The goal of a linkage study is to identify a set of loci or a chromosomal region that show linkage to a specific trait. Evidence of linkage, however, does not necessarily mean that we have located the functional gene that influences the trait but merely that we are likely to be in close proximity of the functional gene.

Clearly, it is desirable to uncover the genetic basis of a quantitative trait or a disease in order to use this information to for example increase milk production of cows, to produce effective treatments or merely to identify individuals with high risk of developing a disease.

Association mapping are based on the assumption that we have a set of genetic markers available and that the markers either represent actual genes (or variants within genes) that are of interest for the trait we are analyzing or that the markers are so close to the actual functional genes that they not only segregate together but that they also are in linkage disequilibrium. This implies that we can estimate the exact effect of the various alleles/genotypes at the trait locus, and — since we are examining the association between the trait and the actual observed genotypes — that we do not require pedigree data.

In human genetics it has been suggested that genome wide association studies are going to be more successful in locating genes for complex diseases than linkage studies. Two reasons for this is that the abundance of single nucleotide polymorphism markers (SNPs) in the human genome makes it possible to saturate the genome with genetic markers such that at least one marker should be in linkage disequilibrium with the disease gene and the fact that the cost of genotyping has dropped dramatically in recent years. Also, it is usually much cheaper to sample unrelated individuals required for association studies than to sample (larger) pedigrees for linkage studies (not to mention that human multi-generational pedigrees can be especially difficult to sample before the investigator dies!).

The basis of genetic association

Let us consider one way that genetic association may originate. When a (rare) disease mutation first appears it occurs at random in one of the haplotypes in the population. If we examine a marker close to the mutation locus and compare the disease subpopulation (i.e., the population of individuals carrying the mutation) to the non-carriers the carriers must have an increased frequency of the alleles found on the original disease haplotype. Recall from lecture 5, that this increased association between the mutation and a marker allele will break down over time due to recombination:

If \( P_{A_i,B_j}(t) \) is the frequency of gamete \( A_i, B_j \) \((A\) can be viewed as the mutation locus and \( B\) as the marker locus\) at generation \( t \) and \( c \) is the recombination fraction between \( A \) and \( B \) then we have that

\[
P_{A_i,B_j}(t + 1) = (1 - c)P_{A_i,B_j}(t) + cP_A P_B,
\]

where \( P_A \) and \( P_B \) are the marginal allele frequencies. Rearranging (1) we get (see formula (2.6) from lecture 2) that the disequilibrium

\[
D(t) = D(0)(1 - c)^t
\]

decreases exponentially in time. So it should be possible to locate causative genes even in old populations if we “look close enough” (i.e., have a marker allele close to the disease locus).

Association mapping

To test for association between a marker and a trait we can use standard statistical methods like one-way ANOVA (for Gaussian distributed quantitative traits) or logistic regression (for binary traits).

If the between group (i.e., between genotype) variance is significantly greater than zero for quantitative traits there is evidence for association between the marker and the trait. For logistic regression a likelihood ratio test can be used to test for no difference between genotypes. Both of these approaches have the advantage that the linear models can be extended to included additional risk factors.

Alternatively, a traditional case-control approach can be used to test for association to at binary trait, and this is corresponds to comparing the multinomial distribution of allele frequencies between the case
and the control populations. Significant differences in allele frequencies can be taken as evidence for association.

For example, the observed number of genotypes for a diallelic marker can be summarized as follows:

<table>
<thead>
<tr>
<th></th>
<th>MM</th>
<th>Mm</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
</tr>
<tr>
<td>Control</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
</tr>
</tbody>
</table>

The log-likelihood ratio test statistic becomes

$$\sum_{i=1}^{3} x_{ii} \log(x_{ii}/(x_{ii} + x_{i1} + x_{i2} + x_{i3})) + x_{1i} \log(x_{1i}/(x_{1i} + x_{12} + x_{13})) + x_{2i} \log(x_{2i}/(x_{2i} + x_{22} + x_{23}))$$

Obviously, standard hypothesis testing approaches can be used to test hypotheses successively in order to determine the possible mode of inheritance. Figure 1 shows an example where a case-control study is used to test for association and mode of inheritance of a diallelic marker. The figure shows a hypothesis testing diagram, where the hypotheses and estimated odds ratios are listed under each model and the test statistics are listed on each line. The three possible genotypes for this marker are denoted “wt” for wild type, “het” for heterozygous and “hom” for homozygous. A test for Hardy-Weinberg is found in the lower left corner.

We can see, that we reject the dominant model and that we fail to reject the both the codominant model and the recessive model. The null model (no effect of genotype on disease status) is rejected for both the codominant model and the recessive model suggesting that there is indeed an association between the disease status and the marker locus and that the mode-of-inheritance is not dominant.

Figure 1: Example of testing for association between diallelic marker and type 2 diabetes.

1The output is from WebAssotest found at www.ekstroem.com/assotest/assotest.html
Linkage disequilibrium revisited

Association between marker alleles and disease mutations in case-control studies are clearly related to the concept of population level associations as measured by linkage disequilibrium. Recall that alleles from two loci are in linkage equilibrium if the frequency of a gamete equals the product of the individual alleles constituting the gamete, i.e., if

\[ D_{AB} = P(AB) - P_A P_B = 0 \]

where \(P(AB), P_A\) and \(P_B\) are the gamete frequency, the marginal frequency of allele \(A\) at the first locus and the marginal frequency of allele \(B\) at the second locus, respectively.

So how do we quantify deviation from linkage equilibrium? Several different measures of linkage disequilibrium have been suggested and we will consider two (both defined for two diallelic loci).

**Correlation coefficient** The correlation coefficient is one of the more widely used measures of linkage disequilibrium. It can be calculated by scoring one allele as 0 and the other allele as 1 for each locus and then calculating the Pearson correlation coefficient

\[
\rho_{AB} = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^N (x_i - \bar{x})^2} \sum_i^N (y_i - \bar{y})^2}
\]

\[
= \frac{NP(AB) - N P_A P_B}{\sqrt{(NP_A(1 - P_A))(NP_B(1 - P_B))}} = \frac{D_{AB}}{\sqrt{(P_A(1 - P_A))(P_B(1 - P_B))}}
\]

where we have replaced \(x_i\) with the indicator function that we observed the ‘1’-allele at the first locus and \(y_i\) is the indicator function for the second locus.

The correlation coefficient takes values from -1 to 1 with -1 and 1 indicating complete linkage disequilibrium and 0 meaning linkage equilibrium. Since the measure is symmetric around 0 it is usually squared to give just a single value ‘1’ to represent complete linkage disequilibrium. The squared correlation coefficient is often denoted \(\Delta^2\) or \(r^2\) and has the usual nice interpretation as the proportion of the variance at one locus that is explained by the alleles at the other locus.

**Scaled disequilibrium** The scaled linkage disequilibrium measure is defined as

\[
|D'| = \begin{cases} 
\min(P_A(1 - P_A), P_B(1 - P_A)) & \text{if } D_{AB} < 0 \\
\min(P_A P_B, (1 - P_A)(1 - P_B)) & \text{if } D_{AB} > 0 
\end{cases}
\]

and has a minimum of 0 (when \(A\) and \(B\) are uncorrelated) and a maximum of 1. \(|D'|\) indicates the degree to which gametes exhibit the maximum potential disequilibrium for a given set of allele frequencies.

From a mapping standpoint, \(|D'|\) is said to be preferable to \(r^2\) because differences in \(|D'|\) more closely reflect the extent of recombination. However, \(|D'|\) has two drawbacks: \(|D'|\) will always be one when there are only three of the four possible gametes present for a pair of diallelic loci, and \(|D'|\) is very unstable for small sample sizes.

**Problems with association studies**

One of the problems with the above tests for association is that several factors can cause association without linkage being present. One example is unmatched (or badly-matched) case-control studies where disease status is influenced by a covariate that is not properly accounted for in the matching (e.g., a longevity-related gene may appear to be associated with any late-onset disease because individuals simply have to live longer before they the disease). Other cases may be inbreeding, genetic drift or selection. The most quoted example, however, is that of population admixture where two or more sub-populations make up the total population. If two or more subpopulations exists within a population and the allele frequencies differ between population then there will be evidence for association even though each sub-population shows no sign of association.
Figure 2: Effect of population stratification on an association study of a quantitative trait. Two subpopulations constitute the population and while the mean (dots) is constant for each genotype within a population, the means are different for the two populations. If the genotype frequencies differ between subpopulations (shown by different sizes of standard errors) and their overall quantitative trait level is different (for example due to environmental factors like food intake, smoking or geographical location etc.) the total population will show association between genotype and average trait level. Association can also be found if there is significant variance heterogeneity for the six populations and genotypes.

For example, consider a diallelic marker and a population consisting of two subpopulations found in proportions 1/10 and 9/10. Assume that we want to test for association between marker alleles $A_1$ and $A_2$ and disease status, and that the frequencies within the subpopulation are as follows:

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>1/9</td>
<td>2/9</td>
<td>Case</td>
<td>1/121</td>
<td>10/121</td>
</tr>
<tr>
<td>Control</td>
<td>2/9</td>
<td>4/9</td>
<td>Control</td>
<td>10/121</td>
<td>100/121</td>
</tr>
</tbody>
</table>

It is clear that no association exists in either of the populations (OR is 1 in both populations) but the combined population will have a status/marker association that is

\[
\begin{array}{c|cc|c|cc}
 & A_1 & A_2 & & A_1 & A_2 \\
\hline
\text{Case} & 101/5445 & 526/5445 & & 526/5445 & 4292/5445 \\
\text{Control} & 526/5445 & 4292/5445 & & & \\
\end{array}
\]

which has an OR of 1.567. Any level of significance can be reached if the difference in allele frequencies between subpopulations is large enough! The same argument can be used for quantitative traits as figure 2 shows.

The problem in the example is caused by confounding but other factors may seriously influence the power of association studies. One such problem is allelic heterogeneity where multiple susceptibility alleles exist at a disease locus. Obviously, allelic heterogeneity is in direct contrast to the basic idea behind association analysis and can make it impossible to identify the influential genes.

Other issues that may reduce the usefulness of association studies include multiple contributory loci, low penetrance and environmental effects.
Transmission/disequilibrium-test

The transmission/disequilibrium-test (TDT) was “invented”\(^2\) to address the problem with population admixture for binary traits. The standard TDT design requires trio families (two parents and an affected offspring). The family is genotyped at a locus and if that locus does not influence the disease status, each allele from a parent should have equal probability of being transmitted to the offspring (see notes for lecture 5 for formulas and an example).

There are several factors that reduce the power and efficiency of the TDT. First, it requires sampling and genotyping of three individuals to get a data point. Second, obtaining parental genotypes can be difficult, particularly for late-onset diseases such as Alzheimers. Finally, in order to be informative at a locus, parents need to be heterozygous at the locus. Although efforts can be made to use loci with high heterozygosity, a significant fraction of trios will always be uninformative. Some of these problems can be reduced by using alternative TDT designs with siblings rather than parents.

The original TDT was designed for diallelic markers. Markers with more than one allele are usually analyzed by lumping together all alleles but one, i.e., doing a TDT test with allele \(X\) vs. non-\(X\). While multi-allele markers are more informative than diallelic markers this approach increases the number of tests and give rise to at multiple testing problem.

Common disease — common variants or rare alleles?

One of the key assumption for association mapping of complex diseases is the common disease common variant hypothesis postulating that common diseases are influenced by a few loci, each contributing to the disease with a single major allele. Each allele may have low penetrance but if enough data are sampled association should be detected.

The reason for this assumption — that may or may not be true — is that if many rare alleles influence disease status we may end up with some of the problems rendering association studies useless. If there are many rare alleles, we have strong locus heterogeneity, and we will observe allelic heterogeneity (some of the individuals with disease do not have the disease mutation).

Arguments for the CDCV hypothesis in humans are that humans can be traced back to a small and relatively recent founder population and that most of the variation dates back to this founder population. If a disease allele that has nothing to do with natural selection (as is the case for most of the “popular” complex diseases) and if it was common in the founder population, it takes a many generations before it is diluted out of by new alleles generated by population growth.

Another reason for the attractiveness of the CDCV hypothesis is that if a disease is indeed caused by a common variant (with an allele frequency of, say, >15%) it will have much bigger clinical impact on the disease. If there are many rare alleles influencing a trait and we find one, we are only able to explain a small fraction of the observed cases.

Pritchard (2001) tried various models to simulate human evolution and found that at a random susceptibility locus the alleles are quite likely to have a frequency near 0 or 1 (suggesting that rare alleles might cause diseases). For any given disease, however, the disease susceptibility loci that make any substantial contribution to the genetic variance are likely to be more polymorphic, i.e., have allele frequencies that are more common. Although this is hardly conclusive evidence for the CDCV hypothesis it does suggest that the CDCV hypothesis may not be too unrealistic.

Haplotype tests

Until now we have considered association mapping where association between a single marker and a trait was examined. If we have several hundred or thousand markers (as is not unrealistic with current genotyping technologies) there will obviously be a multiple testing problem if we make hundreds or thousands of association tests.

Another problem that we briefly touched upon in the beginning is that a marker used for association mapping is not necessarily the disease-causing polymorphism but it may be in linkage disequilibrium with the causal locus. Instead of trying to find association between a single marker and a trait it may be easier to find association to a small chromosomal region or to a more polymorphic marker.

\(^2\)Actually, the TDT is just the well-known McNemar’s test.
Haplotype analysis may be one way to reduce the multiple testing problem and improve the information from non-causal markers. In haplotype analysis, two or more closely linked markers (typically SNPs that are abundant in the human genome) are combined to provide more information about the chromosomal region and to create a (potentially) more polymorphic marker. Combining information from several polymorphisms in the same gene not only has a statistical advantage: if several variants within a single gene is genotyped any interaction between polymorphisms (e.g., regulation) can be addressed by considering haplotypes instead of single polymorphisms.

If we start by assuming that we instead of alleles at a locus have information about haplotypes from a small chromosomal region, there are three approaches that are widely used for analysis of haplotype data, and we will only touch upon them briefly:

1. Haplotypes are treated in the same way as alleles for a single locus. We used the methods discussed above and use the haplotypes as multi-allelic markers.

2. Test for tendency that “similar” haplotypes will have “similar” effects on disease susceptibility. One such method for case-control studies is due to Cuzick and Edwards (1990). They construct a test statistic, \( T_k \) as the sum over all cases of the number of \( i \)'th \( k \) neighbors that are also cases,

\[
T_k = \sum_{i=1}^{N} I(i \text{ is a case})I(k \text{'th nearest neighbor to } i \text{ is a case}).
\]

In order to define neighbors a similarity model is used to construct a similarity matrix that shows the similarity or “distance” between two haplotypes.

When cases are clustered the nearest neighbor to a case will tend to be another case and \( T_k \) will be large and \( T_k \) will be zero when all of the cases have only controls as nearest neighbors.

Under the null hypothesis of no spatial clustering the neighbors of a case should be a random sample of the remaining cases and controls. The expected value of the test statistic under the null hypothesis is

\[
E(T_k) = N P(\text{case})E(X|\text{case}) = N \frac{n}{N} \frac{n-1}{N-1}
\]

where \( n \) is the number of cases, \( N \) is the total number of observations and where \( X \) is a stochastic variable that — conditional on an observation being a case — follows a hypergeometric distribution with parameters \( k \), \( n \) and \( N \).

Large values of the test statistic, \( T_k \) are considered significant and indicate that there is spatial clustering, i.e., that disease will cluster around a specific set of haplotypes. The variance for the test statistic, \( T_k \) can be found in Cuzick and Edwards (1990).

3. Use a log-linear model approach to take account of the haplotype structure. With a log-linear model we model the simultaneous distribution of the markers, the disease and their association. If \( M_1, \ldots, M_K \), represent \( K \) genetic markers and \( D \) is disease status a model with \( D \) and \( M_1 \times \cdots \times M_K \) corresponds to the situation where there is a parameter for each haplotype but no association to disease. The model corresponding to 1) above is \( D \times M_1 \times \cdots \times M_K \) where each haplotype has its own relative risk parameter. More refined models can be applied, for example a Markov-type model with disequilibrium between neighboring markers but no association to disease would include model terms \( D, M_1 \times M_2, M_2 \times M_3, \ldots, M_{K-1} \times M_K \). The problem with the log-linear approach, obviously, is that the optimal model may not be known.

Before haplotype analysis can be used we need to get the haplotypes. Not only do we need the genotypes at each loci but we also need to determine the phase of the genotypes. With some designs, e.g., \( F_2 \)- or back crosses, it is easy to determine haplotypes, and for some pedigree it is possible to infer the haplotypes if the parents or siblings are sufficiently informative, but association mapping studies relies on samples on independent sampling units so these methods of inferring haplotypes can not be applied.
Reconstructing haplotypes

We need to reconstruct haplotypes from a sample of \( N \) unrelated sampling units that are genotypes at \( K \) marker loci. If we knew the frequencies of the haplotypes in the population we would assign the set of haplotypes to an individual that would make his observed genotypes most likely.

As an example consider two diallelic loci in Hardy-Weinberg equilibrium with population haplotype frequencies

\[
\begin{align*}
P(AB) &= .5 \quad P(\text{ab}) = .2 \\
P(aB) &= .1 \quad P(ab) = .2
\end{align*}
\]

If we observe an individual with genotypes \( AaBB \) we know — since that person is homozygous at locus 2 — that the haplotypes must be \( AB|\text{ab} \). An individual with genotypes \( AaBb \) could either be \( AB|ab \) or \( Ab|aB \). The probability of observing an \( AB|ab \) individual in the population is \( 2 \times .5 \times .2 = .2 \) while an \( Ab|aB \) individual has probability \( 2 \times .1 \times .2 = .04 \). In this case we would estimate the individual to have haplotypes \( AB|ab \).

All we need is therefore estimates of the population haplotype frequencies. These can be estimated by the EM algorithm where the actual haplotype configuration plays the role of the “missing” part in the algorithm. Let us again consider the case with two diallelic loci genotypes for \( N \) individuals and assume that both marker loci are in Hardy-Weinberg equilibrium. The number of observed individuals with different combinations of genotypes at the two loci can be summarized in a three by three contingency table.

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>( n_{13} )</td>
</tr>
<tr>
<td>Aa</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>( n_{23} )</td>
</tr>
<tr>
<td>aa</td>
<td>( n_{31} )</td>
<td>( n_{32} )</td>
<td>( n_{33} )</td>
</tr>
</tbody>
</table>

Note that the \( n_{22} \) individuals with genotypes \( AaBb \) are the only ones where we cannot determine the phase. In the expectation step of the EM algorithm we need to assign these \( n_{22} \) individuals to either the \( AB|ab \) group or to the \( Ab|aB \) group. This assignment is done according to the current estimate of the haplotype frequencies. The EM algorithm can be summarized as follows:

1. Start by calculating the proportion of the \( n_{22} \) individuals that have haplotypes \( AB|ab \) given the current haplotype estimates:

\[
X = \frac{P^k(AB)P^k(ab)}{P^k(AB)P^k(ab) + P^k(\text{Ab})P^k(aB)}
\]

2. Calculate new estimates of the haplotype frequencies by simple counting of the number of observed haplotypes, e.g.,

\[
P(AB) = \frac{2n_{11} + n_{12} + n_{24} + Xn_{22}}{2N}
\]

and likewise for \( P(aB), P(\text{Ab}) \) and \( P(ab) \).

Iterate back and forth between the two steps in the EM-algorithm to obtain maximum likelihood estimates of the population haplotype frequencies.

With more than a few loci the contingency table quickly becomes very sparse which makes it possible to get good estimates of the population haplotype frequencies. Also the number of parameters increases rapidly with increasing number of marker loci so much more than a few marker at a time can become computational difficult to handle.

As a concluding note, it should be mentioned that while both the haplotype frequencies and the most likely haplotypes are estimated and therefore prone to error it is very common to see the estimated haplotypes used for analysis as if they were determined without error. This is clearly an area where more research is needed.

DNA pooling

Although the cost of genotyping has dropped dramatically it can still be costly to genotype a large number of markers. Since association studies look for differences in allele frequencies between groups
of individuals (e.g., cases and controls) molecular biologists have constructed a quick way to get a rough estimate of the allele frequencies for a group of individuals.

DNA pooling relies on a method to detect the relative strength of alleles in a DNA sample. This differentiation between signal strength can be achieved through cleavage of the SNP to generate PCR product of differing sizes or through fluorescence-tagged primers that are specific to the various alleles.

In a two-pool design, equal amounts of DNA is taken from each individual and the DNA from the cases are combined to create one pool while the DNA from the controls form another pool (for quantitative traits two groups can be created by considering individuals with trait values at the two extremes of the quantitative trait). Since we can use DNA pooling techniques to get an estimate of the relative frequencies of the alleles in each of the two groups we only have to do two lab experiments — one for the pool of cases and one for the controls — which reduces the genotyping considerably. Also, we can include any number of individuals in the two pools — the amount of work to estimate allele frequencies for 100 individuals is the same as for 1000 individuals.

Statistical comparison of allele frequencies can be done by considering the difference in allele frequencies relative to the standard error of the difference under the null,

\[ Z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}(\hat{p})} \]

However, since the variance is likely to be influenced by factors other than sampling variation such as experimental errors (e.g., adding unequal amount of DNA or inaccuracies in determining allele frequencies) that will result in measurement errors, the standard error under the null is often inflated to be

\[ \text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})(1 + \tau)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + 2\varepsilon} \]

where \( \tau \geq 0 \) is the variance caused by unequal amounts of DNA included for each individual and \( \varepsilon \geq 0 \) is the variance related to inaccuracies in allele frequency estimation. The order of magnitude for \( \tau \) and \( \varepsilon \) need to be determined through additional experiments and cannot be identified from the pooled study itself.

The power of pooled DNA analysis depends on the expected effect size, which in turn depends on the difference in allele frequencies. Small relative risks result in smaller allele frequency differences and will therefore be more difficult to detect. Part of the variation, \( \hat{p}(1 - \hat{p})(1 + \tau)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \) will decrease with increasing sample sizes but the measurement variation related to allele frequency estimation, \( \varepsilon \) will remain constant with increasing sample size. Small measurement errors can therefore substantially reduce the power to detect risk alleles of modest effect.

Pooled DNA analysis can be less powerful than individual genotype analysis simply because it is impossible to correct for environmental risk factors (e.g., sex, age, BMI, or smoking status) that may influence the trait. This problem can be overcome by placing individuals in subgroups that have similar environmental risk factors. However, this approach will increase the number of pools and hence the number of experiments and will increase the cost of doing pooled DNA samples.

**Empirical patterns of linkage disequilibrium**

Over the last few years there have been several large-scale studies to determine linkage disequilibrium in the human genome (e.g. Huttley et al. (1999); Abecasis et al. (2001); Reich et al. (2001). These empirical studies found that LD extended over large genomic regions up to the scale of 100kb. This is interesting from an association mapping viewpoint as the extension of LD can be used to determine the marker density required for genome-wide association mapping. In this case, a marker every 10-50 kilobases may be sufficient for association mapping.

Even more interesting are recent findings that there is substantial differences in the distribution of LD along the chromosomes. Several studies (for example Daly et al. (2001)) have demonstrated that the human genome consists of blocks of strong haplotype structure (denoted “deserts” and characterized by low haplotype diversity, strong associations between alleles and rare recombination), separated by shorter regions of high haplotype diversity (characterized by weak allelic associations and multiple recombination events and denoted “jungles”). The identification of haplotype blocks means that it is not
necessary to genotype all markers but that it should be sufficient to genotype a few markers to determine
the haplotypes in the deserts while more markers should be genotyped in the jungle regions where allelic
association is low.
References


