Sarah Ennis
is undertaking a PhD in the Genetic Epidemiology and Bioinformatics Research Group at Southampton University.

Nikolas Maniatis
is involved in post-doctoral research in the Genetic Epidemiology and Bioinformatics Research Group at Southampton University.

Andrew Collins
is head of the Genetic Epidemiology and Bioinformatics Research Group at Southampton University. The group's interests include genetic and linkage disequilibrium map integration, mapping of genes involved in complex traits and genetic epidemiology of the fragile X region.

Keywords: *linkage disequilibrium, haplotype and diplotype analysis, disease gene mapping, Malecot model*

Andrew Collins,
Genetic Epidemiology,
Human Genetics,
Duthie Building (808),
Southampton General Hospital,
Tremona Road,
Southampton, SO16 6YD,
UK

Tel: +44 (0) 23 8079 6939
Fax: +44 (0) 23 8079 4264
E-mail: arc@soton.ac.uk

Allelic association and disease mapping

Sarah Ennis, Nikolas Maniatis and Andrew Collins

Date received (in revised form): 3rd September 2001

Abstract

The application of allelic association to map genes for complex traits, particularly using high-density maps of single nucleotide polymorphisms in candidate regions, is an area of very active research. Here we present some aspects of the methodology and applications to both major gene mapping, which illustrates the effectiveness of the method, and oligogenes, where methods are still in flux and for which there have been relatively few successes to date. Several important considerations emerge, including the selection of the optimal metric for measuring association and the importance of modelling the decline in association with distance given the variability in association in a candidate region. The Malecot model of association with distance is shown to have a resolution of greater than 50 kilobases but the available evidence suggests that considerably higher resolution might be achieved with dense single nucleotide polymorphism (SNP) maps.

INTRODUCTION

One of the outstanding success stories of the 1980s and 1990s was the linkage mapping of disease genes showing a Mendelian pattern of inheritance which have a large phenotypic effect (major genes). Many such genes have been mapped by positional cloning exploiting linkage analysis to localise a gene to a region of several megabases (Mb) and then employing a variety of techniques to isolate and clone the gene. However this step of refining a candidate region defined by linkage is often time-consuming and tedious because of many possible causal sites in a (still large) genomic region. Allelic association, also called linkage disequilibrium (LD) mapping, offers a strategy for reducing the target region and has been applied, with some success, to major genes. Furthermore this approach is considered to be particularly useful for localising genes for common diseases (oligogenes) which show non-Mendelian patterns of inheritance and have small individual phenotypic effects. Linkage disequilibrium is the non-random association of alleles at linked loci. Linkage mapping relies on the co-inheritance of adjacent disease and marker alleles over a small number of generations within families/pedigrees. By contrast linkage disequilibrium relies on the retention of association between specific alleles over many generations. When a disease mutation first appears in a population it is located on a single haplotype with one set of linked marker alleles. At this point there is complete disequilibrium between the markers and the disease mutation, ie the disease mutation is found only in the presence of a specific set of marker alleles. Over generations recombination occurs between the disease and marker alleles and disequilibrium is gradually lost. The rate at which this declines is a function of the recombination frequency but is also subject to other influences, such as mutation, selection and drift. For mapping a disease, the hope is that recombination dominates these other forces, therefore markers close to the disease gene exhibit higher levels of disequilibrium than those further away. As there are relatively few opportunities for recombination to occur in pedigrees, disease gene regions identified by linkage are often large, even in the most ideal case spanning at least 1 Mb, and encompassing hundreds of genes. In contrast association exploits larger numbers of recombination events which may give small disease associated regions (50 kilobases, Kb, or less).

Current interest is focused on mapping genes for common diseases and the favoured strategy is to exploit LD in maps of single nucleotide polymorphisms (SNPs), which are very abundant single base variants found throughout the genome. However, difficulties with this strategy arise through the variability of LD in different populations and chromosome regions. The variability is introduced by, among others, mutation and drift (random variations in allele frequencies over generations), ethnic differences and population admixture (population subgroups which have different allele frequencies). Recent evidence also suggests that gene conversion may play a significant role in the breakdown of LD over short genomic distances.[1] For these reasons LD must be carefully modelled within a candidate region. Strategies for mapping both major genes and oligogenes by allelic association including study designs, different association metrics and an approach to modelling association with distance are reviewed here. Important issues in gene mapping such as the extent of LD in the genome, the use of haplotypes or diplotypes and development of an LD map are also considered.

Variability of LD in different populations and chromosome regions

Efficiency of TDT trios is substantially less than a sample-based design

CASE-CONTROL AND FAMILY-BASED SAMPLING

A great deal of the methodology for association studies has focused on family-based as opposed to sample-based or case-control designs. The main reason for this has been concern over the problems posed by population stratification in case-control sampling. It is thought that population subgroups represented within the sample will have different underlying disease and marker allele frequencies and that unless a control population is perfectly matched to the case population, spurious association may arise. The most common approach is the transmission disequilibrium test (TDT)[2] which has many derivatives, including extension to quantitative traits.[3,4] The TDT as originally developed ‘considers parents who are heterozygous for an allele associated with disease and evaluates the frequency with which the allele or its alternate is transmitted to affected offspring’.[2] By focusing on only heterozygous parental genotypes it provides a test of association between linked loci and therefore avoids spurious associations between unlinked loci in the presence of population stratification. The test does, however, have a number of drawbacks. There is a loss of genotypic data through exclusion of homozygous parents and greater cost and difficulty when sampling family material. For late onset diseases determining parental genotypes may not be possible. Morton and Collins[5] showed that because of these issues the efficiency of TDT trios (two parents and an affected child) is substantially less than a sample-based design. Compared to a design with normal controls the efficiency of the TDT is 2/3 and is further reduced to 1/6 when using a case-control design with hypernormal controls (for example, where cases are affected individuals with early onset and controls are elderly unaffected individuals).

Problems posed by population stratification

The case-control design (or a sample of unrelated individuals with a quantitative trait) clearly has a number of advantages for examining candidate region association with disease but the issue of spurious association remains. Several approaches to detect stratification, by looking at associations between unlinked markers in a sample have been developed[6] and methods to correct for it have also been proposed. It could be argued that admixture problems are countered by careful selection of controls and that there is little evidence in the literature of analyses with admixture problems. Perhaps more important than these concerns is the poor understanding of patterns of LD within large candidate regions, a reliance on association with a small number of markers where the pattern of LD in a region cannot be elucidated, and large numbers of markers where many tests can generate false positives by chance. It is often assumed that LD between marker and disease declines with distance but in reality this is an oversimplification. Given the complex evolutionary history of haplotypes there may be islands of LD separated by regions with relatively little LD. The pattern of LD is further complicated by having polymorphic markers of different duration and allele frequencies with different mutation rates. For this reason it is essential to model the pattern of LD in a candidate region to separate 'signal' from 'noise'.

There may be islands of LD separated by regions with relatively little LD

MODELLING ASSOCIATION

Devlin *et al.*[7] modelled linkage disequilibrium using composite likelihood. Composite likelihood estimators are useful in situations where the full likelihood is very difficult to specify. Such models are generally consistent in that the parameter estimates converge to the true parameter estimates in large samples. A feature of such models is that the individual log-likelihood terms in the summation, in this case representing pairwise disequilibrium measures between pairs of markers or between marker and disease, need not be independent. Collins and Morton[8] developed a composite likelihood model to represent both disease by marker association and also marker by marker association in a region.[9] Association is defined in terms of a 2 × 2 haplotype table (Table 1). For major genes, disease and normal haplotypes may be used to model association with distance. In the case of biallelic markers and disease haplotypes, the cells of the table give counts for a given marker allele where '*a*' represents disease haplotypes with the allele, '*b*' gives disease haplotypes without the allele and '*c*' and '*d*' represent the corresponding normal haplotypes. When a disease mutation first arises the frequency of disease haplotypes with the associated allele is given by Q which is also the disease allele frequency. At this point there is complete disequilibrium as the frequency of haplotypes with both the disease allele and the other (non-associated) marker allele is 0. This is before the processes that reduce association, principally recombination, have begun to operate. In contrast, at equilibrium, the haplotype frequencies are simply the products of the individual

Composite likelihood is useful when the full likelihood is difficult to specify

Table 1: Haplotype frequencies by population

Disease	Population	Marker		Total
		Disease-associated allele +	Non-associated allele −	
Disease allele	Founders	Q	0	Q
+	Equilibrium	QR	$Q\,(1-R)$	Q
	Cohort	$Q\rho + QR(1-\rho)$	$(1-\rho)\,Q\,(1-R)$	Q
	Case-control	$\pi_{11} = \omega Q[\rho + R(1-\rho)]/\Sigma$	$\pi_{12} = \omega(1-\rho)Q(1-R)/\Sigma$	$\omega Q/\Sigma$
	Observed (counts)	a	b	
Normal allele	Founders	$R-Q$	$1-R$	$1-Q$
−	Equilibrium	$R(1-Q)$	$(1-R)\,(1-Q)$	$1-Q$
	Cohort	$(\mathrm{R}-Q)\rho + R(1-Q)(1-\rho)$	$(1-R)[\rho + (1-Q)(1-\rho)]$	$1-Q$
	Case-control	$\pi_{21} = [(R-Q)\rho + R(1-Q)(1-\rho)]/\Sigma$	$\pi_{22} = (1-R)[\rho + (1-Q)(1-\rho)]/\Sigma$	$(1-Q)/\Sigma$
	Observed (counts)	c	d	
Total (except case-control)		R	$1-R$	1
				$\Sigma = 1 + (\omega - 1)Q$

For the *i*th marker locus all parameters are subscripted by i $\ln lk = a \ln \pi_{11} + b \ln \pi_{12} + c \ln \pi_{21} + d \ln \pi_{22}$; Q = disease gene frequency; R = disease-associated marker allele frequency; ω = sample enrichment factor; ρ = association.

Malecot model describes association as a function of distance

allele frequencies QR, $Q(1-R)$, $R(1-Q)$, $(1-R)(1-Q)$ for the four cells respectively. The situation that is usually observed between tightly linked loci, however, is neither complete disequilibrium or complete equilibrium but an intermediate stage at which disequilibrium is characterised by ρ, the association metric. Haplotype frequencies as a function of ρ in both cohort and case control samples are given in Table 1.

For major genes, cases are much enriched, hence an adjustment to rescale Q is given by ω, the sample enrichment factor which is the ratio of the number of cases to controls divided by the ratio of disease frequency to normal in the population of haplotypes. The 2 × 2 table is rearranged such that $ad > bc$ and $Q < R$ which ensures that $0 \leqslant \rho \leqslant 1$. The ith pair of marker by disease measures of association, ρ, has information K_i and these are entered into composite log likelihood (lk) as $\ln lk = -\sum K_i(\rho_i - \hat{\rho}_i)^2/2$, where $\hat{\rho}_i$ are the fitted values in the Malecot model, which describes association as a function of distance. The model is defined as $\hat{\rho}_i = (1-L)Me^{-\varepsilon d_i} + L$, where d_i is the distance between a disease and marker or between a pair of markers. $M = 1$ indicates monophyletic origin, a single founding haplotype for the disease. Values less than 1 suggest polyphyletic origin, frequently seen when modelling association between pairs of markers. The L parameter represents association between unlinked markers, perhaps due to admixture. It also models the bias due to the constraint $\rho_i \geqslant 0$. Preliminary results indicate that L may be increased in small samples. The ε parameter represents recombination and the number of generations during which the haplotypes have been approaching equilibrium.

Table 2: Measures of allelic association (ψ) under H_0

Definition	Symbol	Estimate $\hat{\Psi}$	Synonyms
Covariance	D	$D = \pi_{11}\pi_{22} - \pi_{12}\pi_{21}$	LD[10]
Association	ρ	$\lvert D\rvert/Q(1-R)$	
Correlation	r	$\lvert D\rvert/\sqrt{Q(1-Q)R(1-R)}$	Δ[11]
Regression	b	$\lvert D\rvert/R(1-R)$	β[4]
Frequency difference	f	$\lvert D\rvert/Q(1-Q)$	D[12]
Delta	δ	$\lvert D\rvert/Q(1-R-Q+RQ)$	*P* excess[13]
Yule	y	$\lvert D\rvert/[2Q(1-Q)R(1-R)]$	*Q*[14]

Figure 1 gives a graphical representation of the model. For a given number of generations, and by

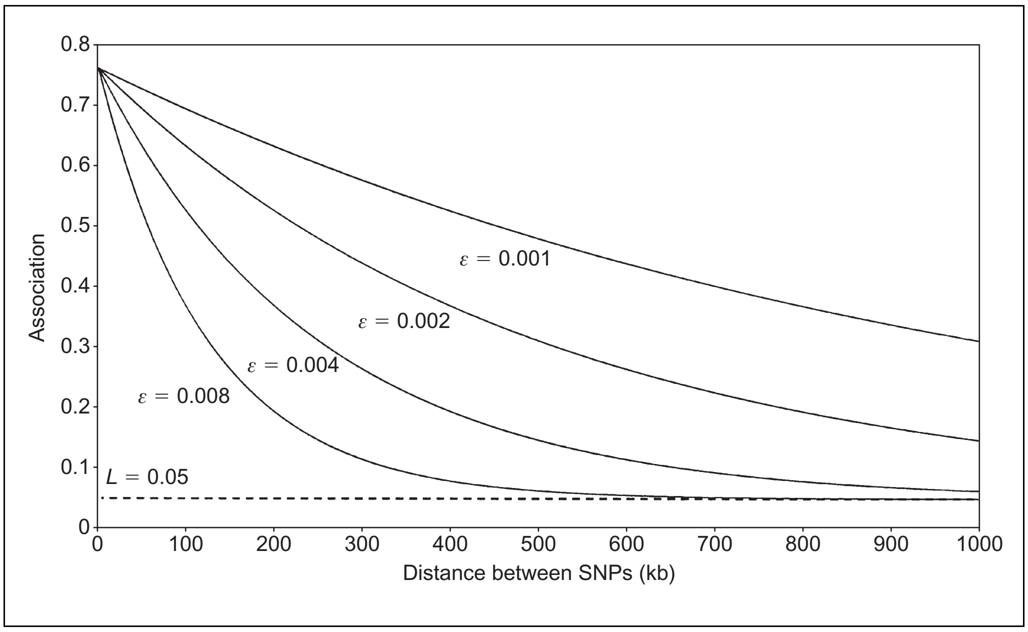

Figure 1: The Malecot model for $M = 0.75$, $L = 0.05$ and a range of values for ε

representing distance on the physical scale, high values of ε indicate a rapid decline in linkage disequilibrium, meaning LD can only be usefully detected and exploited in very close proximity to the disease gene. This might be particularly true in regions of the genome where there is a high recombination rate. Conversely, low values of ε imply that LD extends to large distances. A useful measure of the extent of LD has been termed the 'swept radius' and is defined as $1/\varepsilon$, the distance at which LD declines to $e^{-1} = 0.37$ of its original value. This allows comparison of the extent of LD in different genomic regions and populations.

The swept radius is a useful measure of the extent of LD

ASSOCIATION METRICS

Discussion thus far has assumed the adoption of the ρ metric as the measure of allelic association. However, there exists a veritable catalogue of metrics which may be applied,[10] all of which are functions of the covariance D (Table 2), which is expressed in terms of the four haplotype frequencies (π_{11}, π_{12}, π_{21} and π_{22}) in a 2×2 table. None of these metrics can boast independence from allele frequencies, but some, and ρ in particular, show reduced dependence and exhibit superior efficiency when the pattern of association is modelled. The literature on this matter is confused, and researchers arbitrarily apply their local metric of choice with little or no explanation of either the metric or the reason for its selection. This inconsistent approach to measuring allelic association has meant that results from different groups are incomparable, even for the same genomic regions, thus confounding the confusion.

There are many association metrics; ρ is the most efficient

In an effort to address this problem Morton *et al.*[16] determined the metric of best fit to the Malecot model. This model is justified on population genetics and theoretical grounds. LD was assessed in a number of large haplotype samples using each of the metrics in Table 2. The relative efficiency of each metric was then calculated based on the residual sums of squares after fitting the Malecot model. In each sample, ρ was consistent in exhibiting the highest relative efficiency and the lowest error variance. The best of the other metrics showed a loss of over 20 per cent of the information, and others showed losses of up to 70 per cent (Figure 2).

A current goal in genetic research is the development of an LD map of the entire genome. Such a map will trace

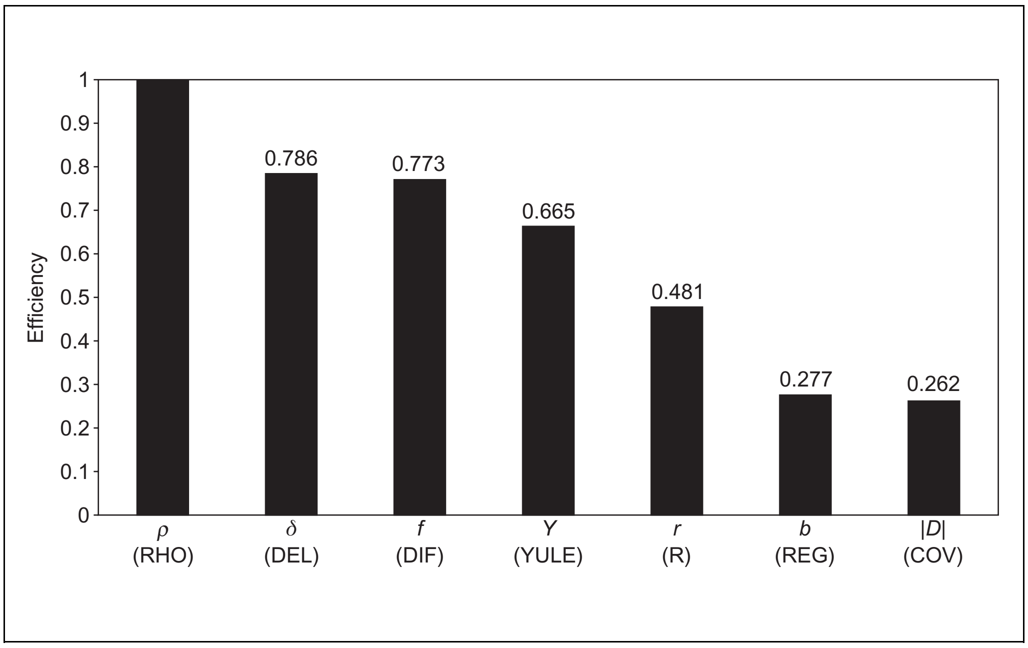

Figure 2: Efficiency relative to association (ρ) under H_0 (from residual sums of squares after fitting the Malecot model). Values shown are the combined results over all eight samples. Efficiency relative to ρ is extremely low for D (absolute value used for analysis) and b and intermediate for r. It should be noted that although the order of the relative efficiencies varied between samples, ρ performed consistently throughout all samples in giving the best fit to the model

recombination across the genome, providing core information regarding marker density required for mapping disease genes, and will pinpoint regions of extensive LD attributable to selective sweeps. The expedient building of such a map requires an integrated approach that facilitates comparison and verification from all research groups.

THE EXTENT OF LD IN THE GENOME

Knowledge of the extent of LD is essential to mapping disease genes

An understanding of the extent of LD in different genomic regions, and indeed populations, is essential to the development of strategies to map disease genes in the newly emerging dense maps of SNPs. This topic has therefore received a good deal of attention. [9,17–19] Lonjou *et al.*[20] were among the first to suggest that LD might be rather extensive in SNP maps and that this might be true of both isolated and large populations. From the extensive data available on biallelic polymorphisms in the Rh blood group and the MNS systems the authors found detectable LD out to at least 0.16 cM and little variation in linkage disequilibrium among isolated and large populations, with the exception of sub-Saharan Africa where LD was less extensive. However a simulation study by Kruglyak[17] suggested that a useful level of LD would be unlikely to extend beyond an average distance of 3 Kb in the general population. Subsequent studies in SNP maps have shown this to be a considerable underestimate. Figure 3 shows the swept radius for both large populations (UK, USA) and more isolated populations (Finland, Sardinia) based on data from chromosome Xq[21] and chromosome 18.[22] Disequilibrium is found to extend to between 385 and 893 Kb, with little distinction between the two sets of populations. This finding of extensive LD has been confirmed recently in a large-scale study of 19 chromosome regions.[18]

The extent of LD in the genome seems surprising given the likely duration of single nucleotide polymorphisms which may include many polymorphisms contributing to common diseases. Presumably the observed extent reflects population bottlenecks when populations became so small that a relatively small number of founding haplotypes were available to give rise to most of the haplotypes seen today.

MAPPING MAJOR GENES BY ALLELIC ASSOCIATION

There exist many methods for gene mapping which exploit LD. The more

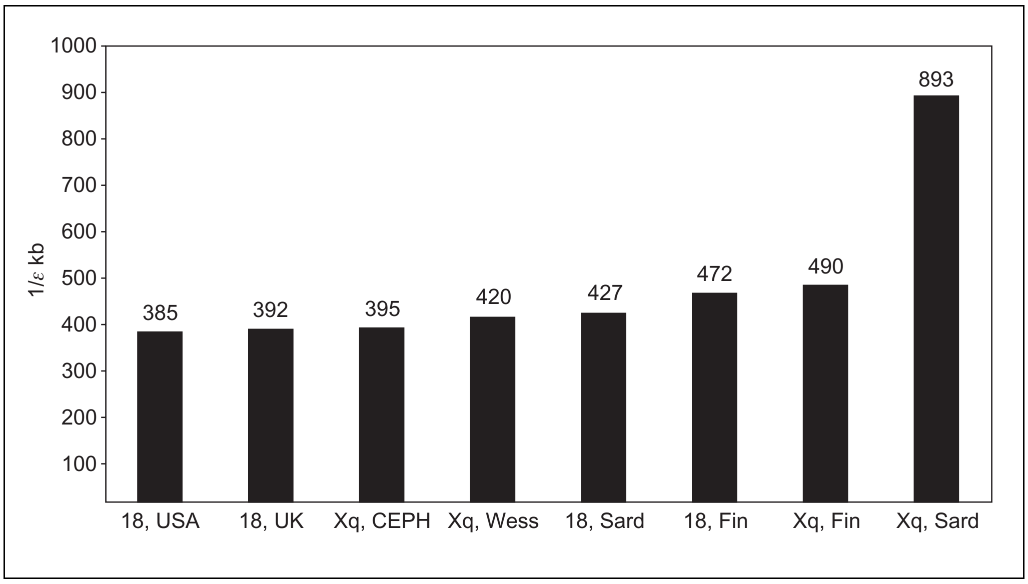

Figure 3: Swept radius (1/ε) in Kb under the Malecot model with ρ

commonly applied models are reviewed by Jorde.[23] We describe here a case-control method implemented in the ALLASS program.

LD mapping of major genes is facilitated by family studies

LD mapping of major genes is facilitated by family studies permitting the determination of disease case and normal control haplotypes. Given a set of case and control haplotypes covering a disease gene region, the pattern of decline in association with distance from the disease gene can be used to delimit and refine the disease gene location. By modelling this pattern high-resolution mapping has been achieved for several major genes. Collins and Morton[8] examined data from the CFTR (cystic fibrosis transmembrane conductance regulator) gene region. These data included those of Kerem *et al.*[24] who reported 23 polymorphisms defining 77 haplotypes carrying the ΔF508 mutation in CFTR and 149 other haplotypes. Also, Morral *et al.*[25] examined three intragenic microsatellites in the region. For mapping in this 1.77 Mb region, Collins and Morton[8] treated negatively and positively associated alleles at a locus as separate 'loci', there were thus a total of 27 markers. A distinct advantage of the ρ metric is that it uniquely specifies the adjustment ω, which is important where the disease haplotype representation is greatly enriched. Using these data the ΔF508 mutation was mapped at position 0.834 Mb in the 1.77 Mb region with Malecot parameters $M = 1$ and $L = 0$, consistent with monophyletic origin and no evidence for a bias due to association between markers at large distance. The known location of this mutation is 0.88 Mb so localisation to within 50 Kb was achieved. The value of ε was 1.019 corresponding to a swept radius of 0.98 Mb, consistent with the extensive LD expected around a disease mutation of relatively recent origin.

Close relationship between the genetic linkage and LD maps

Major genes localised to within 50 Kb

To further evaluate LD mapping of major genes Lonjou *et al.*[26] considered the Huntington disease (HD) gene region as an example of a disease of polyphyletic origin (multiple mutations having arisen at different times) and haemochromatosis (HFE) which maps to a region with greatly reduced recombination relative to the physical map. The HD analysis considered all published data on allelic association with HD. Polyphyletic origin was reflected in the magnitude of the Malecot M parameter ($M = 0.282 \pm 0.037$) but there was no evidence against $L = 0$. Despite polyphyletic origin the disease was localised to 3.688 ± 0.095 Mb which is within its assigned interval ($3.635 - 3.804$ Mb).

The HFE gene is particularly interesting representing a disease which maps to a region with considerable heterogeneity in recombination. In a region of 7.2 Mb the recombination rate varies from 0.97 Mb/cM distally to 6.14 Mb/cM proximally, indicative of a recombination cold spot in the region. Failure to recognise the discrepancy between the physical and genetic maps may have delayed the cloning of this disease gene. The impact of this phenomenon on LD mapping in the region is considerable. As might be expected the close relationship between the genetic linkage and LD maps means that if genetic rather than physical distances are used when modelling the pattern of LD a considerably better fit is obtained. The disease is localised to within 0.035 cM using the genetic map but only 2.3 Mb using the physical map. The importance of recognising the relationship between genetic and physical map distances in candidate regions is clear and the value of an LD map of the genome which highlights regions with extensive LD in a population-specific way is also evident.

MAPPING OLIGOGENES BY ALLELIC ASSOCIATION

Oligogenes are genes with a relatively small individual effect on phenotype which are involved in common diseases. Mapping of these genes by allelic association poses a number of problems including the inability to identify disease

Oligogenes are usually sufficiently common that their associations are efficiently estimated in cohorts

The resolution of LD mapping may be considerably greater than 50 Kb suggested by major genes

haplotypes since it is not possible to know whether 0, 1 or 2 are present. Much of the attention is focused on quantitative traits and Zhang *et al.*[27] present theory and an application of an approach that extends the Malecot model to oligogenes in this situation. In contrast to major genes, where haplotypes determined in families segregating the disease provide the basis for mapping, oligogenes are usually sufficiently common that their associations are efficiently estimated in cohort or case-control studies. Table 3 gives haplotype frequencies and phenotype means for a causal and predictive marker SNP. The table is arranged such that disease allele G is associated with the predictive marker allele P and haplotype frequencies are expressed in terms of ρ. Usually the causal SNP is not identified but in this table μ_G is positively correlated with μ_P for the predictive SNP and an expression for ρ for the ith associated polymorphism can be derived in terms of the regression coefficient b for the ith-associated polymorphism in the QTDT program.[4]

Assuming a single monophyletic causal SNP in the region ($M = 1$), Zhang *et al.*[27] formulated the Malecot model as $b_i R_i = T[(1 - L)e^{-\varepsilon d_i} + L] = z_i$, where the parameter T is $b_G Q$, for the ith associated polymorphism, which can be entered into composite likelihood to allow testing of hypotheses about the position of causal sites in a candidate region. The confidence interval around location estimates determines where a causal SNP should be sought. Application of this model to 10 SNP markers spanning 27 Kb of the angiotensin-1 converting enzyme (ACE) gene[28] localises a causal SNP between G2530A and 4656(CT)3/2 in the 3′ region at a distance of 21.75 ± 1.28 Kb from the most proximal SNP (T-5491C). If this evidence holds up it suggests that the resolution of LD mapping may be considerably greater than 50 Kb suggested by major genes. This is further emphasised by the analysis of association between SNPs in the region (Figure 4) which shows a decline in LD with distance even in a region of only 27 Kb.

HAPLOTYPES OR DIPLOTYPES?

A diplotype is defined as an individual's phase unknown multilocus genotype. The problematic aspects and/or limitations of haplotype reconstruction, either molecularly or probabilistically, have driven research to the conversion of diploidy to haploidy. It seems self-evident that haplotypes always provide much more information than diplotypes in which the phase is unknown. However, this is not necessarily the case. Recently, attention has been drawn again to human/rodent cell hybrids that retain a subset of human chromosomes and generate relatively stable monosomic clones by selective loss of human chromosomes.[29] This is technically quite demanding and relatively expensive so it is worth assessing the relative costs and benefits of this approach over diplotype analysis.

Table 3: Haplotype frequencies and phenotype means

Causal SNP	SNP Marker		Total	Mean
	P	P′		
G	$QR + \rho Q(1 - R)$	$Q(1 - R) - \rho Q(1 - R)$	Q	μ_G
G′	$R(1 - Q) - \rho Q(1 - R)$	$(1 - R)(1 - Q) + \rho Q(1 - R)$	$1-Q$	$\mu_{G'}$
Total	R	$1-R$	1	–
Mean trait value for allele	μ_P	$\mu_{P'}$	–	–

$b_G = \mu_G - \mu_{G'}$ (causal SNP)
$b_i = \mu_P - \mu_{P'}$ (predictive SNP)

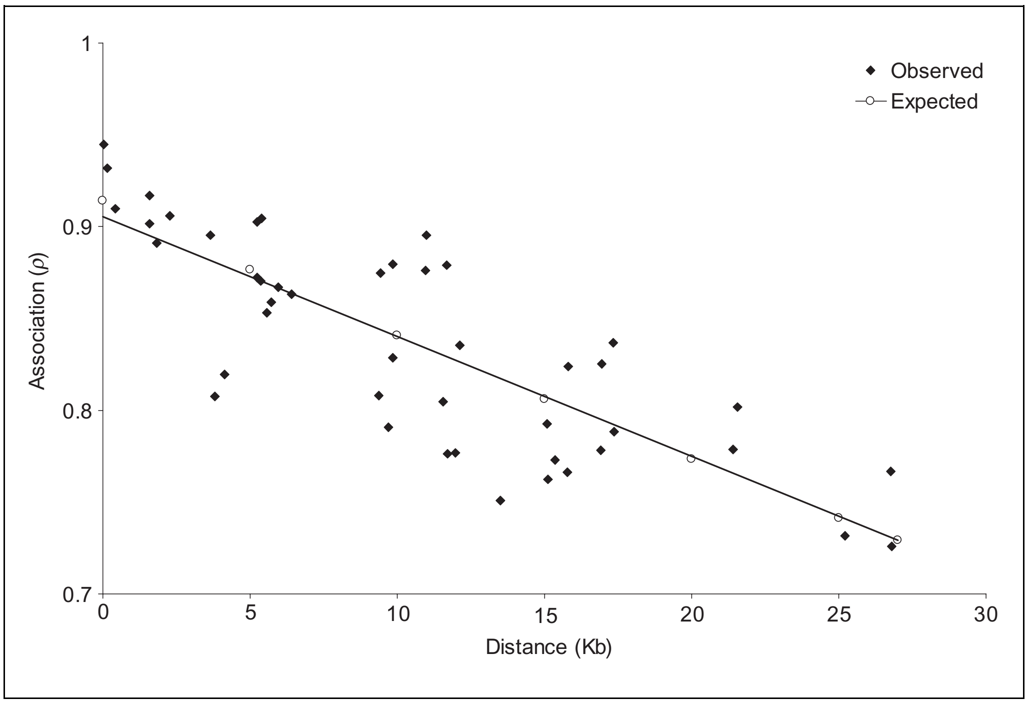

Figure 4: Allelic association for pairs of SNPs in the ACE data

Comparing the efficiency of diplotypes and haplotypes

The essential theory for pairs of codominant loci under random mating was developed by Bennett[30] and Hill[31] in terms of the covariance $D = \pi_{11}\pi_{22} - \pi_{12}\pi_{21}$. This can be extended to derive the association probability ρ, which gives the best fit to the Malecot model (Maniatis *et al.*, paper in preparation). Thompson *et al.*[32] evaluated the number of individuals required to detect strong positive disequilibrium for both diplotypes and haplotypes but did not consider the relative efficiencies under varying levels of association.

Table 4: Diplotype frequencies under random mating.

Younger polymorphism	Older polymorphism			Total
	PP	PP′	P′P′	
GG				
Observed	n_{11}	n_{12}	n_{13}	$n_{1.}$
Expected	π_{11}^2	$2\pi_{11}\pi_{12}$	π_{12}^2	Q^2
GG′				
Observed	n_{21}	n_{22}	n_{23}	$n_{2.}$
Expected	$2\pi_{11}\pi_{21}$	$2(\pi_{11}\pi_{22} + \pi_{12}\pi_{21})$	$2\pi_{12}\pi_{22}$	$2Q(1-Q)$
G′ G′				
Observed	n_{31}	n_{32}	n_{33}	$n_{3.}$
Expected	π_{21}^2	$2\pi_{21}\pi_{22}$	π_{22}^2	$(1-Q)^2$
Total				
Observed	$n_{.1}$	$n_{.2}$	$n_{.3}$	N
Expected	R^2	$2R(1-R)$	$(1-R)^2$	1

In Table 1, association ρ is defined in terms of the four haplotype frequencies in a 2 × 2 haplotype table. Alternatively, let n_{ij} be the count of individuals with genotype (diplotype) *ij*. Then the nine diplotype frequencies for a pair of codominant biallelic loci can be arranged in a 3 × 3 table (Table 4). Hill[31] introduced an iterative maximum likelihood estimate of D. Under the null hypothesis ($D = 0$) the conditional information is $K_D = N/Q(1-Q)R(1-R)$ for both haplotypes and diplotypes. Under the alternative hypothesis ($D > 0$) the information is $K_D = N/[Q(1-Q)R(1-R) + D(1-2Q)(1-2R) - D^2]$[16] and this information from diplotypes must be evaluated by inversion of the 3 × 3 matrix for the genes' frequencies Q and R and the covariance D.[31] D and K_D can be easily transformed to ρ and K_ρ. Employing the above equations, the efficiency of haplotypes and diplotypes can be compared (Figure 5). Using

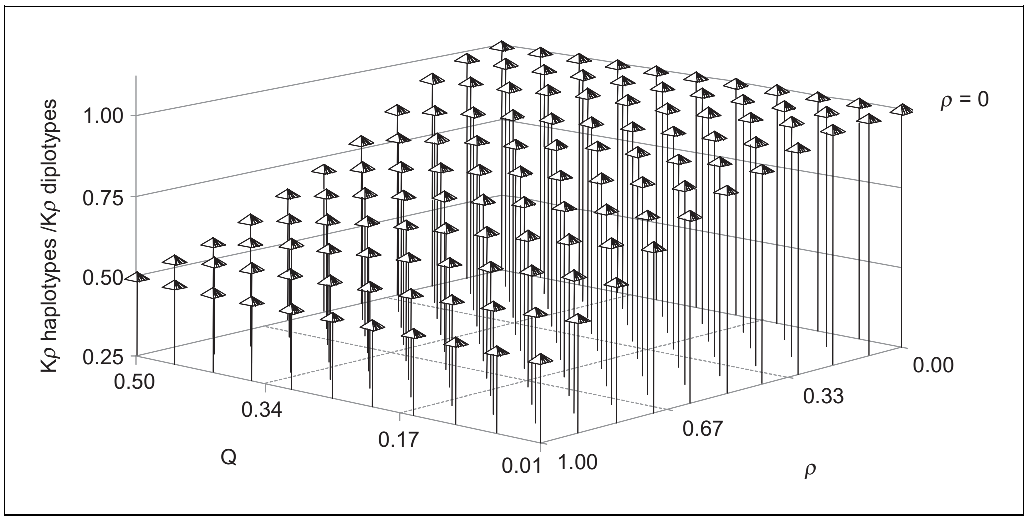

Figure 5: Efficiency of ($N = 10{,}000$) haplotypes relative to N diplotypes

$N = 10{,}000$ haplotypes and constraining R to 0.5, the relative efficiency was derived for different gene frequencies ($\mathrm{Q} = 0.01 - 0.5$) and association probabilities (ρ). Figure 5 demonstrates that under the null hypothesis ($\rho = 0$) the efficiency of N haplotypes compared to N diplotypes is 1. In other words any sample of N haplotypes will contribute the same information as N diplotypes. However, this is the worst case since, as ρ approaches 1 (complete disequilibrium), the efficiency of haplotypes goes to 0.5, N diplotypes have as much information as $2N$ haplotypes.

A 'hot spot' for LD is likely to be in a recombination cold spot

Haplotype and diplotype efficiency was also studied using the X chromosome in males that give credible samples of directly observed haplotypes. The outbred Centre d'Etude du Polymorphisme Humain (CEPH) sample presented by Taillon-Miller *et al.*[21] comprises 39 SNPs typed on the X chromosome, located between 32.6 Mb and 158.5 Mb. These data were subsequently used to simulate a sample of diplotypes by random sampling of haplotypes with replacement. Analysing haplotypes and diplotypes by the Malecot model under the null hypothesis ($\rho = 0$) yields similar parameter estimates for both cases (Figure 6), demonstrating the equivalence of N diplotypes and N haplotypes in a real data example. Therefore, it seems reasonable that LD can be mapped efficiently using diplotype data with less information lost than might be supposed.

DEVELOPING THE LD MAP

Neither the genetic linkage map nor the LD map is proportional to the physical map derived from the DNA sequence. The LD map mirrors the linkage map more closely since LD in a region declines through recombination. For this reason a 'hot spot' for LD is likely to be in a recombination cold spot. Furthermore, since the resolution of linkage maps is low, being limited to a relatively small number of meioses, the pattern of LD in a region might usefully extend the resolution of the linkage map. However, the agreement between the two maps is disturbed by numerous other processes that disrupt LD. We therefore need an LD map to facilitate positional cloning, determine suitable marker densities, compare populations and detect selective sweeps and other events of evolutionary interest. The methodology for constructing such a map is still being developed. One approach that holds promise[33] is to estimate the Malecot parameter ε in an interval between a pair of adjacent markers in a high-density map. To determine ε all pairwise values of

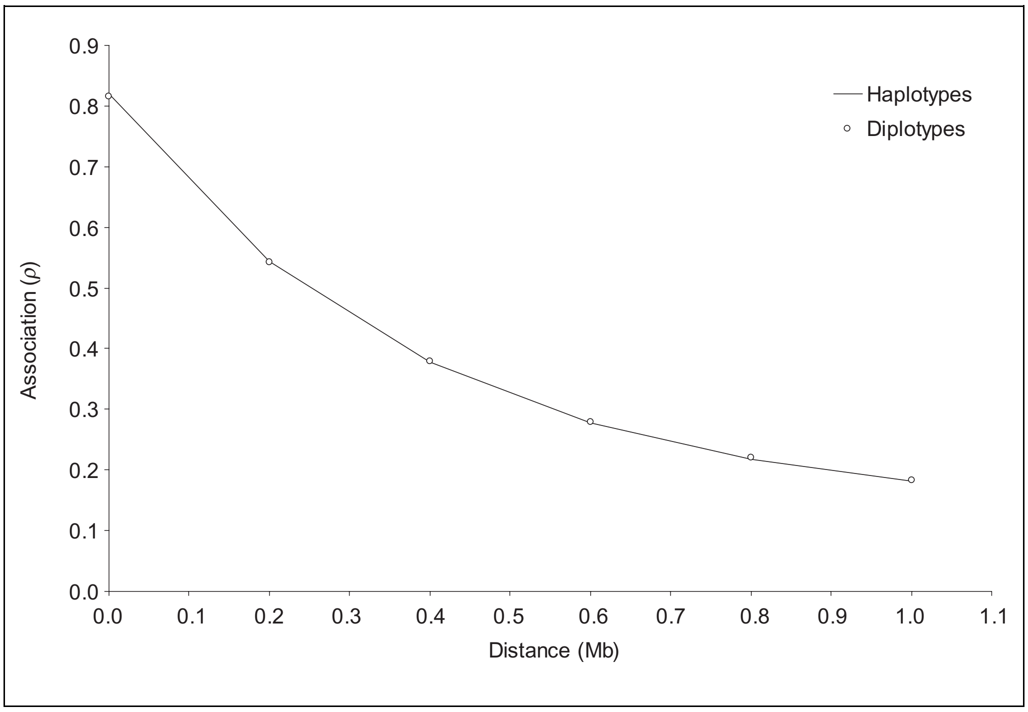

Figure 6: The Malecot model for haplotypes and diplotypes

ρ that include these two markers provide the data for establishing the model. A natural measure of LD is $\varepsilon d = \theta t$ where distance (d) is expressed in Kb, θ is a small frequency of recombination and t is the number of generations. As εd is not biased in favour of the linkage map and is more accurately known than θt, it is a more useful metric for LD. A definitive property of a chromosome map is that its distances are additive. To ensure additivity, reference markers in the linkage map may be useful for integration of LD map distances but further experience in this area is required. The hope is that it will be possible to develop a 'standard' LD map to which population-specific maps are approximately proportional with deviations that mirror mutation, time, selection, gene frequency and drift. To develop such a map for the whole genome requires major effort. Some information about LD in the genome may be obtained from the CEPH genotype database,[34] which has a large number of genotypes although the marker density and sample size is too low to give a definitive map.[35]

Population specific LD maps may be proportional to a 'standard' LD map

CONCLUSIONS

Allelic association or LD mapping offers great promise for the localisation of disease genes. This effort will be greatly facilitated by dense arrays of SNPs in candidate regions and a thorough understanding of both candidate region and population-specific LD patterns. The methodology for the analysis of LD data has been an area of intense research and it has become clear that stochastic variation makes careful modelling of LD crucial for reliable inference. It is not clear that experimentally derived haplotypes are necessary, particularly for mapping oligogenes where it is not even possible to determine which haplotypes include the disease allele. The available evidence suggests that LD is extensive in many populations and genome regions, which should facilitate mapping, and LD maps will be invaluable for this process.

WWW RESOURCES AND SOFTWARE

The ALLASS program which implements the Malecot model is available from:

http://cedar.genetics.soton.ac.uk/public_html/

The genetic linkage analysis web resource at Rockefeller provides a large collection of software and references relevant to this area: http://linkage.rockefeller.edu/

The QTDT program is available from: http://www.well.ox.ac.uk/asthma/QTDT

References

1. Ardlie, K., Liu-Cordero, S. N., Eberle, M. A. *et al.* (2001), 'Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion', *Amer. J. Human Genet.*, Vol. 69(3), pp. 582–589.
2. Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993), 'Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Amer. J. Human Genet.*, Vol. 52(3), pp. 506–516.
3. Allison, D. B. (1997), 'Transmission-disequilibrium tests for quantitative traits', *Amer. J. Human Genet.*, Vol. 60(3), pp. 676–690.
4. Abecasis, G., Cardon, L. R. and Cookson, W. O. A. (2000), 'A general test of association for quantitative traits in nuclear families', *Amer. J. Human Genet.*, Vol. 66, pp. 279–292.
5. Morton, N. E. and Collins, A. (1998), 'Tests and estimates of allelic association in complex inheritance', *Proc. Natl Acad. Sci. USA*, Vol. 95(19), pp. 11389–11393.
6. Pritchard, J. K. and Rosenberg, N. A. (1999), 'Use of unlinked genetic markers to detect population stratification in association studies', *Amer. J. Human Genet.*, Vol. 65, pp. 220–228.
7. Devlin, B., Risch, N. and Roeder, K. (1996), 'Disequilibrium mapping: composite likelihood for pairwise disequilibrium', *Genomics*, Vol. 36, pp. 1–16.
8. Collins, A. and Morton, N. E. (1998), 'Mapping a disease locus by allelic association', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 1741–1745.
9. Collins, A., Lonjou, C. and Morton, N. E. (1999), 'Genetic epidemiology of single nucleotide polymorphisms', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 15173–15177.
10. Devlin, B and Risch, N. (1995), 'A comparison of linkage disequilibrium measures for fine-scale mapping', *Genomics*, Vol. 29(2), pp. 311–322.
11. Lewontin, R. C. and Kojima, K. (1960), 'The evolutionary dynamics of complex polymorphisms', *Evolution*, Vol. 14, pp. 458–472.
12. Hill, W. G. and Weir, B. S. (1994), 'Maximum likelihood estimation of gene location by linkage disequilibrium', *Amer. J. Human Genet.*, Vol. 54, pp. 705–714.
13. Kaplan, N. and Weir, B. S. (1992), 'Expected behaviour of conditional linkage disequilibrium', *Amer. J. Human Genet.*, Vol. 51, pp. 333–343.
14. Lehesjoki, A-E., Koskiniemi, M., Norio, R. *et al.* (1993), 'Localisation of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping', *Human Mol. Genet.*, Vol. 2, pp. 1229–1234.
15. Yule, G. U. (1900), 'On the association of attributes in statistics', *Phil. Trans. R. Soc. London*, Vol. 194, pp. 257–319.
16. Morton, N. E., Zhang, W., Taillon-Miller, P. *et al.* (2001), 'The optimal measure of allelic association', *Proc. Natl Acad. Sci. USA*, Vol. 98(9), pp. 5217–5221.
17. Kruglyak, L. (1999), 'Prospects for whole-genome linkage disequilibrium mapping of common disease genes', *Nature Genet.*, Vol. 22, pp. 139–144.
18. Reich, D. E., Cargill, M., Bolk, S. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature*, Vol. 411, pp. 199–204.
19. Abecasis, G. R., Noguchi, E., Heinzmann, A. *et al.* (2001), 'Extent and distribution of linkage disequilibrium in three genomic regions', *Amer. J. Human Genet.*, Vol. 68, pp. 191–197.
20. Lonjou, C., Collins, A. and Morton, N. E. (1999), 'Allelic association between marker loci', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 1621–1626.
21. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L. *et al.* (2000), 'Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28', *Nature Genet.*, Vol. 25, pp. 324–328.
22. Eaves, I. A., Merriman, T. R., Barber, R. A. *et al.*(2000), 'The genetically isolated populations of Finland and Sardinia may not be panacea for linkage disequilibrium mapping of common disease genes', *Nature Genet.*, Vol. 25, pp. 320–323.
23. Jorde, L. B. (2000), 'Linkage disequilibrium and the search for complex disease genes', *Genome Res.*, Vol. 10(10), pp. 1435–1444.
24. Kerem, B., Rommens, J. S., Buchanan, J. A., *et al.* (1989), 'Identification of the cystic fibrosis gene: genetic analysis', *Science*, Vol. 245, pp. 1073–1080.
25. Morral, N., Bertranpetit, J., Estivill, X. *et al.* (1994), 'The origin of the major cystic fibrosis mutation (delta F508) in European

populations', *Nature Genet.*, Vol. 7, pp. 169–175.

26. Lonjou, C., Collins, A., Ajioka, R. S. *et al.* (1998), 'Allelic association under map error and recombinational heterogeneity: A tale of two sites', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 11366–11370.

27. Zhang, W., Collins, A., Abecasis, G. R. *et al.* (2001), 'Mapping quantitative effects of oligogenes by allelic association' (*Ann. Human Genet.*, submitted).

28. Keavney, B., McKenzie, C. A., Connell, J. M. *et al.* (1998), 'Measured haplotype analysis of the angiotensin-1 converting enzyme gene', *Human Mol. Genet.*, Vol. 7, pp. 1745–1751.

29. Yan, H., Papadopoulos, N., Marra, G. *et al.* (2000), 'Conversion of diploidy to haploidy', *Nature*, Vol. 403, pp. 723–724.

30. Bennett, J. H. (1965), 'Estimation of the frequencies of linked gene pairs in randomly mating populations', *Amer. J. Human Genet.*, Vol. 17, pp. 51–53.

31. Hill, W. G. (1974), 'Estimation of linkage disequilibrium in randomly mating populations', *Heredity*, Vol. 33, pp. 229–239.

32. Thompson, E. A., Deeb, S., Walker, D. and Motulsky, A. G. (1988), 'The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CII apolipoprotein gene', *Amer. J. Human Genet.*, Vol. 42, pp. 113–124.

33. Collins, A., Ennis, S., Taillon-Miller, P. *et al.* (2001), 'Allelic association with SNPs: metrics, populations, and the linkage disequilibrium map', *Human Mutat.*, Vol. 17, pp. 255–262.

34. Dausset, J., Cann, H., Cohen, D. *et al.* (1990), 'Centre d'etude du polymorphisme humaine (CEPH): collaborative genetic mapping of the human genome', *Genomics*, Vol. 6(3), pp. 575–577.

35. Huttley, G. A., Smith, M. W., Carrington, M. and O'Brien, S. J. (1999), 'A scan for linkage disequilibrium across the human genome', *Genetics*, Vol. 152, pp. 1711–1722.