Linkage mapping for complex traits : a regression-based approach

Lebrec, J.J.P.

Chapter 1

Introduction

Once the heritable character of a trait has been established, the strategies available for gene mapping may be split into two classes. In the first 'candidate gene' approach, prior biological knowledge is available about the function of one or several genes, the scientific question to be tested is whether this limited number of pre-identified genes influences the trait of interest. Subsequently, researchers are usually interested in quantifying those effects. Although the field of genetics offers some peculiarities, well known epidemiological methods are suited to answer this type of questions. The second 'positional mapping' approach requires, in principle, no prior biological knowledge but its purpose is perhaps less ambitious: it aims at identifying chromosomal regions which contain genes influencing a trait. As far as the search for genes is concerned, the first approach therefore is an hypotheses-testing exercise while the second approach generates hypotheses. linkage as well as association studies fall into the positional mapping category. The former relies on the biological process of recombination (see 1.1) and the latter on the presence of linkage disequilibrium (see also 1.1) in populations. In the traditional gene-mapping paradigm, positional mapping precedes candidate gene-mapping but the frontiers between the two categories are sometimes fuzzy. Indeed nowadays, association scans often attempt to combine the two steps together. This thesis only deals with issues related to linkage mapping.

1.1 Some basics in genetics

This section introduces some basic concepts of genetics that are a pre-requisite to the understanding of the problem of linkage.

A gene is defined as a sequence of desoxyribonucleic acid (DNA) that codes a protein; most of our DNA is non-coding. Despite this formal definition, the term gene

is often loosely used to refer to a piece of DNA or genetic material, whether coding or not. This imprecision in terminology is often a hurdle for statisticians willing to enter the realm of genetics. Nevertheless, I will adhere to this practice. The genetic material of human beings is stored in 23 pairs of chromosomes, 22 pairs of autosomes and 1 pair of sex chromosomes. The transmission of this material from parents to offspring occurs independently at each chromosome: each parent contributes one copy of his/her two genes at random to an offspring via their gametes, this is known as the law of segregation or Mendel's first law. Parents, however, rarely transmit an entire copy of one of their two chromosomes (termed grand-paternal and grand-maternal). Instead, their transmitted chromosome is made up of alternating segments from the grand-paternal and grand-maternal chromosomes. This exchange of genes between the grand-paternal and grand-maternal chromosomes occurs during the formation of gametes or meiosis at points called crossovers, as a result chromosomes in gametes and resulting offspring are made up of recombinant chromosomes (see Fig.1).

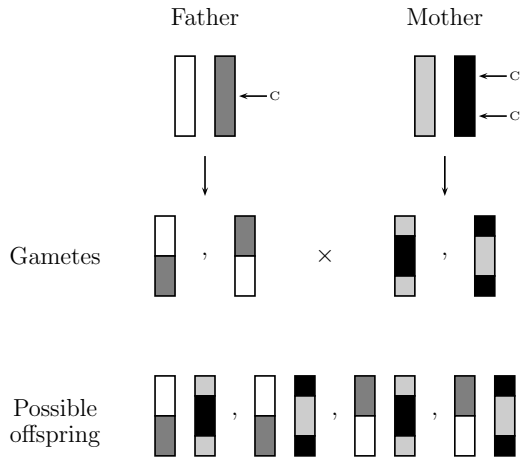

Figure 1.1: Chromosomes in gametes and offspring after recombinations - C indicates a crossover event

This recombination process ensures genetic diversity, it is also the phenomenon that makes linkage analysis possible because it introduces variation in genetic similarity between relatives across one single chromosome. A recombination event between two chromosomal positions or loci is equivalent to an odd number of crossovers

between those two loci in one meiosis, this happens at a certain rate called the recombination fraction θ. The recombination fraction increases with physical distance, however the relation between the two varies across the genome. If two loci are close together on the same chromosome, they are said to be linked; if they are very far apart, on the same chromosome or on different chromosomes, they are unlinked and the law of segregation implies that $\theta = 0.5$. The genetic distance d_{AB} (unit=Morgan) between two loci A and B is defined as the average number of crossovers between them per meiosis, by linearity of the expectation $d_{\mathrm{AC}} = d_{\mathrm{AB}} + d_{\mathrm{BC}}$ (if B lies between A and C). This additive property of the genetic distance scale is extremely convenient but obviously does not apply to recombination fractions although this is the probabilistic quantity needed for computations in linkage testing. Mapping functions that convert recombination fraction θ into genetic distance m, or conversely, are therefore available. One slightly simplistic but practically important such function is given by Haldane's function $\theta = \frac{1}{2}(1 - \mathrm{e}^{-2m})$ which is obtained by assuming that the number of crossovers between two loci follows a Poisson distribution with mean proportional to the genetic distance between loci.

Since the genetic similarity between relatives extends over relatively large chromosomal segments, it would be far too costly and inefficient to sequence the whole genome of each individual. Geneticists have identified DNA polymorphisms (so called markers) which can be seen as genes (in the loose sense) whose alleles (the different forms that a gene can take) can easily be identified by modern molecular biology techniques. It must be stressed that this technology can only determine the unordered pair of alleles (or genotype) at each marker for the two paired chromosomes of an individual. Classically, a few hundreds highly polymorphic genetic markers known as micro-satellites are scattered more or less evenly across all chromosomes. Since they have many and therefore relatively rare alleles, those markers allow one to tell whether relatives share the same genes at that location with little uncertainty. Those markers are usually taken in non-coding regions of the genome and are therefore believed, due to lack of selective pressure, to be neither related with each other nor with the potentially causing genes, in the overall population. In genetic jargon, the markers are said to be in linkage equilibrium with each other and with the genes [1]. Another

[1] In statistical terms, considering the one-allele genotypes of gametes at different loci as random

type of (bi-allelic) markers known as single nucleotide polymorphisms (SNP) is now routinely used in gene-association studies, these markers are more densely available across the genome and they can be cheaply typed in chips called SNP-arrays. They are now being used in linkage analysis too although their use is more problematic due to linkage disequilibrium between them. Despite the intensive computations involved in their use in linkage analysis, they offer the promise of a cheap and evenly distributed linkage information map across the genome.

1.2 Overview of linkage methods

The first traits to be mapped by linkage methods were Mendelian i.e. they were rare and determined in an almost one-to-one relation by the genotype at a single location. With such strong genetic effects, the actual mode of inheritance (i.e. genetic model) was fairly well known via segregation analysis (which only requires phenotypic data in families). This type of traits lent itself very well to the so-called parametric linkage methods. In its simplest version, this methodology postulates a genetic model for the trait values Y given the genotype at the causing locus with genotype G via a penetrance function $\mathbf{P}(Y \mid G)$. The likelihood $L(M \mid Y; \theta)$ of the data at a marker M given the recombination fraction θ between marker M and true locus can be computed and the corresponding likelihood ratio test $\sup_\theta \frac{L(M \mid Y; \theta)}{L(M \mid Y \ ; \ \theta=0.5)}$ provides a test for linkage.

This model for linkage was appealing for Mendelian traits and did yield an unprecedented harvest of genes for those rare diseases but it is much less suited for the analysis of complex traits. The methodological emphasis has long switched to biometrical models and to the so-called non-parametric linkage methods. This other branch of methods is essentially based on identifying chromosomal regions where phenotypic similarity coincides with genotypic similarity. The concept of identity-by-descent (IBD) formalizes the idea of genetic similarity between relatives: two genes are said to be IBD if they are copies of the same ancestral gene. The IBD configuration at different loci in a pedigree is not observable directly but it can be conceived of as

variables (a haplotype is a possible value of the resulting multivariate random variable), two loci are said to be in linkage equilibrium if the genotypes at those two loci are independently distributed, if not they are said to be in linkage disequilibrium

a hidden Markov process whose transition probabilities depend upon the recombination fractions [Lander and Botstein, 1989] between loci. The observations at the markers are used to calculate the IBD distribution at any arbitrary position on the chromosome [Kruglyak et al., 1996; Abecasis et al., 2002].

Continuous traits

For a quantitative trait, a Gaussian distribution naturally arises from the view that many factors, whether environmental or genetic, with equally small individual effects contribute to the trait. By further assuming a random mating population, one obtains the so-called variance components model [Lange et al., 1976; Amos, 1994; Almasy and Blangero, 1998]. In a simple additive version of the model, the total trait variance is decomposed into three sources: familial or common environment, additive genetic and measurement error or unique environment. The covariance of two relatives turns out to be the sum of the common environment variance and the additive genetic variance times a kinship coefficient which is proportional to the average proportion of genes that the relatives share. The model is often used in heritability and segregation analysis where the purpose is to establish the genetic character of a trait and to further characterize its mode of inheritance. Monozygotic twins have the same genes while dizygotic twins share only half of them but the degree to which the environment is shared by individuals in the two types of twinships is identical. Twin studies therefore provide a simple design for testing for a purely genetic component.

If IBD was measured exactly at a causative additive gene, the covariance for two relatives in the variance components model would include a term equal to the product of kinship coefficient by the gene attributable variance σ_q^2 times the IBD sharing. The test for linkage at any putative position is therefore based on rejecting the null hypothesis that $\sigma_q^2 = 0$ in favor of the alternative $\sigma_q^2 > 0$. In unselected families, this is traditionally done using a likelihood ratio test statistic. In practice, IBD is measured at locations nearby the causing gene(s) and the estimated attributable variance will be a deteriorated version of σ_q^2, nevertheless the test statistic will tend to be maximal at positions closest to the true gene location. The popularity of the variance components model in quantitative trait locus (QTL) mapping is undoubtedly due to its extreme flexibility: variance components corresponding to non-additive

(dominant) gene effects, gene-gene interactions, gene by covariate interactions can be accommodated, the model mean can be corrected for important covariate effects, multivariate phenotypes can be conjunctly analyzed, the method can be adapted for analysis of the sex-chromosomes [Ekstrøm, 2004] and mixtures of variance components models can be used to face the problem of locus heterogeneity (see 1.3) [Ekstrøm and Dalgaard, 2003]; these extensions are only hindered by the computations required for fitting the corresponding models.

The much less computationally greedy regression-based methods for linkage analysis stem back to the work of Haseman and Elston [1972] who proposed to regress the squared difference in phenotypic values of siblings on their IBD sharing. In 30 years, many variations have appeared on the theme and they are all based on the regression of some form of phenotypic similarity statistic on the IBD sharing. It is only recently that light has been shed on the relation between Haseman-Elston regressions and the score test of the linkage parameter $\sigma_q^2 = 0$ in the variance components model [Tang and Siegmund, 2001; Putter et al., 2002; Wang and Huang, 2002a]: some optimal form of Haseman-Elston regression happens to coincide with such a score test in an additive variance components model for sibling pairs. The conceptualization of those regression methods as score tests in the flexible variance components model frameworks has opened the way to fruitful generalizations of the regression-based methods e.g. to arbitrary pedigrees. In addition to their light computational burden, regression-based or score test based methods are appealing because of their potential robustness (in terms of false positive rate) to normality and to outliers. Finally, by inverting the regression i.e. IBD is regressed on a function of phenotypic similarity, the method can in principle be used to make valid inference in families sampled using their trait values [Sham et al., 2002].

Qualitative traits

For qualitative traits, which for linkage studies is almost synonymous of binary traits (i.e. disease in the medical field), non-parametric testing for linkage is usually done by comparing the average observed IBD sharing with its expected value under the assumption of no linkage. In designs where only one type of independent relative pairs is collected (e.g. affected sib-pair designs, ASP), this test based on deviation of

IBD sharing uses 1 degree of freedom (df), while a totally model-free ASP analysis necessitates a 2-df test [Risch, 1990]. Although the recognition of constraints for the parameters reduces the space of alternatives [Holmans, 1993], the higher level of significance required for the 2-df test often annihilates the gain in non-centrality parameter and the 1-df test appears to be a good testing strategy for a wide range of genetic models. Different types of independent relative pairs (e.g. affected sib pairs, discordant sib pairs, affected cousins) can be combined by using a weighted average of the excess IBD sharing of each kind; whatever the weights, provided markers segregate in a Mendelian fashion, the test will have adequate type I error, however its optimality will depend on how close the chosen relative weights are from the true relative excesses in IBD sharing at the causative locus [Teng and Siegmund, 1997].

Although less attractive than when disease inheritance is clearly Mendelian, larger families are sometimes sampled in linkage studies for complex traits. In that case, IBD-based tests can be generalized by the use of sensible scoring functions of the different IBD configurations in a pedigree [Whittemore and Halpern, 1994; Kong and Cox, 1997]. Alternatively, locally optimal tests based on the likelihood of the IBD configuration in each pedigree may be derived. The tests are pedigree-specific and only optimal if the true relative weights of the different parameters are known but sensible guesses provide decent efficiency across a wide range of genetic models [Teng and Siegmund, 1997]. As in the case of families consisting only of pairs of relatives, combining families of different types is a matter of assigning relative weights to the family-specific tests.

The incorporation of covariate information into disease linkage studies has been an active area of research in the past few years [Schaid et al., 2003]. The usual approach amounts to regressing the IBD sharing on the covariates of interest in a linear or non-linear fashion [Olson, 1999]. At least for categorical covariates, the approach can be made non-parametric at the cost of an increase in the number of parameters, however parsimonious models are needed in order to carry out efficient inference. Age is a crucial covariate to take into account in order to include unaffected individuals in a linkage study. Another way to approach the problem is to use the disease age of onset as the possibly censored endpoint.

Significance level

Since the position of the true locus is often completely ignored, the whole genome is scanned using a linkage statistic on a grid of chromosomal positions, this multiplicity of tests increases the false positive rate. The tests at neighboring positions are highly correlated so a Bonferroni correction of the α level of each test is too conservative. Asymptotic arguments based on the theory of Gaussian processes leads to approximate thresholds for the non-parametric methods statistics [Lander and Green, 1987; Feingold et al., 1993]. These thresholds rely on the Haldane's mapping function, they depend on the type of families studied (which determines the correlation structure of the process) and the degrees of freedom for the test; although they are derived under the idealized assumption of a dense map of completely informative markers, the thresholds seem to be only slightly conservative when applied to discrete evenly distributed maps of partially informative markers [Teng and Siegmund, 1998]. Due to a tradition dating back to the early days of parametric linkage [Morton, 1955], statistical significance of linkage tests is usually presented as a LOD score (originally a $\log_{10}$ of the odds that a locus is linked versus unlinked) which is obtained by dividing a $\chi^2_{[1]}$-distributed statistics by $2 \times \ln(10)$. In current practical situations of human sib-pair linkage studies, a LOD score of 3 or higher gives a rule of thumb for declaring that a 1-df statistics based on average IBD sharing is significant.

In practice, various types of families are often combined, marker information varies across the genome and the assumptions underlying the linkage model (eg. normality in variance components model) might not be fulfilled. Nowadays, researchers tend to base their assessment of significance on simulations. Given the 'experimental conditions' of a study (marker map characteristics, pedigree structures and patterns of genotype missingness), marker genotypes can be simulated under the null hypothesis of no linkage i.e. by simply obeying the rules of Mendelian segregation. In that way, provided the linkage statistic can be quickly computed, the null distribution of the statistic may be obtained at any point on the genome. This method, sometimes called gene-dropping, therefore yields point-wise empirical p-values. The number of times the statistic exceeds a certain threshold on a given chromosome can be counted (note that this entails the choice of a minimal distance for considering two consecu-

tive peaks as separate). By combining the corresponding independent p-values on all chromosomes, one can obtain a genomewide assessment of significance.

1.3 Issues in linkage mapping

Linkage analysis has been successful in the gene mapping of hundreds of mendelian diseases, however application of the same methodology in the search for genes responsible for complex traits has proved extremely disappointing. Most studies often provide only suggestive evidence for linkage, and when clearly significant, replication of the findings appears to be the exception rather than the rule.

Failure of the linkage approach to gene-mapping of complex traits is often attributed to locus heterogeneity i.e. the fact that the loci influencing a trait differ across families or groups of families [2]. This is indeed a problem likely to be more acute in linkage studies of complex traits where data from numerous small families are gathered as opposed to a small number of large families. A direct corollary of locus heterogeneity is that linkage studies are under-powered. In fact, due to the polygenic nature of complex traits, most studies probably lack the sample size to detect the inherent small gene effects.

One obvious way to tackle the problem of heterogeneity is to refine the definition of a phenotype by defining more homogeneous clinical subgroups, so instead of sampling breast cancer patients, geneticists successfully selected families with early-onset breast cancer. Researchers also try to select phenotypes that are likely to be more closely related to a biological mechanism than a broadly defined disease itself. For instance different plasma lipid levels can be measured in the search for genes involved in obesity.

One strategy for improving power is to resort to selective genotyping [Risch and Zhang, 1995] i.e. to only genotype families whose extreme phenotypic values promise to deliver high linkage information. Another natural route for solving the issue of power is by a sufficient increase of the sample size. Collaborative efforts such as the GenomEUtwin project (`http://www.genomeutwin.org/`) are being set up in order to gather sufficient data from different centers. This obviously calls for meta-

[2] Another type of heterogeneity called allelic heterogeneity refers to a situation where different allelic mutations at the same locus contribute to a phenotype, however, linkage analysis is immune to this type of heterogeneity

analytic methodologies routinely used in the field of clinical trials.

It is also felt that the models underlying the linkage methods are too simplistic, for instance, important covariates or interactions are often ignored. Although biologically plausible, incorporation of gene-gene interactions in models for linkage analysis is unlikely to yield substantial benefit [Tang and Siegmund, 2002; Purcell and Sham, 2004]. Using covariate information appears to be a more promising path towards a refinement of the methods [Peng et al., 2005].

1.4 This thesis

This thesis presents some attempts to improve the current design and analysis of linkage studies for complex traits. The statistical methodology adopted is driven by the fact that genes involved in complex traits have small effects, it therefore seems legitimate to use score tests [Cox and Hinkley, 1974] because of their local optimality properties. In addition, score tests often give rise to tractable expressions, in the context of linkage these can be meaningfully interpreted in terms of regressions and quickly computed which is a crucial feature in genetics.

Chapter 2 deals mainly with the analysis of quantitative traits in families that have been selected based on their trait values. We derive a general score test for linkage in arbitrary pedigrees which is based on the likelihood conditional on the phenotypic values. Although the derivation of the test relies on the normally distributed variance components model, its size is robust to deviations from normality. Under local alternatives and assuming the variance components model correctly specifies the distribution of the phenotype, the test has some optimality properties. In addition, the value of the test's Fisher information provides an indication of the informativeness of each family and can be used as a criterion for genotyping selection. The test is adapted to the case of binary data via a liability threshold model.

Chapter 3 advocates the use of selected families in the mapping of complex traits using twins. The methodology relies on the informativeness criterion derived in chapter 2, but we quantify the potential gains obtained using a series of examples of quantitative and qualitative phenotypes that are relevant to the GenomEUtwin project.

Chapter 4 addresses the issue of genotyping error in linkage analysis. We first

study analytically the impact of genotyping error on linkage and provide formula for the bias incurred. These results provide insights into some empirical findings, in particular, we are able to explain the differences in impact of genotyping error in random and selected designs. Finally, we suggest a robust modification of the usual linkage test based on a genomic control of the excess IBD sharing, it provides robustness against genotyping error as well as against other processes whose effect is to distort the expected value of the IBD sharing.

Chapter 5 is concerned with the (in)validity of a range of standard methods when marker information is incomplete, in particular circumstances where the generalized estimating equations method for gene localization [Liang et al., 2001] fails are identified.

Chapter 6 transfers standard meta-analytic techniques to the field of QTL mapping. The field has some specificities that can be accommodated, in particular, the problem of genetic locus heterogeneity is looked at carefully. In absence of covariate observations at the individual level and under a homogeneous model, the meta-analytic approach is asymptotically equivalent to an analysis of a pooled data set but it is logistically much easier to carry out.

Finally, in chapter 7, we develop an approximate score test for linkage in the rich class of generalized linear models. It is based on a pseudo-likelihood of the data and although unlikely to be optimal in all situations, the test has the advantage of being tractable and to have a robust type I error. It provides a simple way to incorporate known covariate effects into linkage analysis and is applicable to arbitrary pedigrees.

The last chapter is a conclusion where I draw a perspective of the role of linkage in gene mapping.