linkage disequilibrium and association mapping

Gil McVean

Department of Statistics, University of Oxford

March 14, 2002

Contents

1 Complex disease

Complex disease is perhaps the biggest challenge facing medical genetics. The clustering of diseases such as heart disease, cancers, asthma, diabetes and schizophrenia, among related individuals indicates that almost all common diseases have some genetic basis to susceptibility. If the genetic basis of susceptibility could be identified, efficient treatments for disease could hopefully be engineered, or risk factors identified and eliminated from at risk individuals. Although understanding the genetic basis of susceptibility is not an end in itself, an understanding of the genetic factors influencing the etiology of a disease is a critical component of any program to reduce disease incidence.

1.1 Features of complex disease

In the previous lecture we considered the population genetics of Mendelian disease mutations, such as those causing cystic fibrosis or Dyastrophic Dysplasis. Mendelian diseases typically result from mutations in a single gene that have high penetrance, and which are rare at the population level. In contrast, complex, or multifactorial (to emphasise that non-genetic factors are also important in the disease) diseases are probably influenced by genetic variation at several loci in the genome, each of low penetrance. In addition, factors such as environmental influences, interactions between genes and the environment (G $\times$ E interactions), and interactions among genetic variants at different loci (epistasis) are all likely to be important for complex disease. Unlike Mendelian disease, complex diseases can be very common, and have important consequences for the economics (and politics) of medical provision.

1.2 Limitations of pedigree studies

The classical paradigm for identifying loci contributing to disease is the pedigree study. In a family where there are affected and unaffected individuals, the coseg-

regation of the disease phenotype and alleles at marker loci is used to detect associations. However, there are several important limitations to the power of pedigree studies for complex disease. Most importantly, the low penetrance of complex disease means that vast amounts of information are required to identify associations, and pedigrees of adequate size simply are not available. Even if candidate regions can be identified from pedgirees, the resolution of linkage studies is typically on the order of a few cM, which in terms of the human genome may correspond to several Mb of DNA, and 100s, or 1000s of genes. Furthermore, if pedigree studies can identity complex disease loci to the gene level, there is a strong ascertainment bias towards variants that cause Mendelian forms of complex disease (e.g. mutations in *BRCA1* associated with early-onset breast cancer) which actually contribute relatively little to the disease phenotype on a population scale.

1.3 Limitations of family-based methods

For these reasons, a population view of complex disease is highly preferable. In the mid-1990s, the paradigm for disease mapping became family-based population methods, of which the exemplar is the transmission disequilibrium test (TDT Spielman et al., 1993). The TDT consider triads of affected individuals and their parents. Associations at the genetic level are identified by comparing the proportions of transmitted and untransmitted alleles. Untransmitted alleles therefore act as an internal control for transmitted alleles, leading to a very robust experimental design. In addition, because the TDT considers a population sample of the disease, it will not bias towards locating rare, Mendelian forms of complex disease. Allelic heterogeneity (multiple susceptibility alleles at a disease locus), multiple contributory loci, low penetrance and environmental effects will all act to reduce the power of population methods, but if a significant result is obtained with the TDT, it is unlikely to be a false positive.

However, there are several factors that reduce the power and efficiency of the TDT. First, it requires a lot of effort; at least three individuals have to be genotyped

for each data point. Second, obtaining parental genotypes can be difficult, particularly for late-onset diseases such as Alzheimers, although alternative designs using sibs, rather than parents, can be constructed. Finally, in order to be informative at a locus, parents have to be heterozygous at a locus. Although efforts can be made to use loci with high heterozygosity, a significant fraction of triads will always be uninformative.

1.4 Critical assumptions in association mapping

The alternative to the TDT at the population level is historically the first type of mapping approach; the case-control study. Case-control studies compare allele frequencies at loci among disease and matched control populations (matched for factors such as race, sex, age and lifestyle). The reasoning behind case-control studies is that disease susceptibility mutations will show strong differences in frequency between cases and controls. While it is unlikely that the causative mutation will be included among the marker loci, linked neutral variants will also be associated with the disease phenotype through linkage disequilibrium (population level associations) with the causal variant. Significant differences in allele frequency can be taken as evidence for linkage, although there are several factors that can cause association without linkage (see below).

The critical assumptions of association mapping are embodied in what has become known as the common-disease common-variant (CDCV) model for complex disease (Cardon and Bell, 2001). The key feature of the model is that disease susceptibility is influenced by a few loci, each of which has a single major allele contributing to the phenotype . Each allele may have low penetrance, but if enough data can be gathered, associated should be detectable. Implicit is the assumption that alleles contributing to disease susceptibility are at an appreciable frequency in populations (¿¿1%). While the biological reality of these assumptions is unknown, they are critical to the success of association mapping experiments. If allelic heterogeneity is high, or disease susceptibility is influenced by many rare

alleles at many loci, association mapping studies will have little power. Under the pessimistic scenario, it is perhaps worth noting that understanding the genetic basis of susceptibility will not lead to widely applicable treatments for disease.

1.5 Case-control genealogies

For the rest of the lecture we will assume that the CDCV model is true. For theoretical population genetics, the key questions become; what factors influence associations between mutations at different loci? And, what is the best design for an association mapping experiment? From the empirical point of view, we need to have some idea of what allelic associations in the human genome look like.

In keeping with the rest of the course, we will take a genealogical perspective to describing the population genetics of association mapping. In particular, we can think of a genealogy of chromosomes at a locus contributing to the disease phenotype for both cases and controls. If there were perfect concordance between the disease-associated mutation and the disease phenotype, the genealogy of the disease chromosomes would be a subgenealogy within the larger population genealogy (just as was described previously for the intra-allelic genealogy for Mendelain disease mutations). However, because of incomplete penetrance, allelic heterogeneity at the locus, alleles at other loci, and non-genetic factors, the genealogies of the case and control chromosomes are interwoven and it is highly likely that considerable history will be shared between the two (particularly the MRCA).

However, there differences between the genealogies of the case and control chromosomes; notably a subset of the disease chromosomes show rapid coalescence under the causative mutation leading to strong identity by descent at the locus. In contrast, the control chromosome genealogy is much less distorted or unbalanced than the disease chromosome genealogy.

1.6 The population genetics of associations

In terms of identifying disease loci, the key point from comparing disease and control chromosomes is that the former should show strong identity by descent at the disease locus relative to the control population. However, as we are unlikely to have sampled the causative mutation as a marker, we must look for excess identity at loci linked to the disease mutation.

In the previous lecture I introduced a crude deterministic model for associations between disease mutations and linked marker alleles. The key point was that when the mutation first appears, it picks a haplotype at random from the population, so marker allele frequencies in the disease subpopulation are all equal to one. Over time, recombination breaks down the association, such that the difference in allele frequency between the disease and control populations, $q_D - q_C$, at a marker locus with recombination distance r from the disease locus decays in an exponential fashion

$$q_D - q_C = (1 - q_C)e^{-rt} \tag{1}$$

Where t is the number of generations since the origin of the disease mutation.

A deterministic model, however, has several limitations. Most importantly, it ignores the stochastic nature of the recombination and drift processes that influence allelic associations. A more realistic way of modelling recombination is in the genealogical context; looking back in time from the present, the history of the disease locus and a linked marker locus are perfectly correlated until recombination events (the waiting times to which are exponentially distributed) decouple the histories of the marker and disease loci. Consequently, while the marker locus genealogy shares some of the increase in identity among case chromosomes seen at the disease locus, the effect is diluted by recombination.

1.7 Measuring associations

How do associations between alleles relate to the power of association mapping? The simplest test we can construct to detect associations considers markers one

at a time, contrasting the frequency of markers alleles among case and control chromosomes. Consider the following result for a marker with two alleles

	Allele frequency	Proportion of sample
Case	q_D	x_D
Control	q_C	x_C

Table 1: Design of association tests. The sample size is n

A simple test we can construct is to assume that under the null the cell counts are Normally distributed, in which case the test statistic

$$X^2_M = n\frac{(q_D - q_C)^2 x_D(1-x_D)}{\bar{q}(1-\bar{q})} \tag{2}$$

where $\bar{q} = q_D x_D + q_C(1-x_D)$, will be approximately χ^2 distributed with one degree of freedom. An alternative approach would be to use a likelihood ratio test.

Clearly, the critical factor influencing test power is the difference in allele frequency between the case and control chromosomes, scaled by the heterozygosity of the marker. When there are more than two alleles at the marker locus, the equivalent test statistic is given by the sum over alleles of the difference in allele frequency scaled by the average frequency of the allele.

$$X^2_M = n x_D(1-x_D)\sum_i \frac{(f_{iD} - f_{iC})^2}{\bar{f}}$$

1.8 Factors affecting test power

What factors will influence the power of the association test? In order to address this question we can consider the deterministic model for differences in allele frequency (1), assuming that the frequency of the allele in the control population has remained constant over time, and that the control population is much greater than the case population, hence $\bar{f} \approx q_C$. The expected value of the test statistic is therefore

$$E[X^2_M] \propto (k-1)e^{-2rt} \tag{3}$$

Where k is the number of alleles at the marker locus (the sample size and proportion of sample represented by disease chromosomes are also obviously important, but are in the control of the researcher). Not surprisingly, the recombination rate strongly influences test power, as associations between tightly linked alleles persist for many generations. Likewise, the age of the disease mutation is important as older alleles will have had more opportunity for recombination. What is surprising from the deterministic solution is that the marker locus allele frequencies are not important. The reason for this is that while rare mutations captured by the ancestral disease mutation are highly informative about associations, they are also less likely to be captured. The only factor at the marker locus that influences power is the number of alleles; more alleles means more power.

To what extent should we trust the deterministic formulation of test power? There are two sources of stochasticity that may be important in determining test power. First, genetic drift among allele frequencies in the control population since the origin of the disease mutation may be important. Second, the variance in associations will be affected by allele frequencies. For example, in the deterministic case with two alleles the variance in X^2_M is proportional to $1 - 2H$, where H is the heterozygosity at the marker locus. In other words, loci with skewed allele frequencies (high homozygosity) will be more variable in their power to detect associations.

1.9 Common SNPs for common diseases?

As an example of how modelling the stochastic nature of associations is important in the design of association studies, consider the question of whether rare or common SNPs are better, i.e. more powerful, for detecting associations. Compare two scenarios, one in which the causative mutation is rare, 5-10% frequency, and the other where the causative mutation is common, 25-75% frequency. Although the WF model does not capture the full complexity of the stochastic influences on associations in human populations, we can get a good idea of the extent and

variability of associations through coalescent simulations. In particular, we can ask what the distribution of test power is for both rare SNPs at marker loci (5-10%) and common SNPs at marker loci (25-75%), where the marker locus is about 2.5kb from the disease locus.

Very different results are obtained under the two scenarios. When the causative mutation is rare, most rare SNPs show no association, but a fraction, about 20% show very strong association. Most common SNPs also show no association, but there is a small fraction, again about 20%, that show weak associations. In contrast, when the causative mutation is common, most rare SNPs show no association, and a few show weak association, but almost all common SNPs show appreciable association.

2 Linkage disequilibrium

Associations between marker alleles and disease mutations in case-control studies are clearly related to the concept of population level associations as measured by statistics of linkage disequilibrium. Alleles are said to be in linkage equilibrium if the frequency of a particular genotype is equal to the product of the frequencies of the individual alleles that make up the genotype. A natural way to measure the deviation from linkage equilibrium is to compare the observed and expected genotype frequencies

$$D_{AB} = f_{AB} - f_A f_B$$

The term linkage disequilibrium is actually an inappropriate name for deviations from this expectation as physical linkage between loci is neither necessary, nor sufficient to generate associations.

For two loci, each with only two alleles, there is only one coefficient of linkage disequilibrium; $D_{AB} = D_{ab} = -D_{Ab} = D_{aB}$. The expectation of this coefficient is zero, hence in order to summarize patterns of association in empirical data, it is necessary to consider measures that always take positive values.

Several different measures of linkage disequilibrium (hereafter LD) have been proposed. One of the most widely used approaches is to consider associations between alleles as correlation coefficients with allelic values represented as 0s and 1s (note that this approach does not make sense for multiple alleles at a locus, except perhaps when there is sense in a scaled metric, e.g. for microsatellites).

$$\rho_{AB} = \frac{D_{AB}}{\sqrt{f_A(1-f_A)f_B(1-f_B)}}$$

Again, the expectation of the correlation coefficient is zero, so it is convenient to consider the square of the correlation coefficient (Hill and Robertson, 1968). This quantity is often referred to r^2 or Δ^2.

The square of the correlation coefficient ranges between 0 and 1. However, unless the allele frequencies at the two loci are identical, perfect correlation cannot be achieved. To solve this problem, Lewontin (1964) proposed to scale the standard measures of linkage disequilibrium by the maximum value it can achieve given the allele frequencies.

$$\begin{aligned} |D'| &= \frac{-D_{AB}}{\min(f_A(1-f_B),(1-f_A)f_B)} \quad D_{AB} < 0 \\ &= \frac{D_{AB}}{\min(f_A f_B,(1-f_A)(1-f_B))} \quad D_{AB} > 0 \end{aligned}$$

While this formulation solves one problem, it does create another; namely that when there are only three of the four possible haplotypes present for a pair of loci (each with two alleles), $|D'|$ will always be one.

A number of other statistics have been suggested for summarizing LD. Recently, a statistic has been suggested which is claimed to be superior for association studies (Devlin and Risch, 1995)

$$\delta_{AB} = \frac{D_{AB}}{f_B f_{ab}}$$

With the constraint that $D_{AB} > 0$. However, because there can sometimes be two possible values that the statistic can take, depending on how alleles are coded, the statistic has not been widely accepted.

There is no *a priori* reason for supposing one summary statistic of LD to be superior, and all have serious weaknesses. Most importantly, the two principal

statistics confuse a lack of information about associations (i.e. a lack of power) with either evidence for absence, r^2, or presence, $|D'|$, of associations.

2.1 LD and power in association studies

With respect to association mapping, however, one statistic, r^2, does bear a direct relationship to the power of association tests. Equation (2) shows how the power to detect associations between a marker and a disease phenotype depends on the difference in allele frequency between case and control populations. This difference in allele frequency depends on associations between the marker alleles and the disease locus alleles, and the strength of association between the disease mutation and the disease phenotype. If we write p_D and p_C for the frequency of the disease-associated mutation in case and control chromosomes, we can write the marker locus test statistic as

$$X^2_M = n\frac{\left[D_{pq} - x_D D^D_{pq} - (1-x_D)D^C_{pq}\right]^2}{\bar{q}(1-\bar{q})x_D(1-x_D)} \times \frac{1}{(p_D - p_C)^2} \qquad (4)$$

Where D_{pq} refers to the standard coefficient of LD between the disease mutation and the marker allele in the whole population, and the superscripts D and C refer to the LD between the disease and marker mutations in the case and control chromosomes (see Pritchard and Przeworski, 2001, for a related derivation).

How does this result relate to the different measures of LD? In the best case scenario, when $p_D = 1$ and $p_C = 0$; i.e. there is a one-to-one correspondence between the presence of the disease mutation and the presence of the disease phenotype, the test statistic is proportional to the square of the correlation coefficient between the marker allele and the disease mutation. Furthermore, when there is incomplete penetrance, and other causative factors, but no linkage disequilibrium between either the disease or marker locus alleles and other genetic factors, the expected value of the test statistic at the marker locus is equal to the test statistic at the disease locus multiplied by the square of the correlation coefficient between the marker allele and the disease mutation (Pritchard and Przeworski, 2001). In

short, under certain conditions there is a one-to-one correspondence between LD as measured by r^2 and the power of association studies.

2.2 LD in Wright-Fisher populations

Given the relationship between LD and the power of association studies, it is clearly important to understand what factors influence LD from a theoretical viewpoint. The starting point for any theoretical population genetics treatment is a WF constant population size model with recombination. Consider two loci in a population of N_e diploids, separated by a recombination fraction of r. Each generation associations between alleles are generated by the stochastic processes of genetic drift and mutation, and broken down by recombination. The key quantity influencing this dynamic equilibrium is the relative rate of recombination and genetic drift as measured by the population recombination rate, $4N_e r$.

Of course, for any particular value of $4N_e r$, there is a huge variation in the degree of LD observed at a given locus. We can simulate the distribution of LD through coalescent methods (Hedrick, 1987); for low recombination rates the r^2 statistic has a highly skewed bimodal distribution, with most values near zero, but with a second peak around one (perfect correlation). For high recombination rates all values of r^2 are low. In contrast, the $|D'|$ statistic is unimodal with a peak at one for low recombination rates, and bimodal for high rates, with most values low, but a large fraction still at one.

What can we say about the moments of the distribution of LD? Analytically, it is not possible to derive the moments of the distributions of either statistic of LD. However, it is possible to derive the expectation of a quantity that is closely related to r^2 (Ohta and Kimura, 1971)

$$\sigma_d^2 = \frac{E[D_{AB}^2]}{E[f_A(1-f_A)f_B(1-f_B)]} \approx E[r_{AB}^2] = \frac{10+R}{22+13R+R^2} \qquad (5)$$

Where $R = 4N_e r$. If rare alleles are excluded, σ_d^2 is an adequate approximation to the expected value of r^2. In WF populations LD is expected to fall off rapidly with recombination, such that for $R > 20$, little LD is expected to be found.

2.3 Estimating $4N_e r$

Given the importance of the scaled recombination rate, $4N_e r$, in determining LD, it would be useful to be able to estimate the parameter from empirical data. Several different methods have been proposed to estimate $4N_e r$, the earliest of which was to attempt to fit a theoretical expectation related to equation (5) to a scatter plot of the values of r^2 for all pairs of alleles at the β-globin locus (Chakravati et al., 1984). However, because the theoretical result relates to the expected value for a single locus, not an ensemble of interdependent pairs, such an approach cannot be justified (Weir and Hill, 1986).

Just as several moment estimators for the population mutation rate can be derived, so it is possible to derive moment estimators of $4N_e r$. Hudson (1987) derived the expectation of the variance in pairwise differences as a function of $4N_e r$ which generates an unbiased estimator, however the variance of the estimator is large and resulting confidence intervals are likewise.

Full coalescent-based likelihood estimation of $4N_e r$ has been developed using importance sampling schemes, or MCMC approaches to propose coalescent histories compatible with the data (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; Kuhner et al., 2000). While coalescent inference is highly desirable, the computational burden of the estimation problem is enormous, and current methods are unfeasible for large data sets with high recombination rates.

A number of *ad hoc* methods have been proposed to estimate $4N_e r$. Hey and Wakeley (1997) combined the coalescent likelihoods for every pair of loci sampled from four chromosomes and Wall (2000) proposed to use an approximation to the likelihood of the summary statistics H (number of haplotypes) and R_M (minimum number of recombination events). Recently, composite likelihood methods have been developed which sum coalescent likelihoods for every pair of segregating sites assuming independence (Hudson, 2001; McVean et al., 2002).

Perhaps the most interesting aspect of estimating $4N_e r$ from empirical data is that we can ask how well the model describes the observed patterns. Goodness-

of-fit tests on empirical data can reveal important biological properties that are missing from the assumed models (Frisse et al., 2001), such as gene conversion, variation in the recombination rate, and demographic influences on LD.

Many different demographic processes influence LD. Population growth tends to decrease LD (Slatkin, 1994; Krugylak, 1999), population bottlenecks increase LD, and so do models of population structure. For any given demographic model it is possible to simulate coalescent histories in order to build up an understanding of the patterns expected (Pritchard and Przeworski, 2001). By comparing such patterns to real data, hypotheses can be made as to the key factors influencing human diversity (e.g. Reich et al., 2001; Weiss and Clark, 2002). Methods for actually testing such hypotheses, however, have yet to be implemented.

2.4 A genealogical view of LD

Allelic associations are obviously a property of mutations, but they must also reflect associations in the genealogical history of the loci. Given the importance of a genealogical understanding in population genetics, it is of interest to ask what aspect of genealogical history allelic associations reflect.

Although it is not possible to derive results about the expectation of r^2 or $|D'|$, it is possible to show that the quantity considered by Kimura and Ohta (1971) has a direct relationship to correlations in coalescence time for pairs of loci (x and y) on chromosomes sampled in different ways (McVean, 2002).

$$\sigma_d^2 = \frac{\rho[\tau_{x(ij)}, \tau_{y(ij)}] - 2\rho[\tau_{x(ij)}, \tau_{y(ik)}] + \rho[\tau_{x(ij)}, \tau_{y(kl)}]}{E[\tau]^2/\mathrm{Var}(\tau) + \rho[\tau_{x(ij)}, \tau_{y(kl)}]} \tag{6}$$

Where $\rho[\tau_{x(ij)}, \tau_{y(kl)}]$ is the correlation in coalescence time at loci x for chromosomes i and j, and locus y for chromosomes k and l. The key point about this result is that correlations in genealogical history will be most affected by recombination events in the history of the pair of loci, whereas the expectation and variance of coalescence times are strongly affected by demographic processes.

2.5 Empirical patterns of LD

A theoretical understanding of LD helps us to understand what the important influences are on LD, and how to interpret empirical patterns. From the point of view of the design of association mapping studies, the critical question, however, is what does LD look like in real populations?

For this reason, over the last few years there have been several large-scale surveys of LD in the human genome (e.g. Huttley et al., 1999; Abecasis et al., 2001; Reich et al., 2001). The major conclusion of these studies has been that in Caucasian populations, LD extent over large genomic regions, even up to the scale of 100kb. These finding are in direct contrast with an early, theoretical prediction based on a model of extreme population growth in the human population (Krugylak, 1999; Pritchard and Przeworski, 2001). The critical importance of this result is that in order to capture associations through genome-wide surveys of association, marker densities on the scale of 10s of kbs may be sufficient.

Other notable findings about LD relate to differences between populations. For example, it is fairly clear that LD in African populations tends to be lower than in Caucasians (e.g. Frisse et al., 2001). This finding agrees with the evidence for slightly higher genetic diversity in African populations, which has been taken as evidence for the out-of-Africa model. Under this scenario, African populations would reflect relatively stable populations of considerable age, from which smaller, founder populations arose to colonize the rest of the world.

In line with the idea that population size is an important determinant of LD, very small, isolated populations, exemplified by the Saami (Laan and Pääbo, 1997), tend to have high levels of LD. However, the extent of LD in the Saami is so high, on the order of cMs (Laan and Pääbo, 1997), that admixture events must also have contributed to the effect. Indeed, several populations, such as Finland and Sardinia, which were assumed to have high LD due to their relatively small size, and founder history, do not appear to have greatly increased LD (e.g. Eaves et al., 2000). In short, the complexity patterns of colonization and gene-flow has

been critical in determining current patterns of LD. Dissecting out the demographic history of populations is a major challenge to modern population genetics.

In the last couple of years a particular feature of LD in human populations has attracted much attention; the heterogeneous distribution of LD along chromosomes. Several studies (e.g. Taillon-Miller et al., 2000; Jeffreys et al., 2001; Daly et al., 2001; N. et al., 2001) have demonstrated that the genome is broken into blocks of strong haplotype structure, characterised by low haplotype diversity, strong associations between alleles and rare recombination, separated by shorter regions of shattered haplotype structure, characterised by high haplotype diversity, weak allelic associations and multiple recombination events. Although the data have been interpreted mainly in terms of variation in the recombination rate, with most recombination events occurring in hotspots, after the manner of the MHC locus citepJeffreysetal01, it is not yet clear whether the pattern is consistent across different populations. If so, it may be possible to use a much lower marker density that previously thought to capture associations between mutations within the blocks of strong haplotype structure.

2.6 Which population?

The question of which population to focus on for association mapping studies has attracted much debate. There are two opposing issues; marker densities and biological relevance. Given that high LD means associations between alleles can be found over considerable physical distances, populations with high LD are economically attractive for association studies. In addition, in small populations (e.g. the Saami), or those with relatively few founders (e.g. Iceland), the genetic basis of disease susceptibility is likely to be less complex than in large populations. In other words, small and/or founder populations may well have a low number of high frequency mutations contributing to disease susceptibility.

The downside of focusing on small, relatively isolated populations, is that in doing so, the biological relevance to the larger populations of interest may be com-

promised. Two factors complicate the biological picture. First, founder effects and genetic drift may generate genetic homogeneity within the mapping population, but the important alleles of large effect are not guaranteed to be the same in the mapping population and population of real interest. Second, mapping and target populations may have considerably different environments, so if G$\times$E effects are important in the disease phenotype, using a different population to map genes than the one you are ultimately interested in may mean that key effects are missed.

A related set of problems apply to the idea of using admixed populations for mapping. Admixture generates linkage disequilibrium even between unlinked markers, because of differences in allele frequency between the two source populations. Using a simple deterministic model for admixture, the coefficient of disequilibrium between alleles separated by a recombination fraction r, t generations after the admixture event is

$$D_{AB}(t) = \delta_A \delta_B x_1 x_2 (1-r)^t$$

Where δ is the difference in allele frequency between the source populations. The key point is that admixture generates long range LD, which can persist for several generations. Mapping by admixture linkage disequilibrium (MALD Chakraboty and Weiss, 1988; McKeigue, 1997) aims to make use of this demographic LD in order to reduce the density of markers required for detecting linkage. The paradigms for admixed populations are places such as Jamaica and South Africa, however most countries, e.g. the UK and even Iceland, show some evidence of admixture. Given the fluid nature of human settlement, as revealed by archaeology, we should not be surprised at the levels of population mixing indicated by genetic data.

There are several complexities for MALD. First, admixture generates spurious associations, which must be accounted for in assessing the significance of associations. The background level of association can be assessed from LD between alleles at unlinked markers (Pritchard and Rosenberg, 1999), and under certain conditions it may be possible to correct explicitly for admixture in the data (Pritchard et al., 2000). Second, the obvious target for MALD studies are disease that show

differences in frequency between the source populations. However, because the environments of the source populations differ, it may be more parsimonious to presume that differences in disease prevalence are the result of differences in environment. Finally, if genetic disease is considerably influenced by interactions between mutations, or partially recessive mutations, then admixed populations may actually show a reduction in the frequency of the disease phenotype, making the location of disease mutations problematic.

2.7 Big questions for association mapping

What are the prospects for association studies? There are a number of big questions

- Are common variants responsible for common disease? Or are complex, multifactorial diseases influenced by many rare mutations at many loci?
- Are single mutations at disease loci responsible for most variation? Or is allelic heterogeneity a serious problem.
- Does demographic LD aid association mapping? Or are the complexities introduced by admixture, structure, and environmental effects more hindrance than help?

And a number of empirical issues relating to the distribution of LD

- Is a marker spacing of Xkb (3-50kb) sufficient to capture associations? Or is a much finer map required to allow for the stochastic nature of associations?
- Can global haplotype diversity be captured by a few well-chosen markers? Or are population differences in block haplotype structure important?

Given the current attention focused on association mapping, and the economic interest associated with medical genetics, the answers to these questions will be known remarkably soon.

References

Abecasis G, Noguchi E, Heinzmann A, Traherne J, Bhattacharya S (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. 68:191–197.

Cardon L, Bell J (2001). Association study design for complex diseases. Nature Rev. Genet. 2:91–99.

Chakraboty R, Weiss KM (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. U.S.A. 85:9119–9123.

Chakravati A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH (1984). Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. 36:1239–1258.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001). High-resolution haplotype structure in the human genome. Nature Genet. 29:229–232.

Devlin B, Risch N (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, et al (2000). The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nature Genet. 25:320–323.

Fearnhead P, Donnelly PJ (2001). Estimating recombination rates from population genetic data. Genetics 159:1299–1318.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001). Gene conversion and difference population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69:831–843.

Griffiths RC, Marjoram P (1996). Ancestral inferences from samples of DNA sequences with recombination. J. Comput. Biol. 3:479–502.

Hedrick PW (1987). Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341.

Hey J, Wakeley J (1997). A coalescent estimator of the population recombination rate. Genetics 145:833–846.

Hill WG, Robertson AR (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hudson RR (1987). Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50:245–250.

Hudson RR (2001). Two-locus sampling distributions and their application. Genetics 159:1805–1817.

Huttley G, Smith M, Carrington M, O'Brien S (1999). A scan for linkage disequilibrium across the human genome. Genetics 152:1711–1722.

Jeffreys AJ, Kauppi L, Neumann R (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genet. 29:217–222.

Krugylak L (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. 22:139–144.

Kuhner MK, Yamato J, Felsenstein J (2000). Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393–1401.

Laan M, Pääbo S (1997). Demographic history and linkage disequilibrium in human populations. Nature Genet. 17:435–438.

Lewontin RC (1964). The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67.

McKeigue PM (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am. J. Hum. Genet. 60:188–196.

McVean G, Awadalla P, Fearnhead P (2002). A coalescent-based method for detecting and estimating recombination rates from gene sequences. Genetics in press.

McVean GAT (2002). A genealogical interpretation of linkage disequilibrium. unpublished.

N. P, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, et al (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Ohta T, Kimura M (1971). Linkage disequilibrium between two segregating nucleotide sites udner the steady flux of mutations in a finite population. Genetics 68:571–580.

Pritchard J, Przeworski M (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69:1–14.

Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 63:1839–1851.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000). Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, et al (2001). Linkage disequilibrium in the human genome. Nature 411:199–204.

Slatkin M (1994). Linkage disequilibrium in growing and stable populations. Genetics 137:331–336.

Spielman RS, McGinnis RE, Ewens WJ (1993). Transmission test for linkage disequilibrium: he insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–513.

Taillon-Miller P, Bauer-Sardina I, Saccone N, Putzel J, Laitinen T, Cao A, Kere J, et al (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nature Genet. 25:324–328.

Wall JD (2000). A comparison of estimators of the population recombination rate. Mol. Biol. Evol. 17:156–163.

Weir BS, Hill WG (1986). Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. 38:776–778.

Weiss KM, Clark AG (2002). Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 18:19–24.