Pacific Symposium on Biocomputing 5:636-647 (2000)

LINKAGE DISEQUILIBRIUM MAPPING USING SINGLE NUCLEOTIDE POLYMORPHISMS -WHICH POPULATION?

A. COLLINS
Department of Human Genetics
University of Southampton
Duthie Building (808)
Southampton General Hospital
Tremona Road, Southampton SO16 6YD
arc@soton.ac.uk

Abstract

There is considerable interest in the potential of single nucleotide polymorphisms (SNPs) for mapping complex traits which are determined by genes of small individual effect (oligogenes). It is thought likely that many oligogenes are themselves common polymorphisms, perhaps biallelic, for which there is effectively neutral selection reflected in late age of onset. The extent of detectable linkage disequilibrium between SNP x SNP pairs and SNP x oligogene pairs is of considerable interest, particularly in the context of identifying 'favourable' populations. Unfortunately data are sparse and few populations have been extensively sampled. Polymorphisms with the appropriate characteristics that have been studied are blood groups in the Rhesus and MNS systems for which there are extensive data on four pairs of biallelics. These might be regarded as surrogates for SNP-SNP or SNP-oligogene pairs. By developing and applying an approach, previously used for major genes, to evaluate association (ρ) in SNP haplotypes, it is evident that, with some exceptions, there is little difference between isolates and large populations. Furthermore it is apparent that there is useful linkage disequilibrium even for the MN-Ss locus pair (0.195 cM apart), in both large populations and isolates. This is somewhat more favourable to linkage disequilibrium mapping than a recent simulation suggests.

Introduction and theory

Numerous 'major' genes associated with Mendelian disease have been identified and each involves a rare mutation which has arisen relatively recently (within the last 2000 years or so). Characteristically, the mutation is both necessary and sufficient to produce the disease phenotype. For common diseases (complex traits) it is likely that multiple genes (oligogenes) are involved each of which has a small effect that is insufficient to cause disease individually. For many complex traits common polymorphisms are candidate disease genes and many of these alleles are very old (perhaps exceeding 10,000 years?). The detection of genes of small individual effect is difficult. The positional cloning strategy, so successfully applied to major genes, has been rather ineffective for complex diseases. Even if approximate localisation to a region can be achieved confirmation and replication of findings is difficult and

refinement of the region leading to cloning of the gene has been an even greater challenge. The interest in SNPs and technologies related to SNPs has suggested alternative approaches one of which is to directly test SNPs in the coding regions of genes for association with a disease phenotype. However, there are some uncertainties about the density of SNPs required for this to be successful and the assumption that cSNPs, common variants in coding regions, are themselves aetiological sites for complex diseases. It is also likely that many of the relevant mutations are in regulatory regions[1]. For complete generality this direct approach requires the characterization of all of the human genes and their polymorphisms and is therefore only a prospect for the longer term.

A second approach to identifying oligogenes is an indirect strategy using genome wide high-resolution maps of SNPs and other polymorphisms to scan for marker-disease associations. SNPs are particularly useful as they are abundant and have low mutation rates. Map-based association studies depend on allelic association or, more correctly, linkage disequilibrium for linked loci. However, patterns of association are highly complex and influenced by recombination, mutation and evolutionary factors. It is known that major genes can be fine-scale mapped by exploiting the greater resolution offered by linkage disequilibrium, as opposed to linkage. Linkage disequilibrium mapping takes advantage of recombination over multiple generations rather than a single generation in family material. However, for oligogenes it is still not clear that this is a useful approach given their presumed greater duration. An important concept in this respect is that whilst for major genes founder effects predominate and disequilibrium is decreasing with time, dominated by recombination, a proportion of oligogenes may be at quasi-equilibrium more dependent on the founding population size. Thus the mapping of genes using linkage disequilibrium generated by genetic drift ("drift mapping"[2]), may be more relevant.

This phenomena of genetic drift generating linkage disequilibrium between loci that show effectively neutral selection has been studied extensively, following on from work by Hill and Robertson[3]. Assume that a gene of interest is represented by an SNP with a disease allele, D, which increases risk of disease, and an allele d not implicated in disease. Markers located very close to the disease locus will have effectively zero recombination between them. Under these conditions the region around the D allele will be isolated from the d allele region and over this small distance disease and non-disease chromosome populations will evolve independently. This is sufficient to generate substantial differences over time in marker allele

frequencies for the disease and normal chromosome populations. Thus linkage disequilibrium can exist in the absence of a founder effect. In these circumstances population size is a critical factor with slower changes in allele frequencies arising in large populations. Typically, the disease chromosome population will be smaller resulting in more rapid changes. The process will not tend to generate a single predominant 'disease haplotype' since there is likely to be considerable haplotypic diversity.

With a founder effect there is fixation of alleles in the D population through the creation of a single disease-associated haplotype. Even if there were more than one founder drift is likely to cause one or a few to predominate in time. The founder effect can be generated over time as long as the copy number of disease alleles remains small. Recombination acts to equilibrate allele frequencies between disease and normal chromosome populations but its impact depends on the rate at which disequilibrium can be generated by drift which in turn depends on the effective population size of the disease chromosomes. Drift will dominate recombination where the population frequency of disease chromosomes is small.

Sved[4] developed theory to describe the increase or decline in linkage disequilibrium in a finite population. The Sved equation defines kinship ϕ, between pairs of loci, which is the probability of joint identity by descent at one locus conditional on identity by descent at the other locus. The general theory which is a function of the effective population size (N_e) can be expressed as

$$\phi_t = \phi_{rt} + \phi_{ct} \text{ where} \quad [1]$$

$$\phi_{rt} = \phi_o e^{-(1/2\,Ne + 2\theta)t} \text{ and} \quad [2]$$

$$\phi_{ct} = \phi_\infty\ (1-e^{-(1/2\,Ne + 2\theta)t}) \text{ and} \quad [3]$$

$$\phi_\infty = 1/(1 + 4\,N_e\,\theta). \quad [4]$$

Here remote kinship (ϕ_{rt}) is diminishing with the number of generations (t) from an initial value ϕ_o as t approaches ∞, reflecting a founder effect. Close kinship ϕ_{ct} builds up by drift from an initial value at zero to some equilibrium ϕ_∞ depending on the effective population size, N_e, which is assumed to be constant. Sved equated ϕ_t to the

expected value of r^2, where r is the correlation of gene frequencies in a 2 x2 table in which a, b, c, and d are haplotypic counts for a pair of loci such that

$$r = (ad-bc) / \sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)} \qquad [5]$$

Lonjou et al[5] in their analysis of blood group polymorphisms adopted $r^2 \approx \phi_t$ however r is not independent of the marginal gene frequencies[4] and it is preferable to use ρ[6] which takes the value 1 for a monophyletic allele (Table 1). The Malecot model for allelic association[6] was developed to describe ρ as a function of mono or polyphyletic origin, the number of generations and the bias due to spurious association.

Methods

Table 1. Haplotype frequencies

		'Older' SNP_2		Totals
		allele 1	allele 2	
'Younger'	allele 1			
SNP_1	number:	a	b	a + b
	founders:	Q	0	Q
	frequency:	$Q\rho + QR\,(1-\rho)$	$(1-\rho)\,Q\,(1-R)$	Q
	equilibrium:	QR	Q (1-R)	Q
	allele 2			
	number:	c	d	c+d
	founders:	R-Q	1-R	1-Q
	frequency:	$(R-Q)\,\rho+R\,(1-Q)\,(1-\rho)$	$(1-R)\,[\rho+ (1-Q)\,(1-\rho)]$	1-Q
	equilibrium:	R (1-Q)	(1-R) (1-Q)	1-Q
Totals		R	1-R	n

Q - frequency of allele 1, SNP_1; R - frequency of allele 1, SNP_2; ρ - association; a, b, c, d - observed haplotypic counts, n - total number of haplotypes

Table 1 gives haplotype frequencies for SNPs[6]. It is assumed that, by analogy to disease gene: marker haplotypes, there are 'younger' and 'older' SNPs. The cells of Table 1 may be arranged by interchanging alleles for either or both SNPs which may

themselves be interchanged . This ensures that $ad \geq bc$ and $b \leq c$. Allele frequencies are assumed constant over time. If each SNP has alleles coded 1,2 then allele 1 in the 'younger' SNP_1 is analogous to the disease allele, assumed to be present in founders only in conjunction with allele 1 of SNP_2 (1,1 haplotype). Thus the frequency of 1,1 haplotypes in founders is Q and which point there are no 1,2 haplotypes. At equilibrium it is assumed that the 1,1 haplotype frequency is simply a product of the corresponding allele frequencies, namely QR. The present day frequency , $Q\rho + QR(1-\rho)$, is therefore a function of association ρ which varies between 0 (no association) and 1 (complete association).

The Malecot model was originally developed for populations isolated by distance but adapted[6] to model the decline of allelic association with distance from a disease gene or 'younger' SNP of a pair . The log likelihood of the multiple pairwise observations summed over i pairs is

$$- \sum K_i (\hat{\rho}_i - \rho_i)^2/2 \quad \text{where } \rho_i = (1-L)\, M e^{-\epsilon d} + L \qquad [6]$$

K_i is the information about ρ (Table 1), $\hat{\rho}_i$ are the values of ρ fitted to the Malecot model, M=1 if SNP_1 is monophyletic and <1 otherwise, $d \geq 0$ is the distance on the genetic map between the SNP-SNP pair and L is the bias due to the constraint $\rho \geq 0$. The parameter $\epsilon \geq 0$ depends on the number of generations during which haplotypes have been approaching equilibrium.

Analysis of population data

Simulation has been used to attempt to answer questions about the extent of linkage disequilibrium in SNP maps and therefore the densities required to localise disease genes and the types of populations for which this is best suited[7]. The analysis of real data has obvious advantages but extensive studies of different populations for SNPs are not yet available. Common biallelic polymorphisms for which there are extensive data are blood group polymorphisms[8,9,10], and these have been examined as 'surrogates' for common disease genes[5].

Four locus pairs were considered by Lonjou et al[5]:

1. The glycophorin loci GYPA and GYPB which control the MN and Ss blood groups and are approximately 0.195 cM apart corresponding to only 80Kb on

the physical map. There is known to be a relatively high frequency of recombination and gene coversion in the GYP cluster.

2. The RHCE locus contains two sites coding for Cc and Ee and these are approximately 30kb apart (≈ 0.03 cM).
3. The RHD locus is a close homologue of RHCE and is associated with the D antigen in man. The precise orientation with respect to RHCE is uncertain and evidence conflicting[11]. Based on kinship Lonjou et al[5] suggested 0.02 cM as the distance between Rh C and Rh D sites.
4. Based on kinship, Lonjou et al[5] suggested 0.16 cM for the Rh D- Rh E locus pair.

In the light of recent discussion[11] and simulation[7] it is of interest to examine these data with regard to the extent of linkage disequilibrium in different populations and the potential importance of isolates for linkage disequilibrium mapping.

Results

Table 2 Association ρ and information K from pooled haplotype frequencies for large populations and isolates

	Rh C-D		Rh C-E		Rh D-E		MN-Ss		Mean ρ
	ρ	K	ρ	K	ρ	K	ρ	K	
Large populations									
Amerindians	.964	1770	.830	31539	.776	877	.272	7614	.771
Far East	.674	1276	.916	7407	.106	233	.272	1532	.492
Europe	.906	15404	.934	3839	.888	2752	.428	19453	.789
India and Pakistan	.894	9568	.913	4117	.823	694	.153	3718	.696
Oceania	.250	888	.964	20464	.793	51	.503	937	.628
Near East	.958	1546	.967	741	.915	296	.338	2983	.795
North Africa	.931	966	.984	238	.863	251	.405	996	.796
Sub-Saharan Africa	.381	267	.780	66	.648	197	.218	4580	.507
Means:	.745	-	.911	-	.727	-	.324	-	.676
Isolates									
Ainu	.902	341	1.000	1060	.679	376	.798	269	.845
Eskimos	1.000	285	.946	2269	.916	76	.442	253	.826
Basques	.930	1066	1.000	81	.928	125	.287	66	.786
Lapps	.981	955	1.000	1047	1.000	151	.218	1158	.800
Jews	.944	2707	.859	865	.957	389	.233	1291	.748
Tristan da Cunha	.518	88	1.000	210	1.000	153	.322	132	.710
Means	.879	-	.968	-	.913	-	.383	-	.786

Table 2 shows association ρ and information K, calculated from the pooled haplotype frequencies. Assuming that blood group antigens are representative of SNP x oligogene pairs then these results offer insight into the expected level of linkage disequilibrium in a range of different populations. With some exceptions there is surprisingly little difference in ρ between large populations (mean = 0.676) and isolates (mean = 0.786). The Ainu (mean = 0.845) and the Sub-Saharan African population (mean = 0.507) and Far Eastern populations (mean = 0.492) have the most extreme maximum and minimum values. Four of the 6 isolates have mean values of ρ that lie within the range for large populations. Tristan da Cunha gives discrepant results, particularly for the Rh C-D pair where ρ is only 0.518, although this is based on a small sample. This pair gives similar anomalous results (given close linkage) with sub-Saharan and Oceania populations.

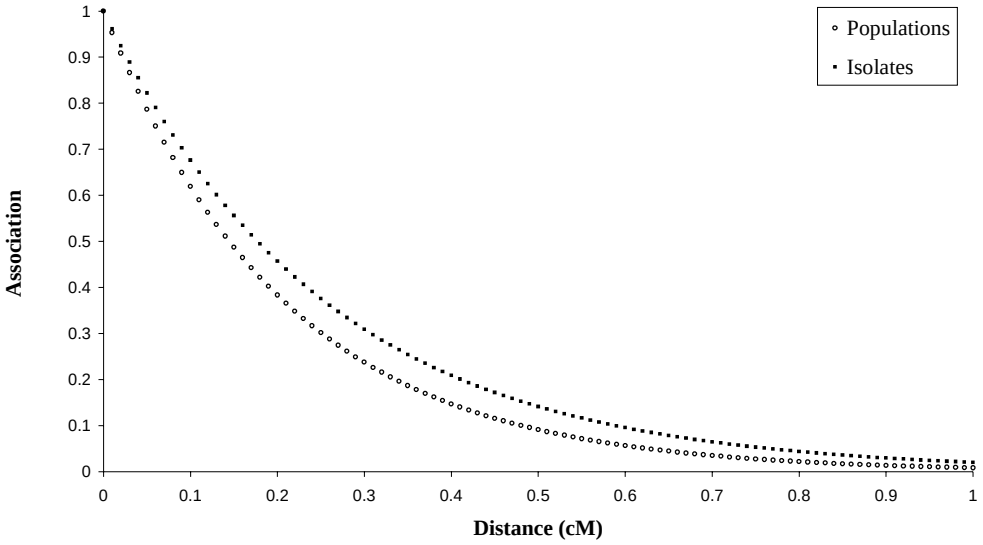
Recently Kruglyak[11] has suggested that only the very tightly linked Rh C-D and Rh C-E pairs show sufficiently broad ranges of linkage disequilibrium for valid comparisons. As expected association is generally lower for the more loosely-linked pairs, although it is evident that there is detectable linkage disequilibrium even at 0.195 cM for the MN-Ss pair.

Although there is substantial variation the mean association ρ is only 14% greater in the isolates sample than in the large populations. However, this relatively small difference is more clearly demonstrated when the Malecot model is fitted and each sample is thus weighted by its information (Table 3, Figure 1).

Table 3 Estimates of Malecot model parameters ($\hat{L} = 0$, $\hat{M} = 1$)

	$\in$	**Residual χ^2**	**df**
Populations	4.792	2174	29
Isolates	3.914	408	21
All samples	4.702	2609	53

Figure 1. Association with distance for large populations and isolates (Malecot model)



It should be noted that the fitted model (Figure 1) is extrapolated beyond 0.195 cM and firm conclusions at distances greater than this will require a larger range of pairs of loci. In both separate and combined samples M takes a value of 1, and the bias (L) due to the assumption that $\rho \geq 0$ is zero due to the lack of information for more widely spaced pairs. The relatively slightly greater extent of linkage disequilibrium in isolates is reflected in a lower value of $\in$ and Figure 1. The residual χ^2 after fitting the model is large in every case reflecting heterogeneity between populations within the two major groups. However, the fitted model is entirely consistent with a large sample of SNP-SNP pairs[12] from RFLP haplotypes which suggests detectable linkage disequilibrium at a distance of 350-500 Kb.

Discussion

The study of common blood group biallelic polymorphisms as surrogates for SNP x oligogene or SNP x SNP pairs gives some insight into the extent and magnitude of

linkage disequilibrium in human populations. The study suggests that many isolates have no more extensive disequilibrium than larger populations. There are three areas for discussion:

The extent of linkage disequilibrium

Krugylak[7] performed a backward simulation from a sample of present-day gametes to common ancestors assuming that the general human population had a constant effective size of 10,000 until 5000 generations ago, followed by exponential expansion. Kruglyak concludes that essentially no linkage disequilibrium is observed at $\theta = 0.0003$ or greater (≈30 Kb) and that a marker within $\theta = 0.00003$ (≈ 3 Kb) of every variant is required to avoid large increases in sample size. This corresponds to a map of 500K SNPs with an average spacing of 6 Kb. The results presented here suggest that for most populations these findings are unduly pessimistic since there is detectable linkage disequilibrium to at least c.0.2 cM and perhaps up to 0.4 cM (≈ 200-400 Kb). Recent analysis[12] of SNP-SNP pairs in random haplotypes shows that there is some "spurious" association extending to greater distances, reflected in the parameter L in the Malecot model taking a value of 0.18. From Figure 1, $\rho > 0.18$ corresponds to approximately 0.35 - 0.45 cM as the maximum extent of 'useful' linkage disequilibrium suggesting that lower SNP densities than simulation indicates are needed for adequate coverage of a region. As with most simulations the simplifying assumptions presumably have a big influence. Principal amongst these must be the treatment of the human population as panmictic and the assumption of exponential expansion, for the purposes of reconstructing the genealogy. Other limitations recognised[7] include neglect of non-uniform recombination and linkage disequilibrium created by selection either by 'hitchhiking' or background selection which reduces the effective population size.

The role of population expansion.

Terwilliger et al[2] distinguishes expanded populations, where disequilibrium declays over time, and small populations of constant size where disequilibrium by drift is more profound. The advantages of such populations are that they should generate new disequilibrium faster than recombination can deplete it. Lonjou et al[5] argue that this argument does not hold for allelic association over small distances with disequilibrium largely determined by ϕ_0 and N_e at the time of the bottleneck implying subsequent expansion is less important. The data support this view with relatively

little distinction between most isolates with small stable populations and large expanded populations, implying a large effect of an ancient bottleneck.

Which population for localising oligogenes?

The results from blood groups, as far as they go, suggest that isolates with small stable populations might not necessarily be advantageous for mapping common disease genes. However, since some specific populations may show increased linkage disequilibrium (eg the Ainu) the relative costs and difficulties of obtaining samples should be weighed against the potentially modest advantages for mapping and the advantages gained from a less heterogeneous environment and reduced allelic heterogeneity. Clearly some populations do show more extensive linkage disequilibrium than others and it would be unwise to draw very general conclusions from a limited number of marker x marker pairs. Linkage disequilibrium measured for a larger number of pairs in a range of populations will obviously be useful. Furthermore differences in the extent of linkage disequilibrium will also depend on the region of the genome sampled, since wide variations in recombination rates occur. Kruglyak[10] has suggested that through collaboration and integration of the Human Genome and Genome Diversity Projects a linkage disequilibrium map that identifies populations and genomic regions with strong and weak linkage disequilibrium should be constructed. This would clearly be useful for identifying suitable populations for oligogene mapping in a particular candidate region and may well prove to be an essential tool in the longer term.

References

1. A. Chakravarti, "Population genetics - making sense out of sequence". Nat. Genet. supplement 21, 56-60 (1999).

2.J.D. Terwilliger, S. Zollner, M. Laan, S. P≅≅bo, "Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion". Hum. Hered. 48, 138-154 (1998).

3.W.G. Hill and A Robertson, "Linkage disequilibrium in finite populations". Theor. Appl. Genet. 38, 226-231 (1968).

4.J Sved, "Linkage disequilibrium and homozygosity of chromosome segments in finite populations". Theor. Pop. Biol. 2, 125-141 (1971).

5.C. Lonjou A. Collins, N.E. Morton, "Allelic association between marker loci". Proc. Natl. Acad. Sci. USA 96, 1621-1626 (1999)

6.Collins A & Morton NE, "Mapping a disease locus by allelic association". Proc. Natl. Acad. Sci. USA 95, 1741-1745 (1998).

7.L. Kruglyak, "Prospects for whole genome linkage disequilibrium mapping of common diseases". Nat. Genet. 22, 139-144 (1999)
8.A.E. Mourant, A.C. Kopec, K. Domaniewska-Sobczal, "The distribution of human blood groups". Oxford University Press, London (1976)
9.D. Tills, A.C. Kopec and R.E. Tills, "The distribution of the human blood groups", supplement 1. Oxford University Press, London (1983)
10 A.K. Roychoudhury and M. Nei, "Human polymorphic genes: World distribution" . Oxford University Press, London (1988)
11.L. Kruglyak, "Genetic isolates: separate but equal?" Proc. Natl. Acad. Sci. USA 96:1170-1172 (1999)
12. Collins A, Lonjou C & Morton NE, “Genetic epidemiology of single nucleotide polymorphisms”. Submitted to Proc. Natl. Acad. Sci. USA