Chapter 16

SINGLE NUCLEOTIDE POLYMORPHISMS AND THEIR APPLICATIONS

Rudy Guerra and Zhaoxia Yu
Department of Statistics, Rice University, Houston, Texas, USA

1. Introduction

Biotechnology has had a tremendous impact on modern biology, especially molecular biology and genetics. Available now are genetic data of various types and resolution, including DNA sequence, genotype, haplotype, allele-sharing, gene expression, and protein expression. In addition to making these data available, biotechnology has also made possible high-throughput assays that can generate a large amount of data. A prime example is microarray technology that allows for RNA expression measurement on thousands of genes simultaneously. An important use of measured genetic data is in finding polymorphisms that underlie human disease, such as asthma, diabetes, and Alzheimers, among others. To this end, one of the most popular types of genetic data comes in the form of single nucleotide polymorphisms (SNPs), which are estimated to occur every 600-1000bp in the human genome. In this chapter we give general background on SNPs and their application in genetic association studies and haplotype reconstruction. References to SNP databases and application software are also given. The main context is human genetics.

A few of the more important definitions and concepts necessary for the discussion are given below for easy reference.

Allele Variant of a DNA sequence at a specific locus on a single chromosome, usually in reference to a gene or genetic marker.

Genotype Paired alleles of a fixed locus on homologous chromosmes.

Haplotype Allelic combination of different loci (markers, genes) on the same chromosome.

Genetic Polymorphism DNA sequence variation, typically at loci such as SNPs and sequence repeats across individuals.

Genetic Marker A segment of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. A marker can be a gene, or it can be some section of DNA with no known function. [NHGRI]

Genetic Linkage The association of genes on the same chromosome. Two genes on the same chromosome are said to be *linked*. A common measure of association between two linked genes is the recombination fraction.

Understanding the genetic basis of phenotypes, including diseases, requires an appreciation and understanding of genetic polymorphisms. Indeed, polymorphisms known as mutations contribute to many diseases and being able to detect sequence differences between normal and diseased individuals is an important step in understanding the genetics of diseases. The Human Genome Project has generated and catalogued various classes of genetic polymorphism which can be used to investigate genomic variation across individuals or groups of individuals. The largest class is the single nucleotide polymorphism, accounting for approximately 90% (Collins et al. 1998) of naturally occurring human DNA variation. The following definition of a SNP is due to Brookes (1999):

SNP: Single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater.

The popular working defintion of a SNP is a diallelic marker, but according to the above definition this is somewhat misleading (Brookes 1999). Nevertheless, there seems to be little harm in using the popular definition, by which a SNP is defined by two alleles, say A and G, at a specific location. Individuals would therefore be one of three genotypes, AA, AG, GG. It is important to remember that DNA is double-stranded and, thus, the complementarity of DNA would seem to indicate that all four nucleotides are present at the base-pair location since A and G individuals carry a T and G, respectively, on the complementary strand. However, in defining a SNP only one of the two complementary strands (Watson or Crick) is used.The typical frequency with which one observes single base differences in genomic DNA from two equivalent chromosomes is on the order of 1/1000bp (nucleotide diversity). The typical frequency of SNPs in a whole population is about 1/300bp. By screening more individuals (more chromosomes) more base differences can be found, but the nucleotide

diversity index remains unchanged. SNPs are estimated to occur about once every 1000bp although SNP density can vary by as much as 100-fold across the human genome (Brooks 1999). SNPs are found throughout the entire genome, including pormotors, exons and introns. However, since less than 5% of a person's DNA codes for proteins, most SNPs are found in non-coding sequences. The most common (2/3) SNP type is defined by alleles C and T.

Although there is much promise and hope in using SNPs for the identification of genes that determine (complex) diseases, many questions remain as to how to best use SNP data for this purpose. One common approach is the case-control study design in which the respective SNP patterns are compared (see below). This is reasonable for candidate genes or a relatively small collection of SNPs, but the ideal situation would be to compare large numbers of cases to controls with a dense set of SNPs over the entire genome. To achieve this, however, will require ultra-high-throughput assays (Isaksson et al. 2000) to discover, score, and genotype SNPs and such technologies are now being introduced, including Dynamic Allelic-Specific Hybridization (DASH) (Prince et al., 2001), reduced representation shotgun (RRS) sequencing (Altshuler et al., 2000), MALDI-TOF mass-spectrometry (e.g., Bray, Boerwinkle, and Doris, 2001), and SNP microarrays (e.g., Matsuzaki, 2004).

The identification and cataloguing of SNPs began in earnest with the Human Genome Project (Collins et a., 1998) and continues today. Beginning in the late 1990's several private efforts, such as Genset, Incyte and Celera, began to identify and build SNP databases. To ensure publicly available data for the scientific community major efforts were initiated by the National Institutes of Health (NHGRI), The SNP Consortium (TSC), and the Human Genome Variation database (HGVbase) (Brookes et al., 2000, Fredman et al., 2002). In 2001 The SNP Consortium (The International SNP Map Working Group, 2001) reported 1.4 million SNPs in the human genome, with an average density of one SNP every 1.9kb. Along with many other sources, these data were eventually deposited to a public SNP database, dbSNP (Sherry et al. 2001), maintained by the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/SNP/). As of this writing dbSNP reports (Build 122) 8.3 million SNPs in the human genome with a density of about 28 SNPs per 10kb. dbSNP also provides databases for other organisms, including mouse, rat, chicken, zebrafish, and the malaria parasite, among others. To facilitate research NCBI provides cross-annotation with resources such as PubMed, GenBank, and LocusLink. dbSNP is also included in the Entrez retrieval system for intergrated access to a number of other data databases and software tools. Several SNP databases continue to be available (e.g, HGVbase), but it appears that dbSNP will serve as the main public repository of SNPs.

The information provided by the SNP databases is a very important and valuable resource for research. Nevertheless, their usefulness is determined by quality and coverage. Since there are literally hundreds of sources that deposit SNP information into the databases, issues of quality are particularly important. The heart of the matter is whether or not a submitted SNP is real. Related issues concern SNP distribution relative to genic and non-genic regions, the allele-frequency spectrum of the SNPs, and frequency differences among racial groups. To address these and other related issues the public databases make some provisions for quality control and validation. In an independent study of public databases, Jiang et al. (2003) compared records in dbSNP and HGVbase to their own database of SNPs in pharmaceutically relevant genes (promotors, exons, introns, exon-intron boundary regions). Of the 126,300 SNPs in 6788 genes from their database, Jiang et al. matched 22,257 SNPs to HGVbase and 27,775 SNPs to dbSNP. The Jiang et al. SNPs were found by resequencing a standard cohort of 70 unrelated individuals comprising four major ethnic groups. Jiang et al. were able to verify that 60% of the public SNPs with minor allele frequencies greater than 1% were real. The remainder are thought to be of very low frequency, mismapped, or not polymorphic. No sampling bias was found with respect to ethnicity at high frequency SNPs. Jiang et al. report on seven other similar studies with confirmations ranging from 45% to 95%. Factors that differed among the studies include sample size (number chromosomes), SNP detection methods, and genome coverage. Remarkable are two studies (Mullikin et al. 2000, Altshuler et al. 2000) that each report 95% confirmation (Mullikin: 74/78 SNPs; Altshuler: 216/227 SNPs) using the reduced representation shotgun sequencing technique (Altshuler et al. 2000). The general advice is to carefully consider the sources of any SNPs extracted from the public databases. With careful scrutiny these public databases serve to provide a wealth of polymorphisms.

SNPs can be used for a variety purposes in both basic and applied research, ranging from theoretical population genetics to genetic counseling. In this article we focus on their use in genetic association studies and haplotype block reconstruction, both of which are used to help localize disease susceptibility genes. Readers interested in recombination, linkage disequilibrium, mutation, population admixture, estimation of population growth rates, and other aspects of population genetics and evolutionary history are referred to papers by Nielsen (2000), Pritchard and Przeworski (2001), Li and Stephens (2003), and Zhao et al. (2003), among others.

2. SNPs and Genotype-Phenotype Association

In many situations geneticists understand enough about a biological process that they can identify specific genes that may in part determine the

trait. Total cholesterol, for example, is a genetic trait and much is known about how cholesterol gets in and out of the bloodstream, its role in coronary artery disease, and its relationship with environmental factors (e.g., diet and exercise). See, for example, Rader, Cohen and Hobbs (2003) for a recent review of developments in the molecular pathogenesis and treatment of monogenic forms of severe hypercholesterolemia. In cases such as this biomedical researchers can statistically analyze SNPs within candidate genes for their possible association (correlation) with the trait of interest.

The possible relationship between a fixed locus and a phenotype has traditionally been studied with family data using genetic linkage analysis (Ott, 1999). In this approach the idea is to estimate through genetic recombination the physical or genetic distance between a given locus and a putative trait locus: genetic markers that co-segregate with disease status provide evidence of a nearby trait locus. One limitation of genetic linkage is that the resolution of physical distance is on the order of megabases, whereas genetic association methods are believed to have detection levels on the order of kilobases (Risch and Merikangas, 1996; Kruglyak, 1999). In contrast to genetic linkage analysis, association analysis can be based on population data (unrelateds) or family data. In a population based case-control design an association analysis evaluates the null hypothesis of equal genotypic distributions between cases and controls. If the trait is continuous a one-way analysis-of-variance can be used to evaluate a null hypothesis of equal means across the three genotypes at a SNP locus. Analogous tests can be conducted using alleles or haplotypes. If a significant association is detected the conclusion is that the given marker locus is in linkage disequilibrium (Section 4.1 below) with a susceptibility locus.

Genetic Case-Control Study Assumptions

a. There is a binary trait that defines cases and controls.

b. A random sample of n unrelated cases and a random sample of m unrelated controls are collected. The cases are independent of the controls.

c. A discrete risk factor that is obtained retrospectively is available for each case and control. In genetic studies the risk factor is usually defined by a set of alleles or genotypes of a genetic marker or gene.

To formalize the genetic case-control study we make the following typical assumptions for a binary trait that defines cases and controls.

Detailed discussion of case-control study design and analysis for genotype-phenotype associations are given in Khourny et al., (1993).

In this setting the numbers of cases and controls are fixed and the allele or genotype counts within cases and controls are random. The comparison can be viewed as a test of homogeneity, whereby the distributions of case and control counts are assumed equal under the null hypothesis. Table 16.1 shows a generic contingency table for genotype counts corresponding to a SNP with alleles, A and B. This approach is not unique to SNPs since any genetic marker defined by two alleles or two allelic classes can be similarly analyzed. The first row shows case counts r_0, r_1, and r_2 corresponding to genotypes with 0, 1, or 2 B-alleles, respectively, and sample size $n = r_0 + r_1 + r_2$. Similarly, $m = s_0 + s_1 + s_2$ for the control sample in the second row. The total sample size is $N = n + m$.

The natural pairing of alleles as genotypes makes the above case-control genotype table a basis for preliminary analysis. However, other table structures are possible. If one considers the fact that the alleles at a given locus are inherited from two different parents the data may be viewed as $2N$ independent observations, instead of the usual sample size of N. In this case, each individual contributes two alleles to the counts as in Table 16.2. Cases have $2r_0 + r_1$ A-alleles since each AA individual contributes two A-alleles and each AB individual contributes one A-allele. The other counts are similarly calculated. This distribution of counts also corresponds (Lewis 2002) to a multiplicative model with a k-fold increased risk for AB and an k^2 increased risk for BB.

Table 16.1. Case-control genotype counts.

	AA	AB	BB	Total
Case	r_0	r_1	r_2	n
Control	s_0	s_1	s_2	m
Total	n_0	n_1	n_2	N

Table 16.2. Case-control allele distribution.

	A	B	Total
Case	$2r_0 + r_1$	$2r_2 + r_1$	$2n$
Control	$2s_0 + s_1$	$2s_2 + s_1$	$2m$
Total	$n_1 + 2n_0$	$n_1 + 2n_2$	$2N$

Table 16.3. Case-control dominance distribution.

	AA	$AB + BB$	Total
Case	r_0	$r_1 + r_2$	n
Control	s_0	$s_1 + s_2$	m
Total	n_0	$n_1 + n_2$	N

A third tabulation of the counts is given in Table 16.3, where interest is in having allele B or not. Therefore, AB and BB genotypes are indistinguishable as far as the analysis is concerned. This exactly corresponds to an assumption of the B-allele being dominant to the A-allele, or equivalently, the A-allele being recessive to the B-allele. It is also an appropriate formulation when one is unable to distinguish the BB homozygote from the AB heterozygote, as was common, for example, when HLA typing was done by serology (Sasieni, 1997).

In all three approaches a standard chi-square test can be used to compare the observed counts to expected counts under a null hypothesis of equal distributions between cases and controls. A test of homogeneity to compare the case genotype distribution to the control genotype distribution in Table 16.1 is accomplished with the statistic,

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2$$

where the summation is over all six cells, O_i are the observed counts, and E_i are the expected counts. If cell i is defined by row j and column k, then E_{jk} is calculated as $R_j C_k/(n+m)$, where R_j and C_k are row j and column k counts, respectively. The X^2 statistic has an asymptotic chi-square distribution with two degrees of freedom. As an asymptotic distribution the usual precautions regarding sample sizes should be kept in mind (Agresti, 1990). The genotype data in Table 16.1 can also be analyzed under an additive model assumption where an r-fold increase in risk is associated with the AB genotype and a $2r$ increased risk with the BB genotype (Lewis 2002). This is also viewed as a "dosage-effect" analysis and the test can be performed with Armitage's trend test (Armitage 1955; Agresti 1990). The other two scenarios can be similarly analyzed with Tables 16.2 and 16.3 having 1 df in the χ^2 test.

Sasieni (1997) gives an excellent discussion of the interpretation and comparison of these three approaches to analyzing genetic case-control studies. Using the AA homozygote as the reference group, the genotype

Table 16.4. Case-control odds ratios.

Table	Odds ratio	Estimator
16.1	ψ_{het}	$\frac{r_1 s_0}{r_0 s_1}$
16.1	ψ_{hom}	$\frac{r_2 s_0}{r_0 s_2}$
16.2	ψ_{allele}	$\frac{(2r_2+r_1)(2s_0+s_1)}{(2r_0+r_1)(2s_2+s_1)}$
16.3	ψ_{sero}	$\frac{(r_1+r_2)s_0}{r_0(s_1+s_2)}$

counts allow for two 2-by-2 table odds ratios, another is available from the allele counts, and the "serological" table provides a fourth odds ratio as shown in Table 16.4.

The distinction between ψ_{allele} and ψ_{sero} is particularly important. The serological odds ratio, ψ_{sero} compares the odds of disease in subjects with either an AB or BB genotype, that is, exposure to at least one B allele, to that in AA subjects. The comparison is ultimately at the genotype level and there is no need to make a dominance assumption, equating risk of disease between BB homozygotes and AB heterozygotes. The allelic odds ratio, on the other hand, is a comparison at the allelic level. In the situation where the allele of interest (B) is rare, ψ_{allele} approximates the relative gene frequency in cases and controls. Sasieni (1997) remarks about the difficulty of interpreting ψ_{allele} as it is hard to imagine the risk of an allele developing the disease, and thus generally recommends against using the allelic odds ratio. Under the null hypothesis of no association betweeen the disease and the genetic locus, chi-square statistics associated with genotype and serological data are both asymptotically χ^2 with 1 df, and they are locally most powerful under certain assumptions. Factors that affect the chi-square test include Hardy-Weinberg equilibrium and co-dominance of the allele of interest. Additional discussion is given by Lewis (2002), including the important topic of population stratification that can lead to false-positive associations.

3. SNPs, Haplotypes and Genetic Association

Association studies attempt to take advantage of linkage disequilibrium, which exists at small genetic distances. Thus, one has to be quite lucky to come upon a single SNP that is in LD with a trait locus. Haplotypes of SNPs, on the other hand, cover more genetic distance and may have more statistical power to detect trait loci than a single SNP. Figure 16.1 shows an example of a two-SNP haplotype; each SNP has alleles 0 and 1.

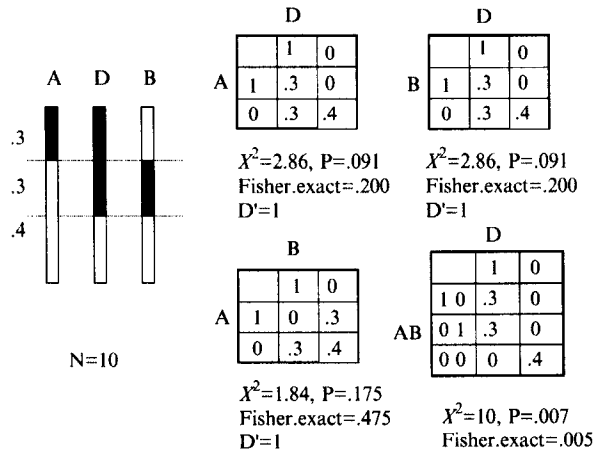

Figure 16.1. Single-SNP vs haplotype-SNP association. Each SNP has two alleles represented at black (1) and white (0) in the vertical bars. Haplotypes in the bar-figure and tables are read similarly; for example, 30% of the population has haplotype $AD = 11$. See text for additional explanation.

The vertical bar for SNP A shows that 30% of the population has allele 1, while 70% (30% + 40%) has allele 0. Considering SNP A with the disease locus, D, 30% of the population has haplotype $AD = 11$, 30% has $AD = 01$, and 40% has $AD = 00$. This information is also shown in the AD contingency table. The data for SNP B is read similarly. Using either the (asymptotic) χ^2-test or Fisher's exact test a sample of 10 individuals does not yield a significant association between AD or BD. However, when the haplotype AB is considered we do obtain a significant result; D is positive ($D = 1$) when either A or B is positive, and zero otherwise. Figure 16.2 shows another example with different allelic frquencies and sample size ($n = 50$). Here A is statistically associated with D, while B is not. The haplotype AB is strongly associated with disease. However, the results are difficult to interpret as the homozygotes of AB (11, 00) are associated with a positive disease status, while the heterozygotes are not. Of course, if one is interested in looking at interactions between SNP loci, the combinations can be evaluated regardless of linkage disequilibrium between the loci. The lesson is that there must be some basis for constructing haplotypes for association, or a more appropriate test than something like the generic χ^2-test should be considered; that is, the biology of the situation should motivate the analysis.

Several authors have considered the relative performance of single-SNP and haplotype-SNP association studies. In general, haplotypes are shown

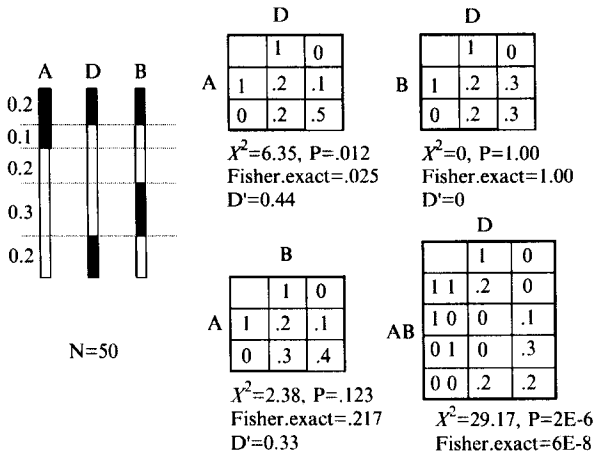

Figure 16.2. Single-SNP and haplotype-SNP association. See Fig. 16.1 for explanation.

to have better power, but see Yu and Guerra (2004) for a discussion on interpretation. Simulation studies by Service et al. (1999) and Zollner and von Haeseler (2000) . Akey and Xiong (2001) provide theoretical power calculations based on standard chi-square statistics. Morris and Kaplan (2002) consider the relative power between single-SNP and haplotype-SNP analyses in the situation where the disease locus has multiple susceptibility alleles. The answer depends on the degree of nonrandom association among the component SNPs of the haplotype; the weaker the linakge disequilibria among the SNPs, the better the haplotype analyses performs. One important class of haplotypes are those located within functional regions since they may be able to capture cis-acting susceptibility variants interacting within the same gene (Epstein and Satten 2003; Neale and Sham 2004). Related discussions are given by Hirschhorn et al. (2002)and Pritchard and Cox (2002). It is also worth mentioning that several other efforts use multiple SNP data for association, but necessarily through haplotype analysis, for example: logistic regression (Cruickshanks et al., 1992); Bayesian genomic control (Devlin and Roeder 1999); logic regression (Kooperberg et al. 2001); sums of single-SNP statistics (Hoh and Ott 2003; Hao et al. 2004); Hotelling's T^2 (Xiong et al. 2002; Fan and Knapp 2003).

3.1 Haplotype Methods for Genetic Association

As with single SNPs, when working with binary traits, a sample of haplotypes or estimated haplotype frequencies can be analyzed with a χ^2-test. If

there are only a handful of common haplotypes this approach is potentially useful; however, with more haplotype variation one encounters small cell counts which can lead to wrong p-values.

Work on quantitative traits by Long and Langley (1999) compared phase-known haplotype association to single-SNP association. Their simulations showed that haplotype-SNP tests performed no better than single-SNP tests. However, the nature of their single-SNP test (HMP: haploid marker permutation test) is not exactly based on one-marker-at-a-time methodology. The single-SNP approach is based on the SNP (among many) that shows the highest ANOVA F-statistic and its significance is evaluated by a permutation distribution obtained by permuting the (quantitative) phenotypes over the observed marker data (e.g., Churchill and Doerge 1994; Wan, Cohen, and Guerra 1997). The HMP test, therefore, is in fact a multiple-marker test which by construction only allows evaluation of a single SNP. If the collection of markers define a haplotype, then this approach may be viewed as an approximate haplotype analysis since all SNPs are being used through the identification of the strongest correlated SNP. A bona fide single-SNP analysis allows for the possibility of multiple significant SNPs. Long and Langley (1999) also consider a haploid haplotype one-way ANOVA test (HHA) whereby the means associated with haplotypes from a sample of haploid individuals are compared,

$$Y_{ij} = h_i + \epsilon_{ij}, i = 1, \ldots, H;\ j = 1, \ldots, n_i,$$

where H distinct phase-known haplotypes are observed, each with a population mean of h_i. If the error terms (ϵ_{ij}) follow a standard Gaussian distribution with equal variances, then an F-test is used, otherwise a permutation test is suggested.

ANOVA models are at the basis of other proposals. For example, Zaykin et al. (2002) propose haplotype methods for diploid individuals. For N individuals one can view the $2N$ haplotypes as $2N$ observations and analyze them according to the HHA model of Long and Langley (1999); Zaykin et al. (2002) call this approach the "$2N$-based ANOVA model." An alternative approach, the haplotype trend regression (HTR) method, maintains the paired haplotypes within each individual. The regression model is, $Y_i = D_i\beta + \epsilon_i$, where D_i is a design vector indicating the two haplotypes of individual i; D_{ij} is 1 for homozygous haplotype j, 1/2 if the heterozygous individual posseses haplotype j, and 0 otherwise. The authors suggest using permutation methods to test the null hypothesis of no haplotypic effect, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_H$, for H distinct haplotypes. Individual haplotype effects can also be tested. Phase unknown data (genotypes) are analyzed by using the Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977) as described below. In the case

of inferred haplotypes, the design matric is replaced by conditional probabilities of haplotypes, given the observed genotype data. Zaykin et al. (2002) conclude that haplotype analysis based on the HTR method can be more powerful than both the $2N$-based ANOVA method and single-SNP studies.

A generalization of the above ANOVA models with ambiguous phase SNP data is given by Schaid et al. (2002). They propose generalized linear models to accommodate qualitative or quantitative phenotypes and non-genetic covariates. Score tests are used to evaluate hypotheses of global haplotype association or haplotype-specific association.

3.2 Estimating Haplotypes with SNPs

Haplotypes ultimately determine the genetic variation in a population and thus are of interest beyond genetic association studies. One limitation of haplotype-based studies, however, is the fact that haplotypes are difficult to collect, much more so than genotypes. Although laboratory methods are available for obtaining haplotypes from multi-site genotype data (Saiki et al. 1985; Scharf et al. 1986; Newton et al 1989; Wu et al. 1989; Ruano et al. 1990) or by genotyping family members (Perlin et al. 1994, Sobel and Lange 1996) the process is still time and labor consuming. As such, inferential methods for reconstructing haplotypes from more easily acquired genotype data are desired. The nature of the combinatorial problem is easily appreciated by simply considering the fact that an individual heterozygous at n phase-unknown loci has one of the possible 2^{n-1} haplotypes. One of the major uses of SNPs is to infer haplotypes from unphased genotype data.

The first major statistical method for inferring haplotypes was introduced by Clark (1990). In this approach, genotypes of individuals with unambiguous phases are first haplotyped. The unambiguous phases are defined by sites where no more than one of the sites is heterozygous. A sequential iterative method is then used to haplotype the remaining individuals based on the haplotypes that have already been identified. The basic idea is to determine if an ambiguous haplotype could have arisen from one of the known haplotypes. Following Clark (1990), suppose one of the known haplotypes is $ATGGTAC$ and that we have a 7-site sequence $AT\{G, C\}G\{C, T\}AC$ with ambiguous third and fifth positions. This gives four possible haplotype assignments, one of which matches the known sequence $ATGGTAC$. Therefore, another count is added to $ATGGTAC$, and the homologous haplotype $ATCGCAC$ is also added to the evolving haplotype list since we must account for the observed genotype. This process continues until all data are exhausted. Matching

ambiguous haplotypes to the "known" group minimizes the total number of inferred haplotypes and thus is a type of parsimony method in the spirit of Ockam's razor. Two limitations of Clark's method are the possible lack of unambiguous haplotypes to start the process and that the results can be sensitive to the order in which individuals are added to the "known" haplotype group.

Excoffier and Slatkin (1995) proposed a maximum likelihood approach to estimate haplotype *frequencies*. See also Hawley and Kidd (1995) and Long et al. (1995). Following the notation of Stephens et al. (2001a), let $G = (G_1, G_2, ..., G_n)$ denote the observed genotypes of n sampled individuals; $H = (H_1, H_2, ..., H_n)$ their (unknown) haplotype pairs, $H_i = (h_{i1}, h_{i2})$; M the number of possible haplotypes in the population; $F = (F_1, F_2, ..., F_M)$ the set of unknown population haplotype frequencies; $f = (f_1, f_2, ..., f_M)$ the unknown sample haplotype frequencies. In this application the likelihood is a function of F, given the observed genotype data (G),

$$L(F) = Pr(G|F) = \prod_{i=1}^{n} Pr(G_i|F) = \prod_{i=1}^{n} \sum_{(h_1,h_2)\in H_i} F_{h_1} F_{h_2},$$

where H_i is the set of all ordered haplotype pairs consistent with the multisite genotype data G_i. The likelihood calculations are based on the EM algorithm under an assumption of Hardy-Weinberg equlibrium (HWE). Let p_i denote the population frequency of haplotype h_i. In the E-step the current haplotype frequencies are used to calculate phased genotype probabilities. In the M-step the haplotype frequencies that maximize the likelihood are calculated based on the updated phased genotype probabilities from the previous E-step. The haplotype frequencies are easily found by counting if the gametic phases of the observed genotype data are known, which is precisely the essence of the "missing data" in the E-step. It is important to emphasize that the EM method as described above is for estimating haplotype frequencies per se; estimation of the haplotypes themselves is not obvious (Stephens et al., 2001a).

The performance of the EM algorithm is excellent for large samples regardless of the recombination rates among the loci. Simulation studies by Fallin and Shork (2000) considered the accuracy of the EM algorithm as a function of sample size, number of loci, allele frequencies, and Hardy-Weinberg proportions. They found that the performance under diallelic diploid genotype samples was generally quite good over a wide range of parameter configurations. The largest source of error in estimating haplotype frequencies appears to be sampling error, and those thought to have a big effect - such as departures from Hardy-Weinberg proportions - were

relatively minimal. To avoid convergence to local maxima, Treqouet et al. (2004) introduced a stochastic version of the EM algorithm.

The number of loci that can be efficiently haplotyped by the EM algorithm approaches is limited. As the number of loci increases the computer space needed to store the haplotypes grows exponentially. More recent estimation methods gain on computational efficiency by using population genetic theory. For example, the coalescent (Kingman 1982) is used to better predict haplotype patterns in natural populations. By exploiting such *a priori* expectations about haplotype patterns Stephens, Smith, and Donnelly (2001a) proposed a novel class of Bayesian models for haplotype inference. In addition to significantly reducing error rates, the proposed method has the added bonus of providing a meausre of uncertainty in the phase calls. Like the EM algorithm their approach views the inference as a "missing" data problem whereby the haplotypes are treated as unobserved variables whose conditional distribution is estimated given the unphased genotype data. The method is based on a Gibbs sampler whereby an estimated stationary distribution, $Pr(H|G)$, of a Markov chain is sampled to obtain the reconstructed haplotypes.

Two other notable Bayesian contributions have been made following the ideas of Stephens et al. (2001a). Niu et al. (2002) introduced prior annealing and partition ligation to enhance the speed of haplotype reconstruction. Prior annealing protects the algorithm from converging to a local maximum, while partition ligation addresses the difficult problem of estimating haplotype frequencies over a large number of contiguous sites. To this end, the whole region is partitioned into several mutually exclusive shorter segments each of which can be be efficiently analyzed. Estimation of haplotype frequencies within each segment, as well as the reassemblage of the entire segment are accomplished by Gibbs sampling. Lin et al. (2002) proposed a version of the Stephens-Smith-Donnelley algorithm by suggesting two modifications. First, they account for missing data. Second, they avoid the problem of guessing haplotypes at random in the situation where an individual does not match to any known (or already inferred) haplotypes in the smaple; recall Clark's methods. To maintain the basic idea that future-sampled haplotypes should resemble what's already been observed, Lin et al. (2001) suggest looking for matches only at sites where the individual is heterozygous; clearly, homozygous sites already help to fix the haplotype reconstruction. Both simulated (Stephens et al. 2001a) and real data (Stephens et al. 2001b) demonstrated that the Stephens-Smith-Donnelly algorithm performaned better than existing methods of the time. More recently, Stephens and Donnelly (2003) compared the Stephens-Smith-Donnelly algorithm to those of Niu et al. (2002) and Lin et al. (2002). In addition, they introduced a new

algorithm that incorporates the modeling and computational strategies from all three Bayesian approaches. The new algorithm (PHASE 2.0) outperforms the individual algorithms when analyzing phase unknown population data. It is worth noting that the maximum likelihood estimates of haplotype frequencies coincide with the mode of a posterior distribution using a Dirichlet prior. Thus, the EM algorithm approach (which is one of many ways to find MLEs) to estimating haplotype frequencies is an instance of a Bayesian method, albeit with an unrealistic prior as discussed above (Stephens et al. 2001b).

The above methods (Clark, EM, Stephens-Smith-Donnelley) seem to form the basis of many algorithms for haplotype reconstruction. Several other approaches, however, are available. The following are not exhaustive but give a flavor of the different types of approaches. The partition ligation approach has been also been applied (Qin et al. 2002) with the EM algorithm. To accomodate a large number of loci, Clayton's SNPHAP program implements a sequential method that starts with two loci and then adds one locus one at a time until completion. Researchers from computer science and engineering have also considered haplotype reconstruction. Gusfield (2002) and Eskin et al. (2003) propose phylogenetic approaches and Eronen et al. (2004) introduce a Markov chain approach aimed at long marker maps with variable linkage disequilibrium. Unlike other methods, the Markov approach does not treat the entire haplotype as a unit; instead, it specifically accommodates recombination and works with "local" haplotype fragments that may be conserved over many generations. A selection of software packages for haplotype reconstruction are listed in Section 5 of this paper.

4. SNPs and Haplotype Blocks

One problem in genomewide mapping studies is multiple testing of many markers, which can lead to a large number of false-positive genotype-phenotype correlations. Adjustments for mutliple testing can in turn be affected by correlated test statistics associated with neighboring markers. In this section we consider the problem of finding regions of low disequilibria over chromosomal segments, the so-called *haplotype block* reconstruction problem. Haplotype block structures can minimize correlation among tests and reduce the problem of multiple testing by treating each block as a single multilocus marker.

Daly et al. (2001), Patil et al. (2001), and Gabriel et al. (2002) were among the first to formalize the idea that haplotype blocks prevail over the human genome and that this structure could be found, with historical recombination a natural source of boundaries between blocks. Although

debate still exists regarding the biological plausibility of haplotype blocks, more evidence in the direction of their existence is available than not. Wall and Pritchard (2003b) provide an excellent review of these and several other issues.

4.1 Linkage Disequilibrium

To estimate the haplotype block structure on a chromosomal segment a measure of correlation between alleles on the same chromosome is needed and the one most commonly used is linkage disequilibrium (Lewontin and Kojima, 1960; Weir, 1990), also called gametic phase disequilibrium or allelic association. Consider two diallelic loci A and B with alleles A_i ($i = 1, 2$) and B_j ($j = 1, 2$), respectively. Denoting the haplotype with alleles A_i and B_j as A_iB_j, linkage disequilibrium is defined as the difference between the observed frequency (p_{ij}) of the haplotype and its expected value under the assumption of complete linkage equilibrium: $D = p_{ij} - p_{A_i}p_{B_j}$. It is well known that the range and magnitude of D is highly sensitive to the underlying allele frequencies. It is therefore more common to work with a normalized version, D', which is bounded by -1 and 1 (Lewontin, 1964), but see Lewontin (1988) for further discussion:

$$D' = \begin{cases} D/min\{p^{A_1}(1-p^{B_1}),\ p^{B_1}(1-p^{A_1})\} & D > 0 \\ D/min\{p^{A_1}p^{B_1},\ (1-p^{A_1})(1-p^{B_1})\} & D \leq 0. \end{cases} \quad (16.1)$$

The superscripts should be subscripts.

Over many generations LD between two loci diminishes due to recombination. However, the rate of erosion depends on the genetic distance between the two loci, with closer loci maintaining their LD longer than loci farther apart. Letting D_t denote LD at generation t, the relationship between LD and recombination is $D_t = (1-\theta)^t D_0$, where θ is the recombination rate per site per generation, and D_0 is the initial LD (Ott, 1992). In practice the sign of D' is of little interest and thus $|D'|$ is often reported. When $|D'|$ is 0, the two loci are said to be in linkage equilibrium or completely independent; when $|D'|$ is at its maximum value of 1, the two loci are said to be completely linked. Values of $|D'|$ between 0 and 1 are not so easily interpreted. The natural estimator of D ($|D'|$) is obtained by substituting observed haplotype and allele frequencies from the sample. Although D (D') has an appealing definition and simple estimator, several shortcomings are recognized. The distribution of $|D'|$ can only be roughly approximated, and estimates of $|D'|$ show high variation even between pairs of sites that are far from each other (Hudson 1985). Devlin and Risch (1995) give an overview LD and many alternative measures.

is a

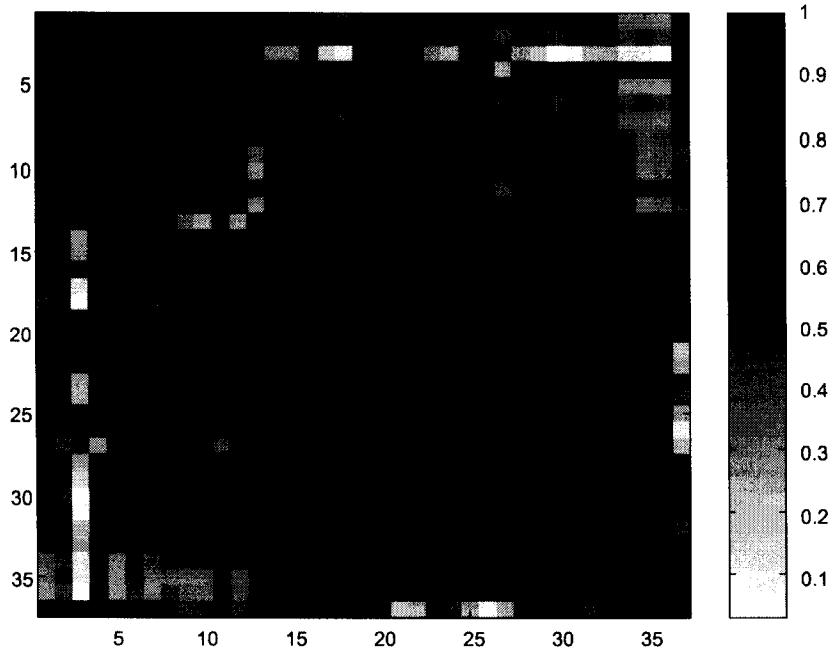

Figure 16.3. LD heatmap as measured by D' on 37 SNPs from dataset 10aA collected by Daly et al. (2001). The coordinate $(x, y) = (18, 1)$ shows LD = 0.4 between SNPs 1 and 18. The heatmap is symmetric and the four dark square regions along the diagonal represent sets of SNPs that are in high LD. Unrelated individuals and loci with MAF> 0.1 were used.

Figure 16.3 shows a heatmap of pairwise LD using D'. Represented are 37 ordered SNPs along a chromosomal segment. LD between a pair of SNPs is represented by a colored pixel. The SNPs are from dataset 10aA of Daly et al. (2001). Genotype data were collected on unrelated individuals and LD calculations were retricted to SNPs with a minor allle frequency (MAF) of at least 10%. The axes show the relative locations of the SNPs and do not account for the physical distance between SNPs. Visual inspection of the plot shows a high degree of LD in four "blocks" as indicated by four dark squares, the first covering SNPs 1-12. The block structure in heatmaps is not always so well defined and formal statistical methods are needed to objectively select the block boundaries.

Two packages that can calculate pairwise LD are GOLD (Abecasis and Cookson 2000; http://www.sph.umich.edu/csg/abecasis/GOLD/) and EMLD (Q. Huang (https://epi.mdanderson.org/ qhuang/Software/pub.htm).

4.2 Haplotype Blocks

As indicated by Fig. 16.3 there appear to be physical contiguous stretches of DNA where recombination events are relatively rare and result in blocks of high linkage disequilibria. Initial studies of this phenomenon (Daly et al., 2001; Jeffery et al., 2001; Johnson et al., 2001; Patil et al., 2001) further characterized these regions by their haplotype variation. For example, in a discussion of the block structure along a 500kb region at 5q31, Daly et al. (2001) estimate that only two haplotypes are responsible for 96% of sampled chromosomes along an 84kb stretch covered by 8 SNPs. There were ten other blocks found in the 500kb region and each one accounted for > 90% of sampled chromosomes with only 2-4 haplotypes. Moreover, within each block, none of the common haplotypes showed evidence of being derived from the others by recombination, which strongly indicates existence of a few ancestral haplotypes that tended to be inherited without recombination.

One advantage of having a block structure would be in their application to genetic association studies, which have traditionally been plagued with low polymorphism (per SNP) and multiple testing. Haplotype blocks could be treated as individual markers with a higher degree of polymorphism, while reducing the problem of multiple comparisons by minimizing redundant testing arising from markers in linkage disequilibrium.

The evidence for haplotype blocks provided by these earlier authors is quite compelling. Later research tempered the idea of general block-like structures over the entire human genome. In particular, Wall and Pritchard (2003a) and Stumpf and Goldstein (2003) discussed the role of recombination rate heterogeneity in determining block structures. The general conclusion being that block structures are likely present over parts of the genome, but not all, and that recombination occuring in narrow hotspots is likely responsible for the block structures. Given that a block-structure exists in a given chromosomal region, it nevertheless can be difficult reconstructing (estimating) the block structure. This is especially true of studies that depend on unphased genotype data on unrelated individuals. As discussed above, the haplotypes themselves have to be estimated from the genotype data, which in turn adds another level of uncertainty to the block reconstruction problem. The problem thus seems circular in that knowing the block structure would inform us on where to estimate haplotypes, but estimating the block structure requires us to have haplotype information. One way around the problem is to work with pairwise LD information over a set of SNPs in the region of interest. Ideally, a good definition of a haplotype block would depend on a measure of "regional" linkage disequilibrium (Wall and Pritchard 2003b), but in practice blocks are contructed by pairwise LD. The SNP data from Daly et al. (2001) were based on family

largely

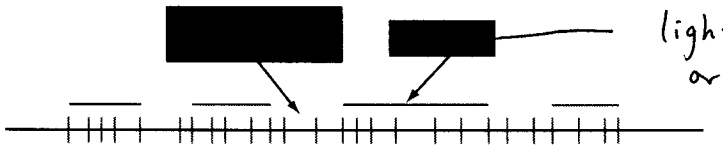

Figure 16.4. Haplotype block model. The simplest model assumes that a chromosome is naturally divided by historical recombination hot spots. The regions between hot spots are haplotype blocks, each of which exhibits limited haplotype variation.

trios, which therefore had parental transmission information not common in population based studies. Figure 16.4 shows a generic haplotype block structure. The simplest model can be developed by assuming that recombination hot spots determine the boundaries of the blocks. The success of this approach depends on the ability to directly measure or accurately estimate recombination, which can be difficult in humans.

Accurate reconstruction of haplotype blocks depends on a number of factors which should be accounted for in either the estimation method or the interpretation of results. One factor already noted is estimation of haplotype frequencies from unphased genotype data. Perhaps the most important factor is population admixture, which is known (Ewens and Spielman, 1995) to lead to false-positive results in association studies of even single markers. Gabriel et al. (2002) and Yu and Guerra (2004) show that older populations (e.g. African-Americans) tend to have more blocks of shorter length than relatively younger (Caucasian-Americans) populations. Estimated block topologies will therefore depend on the underlying population heterogeneity represented in the sample (Pritchard and Przeworski 2001; Wall and Pritchard 2003b). Additional factors include (Yu and Guerra, 2004) sample size (number of chromosomes), SNP density, minor-allele-frequency thresholds, measures of LD, and any assumptions of statistical procedures or models used to estimate the block structure. The treatment of "rare" SNPs is also important. For example, if certain SNPs are excluded from analysis due to a MAF threshold, then the inferred haplotype variation within blocks is underestimated since the rare SNPs are not represented in the estimation of haplotypes.

Several block definitions and corresponding search algorithms have been proposed. Recent reviews are given by Wall and Pritchard (2003), Yu and Guerra (2004), and Ding et al. (2005). One major feature that is common among the various block definitions is that blocks should contain limited genetic variation in that only a few haplotypes should account for most of the sampled chromosomes. Ding et al. (2005) evaluate the impact of three major operational definitions for haplotype blocks shown in the box below.

Operational Haplotype Block Definitions

Diversity Block structure defined to have low sequence "diversity" within blocks (Patil et al., 2001).

LD Requires high (low) pairwise LD within (between) blocks (Gabriel et al., 2002).

Recombination Define blocks as recombination free regions (Wang et al., 2002).

To complete the operational definitions algorithms must be defined to implement the methods. Thus, for the diversity-based method low sequence diversity may be taken to mean that a minimum fraction of all observed haplotypes are represented beyond a certain threshold in the sample; for example, Ding at al. (2005) require that within blocks 80% of the observed haplotypes occur in at least 5% of the sample. The dynamic-programming algorithm of Zhang et al. (2002) can be used to implement the method. For the LD-based approach one can define a block as a region where a minimum percentage (e.g., 90%) of all pairwise SNPs are in strong LD, which in turn must be defined using some measure of LD (e.g., D'). The recombination-based method has been implemented with the four-gamete test of Hudson and Kaplan (1985).

The first approaches to haplotype block reconstruction were proposed by Daly et al. (2001), Patil et al. (2001), Gabriel et al. (2002), and Zhang et al. (2002). Daly et al. (2001), Patil et al. (2001), and Zhang et al. (2002) propose methods based on haplotype diversity with multiple (> 2) SNPs jointly analyzed. Daly et al. (2001) define haplotype blocks as regions which have only a few (2-4) haplotypes that account for $> 90\%$ of the sampled chromosomes. Patil et al. 2001 and Zhang et al. 2002 propose methods similar to Daly et al. (2001), as well as use the idea of tagging SNPs to uniquely identify haplotypes within blocks(see below) with a minimal set of SNPs.

Gabriel et al. (2002) propose a pairwise LD approach. The pairwise LD approaches are highly intuitive and relatively easy to implement and thus seem to be the most popular of the approaches. The method is sequential by starting at one end of the chromosome (or chromosomal segment) and building the haplotype blocks one SNP at a time. If the first two (ordered) SNPs are found to be strongly correlated, the algorithm then considers addition of the third SNP, and so on. With the pairwise approach there is much flexibility in how one decides whether to add the third and later

SNPs. Gabriel et al. (2002) use pairwise linkage disequilibrium as measured by D' to assess the correlation between two SNPs. To assess the degree of LD for a give pair of SNPs they calculate 5% and 95% confidence bounds for $|D'|$ and classify the pair as: (a) *strong* LD if the lower bound is greater than 0.7 and higher bound is greater than 0.98, (b) *weak* LD if the higher bound is less than 0.9, or (c) *ambiguous*, otherwise. A haplotype block is defined to be a set of contiguous loci such that the ratio of the number of strongly linked pairs to weakly linked pairs is no less than 19-fold, ignoring the ambiguous pairs. Due to the relatively high level of variation in estimates of $|D'|$ many pairs are found to have ambiguous linkage resulting in a potential loss of information. It appears (Yu and Guerra, 2004) that many of the ambiguous cases are characterized by large differences in minor allele frequencies at the two SNPs. One possible consequence (Yu and Guerra 2004) of having too many ambiguous cases is failing to include SNPs in blocks when they should be included. In turn, this may result in more and/or shorter blocks than necessary.

No single numerical measure seems to capture the different aspects of linkage disequilibrium. Therefore haplotype block reconstruction based on the pairwise LD approach may yield different results depending on the LD measure used. Recognizing such limitations Yu and Guerra (2004) proposed a new composite summary of LD for LD-based haplotype block reconstruction. The first part of the composite LD measure is a marjority counting summarization of LD. Consider two diallelic loci A and B with two alleles A_1, A_2 and B_1, B_2. Without loss of generality, assume that A_1 and B_1 are the minor frequency alleles. We thus have two classes of haplotypes: coupling (A_1B_1, A_2B_2) and repulsion (A_1B_2, A_2B_1). Strong evidence of LD is indicated if either of the two classes is observed in high proportion. A measure of LD is thus defined and called "proportion disequilibrium" (PLD):

$$PLD = max(p_{A_1B_1} + p_{A_2B_2},\ p_{A_1B_2} + p_{A_2B_1}),$$

with the constraint $p_{A_1B_1} + p_{A_2B_2} + p_{A_1B_2} + p_{A_2B_1} = 1$. The idea underlying PLD is the same as that in finding haplotype blocks, whereby definition of low disequilibria across a block we expect a "few" haplotypes within a block to account for the vast majority of chromosomes. PLD can be used as a measure of LD and, in fact, has characteristics similar to the LD measure r (Hill and Wier, 1994; Devlon and Risch, 1995). However, PLD has much smaller variation than D'. On the other hand, like r, PLD cannot attain the value 1 (complete linkage disequilibrium) unless there are only two haplotypes; D' can attain the value of 1 when only one of the haplotype frequencies is zero.

To complete the composite summary of LD, Yu and Guerra (2004) summarize the 2 × 2 table of haplotype counts by combining PLD with another term defined as $\alpha = min\{p_{A_1B_1}, p_{A_1B_2}, p_{A_2B_1}, p_{A_2B_2}\}$. When α is very small, there are practically three haplotypes and, thus, no strong evidence of recombination between the two loci. Note that PLD on its own may only be, say, 60% indicating low LD. However, if one of the minority haplotypes is very small, we then have additional evidence in favor of strong LD. The composite LD summary is thus defined: a pair of loci are in (a) strong LD, if either $PLD > 0.95$ or $\alpha < 0.015$; (b) weak LD, if $PLD < 0.9$ or $\alpha > 0.03$; (c) ambiguous, otherwise. Let S = number of pairs in strong LD and W = number of pairs in weak LD. A haplotype block is defined as a region where $S/(S+W) > 0.98$. This algorithm starts with the first SNP and finds the longest block that satisfies the criterion; the longest block may be of length one, a single SNP. The process starts again with the first SNP following the first estimated block, and continues until the entire chromosomal segment has been analyzed.

Several other methods have been proposed for haplotype block reconstruction (Nothnagel et al. 2002; Zhu et al. 2003; Kimmel and Shamir 2004; Koivisto et al. 2004; Greenspan and Geiger, 2004). However, given the relatively short time since the first proposals appeared in 2001, it appears that the methods by Daly et al. (2001), Patil et al. (2001), and Gabriel et al. (2002) continue to be the most popular. Therefore, it is of interest to know the relative advantages and disadvantages of any new proposal with at least one of these approaches. Assessing the performance of different block definitions and search algorithms, however, is challenging due to the variety of data sets reported by different authors. When sampled populations, marker densities, sample sizes, minor allele frequency thresholds, and other factors differ from paper to paper it is indeed quite difficult to compare methods. One difficulty in assessing the relative performance of methods is defining appropriate criteria for comparison. For example, several methods are algorithmic and do not provide measures of statistical accuracy. Indeed, each reconstructed block topology is an estimate; some blocks will be statistically significant and others will not. From a statistical viewpoint measures of accuracy are needed not just to evaluate method performance, but also for practice. Simulations are also useful for comparison, but in this case much care is required to estimate realistic evolutionary forces that determine haplotype variation. One very useful model is the coalescent (Hudson, 2002), which has been used in haplotype block model simulations (e.g., Wall and Pritchard 2003). Fallin and Schork (2000) used Dirichlet distributions to simulate haplotype frequencies. Issues in comparing alternative block estimation methods have been discussed by Wall and Pritchard (2003), Schwartz et al. (2003), Yu and Guerra (2004), and Ding et al. (2005).

Wall and Pritchard (2003a) proposed three criteria as a test of fitness for a block model; they can also be used to assess the performance of block identification algorithms. The criteria are: (1) Coverage, the percentage of physical distance covered by blocks. A region under a true block structure should largely be covered by blocks. (2) Holes. If loci A, B and C are physically ordered as $A-B-C$, then a hole occurs if A and C are in strong LD but either A and B or B and C are in weak LD. A true block structure should show few holes. (3) Overlapping blocks. Two blocks are said to be overlapping if at least one site can be identified with both blocks. There should be no overlapping blocks in a true block structure. Schwartz et al. (2003) consider the problem of comparing two competing block partitions from two different methods. They use the number of shared boundaries between the two partitions as a statistic for comparison. To this end, they provide a P-value formula for the null hypothesis that two block partitions were determined randomly and independently from one another. They also suggest using this statistic in testing the robustness of a particular block partition method since most methods can give several equally good solutions. The proposed statistic is intriguing and as far as we know is the first approach suggested for formally comparing competing block partitions. This is a very important problem, especially as interest grows in using haplotype blocks for association studies. One point that warrants further investigation is the definition of "shared boundaries," which in Schwartz et al. (2003) appears to be absolute concordance at the SNP-location resolution. Yu and Guerra (2004) propose a more traditional Type I and Type II error rate approach to compare methods in the context of simulation studies; see below. For the region of interest, each SNP is either correctly captured by a block or not and these binary results can be used to calculate the error rates, as well as the Wall and Pritchard (2003) criteria.

Sun, Stephens, and Zhao (2003) report results on the impact of sample size and marker density on block partitions. Using real data sets (Daly et al., 2001) from African-American (group B), Janpanese, and Chinese (group C) populations, these authors show that both sample size and marker density have a substantial impact on block partitioning and tagging SNPs. Both the number of blocks and the number of tagging snps increase with sample size and/or maker density, at least with respect to the size and density in the samples, but this behavior is also reported by Yu and Guerra 2004).

4.3 Simulations

Below we briefly summarize some results from simulation studies conducted by Yu and Guerra (2004) to compare the their new approach based on the composite summary of LD with Gabriel's LD method for block

reconstruction. In both cases blocks are constructed via pairwise LD and the difference lies in how one decides on the strength of disequilibrium. Readers are referred to the original paper for simulation details and more extensive results. The coalescent (Hudson, 2002) was used to simulate haplotype data, which in turn were randomly paired to ultimately analyze unphased genotype population data. We simulated haplotype data in part based on the characteristics of real data from Gabriel et al. (2002). We assumed a constant population size $N = 10,000$ and a mutation parameter $\theta(= 4N\mu)$ set to 7.836×10^{-5} per bp (Wall and Prtichard, 2003). In the first simulation we modeled block length as an exponential distribution with mean 30kb with blocks separated by recombination hot spots of 1kb in length. In other simulations we changed the parameters to generate two hot spots each 10kb in length and one hotspot of 1kb. In each simulation we ran 50 data sets with 200 chromosomes. This resulted in 100 unphased individuals whose genotypes were subjected to haplotype block inference, using the EM algorithm (Excoffier and Slatkin 1995) for haplotype frequency estimation. Results from two simulation experiments are summarized in Tables 16.5; Figs 16.5 and 16.6 correspond to their respective true block models and instances of simulated data with estimated block structure.

of Yu and Guerra (2004)

The striking difference between the two methods is the increase in coverage and decrease in Type II error rates by the Yu and Guerra approach. The statistical power (1− Type II) of the Yu and Guerra approach is especially encourgaing with a view toward association studies. Still problematic for both methods are the hole and overlap frequencies, which in the

Table 16.5. Simulation results for block reconstruction. The true model included 50 SNPs over a 100kb region with four haplotype blocks separated by three recombination hot spots. Cell values are mean percentages over 50 simulated data sets, each with 200 chromosomes. Only SNPs with observed MAF>0.1 were used. (A) Hot spot lengths (1kb, 1kb, 1kb); see Fig. 16.5. (B) Hot spot lengths (10kb, 1kb, 10kb); see Fig. 16.6.

Simulation	Coverage	Holes	Overlap	Type I	Type II
A					
Gabriel et al.	33.3%	23.5	17.3	0.47	21.77
Yu and Guerra	50.0	16.4	16.3	1.83	6.78
B					
Gabriel et al.	24.6%	25.9	15.31	0.57	22.03
Yu and Guerra	41.8	22.6	19.1	6.37	7.86

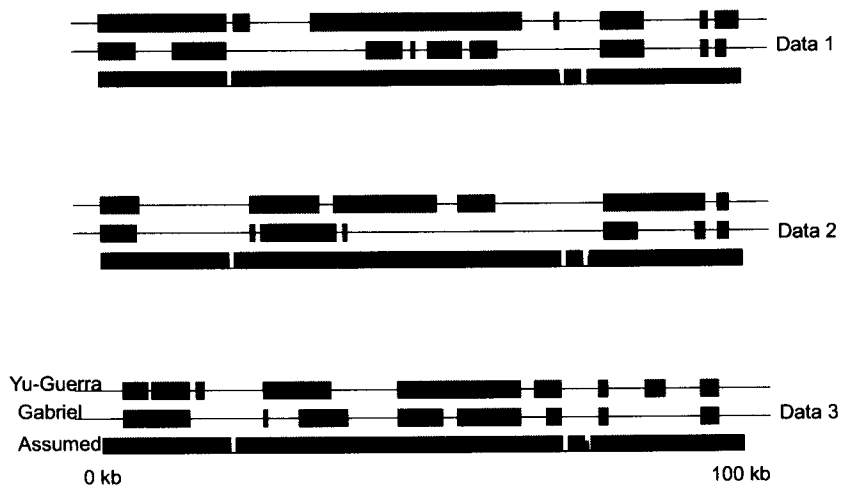

Figure 16.5. Three replications of Simulation A blocks. Hashmarks show SNP locations. See Table 16.5 caption and text for details.(Please turn to the color section in this book to view this figure in color.)

Can the heatmap (Fig. 16.3) also be shown in color? It is actually more important to show the heatmap in color than these figures. -RG

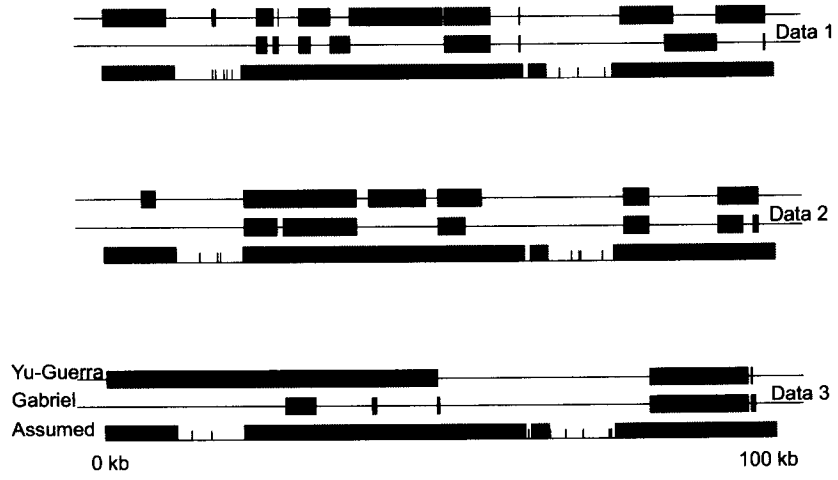

Figure 16.6. Three replications of Simulation B blocks. Hashmarks show SNP locations. See Table 16.5 caption and text for details. (Please turn to the color section in this book to view this figure in color.)

case of a true block-structure may not be expected to be at these levels. The results, however, are consistent with those of Wall and Pritchard (2003a).

4.4 Applications

We also applied the block methods of Gabriel et al. (2003) and Yu and Guerra (2004) to the four populations from Daly et al. (2001). We only used unrelated individuals and markers with MAF ≥ 0.1. The coverages for the Utah CEPH, African Americans, East Asians, and Nigerian samples using the Gabriel et al. method were 46%, 24%, 41% and 25%, respectively, which increased to 58%, 41%, 54% and 45% using the Yu and Guerra method. Hole and overlap frequencies were about the same. The apparent low coverage values may indicate a weak to moderate block-like pattern in these data and/or chromosomal region (Wall and Pritchard 2003a).

There is now much interest in haplotype-based association studies using SNPs, especially in the context of complex traits including quatitaive traits. As discussed in Section 3.1 there is some debate concerning the relative merits of single-SNP and haplotype-SNP approaches for genetic association. Intuitively, one might expect haplotypes to provide higher power. The issue is not so simple, however, especially in the case of unphased SNP genotypes where one must first estimate haplotype frequencies and/or haplotype blocks. Yu and Guerra (2004) considered the question in the context of real genes where *all* the SNPs have been identified. The real data are part of a larger study, the Dallas Heart Study (Victor et al., 2004), to investigate cardiovascular disease. We selected one of the genes being studied and simulated a continuous phenotype. We assume there is a single trait locus in complete LD with one of the SNPs (dotted line in Fig. 16.7) and model the continuous trait as a mixture of three normal distributions with means 50, 55 and 60 corresponding to genotypes CC, CA and AA, and homogeneous standard deviation of 20. Haplotype blocks were estimated using the approach of Yu and Guerra (2004). To test the single SNPs we used an additive (allelic) model defined by a predictor that counts the number of *A* alleles. The block haplotypes were also tested with an additive model using Haplo.Stat (Schaid et al., 2002). Fig. 16.7 shows the results. There were nine blocks found, including a very thin one at 30k comprised of two SNPs and a very short one at 38kb with three SNPs. There was significant single-SNP association at the assumed disease locus, as well as at neigboring SNP loci. There is a general decreasing trend in significance as the SNPs are farther away from the trait SNP. A significant global-block p-value was also found at the block covering the disease locus, although the block p-value is less than the single-SNP

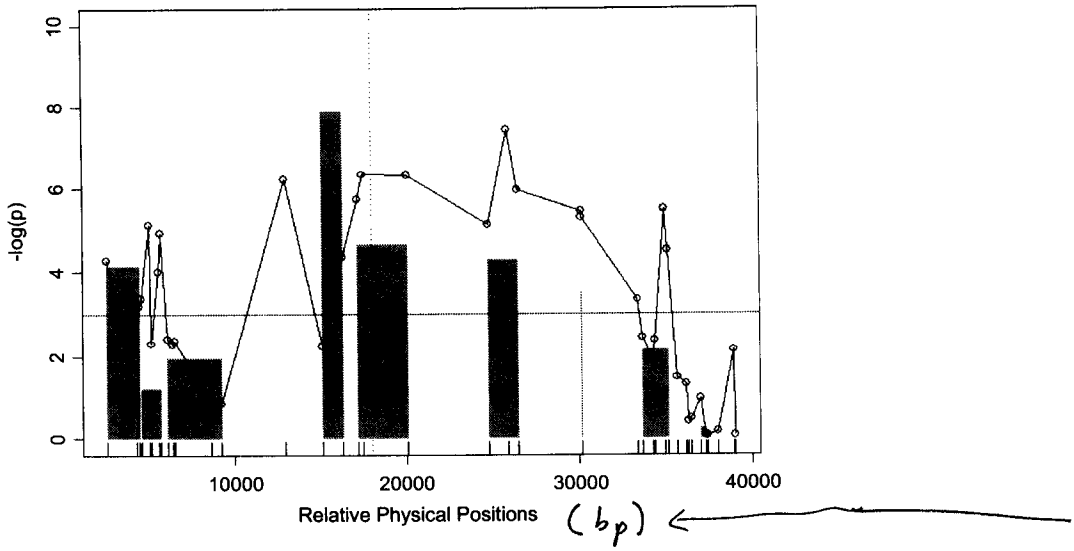

Figure 16.7. Single-SNP vs. haplotype-SNP association. The y-axis is -log(p) with $y = 3$ corresponding to a significance level of 0.05 shown as the solid horizontal line. The x-axis is physical position; SNPs are shown as hashmarks. Individual dots correspond to single-SNP tests. The eight shaded rectangles are estimated haplotype blocks and their heights correspond to a global p-value. (Please turn to the color section in this book to view this figure in color.)

p-value. In this application there is concordance (vis-a-vis significance) between single-SNP and haplotype-SNP analysis at the trait locus, as well as at the third, sixth, and seventh blocks. The other blocks, for example, block 2, show conflicting results with the single-SNPs that are within the blocks. Block 4 is highly significant with two SNPs, one significant and one not. The causes of such discrepancies are still being investiagted by various authors. See Yu and Guerra (2004) for further discussion, including results on specific haplotypes which can help explain some of the discrepancies.

4.5 Tagging SNPs

The idea of tagging SNPs (Johnson et al. 2001) is to identify a few SNPs that capture most of the variation at a haplotype locus. In a haplotype block model there would be a set of haplotype tagging SNPs (htSNPs) for each block. One advantage of htSNPs is that future sampled chromosomes need only be typed at the tagging SNPs, saving time and resources. In Fig.16.8 there are four SNPs giving four haplotypes with frequency 0.3, 0.3, 0.2 and 0.2 respectively. SNP1 and SNP2 are redundant and SNP 3 has 80%

Haplotype	SNP1	SNP2	SNP3	SNP4	Frequency
hap1	A	T	G	C	0.3
hap2	A	T	G	T	0.3
hap3	C	A	G	C	0.2
hap4	C	A	C	T	0.2

Figure 16.8. An illustration of tagging SNPs.

of the population with allele G. We can thus select as tagging SNPs (1,4) or (2,4) without loss of information.

A standard measure of haplotype variation is haplotype diversity,(D), which is defined as the total number of base differences across all possible pairwise comparisons of the observed haplotypes. If $z = (z_1, ..., z_S)$ represents a haplotype of S linked SNPs, each represented by alleles 0 and 1, the haplotype diversity based on a sample of n haplotypes is

$$D = \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i - z_j)^T (z_i - z_j).$$

Chapman et al. (2003) proposed using D to find htSNPs. Any subset of k SNPs partitions the observed haplotypes into no more than 2^k groups. By computing D within each group, we can define the residual diversity (R) of the subset as the sum of within-group diversities. The proportion of diversity explained (PDE) by the selected set of SNPs is thus defined as $PDE = 1 - R/D$ and can be used to measure the informativeness of the selected SNPs. SNP sets with a high PDE assure that not much information will be lost. Other authors propose methods that simultaneously estimate haplotype blocks and tagging SNPs. In a greedy block partition algorithm used by Patil et al. 2001, the partition that has the minimal number of SNPs that can distinguish the common haplotypes is selected. Zhang et al. (2002) note that the greedy algorithm by Patil et al. (2001) does not guarantee an optimal solution and thus propose an extended dynamic programming solution. The tagging methods discussed above all require phase information or estimated phase information, which is computationally expensive when there are a large number of SNPs. Meng et al. (2003) introduced a method based on the spectral decomposition of the LD matrix for dimension reduction. The benefit is that phase information is not needed, which makes the algorithm more efficient and eliminates the extra degree of uncertainty due to estimation of haplotype frequencies from unphased data. Section 6 gives a list of software programs that offer tagging SNP calculations.

5. Conclusions

SNPs are highly abundant in the human genome, explaining most of sequence variation. This make them a valuable resource for population genetics, evolution, and gene mapping. In this article we have given on overview of the major issues arising in their application to haplotype and haplotype block estimation and genetic association. The discussion should make clear that many statistical methods have been developed for these problems, but there is still much more to understand about the relative merits of the competing methods. Perhaps more important is further understanding of the practical utility of the methods.

6. Resources

A generally useful website for literature and software on statistical genetics, including SNPs and haplotypes, is http://linkage.rockefeller.edu/. Haplotype data from unrelated individuals can be simulated based on coalescent theory (Kingman, 1982). The *MS* package (Hudson 2002; http://home.uchicago.edu/ rhudson1/source/mksamples.html) can simulate sequences based on evolutionary forces such as mutation, recombination, and migration. In order to generate sequences with high recombination, Li and Stephens (2003) introduced a distance shrinking method to produce the effect of recombination hotspots. The package *hdhot* with block structure is available at http://www.biostat.umn.edu/ nali/SoftwareListing.html. The HapMap Project (The International HapMap Consortium, Nature 426: 789-796, www.hapmap.org) is a public database of haplotype resources. The goal of this project is to compare the genetic sequences of several hundred individuals to identify common variation; DNA samples come from populations representing African, Asian, and European ancestry.

6.1 Haplotype Reconstruction Software

If available, published papers introducing the software are given below. In all cases an internet URL is provided.

Arlequin
Excoffier and Slatkin (1995)
cmpg.unibe.ch/software/arlequin/

EH
Xie and Ott (1993)
linkage.rockefeller.edu/software/eh

EH+
Zhao, Curtis, and Sham (2000)
www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.shtml

EM-decoder of **Haplotyper**
Niu et al. (2002)
EM-decoder: www.people.fas.harvard.edu/~junliu/em/em.htm
Haplotyper: www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm

PHASE
Stephens, Smith, and Donnelly (2001a); Stephens and Donnelly (2003)
www.stat.washington.edu/stephens/software.html

PLEM
Qin, Niu, and Liu (2002)
www.people.fas.harvard.edu/~junliu/plem

SNPHAP
Clayton, D.
www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt

HAP
Eskin, Halperin, and Karp (2003)
www.cs.columbia.edu/compbio/hap/

Haplo.STAT
Schaid, et al. (2002)
www.mayo.edu/hsr/people/schaid.html

HaploBlock
Greenspan and Geiger (2003) bioinfo.cs.technion.ac.il/haploblock/

6.2 Tagging SNP Software

htSNP
Chapman, et al. (2003)
www-gene.cimr.cam.ac.uk/clayton/software/stata

HaploBlockFinder
Zhang and Jin (2003)
cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi

SNPTagger
Ke and Cardon (2003)
www.well.ox.ac.uk/~xiayi/haplotype

TagSNPs
Stram et al. (2003)
www-rcf.usc.edu/~stram/tagSNPs.htm

Hapblock
Zhang et al. (2002)
www.cmb.usc.edu/msms/HapBlock

ldSelect
Carlson et al. (2004)
droog.gs.washington.edu/ldSelect.html

Acknowledgments

Drs. Jeffrey Wall and Jonathan Pritchard kindly provided formatted data from the Daly datasets, as well as their C simulation program for the coalescent (Wall and Pritchard 2003), whose foundation is due to Li and Stephens (2003). Gudmundaut Arni Thorisson, Haplotype Map Project, provided helpful information on the databases dbSNP and hapmap. Drs. Helen Hobbs, Jonathan Cohen, and Alex Pertsemlidis provided SNP data and helpful discussion. Zhaoxia Yu is supported by a training fellowship from the W.M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology, Rice University.
Rudy Guerra is partially supported by NIH BAAHL0204 and NSF EIA0203396.

References

Abecasis GR and Cookson WO (2000) GOLD–graphical overview of linkage disequilibrium. Bioinformatics, 16(2):182-183.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Akey J, Jin L, and Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur.J.Hum.Genet. 9(4):291-300.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., and Lander, E. S.(2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature, 407:513-516.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375-386.

Bray, M.S., Boerwinkle, E., Doris, P.A. (2001). High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum Mutat.*, 17(4):296-304.

Brookes, A,J. (1999) The essence of SNPs. *Gene* 234(2):177-186.

Brookes, A.J., Lehvslaiho, H., Siegfried, M., Boehm, J.G., Yuan, Y.P., Sarkar, C.M., Bork, P., Ortigao, F. (2000). HGBASE : A Database of SNPs and Other Variations in and around Human Genes. *Nucleic Acids Research*, 28:356-360.

Carlson C.S., Eberle M.A., Rieder M.J., Yi Q., Kruglyak L., Nickerson D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *American Journal of Human Genetics*, 74:106-120.

Chapman, J.M., Cooper, J.D., Todd, J.A., and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity*, 56:18-31.

Churchill G.A. and Doerge R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963-971.

Clark, A.G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, 7:111-122.

Collins FS, Brooks LD, and Chakravarti A (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, 8(12):1229-1231.

Cousin, E., Genin, E., Mace, S., Ricard, S., Chansac, C., del Zompo, M., and Deleuze, J. F.(2003). Association studies in candidate genes: strategies to select SNPs to be tested. *Hum.Hered.*, 56:151-159.

Cruickshanks, K. J., Vadheim, C. M., Moss, S. E., Roth, M. P., Riley, W. J., Maclaren, N. K., Langfield, D., Sparkes, R. S., Klein, R., and Rotter, J. I. (1992). Genetic marker associations with proliferative retinopathy in persons diagnosed with diabetes before 30 yr of age. *Diabetes*, 41: 879-885.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S.(2001). High-resolution haplotype structure in the human genome. *Nat.Genet.*, 29:229-232.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39:1-22.

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

Devlin, B. and Roeder, K. (1999). *Biometrics*, 55(4):997-1004.

Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X., and Shen, Y.(2005). The Effect of Haplotype-Block Definitions on Inference of Haplotype-Block Structure and htSNPs Selection. *Mol. Biol. Evol.*, 22:148-159.

Epstein, M. P. and Satten, G. A.(2003). Inference on haplotype effects in case-control studies using unphased genotype data. *Am.J.Hum.Genet.*, 73:1316-1329.

Eronen, L., Geerts, F. and Toivonen, H. (2004). A Markov Chain approach to reconstruction of long haplotypes. *Pacific Symposium on Biocomputing*, 9:104-115.

Eskin, E., Halperin, E., Karp, R.M. (2003). Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinform. Comput. Biol.*, 1(1):1-20.

Ewens, W. J. and Spielman, R. S.(1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, 57: 455-464.

Excoffier, L. and Slatkin, M. (1995). Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Mol. Biol. Evol.*, 12(5):921-927.

Fallin, D. and Schork, N. J.(2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am.J.Hum.Genet.*, 67:947-959.

Fan, R. and Knapp, M.(2003). Genome association studies of complex diseases by case-control designs. *Am.J.Hum.Genet.*, 72:850-868.

Fredman, D., Siegfried, M., Yuan, Y.P., Bork, P., Lehvslaiho, H., Brookes, A.J. (2002). HGVbase : A human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research*, 30:387-91.

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, Copper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science*, 296:2225-2229.

Garner C., and Slatkin M. (2003). On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. *Genet Epidemiol.*, 24(1):57-67.

Greenspan, G. and Geiger, D. (2004). Model-based inference of haplotype block variation. *Journal of Computational Biology*, 11(2-3): 493-504.

Gusfield, D. (2002). Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *Proceedings of the Sixth Annual Inter-*

national Conference on Computational biology, 166-175. ACM Press.

Hao, K., Xu, X., Laird, N., Wang, X., and Xu, X.(2004). Power estimation of multiple SNP association test of case-control study and application. *Genet.Epidemiol.*, 26:22-30.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., Clegg, J.B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, 60:772-789.

Hawley, M. E. and Kidd, K. K.(1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J.Hered.*, 86:409-411.

Hedrick, P.W. (1987) Gametic disequilibrium measures: Proceed with caution. *Genetics*, 117:331-341.

Hill, W. G. and Weir, B. S.(1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am.J.Hum.Genet.*, 54:705-714.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K.(2002). A comprehensive review of genetic association studies. *Genet.Med.*, 4: 45-61.

Hoh, J. and Ott, J.(2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat.Rev.Genet.*, 4:701-709.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183-201.

Hudson, R. R.(1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, 109:611-631.

Hudson, R.R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159: 1805-1817.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, 18:337-8.

Hudson, R. R. and Kaplan, N. L.(1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147-164.

Isaksson, A., Landegren, U., Syvanen, A. C., Bork, P., Stein, C., Ortigao, F., and Brookes, A. J.(2000). Discovery, scoring and utilization of human single nucleotide polymorphisms: a multidisciplinary problem. *Eur.J.Hum.Genet.*, 8:154-156.

Jiang, R., Duan, J., Windemuth, A., Stephens, J. C., Judson, R., and Xu, C.(2003). Genome-wide evaluation of the public SNP databases. *Pharmacogenomics*, 4:779-789.

Johnson, G., Esposito, L., Barratt, B., Smith, A., Heward, J., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29:233-237.

Ke, X. and Cardon, L.R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, 19(2):287-8.

Khoury, M.J, Beaty, T.H., Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press.

Kimmel, G. and Shamir, R.(2005). GERBIL: Genotype resolution and block identification using likelihood. *Proc. Natl. Acad. Sci.*, 102: 158-162,

Kingman, J.F.C. (1982). The Coalescent, *Stochastic Proc. Appl.*, 13: 235-248.

Koivisto, M., Perola, M., Varilo, T., Hennah, W., Ekelund, J., et al. (2003). An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. In: R.B. Altman, A.K. Dukner, L. Hunter, T.A. Jung, and T.E. Klein (eds.), *Proceedings of the Eighth Pacific Symposium on Biocomputing (PSB'03)*, pp. 502-513, World Scientific.

Kooperberg, C., Ruczinski, I., LeBlanc, M., and Hsu, L. 2001. Sequence analysis using logic regression. *Genetic Epidemiology*, 21(S1):626-631.

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

Lewis, C. M.(2002). Genetic association studies: design, analysis and interpretation. *Brief.Bioinform.*, 3:146-153.

Lewontin, R.C., and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:458-472.

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

Lewontin R (1988) On measures of gametic disequilibrium. Genetics 120: 849-852.

Long, A. D. and Langley, C. H.(1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.*, 9:720-731.

Long, J. C., Williams, R. C., and Urbanek, M.(1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am.J.Hum.Genet.*, 56:799-810.

Li, N and Stephens, M (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213-2233.

Lin, S., Cutler, D.J., Zwick, M.E., and Chakravarti, A. (2002). Haplotype inference in random population samples. *Am. J. Hum. Genet.*, 71: 1129-1137.

Matsuzaki, H., Loi, H., Dong, S., Tsai, Y., Fang, J., et al. (2004). Parallel Genotyping of over 10,000 SNPs using a One Primer Assay on a High Density Oligonucleotide Array *Genome Research*, 3:414-25.

Meng, Z., Zaykin, D. V., Xu, C. F., Wagner, M., and Ehm, M. G.(2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am.J.Hum.Genet.*, 73:115-130.

Morris, R. W. and Kaplan, N. L.(2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet.Epidemiol.*, 23:221-233.

Mullikin, J. C., Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., et al. (2000). An SNP map of human chromosome 22. *Nature*, 407: 516-520.

Neale, B. M. and Sham, P. C.(2004). The future of association studies: gene-based analysis and replication. *Am.J.Hum.Genet.*, 75:353-362.

Newton, C. R., A. Graham, L. E. Heptinstall, S. J. Pow-Ell, C. Summers, et al. (1989). Analysis of any point mutation in DNA: the amplification refractory mutation system (ARMS). *Nucleic Acids Res.*, 17:2503-2516.

Nielsen, R.(2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931-942.

Niu T, Qin ZS, Xu X, Liu JS (2002), Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, 70(1):157-169.

Nothnagel, M., Furst, R., and Rohde, K.(2002). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum.Hered.*, 54:186-198.

Ott, J. (1999). *Analysis of Human Genetic Linkage, Third edition*. Johns Hopkins University Press.

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

Perlin, M. W., Burks, M.B., Hoop, R.C. and Hoffman, E.C. (1994). Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular distrophy. *Am. J. Hum. Genet.*, 55:777-787.

Prince, J. A., Feuk, L., Howell, W. M., Jobs, M., Emahazion, T., Blennow, K., and Brookes, A. J.(2001). Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res.*, 11: 152-162.

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

Pritchard, J. K. and Cox, N. J.(2002). The allelic architecture of human disease genes: common disease-common variant...or not? *Hum.Mol.Genet.*, 11:2417-2423.

Qin Z.S., Niu T., Liu J.S. (2002). Partition-Ligation Expectation-Maximisation algorithm for haplotype inference with singlenucleotide polymorphisms. *Am. J. Hum. Genet.*, 71(5):1242-1247.

Rader, D.J., Cohen, J.C. and Hobbs, H.H. (2003). Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. *J. Clin. Invest.*, 111:1795-1803.

Risch, N., Merikangas,K. (1996). The future of genetic studies of complex huuman diseases. *Science*, 273:1516-1517.

Ruano, G., K. K. Kidd, and J. C. Stephens. (1990). Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc. Natl. Acad. Sci.*, 87:6296-6300.

Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. (1985). Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230: 1350-1354.

Sasieni, P. D.(1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53:1253-1261.

Schaid D.J., Rowland C.M., Tines D.E., Jacobson R.M., Poland G.A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, 70:425-34.

Scharf S. J., G. T. Horn, and H. A. Erlich. (1986). Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*, 233: 1076-1078.

Schwartz, R.S., Halldórsson, B., Bafna, V., Clark, A.G., and Istrail, S. (2003) Robustness of inference of haplotype block structure. *J. of Comp. Bio.*, 10: 13-19.

Schork, N. J., Fallin, D., and Lanchbury, J. S.(2000). Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin.Genet.*, 58:250-264.

Schork, N. J., Nath, S. K., Fallin, D., and Chakravarti, A.(2000). Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. *Am.J.Hum.-Genet.*, 67:1208-1218.

Service, S. K., Lang, D. W., Freimer, N. B., and Sandkuijl, L. A.(1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.*, 64: 1728-1738.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 2000, 28:352-355

Sobel, E., Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.*, 58:1323-1337.

Stephens M. and Donnelly P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, 73: 1162-1169.

Stephens M., Smith N.J., and Donnelly, P. (2001a). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68:978-989.

Stephens M., Smith N.J., Donnelly, P. (2001b). Reply to Zhang et al., *Am. J. Hum. Genet.*, 69:912-914.

Stram, D.O., Haiman, C.A., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E., Pike, M.C. (2003). Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study, *Hum. Hered.*, 55(1):27-36.

Stumpf, M. P. and Goldstein, D. B.(2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr.Biol.*, 13:1-8.

Sun, X., Stephens, J.C., Zhao, H. (2004). The impact of sample size and marker selection on the study of haplotype structures. *Human Genomics*, 1: 179-193.

Terwilliger, J.D., Haghighi F., Hiekkalinna T.S. and Goring H. H. (2002). A bias-ed assessment of the use of SNPs in human complex traits. *Current Opinion in Genetics and Development*, 12: 726-734.

The International SNP Map Consotium (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature*, 409:928-933.

Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A., and Golmard, J. L.(2004). A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Ann. Hum. Genet.*, 68:165-177.

Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, et al. (2004). The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am. J. Cardiol.*, 93(12):1473-80.

Wall J.D. and Pritchard J.K. (2003a). Assessing the performance of the haplotype block model of linkage disequilibiurm. *Am.J.Hum.Genet.*, 73:502-515.

Wall, J,D, and Pritchard, J.K. (2003b). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Review Genetics*, 4:587-597.

Wan, Y., Cohen, J., and Guerra, R.(1997). A permutation test for the robust sib-pair linkage method. *Ann. Hum. Genet.*, 61:79-87.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L.(2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am.J.Hum.Genet.*, 71:1227-1234.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., et al. (1998). Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280: 1077-1082.

Weir, B.S. (1990). *Genetic Data Analysis. Methods for discrete population genetic data*. Sinauer, Sunderland, Mass.

Wu, D.Y., Ugozzoli, L., PAL, B.K. and WALLACE, R.R. (1989). Allele-specific amplification of P-globin genomic DNA for diagnosis of sickle-cell anemia. *Proc. Natl. Acad. Sci.*, 86:2757.

Xie, X. and Ott, J. (1993). Testing linkage disequilibrium between a disease gene and marker loci. *Am. J. Hum. Genet.*, 53:1107.

Xiong, M., Zhao, J., and Boerwinkle, E.(2002). Generalized T^2 test for genome association studies. *Am. J. Hum. Genet.*, 70:1257-1268.

Yu, Z. and Guerra, R. (2004). Haplotype blocks and association: Operating performance and new methods, TR2004-07, Department of Statistics, Rice University.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J., and Ehm, M. G.(2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum.Hered.*, 53:79-91.

Zhang, K., Calabrese, P., Nordborg, M., and Sun, F.(2002). Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.*, 71:1386-1394.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F.(2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci.*, 99:7335-7339.

Zhang, K. and Jin, L. (2003). HaploBlockFinder: haplotype block analyses. *Bioinformatics*, 19(10):1300-1301.

Zhao, J. H., Curtis, D., Sham, P. C. (2000). Model-free analysis and permutation tests for allelic associations. *Hum. Hered.*, 50(2):133-139.

Zhao, H., Pfeiffer, R., and Gail, M. H.(2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4:171-178.

Zollner, S. and von Haeseler, A.(2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, 66:615-628.

Zou G. and Zhao, H. (2003). Haplotype frequency estimation in the presence of genotyping errors. *Hum. Hered.*, 56:131-138.

Zhu, X., Yan, D., Cooper, R. S., Luke, A., Ikeda, M. A., Chang, Y. P., Weder, A., and Chakravarti, A.(2003). Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Res.*, 13:173-181.