Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 8

Summary

Complex traits are caused by multiple genetic and environmental factors, and are therefore difficult to study compared with simple Mendelian diseases. The modes of inheritance of Mendelian diseases are often known. Methods to dissect such diseases are well described in literature. For complex genetic traits, the inheritance pattern is not clear and difficult to understand, and genetic variants contributing to such traits probably have small effect sizes. Hence, searching for genes responsible for complex traits requires different strategies as well as new methods. A common strategy for mapping complex traits is as follows: (1) perform a genome-wide linkage analysis using dense genetic markers, and identify regions showing evidence of linkage, then (2) perform association analysis to refine these regions. Along these lines we propose several methods for detecting disease genes in this thesis.

In **chapter 1** we provide a general introduction to statistical tools for mapping genes responsible for complex genetic traits. Familial clustering, ascertainment issues, linkage analysis, and association analysis are considered. Furthermore, the aim and the outline of this thesis are described.

A first step in studying the genetic background of any trait is to verify whether it clusters within families. To this end, families ascertained through one or more members are often collected, especially when the trait is rare. Therefore, statistical methods that assume random families are not appropriate to this kind of data. In **chapter 2** we address this issue and develop a score statistic for testing familial clustering that takes into account the ascertainment scheme. The method does not assume any particular scheme, however, the set relevant for ascertainment (probands) should be known (Ewens and Shute, 1986). Familial correlation is modeled via a random effect parameter in the context of a generalized linear mixed model (Houwing-Duistermaat et al., 1995; le Cessie and van Houwelingen, 1995). The score statistics has the following features. It measures the proband-relative correlation as well as the relative-relative correlation. It allows for adjusting of covariates. No assumption about the distribution of the random effects is made. Furthermore, by

conditioning on the trait value of all individuals related to ascertainment the method is robust to the ascertainment scheme. The score test can also be used to test the presence of correlation that is partly due to excess sharing of alleles identical by descent (IBD) in a candidate region. The method is applied to a candidate region in families of probands with type 2 diabetes mellitus.

In **chapter 3** we develop an allele sharing method to test for linkage along the genome or in a candidate region. The method considers all genetic markers jointly and hence multiple testing problems are avoided. At the disease locus, the increase of the mean of the number of alleles shared identically by descent (IBD) is modeled as a function of the relative risk ratio (Risch, 1990a). At each marker locus, the average of mean IBDs is calculated over the number of affected sibling pairs and the resulting variables behave approximately as a Gaussian Markov process along each chromosome (Feingold et al., 1993). Furthermore, the method sums the conditional likelihood of data given that a certain marker is a disease locus, over the marker positions. Either the likelihood ratio or the score statistics can be used to test for global linkage in the region of interest. Both statistics have known asymptotic distributions. For a genome-wide scan, likelihood ratio should be used, while for candidate region the score test is appropriate since it is comparable to the likelihood ratio when the assumed model is true, and it may perform better when the gene effect size is very small or the model is incorrect. However, for genome-wide scans the sample size required for detecting positive linkage is very high, especially for small effect sizes.

In **chapter 4**, **5**, and **6** we propose new statistical tools for association analysis. For testing disease association with a single nucleotide polymorphism (SNP), standard methods such as Pearson's χ^2 can be used. However, standard methods may not be suitable when testing disease association with a multi-marker. As an alternative, Terwilliger (1995) proposed a likelihood ratio test (TLR). The likelihood of the data is the weighted sum of the conditional likelihoods given that an allele is associated with the disease over all marker alleles with weights equal to the allele frequencies. In these chapters we consider testing for genetic association between a disease and either multi-allelic markers or haplotypes constructed from several flanking SNPs. However, haplotypes must be constructed with certainty.

In **chapter 4** we propose the score test corresponding to the Terwilliger's likelihood ratio. Under the null hypothesis of no genetic association the expectation and variance of the score statistic and Pearson's χ^2 are approximately equal. Under the alternative hypothesis of the presence of one associated al-

lele, the score statistic has higher expectation than Pearson's χ^2 when the associated allele is common, which may imply higher power in favour of the score test. Furthermore we provide heuristic as well as empirical comparisons between the score test and other test statistics. We conclude that the score test has the highest power on average. Furthermore, we illustrated the methods based on a real data example.

In **chapter 5** we apply, the score test proposed in chapter 4, TLR and Pearson's χ^2 to candidate regions that initially showed evidence of linkage with Kofendrerd Personality Disorder (KPD), a fictional disorder. Data are from the GAW14 simulated micro-satellite data problem (Bailey-Wilson et al., 2005; Greenberg et al., 2005). All test statistics suggest the presence of association between KPD disease and the multi-allelic marker, D03S0127, in this region. Testing a dense map of SNPs reveal strong evidence of genetic association with SNPs B03T3056 and B03T3057, which are in linkage disequilibrium with a disease locus located at the end of chromosome 3 (Greenberg et al., 2005). The analyses are done without prior knowledge of the answers.

In **chapter 6**, we generalise the method proposed by Terwilliger (1995). Instead of allele frequencies we propose weights, which do not depend on actual data. In this light, we derive a new score statistic, which is easy to compute. The new score statistic has a power comparable to TLR statistic and sometimes even better, especially when the excess of associated allele is small and the allele frequency is relatively common. However, the score statistic may not have a reasonable power if the frequency of the associated allele is equal to the inverse of the number of marker alleles, or if there is one positively and one negatively associated allele with the same amount of allele excess. The score statistic is successfully applied to three candidate genes studied in the Leiden Thrombophilia Study (Uitte de Willige et al., 2005).

Chapter 7 describes a study of the relevance of adding information of a second informant when children are diagnosed with Attention Deficit Hyperactivity Disorder (ADHD). Especially, two points are addressed: (1) The usefulness of the use of two informants in genetic studies, and (2) whether phenotypic subtypes of ADHD cluster similarly in families. The study is conducted in a genetically isolated population in the Southwest of the Netherlands. Genealogical information is available for 22 generations. The study finds that patients that are diagnosed based on two informants are more closely related than those diagnosed based on one informant, which may be relevant for further study of the genetic background of ADHD. Moreover, the study confirms the familial clustering of ADHD, which was previously suggested by other

studies. Furthermore, patients with the inattentive subtype of ADHD are found to be more closely related. The study suggests that recessive genes may be involved in causing ADHD. In order to study the closeness of familial relationship between two patient groups we propose a test statistic, which compares the mean kinship coefficients of the two groups. For large sample size, the test statistic follows the normal distribution. However, in this study all patient groups are small and hence the asymptotic distribution of the test statistic cannot be used. Nevertheless the significance level could be obtained by means of Monte Carlo simulation.