Williams, J.T., Van Eerdewegh, P., Almasy, L. and Blangero, J. (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. *Am. J. Hum. Genet.* **65**: 1134–1147.

Zonderman, A.B. (1986) Twins, families, and the psychology of individual differences: the legacy of Steven G. Vandenberg. *Behav. Genet.* **26**: 11–24.

13 Factors affecting type-I error and power of linkage analysis

Manuel A.R. Ferreira

As with many statistical tests, the performance of the linkage methods described in previous chapters is influenced by a plethora of experimental factors. Therefore, in order to appropriately interpret linkage results, either significant or not, it is important to identify factors that may have introduced any biases in the analysis and, if possible, adjust for it accordingly.

This chapter describes common factors which are known to influence the power and type-I error rate of linkage analysis. The statistical concepts of power and type-I error were introduced in Chapter 5, and so here we shall only focus on how these relate to the specific case of linkage tests. In linkage analysis, we may be interested in determining the required sample size to detect a QTL with a given power or, alternatively, we may be interested in determining the power to detect a QTL with a given sample size. However, the statistical parameter power itself (i.e., the probability of rejecting the null hypothesis when the alternative hypothesis is true; defined as $1 - \beta$, where β is the type-II error) is not a convenient quantity, since it is not linearly related to sample size. A more appropriate measure of the power of linkage analysis is the noncentrality parameter (NCP, λ). The test statistic of many linkage methods has a central χ^2 distribution when the null hypothesis is true (with mean = degrees of freedom (df) of the test) and a noncentral χ^2 distribution when the alternative hypothesis is true (Kendall and Stuart 1979, Sham and Purcell 2001). The mean of the noncentral χ^2 distribution is given by df + NCP of the test (Figure 13.1). Thus, the NCP can be thought of as the displacement between the expected distribution of the test statistic under the null (no linkage)

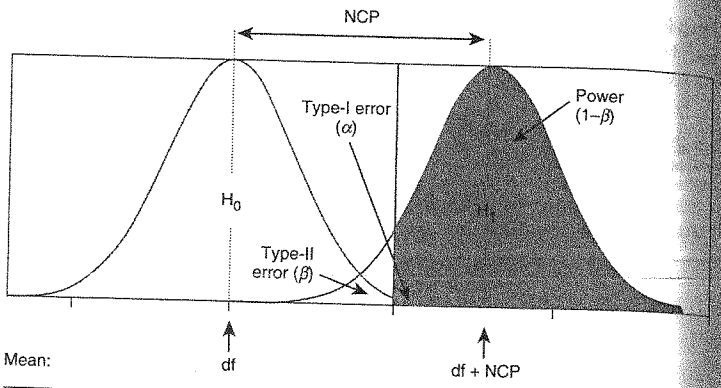

Figure 13.1 Graphical representation of the noncentrality parameter (NCP) of a test. H_0 represents the distribution of the test statistic when the null hypothesis is true (e.g., no linkage between marker and trait). H_1 represents the distribution of the same test statistic when the alternative hypothesis is true (e.g., linkage between marker and trait). The vertical solid line defines the critical threshold for significance; this determines the type-I error (α), the type-II error (β), as well as the power ($1 - \beta$) of the test. Type-I error is the probability of rejecting the null hypothesis when the null hypothesis is true; corresponds to the filled area of the H_0 curve. Power is the probability of rejecting the null hypothesis when the alternative hypothesis is true; corresponds to the filled area of the H_1 curve. df: degrees of freedom of the test. The shape of the two distributions was drawn for illustration purposes only.

and the alternative hypothesis (linkage). The greater the NCP, the smaller the overlap between the null and the alternative hypothesis and, therefore, the greater the power of the test.

A number of factors influence the magnitude of the NCP (hence, the power) and/or the type-I error of a linkage test statistic. These include, but are not restricted to, selective sampling, sample size, deviations in trait normality, outliers, pedigree errors, genotyping errors, marker informativeness, choice of marker density and genetic map. Specifically, this chapter focuses on practical approaches to detect and minimize the impact of these factors.

13.1 Selective Sampling

Description

As noted by Fisher (1934), 'the interpretation of a body of data requires a knowledge of how it was obtained [...]. Nevertheless, in human genetics especially, statistical methods are sometimes put forward, and their respective claims advocated with entire disregard of the conditions of ascertainment'.

In the context of linkage analysis, random sampling refers to a study design in which all families in the population have equal probability of being tested. In contrast, selective sampling refers to a design that collects data on a restricted group of families from the general population. Since most linkage studies ascertain families with two or more siblings, strictly speaking, they should be by default considered to use selective sampling. However, more commonly, a linkage study is considered to use a selective sampling design if the ascertainment of families imposed additional selection criteria based upon the phenotypes of the sibs.

Common selective sampling designs used in linkage analysis include the following: (i) affected proband design, in which a family is selected on the basis of having at least one offspring exceeding a high phenotypic threshold (z_H); (ii) affected sib-pair (ASP) design, in which at least two offspring exceed z_H; (iii) concordant sib-pair (CSP) design, in which a family contains at least two offspring with phenotypes both above z_H or both below a low phenotypic threshold (z_L); (iv) discordant sib-pair (DSP) design, families with at least one offspring exceeding z_H and another below z_L; and (v) extreme discordant and concordant sib-pair (EDAC) design, families with either DSP or CSP. Note that the thresholds z_H and z_L can represent either an observed quantitative trait or a latent continuous variable underlying an observed discrete trait. If appropriately analyzed, study designs based on selective sampling tend to provide higher power to detect linkage than random samples. However, they may result in biased results (described below) if inappropriate analyzed.

Impact on linkage analysis

It is well known that sib pairs are not all equally informative for linkage. For example, sib pairs with IBD = 1 are not informative at all, whereas sib triplets in which one of the possible sib pairs is IBD = 2 while the other two pairs are IBD = 0 are the most informative (Dolan *et al.*, 1999). This forms the theoretical basis of the selective sampling designs described above: by selecting families with siblings that are either highly similar or highly dissimilar for our trait of interest, we hope to enrich our sample with the most informative sib pairs, that is those that are either highly similar (IBD = 2) or dissimilar (IBD = 0) for a QTL that regulates the trait (Cardon and Fulker, 1994; Carey and Williamson, 1991; Eaves and Meyer 1994; Gu *et al.*, 1996; Risch and Zhang, 1995). In this way, the choice of the thresholds z_H and z_L determines the likelihood of ascertaining an informative sib pair and, thus, it influences the power of linkage analysis. With some notable exceptions, the more

extreme the selection is, the greater the improvement in power it achieves (Allison *et al.*, 1998).

Therefore, if appropriately analyzed, selected samples can in most cases improve the power to detect linkage when compared to randomly ascertained samples. Appropriate methods which are robust and powerful when analyzing such selected samples include allele-sharing statistics (Kong and Cox, 1997; Risch and Zhang, 1995, 1996; Whittemore and Halpern, 1994), the classic Haseman and Elston (1972); HE method, the Sham and Purcell (2001) HE method (HE-COM), the Forrest and Feingold (2000) composite statistic, MERLIN-regress (HE-R; Sham *et al.*, 2002) and reverse variance components analysis (VC-R; Sham *et al.*, 2000b).

In contrast, some linkage methods which are robust and particularly powerful when applied to random samples do not perform so well in the presence of selective sampling. The maximum-likelihood variance components statistics described in Chapters 10 and 12, henceforth referred to simply as VC, are one such group of methods. The main reason for this is that VC methods estimate individual pedigree likelihoods under the assumption that the entire population was sampled. However, as some families from the population are excluded by selective sampling, so the likelihood of observing the remaining families increases. If this effect is ignored or not accounted for in VC, selective sampling results in biased estimation of pedigree likelihoods which leads to an inflation of the type-I error rate (Sham *et al.*, 2000b).

Detection

For all experimental designs other than random sampling, the question of detecting whether selective sampling took place or not is rarely an issue: the designs were chosen in the first place because they were selective. However, in this case, it may still be important to estimate the proportion of the population that was actually ascertained if VC methods are to be used for linkage analysis.

To estimate the proportion of the population that was ascertained it is necessary to know in detail the conditions of ascertainment used. For simple selection strategies (e.g., single probands or sib pairs), the procedure described by Neale et al. (1994) and implemented in Mx (page 148 of the Mx manual, (Neale et al. 2002)) can be used.

A more practical approach to assess how severe selection was is to compare sample parameters such as the mean, variance and correlation with the respective parameters estimated from a random sample of the population. The larger the deviation between sample and population parameter estimates, the larger the degree of ascertainment.

Correction

If selective sampling took place, traditional VC methods can still be used for analysis provided that an appropriate correction is implemented to maintain the type-I error rate at nominal levels. Below, we consider four alternative methods to correct for selective sampling in VC linkage analysis.

Following de Andrade and Amos (2000), the first two methods are henceforth referred to as 'point-probability' and 'cumulative-probability' corrections. Both methods calculate a corrected pedigree likelihood by dividing the likelihood of observing a pedigree given the estimated IBD status, $L(x \mid \hat{\pi})$ (provided by most linkage computer programs), by a specific amount which reflects the ascertainment scheme used. Note that either correction should always be applied when analyzing a selected sample, whether the trait being analyzed has a normal distribution or not (Sham *et al.*, 2000b).

The point-probability correction was proposed by Hopper and Mathews (1982) and Sham *et al.* (2000b). The idea is that we can correct the likelihood of a pedigree i given the observed IBD status by additionally considering the probability of observing that particular family i in the selected sample (and not in the population). Therefore, in this case, the quantity used to reflect the ascertainment scheme is the point-probability of observing pedigree i in the selected sample, that is

$$\ln L_i = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln V_P - \frac{1}{2}\frac{(\mathbf{y}_i - \mu)}{V_P} \qquad (13.1)$$

where $\mathbf{y}_i$ is a vector of observed values for the ith pedigree, μ is the sample mean and V_P is the total variance of the trait. Thus, the point-probability correction does not require any knowledge of the ascertainment conditions used but, rather, it uses the overall mean and total variance of the selected sample to calculate the probability of observing the trait values of a given pedigree i. This correction needs to be calculated for each pedigree individually.

A similar but alternative method is the cumulative-probability correction, the classical conditioning method adapted from segregation analysis (see Sham, 1998, page 34). In this case, the term used to correct the likelihood of pedigree i is not the probability of observing the vector of phenotypes in the selected sample but instead the probability that such vector of phenotypes was originally ascertained. As mentioned above, this depends on the

type of ascertainment strategy used. For example, for an affected sib-pair design, in which a family with two offspring is selected if both sibs exceed a given threshold t, this likelihood is obtained by integrating the probability density function from the thresholds t_1 and t_2 to positive infinity. For other ascertainment schemes, the interval for integration will vary accordingly. Since the selection criterion is the same for all families, this correction need only be calculated once. Formally, the correction term is calculated as

$$LL_i = \ln\left[1 - \int_{-\infty}^{t}\int_{-\infty}^{t}\varphi(x_1, x_2), dx_1 dx_2\right] \qquad (13.2)$$

where the limits t of the integrals are chosen so that $\int_{-\infty}^{t}\int_{-\infty}^{t}\varphi(x_1, x_2), dx_1 dx_2$ represents the probability that a family is not ascertained. Irrespective of the correction term used, the overall corrected likelihood for a sample of n pedigrees is obtained by

$$LL_i = \sum_{n}(LL_{1i} - LL_{2i})$$

where LL_{1i} is the log-likelihood of pedigree i conditional on the IBD status and LL_{2i} is the correction term as specified by Equations 13.1 or 13.2.

The point-probability and cumulative-probability methods were compared by de Andrade and Amos (2000) and by Sham *et al.* (2000b) with simulated datasets of selected samples. de Andrade and Amos demonstrated that for a trait not departing excessively from normality, both methods gave comparable but improved results when tested against uncorrected VC. The authors noted that the cumulative-probability correction is the most efficient approach but it may not be feasible when the ascertainment process is very complex or unknown. In contrast, the point-probability correction is always possible but may not be fully efficient. Similarly, Sham *et al.* (2000b) demonstrated that uncorrected VC is not adequate for the analysis of selected sib-pair samples, whether the trait follows a normal distribution or not. Additionally, Sham *et al.* showed that when analyzing a normal trait in a selected sample, both ascertainment corrections result in equivalent type-I error rate and power. However, with non-normal trait data, the point-probability approach maintains the type-I error rate at nominal levels (albeit providing very low power to detect linkage), whereas the cumulative probability does not. Thus, the point-probability approach is likely to be more robust but possibly not as powerful (Feingold, 2001). Finally, both methods are easily implemented in the statistical package Mx (Neale *et al.*, 1994, 2002).

Two other approaches may be used when analyzing data from selected samples with VC. First, if the selected sample that was genotyped is nested within a larger random sample of the population with available phenotypes, then VC analysis of the entire dataset (i.e., genotypes for the selected sample plus phenotypes for the entire sample) should result in a robust and powerful test for linkage, as long as the algorithm used for IBD estimation appropriately accounts for missing data, which is the case for most commonly used linkage packages (Abecasis et al. 2004a). Second, since the main concern of VC analysis of selected samples is the potential increase of type-I error, we can simulate genotypic data for the selected sample under the null hypothesis of no linkage and obtain an empirical estimate of the type-I error rate of our analysis (for details and examples see Abecasis *et al.*, 2004b; Ferreira *et al.*, 2005; Kruglyak and Daly, 1998; Lander and Kruglyak, 1995). In some situations, this is an efficient alternative to estimate linkage significance while accounting for the potential bias introduced by selective sampling.

13.2 Sample Size

Description

Irrespectively of the study design used, all linkage studies are based on the analysis of data collected from families. Indeed, linkage analysis combines the individual evidence for linkage provided by each family to construct an overall test for linkage between a marker locus and a trait across the entire sample. Generally speaking, and assuming that there is a true trait locus, the larger the number of families tested, the larger the overall evidence for linkage. However, this relationship between sample size and power is influenced by a number of additional factors. For convenience, we discuss here five factors that interact closely with sample size to determine the power of linkage tests: disease prevalence, QTL heritability, residual shared variance, incomplete linkage and pedigree size. Other factors, such as pedigree errors, genotyping errors and marker informativeness, are individually discussed in subsequent sections.

Impact on linkage analysis

The first factor to be considered here is disease prevalence, which is particularly relevant when analysing discrete traits, such as disease status. For a study to have enough power to detect linkage between a marker and a disease, the sample being analysed must include an appropriate number of affected individuals. This problem is similar to the effect on power of the phenotypic ascertainment threshold for a continuous trait (Section 13.1 above). For rare

diseases and random ascertainment, a large proportion of the population must be sampled to include a significant number of cases and, thus, to provide a powerful test of linkage (Williams and Blangero 2004). Often, however, this strategy is unfeasible. A more efficient alternative for gene mapping of rare diseases is to ascertain families with multiple affected relatives, for example affected sib-pairs (ASPs). Blangero et al. (2001) showed that for rare diseases and for a range of QTL heritabilities, the power obtained by genotyping a sample of ASPs can be similar to the power obtained by genotyping a 10-100 larger random sample of sib-pairs. The authors showed that relatively small to medium-sized (<1000 sib-pairs) selected samples may provide powerful linkage tests for rare diseases under a wide range of QTL heritabilities (>15%). By contrast, for common diseases, a randomly selected sample is likely to have sufficient affected individuals to provide a powerful test for linkage, and so selective sampling may not be justified. In this case, the sample size required to detect a QTL will depend to a great extent on the QTL heritability (Williams and Blangero 2004).

A second factor that influences the power of linkage analysis is the proportion of phenotypic variance that is attributable to the QTL. Sham et al. (2000a) and Sham and Purcell (2001) showed that the noncentrality parameter (NCP, see Introduction above) of two HE methods and VC methods for random samples of sib-pairs is proportional to the square of both the additive and dominance QTL variance components. Thus, the power of linkage will tend to be small for a QTL with low heritability and, hence, a larger sample size will be required to localize it. The impact of the QTL heritability on the power of variance components linkage analysis is shown in Figure 13.2 for different study design scenarios. Similar results apply for regression-based methods. For allele-sharing methods, Risch and Zhang (1996) concluded that the smallest genetic effect detectable with a realistic sample of extreme discordant sib-pairs was approximately 10%.

A third factor that influences the power of linkage is the proportion of variance due to shared residual factors, such as polygenic effects or common environment. Sham et al. (2000a) demonstrated both theoretically and empirically that the power of VC increases with increasing residual shared variance. You can visualize this effect by comparing panels A and B against C and D of Figure 13.2, respectively. The reason for this seems to be the following. The crucial parameter in QTL linkage analysis is the proportion of the within-family variance that is explained by the QTL. Following VC notation, there are three components that determine the proportion of within-family variance for a given trait: the variance due to the

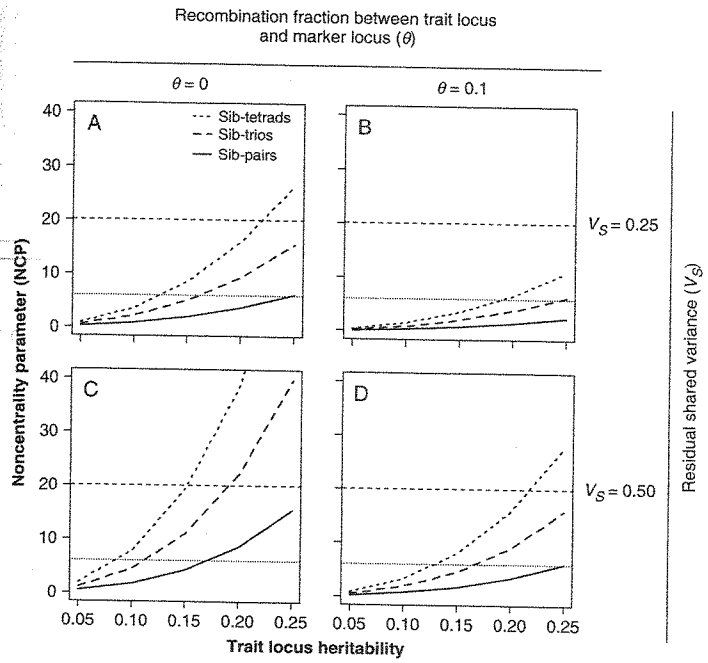

Figure 13.2 Noncentrality parameter (NCP) for the univariate variance components linkage test. The NCP has a direct (but not linear) relation to the power of the test: the larger the NCP, the greater the power. The NCP was estimated for a fixed sample size of 2000 genotyped individuals (i.e., 500 families with a sib-pair, 400 families with a sib-trio and 333 families with a sib-tetrad; both parents were also assumed to be genotyped). The NCP threshold for a study to have 80% power to detect linkage is for example 6.17 (for a type-I error $\alpha = 0.05$; horizontal dotted line) or 20.8 ($\alpha = 0.0001$; horizontal dashed line). A recombination fraction of 0.1 corresponds to ~11cM genetic distance, using Haldane's mapping function.

QTL (V_Q), the variance due to shared factors (V_S) and the variance due to non-shared factors (V_N). Shared factors always decrease the proportion of within-family variance (i.e., they increase the similarity between sibs), whereas non-shared factors always increase it (i.e. they decrease the similarity between sibs). Thus, the presence of shared factors other than the QTL will tend to increase the proportion of the within-family variance that is explained by the QTL and, consequently, it will improve the power to detect the QTL itself.

In this way, increased residual shared variance reduces the sample size required to detect a QTL. Similar results were reported by Dolan et al. (1999) and by Visscher and Hopper (2001) for the classic HE method and for their own HE extension. Together, results from these studies indicate that a design that decreases the residual non-shared variance and increases the residual shared variance, for example by ascertaining DZ twin pairs instead of ordinary sib-pairs, will potentially improve the power of the test.

Fourthly, incomplete linkage between a marker locus and a trait locus has been shown to attenuate the power of linkage (Sham et al., 2000a). This situation occurs when the analysis is performed at a marker locus that is not the trait locus itself but it is linked to it. The more distant the marker locus is from the trait locus, the weaker the evidence for linkage (compare panels A and C against B and D of Figure 13.2, respectively). Indeed, for an additive QTL, the attenuation of the NCP of the VC linkage test with increasing incomplete linkage is by a factor of $(1 - 2\theta)^4$ (Sham et al. 2000a), where θ is the recombination fraction between the marker locus and the trait locus. Thus, theoretically, for the same model parameters and the same power, one requires increasingly larger sample sizes as the density of markers used for linkage analysis decreases. The problem of marker density will be discussed more thoroughly in the last section of this chapter. Nonetheless, we can already appreciate that if a trait locus lies midway between two markers of a sparse map, and if the sample size and the QTL heritability are relatively small, then linkage analysis based on singlepoint IBD estimation may very well fail to detect such QTL. As we shall see (Section 13.7), one possible solution to this limitation is multipoint IBD estimation.

Finally, together with the number of families used, the pedigree size plays an important role in determining the power to detect linkage. Larger pedigrees – that is, families with many relatives per generation (typified by the sibship size) and/or many generations represented – will nearly always provide greater power than smaller pedigrees. Several studies have shown analytically that the power of variance components linkage analysis is a function of the sibship size (Dolan et al. 1999, Williams and Blangero 1999, Sham et al. 2000a). For the same total number of genotyped individuals, larger sibships provide greater power than smaller sibships (Figure 13.2). In addition, for a given sibship size, some sib-pairs will be more informative for linkage than others, based on their IBD configurations and, therefore, on their trait scores. Following from this, Sham and Purcell (2001) suggested a simple method to rank sib-pairs according to their phenotypic values in terms of their potential informativeness for regression-based linkage analysis. Purcell et al. (2001) derived an identical strategy for selective genotyping under a VC-based analysis, extending it to sibships of any size. Their method allows the estimation of the sibship NCP according to sibs' trait values, under a given set of QTL heritabilities and trait correlation. In this way, by summing individual NCPs one can effectively assess if more sibships should be genotyped to achieve a given power level to detect linkage (noting that an overall NCP of 20.8 is indicative of 80% power for $\alpha = 0.0001$, see next section). A module for the calculation of this index of sibship informativeness is available in the GPC (Purcell et al., 2003). Lastly, it is important to highlight that power can be boosted not only by ascertaining large, informative sibships, but also by ascertaining pedigrees with multiple generations represented. See Williams and Blangero (1999) for a good discussion on this topic.

Detection

In linkage analysis we will often be required to determine the power to detect a QTL for a given study design or, alternatively, to determine the optimal study design to detect a QTL with a given power. There are two general approaches that can be used to estimate the power of linkage analysis: analytical and empirical approaches. Analytical approaches provide fast approximations to the power of a test. Most often, however, these approximations have only been derived for a few linkage methods and for "typical" study designs, i.e., designs that involve random ascertainment, complete parental genotypes, normally distributed traits, etc. On the other hand, the empirical approach is in principle applicable to any study design and method of analysis, as long as data can be simulated that closely reflects the design used. However, this greater flexibility comes at a cost: computational time. Both approaches are summarised below.

Sham and Purcell (2001) derived analytical formulas for the calculation of the NCP for five HE regression methods and for variance components, VC. They demonstrated that the NCP (and, hence the power) is equivalent between two modified versions of HE (HE-W and HE-COM) and VC methods for random samples of sib pairs, and Sham *et al.* (2000a) showed that it can be approximated by

$$NCP \approx \frac{s(s-1)}{2}\frac{(1+r^2)}{(1-r^2)^2}\left[V_A^2 Var(\hat{\pi}) + V_D^2 Var(z) + V_A V_D Cov(\hat{\pi}, z)\right] \qquad (13.3)$$

Thus, the NCP is proportional to the square of the number of pairs in the sibship (s), to the sib correlation (r), to the squared variance

due to the additive QTL component (V_A), to the marker informativeness (as reflected in the variance of $\hat{\pi}$ and z; see Section 13.8 below), and to the squared variance due to the dominance QTL component (V_D). Equation 13.3 calculates the NCP per sibship for a specific set of model parameters, and it is implemented in the GPC (Purcell *et al.*, 2003; Sham *et al.*, 2000a). Once the theoretical NCP for a randomly ascertained sibship has been calculated, there are two approaches that can be used to calculate the power of an experimental design. The first applies when dealing with sibships of constant size, in which case the theoretical NCP is the same for all sibships. It is a useful approach when we wish to investigate the power of different study designs (e.g., different sibship size, QTL heritability, marker informativeness) to detect linkage. For linkage, the level of significance required is traditionally set at a LOD score of 3. This is the logarithm of the likelihood ratio (1000) that is necessary to convert the odds in favor of linkage from 1:50 (prior probability) to 20:1, the latter corresponding to the conventional 0.05 threshold for statistical significance (Ott, 1991). A LOD score of 3 is equivalent to a central χ^2 statistic of 13.8 ($\chi^2 = 2\ln(10) * \text{LOD}$), which, for a 1 df test, ordinarily corresponds to $\alpha = 0.0002$. However, since under the null hypothesis of no linkage the LRT is asymptotically distributed not as a simple χ^2_1, but as the mixture $\frac{1}{2}\chi^2_1 : \frac{1}{2}\chi^2_0$ (Hopper and Mathews, 1982), the correct type-I error associated with a χ^2 statistic of 13.8 is 0.0001 (Williams and Blangero, 1999). Note that under the alternative hypothesis of linkage, the univariate LRT still has a 1 df χ^2 distribution. If 13.8 is adopted as the critical central χ^2 statistic, the corresponding noncentrality parameter for 80% power and 1 df is 20.8 (easily calculated with the GPC). In other words, a linkage study design will have 80% power to detect a QTL under any range of conditions (heritability, recombination fraction, etc.) if it results in an overall theoretical NCP of 20.8. Thus, the required number of sibships to achieve 80% power can be obtained by dividing 20.8 by the theoretical noncentrality parameter per sibship. For example, for a QTL heritability of 0.15, no QTL dominance, residual shared variance of 0.25 and assumed recombination fraction between the marker locus and the trait locus of 0.05, one sib pair has the theoretical NCP of 0.002559. Dividing 20.8 by this value results in a required sample size for 80% power of 8128 sib pairs.

The analytical approach can also be used to calculate the power of an experimental design that includes a mixture of sibships of different sizes. In this case, a different NCP has to be calculated for each class of sibship. The overall NCP of the linkage test is simply

obtained by the weighted sum of the different NCPs for all sibships, where the weights are the number of sibships collected for the respective sibship class. If this overall NCP is greater than 20.8, the study design used has greater than 80% power to detect the QTL given the appropriate range of conditions. Conversely, one can calculate the power associated with such NCP by referring to the appropriate noncentral χ^2 distribution function. Again, the "Probability Function Calculator" module of the GPC can be used for this. For example, if the overall NCP is found to be 18.2, the proportion of this noncentral χ^2 distribution with 1 df that lies after the central χ^2 critical value of 13.8 is 71%. This would be the approximate power of our test. The NCP formula implemented in the GPC is only applicable to one-generation families and for continuous traits. Chen and Abecasis (2006) derived an approximation for the VC NCP for continuous traits that accommodates multigenerational pedigrees. This analytical approach is implemented in their software POLY. Williams and Blangero (2004) derived an analytical formula for the power of VC for discrete traits which can also be applied to arbitrary pedigrees.

It is important to note that the parameters included in NCP approximations concern the true values and not the estimated values, which may be biased when an incorrect model is assumed. Thus, when using maximum likelihood estimators of such parameters, the power calculations should be interpreted with care.

When analysing a trait measured in a randomly selected sample with VC or HE methods, the analytical methods described above provide a good approximation of the power of a study to detect a QTL. However, when analysing a discrete trait with allele sharing methods, or more complex study designs (e.g., ascertained sample, non-normal data, etc), the previous methods are of limited application. In this case, one may need to estimate power empirically. Briefly, this approach consists of (i) simulating trait data for pedigrees according to a specific genetic model (e.g., QTL heritability, residual shared variance, etc) and study design; (ii) analysing this simulated dataset with your linkage test statistic of choice; (iii) recording the observed test statistic (e.g., as a p-value, LOD score or χ^2) for the simulated dataset; (iv) repeating (i) to (iii) a large number of times; (iv) estimate the power to detect a QTL under the simulated conditions as the proportion of datasets for which the recorded test statistic exceeded a given critical value (e.g., $p = 0.0001$, or LOD = 3, for $\chi^2 = 13.8$). With simple shell scripts, simulated datasets can be generated using software such as R and then analysed with any desired linkage software (e.g., MERLIN). Note that when using this empirical approach, it is important to

also simulate datasets where the QTL heritability is set to zero; this checks whether the type-I error of the analysis is or not close to nominal values. In this case, the proportion of datasets that exceed a given critical value (e.g., $p = 0.05$) will provide an empirical estimate of the type-I error of the analysis.

Correction

Linkage studies should be planned to collect the appropriate sample required to detect the predicted QTL with a power >80%. This means choosing the appropriate ascertainment strategy, the number of families to be ascertained, marker density and anticipating which methods should be used to extract the most linkage information from the dataset. This necessarily depends on the study design used. If a sample has already been collected and a quality control assessment indicates that linkage analysis may be underpowered to detect a QTL of interest, two steps may be necessary. First, identify the factors which explain the low analytical power, and second, attempt to correct them. From what was discussed in this section, it is very likely that this may involve data collection from an additional number of families. If that is the case, these families should be judiciously chosen based on their potential linkage informativeness under the study design used.

13.3. Deviations in Trait Distribution

Description

A standard normal distribution has mean = 0, skewness = 0, kurtosis = 3 and variance = 1. Respectively, these parameters measure the position, symmetry, tail heaviness and width of a distribution. Continuous phenotypic traits collected as part of linkage studies may show considerable deviations in skewness and kurtosis from the expected normal values. If that is this case, they are said to be non-normally distributed, which violates an assumption underlying the calculation of many linkage test statistics. Additionally, the mean and variance of a given sample may be different from the observed values in the general population. Such distribution deviations can also introduce biases in linkage tests.

Trait non-normality may arise for several reasons. First, the trait may be a discrete variable, either binary or polychotomous; in either case, the trait is clearly non-normally distributed. Secondly, if the trait of interest is determined by the action of a major QTL, the trait will follow a mixed non-normal distribution (Schork *et al.*, 1996). A third factor that may induce non-normality is selective sampling: by ascertaining families with probands which are above or below a given threshold, the selected trait sample is unlikely to

be normal. Fourthly, some traits are not normally distributed in a random sample of the population. For example, total serum immunoglobulin E is known to have a characteristic logarithmic distribution. Finally, normality violations may be introduced by trait censoring (e.g., assay sensitivity, selective sampling): trait values less than some threshold t are observed to be at t or they are not observed at all. As a consequence, the distribution will show significant positive kurtosis (kurtosis > 3, that is, greater than expected for a normal distribution) and skewness.

Impact on linkage analysis

Regression-based methods, such as the classic Haseman and Elston (1972) approach (HE), are robust to violations in trait normality. However as discussed below, the power to detect linkage may be sensitive to significant deviations from normality. Unlike regression-based methods, VC linkage tests are only valid if the trait being analysed is normally distributed. When the true underlying distribution of the phenotype is multivariate normal, several important conditions are met that are critical for the LRT to have its usual asymptotic properties (Blangero *et al.*, 2000). If a trait violates the distributional assumptions, the results from VC methods may be biased and/or invalid. The impact of trait distribution violations for both types of methods are summarized next.

Regression-based methods

Deviations in trait normality (i.e., skewness and kurtosis). Most regression-based linkage methods have a type-I error rate that is robust to departures from normality for large sample sizes and fully informative markers (Feingold, 2002; Palmer *et al.*, 2000; Sham and Purcell, 2001; Sham *et al.*, 2002). However, different methods have considerably different power to detect linkage. Indeed, both the classic (HE-SD) and the trait cross-product (HE-CP) methods provide considerable power even when the assumption of multivariate normality is grossly violated (Forrest and Feingold, 2000). On the other hand, both the Forrest (2001) and Visscher and Hopper (2001) weighted extensions (HE-W), and Sham *et al.* (2002) MERLIN-regress method (HE-R) have very low power to detect a major QTL when analyzing non-normal traits. In the latter case, the detection of a QTL is still possible but it is restricted to large samples with informative markers and high QTL variance (e.g., 50%). Sham and Purcell (2001) did not report the behavior of their weighted method (HE-COM) with non-normal data, although it was shown to be robust for the analysis of selected samples.

Deviations in trait mean, variance and the effect of residual correlation. Another issue that may influence both type-I error and the power of regression-based methods is misspecification of the population mean, variance and correlation. This commonly happens when analyzing selected samples. Palmer *et al.* (2000) showed that the type-I error rate of both the HE-SD and the HE-CP methods was insensitive to deviations in the mean. Additionally, the HE-SD method consistently provided high power to detect linkage for different mean values. The HE-CP method, however, was considerably sensitive to misspecification of the mean, with power being highest at a mean value between the true population mean and the assumed sample mean. Sham and Purcell (2001) tested the power of their HE-COM method under different residual sibling correlations (i.e., the correlation obtained after removing the effects of the QTL) and concluded that although it may be reduced for low sibling correlations, it approaches that of variance components as the sib correlation increases. Their method requires the knowledge of the population mean and variance, although it is unclear how robust the method is to misspecifications of these parameters. Finally, Sham *et al.* (2002) provided a series of simulations to test the robustness of their HE-R method to misspecification of the population mean, but also to misspecifications of the variance and correlation. The nominal type-I error rate was maintained for different values for the mean, variance and heritability. However, misspecification of the mean considerably reduced the power to detect linkage. No effect on power was observed for the variance or heritability.

Thus, in summary, most regression-based methods seem to provide statistically valid tests for linkage even when analyzing non-normal traits. However, most extensions to the classic HE method (e.g., HE-CP, HE-W, HE-R) seem to lose much of their power when analyzing extremely non-normal data, or when the mean is misspecified. The exception to this is the original HE-SD method and possibly HE-COM (Feingold, 2002), which seem to be quite robust to these violations.

Maximum-likelihood variance-components methods

Deviations in trait normality (i.e., skewness and kurtosis). The type-I error rate of standard variance components analysis has been shown to be extremely sensitive to positive kurtosis (Blangero *et al.*, 2001). Similarly, with a range of simulations, Allison *et al.* (1999) showed that increasing kurtosis and skewness, while retaining similar trait means and variances, resulted in increased type-I error rate. This effect was aggravated in the presence of significant residual sibling correlation. Finally, Sham *et al.* (2002) demonstrated that VC consistently produces inflated type-I error when analyzing non-normal data. Thus, in practice, incorrectly assuming normality when there is marked kurtosis or skewness leads to an unreliable test for linkage: typically, the LOD score obtained is unrealistically large. The QTL effect size estimate, however, seems to be relatively robust to modest departures from multivariate normality (Blangero *et al.*, 2001).

Deviations in trait mean, variance and the effect of residual correlation. Frequently, we might have data on a normal distributed trait in which the sample mean, variance and correlation differ from the population estimates. This effect is closely associated with (i.e., reflects) selective sampling: the impact on variance components includes biased estimates of heritability (both due to the QTL and to residual genes) and reduced power. The latter reflects the fact that by misspecifying these distribution parameters, we are effectively influencing the estimated likelihood of observing a specific set of trait values and, in this way, we bias the evidence that each family provides for linkage.

Detection

One way to screen for normality deviations is to examine residuals by graphical methods (box and normal probability plots). If the residuals are symmetrically distributed around a mean value of zero, normality can roughly be assumed. Another way to screen for normality is to examine the distribution of the variable itself by either graphical or statistical methods. Graphical inspection of histograms allows for a quick inspection of deviations in skewness and kurtosis, but this may be misleading. Statistically, one can test if the values of skewness and kurtosis are significantly different from zero using the z distribution,

$$z = \frac{x - 0}{s_x}$$

where, x represents the sample skewness or the kurtosis excess value that can be calculated by most statistical packages and s_x is the respective standard error of x. The 0 is included in the formula only to illustrate the point of the test. The standard error for the skewness is approximately $\sqrt{6/n}$, and for the kurtosis $\sqrt{24/n}$, where n is the number of observations. Conventional but conservative (0.01 or 0.001) alpha levels can then be used to evaluate the significance of skewness and kurtosis with small to moderate samples. Some statistical packages such as SAS provide additional statistical tests for normality, namely the Kolmogorov-Smirnov

and the Shapiro-Wilk tests. The latter tends to be more powerful but it is restricted to sample sizes of up to 2000. All these tests should be taken with a grain of salt and be seen as complementary sources of information when investigating the normality of a variable.

Finally, one should test if the sample mean, variance and correlation (in the case of sib-pair data) differ from the population parameters. If the sample parameters are shown to be different from the parameters estimated from a random sample of the population, then ascertainment correction methods, as discussed in Section 13.1, may be necessary.

Correction

In this section, we discuss four possible approaches that may be used when analyzing a non-normally distributed trait. First and intuitively, a simple transformation will often be sufficient to normalize a non-normal distribution to such an extent that inference about linkage based on a likelihood approach becomes valid. In practical terms, Blangero *et al.* (2001) suggested that distributions with kurtosis (k) < 2 could be reasonably analyzed under the assumption of normality for pedigree trait values. For traits where $k \geq 2$, a general rule for transformation can be found in Box and Cox (1964), but this is not guaranteed to always work and does not necessarily ensure multivariate normality. Allison *et al.* (1999) further pointed out the additional problem of the residual sib correlation (r). In the presence of high residual correlation ($r = 0.5$), even small deviations of skewness and kurtosis ($0.5 < k < 1$) can lead to increased type-I error rate. Thus, in face of this, a wise approach may be to find the best transformation using the Box and Cox method for all variables with $k \geq 2$. Then, for all transformed and original variables with $k < 2$, perform linkage analysis using standard VC if the residual sib correlation can be shown to be ≤ 0.2. This should provide a reliable test for linkage, assuming that we are not analyzing a selected sample (Sham *et al.*, 2000b). On the other hand, if, even after transformation, a variable has kurtosis > 2 and the residual sib correlation is > 0.2, or the variable was measured in a selected sample, then we may use one of the approaches described next.

A second possible approach is to define a more robust LRT. Different methods have been developed that explore this. Blangero *et al.* (2000) derived the distribution of the LRT under an incorrect probability model and proposed an accurate robust LOD score applicable to any pedigree structure and for any trait distribution. This correction is implemented in the computer package SOLAR (Almasy and Blangero, 1998). Another robust VC approach has been developed by Sham *et al.* (2000a). By considering the trait values as the dependent variable, this 'reverse' VC method is no longer bound to the tight distributional assumptions and leads to a LRT that is valid in selected samples and non-normal data. However, although valid, this method seems to provide very low power to detect linkage when analyzing non-normal data. Finally, Epstein *et al.* (2003) extended the traditional VC method to accommodate censored data. This was done by modifying the likelihood calculation so that it accounts for the censoring event. Their simulations suggest that when analyzing a 25% censored trait, their method returns the appropriate type-I error rate and unbiased parameter estimates (unlike traditional VC). However, as with other methods, the power is only significant for large QTL effects (25%).

A third approach to analyze non-normal data is to return to more robust procedures which do not require any distributional assumptions, such as the classic Haseman-Elston method. Due to the low power of this method when compared to variance components, this choice may be wise but suboptimal.

The final approach deals with the case in which the sample mean, variance and correlation differ from the population parameters. As mentioned above, when using variance components to analyze such sample, it may be appropriate to alter the likelihood calculation as described in Section 13.1. If successfully applied, these ascertainment correction methods will give the desired result of shifting the sample parameters closer to the population parameters.

13.4 Outliers

Description

One potential source of bias in linkage studies is the presence of phenotypic outliers. Although these are closely related to deviations in trait distribution, special attention is given to them here. A univariate outlier can be defined as a value from a continuous variable which falls on or beyond a predefined threshold t of the distribution. The threshold t should be chosen according to the variation that the trait exhibits in a random sample of the population; often, t is set at $\pm$ 3 standard deviations (SD) from the mean. An outlier can represent an error in measurement, in data recording, in data entry, or it could represent a legitimate value that just happened to exceed t. For example, as part of data cleaning for a large asthma study performed at the Queensland Institute of Medical Research, 38 continuous variables were screened in 100

randomly selected asthma questionnaires, a total of 3800 fields. Of these, 11 (0.3%) were found to contain data entry errors of some sort and two of these (0.05%) resulted in an outlier ($\geq$ 3 SD) for the respective variable. Outliers can also be present in a multi-dimensional space. Although individual values for a continuous trait may all be within 3 SD from the mean, the *joint* distribution of two or more related values may be beyond what would be expected from a multivariate normal distribution. For example, consider two sibs with values -2.8 and +2.8 for a normally distributed trait, with mean 0 and SD of 1. Although individually both observations may not be considered outliers, their joint distribution would be very unlikely, being only expected to occur under multivariate normality with a frequency of 10^{-5}. Both univariate and multivariate outliers can have a large impact on linkage analysis.

Impact on linkage analysis

Univariate outliers can heavily influence both the skewness and kurtosis of a distribution, as well as the mean and variance. In this way, they can increase the type-I error rate of VC analysis and reduce the power of regression-based methods, such as Forrest's (2001) weighted extension and Sham *et al.*'s (2002) MERLIN regress method. Possibly more important than the effect of individual outliers is the impact of multivariate or "familial" outliers (Barnholtz *et al.*, 1999). If the joint distribution of the phenotypes of two or more relatives in a family constitutes an outlier in the multivariate space, the likelihood contribution of that family can be large and, thus, can inflate the total likelihood value or LOD score (de Andrade *et al.*, 2003). This may provide misleading results, unless the family's trait values can be shown to be valid.

Detection

As with normality violations, individual outliers can be detected graphically by means of various plots of the quantitative variable and of the residuals. Alternatively, the sample can be screened to identify values which fall beyond a predefined threshold *t*.

To detect families that may harbor multiple phenotypic outliers, one can calculate for each family the mean trait deviation from the sample mean; a normal probability plot can then be constructed using this quantity to identify families which on average deviate excessively from the sample mean. Additionally, when using VC analysis, familial outliers may be detected by investigating individual family likelihoods, log-likelihoods or LOD scores. Programs such as MERLIN and Mx provide an option ("--perFamily" and "%P", respectively) to output these. Plots of these values for each family may help to identify families that have a large contribution to the overall linkage signal and, therefore, that may warrant further investigation (de Andrade *et al.*, 2003).

Correction

If an outlier is detected, one should first check if it has arisen as a result of an error in measurement, in data recording or in data entry. If so, and when possible, it should be replaced by the correct value. If, on the other hand, the outlier may in fact be biologically plausible, the following two options may provide a more robust analysis. The first alternative is winsorization (Fernandez *et al.*, 2002), that is, any observations that are > *t* SD from the mean should be recoded to precisely *t* SD from the mean (for a reasonable choice of *t*). A second alternative when using VC is to use a more robust LRT as mentioned above (Blangero *et al.*, 2000; de Andrade *et al.*, 2003).

13.5 Pedigree Errors

Description

The pedigree structure represents the true familial relationship between every pair of individuals in a given pedigree. This not always matches the relationships reported by subjects in questionnaires or interviews. Whenever the assumed familial relationship for a given pair of individuals differs from the real relationship, we are in the presence of a pedigree error. For small pedigrees, typical errors are derived from an incorrect report of zygosity status (assumed full sibs which are in fact monozygotic twins, or *vice-versa*), false paternity (assumed full sibs which are in fact half sibs) and unknown adoption (assumed full sibs which are in fact unrelated).

Impact on linkage analysis

Both for regression methods and variance components, pedigree errors can increase the type-I error rate and decrease the power to detect linkage. False paternity and unreported adoption are likely to result in allele segregation patterns for parent-offspring pairs that are inconsistent with Mendelian laws of inheritance; these errors can result in the exclusion of families from analysis and, thus, may lead to a loss of power to detect linkage. Additionally, even if they do not result in Mendelian inconsistencies, false paternity and unreported adoption can produce biased estimates of IBD sharing for alleged full sibs. If present, this bias will always tend to overestimate IBD for sib pairs which should otherwise show low

IBD (half sibs cannot share more than 1 allele IBD, whereas an adopted sib will always have 0 alleles IBD with nonadopted sibs). Since these pairs are likely to have low phenotypic correlation, they will provide false evidence against linkage.

Incorrect report of zygosity status can influence the power of linkage in two ways. Monozygotic (MZ) sibs have identical genotypes and, typically, high phenotypic correlations. Because of this, they cannot provide evidence for linkage to an arbitrary marker *M*; an MZ pair will always provide the same strong signal in favor of linkage at all markers across the genome. Thus, if an assumed dizygotic (DZ) twin pair is in fact MZ, this will inflate the linkage test statistic at all markers tested, which provides false evidence for linkage. Alternatively, assumed MZ twins can in reality be DZ twins. In this case, by not including such misclassified pairs in the analysis as DZ twins, the power to detect linkage may be reduced. Thus, a simple rule to remember is that if a class of relative pair is incorrectly assigned to a class which on average has higher IBD, it will result in decreased power to detect linkage. If, on the other hand, the same class of relative pair is incorrectly assigned to a class which on average has lower IBD, it will result in increased type-I error. Since we are discussing MZ twins in the context of linkage analysis, it is worth pointing out that although an MZ twin pair alone does not provide evidence for linkage, when there is an additional sibling in the family, the appropriate inclusion of both MZ twins in the analysis may increase power (Evans and Medland 2003).

Detection and correction

Both likelihood-based methods (Boehnke and Cox, 1997; Epstein *et al.*, 2000; Goring and Ott, 1997; McPeek and Sun, 2000) and allele-sharing methods (Ehm and Wagner, 1998; McPeek and Sun, 2000; Olson, 1999) have been proposed for pairwise relationship estimation using genotype data. More recently, Abecasis *et al.* (2001b) developed a practical allele-sharing approach to graphically verify if individuals with a given specified relationship have the expected pattern of allele sharing. Specifically, it identifies pairs of individuals where the pattern of their identity-by-state (IBS) sharing across the genome is inconsistent with their assumed relationship. The method is easily applicable to genome-wide genotyping data and it involves plotting for each pair of relatives (or individuals, for that matter) the mean IBS across all markers against its standard deviation, and locating outliers from this bivariate distribution (Cherny *et al.*, 2001). This method is implemented in GRR (Abecasis *et al.*, 2001b). Once identified, the correction of pedigree errors is in most cases straightforward: the misclassified relative pairs are recoded and a new corrected pedigree file for linkage analysis is created.

13.6 Genotyping Errors

Description

A genotyping error is defined as a discrepancy between the true genotype of an individual for a given locus and the assumed genotype as inferred by molecular biology techniques. Even with the most modern techniques, the inferred genotype does not always match the true genotype, and this has been shown to occur at a rate of 0.5–7% (Brzustowicz *et al.*, 1993). Factors that may contribute to this mismatch between the assumed and the real genotype include sample swapping, data processing and data entry.

Impact on linkage analysis

If a genotyping error leads to segregation patterns inconsistent with Mendelian laws of segregation (see below), then it will reduce the power to detect true linkage by reducing the number of families available for analysis. If, however, it does not lead to violations of Mendelian inheritance, which is the case for many genotyping errors, then the impact on linkage analysis is more complex, depending on factors such as the method used for IBD estimation (singlepoint or multipoint), the study design used and the method of analysis.

Genotyping errors affect multipoint IBD estimation more dramatically than singlepoint estimation (Goring and Terwilliger, 2000a). The reason for this seems to be that multipoint analysis assumes that IBD distributions are linearly related along the genome. Genotyping errors will introduce deviations from this linearity which may lead to a systematic underestimation of IBD along the genome. In other words, with singlepoint analysis, a genotyping error at a marker *I* will only impact the IBD estimation at that locus (typically underestimating it); on the other hand, with multipoint analysis that same genotyping error not only biases the IBD estimation at the marker *I*, but also at all other markers along the same chromosome. Hence, in practical terms, singlepoint analysis may result in higher LOD scores than multipoint analysis, despite the use of more information in the multipoint analysis. More information will only lead to a more powerful test if this additional information is accurate (Goring and Terwilliger, 2000a).

Thus, undetected genotyping errors often lead to IBD estimates (singlepoint or multipoint) between relative pairs being underestimated. The effect that this may have on linkage analysis differs between different study designs. For affected sib-pair designs (ASP),

a moderate amount of error (0.5–1%) was shown to result in a loss of linkage signal between 10–58% (Douglas *et al.*, 2000). Abecasis *et al.* (2001a) reported similar results (10–25% reduction in the LOD score). By contrast, for discordant sib-pair designs (DSP), which test for lower than average IBD sharing between sibs, genotyping errors may have the opposite effect and provide false evidence for linkage. Finally, for a quantitative trait analyzed in a random sample of sib pairs using a multipoint VC approach, genotyping error rates of 0.5–1% resulted in only 2–4% reduction in the evidence for linkage (Abecasis *et al.*, 2001a). Together, these results show that methods using the full allele-sharing distribution (VC, HE) are more robust in the presence of genotyping error than methods which rely exclusively on mean allele sharing (IBD sharing statistics).

Detection

We discuss here two possible methods for the detection of genotyping errors. The first of these methods (O'Connell and Weeks, 1998) focuses on the detection of conspicuous inheritance inconsistencies and, therefore, it requires parental genotypes. Similar methods have been proposed by others (Lathrop *et al.*, 1983; Ott, 1993; Stringham and Boehnke, 1996). The method proposed by O'Connell and Weeks (1998) involves four error-checking algorithms which handle different types of datasets and different degrees of difficulty in the identification of an error. The first algorithm (nuclear-family algorithm) uses the known genotypes for each marker to check for inconsistencies between parents and offspring. This algorithm flags simple errors but does not involve genotype elimination; therefore, it is very efficient to check raw laboratory datasets that have not been checked previously and which might have numerous data-entry or genotype-scoring errors. If there are no errors or if the errors have been corrected, then there may still be inconsistencies due to the revised information or to the genotype information given by other members of the pedigree other than first-degree relatives. The second algorithm (genotype-elimination algorithm) detects these additional more subtle errors. The algorithm determines if all the genotype information available for a given pedigree (and not just that of individual nuclear families) is consistent or not with Mendelian laws of inheritance. However, in some cases, this algorithm may not be able to pinpoint the individual whose genotype is the source of the problem. The two final algorithms may be useful to find the source of these even more subtle errors. The critical-genotype algorithm identifies individuals that eliminate the pedigree inconsistency when the respective genotype is removed from the data. There may be none, one or more of these critical genotypes. If there are none, the effect of simultaneously removing more than one genotype may be tested. If there is only one, then that genotype is very likely to represent the error (though that may not always be the case) and should be eliminated. If there is more than one critical genotype, then one should invoke the fourth and final odds-ratio algorithm to determine which is more likely to be the erroneous genotype. This procedure is implemented in PedCheck (O'Connell and Weeks, 1998).

The second method was proposed by Abecasis *et al.* (2002) and is applicable even when parental genotypes are not available. This multipoint likelihood-based method identifies genotypes that imply a recombination pattern that is not supported by neighboring markers and which are likely to be erroneous. Abecasis *et al.* showed that this significantly improves the rate of error detection from 16% with only sib-pair genotypes, to greater than 60% with four genotyped siblings and over 90% when both parents and at least two offspring are genotyped. It is also likely to reduce the false-positive rate of error detection, that is, the frequency with which errors are flagged when in reality they do not exist (e.g., true rare recombinant genotypes). This method detects subtle errors which do not result in Mendelian inheritance incompatibilities and should be used when a dataset has already been cleaned of such errors. This algorithm is implemented in MERLIN (Abecasis *et al.*, 2002).

Correction

Controlling the impact of genotyping errors on linkage can take place prior to or during analysis. At the pre-analytic stage, the first step is the correction of genotyping errors that lead to Mendelian inconsistencies. If a dataset consists of a small number of large pedigrees, the individuals with the possible genotyping errors should be rescored and/or retyped. In this scenario, the gain in linkage information may outweigh the cost and time involved. On the other hand, if a dataset consists of a large number of small pedigrees or whenever rescoring/retyping is not feasible, the problematic genotypes should be excluded from analysis for the marker with the observed inconsistencies. The second pre-analytic step involves the correction of errors which do not lead to Mendelian incompatibilities but which imply the occurrence of unlikely recombination patterns. Unlike with Mendelian errors, linkage analysis can still be performed in the presence of these more subtle errors. In fact, as mentioned above, multipoint VC analysis of

random sib-pair samples has been shown to be quite robust to an error rate of up to 5% (Abecasis *et al.*, 2001a). Thus, correcting these errors is not essential, but it has been shown to improve the power of linkage when using other study designs such as affected sib pairs and, potentially, affected probands (see Abecasis *et al.*, 2001a; Cherny *et al.*, 2001; Douglas *et al.*, 2000). If correction is in fact desired, then it consists of removing the critical genotypes from a particular family at the problematic marker, that is, the genotypes of individuals that once removed eliminate the pedigree inconsistencies.

Finally, analytical methods have been proposed that allow for a more robust analysis of linkage in the presence of genotyping errors. Goring and Terwilliger (2000a) provided a framework by which multipoint likelihoods can be computed while modeling for genotyping errors. A similar approach has recently been implemented in the latest version of MERLIN (Abecasis and Wigginton, 2005).

13.7 Marker Informativeness, Density and Genetic Map

In theory, if linkage analysis was performed for a large number of markers across the genome in an appropriate sample, and if all those markers were perfectly informative, singlepoint IBD estimation would provide a powerful basis for nonparametric linkage analysis. Furthermore, even if not all markers were fully informative, multipoint IBD estimation with an accurate genetic map would still provide a robust basis for linkage. In practice, however, data are usually collected for a small number of markers (i.e., smaller than optimal), these markers are not perfectly informative and some markers may not be mapped accurately in available human genetic maps. As discussed below, marker informativeness, the choice of marker density and genetic map all have influential effects on QTL mapping. These three factors are discussed under the same section since they often interact to determine the power of linkage.

Description

Marker informativeness

Detecting linkage by nonparametric methods requires the estimation of IBD sharing between relatives at different locations along the genome. The more accurate the estimation is at any given location (i.e., the closer the estimator $\hat{\pi}$ is to the true π), the more powerful the analysis will be. The accuracy of the estimation depends on the informativeness of the marker. Hence, a marker is said to be perfectly informative if it is possible to determine with certainty if a relative pair shares zero, one or two alleles IBD at that locus.

For a sib-pair this is always the case when both parents are heterozygous, unless both parents and at least one child have the same heterozygous genotype (Guo and Elston, 1999). A marker is said to be uninformative for linkage if the information it provides about allele sharing is the same as when no genotyping data are available for that locus; for example, when it implies that a sib pair can share zero, one or two alleles IBD with probabilities ¼, ½, ¼. This is only the case when both parents are homozygous, for the same or different alleles. Finally, a marker is said to be partially informative for linkage if the information it provides does not allow for the unambiguous classification of allele sharing but it concentrates the allele sharing possibilities into only two possibilities: for a sib-pair, these can be zero or one, zero or two, and one or two alleles IBD. This is the case when only one parent is homozygous and when both parents and at least one sib are heterozygous for the same alleles.

Marker density

If IBD sharing for a given sib-pair was exactly the same at all locations along the genome, only one marker would have to be typed per sib-pair. However, the pattern of IBD sharing varies across the genome: if recombination events are common between two markers, then the pattern of IBD sharing is expected to differ between them. On the other hand, if two markers are so close together that recombination events between the two are rare, then IBD sharing is expected to be the same or very similar for the two markers. The question of marker density refers to the assessment of the optimal number of markers necessary to genotype in a given region to extract nearly all the IBD information present in such region. In other words, the marker density that is required to maximize the probability that the sib-pair IBD distribution at a trait locus may be captured (and not missed) by the IBD distribution estimated at flanking marker loci.

Genetic map

Multipoint linkage analysis requires the specification of the recombination fractions between markers, which are often provided as intermarker genetic distances (recall from Chapter 7 that a recombination fraction can be converted to a genetic distance and *vice-versa* using the Kosambi or Haldane mapping functions). Such recombination fractions are estimated empirically by observing the frequency of recombination events in human meiosis from reference pedigrees, such as the families of the Centre d'Etude du Polymorphisme Humain (CEPH; Yu *et al.*, 2001) or the Icelandic

sample (Kong *et al.*, 2002). The recombination fractions are then converted to genetic distances which are used for establishing genetic maps of the human genome. However, the accuracy with which these distances are estimated is greatly dependent on the total number of meiosis investigated and the polymorphism content of the markers (Leal, 2003). Additionally, genetic distances differ between males and females. Thus, genetic maps are likely to differ between mapping consortiums which analyze different reference families. For example, the Marshfield map (Broman *et al.*, 1998) incorporates a total of 8325 markers mapped by analyzing only 188 meioses (eight large CEPH families). The other map currently used by most researchers (deCODE Genetics) consists of 5136 markers genotyped in 146 nuclear families containing 1257 meioses (Kong *et al.*, 2002). There are 5012 markers which have been mapped by both consortiums; on average, the accuracy of the intermarker distances is about five times better in the deCODE map. In fact, with the Marshfield map, the order of markers separated by less than 3 cM is somewhat unreliable, whereas the average resolution of the deCODE map is about 0.5 cM. Recently, Kong *et al.* (2004) combined the CEPH and deCODE data to produce the genetic map with highest resolution to date.

Impact on linkage analysis

Marker informativeness

Two sibs can share zero alleles IBD (with frequency ¼), one paternal allele IBD (½), one maternal allele IBD (½) or two alleles IBD (¼). Therefore, if the true value of π could be observed at a marker, it would take the values of 0, 0.5, 0.5 and 1, with equal frequencies; from this, it follows that $\text{Var}(\pi) = \frac{1}{8}$. When analyzing a perfectly informative marker, $\hat{\pi}$ matches the true π and $\text{Var}(\hat{\pi}) = \frac{1}{8}$. However, when analyzing partially informative markers, $\hat{\pi}$ will never assume the values of 0 or 1, even when they are the true π values (see also *Table I* in Rijsdijk and Sham, 2002). Thus, in this case, $0 < \hat{\pi} < 1$ and $\text{Var}(\hat{\pi}) < \frac{1}{8}$. The extent to which the $\text{Var}(\hat{\pi})$ deviates from ⅛ indicates how inaccurate the IBD estimation is, that is, it reflects marker informativeness. For example, when a marker is completely uninformative for all families, $\hat{\pi}$ will always be 0.5 and, hence, $\text{Var}(\hat{\pi}) = 0$. The lower the IBD information provided by a marker, the lower the $\text{Var}(\hat{\pi})$ and the lower the power to detect linkage. For example, the classic Haseman-Elston method regresses the pairs' squared trait difference on $\hat{\pi}$, the test for linkage being a one-sided t test of the regression slope. However, the regression slope is determined by observations in which $\hat{\pi}$ deviates from 0.5, namely those that approach 0 or 1. Under the hypothesis of linkage, the less informative a marker is, the fewer the number of sib pairs that approach the extreme $\hat{\pi}$ values of 0 and 1 and, hence, the flatter the estimated regression slope will be. Low marker informativeness reduces the power of VC and allele-sharing methods in a similar fashion (Halpern and Whittemore, 1999). Finally, Sham *et al.* (2000a, and Sham and Purcell 2001) demonstrated that the NCP of VC and different HE methods is expressed as a function of marker informativeness (see Equation 13.3).

Marker density

The number of markers used in a linkage study determines the amount of inheritance information that is extracted along the genome. The more information extracted, the more powerful the test for linkage will be. For an infinitely dense map of perfectly informative markers, all inheritance information would be captured (100%). This situation, however, is unrealistic. Kruglyak (1997) showed that a 1 cM map of microsatellites with 10 equally frequent alleles extracts 97% of the available information, whereas a similar dense map of SNPs with 0.5-0.5 allele frequencies captures 88% of the information (*Table 13.1*). Kruglyak showed that both scenarios would be expected to yield a powerful linkage test, although one would have to type over 3600 markers to create such a dense map. This was unfeasible for most large-scale projects until very recently, but it is now viable with the advent of high-throughput SNP genotyping.

Until recently, most linkage studies used a 10 cM map, typically with 300–400 microsatellite markers. In this situation, and assuming that the markers used had between 3 and 10 equally frequent alleles, a 10 cM map was expected to extract only 45–68% of all

Table 13.1 Proportion of inheritance information extracted by different genetic marker sets

Marker spacing (cM)	Number of markers	Microsatellites When no. alleles/marker = 10	5	3	SNP When MAF = 0.5	0.3	0.1
10	362	0.68	0.58	0.45	0.29	0.26	0.15
5	723	0.82	0.78	0.68	0.50	0.46	0.27
3	1205	0.90	0.87	0.80	0.65	0.63	0.42
2	1808	0.94	0.91	0.87	0.75	0.73	0.55
1	3615	0.97	0.96	0.93	0.88	0.87	0.73

Adapted from Kruglyak (1997)

the inheritance information available (assuming genotypes are available for both parents; Kruglyak, 1997). The question then is whether an initial screen for linkage requires higher information content than this. In this matter, Kruglyak concluded that a marker set which extracts between 69–88% information is acceptable for initial linkage studies. This can take the form of a 1–2 cM map of moderately polymorphic (0.5-0.5 to 0.8-0.2) SNPs or, for typical microsatellite markers (heterozygosity $0.65 < H < 0.8$, see section about detection), of a 3–7 cM map.

Thus, for initial screening studies whose main goal is to detect a significant linkage signal, increasing the density of markers beyond these rough guidelines may not be cost effective, since most of the inheritance information has been extracted. However, once linkage has been found in a given region, a fine mapping study may be designed to increase the precision of the QTL location estimate. The effect of increasing marker density on the precision to detect a human QTL in a previously linked region was investigated by Atwood and Heard-Costa (2003). The authors simulated genotypic data for 16 polymorphic markers at a 10 cM interval on a 150 cM chromosome, with a QTL positioned between markers 7 and 8, at 75 cM. This 10 cM map was simulated for different sample designs, which always included 200 nuclear families of size 7. For example, when the total phenotypic variation was partitioned into a QTL component (20%), a polygenic component (20%), a covariate component (25%) and to error (35%), the maximum LOD score obtained by VC analysis for the 10 cM map was on average 6.72 cM away from the true location (75 cM), reflecting the poor accuracy of this map. The QTL location predicted by a 2 cM fine map was on average only ~10% (0.65 cM) closer to the true location; although small, the improvement was reported to be significant. When compared to the 2 cM map, no significant improvement in accuracy was obtained with the 1 cM and 0.5 cM maps (1% and 2% closer to the true location). These results provide weak support for fine mapping when the QTL effect is small and no justification for it below a resolution of 2 cM.

In summary, initial screening linkage studies that use a 10 cM microsatellite with an average marker heterozygosity (H) <0.75 may not only extract insufficient inheritance information to provide a powerful test for linkage but, additionally, they may provide poor QTL mapping precision. Based on Kruglyak (1997) and Atwood and Heard-Costa (2003), a cost effective approach to linkage analysis may be to perform: (i) an initial 5 cM microsatellite scan (~700 markers) with markers having an average $H \geq 0.75$, followed by a 2 cM fine mapping of linked regions with microsatellites or SNPs; or (ii) an initial 2 cM scan with moderately polymorphic (0.5-0.5 to 0.8-0.2) SNPs (~1800 loci). As pointed out by Terwilliger *et al.* (1992) and Kruglyak (1997), when using fine maps of the order of 2 cM or less, markers with low heterozygosity can extract much of the inheritance information (since markers are very close together). Hence, even SNPs with a minor allele frequency of 0.2 can be selected for genotyping. One brief last point to make here relates to the impact of using a dense panel of SNPs for linkage analysis, such as those currently being designed for genome-wide association studies. In such dense panels of SNPs, many markers will be in strong linkage disequilibrium (LD), a topic that will be covered in greater detail in subsequent chapters. Briefly, if two markers are in strong LD, their genotypes are strongly correlated. Most algorithms currently used for multipoint IBD estimation do not account for this effect, which can lead to biased estimates of IBD sharing and consequent increase in the type-I error of linkage analysis (Abecasis and Wigginton 2005).

Genetic map

Errors in intermarker recombination fractions as specified by genetic maps have no impact on singlepoint linkage analysis. Effectively, in this case, IBD estimation at each marker only uses the available genotypic data for that marker and disregards any information regarding genetic distances. By contrast, in addition to the genotypic data at an arbitrary marker l, multipoint linkage analysis uses the genotypic data at other markers together with the recombination fractions between markers to estimate IBD at marker l. Thus, as shown by Daw *et al.* (2000), multipoint analysis can have increased type-I error rate if a genetic map is used that misspecifies the recombination fraction between markers. In addition, Halpern and Whittemore (1999) showed that for a variety of situations, misspecification of the intermarker distance by 4 cM reduces the power of multipoint analysis to detect a QTL on average by 28% and, as a result, multipoint analysis was always less powerful than singlepoint analysis.

An additional problem is the difference between female and male recombination rates which, in most regions of the genome, results in sex-specific genetic distances. Nonetheless, nearly all published linkage scans are based on sex-averaged maps. Daw *et al.* (2000) showed that a multipoint analysis that assumes the same intermarker distances for males and females when this is not true, incurs a small loss of power but, more importantly, a significant increase in type-I error rate. Indeed, their simulations suggest that when using a sex-averaged map, false-positive findings will

increase at a rate which is proportional to the magnitude of the true difference between female (F) and male (M) genetic distances. For example, in regions of the genome where the true F:M distance ratio is 2:1, the LOD score for an unlinked locus is not expected to be significantly upwardly biased by assuming a sex-averaged map instead of sex-specific maps. However, in regions where the true F:M distance ratio is 5:1 (which appears to exist over large portions of the human genome), using a sex-averaged map results in a 5–15% upward bias in the LOD score recorded under the null hypothesis of no linkage. Thus, although it is unlikely that we are missing true evidence for linkage by using sex-averaged instead of sex-specific maps, as discussed in Chapter 11, we must be careful when reporting significant linkage results in a region where large differences in recombination rates have been reported for males and females.

Detection

Marker informativeness

There are at least two situations in which assessing the degree of marker informativeness can be useful for linkage analysis. One is at the stage when a linkage study is being designed and a choice has to be made about which and how many markers to genotype. This is the situation in which the indexes described next better apply: reported marker allele frequencies and simulated map densities can be used to select the most cost-effective marker set. The second situation occurs when genotypic data has already been collected and we wish to investigate the amount of inheritance information that was extracted along the genome. This is a typical quality control issue. In the latter case, we can use the sample allele frequencies and the intermarker distances to assess the quality of our marker set. Although the estimates may be biased, they will be consistent.

As mentioned above, marker informativeness reflects the amount of information it provides about allele sharing in a pedigree. A common measure of average informativeness of a marker is the polymorphism information content (PIC), which is simply the probability that in a nuclear family drawn at random from the population the marker will be perfectly informative. In other words, the PIC is the probability that both parents will be heterozygous and that both parents and the child will not have the same heterozygous genotype. Therefore, the PIC is defined as

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2 = H - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$$

where, p_i is the population frequency of the ith allele, $\sum_{i=1}^{n} p_i^2$ represents the probability of a given individual being homozygous for any i of the n possible alleles (hence, $1 - \sum_{i=1}^{n} p_i^2$ is H, the heterozygosity index), and $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$ represents the probability that both parents and the child have the same heterozygous genotype. The PIC was originally proposed by Botstein *et al.* (1980) for a rare dominant disease but later shown not to depend on any trait characteristics but only on the number of alleles and the frequency of each allele at the marker locus (Guo and Elston, 1999). For example, for a marker with 10 equally frequent alleles, $H = 0.9$ and PIC = 0.891.

As mentioned in above, marker informativeness will be reflected on the accuracy of IBD estimation and, hence, on Var $(\hat{\pi})$. For perfectly informative markers, Var $(\hat{\pi})$ = 1/8, whereas for partially informative markers, Var $(\hat{\pi})$ < 1/8. More formally, Rijsdijk and Sham (2002) demonstrated that the Var $(\hat{\pi}) = PIC/8$, which is equivalent to Var $(\hat{\pi}) = PIC \cdot \mathrm{Var}(\pi)$. In other words, the PIC can be thought of as the ratio of the variance of $\hat{\pi}$ to the variance of π. But, by analogy to singlepoint IBD estimation, the PIC is only useful to assess the amount of inheritance information that will be extracted at a given position l if we used the genotypic data from that locus alone. However, most linkage studies now use multipoint IBD estimation, which means that the amount of inheritance information extracted at a position l depends not only on the informativeness of the marker at that location but also on the informativeness of all other neighboring markers. This observation led Rijsdijk and Sham (2002) to propose another measure of average marker informativeness, the multipoint polymorphism information content (MPIC), which can be simply defined as the ratio of the variance of the multipoint $\hat{\pi}$ to the variance of π. A more commonly used measure of marker informativeness that is applicable to both singlepoint and multipoint analysis is the entropy-based approach proposed by Kruglyak *et al.* (1996) that is implemented in MERLIN. Lastly, Nicolae and Kong (2004) developed specific measures of information content for the analysis of discrete traits with allele sharing statistics.

Marker density

As described above, either the MPIC or another index of multipoint marker informativeness can be used to assess the power to detect linkage for a given marker density (either prior or after data collection). This is an essential step for the design of powerful linkage

studies and/or for the interpretation of linkage results. A separate but important issue also related to marker density is the degree of linkage disequilibrium (LD) found between markers. Very dense marker sets are likely to have many markers in strong LD which, as outlined above, can inflate linkage statistics. Therefore, when conducting multipoint linkage analyses on dense marker sets it is important to estimate the degree of LD between markers, which can be done by calculating one of the two common LD metrics, D' and r^2. These will be described in great detail in subsequent chapters.

Genetic map

Two points to note here: first, it is important to investigate how accurate the intermarker distances being used for analysis are. To do this, the predicted genetic location for all markers should be compared between different genetic maps (e.g., Marshfield and deCODE) and against their respective physical positions. In addition, we may want to compare the number of recombinant meiosis observed between two given markers in our sample with the number that would be expected based upon their reported genetic distances. GENEHUNTER can be used for this. For markers with large inconsistencies, analysis should proceed with caution (see below). The second point concerns sex-specific genetic distances. It is wise to compare a sex-average map for the genotyped markers with the correspondent sex-specific maps. This will provide a hint to potential regions where false-positive linkage may result as a consequence of using sex-average maps.

Correction

Marker informativeness

Even when a linkage study has carefully chosen highly polymorphic markers to be genotyped, it is always possible that some markers will not be perfectly informative for some families. In this situation, multipoint IBD estimation provides a more powerful linkage test than singlepoint estimation. However, this is only true if the intermarker distances are correctly specified (Halpern and Whittemore, 1999). Thus, a useful approach may be to begin by performing multipoint analysis for all markers; then, for markers where there are inconsistencies in the location reported by different genetic maps, between genetic distances and physical positions, or between observed and expected number of recombinant meiosis, perform singlepoint analysis and compare with the multipoint results.

Marker density

Low marker density limits the power of linkage analysis by reducing the amount of inheritance information extracted from a given genomic region. In principle, if correct intermarker distances are available, multipoint analysis can attenuate this limitation. In fact, with multipoint analysis it is possible to estimate IBD at any position between consecutive markers, although these estimates in some situations may not be very accurate. For example, when using a 10 cM genome scan, most linkage packages such as MERLIN, SOLAR or GENEHUNTER, give the user the option to analyze n equally spaced locations in each 10 cM interval, hence effectively increasing the density of the analysis by a factor of $n + 1$. The information that is extracted at each of these n locations (for which there is no actual genotype) is dependent on factors such as the initial marker spacing, marker polymorphism content and the accuracy of the intermarker distance. Thus, plots of multipoint information content may be used to assess how much information is in fact being used between consecutive markers. When strong LD is detected between markers, then it is important to account for this when estimating multipoint IBD sharing. One approach may be to simply prune the original genotyping data to leave only markers that are not in strong LD. It can be argued that not much information is lost, since the removed markers will still be 'tagged' by markers kept for analysis. Recently, a more efficient approach was described that actually models the observed LD between markers when estimating multipoint IBD, resulting in a more robust test for linkage (Abecasis and Wigginton 2005). This approach is implemented in the software MERLIN.

Genetic map

The first point to note when accounting for the potential impact of genetic map errors on linkage refers to misspecified intermarker distances. To minimize this, at this stage the best bet may be to use a genetic map that is based upon the combined linkage-physical map recently made available by Kong *et al.* (2004). David Duffy's webpage (http://www.qimr.edu.au/davidD/index.html) provides an interpolated genetic map based upon this information that is regularly updated to account for the new NCBI build releases. In addition, it may be useful to perform both multipoint and singlepoint linkage analysis. Regions where both provide conflicting evidence for linkage should be interpreted with care. The second point refers to the potential bias introduced by using sex-average maps. In regions of the genome where the F:M distance ratio is

large (e.g., 5:1), using sex-average maps will tend to significantly inflate the LOD scores for unlinked loci. In face of this, one possible approach to account for such bias may be the following: first, perform multipoint analysis using sex-average maps as a first screen to identify regions of significant linkage. Some of these regions may be inflated by the incorrect use of sex-averaged maps. Then, follow this initial screen by analyzing the regions of significant linkage with more accurate sex-specific maps, for example as reported by Kong et al. (2004). The peaks that persist after such analysis are more likely to be true positive results.

Finally, analytical methods which intrinsically account for map uncertainty have been proposed (Goring and Terwilliger, 2000b; Stringham and Boehnke, 2001). These, however, are not yet commonly used. Thus, from all that was mentioned in this section, the message seems to be to perform multipoint analysis as a first screen for linkage. This may minimize the impact of low polymorphism information content and/or of a low marker density, a problem that may still afflict most studies. Then, for regions with markers that have unreliable genetic locations, perform singlepoint analysis and compare the results. In the case of linked loci which lie in regions of the genome of high F:M distance ratio, perform multipoint analysis using sex-specific maps.

13.8 Quality Control Guidelines

Table 13.2 summarizes the main issues discussed in this chapter. Based on this, we suggest here some basic guidelines which may help to improve the robustness and reliability of linkage analysis.

First, describe to the best extent the ascertainment scheme. This includes estimation of the proportion of the population that was ascertained and detailed characterization of the sample (family number, size and structure).

Second, investigate the phenotypic data. This includes (i) testing deviations from normality (rule of thumb: if kurtosis > 2, apply Box-Cox transformation; if still non-normal, use a robust test statistic for linkage analysis), (ii) comparing the sample mean, variance and correlations with parameters from a random sample of the population, and (iii) identification of outliers and appropriate correction (replacement with correct value or winsorization).

Third, check pedigree relationships with all genotyped data available using an allele-sharing method (e.g., GRR). Reclassify any errors.

Fourth, investigate the genotypic data. Identify families which result in Mendelian inheritance inconsistencies for a given marker

Table 13.2 Summary of the impact of different factors on the type-I error and power of linkage analysis

Factor	Impact on linkage analysis	Detection	Correction
1. Selective sampling	*Increases number of informative sib-pairs.* If properly analysed, increases power. If not, increased type-I error rate with VC.	Estimate degree of ascertainment (Mx). Compare sample mean, variance and correlation with population parameters.	Use robust methods (e.g., HE, MERLIN-regress) or apply ascertainment correction in VC.
2. Sample size			
QTL heritability	*NCP proportional to* $V^2_{Q_L}$ and $V^2_{D_{QL}}$. Low heritability ⇒ lower power ⇒larger sample sizes.	Analytical: calculate the expected NCP under a variety of model parameters (GPC, POLY). Empirical: simulate data under a specific genetic model and study design. Estimate power as proportion of datasets that exceed a given critical value.	Study design stage: choose the adequate sampling design, sample size and sibship size to achieve > 80% power using NCP calculations. Analytical stage: if shown to be underpowered, identify limiting factors, judiciously collect additional data.
Disease prevalence	*Determines number of cases in the sample.* Common diseases, random sampling, large samples. Rare diseases, selective sampling, smaller samples.		
Residual correlation	*NCP proportional to the squared residual shared variance.* Increased residual correlation (e.g., DZ) ⇒ smaller samples.		
Incomplete linkage	*NCP proportional to* $(1 - 2\theta)^4$. Trait locus distant from marker locus ⇒ lower power ⇒larger samples.		
Pedigree size	*NCP proportional to pedigree size and informativeness.* Increased power with larger sibships and/or multi-generational data		
3. Deviations in trait distribution	*Violates normality assumption.* Reduces power of HE and VC, increases type-I error of VC.	Examine skewness, kurtosis, residuals, multivariate normality. Compare sample mean, variance and correlation with population parameters.	Transformation (e.g., Box-Cox) or use robust test.

Continued

Table 13.2 Summary of the impact of different factors on the type-I error and power of linkage analysis—cont'd

Factor	Impact on linkage analysis	Detection	Correction
4. Outliers	*Influence normality, shift sample moments.* Increase type I error, reduce power.	Identify both univariate and multivariate outliers. Examine residual plots and individual family likelihoods.	If applicable, replace with correct value. Winsorization or robust test. Investigate outlier families.
5. Pedigree errors	*May bias allele segregation patterns.* Increased type-I error, decreased power	Graphical inspection of relationships based on allele sharing (GRR).	Recode misclassified relatives.
6. Genotyping errors	*May bias allele segregation patterns.* Reduces power of ASP studies. Increase type-I error of DSP. Reduces power of VC.	Identify genotypes inconsistent with Mendelian inheritance (PedCheck, Sibmed, SIBPAIR). Identify genotypes consistent with unlikely recombination events (MERLIN).	If applicable, rescore or retype problematic individuals. Exclude problematic genotypes. Alternatively, model genotyping error in IBD estimation.
7. Marker informativeness	*Accuracy of IBD estimation.* Low marker information ⇒ lower accuracy ⇒ lower power.	Calculate H, PIC, MPIC or entropy index (GPC, MERLIN).	If reliable genetic map available, use multipoint IBD estimation. For unreliable marker distances, perform singlepoint analysis and compare.
8. Marker density	*Inheritance information extracted.* Sparse maps may not extract sufficient information for initial screen for linkage. Dense maps are costly. LD between markers may inflate multipoint tests.	Simulate/calculate MPIC (GPC) to choose/assess the power for a given marker set.	If reliable genetic map available, use multipoint IBD estimation. Model LD in multipoint tests, if required.
9. Genetic map	*Determines inter-marker distances.* Incorrect map increases type-I error and reduces power of multipoint analysis.	Identify regions with unreliable marker distances and/or different male/female recombination ratios.	Use updated map with genetic positions interpolated from physical positions. Compare multipoint with singlepoint analyses. In regions of large F:M distance ratio, perform analysis with both sex-averaged and sex-specific maps.

and correct genotypes appropriately (retype/rescore or delete problematic genotypes). Use software such as PedCheck or SIBPAIR to do this. Once Mendelian errors have been detected and cleaned, identify families which imply unlikely recombination events between markers and, again, correct accordingly (e.g., MERLIN). The final pedigree file should be cleaned to the best extent of any possible genotyping errors. Alternatively, instead of flagging and correcting/deleting problematic genotypes, we may chose to leave them in the dataset and use appropriate software that estimates IBD using an algorithm that models genotyping error (e.g., latest version of MERLIN). Based on the sample allele frequencies, calculate the PIC for each marker.

Fifth, investigate the power of the sample to detect a QTL under different parameter models. Based on the actual family number and size, and on previous biometric analysis (heritability and correlation calculations), derive the expected NCP for different QTL heritabilities, residual correlation and incomplete linkage. Use the GPC to do this.

Sixth, extract marker chromosomal locations from a genetic map based upon the most recent physical and linkage maps available. Flag markers that may have unreliable genetic positions or that may lie on regions of the genome where recombination rates differ significantly between males and females. Characterize the average map density.

Seventh, calculate a multipoint measure of marker informativeness, such as the MPIC or entropy-based index.

Eighth, based on the ascertainment scheme and on the properties of the phenotypic trait (continuous vs. discrete, normal vs. non-normal), choose the most robust and powerful linkage method. Perform multipoint analysis with sex-averaged maps as a first screen for linkage. Then, for loci which have ambiguous genetic locations, perform singlepoint analysis and compare the results. In the case of linked loci which lie in regions of the genome of high F:M distance ratio, perform multipoint analysis using sex-specific maps. Redo analysis with other robust linkage method(s) and plot the genome scan results for all methods. Identify markers with significant and consistent LOD scores. Superimpose marker informativeness plots to identify regions where IBD sharing information may not have been extracted satisfactorily.

Acknowledgments

This research was supported in part by the National Health and Medical Research Council of Australia Sidney Sax Fellowship 389927.

References

Abecasis, G.R., Burt, R.A., Hall, D., *et al.* (2004b) Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *Am. J. Hum. Genet.* **74**: 403–417.

Abecasis, G.R., Cherny, S.S. and Cardon, L.R. (2001a) The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* **9**: 130–134.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2001b) GRR: graphical representation of relationship errors. *Bioinformatics* **17**: 742–743.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002) MERLIN – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**: 97–101.

Abecasis, G., Cox, N., Daly, M.J., Kruglyak, L., Laird, N., Markianos, K., Patterson, N. (2004a) No bias in linkage analysis. *Am. J. Hum. Genet.* **75**: 722–723; author reply 723–727.

Abecasis, G.R. and Wigginton, J.E. (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77**: 754–767.

Allison, D.B., Heo, M., Schork, N.J., Wong, S.L. and Elston, R.C. (1998) Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Hum. Hered.* **48**: 97–107.

Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. and Blangero, J. (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* **65**: 531–544.

Almasy, L. and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**: 1198–1211.

Atwood, L.D. and Heard-Costa, N.L. (2003) Limits of fine-mapping a quantitative trait. *Genet. Epidemiol.* **24**: 99–106.

Barnholtz, J.S., de Andrade, M., Page, G.P., King, T.M., Peterson, L.E. and Amos, C.I. (1999) Assessing linkage of monoamine oxidase B in a genome-wide scan using a univariate variance components approach. *Genet. Epidemiol.* **17** Suppl 1: S49–S54.

Blangero, J., Williams, J.T. and Almasy, L. (2000) Robust LOD scores for variance component-based linkage analysis. *Genet. Epidemiol.* **19** Suppl 1: S8–S14.

Blangero, J., Williams, J.T. and Almasy, L. (2001) Variance component methods for detecting complex trait loci. *Adv. Genet.* **42**: 151–181.

Boehnke, M. and Cox, N.J. (1997) Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.* **61**: 423–429.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**: 314–331.

Box, G. and Cox, D. (1964) An analysis of transformations. *J. R. Stat. Soc.* **26**: 211–252.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**: 861–869.

Brzustowicz, L.M., Merette, C., Xie, X., Townsend, L., Gilliam, T.C. and Ott, J. (1993) Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am. J. Hum. Genet.* **53**: 1137–1145.

Cardon, L.R. and Fulker, D.W. (1994) The power of interval mapping of quantitative trait loci, using selected sib pairs. *Am. J. Hum. Genet.* **55**: 825–833.

Carey, G. and Williamson, J. (1991) Linkage analysis of quantitative traits: increased power by using selected samples. *Am. J. Hum. Genet.* **49**: 786–796.

Chen, W.M., Abecasis, G.R. (2006) Estimating the power of variance component linkage analysis in large pedigrees. *Genet. Epidemiol* **30**: 471–484.

Cherny, S.S., Abecasis, G.R., Cookson, W.O., Sham, P.C. and Cardon, L.R. (2001) The effect of genotype and pedigree error on linkage analysis: analysis of three asthma genome scans. *Genet. Epidemiol.* **21** Suppl 1: S117–S122.

Daw, E.W., Thompson, E.A. and Wijsman, E.M. (2000) Bias in multipoint linkage analysis arising from map misspecification. *Genet. Epidemiol.* **19**: 366–380.

de Andrade, M. and Amos, C.I. (2000) Ascertainment issues in variance components models. *Genet. Epidemiol.* **19**: 333–344.

de Andrade, M., Fridley, B., Boerwinkle, E. and Turner, S. (2003) Diagnostic tools in linkage analysis for quantitative traits. *Genet. Epidemiol.* **24**: 302–308.

Dolan, C.V., Boomsma, D.I. and Neale, M.C. (1999) A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. *Behav. Genet.* **29**: 163–170.

Douglas, J.A., Boehnke, M. and Lange, K. (2000) A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am. J. Hum. Genet.* **66**: 1287–1297.

Eaves, L., Meyer, J. (1994) Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* **24**: 443–455.

Ehm, M. and Wagner, M. (1998) A test statistic to detect errors in sib-pair relationships. *Am. J. Hum. Genet.* **62**: 181–188.

Epstein, M.P., Duren, W.L. and Boehnke, M. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* **67**: 1219–1231.

Epstein, M.P., Lin, X. and Boehnke, M. (2003) A tobit variance-component method for linkage analysis of censored trait data. *Am. J. Hum. Genet.* **72**: 611–620.

Evans, D.M., Medland, S.E. (2003) A note on including phenotypic information from monozygotic twins in variance components QTL linkage analysis. *Ann. Hum. Genet.* **67**: 613–617.

Feingold, E. (2001) Methods for linkage analysis of quantitative trait loci in humans. *Theor. Popul. Biol.* **60**: 167–180.

Feingold, E. (2002) Regression-based quantitative-trait-locus mapping in the 21st century. *Am. J. Hum. Genet.* **71**: 217–222.

Fernandez, J.R., Etzel, C., Beasley, T.M., Shete, S., Amos, C.I. and Allison, D.B. (2002) Improving the power of sib pair quantitative trait loci detection by phenotype winsorization. *Hum. Hered.* **53**: 59–67.

Ferreira, M.A.R., O'Gorman, L., Le Souef, P., Burton, P.R., Toelle, B.G., Robertson, C.F., Visscher, P.M., Martin, N.G. and Duffy, D.L. (2005) Robust estimation of experiment-wise P-values applied to a genome-scan of multiple asthma traits identifies a new region of significant linkage on chromosome 20q13. *Am. J. Hum. Genet.* **77**: 1075–1085.

Fisher, R. (1934) The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.* **6**: 13–25.

Forrest, W.F. (2001) Weighting improves the 'new Haseman-Elston' method. *Hum. Hered.* **52**: 47–54.

Forrest, W.F. and Feingold, E. (2000) Composite statistics for QTL mapping with moderately discordant sibling pairs. *Am. J. Hum. Genet.* **66**: 1642–1660.

Goring, H.H. and Ott, J. (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur. J. Hum. Genet.* **5**: 69–77.

Goring, H.H. and Terwilliger, J.D. (2000a) Linkage analysis in the presence of errors. II. Marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am. J. Hum. Genet.* **66**: 1107–1118.

Goring, H.H. and Terwilliger, J.D. (2000b) Linkage analysis in the presence of errors. III. Marker loci and their map as nuisance parameters. *Am. J. Hum. Genet.* **66**: 1298–1309.

Gu, C., Todorov, A. and Rao, D.C. (1996) Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genet. Epidemiol.* **13**: 513–533.

Guo, X. and Elston, R.C. (1999) Linkage information content of polymorphic genetic markers. *Hum. Hered.* **49**: 112–118.

Halpern, J. and Whittemore, A.S. (1999) Multipoint linkage analysis. A cautionary note. *Hum. Hered.* **49**: 194–196.

Haseman, J.K. and Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**: 3–19.

Hopper, J.L. and Mathews, J.D. (1982) Extensions to multivariate normal models for pedigree analysis. *Ann. Hum. Genet.* **46** (4): 373–383.

Kendall, M. and Stuart, A. (1979) *The Advanced Theory of Statistics*, Vol 2: *Inference and relationship.*, John Wiley & Sons, New York, NY.

Kong, A. and Cox, N.J. (1997) Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61**: 1179–1188.

Kong, A., Gudbjartsson, D.F., Sainz, J., *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241–247.

Kong, X., Murphy, K., Raj, T., He, C., White, P.S. and Matise, T.C. (2004) A combined linkage-physical map of the human genome. *Am. J. Hum. Genet.* **75**: 1143–1148.

Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* **17**: 21–24.

Kruglyak, L. and Daly, M.J. (1998) Linkage thresholds for two-stage genome scans. *Am. J. Hum. Genet.* **62**: 994–997.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**: 1347–1363.

Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**: 241–247.

Lathrop, G.M., Hooper, A.B., Huntsman, J.W. and Ward, R.H. (1983) Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. *Am. J. Hum. Genet.* **35**: 241–262.

Leal, S.M. (2003) Genetic maps of microsatellite and single-nucleotide polymorphism markers: are the distances accurate? *Genet. Epidemiol.* **24**: 243–252.

McPeek, M.S. and Sun, L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**: 1076–1094.

Neale, M.C., Eaves, L.J. and Kendler, K.S. (1994) The power of the classical twin study to resolve variation in threshold traits. *Behav. Genet.* **24**: 239–258.

Neale, M., Boker, S., Xie, G. and Maes, H. (2002) *Mx: Statistical Modeling.* VCU, Richmond, VA.

Nicolae, D.L., Kong, A. (2004) Measuring the relative information in allele-sharing linkage studies. *Biometrics* **60**: 368–375.

O'Connell, J.R. and Weeks, D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **63**: 259–266.

Olson, J.M. (1999) Relationship estimation by Markov-process models in a sib-pair linkage study. *Am. J. Hum. Genet.* **64**: 1464–1472.

Ott, J. (1991) *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, MD.

Ott, J. (1993) Detecting marker inconsistencies in human gene mapping. *Hum. Hered.* **43**: 25–30.

Palmer, L.J., Jacobs, K.B. and Elston, R.C. (2000) Haseman and Elston revisited: the effects of ascertainment and residual familial correlations on power to detect linkage. *Genet. Epidemiol.* **19**: 456–460.

Purcell, S., Cherny, S.S., Hewitt, J.K. and Sham, P.C. (2001) Optimal sibship selection for genotyping in quantitative trait locus linkage analysis. *Hum. Hered.* **52**: 1–13.

Purcell, S., Cherny, S.S. and Sham, P.C. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**: 149–150.

Rijsdijk, F.V. and Sham, P.C. (2002) Estimation of sib-pair IBD sharing and multipoint polymorphism information content by linear regression. *Behav. Genet.* **32**: 211–220.

Risch, N. and Zhang, H. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**: 1584–1589.

Risch, N.J. and Zhang, H. (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *Am. J. Hum. Genet.* **58**: 836–843.

Schork, N.J., Allison, D.B. and Thiel, B. (1996) Mixture distributions in human genetics research. *Stat. Methods Med. Res.* **5**: 155–178.

Sham, P. (1998) *Statistics in Human Genetics*. Arnold, London.

Sham, P.C. and Purcell, S. (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **68**: 1527–1532.

Sham, P.C., Cherny, S.S., Purcell, S. and Hewitt, J.K. (2000a) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**: 1616–1630.

Sham, P.C., Zhao, J.H., Cherny, S.S. and Hewitt, J.K. (2000b) Variance-components QTL linkage analysis of selected and non-normal samples: conditioning on trait values. *Genet. Epidemiol.* **19** Suppl 1: S22–S28.

Sham, P.C., Purcell, S., Cherny, S.S. and Abecasis, G.R. (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.* **71**: 238–253.

Stringham, H.M. and Boehnke, M. (1996) Identifying marker typing incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **59**: 946–950.

Stringham, H.M. and Boehnke, M. (2001) LOD scores for gene mapping in the presence of marker map uncertainty. *Genet. Epidemiol.* **21**: 31–39.

Terwilliger, J.D., Ding, Y. and Ott, J. (1992) On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics* **13**: 951–956.

Visscher, P.M. and Hopper, J.L. (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Ann. Hum. Genet.* **65**: 583–601.

Whittemore, A.S. and Halpern, J. (1994) A class of tests for linkage using affected pedigree members. *Biometrics* **50**: 118–127.

Williams, J.T. and Blangero, J. (1999) Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.* **63**: 545–563.

Williams, J.T., Blangero, J. (2004) Power of variance component linkage analysis-II. Discrete traits. *Ann. Hum. Genet.* **68**: 620–632.

Yu, A., Zhao, C., Fan, Y., *et al.* (2001) Comparison of human genetic and sequence-based physical maps. *Nature* **409**: 951–953.