1 Production of gametes and data for genetic mapping

In order to understand the basis for genetic mapping of disease, we need to consider the events that take place during *meiosis*, the formation of sex cells, also referred to as *gametes*. Sex cells (i.e. egg and sperm cells), unlike other cells in the body (referred to as *somatic cells*) only carry one of each chromosome, either the mother's or father's. Meiosis starts when a cell carrying 2 copies of each gene duplicates all of the chromosomes in the genome. Once the chromosomes have been duplicated, the cell divides into 2 cells with 2 copies of each chromosome. Then these 2 cells both divide one more time, resulting in 4 cells with 1 copy of each chromosome. Each of these 4 cells has a mix of chromosomes from the 2 parental sources. When 2 sex cells fuse, the result then has a pair of chromosomes.

During meiosis, the chromosomes pair up and at certain random locations on the chromosomes contact one another. When the chromosomes touch (which is known as a *crossover*), they exchange segments of DNA, thereby creating a chromosome which is distinct from either of the progenitor chromosomes. This process is called *recombination*, and was first observed in the early twentieth century. The process of recombination is thought to be driven by chromosomal breakage: a break in one chromosome is repaired by using the other chromosome in the pair as a template, and this results in copying DNA from one chromosome to the other in a pair. Actually, since multiple crossovers are possible, not all crossovers will lead to 2 genes on opposite sides of one crossover being on distinct chromosomes, hence recombination is reserved for the cases where 2 alleles end up on different chromosomes than where they started. If 2 loci are close on the chromosome, it is unlikely they will be separated by a crossover since these occur randomly along the genome (although they do not occur uniformly over the genome). The *recombination rate* is the probability of a crossover occurring between 2 genes. This is measured in units called *Morgans* (after the scientist in whose lab crossing over was first observed). For genes on distinct chromosomes this is 0.5, but for genes on the same chromosome this is some quantity less than or equal to 0.5.

The data for genetic mapping in humans are pedigrees, genotypes and phenotypes. Since the underlying random process that gives rise to information regarding gene location is the only observed through matings (i.e. meiosis) we need data on related individuals. For experimental animals, we can design experiments to construct desired pedigrees, but this will not work for humans. Since we are primarily concerned with human genetics in this course, we need to work with the pedigrees we find. There are a number of conventions used in writing down a pedigree. Each individual is represented as either a circle (for females) or a square (for males). Horizontal lines between individuals indicate a mating, and vertical lines drawn down from these horizontal lines indicate offspring. Typically the diseased individuals are indicated by a darkened symbol and a deceased individual has a line drawn through his or her symbol. Monozygotic twins are represented as the lower vertices of a triangle. Numbers next to individuals represent genotypes, and if an individual's genotype is not measured but can be unambiguously identified, it is put in parentheses. If the phase of the genotype is known, a line is drawn between the markers on different chromosomes. Next, a parent who is homozygous for the marker can't tell you anything about linkage, hence the mates of homozygous affected parents are often excluded from the pedigree. Finally, individuals whose parents are not included in the pedigree are referred to as *founders*.

An important distinction from the perspective of computing the recombination rate is between *simple* and *complex* pedigrees. A simple pedigree has no closed loops, whereas a complex pedigree does. Complex pedigrees are quite common in animal populations (especially domesticated, breeding animals like horses) but are less common in human families (although there are cultures in which matings between first cousins are common, and this can give rise to complex pedigrees).

2 Some ideas from population genetics

An important notion from population genetics that we will need in the sequel is the idea of Hardy-Weinberg equilibrium. Put succinctly, genotype frequencies in the population depend only on gene frequencies. To make this statement rigorous requires a number of assumptions. In particular, we need to assume:

1. infinite population,
2. discrete generations,
3. random mating (also known as *no selection*),
4. no migration,
5. no mutation,
6. no association between the genotypes and sex.

That is a long list of assumptions, and they are likely to not hold for many traits of interest (and, of course, the first 2 are pure fiction). In any event, under these assumptions you reach Hardy-Weinberg equilibrium in one generation. To see this, suppose some gene has 2 alleles, A_1 and A_2, and these occur in the population of gametes with probabilities p_1 and p_2. The long list of assumptions given above implies random combination of the gametes, hence the genotypes occur with the following probabilities $P(A_1, A_1) = p_1^2$, $P(A_1, A_2) = 2p_1p_2$, and $P(A_2, A_2) = p_2^2$. Now consider the population of gametes in the next generation. We find $P(A_1) = p_1^2 + p_1p_2 = p_1$ and $P(A_2) = p_2^2 + p_1p_2 = p_2$. This argument easily generalizes to the case of more than 2 alleles. Just from the finiteness of human populations we expect the gene frequencies would change randomly over time, a phenomenon referreed to as *genetic drift*.

There is a set of techniques, referred to as *segregation analysis*, which aims to determine if a disorder is a recessive trait or a dominant trait. We will see that knowledge of the mode of inheritance is important when computing genetic distances, so this question is of considerable interest. If a trait is recessive, only 25% of offspring from heterozygous parents would be affected, while if the trait is dominant 50% of the offspring would be affected. For X-linked traits there are also distinctive patterns of segregation. Based on the expected ratios of affecteds amongst offspring, with experimental animals one can devise experiments to determine which mode of inheritance provides a better model for the data, but for human populations, the situation is more complex. Typically one must model the mechanism by which an affected individual is *ascertained* (i.e. made a member of the study population). Typically families are ascertained on the basis of a single affected individual (the so called *proband*). Since the results will depend on the ascertainment model, many researchers claim relying on the results of a segregation analysis is not helpful. Instead, one should just fit the recessive and the dominant model and let the data select the appropriate model (see, for example, Ott (1999)).

Association analysis is a term used to refer to methods that aim to establish if there is an association between a trait and a marker. The statistical methods used in this sort of analysis are usually familiar to a Biostatistician. For example, one may conduct a retrospective study, put the data in a contingency table and test for association. We will discuss association analysis in more detail later in the course.

3 The idea of linkage analysis

In radiation hybrid mapping, we were able to determine that 2 markers are close by observing how frequently they are on the same fragment of DNA after subjecting a genome to X-ray radiation. Linkage analysis is based on the same idea, except that recombination during meiosis takes the place of X-ray radiation, and we observe the phenotype of each individual. Since we observe the phenotype, we can try to deduce not only the location of a gene, but also its function. Assume each phenotype corresponds to a single allele for some gene. Under this assumption, by examining what markers are shared by individuals with the same phenotype, we can try to locate the gene responsible for the phenotype (since recombination will usually not separate a disease gene from markers nearby it on a chromosome).

4 Quality of genetic markers

As noted above, and observed in simple examples, if a parent is homozygous for a marker, we can not use this parent to help localize a disease gene. For this reason, when we construct genetic maps, we want to use markers which are highly polymorphic. This is the motivation for using micro-satellites. There have been a number of attempts to quantify the polymorphism of a marker. In practice, a marker may be highly polymorphic but may not be useful for other reasons, for example, the genotyping may be difficult due to problems with the genotyping technology for that marker (e.g. the bands in the Southern blot may be difficult to distinguish).

4.1 Heterozygosity

The *heterozygosity*,H, is defined as the probability that an individual is heterozygous at a loci. If p_i represents the probability that an individual has allele i at the loci, then a simple calculation shows

$$H = 1 - \sum_i p_i^2.$$

Inbreeding doesn't alter the p_is (recall we didn't assume no inbreeding in establishing Hardy Weinberg equilibrium), but it does reduce the heterozygosity by a factor (commonly denoted $(1 - F)$). In human populations, inbreeding is uncommon so this can factor can be ignored. In exercise 1 we will see that heterozygosity is highest when all alleles occur with the same frequency. Under this condition we can find how many alleles are necessary to reach a certain level for the heterozygosity.

If we estimate H with the sample proportions, $\hat{p}_i$, then we will obtain a biased estimate of H (denoted $\hat{H}$), although this bias will be negligible in the limit (since plugging in the sample proportions provides the MLE of H and these are asymptotically unbiased). An unbiased estimate can be obtained by using $\frac{n}{n-1}\hat{H}$.

4.2 Polymorphism information content

Another measure of marker polymorphism is the *polymorphism information content* (PIC). It was devised as a measure of polymorphism for the case in which we are studying a rare dominant disease. The PIC is defined as the probability that the marker genotype of a given offspring will allow deduction of which of the 2 marker alleles of the affected parent it had received. The motivation for this measure is we want to measure the quality of a marker with regard to how well it will perform in a linkage study (since we want to determine which marker came from the affected parent). In terms of the allele probabilities, this can be expressed

$$\text{PIC} = 1 - \sum_i p_i^2 - 2\sum_i \sum_{j>i} (p_i p_j)^2.$$

For small p_i this will be approximately equal to H, but in general PIC $< H$.

5 Two point parametric linkage analysis

We will first consider the simplest case of a linkage analysis, namely, there is a single disease gene, the disease is either recessive or dominant, and we have a single marker. This situation is referred to as *two point parametric linkage analysis.* If we consider multiple markers, then we refer to *multipoint linkage analysis.* Non-parametric linkage analysis is used to refer to methods which don't require specification of the mode of transmission (i.e. recessive or dominant), although these methods are more suited to recessive traits.

5.1 Basic linkage analysis

First we will suppose that phase is known. This greatly simplifies the analysis because we can then just count recombinations. We use θ to denote the recombination fraction, and $\hat{\theta}$ to denote its estimate. Since phase is known, we can estimate θ with the proportion of recombinant offspring if this proportion is less than 0.5, and use the convention that if this proportion is greater than 0.5, then $\hat{\theta} = 0.5$. This is known as the *direct method.*

Alternatively, we can use a Bayesian approach. Following Smith (1959), we suppose that prior to observing the data, there is a 1/22 chance that the disease gene is linked to the marker. This is a sensible assumption (especially in 1959) for markers and genes on autosomal chromosomes (since there are 22 of them). Our prior distribution can then be written

$$p(\theta) = \frac{1}{11} 1_{\{\theta < \frac{1}{2}\}} + \frac{21}{22} \delta_{\{\theta = \frac{1}{2}\}},$$

where $\delta_{\{\theta=\frac{1}{2}\}}$ is the *Dirac delta* centered at $\theta = \frac{1}{2}$. Think of δ_x as $\lim_{k \to 0} 1_{\{x-k, x+k\}}/(2k)$. If we use F to represent the family data, then we are interested in computing $p(\theta|F)$ since we can then easily find $P(\theta < 0.5|F)$. If we can compute this probability then we can at least determine if the marker and the gene are on the same chromosome. Using Bayes theorem

$$\begin{aligned} p(\theta|F) &= \frac{p(F|\theta)p(\theta)}{\frac{1}{11}\int_{0 \le \theta < \frac{1}{2}} p(F|\theta)\, d\theta + \frac{21}{22} p(F|\theta = \frac{1}{2})} \\ &= 22 \frac{p(F|\theta)p(\theta)/p(F|\theta = \frac{1}{2})}{2\int_{0 \le \theta < \frac{1}{2}} p(F|\theta)/p(F|\theta = \frac{1}{2})\, d\theta + 21}. \end{aligned}$$

If we then integrate this expression over the region $\theta < \frac{1}{2}$ we find

$$P\left(\theta < \frac{1}{2} \,\middle|\, F\right) = \frac{2\Lambda}{2\Lambda + 21},$$

where

$$\Lambda = \int_{0 \le \theta < \frac{1}{2}} p(F|\theta)/p\left(F \,\middle|\, \theta = \frac{1}{2}\right) d\theta.$$

We mentioned above that we assume phase is known, but we have derived this expression without using that assumption. If phase is known, then the family data is just a count of recombinants out of a certain number of meioses, hence the Binomial model is appropriate for the likelihood. If there are k recombinations out of n meioses

$$p(F|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

This implies

$$\Lambda = 2^n \int_0^{\frac{1}{2}} \theta^k (1-\theta)^{n-k}\, d\theta.$$

The integral here is known as the *incomplete Beta function* and there are a number of ways to compute such expressions. The most direct (since k and $n-k$ are integers) is to just use the Binomial theorem, then integrate the resulting polynomial in θ. If phase is unknown, as is usually the case, then computation of the likelihood is more complicated. The usual approach is to suppose that the 2 possible phases for each individual are equally likely.

5.2 lod scores

Although geneticists use the techniques of maximum likelihood and the associated ideas regarding hypothesis testing (e.g. likelihood ratio tests), the specialized terminology used by geneticists makes this less than evident. A hypothesis of interest to geneticists is the null hypothesis of no linkage

$$H_0 : \theta = 0.5.$$

A natural way to test this is to use a maximum likelihood ratio test. If $L(F|\theta)$ denotes the likelihood for some family data F, the test statistic is then $L(F|\theta)/L(F|1/2)$ maximized over θ. Geneticists don't use this test statistic, instead they use the *lod score*, which is defined to be the base 10 logarithm of the maximized likelihood ratio. We often express the log-likelihood ratio as a function of θ:

$$Z(\theta) = \log_{10} \frac{L(F|\theta)}{L(F|1/2)}.$$

Since we require θ to lie between 0 and 0.5, the usual practice is to plot $Z(\theta)$ against θ (for example, the values 0, 0.001, 0.05, 0.1, 0.2, 0.3, 0.4 are often used) and find the maximizing θ from the plot. This is a sensible procedure since there is the possibility that the MLE is on the boundary of the parameter space. The criterion for linkage is a lod score greater than 3.0. If the χ^2 approximation holds for 2 times the natural logarithm of the log-likelihood ratio, then a lod score of 3.0 implies we use an α-level of 0.0002 for our hypothesis test. While this may seem too strict, we will see it is a sensible procedure given the current practice of genome scans.

There are a variety of algorithms for calculating the likelihood for general pedigrees. The appropriate method depends on the problem at hand. The first general method for calculating the likelihood for general simple pedigrees was devised by Elston and Stewart (1971). The idea of this method is to realize given the genotype information, the phenotypes of distinct individuals are independent (since they only depend on genotypes). In addition, given the parents' genotype information, the grandparents' genotypes don't help predict the childrens' genotype: that is, the childrens' genotypes are conditionally independent of the grandparents' genotypes, given the parents genotypes. Since distinct families are independent, we just consider computing the likelihood for one family (the total likelihood is the product of these factors). If we use g to denote the genotypes and p the phenotype data (this is the phenotype at one or more loci), then we can write the likelihood as

$$L(p|\theta) = \sum_g L(p|g,\theta) L(g|\theta).$$

Here, $L(p|g,\theta)$ can be expressed as a product $\prod_i L(p_i|g_i,\theta)$ over individuals (since phenotypes only depend on genotypes), and, supposing there are N generations $L(g|\theta)$ can be expressed as a product over generations $p(g_0|\theta)\prod_i p(g_i|g_{i-1},\theta)$. Although we have thus far concentrated on the recombination fraction, θ can represent any set of parameters that determine the way genotypes are transmitted and genotypes become phenotypes. Since we evaluate the likelihood by considering each generation in turn, this algorithm is often described as *peeling* the generations off of the pedigree. This algorithm has been extended in a variety of ways, in particular, we can extend this method to handle missing data and complex pedigrees. The computational effort here is substantial (for example, for 2 alleles at 2 loci there are 4 haplotypes, hence there are $\binom{4}{2} + 4 = 10$ possible genotypes, so if there are m family members, there are 10^m possible genotypes to consider). Although we have only considered 2 loci thus far, this algorithm can be used when there are more than 2 loci used to define the genotype. The number of computations increases linearly with the number of pedigree members but exponentially with the number of loci considered. An alternative is the Lander-Green algorithm (1987). This algorithm increases linearly with the number of loci considered and exponentially in pedigree size. Extensions of the Lander-Green algorithm are used in the popular program GENEHUNTER.

Exercises

1. Show that the heterozygosity, H, is maximized when all alleles are present in equal proportions in the population.

2. In this exercise we will compute an approximation to the number of markers necessary to localize a gene to a given resolution with a certain probability.

   (a) Given a map of length L cM, what is the probability of an arbitrary locus being no more than d cM from a randomly located marker (*Hint*: treat the genome as a string of length L, suppose there is a fixed location for some marker, and suppose the locus is uniformly distributed over the genome)?

   (b) Compute the probability, p, that a gene is within distance d of at least one of m markers (*Hint*: use the approximation that the event that a gene is within distance d of marker i is independent of the event that the gene is within distance d of marker j for all i, j).

   (c) Using your expression from part b., express m as a function of p, d and L.

   (d) For humans, averaging over the sexes, $L = 3300$ cM. How many markers do you need to localize a gene if you want to ensure we stand a 95% chance of having a marker within 10 cM of a given gene?

3. Use the method of Smith (1959) to compute the posterior probability of linkage when you observe 2 recombinants out of 10.

4. For X-linked traits, researchers often use a significance level of $\alpha = 0.002$, what value does the lod score need to reach to obtain this level of significance?

References

Lander, E., Green, P (1987), "Construction of multilocus genetic maps in humans", *PNAS*, 84:2363-2367.

Ott, J. (1999), *Analysis of Human Genetic Linkage*, Johns Hopkins Press, Baltimore.

Smith, C. (1959), *Am J Hum Genet*, 11:289-304.