Genetic Epidemiology 27: 321–333 (2004)

The Role of Haplotypes in Candidate Gene Studies

Andrew G. Clark*

Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York

Human geneticists working on systems for which it is possible to make a strong case for a set of candidate genes face the problem of whether it is necessary to consider the variation in those genes as phased haplotypes, or whether the one-SNP-at-a-time approach might perform as well. There are three reasons why the phased haplotype route should be an improvement. First, the protein products of the candidate genes occur in polypeptide chains whose folding and other properties may depend on particular combinations of amino acids. Second, population genetic principles show us that variation in populations is inherently structured into haplotypes. Third, the statistical power of association tests with phased data is likely to be improved because of the reduction in dimension. However, in reality it takes a great deal of extra work to obtain valid haplotype phase information, and inferred phase information may simply compound the errors. In addition, if the causal connection between SNPs and a phenotype is truly driven by just a single SNP, then the haplotype-based approach may perform worse than the one-SNP-at-a-time approach. Here we examine some of the factors that affect haplotype patterns in genes, how haplotypes may be inferred, and how haplotypes have been useful in the context of testing association between candidate genes and complex traits. *Genet. Epidemiol.* © 2004 Wiley-Liss, Inc.

Key words: haplotype inference; haplotype association testing; candidate genes; linkage equilibrium

Grant sponsor: NIH; Grant numbers: GM65509, HL072904.

*Correspondence to: Andrew G. Clark, Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.
E-mail: ac347@cornell.edu

Received 23 June 2004; Accepted 29 June 2004
Published online 14 September 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20025

WHY STUDY HAPLOTYPES?

The primary focus of this review is on the problem of identifying DNA sequence variation that is segregating in a population and has a causal connection to variation in risk of complex disease. One could perform tests of this association without consideration of the fact that the SNPs are not independent of one another, but the determination of critical values must be made in the context of linkage disequilibrium among SNPs. Fortunately, permutation tests accomplish this reasonably well [Churchill and Doerge, 1994; Doerge and Churchill, 1996]. For each SNP, the distribution of phenotypes would be compared among the 2 or 3 genotypes observed in the sample. The statistical inference would need to take into account the multiplicity of tests being performed, and the biological interpretation would need to account for the fact that some SNPs are in linkage disequilibrium with one another. But the basic idea of testing association with individual SNPs is often where the analysis begins. In reality, the DNA sequence variation that is found in a population is the result of the past transmission of that variation through the population, and this historical past produces a structure to the SNP variation that can be of considerable value in trying to solve the primary goal of finding variants associated with disease risk. The three primary reasons for considering the haplotype organization of variation discussed here are: 1) that the unit of biological function, the protein-coding gene, produces proteins whose sequences correspond to maternal and paternal haplotypes, 2) that variation in a population is in fact structured into haplotypes that are likely to be transmitted as a unit, and 3) that regardless of the population genetic reasons, haplotypes serve to reduce the dimensionality of the problem of testing association, and so they may increase the power of those tests. Let's look at these three ideas in more detail.

HAPLOTYPES DEFINE FUNCTIONAL UNITS OF GENES

For each protein-coding gene, regardless of the number of heterozygous amino-acid sites an individual has, he or she will only produce at most two polypeptide chains, one corresponding to the maternal and one to the paternal haplotype (ignoring for now the complexities of alternative splicing, RNA editing, posttranslational modification, etc.). The point is that the folding kinetics, stability, or any other physical properties of the protein may depend on interactions between pairs or higher-order combinations of amino-acid sites. If these interactions are important, then haplotypes are of direct biological relevance. Relatively few cases of directly showing functional interaction exist, but there are convincing ways to demonstrate such interactions from population-level data. ApoE is one example of a protein whose function is influenced by a pair of polymorphic amino acids. The major alleles ϵ2, ϵ3, and ϵ4 differ at two amino-acid residues, but the fourth haplotype is missing from the population [Fullerton et al., 2000]. The two-site haplotypes that do exist have functional differences that are best described by the ϵ2, ϵ3, and ϵ4 allelic classes rather than by the individual SNPs in the gene. Similarly, many transmembrane xenobiotic transporters exhibit structural interactions between two or more residues [Leabman et al., 2003].

GENETIC VARIATION IN POPULATIONS IS INTRINSICALLY ORGANIZED INTO HAPLOTYPES

Each new mutation arises on a particular haplotype background. The haplotype bearing the novel mutation may rise to high frequency by random genetic drift, and it may subsequently be cleaved into segments by recombination. The combination of mutation, drift, selection, migration/population mixing, and recombination results in genetic variation that has a strong segmentwise haplotype structure to it. Population geneticists have a good handle on the determinants of linkage disequilibrium, but predictions about haplotypes go beyond inferences of pairwise LD. Of course, any factor that erodes pairwise LD will also break up haplotype structuring, but the two features are not inseparable. Strong haplotype structure arises from multiple sites having shared ancestry, and this is best imagined by considering the topology of the gene genealogy. We will return to this, but the point here is that haplotypes arise as an intrinsic attribute of population genetic variation.

HAPLOTYPES REDUCE THE DIMENSION OF ASSOCIATION TESTS, AND MAY GAIN STATISTICAL POWER

This is easiest to see by example. Suppose a gene has 8 SNPs, and you want to test for associations in a way that allows any and all of these SNPs to interact in their effects on disease causation. You may start by testing each of the 8 SNPs, asking whether the incidence of cases and controls (or measures of a physiological risk factor) differs among the 3 genotypes for each SNP. You may then test all pairs of SNPs, and for each pair you construct a 3 × 3 table of genotypes {(*AABB*, *AABb*, *Aabb*), (*AaBB*, *AaBb*, *Aabb*), (*aaBB*, *aaBb*, *aabb*)}, and within each cell of this table you assemble the observed phenotypic distribution (or counts of cases and controls). Any of a number of standard statistical tests would let you ask whether the two SNPs impact the phenotype and whether they do so independent of one another or in a synergistic manner. Note that the true genotypic complexity is greater than this, because the test just described pools the *cis*- and *trans*-phase doubly heterozygous genotypes into the *AaBb* class. In other words, it ignores the linkage phase. One can systematically and exhaustively test the null hypothesis of equal phenotypic means for all partitions of the genotype classes, and this approach, known as the combinatorial partitioning method [Nelson et al., 2001], has an appeal for its exhaustiveness. This procedure can be continued: with three SNPs there are now 27 genotypic classes, and so on. By the time you consider 8 SNPs, there are 3^8 possible genotype classes, and the test has a very large number of degrees of freedom.

By the time one considers 4 or more SNPs at a time, most of the genotypic classes will have an observed count of zero, and so testing the significance of interactions is difficult. Rather than thinking of this problem as a contingency table, it makes more sense to realize that the SNPs do not arise independently, but rather there is an intrinsic dependency of SNPs one with another due to the population history of their entry into the population. By explicitly considering the haplotype structure of the SNPs, and how they arose in the population through a genealogical process (a gene tree), one no longer needs to consider this astronomical number of potential genotypes. In this way, using haplotype information may collapse the dimensionality of the statistical test, and thereby gain statistical power over tests that do not reduce dimensions first

[Templeton et al., 1987]. In addition, Morris and Kaplan [2002] showed that haplotype-based tests can have greater power than unphased tests of association in the case when the disease locus has multiple disease-causing alleles.

WHAT ARE HAPLOTYPE BLOCKS AND HOW DO THEY ARISE?

Haplotype blocks are not a topic from classical population genetics theory, although the notion that only a small subset of all possible arrangements of segregating sites will be seen in a population sample was thoroughly understood in both the sampling theory of Ewens [1972] and in the analysis of the infinite-sites model [Watterson, 1975]. What was not clearly established in the classical theory was the discreteness of the boundaries of the regions of low haplotype diversity. Part of the reason that it is difficult to trace the theoretical predictions about haplotype blocks is that the idea arose from empirical observations, and then the term acquired more than one operational definition based on these observations.

EMPIRICAL OBSERVATIONS

Population genetics theory is concerned with the mathematical understanding of variation in a population, whereas the empiricist starts with material that may not reflect the mathematics of perfect ascertainment. In the case of SNPs, the disconnect comes from the need to consider SNPs for which there is any hope of testing association, and the increased uncertainty that rare SNPs are valid. This made human geneticists want to consider only SNPs above some frequency threshold. When this was done, several studies in rapid succession identified a strong pattern of variation that soon received the name "haplotype blocks" [Daly et al., 2001; Gabriel et al., 2002; Dawson et al., 2002; Phillips et al., 2003; Schwartz et al., 2003]. Phased genotype data were obtained in several ways, perhaps most unambiguously by making human-rodent hybrid cells and collecting hybrid cells with only one human chromosome 21 [Patil et al., 2001]. By genotyping in cell lines with only a single human chromosome, these investigators were able to determine the linkage phase across the entirety of chromosome 21. The existence of the blocky pattern was soon seen to be a boon to mapping by linkage disequilibrium, because it meant that common genomic variation was organized in blocks where information about any SNP in the block applied, in a statistical correlation sense, to the whole block. The haplotype block idea has also stimulated theoretical work on the inference of haplotype phase, identification of haplotype blocks, identifying a smaller set of tagSNPs to serve as a proxy to the SNP variation of a whole haplotype block, and testing association between a phenotype and haplotypes.

NEUTRAL COALESCENT AND HAPLOTYPE BLOCKS

A powerful approach to modeling DNA sequence variation within a population sample is the neutral coalescent [Kingman, 1980; Hudson, 1990; Nordborg and Tavaré, 2002]. This theory considers a collection of n alleles sampled today, and asks how many ancestral copies there were in previous generations. As one goes back in time, there will be n ancestral lineages for a while, but at some point, one copy of one of those alleles gave rise to two of the lineages observed today. This event is called a coalescence, because when looking backward in time, the n lineages decreased to $n-1$ lineages. The formal theory of the neutral coalescent derives a simple formula for the distribution of times back to these coalescence events (it turns out to be an exponential distribution). The case of zero recombination produces a perfectly bifurcating genealogy, and this allows an elegant mathematical treatment of many attributes of the relationships among haplotypes and suggests an algorithm for highly efficient haplotype inference [Bafna et al., 2003a,b]. When the gene segment being considered has had recombination, then a convenient way to represent the gene genealogy is an ancestral recombination graph (ARG) [Griffiths and Marjoram, 1997]. Thinking forward in time, a recombination event occurs when two different alleles undergo an exchange event resulting in a single allele that has bits from the two original alleles. Reverse the flow of time and you have the single current allele splitting into the two parental alleles. With the nonrecombining neutral coalescent, the tree shows a monotonic decline in the number of lineages. With recombination, the number of distinct lineages can increase in the short term as one goes back in time. But the increase due to recombinations occurs only as a linear process, and there is a strong exponential decay in the number of lineages due to coalescence (drift), and so eventually, even with high rates of recombination, there will be a single ancestral sequence [Wiuf and Hein, 1997]. Several features of ancestral recombination graphs and their representation of the neutral coalescent in the face of recombination are explained in Figure 1.

HAPLOTYPE STRUCTURE OF CANDIDATE GENES

Before we consider the theoretical predictions of the blockiness of the haplotype structure of human variation, it is useful to first consider the haplotype patterns of candidate genes. Prior to considering the haplotype block structure of long chromosomal segments, several studies had examined by resequencing the variation within candidate genes, including the beta globin studies from John Todd's laboratory [Harding et al., 1997], lipoprotein lipase [Clark et al., 1998], ZFX [Jaruzelska et al., 1999], and apolipoprotein E [Fullerton et al., 2000]. The usefulness of having complete resequencing data became very clear to the whole community, and the National Institutes of Health supported massive resequencing of candidate genes [e.g., Crawford et al., 2004], making the data for dense SNP discovery in these candidate genes as well as inferred haplotype phasing widely available.

In the private sector, Genaissance Pharmaceuticals was especially active in this arena, resequencing and cataloging SNPs and haplotype structure in 3,950 genes [Stephens et al., 2001a; Salisbury et al., 2003]. These data showed striking variation among genes in the degree to which haplotypes could be placed in a gene genealogy. Intragenic recombination makes this task difficult to impossible, depending on the rate of recombination and whether it is tightly clustered into hotspots. The beta globin recombination hotspot shuffled

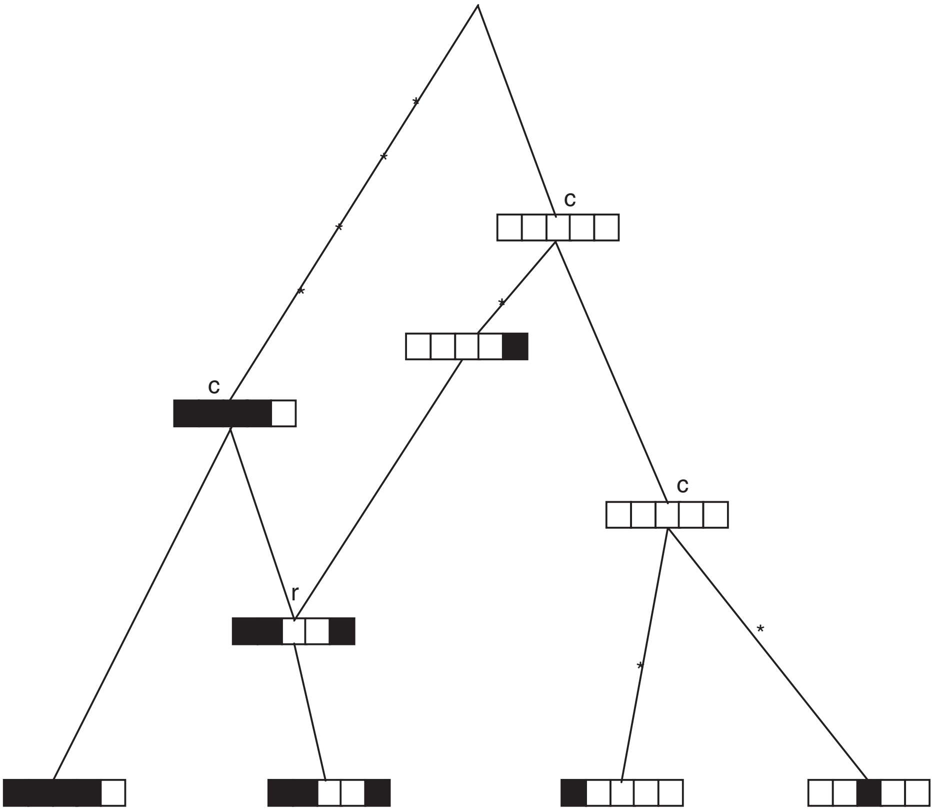

Fig. 1. Ancestral recombination graph is compact representation of ancestral history of segment of a chromosome. At bottom is present-day sample of four haplotypes, each having 5 segregating sites. As one traverses upward, time flows backwards. First event is that second haplotype comes to a node labeled "r." This is a recombination event between two chromosomes drawn along two branches that proceed up from this node. Asterisks indicate mutation events, changing an open (ancestral) square to a solid one. The second event is a joining of two rightmost haplotypes into a single haplotype. This is a coalescence event. There are two additional coalescence events, labeled "c." This ancestral recombination graph altogether has one recombination event and three coalescence events, and seven mutations. Mutations at sites 1 and 3 occur twice on this ARG, in violation of infinite-sites model.

the variation flanking it, but on one side of this hotspot, there was reasonably good cladistic structure, allowing a variety of coalescence-based estimates to be made on a subset of the data [Harding et al., 1997]. *LPL*, on the other hand, appeared to have a more diffuse recombination hotspot, making it difficult to construct any sort of genealogy for the entire gene [Clark et al., 1998], but again, segments of the gene showed good cladistic structure [Templeton et al., 2000]. *ApoE* seemed to have a more uniformly low level of intragenic recombination, because the major allelic classes of $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ were monophyletic in an inferred genealogy, and only a few sites seemed to be responsible for homoplasy due to recurrent mutation [Fullerton et al., 2000]. These different genealogical properties make an enormous difference to the kinds of haplotype-based tests of association that may be applied.

WHAT ARE THE FACTORS THAT DETERMINE BLOCKINESS OF GENETIC VARIATION?

The primary features of an ancestral recombination graph that produces runs of segregating sites with two major haplotypes are long, uninterrupted branches that occur early in the genealogy. By being long, they allow more than one mutation to occur on the same haplotype background, to produce multiple differences. Interruptions in branches may be either coalescence events, meaning that some haplotypes branch off accumulating fewer differences, or recombination events, which also clearly can break up sets of associated differences. So the question of what determines blockiness of haplotypes can be put in a population setting by asking what attributes of a population might make long, deep, uninterrupted branches in an ancestral recombination graph.

Clearly the less recombination, the fewer the interruptions to the ARG because there are fewer recombination ''bubbles.'' But the genomes of organisms clearly do recombine, and classical methods produce genetic maps that have a coherent relationship to the physical genome. But the ARG topology is affected not only by the total amount of recombination, but also by the degree of clustering of that recombination. If recombination were isolated to narrow regions of very high recombination (hotspots), separated by regions of lower recombination, this would result in longer, uninterrupted branches on the ARGs that cover the inter-hotspot regions. In genomic regions that span a hotspot, the ARG would be very complex, with many recombination bubbles. The end result is an expectation that hotspots ought to delineate the termini of haplotype blocks.

HOTSPOTS AND BLOCKS

The empirical observation of recombination hotspots in the human genome began with an inference based on a sudden drop in linkage disequilibrium in the beta globin region [Chakravarti et al., 1984]. The challenge in inferring recombination hotspots in humans is that one needs very large samples to be able to see the rare events of recombination within small regions. Polymerase chain reaction allowed amplification of DNA from single sperm, and so clever experimental designs that allow inference of rare recombinants have identified a number of regions in the human genome, with recombination hotspots having up to 1,000 times the recombination rate of the flanking regions [Huang et al., 1995; Jeffreys et al., 2000]. Even more striking, the boundaries of inferred haplotype blocks in the HLA region correspond beautifully to the locations of recombination hotspots inferred from sperm typing [Jeffreys et al., 2001]. These studies, and the statistical inference that hotspots appear to be widespread [McVean et al., 2004], strongly motivate a broader empirical assessment of the tendency of human recombination to occur in hotspots. Recent comparisons of human hotspots with the pattern of LD in chimpanzees showed that the chimpanzee does not share all human recombination hotspots [Wall et al., 2003]. The transient nature of recombination hotspots makes their underlying mechanism all the more mysterious, since human and chimp are identical at around 99% of nucleotide sites. Acquisition of a hotspot would result in very rapid erosion of local linkage disequilibrium, and loss of a hotspot would result in somewhat slower acquisition of LD through random drift. Compared to the time scale of human-chimp divergence, both of these processes would occur very rapidly, so apart from influences of natural selection, the pattern of LD would be expected to match the local recombination relatively well.

YIN-YANG HAPLOTYPE PAIRS

It is important to note that observation of haplotype blocks does not imply that the ends of the blocks must be recombination hotspots. In fact, even the neutral coalescent with perfectly homogeneous recombination will generate what appears to the eye to be a decidedly blocky pattern of haplotypes [Subrahmanyan et al., 2001]. Another commonly observed feature of haplotypes is the high frequency of haplotype pairs that differ in long runs of SNPs [Labuda et al., 2000]. Such runs can be as long as 20 SNPs or more, and the pattern appears to be exceptionally unlikely to be caused by any neutral process. Because these haplotype pairs differ at every single SNP in the run, they were dubbed ''yin-yang'' haplotypes [Zhang et al., 2003]. Such haplotypes arise as a consequence of gene genealogy, either through chance or population subdivision, having deep lineages that failed to recombine. Zietkiewicz et al. [2003] showed a striking example of ancient haplotype lineages in dystrophin that appear to predate the expansion out of Africa. Despite the striking appearance of the yin-yang haplotype pattern, simulations show that even a panmictic population may produce such high complementarity haplotype pairs by a purely neutral coalescent (Fig. 2).

VARIATION IN RECOMBINATION RATE

The inverse relationship between local rates of recombination and pairwise linkage disequilibrium has been clear since Ohta and Kimura [1971] solved the mutation-drift-recombination balance. They showed that random genetic drift results in changes in gametic frequencies to inflate the variance in LD, which is in turn eroded by recombination. These two forces come to a steady state, such that $E(r^2)=1/(4Nc+1)$, where r^2 here is a metric for LD, N is the effective population size, and c is the recombination rate. This expression suggested that $4Nc$ was an appropriate measure of the population recombination rate, since it is what determines linkage disequilibrium in a finite population. Hudson [2001] provided a means for estimation of $4Nc$, and after correcting for the ascertainment bias of the SNP Consortium SNP genotype data, Clark et al. [2003] showed that there is enormous variation across the genome in the estimate of $4Nc$. Clark et al. [2003] concluded that these estimates confound local recombination rates and local effective population size (which can vary across the genome due to other factors, such as natural selection). More recently, estimates of $4Nc$ were shown to correlate reasonably well with the local recombination rate inferred from pedigree studies [Ptak et al., 2004; McVean et al., 2004]. This remarkable finding suggests that the impact of natural selection and other demographic factors on the effective population size must be sufficiently localized that one can average across larger genomic regions and get reasonable estimates of recombination rates (on average) from patterns of LD among nearby SNPs. For practical purposes, the wide variation in local rates of recombination (and linkage disequilibrium) implies that the density of SNPs needed for equal power in association testing must be quite variable, with closely spaced SNPs in regions of high recombination and greater spacing in regions of low recombination.

NATURAL SELECTION, MATING SYSTEM, AND INBREEDING

Large haplotype blocks imply an ancestral history with constraints that go beyond simply low recombination. As mentioned above, large

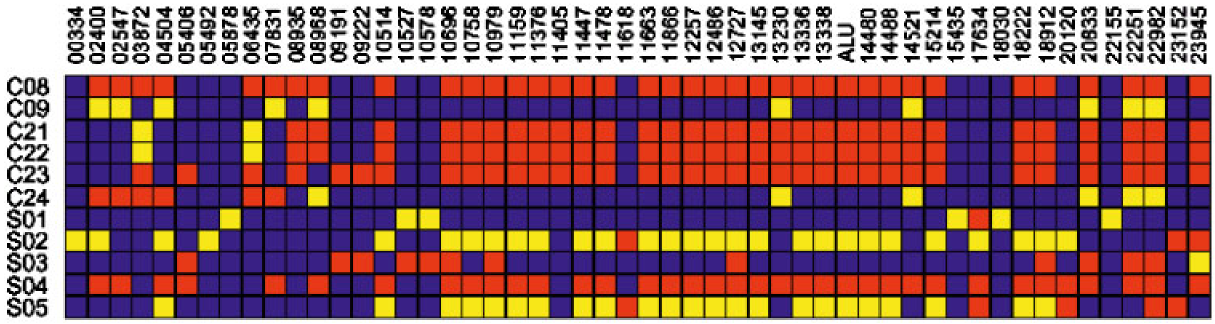

Fig. 2. Example of exceptionally strong pattern of completely mismatching runs of SNPs recently called "yin-yang" haplotypes [Zhang et al., 2004]. Despite striking pattern, these runs are often not statistically overrepresented in data. They arise as a result of coalescent genealogy having such long branches at deepest node (top two branches of Fig. 1). This example is from *DCP1* gene, encoding angiotensin-converting enzyme [Rieder et al., 1999]. Each row represents an individual, and each column is a nonsingleton SNP. The darkest gray shade (blue) depicts homozygotes for common allele; lightest gray shade (yellow), homozygotes for rare allele; medium gray shade (red), heterozygotes. [Color figure can be viewed in the journal's online edition.]

blocks arise from long, uninterrupted branches on the ancestral recombination graph, and few things do this more dramatically than demographic factors. In particular, if a population was subdivided for a very long time with zero migration, and then migration brought the two subpopulations back together, the lineages within these two populations could not undergo exchanges for the duration of the separation, and over time those long, uninterrupted branches will recede deeper into the gene genealogy. The role of natural selection in affecting linkage disequilibrium has been appreciated for many years [Lewontin, 1964]. Natural selection within a subdivided population can accelerate the process of differentiating lineages, and again would increase the chances for yin-yang patterns.

Any departure from random mating may also have a dramatic impact on the pattern and extent of linkage disequilibrium. In plants that mostly self-pollinate, the genome is mostly homozygous, and recombination events between identical (homozygous) stretches of a chromosome do not shuffle the allelic configurations within haplotypes. The end result is a very low effective level of recombination and extensive linkage disequilibrium [Nordborg et al., 2002]. In a similar fashion, some human populations engage in as many as 50% first-cousin marriages [Vardi-Saliternik et al., 2002], and this level of inbreeding is sufficient to dramatically increase LD and haplotype block lengths. Other factors, such as gene conversion and recurrent mutation, also have an impact on haplotype and LD patterns that we will not delve into further [Pritchard and Przeworkski, 2001].

INFERENCE OF HAPLOTYPE PHASE: WHY IS IT RELEVANT?

Most methods for obtaining SNP genotype data do so by testing each SNP in a way that is independent of the genotypes at other SNPs. This means that even for SNPs that are only a short distance apart in a gene, the primary data will not indicate how the pair of alleles is associated in a doubly heterozygous individual. Such data are often called “unphased” genotypes because we know the allelic state of each SNP, but we do not know the haplotype phasing. Unphased data can be used in testing association, but for all the reasons given above, it may be possible that the tests would be improved if the data were instead phased. To the extent that there is linkage disequilibrium in the population, one has some information about phase, and one might be able to do even better by explicitly tackling the statistical inference of the haplotype phase. The earliest statistical inference of the frequency of the four different haplotypes was done for the case of two biallelic loci in a single panmictic population [Hill, 1974]. If one assumes that the frequencies of genotypes made up by these haplotypes are in Hardy-Weinberg proportions, then the haplotype frequencies can be estimated from the genotype counts. For example, the frequency of the *AB* haplotype ought to be the square root of the frequency of the *AABB* genotype. If the genotype frequencies of the sample depart from Hardy-Weinberg, then a composite measure of linkage disequilibrium provides the least biased estimator [Weir and Cockerham, 1989; Schaid, 2004].

Many situations arise in genetics in which not all aspects of the data are visible to the researcher, presenting a problem of solving some aspect of the missing data. Examples include estimation of linkage disequilibrium from unphased data (since one does not have the phase of the double heterozygotes, one cannot simply count the haploid gametic types). Another example is the genotype frequencies underlying the ABO blood groups, since the recessive alleles make the A and B blood groups ambiguous with respect to genotype. Problems of these sorts have been solved with the expectation-maximization (EM) algorithm, a robust and highly efficient approach. Excoffier and Slatkin [1995] and Long et al. [1995] applied the EM algorithm to estimate the frequency of haplotypes in a population when unphased genotype counts are the input data. A limitation of this approach is that for association testing, what one really wants are the haplotype phases of individuals in the study. A nice solution to this problem comes from the realization that one has considerable prior information about phase from the genotype frequencies, and all one wants to do is infer one attribute of the data that is missing. This suggests that Bayesian methods might be the most powerful, and a couple of implementations of this approach are now widely used [Stephens et al., 2001b; Niu et al., 2002; another method for phase interference includes Niu, 2004].

The problem of haplotype phase inference is embedded in the population genetic history of haplotypes, and Clark [1990] tried to underscore this aspect by calculating and simulating the infinite-sites model to assess the efficacy of the inference of haplotype phases, using a sort of parsimony approach. The basic idea was to first identify homozygotes and single-site heterozygotes, since these provide unambiguous haplotypes that exist in the population. Then there follows a chain of "subtractions," where each unphased genotype is queried as to whether any of the phase-known haplotypes could be in the unphased genotype. Actual applications of this algorithm retain all such valid "subtractions" and trace a branching set of admissible solutions. As simple as this algorithm was, with the right combination of sample size and relatively high LD, it can perform surprisingly well. Chung and Gusfield [2003] extended the approach to prove formally that unrecombined data, which fall on a perfectly bifurcating tree, can be readily identified, and they provided a rigorous algorithm for exhaustively testing haplotype phases. There remains considerable interest in improving methods for haplotype phase inference from unphased genotype data, particularly in the context of genome-wide scans where optimal numerical procedures will be crucial [Eskin et al., 2003; Greenspan and Geiger, 2003; Kimmel and Shamier, 2004].

In some circumstances, it is possible to empirically test the phase of pairs (or more) of SNPs. One approach is to design two allele-specific primers for each SNP and to amplify the fragment between a pair of SNPs with the four possible primer pairs. This allele-specific PCR is a bit finicky, which is a problem because failure to amplify could be due to a simple PCR failure, or it could be because the DNA strand with that particular pair of nucleotides in adjacent SNPs is not present in the individual. An effective approach has been to combine phase inference with allele-specific PCR [Harding et al., 1997; Clark et al., 1998; Fullerton et al., 2000]. In some ways, the most appealing approach to haplotype phasing is to obtain sequence or SNP information from clones large enough to directly obtain phasing of multiple SNPs. This can be done on a whole-chromosome basis by constructing human-rodent hybrid cell lines bearing only a single human chromosome. By performing simple tests (like microsatellites) on a series of such cell lines, it is possible to identify cell lines specifically with the maternal chromosome copy and others with the paternal chromosome copy. By using the method of sequencing by hybridization, Patil et al. [2001] obtained phased SNP information across the entire human chromosome 21.

ASSOCIATION TESTING: WITH OR WITHOUT PHASE?

For the same reason that statistical methods for phasing genotype data work reasonably well, they may be unnecessary. That is, the unphased genotype data contain latent information about the frequencies of phased multilocus genotypes, so tests of association may not differ in power, whether or not phase is explicitly estimated. Simulations by several investigators now support the idea that explicit phase inference is not a necessary intermediate step [Kaplan and Morris, 2001; Lu et al., 2003; Morris et al., 2004]. In fact, for SNPs that depart from Hardy-Weinberg equilibrium, for whatever reason, the simplest likelihood ratio test for association, which assumes Hardy-Weinberg equilibrium, is badly biased [Schaid, 2004]. Fortunately, a composite measure of association (linkage disequilibrium) does not suffer from this bias, and has been widely available for estimation of LD in the absence of phased data [Weir and Cockerham, 1989]. For further discussion of the role of haplotype phasing in association testing, see Clayton et al. [2004].

HOW ARE HAPLOTYPES BEING USED FOR FINDING GENES UNDERLYING COMPLEX DISEASES?

TAG SNPS

One of the primary reasons cited for the International HapMap project is to gather information on the linkage disequilibrium structure of variation in human populations to be able to select an optimal (most informative) subset of SNPs for genotyping in association studies. This subset of SNPs is generally called "tag SNPs," and already a welter of methods for selecting them is available [Abecasis et al., 2001; Johnson et al., 2001; Zhang et al., 2002; Bafna et al., 2003a,b; Halldórsson et al., 2004; Weale et al., 2003; Carlson et al., 2004; see also Stram, 2004]. In principle, it makes good sense that the linkage disequilibrium structure could be used to select tag SNPs optimally, and generally the methods demonstrate that they do perform better than a random subset of SNPs. Many methods make explicit use of haplotype blocks, but it is clear that one can select tag SNPs in order to optimize power for association tests without having to first identify haplotype blocks [Bafna et al., 2003a,b; Weale et al., 2003;

Halldórsson et al., 2004]. There is sufficient linkage disequilibrium in humans that some sort of tag SNP approach is likely to result in a gain in efficiency. For candidate gene studies, one might think that identification of tag SNPs is not relevant, but in fact investigators working on candidate genes face the same problem: there are often too many SNPs to be able to afford genotyping all of them. Even for candidate genes, there is a need to select some subset of SNPs to stay under budget, while extracting the most information possible about associations with disease.

However one goes about selecting a subset of all SNPs for genotyping, there must inevitably be an erosion in the power of tests of association compared to having data on all SNPs. The big question is, how much power is lost? Cardon and Abecasis [2003] outlined the four parameters that affect an odds ratio test of association with a single SNP: 1) the odds ratio of true disease-causing SNP, 2) linkage disequilibrium between markers and the causal SNP, 3) the marker allele frequency, and 4) the disease allele frequency. Ideally one can select SNPs with high LD with any other unobserved SNP, and a range of allele frequencies to match the allele frequency of the disease alleles. Several investigators have started to assess the loss in power that occurs when various tag SNP approaches are followed, and in general the picture is fairly discouraging [Wall and Pritchard, 2003; Chapman et al., 2003; Huang et al., 2003; Fullerton et al., 2004; Zhai et al., 2004]. Zhang et al. [2004] produced the most compelling argument that selecting a subset of SNPs that retain haplotype diversity can nevertheless result in considerable loss in power of association tests, especially if risk-enhancing SNPs are low in frequency. The primary reason is that much of the effort of the HapMap project (and tag SNP selection criteria) emphasizes the use of common SNPs, and the statistical association between common SNPs and rare disease-causing alleles is weaker than that for SNPs whose frequencies more closely match the disease allele frequency [Cardon and Abecasis, 2003].

CLADISTIC OR GENEALOGY-BASED APPROACHES

Another application of haplotypes of candidate genes for testing association is to make explicit use of the gene genealogy to organize the statistical testing. This method was first articulated using the example of the alcohol dehydrogenase gene in *Drosophila* [Templeton et al., 1987], and it was extended to human genes with much promise [Haviland et al., 1995; Templeton et al., 2000; Seltman et al., 2003]. The basic idea is to construct a series of hierarchical hypothesis tests, contrasting groups of haplotypes identified by their position on the gene tree. In principle, one could cut the tree in a series of locations, and for each cut, test whether the resulting partitioning of individuals has statistically different mean phenotypes. In fact, rather than partitioning individuals, these cuts in the gene tree result in subdivisions of alleles, and individuals have a pair of alleles. But it should be clear that any cleaving of the gene tree into two classes results in two alleles, which may in turn result in three genotypes. The idea of cladistic analysis works by systematically cutting the tree in all admissible branches [Templeton et al., 1987]. As appealing as these methods are, there remains an issue of how one obtained the gene tree in the first place. Typically this is done by inference based on some model, or it may be done in a model-free way based on a principle such as parsimony [Bandelt et al., 1995]. Whatever the means to obtain the gene tree, there is some uncertainty in the process (especially when there is intragenic recombination), and often hundreds of trees would be virtually equally likely under the data. So a challenge is to adequately incorporate this tree uncertainty into the hypothesis test, and to end up with a test that is as powerful as nontree-based methods that do not need to worry about uncertainty in reconstructing a genealogy. This problem is analogous to incorporating uncertainty in haplotype phases into tests of disease association.

HAPLOTYPES FOR INFERENCE OF PAST EVOLUTIONARY HISTORY

Every disease-causing mutation arises on a single chromosome and so starts its existence in a population in association with SNP alleles on that particular chromosome. Fisher [1954] was interested in this problem, and developed a theory of "junctions" to describe the size of the unrecombined segment of the chromosome that flanks such a mutation. The theory was extended to the case of random mating populations by Stam [1980] and to small and subdivided populations by Chapman and Thompson [2003]. The idea that genetic material flanking a unique mutation would remain identical by descent, until recombination shuffled it, was seized upon by the human genetics community for finding genes associated with rare Mendelian disorders, especially in founder populations. By seeking a unique haplotype associated with cases, this approach was successful in mapping and eventually identifying genes for myotonic dystrophy [Imbert et al., 1993], cystic fibrosis [Kerem et al., 1989], and other diseases.

For monogenic disorders, it is possible to make further inferences about the past population dynamics of the allele through the study of flanking haplotypes. The example of glucose-6-phosphate dehydrogenase deficiency (G6PD) illustrates this point well. G6PD takes the sugar phosphate, G6P, and enters it into the pentose phosphate shunt in order to produce reducing potential in the form of NAPDH. Low-activity G6PD alleles were shown to confer resistance to malaria, and in fact the global distribution of low-activity alleles of G6PD coincides with that of falciparum malaria. Note that these alleles are often called "deficiencies," but in fact a G6PD null is lethal in humans, and the G6PD "deficiencies" have about 8% of normal activity. This association with malaria suggests that natural selection could have maintained G6PD deficiencies in the population, and that there might have been a period of rapid spreading of the G6PD deficiency allele. If selection did drag the G6PD deficiency alleles in, then one expects there to be greater gametic disequilibrium among the deficiency than the nondeficiency alleles, and this is exactly what is seen [Tishkoff et al., 2001; Saunders et al., 2002; Sabeti et al., 2002; Verelli et al., 2002]. Moreover, the span of the haplotypes associated with the deficiency allele is much greater than that of the nondeficiency allele, an observation that is also consistent with strong selection causing a spread in these alleles. A very similar case is found in Thailand, where the hemoglobin E variant is expanding in frequency as a result of its conferral of resistance to falciparum malaria, resulting in a strong pulse of linkage disequilibrium and reduced haplotype complexity around the mutation [Ohashi et al., 2004].

For polygenic disorders, the utility of inference of haplotypes and shared identity is less clear. The biggest challenge to finding genes associated with complex disorders is rare alleles and genetic heterogeneity. If the alleles that cause inflated risk have a frequency below around 1%, then our ability to map them with relatively common SNPs (whose frequency is 10% or greater) will be quite poor. Some investigators argued that it is likely that a large number of risk-elevating alleles will in fact be rare as a consequence of natural selection and recent rapid population growth [Pritchard, 2001]. Others argued that human population expansion and the fact that we have already identified a few relatively high-frequency risk-elevating alleles suggests that this success may not be so uncommon [Reich et al., 2002]. In any event, it is clear that the past evolutionary history of the polymorphism has direct bearing on the ease with which an association will be detected.

Because indirect tests of association rely on linkage disequilibrium, the same forces that impact linkage disequilibrium will influence the power of association tests. Past patterns of human migration established remarkable clines in linkage disequilibrium in the major histocompatibility complex and elsewhere in our genome [Cavalli-Sforza et al., 1994]. The bottleneck that seems to have occurred as ancient humans emerged from Africa to populate the rest of the planet resulted in a remarkable inflation of linkage disequilibrium out of Africa compared to within Africa [Reich et al., 2001, 2002], and we are only beginning to understand the magnitude of among-population variability in LD [e.g., Chattopadhyay et al., 2003; review in Weiss and Clark, 2002]. Population subdivision and local inbreeding can also result in dramatic increases in the size of regions that are identical by descent, otherwise known as haplotype blocks [Chapman and Thompson, 2003]. There is a general tendency for the patterns of LD among the most common SNPs to be shared more strongly among populations than are the LD patterns for more rare SNPs. This variability is a mixed blessing, because it means that the information gathered from the HapMap project will not be entirely universal, but on the other hand, the variation in LD can help in assessing the generality of associations and can improve mapping resolution. In the end, it is clear that nature could be perverse and present us with patterns of genetic variation for complex chronic diseases that will completely evade the approaches that we are bringing to bear on the problem. On the other hand, it is highly unlikely that all undiscovered contributions to complex disorders are so recalcitrant, and we can find solace in the early successes in finding the easier, ApoE-like genes first.

ACKNOWLEDGMENTS

I thank Bret Payseur and Dan Schaid for comments on the manuscript.

REFERENCES

Abecasis GR, Cookson WO, Cardon LR. 2001. The power to detect linkage disequilibrium with quantitative traits in selected samples. Am J Hum Genet 68:1463–1474.

Bafna V, Gusfield D, Lancia G, Yooseph S. 2003a. Haplotyping as perfect phylogeny: a direct approach. J Comput Biol 10:323–340.

Bafna V, Halldórsson BV, Schwartz R, Clark AG, Istrail S. 2003b. Haplotypes and informative SNP selection algorithms: don't block out information. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB '03). The Association for Computing Machinery, p 19–27.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.

Cardon LR, Abecasis GR. 2003. Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton: Princeton University Press.

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

Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31.

Chapman NH, Thompson EA. 2003. A model for the length of tracts of identity by descent in finite random mating populations. Theor Popul Biol 64:141–150.

Chattopadhyay P, Pakstis AJ, Mukherjee N, Iyengar S, Odunsi A, Okonofua F, Bonne-Tamir B, Speed W, Kidd JR, Kidd KK. 2003. Global survey of haplotype frequencies and linkage disequilibrium at the RET locus. Eur J Hum Genet 11:760–769.

Chung RH, Gusfield D. 2003. Perfect phylogeny haplotyper: haplotype inferral using a tree model. Bioinformatics 19: 780–781.

Churchill G, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet 73:285–300.

Clayton D, Chapman J, Cooper J. 2004. The use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol; in press. doi://10.1002.gepi.20032.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622.

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

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544–548.

Doerge RW, Churchill GA. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285–294.

Eskin E, Halperin E, Karp RM. 2003. Large scale reconstruction of haplotypes from genotype data. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular biology (RECOMB 03). The Association for Computing Machinery, p 104–113.

Excoffier L, Slatkin M 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Fisher RA. 1954. A fuller theory of junctions in inbreeding. Heredity 8:187–197.

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881–900.

Fullerton SM, Buchanan AV, Sonpar VA, Taylor SL, Smith JD, Carlson CS, Salomaa V, Stengård JH, Boerwinkle E, Clark AG, Nickerson DA, Weiss KM. 2004. The effects of scale: variation in the APOA1/C3/A4/A5 gene cluster. Hum Genet 115:36–56.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Greenspan G, Geiger D. 2003. Model-based inference of haplotype block variation. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03). The Association for Computing Machinery, p 131–137.

Griffiths RC, Marjoram P. 1997 An ancestral recombination graph. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag. p 257–270.

Halldórsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S. 2004. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res 14:1633–1640.

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789.

Haviland MB, Kessling AM, Davignon J, Sing CF. 1995. Cladistic analysis of the apolipoprotein AI-CIII-AIV gene cluster using a healthy French Canadian sample. I. Haploid analysis. Ann Hum Genet 59:211–231.

Huang MM, Erlich HA, Goodman MF, Arnheim N. 1995. Analysis of mutational changes at the HLA locus in single human sperm. Hum Mutat 6:303–310.

Huang Q, Fu X-Y, Boerwinkle E. 2003. Comparison of strategies for selecting single nucleotide polymorphisms for case/control association studies. Hum Genet 113:253–257.

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

Imbert G, Kretz C, Johnson K, Mandel JL. 1993. Origin of the expansion mutation in myotonic dystrophy. Nat Genet 4:72–76.

Jaruzelska J, Zietkiewicz E, Batzer M, Cole DE, Moisan JP, Scozzari R, Tavaré S, Labuda D. 1999. Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. Genetics 152:1091–1101.

Jeffreys AJ, Ritchie A, Neumann R. 2000. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum Mol Genet 9:725–733.

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

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Kaplan N, Morris R. 2001. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet Epidemiol 20:432–457.

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080.

Kimmel G, Shamir R. 2004. Maximum likelihood resolution of multi-block genotypes. In Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 04). The Association for Computing Machinery, p 2–9.

Labuda D, Zietkiewitz E, Yotova V. 2000. Archaic lineages in the history of modern humans. Genetics 156:799–808.

Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, Johns SJ, Stryke D, Kawamoto M, Urban TJ, Kroetz DL, Ferrin TE, Clark AG, Risch N, Herskowitz I, Giacomini KM. 2003. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci USA 100:5896–5901.

Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 120:849–852.

Long J, Williams RC, Urbanek M. 1995. An EM algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Gen 56:799–810.

Lu X, Niu T, Liu JS. 2003. Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. Genome Res 13:2112–2117.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11: 458–470.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169.

Niu T. 2004. Algorithms for inferring haplotypes. Genet Epidemiol; in press. doi://10.1002/gepi.20024.

Nordborg M, Tavaré S. 2002. Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90.

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193.

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K. 2004. Extended linkage disequilibrium surrounding the hemoglobin e variant due to malarial selection. Am J Hum Genet 74:1198–1208.

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

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson A, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR, Boyce-Jacino MT, Cardon LR. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

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

Ptak SE, Voelpel K, Przeworski M. 2004. Insights into recombination from patterns of linkage disequilibrium in humans. Genetics 167:387–397.

Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D, Lavery T, Kouyoumjian R, Farhadian S, Ward R, Lander E. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135–142.

Rieder MJ, Taylor SL, Clark AG, Nickerson DA. 1999. Sequence variation in the human angiotensin converting enzyme. Nat Genet 22:59–62.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC. 2003. SNP and haplotype variation in the human genome. Mutat Res 526:53–61.

Saunders MA, Hammer MF, Nachman MW. 2002 . Nucleotide variability at g6pd and the signature of malarial selection in humans. Genetics 162:1849–1861.

Schaid DJ. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S. 2003. Robustness of inference of haplotype block structure. J Comput Biol 10:13–19.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58.

Stam P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet Res 35:131–155.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. 2001a. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493.

Stephens M, Smith NJ, Donnelly P. 2001b. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Stram DO. 2004. Tag SNP selection for association studies. Genet Epidemiol; in press. doi://10.1002/gepi.20028.

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA. 2001. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am J Hum Genet 69:381–395.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351.

Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. 2000. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. Genetics 156:1259–1275.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455–462.

Vardi-Saliternik R, Friedlander Y, Cohen T. 2002. Consanguinity in a population sample of Israeli Muslim Arabs, Christian Arabs and Druze. Ann Hum Biol. 29:422–431.

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. 2002. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet 71:1112–1128.

Wall JD, Pritchard JK. 2003. Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–515.

Wall JD, Frisse LA, Hudson RR, Di Rienzo A. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am J Hum Genet 73:1330–1340.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276.

Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551–565.

Weir BS, Cockerham CC. 1989. Complete characterization of disequilibrium at two loci. In: Feldman M, editor. Mathematical evolutionary theory. Princeton: Princeton University Press. p 86–110.

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

Wiuf C, Hein J. 1997. On the number of ancestors to a DNA sequence. Genetics 147:1459–1468.

Zhai W, Todd MJ, Nielsen R. 2004. Is haplotype block identification useful for association mapping studies? Genet Epidemiol 27:80–83.

Zhang J, Rowe WL, Clark AG, Buetow KH. 2003. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet 73:1073–1081.

Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339.

Zhang W, Collins A, Morton NE. 2004. Does haplotype diversity predict power for association mapping of disease susceptibility? Hum Genet 2004; Epub ahead of print.

Zietkiewicz E, Yotova V, Gehl D, Wambach T, Arrieta I, Batzer M, Cole DE, Hechtman P, Kaplan F, Modiano D, Moisan JP, Michalski R, Labuda D. 2003. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. Am J Hum Genet 73:994–1015.