Linkage mapping for complex traits : a regression-based approach

Lebrec, J.J.P.

Chapter 3

Selection Strategies for Linkage Studies using Twins

Abstract

Genetic linkage analysis for complex diseases offer a major challenge to geneticists. In these complex diseases multiple genetic loci are responsible for the disease and they may vary in the size of their contribution; the effect of any single one of them is likely to be small. In many situations, like in extensive twin registries, trait values have been recorded for a large number of individuals, and preliminary studies have revealed summary measures for those traits, like mean, variance and components of variance, including heritability.

Given the small effect size, a random sample of twins will require a prohibitively large sample size. It is well known that selective sampling is far more efficient in terms of genotyping effort.

In this paper we derive easy expressions for the information contributed by sib pairs for the detection of linkage to a quantitative trait locus (QTL). We consider random samples as well as samples of sib pairs selected on the basis of their trait values. These expressions can be rapidly computed and do not involve simulation. We extend our results for quantitative traits to dichotomous traits using the concept of a liability threshold model.

We present tables with required sample sizes for height, insulin levels and migraine, three of the traits studied in the GenomEUtwin project.

This chapter has been published as: H. Putter, J. Lebrec and J.C. van Houwelingen (2003). Selection Strategies for Linkage Studies using Twins. *Twin Research* **6** (5), 377–382.

3.1 Introduction

Genetic linkage analysis (gene mapping) has proved to be a powerful tool for the identification of genes responsible for monogenic inherited diseases such as Huntington disease and cystic fibrosis. The diseases for which the genetic basis has not yet been unravelled do not display a one-to-one correspondence between a single gene and disease status. In these complex diseases, multiple genetic loci are responsible for the disease and these genetic loci may vary in the size of their contribution, they may interact with each other and with external, environmental factors. The effect of any single one of these genes is likely to be small [Risch, 2000].

The GenomEUtwin project comprises a very large source of twins, through the union of a number of large twin registries in different countries in Europe. For the majority of these twins, data on a number of traits of interest have already been recorded. Examples include quantitative traits like height, BMI, risk factors for cardiovascular disease and qualitative traits like migraine, diabetes. Some of these traits are recorded repeatedly over time and require methods for longitudinal data, others can be thought of as having an age of onset and can be treated like survival data.

The first step in unravelling the genetic basis of a disease is to undertake a heritability study. Twin studies are ideally equipped for this purpose, because of the inherent matching for age and other environmental factors, and because of the differential degree of shared genetic variance between monozygotic (MZ) and dizygotic (DZ) twins [Boomsma et al., 2002]. For many quantitative traits of interest, twin studies (or similar studies) have given information on the distribution of the trait in the target population, in particular their mean and variance, and on the heritability.

In the planning phase of a linkage study, one of the important issues is the choice of sib pairs to be included in a scan. The good news is that for large twin registries, the number of phenotypes is in principle adequate even to detect very small genetic effects. Unfortunately, given the anticipated small genetic effect at any one disease locus, a random sample to achieve 80% power is most probably prohibitively large in terms of genotyping effort, even with the current high throughput genotyping technologies. Eaves and Meyer [1994] and Risch and Zhang [1995] showed that similar power to large random samples can be obtained by selecting only a small subset of

extreme discordant pairs. Many studies have later refined these recommendations, giving, under an assumed model, optimal selection strategies for linkage studies. The drawback of these studies is that they typically require simulation and fail to give quick, easy and insightful assessments of the amount of information that a given sib pair is expected to contribute.

In this paper, it is our aim to outline easily computable information content numbers for twins in the context of linkage twin studies for complex diseases. We start in Section 3.2 by considering quantitative traits, with given heritability, mean and variance, assuming that the effect of the quantitative trait locus is small. We replace much of the simulation employed in the above papers by explicit calculation, resulting in particularly easy expressions for the information content for DZ sib pairs. The result is an easy expression closely related to optimal Haseman-Elston regression [Sham and Purcell, 2001] and the score function for the QTL variance in a variance components model [Putter et al., 2002]. We then show in Section 3.3 how the concept of a latent underlying quantitative trait can be used to extend these results to dichotomous traits. Section 3.4 discusses issues like extended pedigrees and dominance variance.

3.2 Selection strategies for quantitative traits

Random sampling

Starting point of our selection procedure for quantitative traits is the variance components model [Schork, 1993; Amos, 1994]. We assume that the traits have been standardised so as to have zero mean and unit variance. For a DZ twin sharing i alleles identical by descent (IBD) at a particular marker locus, the distribution of their phenotypes $\mathbf{x} = (x_1, x_2)$ is assumed to follow a bivariate normal distribution with mean vector 0 and covariance matrix

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \rho + \frac{i-1}{2}\gamma \\ \rho + \frac{i-1}{2}\gamma & 1 \end{pmatrix} .$$

Here ρ and γ represent the proportion of this variance that can be attributed to shared components and the quantitative trait locus respectively. The parameter ρ is half of the heritability (h^2) plus the proportion of common environment variance, c^2. In what follows we consider DZ twins, since MZ twins are not informative for linkage.

We shall refer to DZ twins as sib pairs in the sequel; for our purposes there is no distinction between sib pairs and DZ twins.

The amount of information I at $\gamma = 0$ contributed by one sib pair is given by

$$I = \frac{1}{8}\frac{1+\rho^2}{(1-\rho^2)^2} \ . \tag{3.1}$$

This formula has been derived by Williams and Blangero [1999] and is a special case of our equation (3.5). The factor $1/8$ represents the variance of $\hat{\pi}$ for sib pairs for a fully informative marker [Rijsdijk et al., 2001]. This implies that an estimate of γ based on a random sample of n sib pairs will have a standard error of $\mathrm{se}(\hat{\gamma}) = \frac{1}{\sqrt{nI}}$, in the absence of nuisance parameters. This fact can be used to determine the number of sib pairs required to achieve power $1-\beta$ to detect linkage with a QTL effect size γ, using a significance level α,

$$n = \frac{(z_\alpha + z_\beta)^2}{I\gamma^2} \ . \tag{3.2}$$

Here z_α denotes the $1-\alpha$ percentile of the standard normal distribution. For a power of 80% and a significance level of 0.0001, corresponding to a lod-score of 3, this leads to $n = \frac{20.8}{I\gamma^2}$. Graphs for different values of ρ are shown in Figure 3.1.

For a quantitative trait like height, with an estimated heritability of 0.80 and an estimated common environment variance $c^2 = 0.1$, and hence a value or $\rho = 0.5$, we need to genotype approximately 7500 sib pairs or 15000 individuals to detect linkage with a moderate QTL effect of $\gamma = 0.1$. Clearly, this is not feasible, even with the current high-throughput genotyping technology.

Selective sampling

Risch and Zhang [1995] suggested selecting sib pairs for genotyping on the basis of their trait values and showed that considerably higher efficiency can be obtained by selecting extreme discordant sib pairs. Later, these recommendations have been refined, most of the papers employing simulation to calculate the information content of a sib pair [Dolan and Boomsma, 1998b; Cherny et al., 1999]. A noteworthy exception is the paper by Purcell et al. [2001], where the information content is obtained through an exact calculation that considers all possible genotypes at the quantitative trait locus. We show below a simple approach that can also be used to obtain explicit

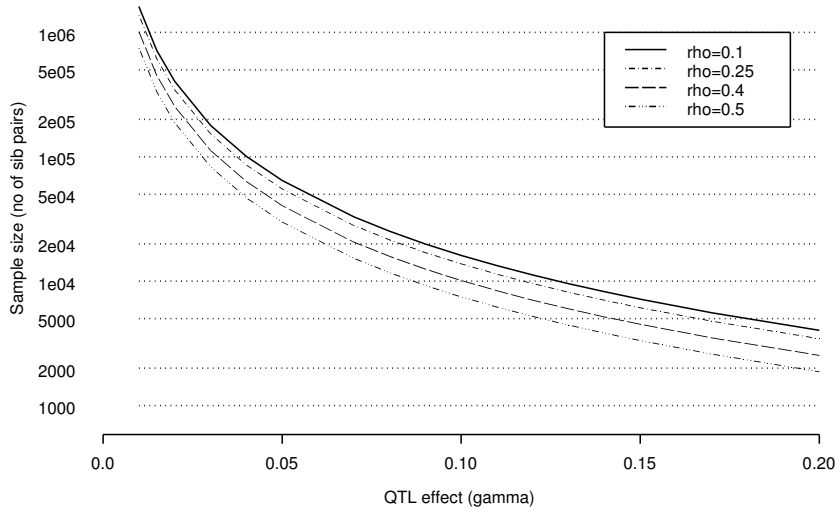

Figure 3.1: Number of sib pairs needed in a random sample to detect linkage to a quantitative trait for different values of ρ and γ. Power is 80%; significance level = 0.0001, corresponding to a lod-score of 3. For 50%, 60% and 70% power respectively, required sample sizes decrease by a factor of 1.50, 1.32 and 1.16 respectively.

expressions for the information content for a number of common designs without the need to do simulations.

The variance components model specifies the conditional distribution of the phenotypes, given the genotypes (IBD-sharing). When dealing with selected samples, it is more natural to invert the reasoning and to think of the phenotypes as given [Sham et al., 2000]. This approach is common for the analysis of dichotomous traits. Let z denote the number of alleles shared IBD by the twins at the marker locus, and $\hat{\pi}$ the proportion of alleles shared IBD. Since it is anticipated that the effect of any single gene is small, we use a linear expansion in γ along with Bayes' theorem to obtain, neglecting terms of smaller order than γ,

$$\begin{aligned} P(z=0|\mathbf{x},\gamma,\rho) &= \frac{1}{4} - \frac{\gamma}{8}C(\mathbf{x},\rho) \ , \\ P(z=1|\mathbf{x},\gamma,\rho) &= \frac{1}{2} \ , \\ P(z=2|\mathbf{x},\gamma,\rho) &= \frac{1}{4} + \frac{\gamma}{8}C(\mathbf{x},\rho) \ , \\ E(\hat{\pi}|\mathbf{x},\gamma,\rho) &= \frac{1}{2} + \frac{\gamma}{8}C(\mathbf{x},\rho) \ . \end{aligned} \tag{3.3}$$

Here,

$$C(\mathbf{x},\rho) = \frac{1}{(1-\rho^2)^2}\left((1+\rho^2)x_1x_2 - \rho(x_1^2+x_2^2) + \rho(1-\rho^2)\right)$$

is the "optimal Haseman-Elston " function [Sham and Purcell, 2001], which was shown to be the score function for the parameter γ in the variance components model [Putter et al., 2002]. Values of $C(\mathbf{x},\rho)$ range from negative to positive. Details of the derivation and extension to general pedigrees can be found in Lebrec et al. [2004].

This observation suggests using a regression method like the Haseman-Elston regression method, as already proposed by Sham et al. [2002], for the analysis of selected samples. The regression for sib pairs amounts to the inverse of the optimal Haseman-Elston regression, namely regressing $\hat{\pi}$ on $C(\mathbf{x},\rho)$. A test for linkage in this setting is a one-sided test for a positive slope in this regression. Indeed, for the case of sib pairs, our results coincide with those found in Sham et al. [2002].

In the context of regression, simple rules are available for selecting samples on the basis of the explanatory variables: since the square of the standard error of the slope of a regression of y on x is inversely proportional to $\sum(x_i-\bar{x})^2$, values of x should be

chosen as widely spaced as possible. This means that sib pairs with extreme values of $C(\mathbf{x}, \rho)$ should be selected for genotyping.

More formally, the optimal Haseman-Elston function $C(\mathbf{x}, \rho)$ determines the information of a sib pair with trait values x_1 and x_2. It is given by

$$I(\mathbf{x}, \rho) = \frac{1}{8} C^2(\mathbf{x}, \rho) \ , \tag{3.4}$$

and was obtained by Sham and Purcell [2001].

This information number is exact (at $\gamma = 0$), in contrast to the approximations used in the conditional distribution of IBD-sharing above. Figure 3.2 shows the distribution of information in a hypothetical population of standardised bivariate normal trait values with $\rho = 0.5$. Pairs are classified according to whether their information content is ranked in the top 5%, between 5% and 10% or in the remainder (i.e., not belonging to the 10% most informative). It clearly shows that both the extreme discordant and the extreme concordant pairs are most informative. The majority of the most informative pairs is discordant; in the top 5%, only about 15% is concordant, in the 5% to 10% category, about 35% is concordant.

For sib pairs chosen such that their trait values lie within a sampling region R, the average information can be computed by integrating over that region, weighted by the probability of the trait values:

$$I(R, \rho) = \int_R I(\mathbf{x}, \rho) \varphi_0(\mathbf{x}, \rho) d\mathbf{x} / \int_R \varphi_0(\mathbf{x}, \rho) d\mathbf{x} \ . \tag{3.5}$$

Here $\varphi_0(\mathbf{x}, \rho)$ denotes the bivariate normal density with mean 0, variance 1 and covariance ρ. Random sampling is a special case of this formula, since it is straightforward to show that when R is the full two-dimensional space, $I(R, \rho) = \frac{1}{8} \frac{1+\rho^2}{(1-\rho^2)^2}$. In order to select e.g. the 5% most informative sib pairs, R is the region of (x_1, x_2)-pairs with $C(x_1, x_2, \rho) \geq C_0$, where C_0 is chosen in such a way that this probability equals 5% under the null hypothesis.

Sampling over a region of sib pair trait values R, the number of sib pairs required to achieve power $1 - \beta$ to detect linkage with a QTL effect size γ, using a significance level α, then equals

$$n = \left(\frac{z_\alpha + z_\beta}{\gamma} \right)^2 / I(R, \rho) \ . \tag{3.6}$$

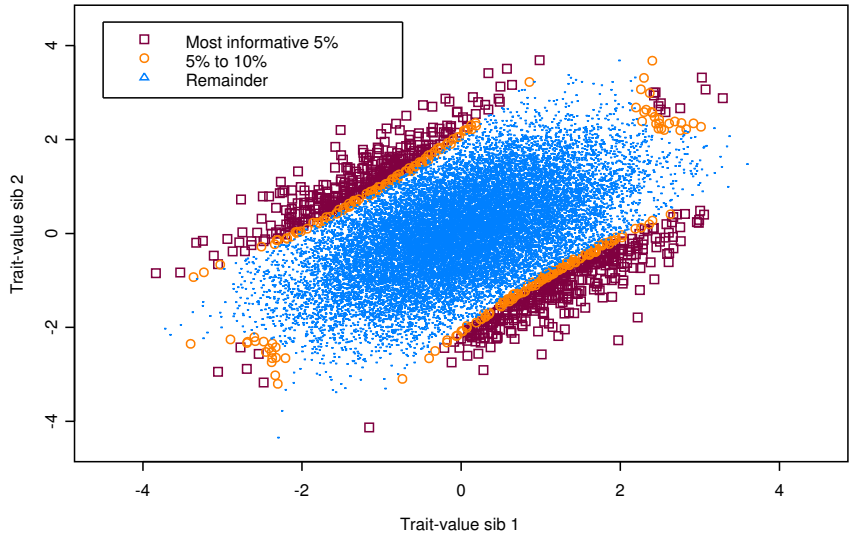

Figure 3.2: Scatterplot of trait values. Pairs are classified according to whether their information content is ranked in the top 5%, between 5% and 10% or in the remainder (not belonging to the 10% most informative).

	Height ($\rho = 0.5$) $h^2 = 0.80, c^2 = 0.10$					Insulin levels ($\rho = 0.35$) $h^2 = 0.40, c^2 = 0.15$				
QTL variance		Selection %					Selection %			
proportion (γ)	Random	10	5	2.5	1	Random	10	5	2.5	1
0.01	748180	105903	66537	43899	27648	1141429	165448	105502	71831	45494
0.02	187045	26476	16634	10975	6912	285357	41362	26375	17958	11373
0.05	29927	4236	2661	1756	1106	45657	6618	4220	2873	1820
0.10	7482	1059	665	439	276	11414	1654	1055	718	455

Table 3.1: The number of sib pairs needed to achieve 80% power to detect linkage to a quantitative trait with a significance level $\alpha = 0.0001$, for different values of γ (proportion of the variance explained by the quantitative trait locus). Height and insulin levels, two traits studied in the GenomEUtwin project are considered.

Table 3.1 shows the impact of these results on the number of sib pairs required for height and insulin levels, two quantitative traits studied in the GenomEUtwin project. For instance, for height, with a QTL variance proportion $\gamma = 0.10$, with a selection percentage of 1%, only 276 sib pairs need to be genotyped, but the trait values of 27,600 sib pairs need to be available, more than 3.5 times the amount needed for random selection. This is one reason not to go for a too restrictive selection percentage. Another, more compelling reason, is that with extreme selection percentages, the normality of the population trait values will become a crucial issue.

3.3 Selection strategies for dichotomous traits

For dichotomous traits it is convenient to think of the disease as being determined by an underlying latent quantitative trait (liability). When the value of this quantitative trait exceeds a threshold t, the individual is affected, otherwise unaffected. The threshold t is determined by the prevalence of disease K in the population of interest, through $t = \Phi^{-1}(1 - K)$, where Φ is the the distribution function of a standard normal variable. In a heritability study using twins, the heritability is estimated from the affection states of the the twins using the tetrachoric correlation of an underlying bivariate normal variable with zero mean and unit variance. The normal liability model is primarily a statistical convenience; if in reality there is no underlying normal liability in risk for an ordinal or dichotomous trait, then the model will be wrong.

The tools of Section 3.2 can be used to determine the information contributed by a twin with two affected (AA), one affected, one unaffected (AU), and two un-

	Trait I			Trait II		
latent QTL variance	$K = 5\%, \rho = 0.5$			$K = 20\%, \rho = 0.5$		
proportion (γ)	AA	AU	UU	AA	AU	UU
0.01	270122	***	***	962936	***	***
0.02	67531	649982	***	240734	403089	***
0.05	10805	103997	***	38517	64494	277326
0.10	2701	25999	***	9629	16124	69331

Table 3.2: The number of sib pairs needed to achieve 80% power to detect linkage to a dichotomous trait with a significance level $\alpha = 0.0001$, for different values of γ (proportion of the variance explained by the latent quantitative trait locus). The prevalence K and heritability approximately match that of migraine in men and women respectively. AA, AU and UU denote sib pairs with two affected, one affected and one unaffected, and two unaffected sibs respectively. *** denotes more than one million sib pairs needed.

affected (UU), given prevalence K, and tetrachoric correlation ρ (determined by the heritability). This information is

$$\frac{1}{8}\left\{\int_R C(\mathbf{x},\rho)\varphi_0(\mathbf{x},\rho)d\mathbf{x} \Big/ \int_R \varphi_0(\mathbf{x},\rho)d\mathbf{x}\right\}^2 , \tag{3.7}$$

where R is the region of (x_1, x_2)-pairs with $x_1 \geq t, x_2 \geq t$ (AA), $x_1 \geq t, x_2 < t$ (AU) or $x_1 < t, x_2 < t$ (UU). From equation (3.3) it can be seen that the expected value of $\hat{\pi}$, conditionally given that $\mathbf{x} \in R$ equals $\frac{1}{2} + \frac{\gamma}{8}\mathbf{E}(C(\mathbf{x},\rho) \,|\, \mathbf{x} \in R)$; the expression in brackets in the above expression is precisely this conditional expectation of $C(\mathbf{x},\rho)$ given $\mathbf{x} \in R$. Power calculations for dichotomous traits are very similar to (but not entirely the same as) quantitative traits using the liability threshold approach; the sampling region is now determined by affection status rather than observed trait values and does not have optimal form as in Figure 3.2. Table 3.2 shows that for dichotomous traits with low prevalence, AA sib pairs are most powerful, for traits with moderate to high prevalence, AU sib pairs however may also be quite informative.

3.4 Discussion

In this paper we have shown a simple approach to obtain explicit expressions for the information that a twin is expected to contribute towards detecting linkage to a quantitative trait. This information is based on the trait values and known values for the variance components of the trait. To achieve a given power to detect linkage to a quantitative trait with a given significance level and an anticipated proportion of the variance explained by the quantitative trait locus, the required number of sib pairs is straightforward to calculate. The expression extends to dichotomous traits through the concept of a liability, a latent underlying quantitative trait.

Earlier work uses simulation to calculate the information content of a sib pair and the number of sib pairs needed to achieve a given power [Dolan and Boomsma, 1998b; Cherny et al., 1999; Purcell et al., 2001]. For sib pairs, simulation can be replaced by calculation, as outlined below. These calculations are well known for random samples [Williams and Blangero, 1999; Rijsdijk and Sham, 2000; Rijsdijk et al., 2001] and have been pioneered for selected samples for the case of sib pairs [Sham and Purcell, 2001] and more implicitly for general pedigrees in Sham et al. [2002]. They have been implemented in `MERLIN` [Abecasis et al., 2002] through the command `MERLIN-regress`. The way they have been derived, by considering the conditional distribution of the IBD-sharing, given the phenotypes [Sham et al., 2000, 2002], also suggests methods for analysing selected samples. This is the subject of ongoing research in our group.

All expressions in Sections 3.2 and 3.3 are valid for DZ twins (sib pairs) only. It is well known however that for random samples sibships of larger sizes can achieve considerably more power than sib pairs [Dolan et al., 1999]. In a sense, a larger sibship constitutes a collection of sib pairs, and indeed the amount of information is roughly proportional to the number of sib pairs [Dolan et al., 1999; Williams and Blangero, 1999] in the sibship. Also for selective sampling, sib pairs could still be collected, even though they belong to a larger sibship. The direction taken in Section 3.2 does not readily extend to larger sibships or general pedigrees. However, the resulting expressions can be generalised more formally using efficient score functions. This approach is followed in Lebrec et al. [2004].

The score approach will also yield information content numbers for general pedi-

grees. These information content numbers can be computed in principle, but in practice the size of the pedigree may limit the calculations. Including parental information may result in a modest increase in power [Williams and Blangero, 1999]; arguably more important is the use of parental genotypes in other stages; it will increase precision of IBD-information, it can be used in quality control, and it may increase power in association studies.

The presence of dominance variance in the variance components model adds a parameter δ specifying the proportion of variance due to dominance variance of the QTL. The standardised traits of a sib pair sharing i alleles IBD will have covariance matrix

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \rho + \frac{i-1}{2}\gamma + (\mathbf{1}_{\{i=2\}} - \frac{1}{4})\delta \\ \rho + \frac{i-1}{2}\gamma + (\mathbf{1}_{\{i=2\}} - \frac{1}{4})\delta & 1 \end{pmatrix} .$$

For complex diseases, both γ and δ will be small, and similar calculations as in Sections 3.2 and 3.3 can be made in this case as well. The number of sib pairs needed to achieve a given power to detect linkage to a quantitative trait with a given significance level α now depends on both γ and δ through the functions $C(\mathbf{x}, \rho)$. In the case of a rare recessive allele, selection based on $C(\mathbf{x}, \rho)$ may no longer be fully informative Purcell et al. [2001]. Otherwise, dominance variance will not have a strong influence on selection, but it can influence the power.

The approach to power calculations that we took in this paper (calculating the Fisher information in an inverted variance components model, where the distribution of IBD sharing given the trait values is considered) is intimately tied to the method of analysis to be used later. As mentioned earlier, this is the subject of ongoing research in our group, but restricting the discussion to sib pairs, we note the following. It is assumed that trait values are normally distributed and have been standardised to have zero mean and unit variance. This standardisation entails subtracting the mean and dividing by the standard deviation, in the absence of covariates. Covariates can also be incorporated into both the power calculations and the analysis. Then in the standardisation the covariate values and the estimated regression coefficients (in the population!) are used instead of a common mean. Covariates can also be incorporated into the analysis of dichotomous traits; in this case not all affected sib pairs for instance will have the same C_{AA} value, but this value will now depend on the

covariate values of the sib pair. When data are not initially normally distributed, a transformation can be used in the population data to obtain approximate normality. Even in populations where the trait values are reasonably normally distributed, we think it is wise to robustify the analysis anyway, by giving sib pairs with extremely high $C(\mathbf{x}, \rho)$ values a lower weight in the inverse regression.