Linkage mapping for complex traits : a regression-based approach

Lebrec, J.J.P.

Chapter 4

Genomic Control for Genotyping Error in Linkage Mapping for Complex Traits

Abstract

It has been suggested that genotyping error could dramatically affect the evidence for linkage, particularly in selective designs. Using the regression-based approach to linkage, we quantify the effect of simple genotyping error models under specific selection schemes for sib pairs. We show for example, that in extremely concordant designs, genotyping error leads to over-pessimistic inference whereas it leads to increased type I error in extremely discordant designs. Perhaps surprisingly, the effect of genotyping error on inference is most severe in designs where selection is least extreme. We suggest a modification of the linkage testing procedure that accounts for genotyping errors based on a genomic estimate of the error rate.

This chapter has been submitted as: J. Lebrec, H. Putter, J.J. Houwing-Duistermaat and H.C. van Houwelingen. Genomic Control for Genotyping Error in Linkage Mapping for Complex Traits.

4.1 Introduction

In the search for genetic determinants of complex traits, the use of selective designs appears to be the only way to gain sufficient power to detect typically small gene effects in linkage studies. A few authors have shown by simulation that the impact of genotyping error on evidence for linkage could be particularly severe in affected sib-pair (ASP) designs [Douglas et al., 2000; Abecasis et al., 2001], virtually masking most of the evidence for linkage. The impact of error on quantitative traits appears to be less dramatic in random samples, however it is unclear whether the same dramatic power losses hold in selected samples.

A method of choice is now emerging for the analysis of quantitative traits arising from selected sib pairs. It boils down to a regression through the origin of excess identical by descent (IBD) sharing on a function of the trait value, whose slope is an estimate of the linkage parameter. It was first proposed by Sham and Purcell [2001] and turns out to be equivalent to a score test [Tang and Siegmund, 2001]. By use of simple genotyping error models (*population frequency error model* and *false homozygosity model*), we show analytically what effects such error generating processes (occurring at rate ϵ per sib pair) induce for an idealized fully informative marker. It is shown that it results in a reduction of the slope estimate (i.e. of the estimated linkage parameter) by a factor $1 - \frac{\epsilon}{2}$ regardless of whether sib pairs are selected or not. Since the genotyping error rate ϵ is typically small, the previous effect on the linkage test is minimal. In addition to this slope effect, the regression's intercept is modified and this may have a much more consequent effect on the test for linkage depending on the sampling scheme used to select sib pairs. Surprisingly, this simple result allows us to predict that in extremely concordant (EC) sib pairs designs and in ASP designs, the effect of genotyping error will be milder as the selection becomes more extreme. In extreme discordant (ED) designs, the effect can in theory be either over-optimistic or pessimistic depending on the definition of discordance, the genotyping error rate and the true linkage effect; in practice however, for small QTL effect, the result will be over-optimistic inference. It is argued that the basic error generating mechanisms assumed provide reasonable approximations of real-life situations. Furthermore, results obtained under the assumption of complete IBD information can be qualitatively

extended to settings where information is incomplete.

Finally, we suggest a simple genomic control for genotyping error which can easily be incorporated into the usual linkage testing procedure. This article is organized as follows: in Section 4.2, we introduce some notations and briefly sketch the inverse regression approach to linkage, in Section 4.3, we describe some common error-generating processes, in Section 4.4, we show analytically what the effect of genotyping error can be on the IBD sharing distribution and its consequence for linkage testing. Section 4.4 is devoted to studying the impact of genotyping error in common selective designs. In Section 4.5, we argue that under certain assumptions regarding the error model, one can easily implement a linkage test that incorporates a genomic control for genotyping error.

4.2 Test for linkage in selected sib pairs

We assume that the sib pair phenotypic data $\mathbf{x} = (x_1, x_2)'$ have been adjusted for any relevant covariates (e.g. sex, age, country, ...) and have been standardized so that the (known) population mean, variance and sib-sib correlation are 0, 1 and ρ respectively. In addition, let's denote by π the proportion of alleles shared identical by descent (IBD) at a certain locus by the two sibs and by $\hat{\pi}$ its estimated value given the marker information available [Kruglyak et al., 1996; Abecasis et al., 2002]. The additive variance components model assumes that $\mathbf{x}$ given IBD information π follows a normal distribution with zero mean and variance-covariance matrix given by

$$\begin{pmatrix} 1 & \gamma(\pi - \frac{1}{2}) + \rho \\ \gamma(\pi - \frac{1}{2}) + \rho & 1 \end{pmatrix} ,$$

where γ denotes the proportion of total variance explained by the putative locus. Sham and Purcell [2001] first proposed the following approach for testing linkage: regression of the estimated excess IBD sharing $\hat{\pi} - \frac{1}{2}$ through the origin of a function of the squared difference and squared sum of sib-pair phenotype values C where

$$C(x_1, x_2, \rho) = \frac{(1+\rho^2)x_1x_2 - \rho(x_1^2 + x_2^2) + \rho(1-\rho^2)}{(1-\rho^2)^2} . \tag{4.1}$$

In a sample of n independent sib pairs with phenotypes $(x_{i1}, x_{i2})_{i=1,\ldots,n}$, the test is based upon the following z statistic

$$z = \frac{\sum_i (\hat{\pi}_i - \frac{1}{2})\ C(x_{i1}, x_{i2}, \rho)}{\sqrt{\sum_i \mathrm{var}_0(\hat{\pi}_i)\ C^2(x_{i1}, x_{i2}, \rho)}}\ ,$$

it is one-sided, only positive values of z being regarded as evidence for linkage. In other words, z_+^2 defined as being equal to 0 if $z \leq 0$ and to z^2 if $z > 0$ is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. For normal data, this is nothing but a score test [Tang and Siegmund, 2001] and therefore constitutes an asymptotically optimal test for linkage with small locus effect γ (see Lebrec et al. [2004] for a generalization of this score test in arbitrary pedigrees). This test is sometimes referred to as the optimal Haseman-Elston regression. In a numerical comparison of methods for selected samples, Skatkiewicz et al. [2003] and Cuenco et al. [2003] showed that this method had good properties in finite samples for extreme proband ascertained sib-pair and discordant sib-pair designs. One important feature of this regression when applied to selected samples (as far as power is concerned) is that it is constrained through the origin and this plays an important role in how genotyping error affects linkage. A different motivation for this regression through the origin was given in Putter et al. [2003] using a first order Taylor's approximation for the three IBD probabilities $\mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{x}, \gamma, \rho)$:

(4.2)

$$\begin{array}{rcl} \mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{x}, \gamma, \rho) & = & \left(\quad \mathbf{P}(\pi = 0 \,|\, \mathbf{x}, \gamma, \rho) \quad , \quad \mathbf{P}(\pi = \frac{1}{2} \,|\, \mathbf{x}, \gamma, \rho) \quad , \quad \mathbf{P}(\pi = 1 \,|\, \mathbf{x}, \gamma, \rho) \quad \right) \\ & \simeq & \left(\quad \frac{1}{4} - \frac{\gamma}{8} C(\mathbf{x}, \rho) \quad , \quad \frac{1}{2} \quad , \quad \frac{1}{4} + \frac{\gamma}{8} C(\mathbf{x}, \rho) \quad \right) \end{array} ,$$

with $C(\mathbf{x}, \rho)$ given by Formula (4.1) which implies $\mathbf{E}(\pi - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \rho) = \frac{\gamma}{8}\ C(\mathbf{x}, \rho)$ when IBD information is known with certainty. This approximation is valid for small quantitative trait locus (QTL) effect γ and will be used in Section 4.4.

4.3 Genotyping error models

We consider two mechanisms for the generation of errors in marker data, namely the *population frequency error model* and the *false homozygosity model* . In those two models, we consider a single marker with m alleles and further assume that a maximum of one allelic error per sib pair can be made and that this happens

with probability ϵ. This restriction to one error per sib pair is just a first order approximation, for small ϵ, of a process where all four alleles would be allowed to be independently erroneous and does not restrict the generalizability of our results.

The *population frequency error model* re-assigns the erroneous allele (chosen at random among the four forming the sib-pair genotype) to one of the possible m alleles with probability equal to population allele frequency. One mathematical advantage of this model is that the marginal distribution of alleles and genotypes is unaltered. The *false homozygosity model* keeps homozygotes unchanged but re-assigns heterozygotes to homozygotes with alleles equal to one of the two original alleles chosen according to probabilities proportional to population allele frequencies.

To our knowledge, *false homozygosity* is a common type of error: fairly rare alleles go un-reported in samples. The *population frequency error model* provides an approximation to a process whereby alleles are misread. Errors at the two alleles of a marker's genotype might be correlated, we do not consider this type of process in details here although the effect on linkage will be qualitatively the same as in the two other models. We refer the reader to Sobel et al. [2002] for a detailed exposé on genotyping error mechanisms. Note that the two models we have chosen have been used successfully in the past in order to identify potential genotyping errors [Douglas et al., 2000; Sobel et al., 2002].

4.4 Impact of genotyping error on linkage

Effect on IBD sharing

Tests for linkage are based on the IBD sharing distribution and although errors as described in Section 4.3 are made at the genotype level (G is read as G^ϵ), the effect of errors on linkage will be entirely mediated via the distortion of the IBD distribution (the true IBD status π of two siblings may be incorrectly inferred as π^ϵ). We are therefore interested in deriving the probability distribution $\mathbf{P}(\pi^\epsilon \,|\, \pi)$, this is done by conditioning on both the true and observed genotypes as follows:

$$\mathbf{P}(\pi^\epsilon \,|\, \pi) = \sum_{G^\epsilon} \mathbf{P}(\pi^\epsilon \,|\, G^\epsilon) \sum_{G} \mathbf{P}(G^\epsilon \,|\, G)\ \mathbf{P}(G \,|\, \pi)\ .$$

Let us consider the case of complete information. This can be conceptualized

by means of an idealized marker whose number of alleles is infinite, in particular identity by state (IBS) status is equivalent to identity by descent (IBD) status. The unordered genotypes of a sib pair can be partitioned into seven exclusive classes denoted ii/ii, ii/ij, ii/jj, ii/jk, ij/ij, ij/ik and ij/kl depending on the number of homozygous sibs in the pair and the number of distinct alleles in the sib-pair genotype. Sharing 0 alleles IBD corresponds to a sib-pair genotype of the ij/kl class, should an error occur according to the *population frequency error model* then one of the four alleles would be transformed into yet another type (since the number of alleles is infinite, the probability that the new allele is read as one of i, j, k or l tends to 0), therefore the sib pair genotype will remain in the ij/kl class and the observed IBD status π^ϵ will still be 0. For the same starting genotype, an error according to the *false homozygosity model* produces an ii/jk class and π^ϵ also equals 0 therefore $\mathbf{P}(\pi^\epsilon = 0 \,|\, \pi = 0) = 1$ whatever the genotyping error mechanism considered in Section 4.3. The same line of reasoning leads to $\mathbf{P}(\pi^\epsilon = 0.5 \,|\, \pi = 0.5) = 1 - \frac{\epsilon}{2}$, $\mathbf{P}(\pi^\epsilon = 0 \,|\, \pi = 0.5) = \frac{\epsilon}{2}$, $\mathbf{P}(\pi^\epsilon = 1.0 \,|\, \pi = 1.0) = 1 - \epsilon$, $\mathbf{P}(\pi^\epsilon = 0.5 \,|\, \pi = 1.0) = \epsilon$. Those results can be summarized by the transition matrix below, where the (i, j) element is equal to $\mathbf{P}(\pi^\epsilon = (j-1)/2 \,|\, \pi = (i-1)/2)$

$$\mathbf{P}(\boldsymbol{\pi}^\epsilon \,|\, \boldsymbol{\pi}) = \begin{pmatrix} 1 & 0 & 0 \\ \frac{\epsilon}{2} & 1 - \frac{\epsilon}{2} & 0 \\ 0 & \epsilon & 1 - \epsilon \end{pmatrix} .$$

The overall effect of genotyping error is thus to reduce the observed IBD sharing. In selected samples of extremely concordant sib pairs (EC) where linkage is evidenced by excess IBD sharing, it therefore seems logical to expect a decrease in power. Conversely, in selected samples of extremely discordant sib pairs (ED) where linkage is evidenced by reduction in IBD sharing, the test might lead to increased type I error. In Section 4.4, we quantify this bias in selective samples schemes for quantitative traits under the usual assumption of a normal variance components model.

Effect on linkage

In this section, we concentrate on the case where IBD information is complete. As exposed in Section 4.2, the test for linkage corresponds to a regression through the

origin of excess IBD sharing $\hat{\pi} - \frac{1}{2}$ on a function of phenotype values $\mathbf{C} = C(\mathbf{x}, \rho)$ with C as defined by Formula (4.1) i.e. it is based on the approximate relation

$$\mathbf{E}(\pi - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = \frac{\gamma}{8}\, C(\mathbf{x}, \rho)\ . \tag{4.3}$$

We show in the appendix that, in presence of genotyping error at rate ϵ, this relation is changed into

$$\mathbf{E}(\pi^{\epsilon} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = -\frac{\epsilon}{4} + (1 - \frac{\epsilon}{2})\ \frac{\gamma}{8}\ C(\mathbf{x}, \rho)\ . \tag{4.4}$$

If we were to know ϵ, we could correct for it in the regression and the loss in efficiency would only be due to the $1 - \frac{\epsilon}{2}$ term preceding γ and would therefore be minimal.

We may ignore genotyping error altogether. In the appendix, we derive a general expression (Equation (4.9)) for the probability of rejecting the null hypothesis of no linkage under this scenario. For small values of the error rate ϵ, the following first order approximation obtains

$$\Phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right) - \epsilon\ \mathcal{I}^{1/2} \left(\frac{\gamma}{2} + 2\frac{\overline{C}}{\overline{C^2}}\right) \times \phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right)\ , \tag{4.5}$$

where α is the nominal type I error rate for the linkage test with a true quantitative trait locus effect γ, $\overline{C}$ is the average of the $C(x_{i1}, x_{i2}, \rho)$ values (given by Equation (4.1)) among a sample of n sib pairs, $\mathcal{I} = \frac{n}{8}\ \overline{C^2}$ is the sample's Fisher's information for the linkage parameter γ, Φ is the cumulative density function of the standard normal distribution and ϕ is the corresponding density function. The first term $\Phi\left(\Phi^{-1}(\alpha) + \gamma \mathcal{I}^{1/2}\right)$ in this expression gives the value of this probability in absence of genotyping error while the second term is the deviation from this reference value; in particular, when $\gamma = 0$, it expresses the actual type I error as a deviation from the nominal type I error rate: $\alpha - 2\epsilon\ \frac{\overline{C}}{\overline{C^2}} \mathcal{I}^{1/2} \times \phi\left(\Phi^{-1}(\alpha)\right)$.

In extremely concordant (EC) designs, $\overline{C}$ is positive while it is negative in extremely discordant (ED) designs, inference will therefore be too conservative in EC designs and too liberal in ED designs. In random samples and under the variance components model, C is a score function hence $\mathbf{E}(C) = 0$ therefore its sample estimate $\overline{C}$ will be small and the effect of genotyping error will be minimal. The same finding would hold for any ascertainment scheme where $\overline{C} = 0$.

We now quantify the effect of genotyping error on power and type I error under specific designs. The distortion of the linkage test in presence of genotyping error

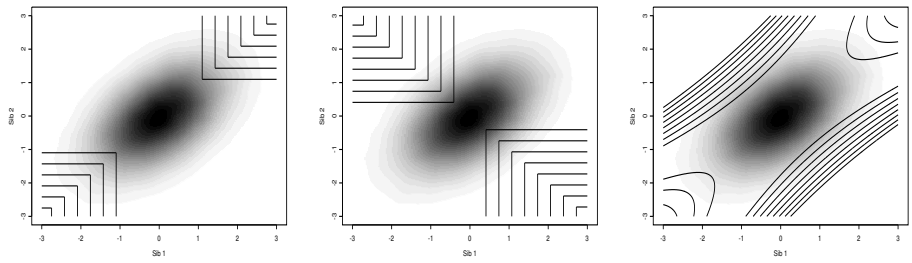

Figure 4.1: Three selective schemes: extremely concordant(EC), extremely discordant(ED) and most informative ($\mathcal{I}$) all for 10%. Joint distribution of sib trait values in gray scale for $\rho = 0.5$ (generated using the scatterplots function of Eilers and Goeman [2004])

depends heavily on the design-specific quantity $\overline{C}/\overline{C^2}$; given an ascertainment scheme corresponding to a certain region of the possible trait values, it is simple to use Monte Carlo methods to determine the expected $\overline{C}/\overline{C^2}$ value in that region. In table 4.1, we considered three different ascertainment schemes: extremely concordant (EC), extremely discordant (ED) and most informative ($\mathcal{I}$) as shown in Figure 4.1. For example, in the $\text{EC}_{10\%}$ scheme with sib-sib trait correlation $\rho = 0.5$, only sib pairs whose trait values (x_1, x_2) fulfill $x_1 > t$ and $x_2 > t$ or $x_1 \leq -t$ and $x_2 \leq -t$ where $t = t_{EC}(10\%, \rho = 0.5) = 0.136$ are retained (the value of t is such that on average 10% of the overall population is sampled). Analogously for ED, sib pairs whose trait values belong to regions defined by $x_1 > t$ and $x_2 \leq -t$ or $x_1 \leq -t$ and $x_2 > t$ are selected. The $\mathcal{I}$ scheme selects the most informative sib pairs determined using the quantiles of Fisher's information ($\mathcal{I} \propto C^2(x_1, x_2, \rho)$) distribution for the linkage parameter γ [Lebrec et al., 2004]. For example, if the percentage selected equals 10% and $\rho = 0.5$ then sib pairs whose trait values fulfill $C^2(x_1, x_2, \rho = 0.5) > 4.36$ would be selected. This sampling scheme combines both EC and ED sib pairs and constitutes a refinement of the so-called EDAC designs [Gu et al., 1996].

Table 4.1 allows us to draw three main conclusions relating to the main bias caused by the intercept mis-specification in the usual linkage testing procedure:

1. It is negative in EC designs and positive in ED designs, positive but without substantial influence for $\mathcal{I}$ designs,

ρ	sel.	EC	ED	$\mathcal{I}$	sel.	EC	ED	$\mathcal{I}$	sel.	EC	ED	$\mathcal{I}$
0.1	1%	0.27	-0.23	-0.07	10%	0.47	-0.40	-0.06	30%	0.65	-0.53	-0.04
0.2		0.29	-0.21	-0.13		0.50	-0.36	-0.11		0.69	-0.46	-0.07
0.3		0.30	-0.19	-0.15		0.52	-0.32	-0.14		0.71	-0.39	-0.09
0.4		0.31	-0.17	-0.14		0.53	-0.28	-0.16		0.69	-0.32	-0.11
0.5		0.32	-0.14	-0.12		0.52	-0.24	-0.17		0.62	-0.25	-0.11
0.6		0.31	-0.12	-0.10		0.47	-0.19	-0.15		0.50	-0.19	-0.10

Table 4.1: Average values for the $\overline{C}/\overline{C^2}$ term determining bias

2. It is more pronounced as the designs becomes less extreme for both EC and ED,
3. It is fairly independent of sib-sib trait correlation ρ for EC designs while it decreases with ρ for ED designs.

Overall, for small QTL effects γ, genotyping error will lead to conservative inference in EC designs and to liberal inference in ED designs. In Figure 4.2, we show the theoretical type I error rate and probability of rejecting the null hypothesis (obtained via Formula (4.9)) for different sampling schemes under perfect IBD information. We have used a QTL explaining 10% of the total trait variance, a trait sib-sib correlation equal to 0.3 and error rates equal to 0.01, 0.02 and 0.05. Although the power is not too badly affected at least for small error rates, genotyping error substantially affects the type I error rate, this may lead to far too liberal inference in ED designs, this deterioration of the size of the test becomes more acute as sample size increases.

Incomplete IBD information

We saw in Section 4.4 that genotyping error not only deteriorated the slope of the linkage signal but also introduced an intercept in the regression of excess IBD sharing on the optimal Haseman-Elston trait function $C(\mathbf{x}, \rho)$. In the case of complete information and at least for the *population frequency error model* and *false homozygosity model* , the perturbation caused by the error processes only depended on the error rate ϵ through the functions given in Equation (4.3). In real-life situations, IBD information is incomplete, but under the usual variance components additive model and

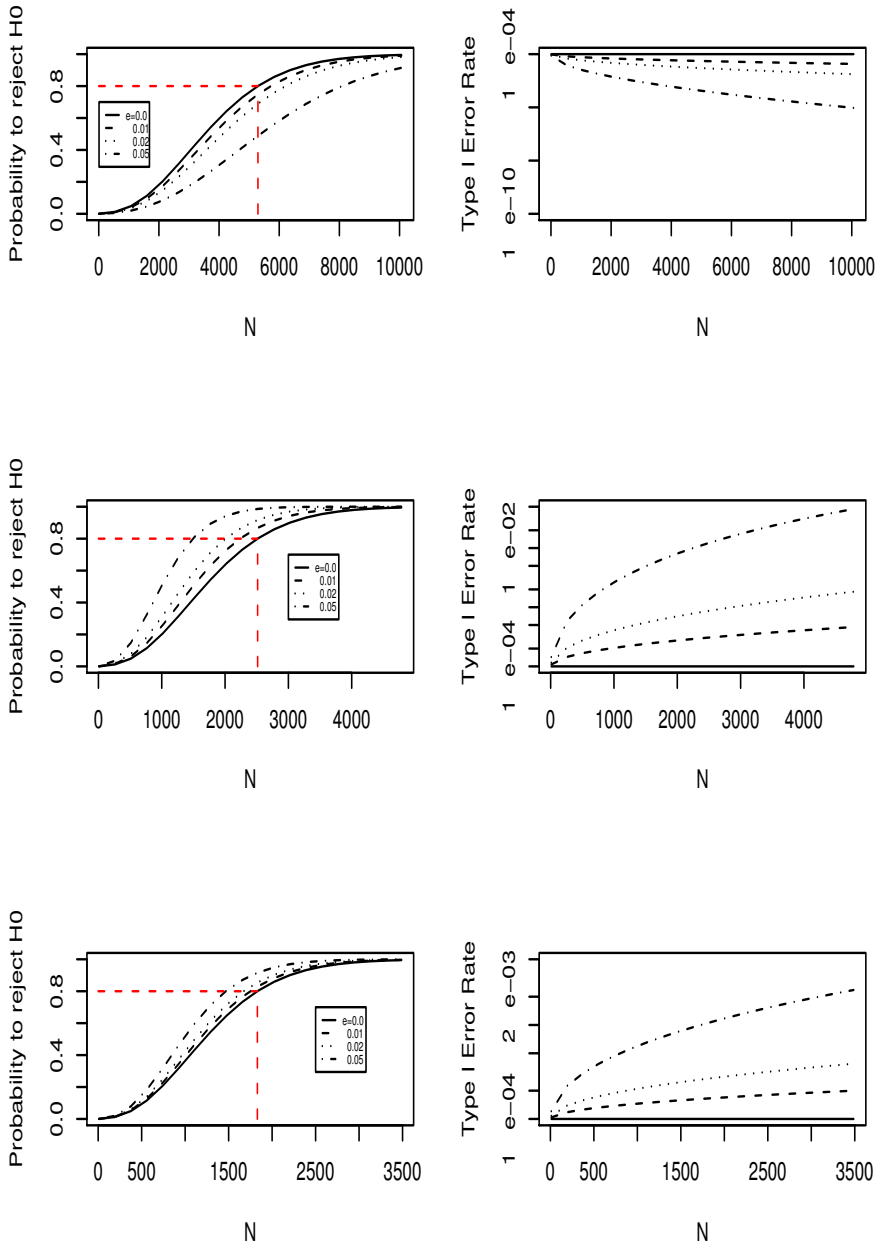

Figure 4.2: Effect of genotyping error on test for linkage in EC (top), ED (middle) and $\mathcal{I}$ (bottom) designs

in absence of genotyping errors, the excess IBD sharing is approximately related to the QTL effect γ and the optimal Haseman-Elston trait function $C(\mathbf{x}, \rho)$ through the regression (this is shown for an approximate additive model as given by Formula (4.2) in the appendix of Lebrec et al. [2006])

$$\mathbf{E}(\hat{\pi} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) \simeq \mathrm{var}_0(\hat{\pi})\gamma \; C(\mathbf{x}, \rho) \;,$$

and the effect of genotyping error is to modify this regression into

$$\mathbf{E}(\hat{\pi}^{\epsilon} - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) \simeq a(\epsilon) + b(\epsilon) \; \mathrm{var}_0(\hat{\pi})\gamma \; C(\mathbf{x}, \rho) \;. \tag{4.6}$$

For simple cases, e.g. a single equi-frequent allele marker, explicit formulae can be derived for a and b; in general though, those functions will depend in a complex manner on the genotyping error mechanism but also on the markers' map and no explicit forms will be available. When multi-point marker data are used to infer IBD sharing, errors tend to propagate around markers and one can expect a more severe effect of genotyping error compared to single-point algorithms. As mentioned earlier, for small QTL effects, most of the impact on linkage in selected samples will be due to the intercept mis-specification in the linkage regression, we therefore focus on this issue.

In random samples or under the null hypothesis of no linkage, the sample mean excess IBD $\overline{\hat{\pi}^{\epsilon}} - \frac{1}{2}$ (averaged across families) provides an estimate of the intercept $a(\epsilon)$. We simulated three different marker map configurations in 10000 sib pairs without parents and quantified by how much IBD sharing was reduced on average under the *population frequency error model* and the *false homozygosity model* (error rates=0.01 and 0.05). MapH and MapL had eleven equi-frequent allele markers located 10cM apart, markers had 10 alleles in MapH and 2 alleles in MapL. MapM only had six markers 20cM apart with 5,2,5,2,2 and 5 alleles on the six markers (from left to right). The results are displayed in Figure 4.3 along with the corresponding map information content as defined in Kruglyak and Lander [1995] (wiggly curves in bottom part of each figure, scale on the right y-axis), for clarity and because results were very similar, we have omitted the curves corresponding to the *false homozygosity model* . One clear trend is that IBD is most affected by genotyping error in areas where marker information is high. Furthermore, even for small error rates, the decrease in

IBD sharing is substantial.

4.5 Genomic control for genotyping error

As we have seen in previous sections, the main effect of genotyping error is to modify the intercept in the regression used to test for linkage. In order to obtain more robust inference, it therefore seems natural to try and constrain the regression through its correct origin a. In this section, we propose a completely data-driven strategy for doing this.

At any position, the sample mean IBD sharing has variance $\mathrm{var}_0(\hat{\pi})/n$ where n is the number of sib pairs available. If we knew that the position is unlinked or if the sample of sib pairs was random then the deviation of this mean from $\frac{1}{2}$ would provide an estimate of the intercept a in the linkage regression. Unfortunately, detection of a position-specific intercept corresponding to typical error rates would require a sample size of order 10^4, a number that is almost never reached in linkage studies. In order to obtain an intercept estimate $\hat{a}$ with sufficient precision, it is therefore essential to combine information across positions. The value of IBD sharing at positions outside of the neighborhood of influencing loci (those positions are subsequently referred to as unlinked) across the genome may serve as control in the test for linkage, this concept of genomic control has been used to robustify the analysis of association studies by Devlin and Roeder [1999].

Ad-hoc method

Let's assume that the proportions of alleles shared IBD $\hat{\pi}$ is inferred at a series of approximately regular positions indexed by t across the whole genome. Let y_t be the sample mean (among families) excess IBD at position t i.e. $y_t \equiv \overline{\hat{\pi}_t^{\varepsilon}} - \frac{1}{2}$. Under the variance components model and for small QTL effect γ, equation (4.6) implies that

$$E(y_t) \simeq \begin{cases} a\ , & \text{if position } t \text{ is unlinked}\ , \\ a + \frac{b}{8}\gamma\overline{C}\ , & \text{if position } t \text{ is linked}\ . \end{cases}$$

In random samples or in any sample where $\overline{C} \simeq 0$, taking the average of y_t across positions provides and estimate of a. In selected samples, we can use a trimmed version of the mean of y, for example a 20%-trimmed mean of the $(y_t)_t$ series (i.e.

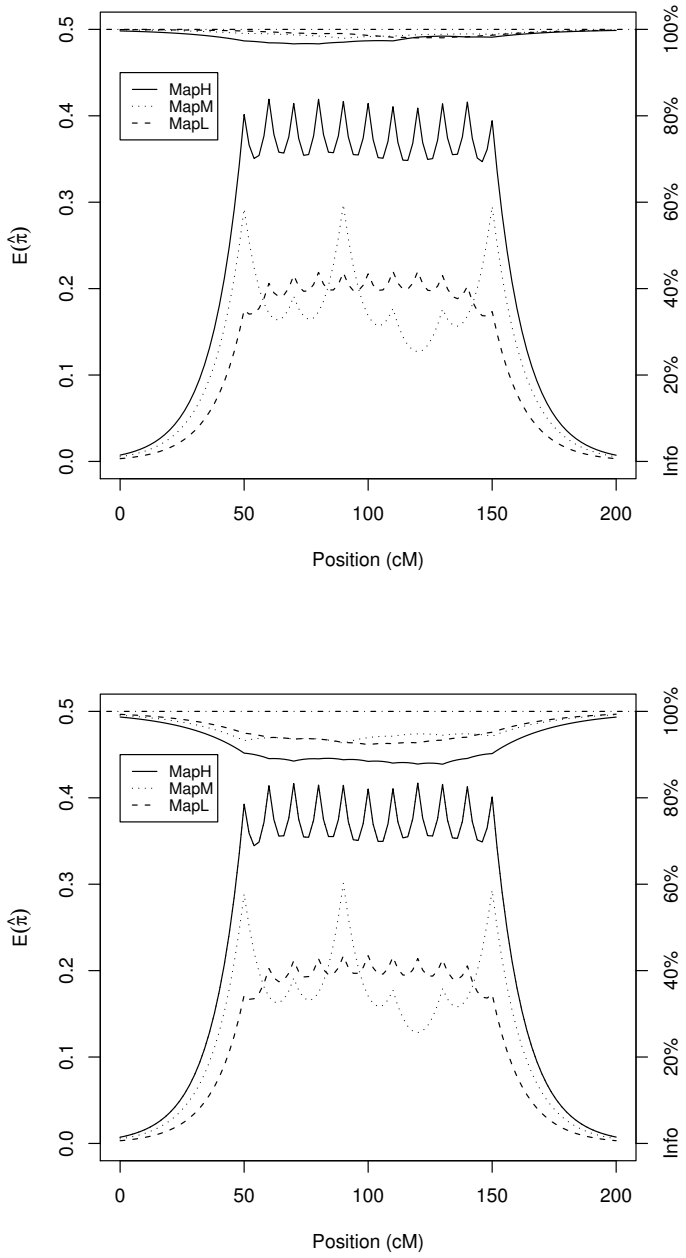

Figure 4.3: Effect of genotyping error on IBD sharing and corresponding map information content in simulated data - Error rates $\epsilon = 0.01$ (top) and $\epsilon = 0.05$ (bottom)

the mean of the y_t values after removing the 20% lowest and and 20% highest values) will provide a robust genomic estimate $\hat{a}$ of a. Because $a \leq 0$ and $\overline{C}$ is positive and negative in EC designs and ED designs respectively, $\hat{a}$ could be refined by trimming off only the 20% highest and lowest y_t values respectively before taking the mean. Of course, how much we trim is arbitrary but 20% can safely be taken as a conservative value for oligogenic traits.

An ad-hoc implementation of the concept of genomic control is then to plug in the estimate of the intercept $\hat{a}$ into the linkage regression (4.6). Since most of the bias in the inference is due to the intercept mis-specification, the precise estimate obtained by pooling across the genome will eliminate it. The implicit assumption that we make in this genomic control approach is that the regression intercept is the same at all positions.

Empirical Bayes

The method in the previous section can be formalized using an empirical Bayes inferential procedure in order to compute the posterior probability that a position is linked. Having set a minimum level of evidence for deciding whether a position is linked, the values of y_t at unlinked positions could be pooled and the estimate thus obtained plugged into the linkage regression as in the previous section. The approach is borrowed from the microarrays literature [Efron and Tibshirani, 2002] and our problem is analogous to the estimation of the proportion of true null hypotheses in false discovery rates testing rules.

We assume that the prior density f of the average excess IBD sharing $y = (y_t)_t$ is given by a mixture distribution

$$f(y) = \alpha_0 f_0(y) + (1 - \alpha_0) f_1(y) \ .$$

Here, α_0 denotes the prior probability that a position is unlinked (a conservative value would be $\alpha_0 = 1$) and $f_0(y)$ is the corresponding prior probability distribution of y, while $f_1(y)$ denotes the prior probability distribution of y at a linked position. Using Bayes' theorem, the following posterior distribution obtains

$$\mathbf{P}(\text{position } t \text{ linked} \,|\, y_t) = 1 - \frac{\alpha_0 f_0(y_t)}{f(y_t)} \ .$$

Non-parametric density estimation techniques such as kernel density estimation may be used to estimate $f(y)$ from the data without having to specify $f_1(y)$. Unless the positions where IBD is inferred are chosen far apart, the observations will not be independent but this does not invalidate the method. It suffers one inherent limitation though: the effective sample size is small in a human genome (choosing positions every 50cM produces only approximately 70 almost independent observations) and this limits our ability to estimate $f(y)$ precisely. Since $\text{var}(y_t) = (8n)^{-1}$, the prior $f_0(y)$ could be chosen as an $N(a_0, (8n)^{-1} + \tau^2)$ where a_0 would reflect our prior knowledge about the intercept a and τ^2 the associated uncertainty.

Instead of applying this empirical Bayesian framework to the average excess IBD sharing $(y_t)_t$, we can apply it directly to linkage statistics such as the QTL effect estimates $\hat{\gamma}_t = \frac{\sum_i (\pi_i^e - \frac{1}{2})\mathbf{C}_i}{\frac{1}{8}\sum_i C_i^2}$ whose expectation is calculated in the Appendix. Since $\text{var}(\hat{\gamma}_t) = (\frac{1}{8}\sum_i C_i^2)^{-1}$, priors $f_0(y)$ of the form $N(a_0, (\frac{1}{8}\sum_i C_i^2)^{-1} + \tau^2)$ are possible although asymmetric versions that favor negative values might be more appropriate. Preliminary simulations give sensible results when the true number of linked positions is not too low ($\geq 5\%$) and the study is adequately powered, however the limited number of independent dimensions in a linkage scan is a serious limitation of this approach.

Alternatives

Alternatives to this genomic-control strategy are possible and they also boil down to constraining the linkage regression through a new origin as in the ad-hoc method, the estimation procedure can be adapted to suit particular circumstances.

Firstly, in random samples, the assumption regarding exchangeability of positions might be relaxed. Indeed, the y_t's may be used as estimates of the position-specific intercepts since a study sufficiently powered to detect linkage in random samples should provide sufficient precision. It must be noted though that the advantage of using a genomic control in random samples is limited because the impact of genotyping error is small in such designs. Secondly, one could use previous lab data to estimate by how much IBD sharing deviates from its expected value, this could also been done at each position separately provided sufficient data are available. In practice, such data might not be available or they might not trustfully reflect current error mechanisms.

4.6 Discussion

Under two basic error models, we were able to predict quantitatively the consequences of genotyping error on inference in linkage analysis. In the idealized situation of complete IBD information, both error models have the same impact on linkage analysis. As we have seen, the effect is due to a decrease in IBD sharing. A contrario, an error process which would increase IBD sharing would produce opposite results. The true error processes involved in practice are complicated mixtures of the models alluded to here. In our experience however, it seems that processes which lower IBD sharing are predominant. Because genotyping error tends to decrease the estimated number of alleles shared IBD, the effect on evidence for linkage is opposite in EC (over-pessimistic) and ED (over-optimistic) designs, it can be dramatic in typical designs and paradoxically less severe for more extreme ascertainment schemes. By analogy, for a dichotomous trait, this means that the effect of genotyping error is less severe in ASP designs for rare diseases than for common diseases. Remarkably, in designs combining both ED and EC pairs like the $\mathcal{I}$ (or EDAC designs), the competing effects of genotyping error tend to cancel each other out. We have considered here only three types of basic selection schemes however the approach can straightforwardly be applied to any arbitrary selection scheme, under a variance components model, the important quantity being $\overline{C}/\overline{C^2}$.

The genomic-control strategy that we have proposed offers a robust method for carrying out linkage analysis but obviously relies on a convenient approximation of a very complex situation. It is probably reasonable to assume that genotyping of markers with a similar degree of polymorphism (number of alleles and frequencies) within the same lab is subject to the same error process. On top of the true underlying error mechanism, in a multi-point setting, not only the number of markers but also the inter-marker distances could have an impact. Ideally, markers should have similar numbers of alleles and respective frequencies and be rather evenly distributed across the genome. Based on results from simulations presented in Section 4.4, it seems appropriate to pool estimates of regression's intercept a which correspond to areas of the genome where marker information is roughly the same. The advent of SNP chip therefore makes us confident of the applicability of our method, indeed this

new technology for linkage data holds the promise of providing marker maps with less variable information content than in classical microsatellites maps [Evans and Cardon, 2004; Schaid et al., 2004].

Elston et al. [2005] have recently pointed out that the implicit assumption made in ASP designs, that randomly sampled sib pairs share half of their alleles IBD, might not hold in practice and have argued for including discordant pairs in such studies. The approach presented here offers an alternative solution to this issue. Finally we note that, although we have only considered designs involving sib pairs, the approach naturally extends to other types of relative pairs.

Acknowledgements

We are grateful to Dr. Bas Heijmans from the section Molecular Epidemiology, Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center for discussions on genotyping error mechanisms.

4.7 Appendix

Effect of genotyping error on linkage

We show how regression (4.3) is modified in presence of genotyping error. We concentrate on the case where IBD information is complete.

By definition $\mathbf{E}(\pi^\epsilon - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = \frac{1}{2}\ \mathbf{P}(\pi^\epsilon = \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) + \mathbf{P}(\pi^\epsilon = 1 \,|\, \mathbf{x}, \gamma, \epsilon) - \frac{1}{2}$. We can then condition on the true IBD status π and use approximation (4.2) in order to evaluate the probabilities involved in the previous expression: $\mathbf{P}(\pi^\epsilon \,|\, \mathbf{x}, \gamma, \epsilon) = \sum_\pi \mathbf{P}(\pi^\epsilon \,|\, \pi)\ \mathbf{P}(\pi \,|\, \mathbf{x}, \gamma)\ \mathbf{P}(\pi^\epsilon \,|\, \pi)$. In the present case of complete information, this yields

$$\mathbf{E}(\pi^\epsilon - \frac{1}{2} \,|\, \mathbf{x}, \gamma, \epsilon) = -\frac{\epsilon}{4} + (1 - \frac{\epsilon}{2})\ \frac{\gamma}{8}\ C(\mathbf{x}, \rho)\ . \tag{4.7}$$

Probability to reject H_0

We derive an approximate formula for the probability of rejecting the null hypothesis of no linkage if we ignore genotyping error.

As we have seen earlier, testing for linkage boils down to regression (4.3). Let's denote by $\hat{\gamma}$, the estimate of the slope in the regression through the origin of a sam-

ple $\left(\pi_i - \frac{1}{2}\right)_{i=1,\ldots,n}$ on the corresponding $C_i = (C(x_{i1}, x_{i2}, \rho))_{i=1,\ldots,n}$ and by $\hat{\gamma}^\epsilon$, the estimate of the slope in the same regression but where the response is replaced by $\left(\pi_i^\epsilon - \frac{1}{2}\right)_{i=1,\ldots,n}$.

$$\hat{\gamma} = \frac{\sum_i (\pi_i - \frac{1}{2})\ C_i}{\frac{1}{8}\sum_i C_i^2} \quad \text{and} \quad \mathbf{E}(\hat{\gamma}\,|\,\mathbf{x}, \gamma) \simeq \gamma$$

i.e. $\hat{\gamma}$ is an approximately unbiased estimate of γ. However it appears that $\hat{\gamma}^\epsilon = \frac{\sum_i (\hat{\pi}_i^\epsilon - \frac{1}{2})\ C_i}{\frac{1}{8}\sum_i C_i^2}$ is biased since

$$\begin{aligned} \mathbf{E}(\hat{\gamma}^\epsilon\,|\,\mathbf{x}, \gamma, \epsilon) = & \quad \frac{\sum_i \mathbf{E}(\pi_i^\epsilon - \frac{1}{2}\,|\,\mathbf{x}, \gamma)\ C_i}{\frac{1}{8}\sum_i C_i^2} \\ \simeq & \quad (1 - \frac{\epsilon}{2})\ \gamma\ - \frac{\epsilon}{4}\ \frac{\overline{C}}{\overline{C^2}}\ . \end{aligned} \tag{4.8}$$

The bias in $\hat{\gamma}^\epsilon$ depends on two factors: the genotyping error rate ϵ and the selection procedure of sib pairs (which determines $\overline{C} = \frac{1}{n}\sum_i C_i$ and $\overline{C^2} = \frac{1}{n}\sum_i C_i^2$). Whatever the ascertainment scheme used (in particular in random samples), the estimate of γ is systematically biased downwards by a factor $1 - \frac{\epsilon}{2}$; then, depending on the sign and value of $\overline{C}/\overline{C^2}$, $\hat{\gamma}^\epsilon$ can be further decreased or increased. For complex traits and thus small QTL effects γ, the intercept mis-specification will have a greater impact than the bias in the slope. The test for linkage is based on the standardized slope estimate $\frac{\hat{\gamma}^\epsilon}{\sqrt{\mathrm{var}_0(\hat{\gamma}^\epsilon)}} = \frac{\hat{\gamma}^\epsilon}{\sqrt{\mathrm{var}_0(\pi^\epsilon)\overline{C^2}}}$, since $\mathrm{var}_0(\pi) = \frac{1}{8}$ is practically unchanged by genotyping error ($\mathrm{var}_0(\pi^\epsilon) = \frac{1}{8} - \frac{\epsilon^2}{16}$), the probability of rejecting the null hypothesis is given by

$$\Phi\left(\Phi^{-1}(\alpha) + (1 - \frac{\epsilon}{2})\gamma\mathcal{I}^{1/2} - \ 8\ \frac{\epsilon}{4}\frac{\overline{C}}{\overline{C^2}}\mathcal{I}^{1/2}\right)\ , \tag{4.9}$$

where $\mathcal{I} = \mathrm{var}_0(\hat{\gamma})^{-1} = \frac{n}{8}\ \overline{C^2}$ is the sample's Fisher's information for the linkage parameter γ, α is the nominal type I error rate for the linkage test with a true quantitative trait locus effect γ and Φ is the cumulative density function of the standard normal distribution. A first order Taylor approximation of (4.9) yields Formula (4.5).