6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Fall 2008 6.047/6.878 Lecture 13

Population Genomics II : Pardis Sabeti

1 Introduction

In this lecture we will discuss linkage disequilibrium and allele correlations, as well as applications of linkage and association studies.

2 Linkage/Haplotypes

Recall Mendel's Law of Independent Assortment, which states that allele pairs from different loci separate independently during the formation of gametes. For example if we have a population of individuals such that at a given locus we find 80% A and 20% G, while at a second locus of interest we find 50% C and 50% T, then independent assortment would imply that the corresponding haplotype frequencies at this double loci in the next generation should be (80% * 50%) = 40% AC, (20% * 50%) = 20% GC, and similarly, 40% AT, 10% GT.

However in real life the separation is not always independent, since the two loci may be linked, by being close together on the same chromosome. For example for the allele frequencies at the two loci given above we might get actual haplotype frequencies of 30% AC, 20% BC, 50% AT, and 0% GT.

This occurs due to "linkage disequilibrium", which is the disequilibrium of the allele frequencies which occur when the loci are close enough that linkage occurs (eventually, recombination occurs and the frequencies settle into equilibrium). Linkage disequilibrium such as in this example, where there are only two alleles observed at each of two loci of interest, is often measured by a quantity, D, which is defined as follows:

$$D = |ObsP(A_1B_1) * ObsP(A_2B_2) - ObsP(A_1B_2) * ObsP(A_2B_1)|$$

where $ObsP(A_iB_j)$ denotes the fraction of observed haplotypes at the A and B loci consisting of allele A_i at locus A and allele B_j at locus B. (Note: in our example, $A_1 = A$, $A_2 = G$, $B_1 = C$, and $B_2 = T$.) In general, A_1, B_1 denote the "major alleles" and A_2, B_2 the "minor alleles" at the two loci.

Note that if there is no linkage disequilibrium, *i.e.,* the allele pairs separate independently, then $ObsP(A_iB_j)$ is simply the product of the allele frequencies, $P(A_i)$ and $P(B_j)$. It follows that in this case,

$$D = |P(A_1) * P(B_1) * P(A_2) * P(B_2) - P(A_1) * P(B_2) * P(A_2) * P(B_1)| = 0.$$

For the observed haplotype frequencies given above, we instead get

$$D = |0.3 * 0 - 0.5 * 0.2| = 0.1.$$

It is useful to compare to D to $Dmax$, the maximum possible value of D given the allele frequencies you have. One can show that $Dmax$ is equal to the smaller of the expected haplotype frequencies $ExpP(A_1B_2)$ and $ExpP(A_2B_1)$ when there is no linkage (see Appendix). In particular, it is convenient to define

$$D' = \frac{D}{Dmax}$$

which in this case equals $0.1/0.1 = 1$, indicating the maximum possible linkage disequilibrium, or "complete skew". (D' varies between 0 when totally unlinked, to 1 when totally linked.)

Three other examples of linkage disequilibrium are given in the slides for this lecture– two with complete skew (D'=1) – in each case where only two gametes are observed, and one with partial (D'=0.6), where all four gametes are observed.

Note that it is unusual to see all four gametes, absent recombination, since the fourth gamete would most probably require a repeat mutation at one of the two loci, which is assumed to never happen by the Inifinite Sites Assumption. Consistent with the Infinite Sites Assumption is a simple test called the Four Gamete Test which says that if there are fewer than four gametes (at a double locus), then recombination does not need to be invoked to explain it, however if there are four gametes then a recombination must have occured. Note that the Infinite Sites Assumption is just a model, which fails for example when the mutation rate is high.

More generally, for n loci, in the case of complete linkage, so that no recombination occurs, we expect at most $n+1$ distinct haplotypes (the ancestral haplotype and a haplotype resulting from a single mutation at each of the n loci). Whereas with no linkage, so that all recombinations occur, any of 2^n haplotypes occur, since any subset of the n loci may have a mutation in that case.

Linkage disequilibrium is governed by recombination rate and the amount of time (in generations) that has passed. With time, linkage disequilibrium drops off, more quickly when the recombination rate is higher. Some parts of the genome have higher recombination rates than others, so it is not always physical distance in the genome which determines recombination rates between two loci. (For example, there is a motif which seems to recruit recombinase, thus raising recombination rates whereever that motif is prevalent.)

History also matters– for example, perhaps one of the mutations hasn't been around as long, and therefore there hasn't been enough time for the linkage disequilibrium to break down.

Correlation between alleles is an important concept related to linkage. It's governed by linkage between the alleles and the allele frequencies. When allele frequencies are very different, correlation is lower. Correlation is measured by the *correlation coefficient* r^2, defined as

$$r^2 = \frac{D^2}{P(A1)P(A2)P(B1)P(B2)}$$

in the case where there are two alleles at each of the two loci of interest. The correlation coefficient is a measure of how predictive the allele at locus A is of the allele at locus B. There are two examples given in the slides– one with complete correlation ($r^2 = 1$), and the other with low correlation ($r^2 = 0.25$).

Correlation is also useful in disease mapping. This slide shows how one disease may be predicted by the SNP at locus B, whereas another disease may not be predicted by any one SNP, but can be predicted by the double SNP at loci A and B. In these cases we say that the SNP at B is a "tag" for the first disease, while the double SNP at A and B is a "tag" for the second disease.

Next we will discuss some linkage applications, but first we give an overview of classic linkage analysis.

3 Classic Linkage Analysis

The classic goal is to find out what causes a disease, or more generally, a phenotype of interest. The classic way to do this was to study pedigrees. Traditionally, squares denote men, and circles women, color denotes affected individuals, and white denotes unaffected individuals. An informative marker in a linkage study is heterozygous in the parent of interest. For example, the the first slide titled "Tracing Inheritance of a Disease Gene" it is easy to see that C is transmitted by the father whenever the child is affected, but in the second slide, it is A that is transmitted by the father whenever the child is affected. Thus in the first case it seems that the causative site for the disease is linked to allele C, whereas in the second case, it seems to be linked to allele A. A classic way to measure linkage given a pedigree is the LOD score, which is the log odds that the observed phenotypes would occur if the recombination rate were some assumed value vs. 0.5 (which corresponds to no linkage).

In this way it took about 20 years to localize and understand the most prevalent known causative mutation for cystic fibrosis (CF). Via DNA sequence comparison, it was discovered that the mutation is in the transmembrane part of a protein affecting function of the chloride channel.

In classic linkage analysis, you can find a many megabase region with a linkage signal, but then how do you find where the causative gene is in that region? It would obviously be much easier if the region were sufficiently small.

Note that in classic linkage analysis we only need linkage with a locus, not with an allele.

4 Population Association

Nowadays we look for allele-associated linkage– *i.e.,* the allele and the causative mutation are close enough that the phenotype is associated, not just linked, to the allele. (The allele and the causative mutation are close enough that recombination has never occurred between them.)

For example, the ApoE4 allele is associated with Alzheimer's disease, HFE with Venous thrombosis, and several other common variants are listed in the slides. Thanks to dramatic technological advances in sequencing we now have a comprehensive catalog of human variation which make genome-wide association studies feasible. Eric Lander et al were responsible for the first genome-wide map, of 4000 SNP's, in 1998. In 1999, The SNP Consortium (TSC) announced their 2-year goal to map 300,000 SNP's across the human genome, but in 2000 already 1.4 million SNP's were mapped (via shotgun sequencing by David Altshuler et al), and the progress continued to escalate, so that by 2004 there were 8 million SNP's mapped.

The International HapMap Project formed in 2002 with the goal of genotyping 3 million well-defined SNP's in 270 people in order to determine correlation among these SNP's, *i.e.,* resulting in a mapping of haplotypes.

Note that in parts of the genome where recombination rates are low, 1 SNP per 5 kilobases of DNA can be sufficient, but in areas where recombination is frequent, considerably more SNP's are needed, for a meaningful association study.

In October of 2005 HapMap met its goal. The slide shows pairwise comparison of the SNP's across the genome, displayed via Haploview by Mark Daly et al. In this graphic red represents linkage and white represents no linkage. Pairs of SNP's across the genome point down to linkage disequilibrium values (D') indicated by color (red for linkage, white for no linkage). Note that the closer two SNP's are, the higher prob. they will point to red. Also note the "haplotype blocks" – red regions showing almost complete linkage disequilibrium between all the SNP's in those regions. Straight white areas correspond to recombination hotspots, where linkage disequilibrium completely drops off.

4.1 Case Control

Roughly 80% of common genetic variations in humans have been identified, but further sequencing still needs to be done. Major technology advances have brought us to the point where we can genotype 1.2 million polymorphisms in 500 people or more, per day.

When studying all of this data, it is essential that cases and controls are well defined. For example, we can calculate the frequencies of disease vs. no disease in the presence of two different alleles at a given locus, and then compute the relative risk of the disease in the presence of each of the two alleles. In particular, calculate the Odds Ratio (OR) for allele 1 as follows:

$$OR(\text{All1}) = \frac{(\text{disease and allele 1}) \times (\text{disease and allele 2})}{(\text{no disease and allele 1}) \times (\text{disease and allele 2})}$$

and analogously, the OR(All2), which is simply the reciprocal of OR(All1). To evaluate the confidence we should have in these numbers we can compute the standard error and the 95% confidence intervals around OR(All1) and OR(All2), for example for the frequencies given in the Case Control slides for this lecture, we get 95% confidence intervals which are significant since they do not span the value 1 (which would correspond to the completely random case). It is also common to compute the Chi-Square value as a measure of the significance of these results.

4.2 Example: Age-related Macular Degeneration

A good example of a disease which has become much more understood thanks to genome-wide association studies is Age-related Macular Degeneration (AMD), which causes loss of vision at the focal point in older people, so that they only have peripheral vision. In 2004 we had no idea why it was happenning. In 2006 at the Broad Institute and elsewhere, using a whole genome association scan, scientists found 3 genes with 5 common variants explaining 50% of AMD occurrence.

For many other diseases (diabetes, schizophrenia, multiple sclerosis, etc.) there are already good candidates for causative mutations, which are in the process of being better understood by similar studies.

4.3 Transmission Disequilibrium Test (TDT)

A commonly done association test which is not subject to population stratification is the Transmission Disequilibrium Test (TDT), within families.

To do TDT, we look at families of three (two parents and a child), in which at least one of the parents is heterozygous for a marker and the child is affected, and note which allele is passed to the affected child from each heterozygous parent. In the absence of association between the marker and the disease, we expect a 1:1 ratio for the allele transmitted from the heterozygous parent to the affected child. Over many families, we can compute

$$X^2_{TDT} = \frac{(n_1 - n_2)^2}{(n_1 + n_2)}$$

where n_i denotes the number of times allele i was transmitted from a heterozygous parent to the affected child, as a measure of the association between this marker and the occurrence of the disease.

4.4 Measuring length of haplotype

Another measure of linkage disequilibrium which is specific not only to the locus of interest but also the particular allele at that locus is the length of the haplotype of that allele at that locus. Sabeti developed Extended Haplotype Homozygosity (EHH), an especially useful measure of the length of a haplotype. This is a bifurcation diagram showing the region of the ancestral chromosome with nodes corresponding to a set of long-range markers in that region, and branches off of those nodes to other sub-branches with corresponding nodes which may branch or not, as well, depending on what mutations have occurred at what markers. The thickness of the branches between each pair of nodes indicates the prevalence of the corresponding allele at the marker corresponding to its endpoint. Thus it captures the decay of association to markers at various distances from the starting locus, giving a fuller picture of the "length" of the haplotype.

5 Appendix

Irwin Jungreis has kindly provided the following derivation of the formula for $Dmax$, the maximum value of the measure of linkage disequilibrium, D, given the allele frequencies.

Since the allele frequencies are computed from the observed data, they impose constraints on the observed values. Let $p = P(A1)$, $q = P(B1)$. Then we have

$$\begin{aligned}
ObsP(A_1B_1) + ObsP(A_1B_2) &= p \\
ObsP(A_2B_1) + ObsP(A_2B_2) &= 1 - p \\
ObsP(A_1B_1) + ObsP(A_2B_1) &= q \\
ObsP(A_1B_2) + ObsP(A_2B_2) &= 1 - q
\end{aligned}$$

If we treat p and q as known and $ObsP(A_iB_j)$ as unknowns we have 4 equations in 4 unknowns. However, they are singular, because the sum of the first two equations equals the sum of the second two. So there is a 1-dimensional family of solutions. Since the expected values are one possible solution, the family of solutions can easily be seen to be:

$$\begin{aligned}
ObsP(A_1B_1) &= ExpP(A_1B_1) + t \\
ObsP(A_1B_2) &= ExpP(A_1B_2) - t \\
ObsP(A_2B_1) &= ExpP(A_2B_1) - t \\
ObsP(A_2B_2) &= ExpP(A_2B_2) + t
\end{aligned}$$

with constraints $t >= -ExpP(A_1B_1), t <= ExpP(A_1B_2), t <= ExpP(A_2B_1), t >= -ExpP(A_2B_2)$ since the probabilities $ObsP(A_iB_j) \geq 0$.
It is easy to see then that

$$\begin{aligned}
max(t) &= min(ExpP(A_1B_2), ExpP(A_2B_1)) \\
max(-t) &= min(ExpP(A_1B_1), ExpP(A_2B_2)) \\
max(|t|) &= max(min(ExpP(A_1B_2), ExpP(A_2B_1)), min(ExpP(A_1B_1), ExpP(A_2B_2))
\end{aligned}$$

Since $A1$ and $B1$ are the major alleles, we have $p \geq 1 - p$ and $q \geq 1 - q$, so $pq >= p(1-q)$, $pq >= q(1-p)$, $(1-p)(1-q) <= p(1-q)$, and $(1-p)(1-q) <= q(1-p)$. Thus

$$min(ExpP(A_1B_2), ExpP(A_2B_1)) >= ExpP(A_2B_2) = min(ExpP(A_1B_1), ExpP(A_2B_2))$$

so

$$max(|t|) = min(ExpP(A_1B_2), ExpP(A_2B_1)).$$

It remains to show that $D = |t|$. By definition,

$$D = |ObsP(A_1B_1) * ObsP(A_2B_2) - ObsP(A_1B_2) * ObsP(A_2B_1)|,$$

but

$$ObsP(A_1B_1)*ObsP(A_2B_2) = ExpP(A_1B_1)*ExpP(A_2B_2)+t(ExpP(A_1B_1)+ExpP(A_2B_2))+t^2$$

and

$$ObsP(A_1B_2)*ObsP(A_2B_1) = ExpP(A_1B_2)*ExpP(A_2B_1)-t(ExpP(A_1B_2)+ExpP(A_2B_1))+t^2$$

so $D = |pq(1 - p)(1 - q) - p(1 - q)q(1 - p) + t| = |t|$, as desired.