Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 1

Introduction

1.1 Introduction

The focus of this thesis is on statistical methods for complex genetic traits. A genetic trait is complex when genetic and environmental factors are involved. Researchers seek to identify genetic factors causing such traits using family as well as population-based approaches. Since many genes are likely to be involved in the etiology of a complex trait, the contribution of each single locus can be small and therefore, such genes are difficult to detect. Many existing statistical methods have been devised to study simple Mendelian disorders and so they are not suitable for studying complex genetic traits. Therefore, studying such traits necessitates the development of new statistical methods. A strategy for studying complex genetic traits is to perform linkage analysis to identify regions, and then carry out association analysis to verify whether a candidate gene is involved.

In this chapter we briefly describe and discuss several statistical methods and analyses used in genetic studies. For a reference book of statistical methods in human genetics we refer to Elston et al. (2002). Recently Forabosco et al. (2005) provided a valuable review of methods for linkage analysis and association studies. Throughout this chapter we also refer to more subject-specific reviews. Finally we close this chapter with an outline of the content of the next chapters.

1.2 Familial Correlation

In genetic epidemiology, researchers aim to identify genetic factors involved in the etiology of traits. Any evidence of correlation between phenotypes and genotypes is suggestive of the presence of such genetic factors. However, before genetic markers are typed, investigators should assure that the trait clusters within families. Closely related individuals tend to have similar phenotypes compared with non related individuals or distant relatives. Although

the presence of such a clustering may be due to shared environmental factors rather than shared genetic factors, the absence of familial clustering implies that it is of no benefit to continue the study. Its presence after adjusting for environmental factors implies that genetic factors may play a role in producing a predisposition to a trait.

Many methods have been proposed in the literature for the analysis of familial correlation. Liang and Beaty (2000) discussed methods of analysis of aggregation for the family case-control design. For binary traits, familial aggregation is often measured by odds ratios using logistic regression models that allow for dependence between family members (Bonney, 1986; Liang and Beaty, 1991). Generalized estimating equations (GEE)(Liang and Zeger, 1986) are often used to address the dependence between family members, when estimating the parameters. For more references about the GEE approach to deal with familial correlation for binary traits see FitzGerald and Knuiman (2000) or Tregouet and Tiret (2000). For quantitative traits, familial correlation is often studied by fitting multivariate normal models to the trait values of family members (Beaty and Liang, 1987; Rao and Wette, 1987). The covariances between relatives depend on their degrees of relationship. For randomly selected families, Houwing-Duistermaat et al. (1995) used a generalized linear mixed model (McCullagh and Nelder, 1989), and derived the score statistic to test for familial clustering within relatives based on their degrees of relationship. The score statistic does not assume any distribution of random effects. However, assessing familial aggregation in non-random pedigrees requires accounting for the sampling scheme.

Ascertainment

Often families are selected based on the phenotype of one or more family members. Family members who caused the family to be entered in a study are referred to as probands. For binary traits, a proband may be an affected subject with a disease. For quantitative traits, a proband may have a phenotype that exceeds a certain threshold. Many studies (Cardon and Fulker, 1994; Carey and Williamson, 1991; Risch and Zhang, 1995; Zhang and Risch, 1996) have shown that the power substantially increases when selected samples are used.

There are different ascertainment schemes e.g. complete ascertainment, single ascertainment and quadratic ascertainment. In complete ascertainment, all families with at least one family member with the required phenotype are equally probable to enter the study. In single ascertainment and quadratic ascertainment families enter the study with probabilities proportional to the

number and the square of the number of family members having the required phenotype, respectively.

When families are selected through probands, the statistical methods used should take into account the ascertainment procedure (Elston and Sobel, 1979; Morton, 1959). For single ascertainment, many authors have proposed to condition on the observed phenotypes of the probands (Beaty and Liang, 1987; Cannings and Thompson, 1977; Hopper and Mathews, 1982). For quantitative traits, Elston and Sobel (1979) and Rao and Wette (1987) proposed to condition on the event that the phenotype of the proband exceeds a pre-specified threshold value. Rao et al. (1988) and de Andrade and Amos (2000) compared the two ways of correcting for ascertainment in the context of the variance components approach. Both studies concluded that the former adjustment for ascertainment is less efficient than the latter provided that the threshold is well known. Whereas conditioning on exceeding the threshold is less efficient if the threshold is not known.

However if single ascertainment cannot be assumed, conditioning on affected probands may not be sufficient. Ewens and Shute (1986) showed that if remaining family members are related to ascertainment, the estimates of genetic parameters may be biased. They proposed to split the family in a set of probands related to ascertainment and a set of relatives not related to ascertainment.

1.3 Linkage analysis

Once familial aggregation is established, researchers pursue the study by gathering genetic materials of family members relevant for the study. Then genome wide linkage analysis is performed to identify regions that may contain susceptibility genes. Methods for linkage analysis rely on the biological phenomenon of recombination. During meiosis, recombination occurs when homologous chromosome pairs exchange genetic material. The probability of a recombination event occurring between loci increases with the physical distance between them. Hence, alleles at close loci are more likely to be transmitted jointly to descendants than alleles at distant loci. Moreover, relatives who have similar phenotypes are expected to have inherited the same genetic materials in the vicinity of the genes that predispose to those phenotypes.

Linkage analysis has been divided into classes: parametric and non-parametric methods. Parametric linkage analysis methods are based on the analysis of the recombination between the unobserved disease locus and observed genetic markers along the human genome. They require specification

of genetic parameters describing the mode of trait inheritance, such as penetrance and disease-allele frequency. When genetic parameters are correctly specified parametric linkage analysis is the most powerful. Very distant loci are expected to recombine with probability $\theta = 0.5$, while close (linked) loci recombine with probability $\theta < 0.5$. To estimate the probability of a recombination, and test the null hypothesis of no linkage, between the unknown predisposing gene locus and a marker locus the LOD score method is usually used. The LOD score is defined as follows

$$Z = \max_{0 \leq \theta \leq 0.5} log_{10}[\frac{L(\theta)}{L(\theta = 0.5)}], \tag{1.1}$$

where $L(\theta)$ is the likelihood of the observed data given assumed parameter values. A test is significant if $Z \geq 3$ (see Lander and Kruglyak (1995) for details). For a comprehensive introduction to parametric linkage analysis we refer to Ott (1999).

Although parametric linkage methods have been successful in localizing genes responsible for simple Mendelian diseases, these methods have achieved only limited success in identifying genes predisposing to complex traits (Risch, 2000). Many complex genetic traits are controlled by multiple genetic factors. For such traits the mode of inheritance is often unknown. In contrast to parametric linkage methods, non-parametric linkage methods do not require the specification of mode of inheritance. Non-parametric methods rely only on the information of sharing alleles identical by descent (IBD) between relatives at a given locus to study whether this locus is genetically linked to the unobserved disease locus. Two alleles are IBD if they are both physical copies of the same ancestral gene (Lange, 2002). Relatives may share 0, 1 or 2 allele IBD at any given locus. According to the Mendelian law of inheritance, the probabilities π_0, π_1 and π_2 of sharing 0, 1 and 2 alleles IBD under random segregation, can be calculated for any two relatives. For example, a pair of siblings shares 0, 1 and 2 alleles IBD with probabilities $\pi_0 = 0.25$, $\pi_1 = 0.5$ and $\pi_2 = 0.25$. Linkage between a disease locus and marker genotypes can be studied by comparing the observed numbers of alleles shared IBD by affected relative pairs to the expected number of alleles IBD under random segregation. An increase in the number of alleles IBD indicates the presence of a susceptibility gene in the region.

A widely used design for binary traits is the affected sibling pair design. To test for linkage between a disease locus and a marker locus, Day and Simons (1976) proposed the proportion test, which compares the observed proportion

of sib pairs that share two IBDs with $\pi_2 = 0.25$, As an alternative, Green and Woodrow (1977) proposed the mean test, which compares the observed mean IBD with its null value of 0.5. Risch (1990b,c) proposed a likelihood ratio statistic that compares the likelihood of observing 0, 1 and 2 alleles IBD with the likelihood under random segregation. For a valuable review about allele sharing based test statistics for affected relative pairs and general pedigrees we refer to Shih and Whittemore (2001).

Since many marker loci should be tested, multiple testing problems arise. A classical method to adjust for multiple testing is Bonferroni correction. However, this method does not take into account the dependency between marker loci. Hence, it yields conservative p-values. Lander and Kruglyak (1995) proposed to reject the null hypothesis of no linkage for loci with p-values smaller than 0.0001.

For a large sample of affected relative pairs, Feingold et al. (1993) used a Gaussian approximation to the IBD process to test for linkage using all marker loci jointly. The authors assumed (1) the presence of at most a single disease locus on any chromomsome, (2) the Haldane's mapping function, and (3) the observed markers are dense and fully informative about the IBD status. They modelled the excess of IBD sharing at the disease locus with a parameter α. For affected sib pairs, α is equal to $\frac{\lambda_s - 1}{\lambda_s}$, with λ_s the relative risk of a sibling of an affected subject (Risch, 1990a). Since the location of the disease gene is unknown, they proposed to use the maximum of the mean IBD sharing over marker loci as the test statistic for linkage. Further, they used a Gaussian approximation based on the central limit theorem to derive its null distribution.

Recently Liang et al. (2001) introduced a GEE approach for affected sib pairs to estimate the location of a single disease gene in a candidate region previously identified by other approaches. The method also uses the IBD information on all markers simultaneously by incorporating the correlation between them. Schaid et al. (2005) extended the GEE approach to other relative pairs, and Biernacka et al. (2005) derived a similar GEE approach when two disease loci exist on the same chromosomal region of interest.

For quantitative traits, many methods have been proposed in the literature. A famous method for modeling excess IBD sharing in relative pairs is the Haseman-Elston (HE) approach (Haseman and Elston (1972), Amos and Elston (1989)). The method regresses the squared trait difference of pair members on their estimated proportion of IBD sharing. Since the square difference does not capture all the information on the linkage (Wright, 1997), Elston (2000) reconsidered the HE method by regressing the product of the trait dif-

ference and trait sum on the estimated proportion of IBD sharing. Another linkage analysis method for quantitative trait-locus is the variance components (VC) approach (Amos, 1994). VC methods fit multivariate normal models to the trait values of family members, and model the covariance between relatives conditional on their IBD sharing. For small effect size the VC approach is equivalent to the HE approach (Putter et al., 2002). A typical VC model is described as follows. Let y_k be a vector of the trait values for the kth family. The variable y_k can be written as

$$y_k = \mu + X_k\beta + q_k + g_k + e_k$$

where μ is the overall mean; β is a vector of the regression coefficients for the covariates; X_k is a design matrix of the covariate values; q_k is a vector of random genetic effects of the locus with $q_k \sim N(0, \sigma_q^2 \hat{\Pi}_k)$; g_k is a vector of random effects representing the sum of residual genetic effects with $g_k \sim N(0, \sigma_g^2 R_k)$; e_k is a vector of uncorrelated variables representing residual environmental factors, with $e_k \sim N(0, \sigma_e^2 I_k)$. The covariance of y_k is

$$COV(y_k|\hat{\Pi}_k, R_k) = \sigma_q^2 \hat{\Pi}_k + \sigma_g^2 R_k + \sigma_e^2 I_k$$

where R_k is the matrix of degree of relationship for the family; $\hat{\Pi}_k$ is the matrix of the estimated proportion of IBD shared by relative pairs in the family; I_k is an identity matrix. Testing for linkage is equivalent to testing the null hypothesis $H_0 : \sigma_q^2 = 0$ versus $H_1 : \sigma_q^2 > 0$. For reviews on linkage methods for quantitative traits see Feingold (2001) and Amos and de Andrade (2001).

1.4 Association studies

Genetic association studies generally aim to narrow candidate regions, which may be chosen because on the basis of their known biological function or because they were initially identified by linkage analysis. Genetic association analysis compares allele or genotype frequencies at markers in candidate regions in affected individuals with those in unaffected individuals. An allele associated with a disease should be over-represented in affected individuals. A marker may be associated with the disease because it is a disease locus, or because one of its alleles is in linkage disequilibrium (LD) with a causal variant at a disease locus. Let A and B be alleles at two distinct loci and let AB be the corresponding haplotype. A and B are said to be in LD if

$$Pr(AB) \neq Pr(A)Pr(B),$$

with $Pr(A)$, $Pr(B)$ and $Pr(AB)$ the probability of the occurrence of A, B, and AB in the population, respectively. LD suggests that two loci may be very close to one another.

Disease-marker association may also be due to population admixture. This may occur if allele frequencies differ among subpopulation groups. To eliminate false evidence for association due to population admixture, family-based designs are used. A simple and common family-based design is the case-parent trio design, which compares the frequencies of alleles transmitted to affected individuals with those non-transmitted. For a review of family-based designs and corresponding statistical methods we refer to Zhao (2000). In this thesis we do not consider further this type of design.

A classical design to perform association study is the case-control design. Unrelated cases (affected individuals) and unrelated controls (unaffected individuals) are ascertained for study. Let q_i and p_i be the frequency of the associated allele i at a marker in cases and controls respectively. Hastbacka et al. (1992) modelled the excess of the associated allele in cases by

$$q_i = p_i + \delta(1 - p_i),$$

with δ the fraction attributable at risk (Clayton, 2000). δ can serve as a measure of linkage disequilibrium. Devlin and Risch (1995) provided a detailed discussion about measures of linkage disequilibrium.

For di-allelic markers, testing for disease-marker association can be carried out using Pearson's χ^2 with one degree of freedom. For a review of linkage disequilibrium methods for di-allelic markers see Lazzeroni (2001). For multi-allelic markers, Pearson's χ^2 has degrees of freedom equal to the number of alleles minus one. When markers with many alleles are considered, sparse data may occur, making the asymptotic distribution of Pearson's χ^2 invalid. Despite the fact that Monte-Carlo simulation can be used to derive the empirical p-values, χ^2 has low power due to large degrees of freedom. As an alternative, the maximum of the chi-squared statistics of 2-by-2 tables, each of which compares one allele against the rest, can be used when at most one allele is associated (Ewens et al., 1992). Sham and Curtis (1995) proposed to use the maximum of chi-squared statistics corresponding to all possible 2-by-2 tables, comparing any combination of alleles against the rest.

An alternative to taking the maximum is to take the sum over all possibilities. When one allele is associated Terwilliger (1995) modelled the excess of the associated allele in cases by the parameter δ. Since it is unknown which marker allele is associated with the disease, the likelihood corresponding to

this model is a weighted sum over all alleles i of conditional likelihoods given that allele i is over-represented in the set of cases. As weights Terwilliger (1995) used the allele frequencies in the overall population. Testing for association can be carried out by comparing the likelihood of the data under the alternative hypothesis of the presence of one associated allele with the likelihood of data under the null hypothesis of no disease-marker association.

When one allele is associated, the maximising approach and the Terwilliger's likelihood approach perform much better than Pearson's χ^2, especially for markers with many alleles. However, if more than one allele is associated with the disease, they may have low power as they are designed to test the alternative hypothesis of the presence of one associated allele (Sham et al., 1996).

When a candidate region contains multiple markers, an alternative to studying single-marker association with a disease is haplotype-based association analysis. A haplotype is a combination of alleles at multiple linked markers inherited together. When haplotypes can be observed directly they can be viewed as variants of a multiallelic maker and then methods for multiallelic markers can be used to study genetic association. However haplotypes are often not observable and they should be inferred from genotype data. Many methods for inferring haplotypes and estimating their frequencies have been developed in the literature (see Niu (2004) for a review). However, when haplotypes can not be determined with certainty, the analysis should take this uncertainty into account. Further discussions of haplotype-based methods are given by Schaid (2004).

1.5 Scope of this thesis

This thesis aims to develop new statistical methods to study genetic backgrounds of complex traits. These methods are presented in a number of subsequent chapters. The order of chapters is motivated by the commonly adopted strategy to identify genes responsible for complex genetic traits, namely starting with aggregation analysis, followed by linkage analysis and accomplished by association studies.

Chapter 2 deals with the ascertainment issue when modelling familial aggregation. A new statistical tool is proposed for testing familial clustering of binary and quantitative traits when families are selected based on the phenotypes of the probands. Familial correlation is modelled using a generalized linear mixed model. To adjust for ascertainment we condition on the phenotypes of the individuals relevant for ascertainment. Further we illustrate

the methods using data on a sample of first-degree relatives of probands with type 2 diabetes mellitus.

Chapter 3 is concerned with linkage analysis. Inspired by Liang et al. (2001), we derived two global test statistics for linkage. The approach is based on allele IBD sharing and uses all markers simultaneously. Excess IBD sharing at disease locus is modelled by a parameter α (Feingold et al., 1993). The likelihood of data is the average of conditional likelihoods of data given that a marker locus is a disease locus. Either the likelihood ratio or score statistics can be used to test for linkage. Results of a simulation study of each method's performance are presented.

In Chapters 4, 5 and 6 new methods for analysis of disease association with multiallelic markers or haplotypes are described. Chapter 4 uses the semi-Bayesian likelihood approach proposed by Terwilliger (1995) to derive a score statistic for testing genetic association. The score test is simple to compute and enables us to derive empirical p-values by means of Monte-Carlo simulations. Further, we present the results of analytic as well as empirical comparisons of the performance of the score statistic and some existing statistics for multiallelic markers including Pearson χ^2. Chapter 5 presents an application of some of the methods described in chapter 4 to simulated data from Genetic Analysis Workshop 14 (Bailey-Wilson et al., 2005). The aim of this chapter is to illustrate how these methods perform when they are applied to candidate regions initially identified by means of linkage analysis. Chapter 6 describes a generalization of the method presented in chapter 4, its performance in comparison with Pearson's χ^2 and likelihood ratio proposed by Terwilliger (1995) and their application to data on thrombosis.

Chapter 7 is concerned with (1) the use of combining information from two sources (parents and teachers) when diagnosing children with attention deficit hyperactivity disorder (ADHD) in genetic studies, and (2) studying familial aggregation among three phenotypic subtypes of ADHD in an isolated population. A test statistic is described that compares the distribution of the kinship coefficient between two samples. The kinship coefficients of all pairs in each sample are obtained using the genealogical information of 22 generations.