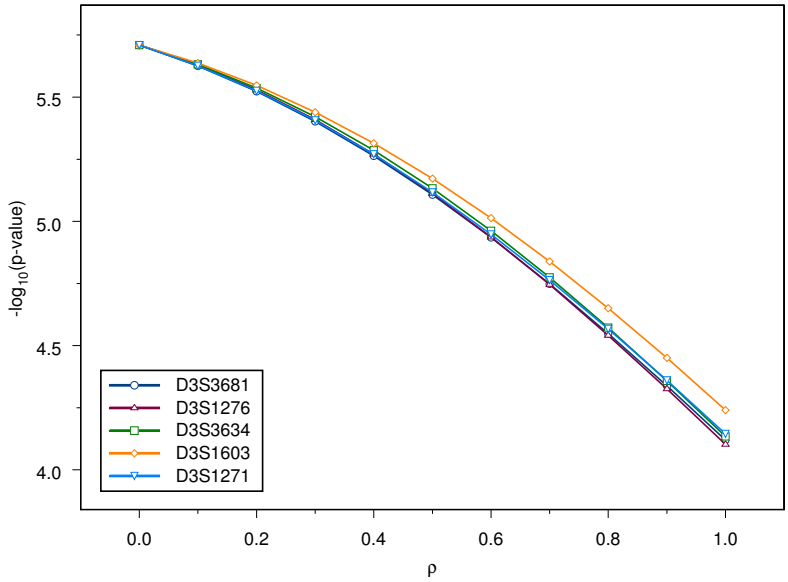

FIGURE 2.2: *For BMI the p-values for testing correlation due to sharing alleles IBD at the five marker positions and residual genetic correlation for* $\rho = 0, 0.1, ..., 1$. *The parameter* ρ *models the proportion of genetic variances explained by IBD sharing.*

In this paper we derived formulae for binomially and normally distributed outcomes. However the score statistic can be used for any distribution belonging to the exponential family. Furthermore we restricted ourselves to correlation due to additive genetic effects because for many complex traits, dominant effects are assumed to be small (Risch, 1990a). If dominant effects do exist the power will be only slightly reduced, because dominant effects only influence the correlation among pairs who can share two alleles IBD.

In addition to correlation due to any genetic component we also considered correlation partly due to excess sharing of alleles IBD in candidate regions. We made plots of ρ, the proportion of genetic variance explained by the locus versus the p-value. A decrease of the p-value at $\rho = 0$ suggests a role of a gene involved in the etiology of the trait linked to the marker locus. Note

that our statistic does not test the null hypothesis of no linkage. A formal test for linkage is a score statistic for $H_0 : \rho = 0$. For quantitative traits, this score statistic corresponds to the statistic derived by Putter et al. (2002). However for binary traits, derivation of the score statistic is complex due to the non linear relation between the outcome and the random effects. Nevertheless for both quantitative and binary outcomes, the test statistic Q provides insight in the underlying correlation structure and should be used before genetic parameters are estimated.

When a trait appears to be significantly correlated, the next step is to estimate the genetic parameters modelling the covariance structure. A natural framework of estimation methods are generalized estimating equations (GEE), because they do not fully specify the distribution (see Tregouet and Tiret (2000) and Ziegler et al. (1998) for reviews on application of these methods to family studies). For quantitative traits observed in random families, Stram et al. (1993) proposed GEE for segregation analysis. For binary traits, Liang and Beaty (1991) used these methods to study the dependence within families under the assumption that the families sampled are geometrically proportional to the number of affected family members. For a case-control family design, Zhao et al. (1998) derived GEE corresponding to the likelihood proposed by Whittemore (1995). To correct for ascertainment, Whittemore (1995) proposed to use different intercepts for probands than for the remaining family members. Further research is needed to extend these methods to the more general selection scheme as considered in this paper.

Methods of estimation should allow for more than one proband per family and also for different relationships between probands and relatives. Furthermore under the alternative, estimation of the parameters modelling the mean from data on the relatives may be biased (Pfeiffer et al., 2001). Hence to adjust for covariates, estimates of parameters should be obtained from other sources. Note that biased estimates of the regression parameters will affect the correlation between residuals (Diggle and Zegger (1994, p. 63-64); Verbeke and Molenberghs (1997, p. 120-122)) and consequently, the estimates of parameters modelling the covariance.

The statistic Q measures deviation from the mean as well as clustering. Hence if the mean is invalid the type I error is inflated. Commenges et al. (1995) proposed to estimate the parameters from the data on the relatives, which is valid under the null hypothesis. However for large sets of probands estimation of the parameters modelling the mean may not be possible. Furthermore for estimation of the genetic parameters, the means should

be known as pointed out above. It is natural to use estimates from other sources when our statistic is applied before models are fitted. Therefore we feel that it is important to know the effects of covariates on the trait in studied populations and to use this knowledge in analysing selected families aiming to elucidate the underlying genetic mechanisms. For our data example obtaining the age and gender distribution of DM2 and BMI from the Rotterdam Study seems to be reasonable since GRIP is a recently isolated population. We conclude that the score statistic is a good tool to study clustering of traits due to genetic factors within families selected via probands.

The analysis was performed using S-plus codes, which are available from: http://www.medstat.medfac.leidenuniv.nl/MS/

2.5 Appendix

For $Y_i \sim N(\mu_i, \sigma^2)$ the expectation of Q given Z^p is

$$E(Q|Z^p) = \sigma^2 trace(R^{rr}) = n\sigma^2,$$

and the variance of Q given Z^p is

$$Var(Q|Z^p) = 2\sigma^4 trace((R^{rr})^2) + 4\sigma^2 \sum_{j=n_p+1}^{n} (\sum_{i=1}^{n_p} Z_i R_{ij})^2.$$

For $Y_i \sim Bin(1, \mu_i)$ the expectation of Q given Z^p is

$$E(Q|Z^p) = \sum_{i=n_p+1}^{n} \mu_i(1-\mu_i),$$

and the variance of Q given Z^p is

$$\begin{aligned} Var(Q|Z^p) &= 4 \sum_{j=n_p+1}^{n} (\sum_{i=1}^{n_p} Z_i R_{ij})^2 \mu_j(1-\mu_j) \\ &+ \sum_{i=n_p+1}^{n} \mu_i(1-\mu_i)(1-6\mu_i+6\mu_i) \\ &+ 2 \sum_{i,j=n_p+1}^{n} (R_{ij})^2 \mu_i\mu_j(1-\mu_i)(1-\mu_j) \\ &+ 4 \sum_{j=n_p+1}^{n} \sum_{i=1}^{n_p} R_{ij} Z_i \mu_j(1-\mu_j)(1-2\mu_j). \end{aligned}$$