1 QTL mapping in Human Populations

There has recently been considerable interest in mapping genes associated with certain quantitative traits. Given the general form for the likelihood for a pedigree, it is conceptually straightforward to allow for incomplete penetrance. While this method allows one to model quantitative traits, there are a number of other more specialized techniques that are often used to map quantitative traits. The primary motivation for these techniques is the difficulty of calculating the likelihood for a general pedigree. There is some evidence that one of these methods, the multipoint variance components method, is comparable in power to multipoint parametric linkage yet it is much easier to test for linkage using this method. A feature of methods that seek QTLs that is a drawback is that they look for a QTL in an interval of the genome. Hence they tend to identify large chromosomal regions that contain one or more QTLs, and so it is difficult to obtain precise location information about a QTL. This is still a very active research area, and there are methods for improving resolution.

1.1 Haseman-Elston regression

The Haseman-Elston regression method works with sib-pair data. For each sib-pair, we determine the proportion of alleles IBD at a given locus. Then we regress the squared difference between sibling's trait values on the proportion of alleles shared IBD. Alternatively, we may regress this squared difference on estimates of the proportion of alleles IBD averaging over a number of markers, or even over a chromosomal segment. The basic idea is that for a marker near a QTL, sibs with similar phenotypes should share many alleles IBD. Since all sibships are assumed independent, we can use the usual theory of least squares to obtain parameter estimates and conduct inference. Interpretation of these results requires a model for the effect of the QTL on the trait, hence we briefly discuss these models next.

When discussing quantitative traits, it is often useful to differentiate between additive genetic effects and dominant genetic effects. Consider a QTL with 2 alleles, A_1 and A_2. We suppose that the mean value of the quantitative trait given the genotype at the QTL is as given in the following table.

genotype	mean of phenotype
A_1 A_1	$\mu + \alpha$
A_1 A_2	$\mu + \delta$
A_2 A_2	$\mu - \alpha$

We call δ the *dominance effect* and α is the *additive effect*. These effects are random effects, i.e. they represent realizations of a random variable, hence we think of such effects with reference to their variances σ^2_α and σ^2_δ. If we assume there is no dominance effect then it can be shown that the expected value of the slope in the regression equation is $-2(1-2\theta)\sigma^2_\alpha$, where θ is the recombination fraction between the marker and the QTL. Hence if we reject the hypothesis that the slope is zero we conclude that $\theta \neq \frac{1}{2}$ and $\sigma^2_\alpha \neq 0$, i.e. the marker is near a QTL which impacts the phenotype. Note that a confidence interval for the regression slope can not be used to estimate the recombination fraction or the additive genetic variance. Despite this, there are methods that allow one to estimate each of these parameters if one needs estimates (for mapping for example).

1.2 Variance components models

In our considerations of extensions of the basic model of parametric linkage we very briefly outlined the use of variance components models for mapping QTLs: here we treat this topic in more depth. If we let y_i

denote the outcome variable for subject i, then we suppose

$$y_i = \mu + \alpha_i + \alpha_i^* + \epsilon_i,$$

where α_i represents the additive genetic contribution from the interval being examined and α_i^* represents the contribution from QTLs outside this interval. We treat α, α^* and ϵ as uncorrelated random variables, hence the phenotypic variance is the sum of the variance of these 3 random terms. The sum of the variances of α and α^*, $\sigma_\alpha^2 + \sigma_{\alpha^*}^2$ is called the *total genetic variance* and the ratio of the total genetic variance to the total variance, $\sigma_\alpha^2 + \sigma_{\alpha^*}^2 + \sigma_\epsilon^2$, is called the *heritability*. The size of the heritability is an indication of the extent to which a trait has genetic determinants.

We estimate these variance components by examing the phenotypic covariance of related individuals. We can express the phenotypic covariance between 2 individuals as

$$\mathrm{Cov}(y_i, y_j) = R_{ij}\sigma_\alpha^2 + 2\Theta_{ij}\sigma_{\alpha^*}^2,$$

where R_{ij} is the proportion of the chromosomal region under investigation shared IBD between individuals i and j and 2Θ is the coefficient of coancestry between the 2 individuals. Below we discuss these 2 terms further. Given the R_{ij}s and the Θ_{ij}s, if we let y denote the vector of measurements $y_1, \ldots, y_n$, then we assume y is multivariate normally distributed with some mean vector and covariance matrix depending on σ_α^2, hence we can express the likelihood of the data as a function of this variance parameter. Then we can use the usual maximized likelihood ratio test to test the hypothesis that $\sigma_\alpha^2 = 0$. If we reject this hypothesis then we conclude that there is a QTL in the chromosomal region of interest. These models can be extended to include dominance effects and effects of covariates. Moreover, the assumption of multivariate normality is inessential. There are other slightly different approaches to variance components models: this is still an area of active research.

1.2.1 Estimating IBD sharing in a chromosomal region

There are a number of methods for estimating R_{ij}, an unobservable random variable. A method due to Goldgar (1990) can be used to estimate this random variable, and then we just treat this estimate as the true value (i.e. we ignore uncertainty in the estimation of this variable). In detail, we assume the number of crossovers is a Poisson random variable and there is no interference. Then the location of crossovers on a chromosome has a uniform distribution given the total number of crossovers. Without loss of generality, suppose the length of the chromosomal segment under study is one. If there are k crossovers and the locations of these crossovers are denoted (in order along the chromosome) $x_1, \ldots, x_k$ then the proportion for which 2 sibs are not IBD, denoted D, is

$$D = \sum_{i=1}^{k/2} (x_{2i} - x_{2i-1})$$

if k is even and positive (k odd can be dealt with similarly). It transpires that D has a Beta distribution with parameters $k/2$ and $k/2+1$ (for k even), and so $R = 1 - D$ has a Beta distribution with parameters $k/2+1$ and $k/2$. We need the distribution of R conditional on IBD status at the markers assuming recombination in the interval or not, i.e.

$$\begin{aligned} \mathrm{E}(R|\text{IBD status and recombination status}) &= \sum_k \mathrm{E}(R|k \text{ crossovers}) \\ &\quad \times P(k \text{ crossovers}|\text{IBD status and recombination status}). \end{aligned}$$

First, E($R|k$ crossovers) is simple to evaluate since given k, R has a Beta distribution. We can then evaluate this sum using the Poisson assumption for the number of crossovers for a given IBD status and recombination status (see Goldgar (1990) for the details). The result will be just a function of the recombination fraction. Guo (1994) provides another approach to these calculations. We can then substitute the resulting value for R_{ij} in the expression for the phenotypic covariance.

1.2.2 Coancestry

The coefficient of coancestry is a measure of the relatedness of individuals. More properly, Θ_{ij} is the probability that a gene selected randomly from individual i and a gene randomly selected from the same autosomal locus from individual j are IBD. This probability depends solely on how related 2 individuals are: for example if i and j are sibs then $\Theta_{ij} = 1/4$. There are a number of algorithms for computing these values (see, for example, Lange (1997) pp. 70–80).

2 Gene mapping in practice

We have ignored several important practical aspects of localizing disease genes. In particular, we have not treated methods for establishing a disease has a genetic component, patient recruitment, genotyping, public databases of marker information, computing or laboratory methods for disease gene localization. All these topics are important for a real study. Haines and Pericak-Vance (1998) provides a good treatment of many of these aspects of a linkage or association study.

While most of the approaches we have discussed are based on a certain data structure (e.g. sib-pairs or trios with an affected child), often methods developed for a certain data structure are adapted to other situations. The most common example of this is the use of sib-pair methods for the analysis of pedigrees. In these cases, the researchers just decompose the pedigree into a number of sib-pairs. If we treat all these sib-pairs as independent (as in the Haseman-Elston approach) then we are overstating our confidence in our results since there will be dependence amongst the sib-pairs.

There are several journals that routinely publish articles related to linkage analysis and association analysis. While mainstream statistics journals do publish manuscripts in this area, and high profile research appears in the usual science journals (e.g. *PNAS* and *Nature*) most of the work appears in the following journals (these are in no particular order):

1. *American Journal of Human Genetics*
2. *Genetic Epidemiology*
3. *Human Genetics*
4. *Human Heredity*
5. *Behavior Genetics*

These journals publish a mix of applied results and theoretical topics. Looking at some of these journals will give you a good idea of what practical research in this area is like.

We have ignored computing in our discussion. Carrying out the computations involved in linkage analysis is complicated by the fact that there is no user friendly package that conducts a variety of statistical methods appropriate for linkage and association analysis. Instead, there are several programs in common use that are freely available on the internet. The text of Ott (see the syllabus) has some information on the programs that are available. The SAS institute is developing capabilities for linkage and association analysis, but they are not very far along yet.

References

Goldgar, D. (1990), “Multipoint analysis of human quantitative genetic variation”, *Am. J. Hum. Genet.*, 47, 957–967.

Guo, S., (1994), “Computation of identity-by-descent proportions shared by two siblings”, *Am. J. Hum. Genet.*, 54, 1104–1109.

Haines, J., and Pericak-Vance, ed.s (1998), *Approaches to Gene Mapping in Complex Human Diseases*, Wiley, New York.

Lange, K. (1997) *Mathematical and Statistical Methods for Genetic Analysis*, Springer, New York.

Exercises

Write a short summary (1 paragraph) of a paper from the March 2004 edition of the *American Journal of Human Genetics*. You should include the problem the researchers sought to address (i.e. what disease is studied or what technical problem does the method address), what sort of data is used (e.g. large pedigrees or unrelated individuals), what methods are used (i.e. association or linkage or both) and the conclusions. (This journal is available in the Biomed library, Deihl Hall, and on the internet-just search under the name of the journal.)