

Algorithms for Inferring Haplotypes

Tianhua Niu*

Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Haplotype phase information in diploid organisms provides valuable information on human evolutionary history and may lead to the development of more efficient strategies to identify genetic variants that increase susceptibility to human diseases. Molecular haplotyping methods are labor-intensive, low-throughput, and very costly. Therefore, algorithms based on formal statistical theories were shown to be very effective and cost-efficient for haplotype reconstruction. This review covers 1) population-based haplotype inference methods: Clark's algorithm, expectation-maximization (EM) algorithm, coalescence-based algorithms (pseudo-Gibbs sampler and perfect/imperfect phylogeny), and partition-ligation algorithm implemented by a fully Bayesian model (Haplotyper) or by EM (PLEM); 2) family-based haplotype inference methods; 3) the handling of genotype scoring uncertainties (i.e., genotyping errors and raw two-dimensional genotype scatterplots) in inferring haplotypes; and 4) haplotype inference methods for pooled DNA samples. The advantages and limitations of each algorithm are discussed. By using simulations based on empirical data on the G6PD gene and TNFRSF5 gene, I demonstrate that different algorithms have different degrees of sensitivity to various extents of population diversities and genotyping error rates. Future development of statistical algorithms for addressing haplotype reconstruction will resort more and more to ideas based on combinatorial mathematics, graphical models, and machine learning, and they will have profound impacts on population genetics and genetic epidemiology with the advent of the human HapMap. *Genet. Epidemiol.* © 2004 Wiley-Liss, Inc.

Key words: haplotype; genotype; phase; single-nucleotide polymorphism; algorithm

Grant sponsor: National Institutes of Health; Grant number: R01 HG002518-01.

*Correspondence to: Dr. Tianhua Niu, Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 900 Commonwealth Ave., Boston, MA 02215. E-mail: tniu@rics.bwh.harvard.edu

Received 8 June 2004; Accepted 24 June 2004

Published online 14 September in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.20024

INTRODUCTION

It was estimated that more than 5 million single-nucleotide polymorphisms (SNPs) with minor-allele frequencies (MAFs) greater than 10% are expected to exist in the human genome [Carlson et al., 2004]. Although each SNP can be analyzed independently of other markers, it is much more informative to analyze markers in a region of interest simultaneously. The combination of marker alleles on a single chromosome is called a haplotype. There is great interest in understanding haplotype structures in the human genome using identified genetic markers because: 1) haplotypes provide snapshots of human evolutionary history because they are "molecular fossils" such that by reconstructing a haplotype network, information contained in haplotype frequency in the sample can be used to infer the topological position of a given haplotype in a cladogram [Crandall and Templeton, 1993]; and 2) haplotype information is crucial in estimating the age and location of

disease mutations relative to a set of multiple linked markers [Liu et al., 2001; Lu et al., 2003]. Recent empirical investigations revealed that the human genome can be partitioned into discrete haplotype blocks, such that haplotype diversity is very constrained within each block. To capitalize on the block-like feature to significantly reduce the amount of genotyping efforts for genetic association studies, the International HapMap Project was launched to produce a genome-wide human haplotype map in several ethnic populations [International HapMap Consortium, 2003].

The challenge of conducting haplotype-centric analyses is that in diploid organisms such as *Homo sapiens*, haplotypes are not directly observable, and only unphased genotype data can be obtained through application of experimental techniques. A spectrum of molecular haplotyping methods was developed, including single-molecule dilution [Stephens et al., 1990], long-range allele-specific PCR [Michalatos-Beloin et al., 1996], diploid-to-haploid conversion [Douglas et al.,

2001], carbon nanotube probing [Woolley et al., 2000], pyrosequencing [Odeberg et al., 2002], intracellular ligation [McDonald et al., 2002], rolling-circle amplification [Zhong et al., 2001], and clone-based systematic haplotyping [Burgtorf et al., 2003]. However, these methods are not widely used, not only because they incur significant costs but also because they are low-throughput, and important technical problems remain unresolved. Therefore, algorithms for inferring haplotypes from unphased genotype data offer practical, accurate, and cost-effective solutions.

POPULATION-BASED HAPLOTYPE INFERENCE

Population-based haplotyping is a major challenge for gene mapping for complex human diseases because most complex diseases are late-onset, and it is not only expensive, but also logistically difficult to recruit the parents of diseased probands into a family-based study. Therefore, it is relatively more cost-efficient and easier to collect affected probands and controls for a complex disease through a case-control design compared to a family-based design. Several large-scale, prospective cohort studies of complex human traits in the US such as the Women's Health Initiative, Women's Health Study, or Physician's Health Study recruited only unrelated individuals and have followed the study participants longitudinally over a long period of time. Analyses of genotype data of DNA samples from such cohorts also require population-based haplotype inference.

CLARK'S ALGORITHM

The earliest algorithm for haplotype reconstruction (from genotype data) was described by Clark [1990], based on the principle of maximum parsimony. This algorithm resolves the haplotypes following three steps: 1) identifying all unambiguous haplotypes (all homozygotes and single-site heterozygotes) and considering them as "resolved;" 2) determining whether each of the resolved haplotypes could be one of the alleles in the remaining yet-to-be-phased genotypes; and 3) each time a new haplotype is identified as one of the resolved ones, this new haplotype is assumed to be known, and the remaining haplotype is added to the resolved haplotype set. The rationale for this algorithm is that homozygous haplotypes are probably common, and that a phase-ambiguous genotype is likely to contain

known common haplotypes. Clark [1990] stated that when all haplotypes are resolved based on maximum parsimony, the solution is unique and correct, and the results of Clark's algorithm based on certain empirical data are shown to be reliable by comparison with haplotypes obtained by direct molecular methods [Clark et al., 1998; Rieder et al., 1999]. Gusfield [2001] reformulated Clark's statement into a "maximum resolution" (MR) problem: given a set of vectors (some ambiguous and some resolved), what is the maximum number of ambiguous vectors that can be resolved by successive application of Clark's inference rule? Gusfield [2001] proved that the MR problem is NP-hard and Max-SNP-complete, which can be reduced to an integer linear programming problem. The advantages of Clark's algorithm are that it is a relatively straightforward procedure, and it can handle a large number of loci when haplotype diversity is rather limited in the population. The disadvantages of Clark's algorithm are that: 1) the algorithm does not start when there are no homozygotes or single-site heterozygotes in the population; 2) the algorithm does not give unique solutions, because the phasing results are dependent on the order of genotypes that need to be phased (therefore, when there is a large number of distinct haplotypes compared to the sample size due to the presence of recombination hotspots, Clark's algorithm sometimes cannot resolve a relatively large fraction of heterozygous individuals); and 3) although Clark's algorithm does not explicitly assume Hardy-Weinberg equilibrium (HWE), its performance is still relatively sensitive to the extent of deviation from HWE [Niu et al., 2002].

STANDARD EM ALGORITHM

The expectation-maximization (EM) algorithm [Dempster et al., 1977] estimates population haplotype probabilities based on maximum likelihood, finding the values of the haplotype probabilities which optimize the probability of the observed data, based on the assumption of HWE. Excoffier and Slatkin [1995] were the first to discuss the use of the EM algorithm in this context. The likelihood function in the EM algorithm can be written as

$$L(\Theta) = P(G|\Theta) = \prod_{i=1}^n \sum_{(a,b): a \oplus b = g_i} \theta_a \theta_b$$

where G denotes the observed unphased genotype data for n individuals, g_i denotes the observed

unphased genotype data for the i th individual, Θ denotes the overall haplotype frequency, θ_a and θ_b denote the respective haplotype frequencies for haplotypes a and b , such that $a \oplus b = g_i$ denotes that the haplotype pair (a, b) is compatible with the i th observed genotype data $-g_i$. The EM algorithm, under the assumption of HWE, is an iterative procedure: $\theta_a^{(k+1)} = E_{\Theta^{(k)}}(n_a|G)/2n$, where $\Theta^{(k)}$ is the current estimate of haplotype frequencies, and n_a is the count of haplotype a that exists in G . It can be important that the initial value of haplotype frequencies are reasonably close to the true population frequencies. The advantages of the EM algorithm are that: 1) it is based on solid statistical theory; and 2) although the EM algorithm makes an explicit assumption of HWE, simulation studies demonstrate that its performance is not strongly affected by the departures from HWE, particularly when the direction of departure is towards an excess of homozygosity [Niu et al., 2002]. The disadvantages are that: 1) the performance is sensitive to the initial value of Θ ; 2) if there exist local maxima, the iteration may lead to locally optimal maximum likelihood estimates (MLEs), which becomes most serious when there are many distinct haplotypes (one sensible way to employ the EM algorithm is to use a good initial guess on Θ [e.g., the product of the allele frequencies, as suggested by Excoffier and Slatkin, 1995]); and 3) the standard EM algorithm cannot handle a large number of loci. Recently, a variant on the EM algorithm, the stochastic-EM algorithm, was applied to the problem of estimation haplotypes from unphased genotypes, by Tregouet et al. [2004]. This algorithm can be useful for avoiding convergence to local maxima.

COALESCENCE-BASED ALGORITHM

Pseudo-Gibbs sampler (PGS) algorithm. Stephens et al. [2001] proposed a coalescence-based Markov-chain Monte Carlo (MCMC) approach: a pseudo-Gibbs sampler (PGS) for reconstructing haplotypes from genotype data. PGS uses Gibbs sampling to obtain an approximate sample from the posterior distribution, $P(Z|G)$, where $Z = (z_1, z_2, \dots, z_n)$ and $G = (g_1, g_2, \dots, g_n)$ denote the phased and unphased (i.e., observed) genotype data for n individuals. The major piece of this iterative sampling algorithm is that at the $(k + 1)$ th iteration, we aim to sample $z_i^{(k+1)}$ from $P(z_i|G, Z_{-i}^k)$, where Z_{-i}^k is a set of phased genotypes for all the remaining $(n-1)$ subjects excluding the haplotype pair z_i for individual i , at

the k th iteration. Then, we have $P(z_i|G, Z_{-i}^k) \propto P(z_i = \alpha \oplus \beta | Z_{-i}^k) \propto \pi(\alpha | Z_{-i}) \pi(\beta | Z_{-i}, \alpha)$, where α and β denote the two respective haplotypes that form z_i . Here, $\pi(\alpha | Z_{-i})$ is essentially a prior for a future sampled haplotype, which is not known to the investigators a priori. Instead of using the Dirichlet prior [based on the “parent-independent mutation” model in Stephens et al., 2001], Stephens et al. [2001] suggested using an approximate coalescent prior: equation 17 of Stephens and Donnelly [2000], which is a stationary Markov chain with transition matrix $T_{\alpha\beta} = \frac{\theta}{r+\theta} P_{\alpha\beta} + \frac{r_\alpha}{r+\theta}$, where r and r_α are the total number of haplotypes and the total number of type α haplotypes in Z , respectively, θ is the scale mutation rate, and P is the transition matrix. The approximate coalescent prior is based on the assumption that “the genetic sequence of a mutant offspring will differ only slightly from the progenitor sequence (often by a single-base change)” [Stephens and Donnelly, 2003]. Recently, Stephens and Donnelly [2003] modified the implementation of the PGS algorithm by incorporating a variant of the partition-ligation (PL) idea [Niu et al., 2002] and by allowing for recombination and decay of linkage disequilibrium (LD) with distance. The key advantage of the PGS is that it incorporates the coalescence theory into its prior, and although the induced Markov chain has a stationary distribution that may depend on the order of g_i s, it was shown to perform well in simulations based on a coalescent model: a constant-size population evolving for a long period of time without recombination or recurrent mutations [Stephens et al., 2001; Stephens and Donnelly, 2003]. The disadvantages are that: 1) PGS is not a fully Bayesian model and it lacks a measure of the overall “goodness” of the constructed haplotypes; 2) because this algorithm makes only local moves in each iteration (i.e., a “piece-by-piece” strategy) to update a new haplotype that closely resembles an existing haplotype, PGS is quite slow and it takes millions of iterations for the algorithm (2 million iterations are suggested as the default value for PHASE version 1.0 in Stephens and Donnelly [2003]) to start to converge to the right answer [e.g., the ACE data from Rieder et al., 1999]; and 3) it remains unclear whether the algorithm performs favorably compared to other algorithms (e.g., standard EM algorithm) for admixed or rapidly expanding populations when the coalescent model does not hold, which is often the case for cosmopolitan US cohorts (e.g., a random sample from downtown Los Angeles).

For example, although Stephens et al. [2001] demonstrated that PHASE outperformed EM by a significant margin under certain conditions, Zhang et al. [2001] and Xu et al. [2002] revealed that this was not the case: PHASE and EM-based methods exhibited similar performances in their simulated datasets.

Phylogeny haplotyping (PPH) algorithm. A phylogeny-based algorithm intends to deterministically deduce haplotype phases based on phylogenetic reconstruction [Gusfield, 2002]. The coalescent model of haplotype evolution tells us that without recombination, the evolutionary history for w distinct haplotypes can be displayed as a “Perfect PHylogeny” (PPH) with w leaves, and each of the SNP sites labels exactly one edge of the tree. Rooted in the coalescent model, there are two key assumptions of PPH: 1) no recombination; and 2) the infinite-sites model. Under these two assumptions, a PPH problem is stated as: given a set S of n genotype vectors, we would like to find a PPH $T(S)$, and a pairing of the $2n$ leaves of $T(S)$ that explains S . A PPH problem can be reduced to a classical problem of recognizing graphic matroids [Tutte, 1960]. Although the general PPH problem is NP-hard [Steel, 1992] and multiple solutions can be possible [Gusfield, 2002], for binary phylogenies, PPH is shown to be linear-time solvable [Bixby and Wagner, 1988]. A program named “Perfect Phylogeny Haplotyper” [Chung and Gusfield, 2003] was developed. More recently, Halperin and Eskin [2004] developed an “imperfect” phylogeny method that extends the framework of PPH by allowing for both recurrent mutations and recombinations. In their approach, the multiple linked SNPs are partitioned into blocks, and for each block, they predict each individual’s haplotype. Then the block-based haplotypes are assembled together to form the long-range haplotype utilizing the PL idea of Niu et al. [2002]. The imperfect phylogeny method appears to be more robust than PPH, and allows the handling of missing data and resolution of a large number of SNPs. Overall, PPH appears to be an interesting application of the graphical model in haplotype inference, although their accuracies remain to be benchmarked by using extensive simulations and real datasets.

PL

In order to phase long-range haplotypes, Niu et al. [2002] introduced a divide-conquer-combine

algorithm: partition-ligation (PL), which can handle a large number of loci. This algorithm is not based on coalescence theory. The essence of PL is that a long-range haplotype is a combinatorial set of atomistic units in nature. This theoretical conception is actually supported by the empirical observations of the “block-like” patterns of human haplotypes. Nevertheless, the atomistic units of PL refer to any arbitrary segments of consecutive loci (typically ranging between 5 and 8) which are more generalized than “high-LD” blocks. It should be noted that partitions at the “high-LD” block boundaries somewhat improve the performance of PL [Niu et al., 2002], and detecting the locations of recombination hotspots before carrying out partition is an interesting thought. Once a long-range haplotype is partitioned into a series of smaller, atomistic units, haplotype phasing for each atomistic unit is tractable, and the long-range haplotype can be considered a ligated product of these atomistic units (Fig. 1). In the ligation step, two strategies can be employed [Niu et al., 2002]: 1) hierarchical ligation; and 2) progressive ligation. Haplotyper [Niu et al., 2002] and PLEM [Qin et al., 2002] implemented PL through hierarchical ligation, whereas variants of the progressive ligation were adopted by a modified version of PHASE version 1.0 and SNP HAP (written by Dr. David Clayton). Here, I illustrate the implementation of PL using either Bayesian or EM implementations through Haplotyper and PLEM.

Bayesian implementation. A Bayesian paradigm, which allows incorporation of prior information into the statistical model, has enjoyed increasing applications in genetics [Beaumont and Rannala, 2004]. Niu et al. [2002] employed a fully Bayesian model to reconstruct haplotype phases, employing two innovative techniques: PL and prior annealing. Instead of using a prior based on coalescence theory, a Dirichlet prior was used in the Gibbs sampling, such that the prior for Θ , $P(\Theta) \sim Dir(\lambda)$, where $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ is the vector of pseudocounts for m distinct haplotypes. The use of the Dirichlet prior assumes that the genetic sequence of a mutant offspring does not depend on the progenitor sequence [Stephens et al., 2001], and therefore is different from the approximate coalescent prior used by PGS. Thus, the prior does not assume that unresolved haplotypes will be *similar* to progenitor haplotypes, and therefore will not make local moves as the PGS. In essence, the Gibbs sampling of

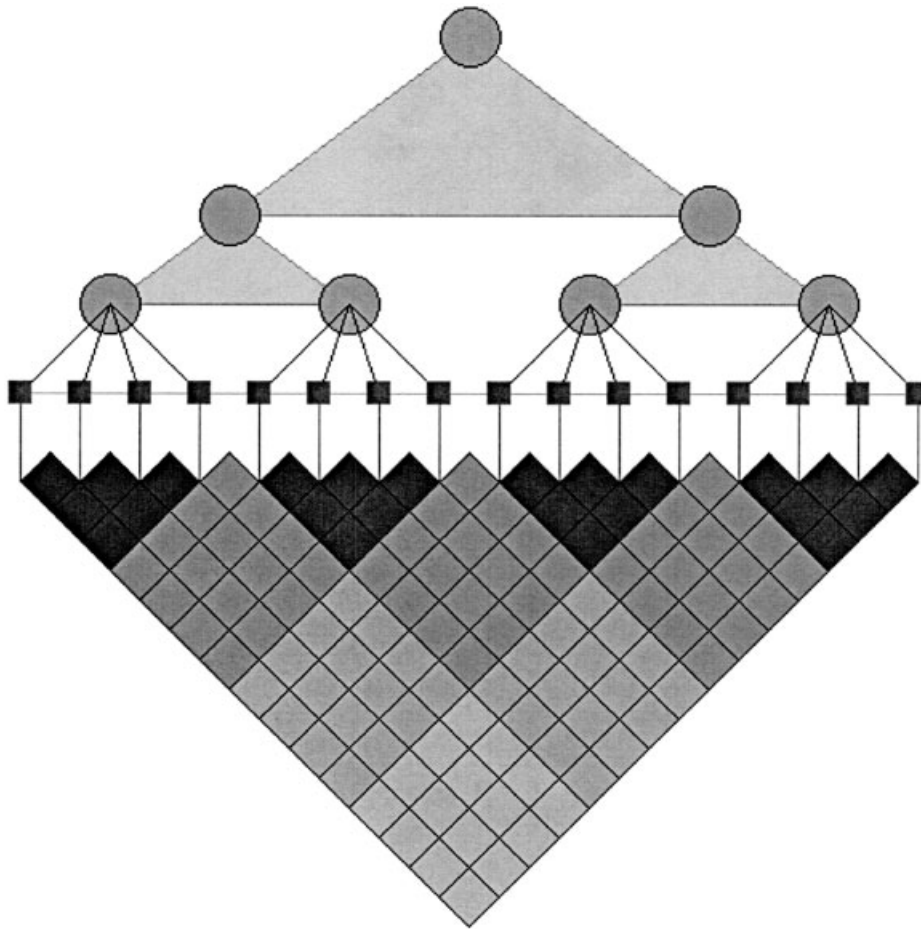


Fig. 1. Schematic diagram of partition-ligation (PL) strategy. Each of 16 squares represents one SNP. Entire haplotype is partitioned into a total of four atomistic units, each comprising four consecutive SNPs. A hierarchical ligation approach was taken to reconstruct entire haplotype. Diamond-shaped linkage disequilibrium (LD) plot is shown at bottom, and for illustrative purposes, each atomistic unit corresponds to a high-LD block (i.e., partition was located at recombination hotspots).

Niu et al. [2002] iterates between two steps until convergence: 1) $P(z_i = (\alpha, \beta) | \Theta, g_i) = \theta_\alpha \theta_\beta / \sum_{(a,b): a \oplus b = g_i} \theta_a \theta_b$, and 2) $P(\Theta | Z, G) = \text{Dir}[\lambda + N(Z)]$. The essence of prior annealing is to dwindle the pseudocounts λ to zero at the end of iteration linearly, to avoid the algorithm being trapped in a local maximum caused by high pseudocounts. The Gibbs sampling algorithm for haplotype reconstruction was performed within each atomistic unit, and at the global level, the Bayesian method employed by Niu et al. [2002] adopted a “block-by-block” strategy in updating long-range haplotypes. The resulting program, Haplotyper, was extensively evaluated using a variety of simulated and real datasets [Niu et al., 2002], and it was shown to be robust to the presence of missing data, to the violation of HWE or coalescent assumption, and is now widely applied for a variety of complex traits [Albagha et al., 2002;

Li et al., 2002; Ferrari et al., 2003, 2004; Kehoe et al., 2003, 2004; Beyzade et al., 2003; Spotila et al., 2003; White et al., 2003; Ertekin-Taner et al., 2004; Katzov et al., 2004; Eriksson et al., 2004; Choi et al., 2004; Mira et al., 2004; Sobacchi et al., 2004].

EM implementation. The computational intensity of haplotype estimation using the standard EM described above becomes unmanageable rather quickly with an increasing number of loci, and it can typically handle a maximum of ~ 20 loci [Qin et al., 2002]. Capitalizing on the idea of PL [Niu et al., 2002], a “PLEM” algorithm was developed that surpassed the limitations of the standard EM. Note that EM is essentially a deterministic procedure based on the maximum likelihood principle. In PLEM, one shall not keep only the MLE answers in resolving partial haplotypes, because the locally optimal solutions

do not necessarily give rise to the globally optimal answer. The way to deal with this problem is to keep a buffer of “runner-up” partial haplotypes to avoid tossing the optimal “partial answers” away prematurely. Simulation studies demonstrate that the larger the buffer size, the more accurate the haplotype estimates will be [Qin et al., 2002]. Besides haplotype frequency estimation and individual-based haplotype phasing, PLEM provides variance estimation for long-range haplotypes based on Fisher information matrix. Recently, the PLEM algorithm was employed in one of the programs developed for haplotype tagging [Zhang et al., 2004].

The advantage of the PL algorithm is that it can handle a large number of loci. The disadvantage is that partitioning outside the recombination hotspots may not give rise to the most optimal answer. However, the PL algorithm appeared to be relatively robust even when the partition was not located exactly at the cutting points. Another limitation of the PL implemented in Haplotyper and PLEM is that neither of the programs incorporated the coalescent model into their statistical framework. However, simulation results demonstrated that compared to EM, Haplotyper performed reasonably better, even when the genotype data were simulated based on coalescence [Niu et al., 2002].

SENSITIVITY ANALYSIS OF HAPLOTYPE PHASING ALGORITHMS

There is an ongoing debate in regard to whether the “best” haplotype phasing algorithm exists in population-based phasing. It has become clearer and clearer that different algorithms perform differently in different populations, with their own strengths and limitations. A “consensus vote” strategy based on multiple algorithms with different underlying statistical theories was suggested to increase the confidence of statistical inferred results. Here, through a simulation study, I would like to illustrate that the performances of various haplotype inference algorithms are sensitive to the haplotype diversity (or equivalently, haplotype variance) of the population under study. I would like to highlight that for populations with relatively low haplotype diversities, most algorithms give robust answers, although some still have slightly better performances; for populations with relatively high haplotype diver-

sities, performance differences across different algorithms become much more evident. In Figure 2, results from a simulation study were used to compare the performances of Clark’s algorithm, Haplotyper, and PLEM in terms of haplotype phasing accuracy at the population (haplotype frequency estimation) level among six different ethnic populations: Beni (Nigeria), Yoruba (Nigeria), Shona (Zimbabwe), African American, European American, and Asian, with different haplotype diversities of the 11-SNP G6PD gene haplotype (Fig. 2a and Sabeti et al. [2002]). As indicated above, PHASE requires at least 2 million MCMC iterations [Stephens and Donnelly, 2003] for phasing each of the $N = 500$ subjects for $M=100$ simulations for each of the six ethnic populations, and it would take nearly 2 months to finish those runs of PHASE alone to get all the results needed; therefore, PHASE was excluded for comparisons. At the individual level, the simulation results showed that Haplotyper and PLEM performed rather closely: PLEM made no single errors (i.e., 100%) for all but the Asian population data (99.94%), and Haplotyper made no single errors (i.e., 100%) for all simulated datasets for the Asian and European American population data, and its average accuracy for the other four ethnic populations was 96.2% (range, 94.2–97.8%). Clark’s algorithm performed best for the Asian population data (accuracy, 93.3%), but made many more individual phasing errors for the other five ethnic populations (accuracy rates were 45.5%, 44.5%, 55.2%, 46.4%, and 83.7% for Beni, Yoruba, Shona, African American, and European American populations, respectively), which demonstrated that for a population with limited haplotype diversity (the Asian population in this example), all three algorithms performed quite well (accuracy, >90%), and the algorithm with the best performance is Haplotyper. For populations with large haplotype diversities (Beni, Yoruba, Shona, and African American), Haplotyper and PLEM performed quite closely, whereas Clark’s algorithm had much higher error rates (the algorithm with the best performance is PLEM).

FAMILY-BASED HAPLOTYPE METHODS

A number of linkage analysis programs, including Genehunter [based on the Lander-Green

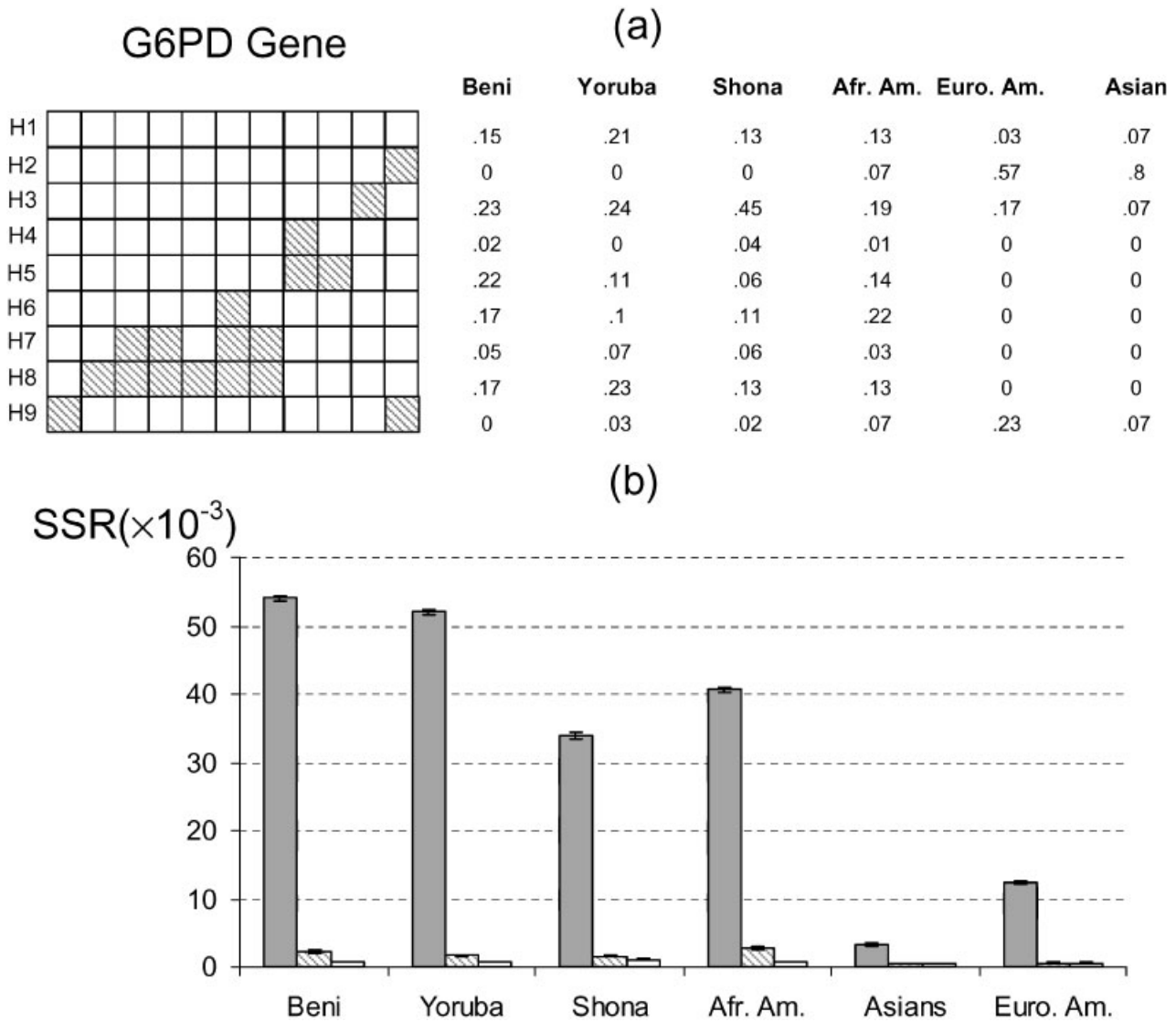


Fig. 2. Simulation study for comparing performances of Clark’s algorithm (shaded bars), Haplotyper (hatched bars), and PLEM (open bars) in terms of haplotype frequency estimation among 6 different ethnic populations: Beni (Nigeria), Yoruba (Nigeria), Shona (Zimbabwe), African American, European American, and Asian, with different haplotype diversities of G6PD gene [Sabeti et al., 2002]. a: Published haplotype frequency data of 11-SNP G6PD haplotype among 6 ethnic populations were used in simulation. Open and hatched squares represent major and minor alleles for each SNP, respectively. b: Y-axis denotes sum of squared error (SSR) defined as $\sum_{i=1}^z (\hat{\theta}_i - \theta_i^{truth})^2$ for $M=100$ simulations for each ethnic population for each algorithm, where $\hat{\theta}_i$ and θ_i^{truth} represent estimated and true [i.e., from Sabeti et al., 2002] haplotype frequency for each respective haplotype. In each simulation, genotype data for $N=500$ subjects were generated by random pairing of haplotypes according to their respective frequencies. Data are presented as mean \pm SE.

algorithm; reviewed in Nyholt, 2002], Simwalk2 [based on MCMC and the simulated annealing technique; Sobel and Lange, 1996], and Merlin [based on a sparse gene flow tree implementation of the Lander-Green algorithm; Abecasis et al., 2002], are capable of haplotype phase reconstruction using pedigree data. However, it should be noted that both the Lander-Green algorithm implemented in Genehunter and Merlin, and the

MCMC/simulated annealing algorithm implemented in Simwalk2, assume that the linked markers are in linkage equilibrium [Schaid et al., 2002; Becker and Knapp, 2003], which implies that all possible haplotype explanations are equally likely in case of phase ambiguity. For Genehunter, the inferred haplotypes may even depend on the order of alleles in the input file [Becker and Knapp, 2003]. Therefore, Genehunter, Merlin, and

Simwalk2 should be applied when there are few haplotype ambiguities (such as for highly polymorphic microsatellite markers) in extended large pedigrees, rather than for low-information-content SNP makers in simplex nuclear family settings [Becker and Knapp, 2003]. Even for pedigree data, individual haplotype reconstruction for multiple linked SNPs by computer programs may not be reliable if there are enough untyped persons in the pedigree, because the space of possible haplotype configurations tends to be too large for procedures such as MCMC to be finished within a reasonable time. Recently, the use of trios has become quite popular due to its relative efficiency and ease of sample collection. A couple of EM-based haplotype inference methods were developed without assuming the linkage equilibrium among the linked markers [Rohde and Fuerst, 2001; Becker and Knapp, 2002] for nuclear family data. Their performances were shown to be superior to that of the Lander-Green algorithm of Genehunter [Rohde and Fuerst, 2001; Becker and Knapp, 2002]. A rule-based “minimum-recombinant haplotyping” (MRH) algorithm, developed by Qian and Beckmann [2002], exhaustively searches all possible minimum-recombinant haplotype configurations in large pedigrees with many markers, and MRH allows missing genotype data to be imputed from identity-by-descent alleles. However, the utility of MRH for genotype data solely obtained from nuclear families remains unknown. It is noted that although trios are more convenient to sample than large, multigenerational pedigrees, for certain complex traits with late onset, the parental DNA samples are often unavailable for typing, and the collection of genotype data from trios can still be extremely difficult. In such scenarios, it is possible to sample additional unaffected siblings of probands. It is therefore critical to assess the contribution of genotype data from additional siblings to family-based haplotype reconstruction.

HANDLING GENOTYPE UNCERTAINTIES IN INFERRING HAPLOTYPES

Because the standard genotyping machines such as the TaqMan assay, the oligonucleotide ligation assay, and MassARRAY typically output a two-dimensional scatterplot, there are often intrinsic uncertainties in the genotype scoring process [Kang et al., 2004]. However, virtually all

genotyping machines directly output deterministic genotype scores. In case of genotype ambiguity, such deterministic genotype scores automatically generated by genotyping machines can introduce genotyping errors.

HANDLING GENOTYPING ERRORS IN DETERMINISTIC CALLS

To date, almost all haplotype inference algorithms assume that the inputs of the genotype data are without ambiguity. However, such an assumption is rarely tenable for real-world genetic markers (including SNPs, microsatellites, and variable numbers of tandem repeats) and, if violated, can severely bias haplotype frequency and reconstruction accuracies [Kirk and Cardon, 2002]. Furthermore, genotyping errors also have significant impact on the accuracy of estimations of genetic distances [Goldstein et al., 1997]. Even with family trios, error detection rates are usually low ($\sim 30\%$) if Mendelian consistency is used as the sole standard for checking errors [Gordon et al., 1999; Kirk and Cardon, 2002; Zou et al., 2003]. Therefore, it is critical to compare the performances of statistical algorithms in the presence of genotyping errors. A simulation study was performed to compare the sensitivities of three haplotype phasing algorithms, Haplotyper, PLEM, and Clark’s algorithms, with regard to sensitivities to genotyping errors. In total, five distinct haplotypes of the TNFRSF5 gene with frequencies $>4\%$ were used in the simulation study, based on Zhou et al. [2004]. In total, 500 subjects were generated in each of 100 simulation runs for each algorithm at genotype error rates of 1%, 2%, 5%, and 10% (denoted as models M1–M4, respectively), and haplotype phasing accuracies at both individual and population levels were compared. It can be seen that Haplotyper and PLEM performed essentially identically in terms of individual phasing accuracy (Fig. 3a), both of which were more robust to the presence of genotyping errors than Clark’s algorithm. In terms of population-based haplotype phasing accuracy, Haplotyper had the best performance among the three algorithms, and PLEM is a very close second (Fig. 3b).

HANDLING RAW GENOTYPE OUTPUTS

Most genotyping errors are caused by the ambiguities of raw genotyping outputs. Therefore, instead of using deterministic calls, an alternative

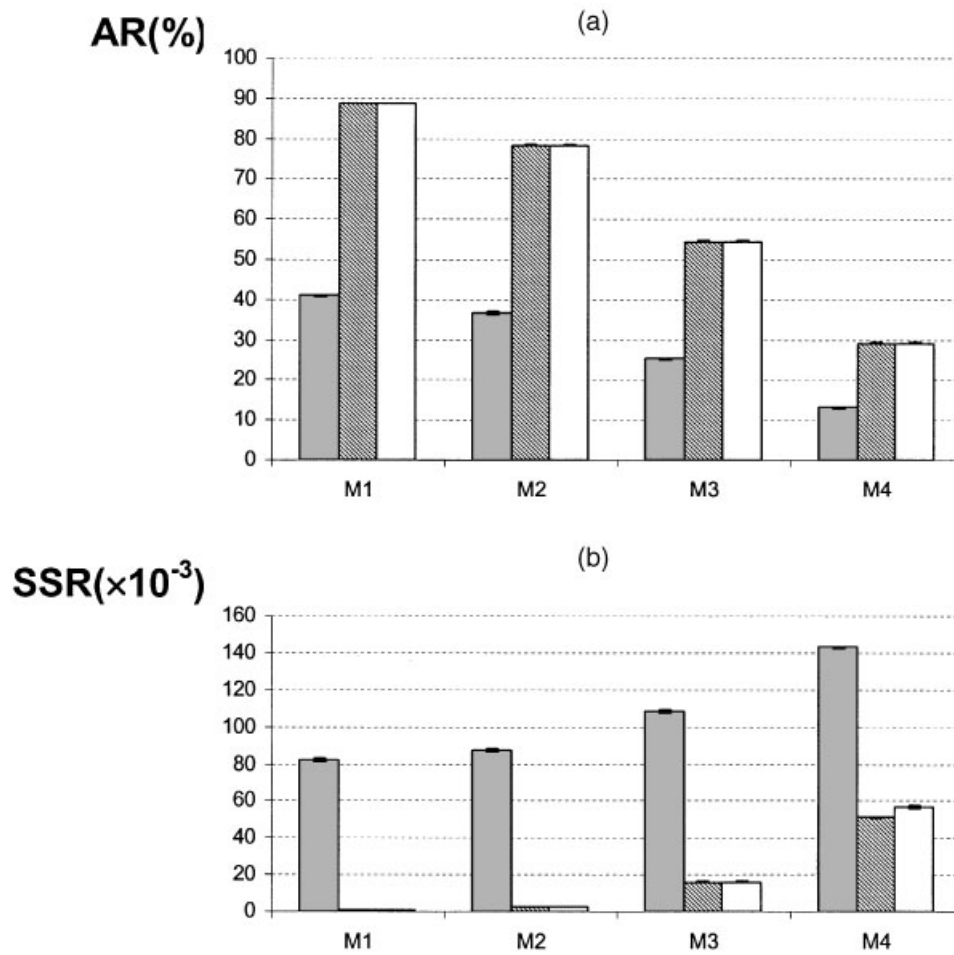


Fig. 3. Simulation study for comparing performances of Clark's algorithm (shaded bars), Haplotyper (hatched bars), and PLEM (open bars) with regard to sensitivities to genotyping errors. TNFRSF5 gene haplotype containing 6 SNPs (g.53031T>C, g.53489A>G, g.53824T>C, g.54068T>C, g.63977A>G, and g.64493T>C; from 5'→3') were used in simulation study, with respective frequencies of 0.417, 0.258, 0.202, 0.051, and 0.043 for 000000, 011111, 10110, 011110, and 000110 (0=major allele; 1=minor allele), according to Zhou et al. [2004]. In total, $N=500$ subjects were randomly generated in each of $M=100$ simulation runs for each algorithm with genotype error rates of 1%, 2%, 5%, and 10% (denoted as models M1–M4, respectively) and were then phased by each respective algorithm. a: Haplotype phasing accuracy comparison at individual level. Accuracy rate (AR)=number of correctly phased individuals/total number of individuals $\times 100\%$. b: Haplotype phasing accuracy comparison at population level, where SSR (defined in Fig. 2) was used as Y-axis. Data are presented as mean \pm SE.

way would be to use probabilistic genotype scores. Kang et al. [2004] introduced a Bayesian clustering algorithm based on a bivariate t -mixture model. The algorithm finds the center and spread of each genotype cluster using the parameter-expanded data augmentation scheme [Liu and Wu, 1999], and assigns each genotype for each SNP with a probabilistic score. Then, for each individual, the probabilistic scores for the multiple linked SNPs can be transformed into a genospectrum, as demonstrated in Figure 4. A novel EM algorithm, called "GS"-EM, is then used to incorporate genospectrum information in

the haplotype inference process: consider the genospectrum for individual i — $(G, \Pi)^i$, with $G^i = (G^{i,1}, G^{i,2}, \dots, G^{i,l_i})$ being the set of possible multi-SNP genotypes with respective probabilities $\Pi^i = (\pi^{i,1}, \pi^{i,2}, \dots, \pi^{i,l_i})$, where l_i denotes the number of possible distinct genotypes for individual i . The likelihood function of GS-EM is

$$L(\Theta) = P(G|\Theta)$$

$$= \prod_{i=1}^n p(g^i|\Theta) \cong \prod_{i=1}^n \sum_{j=1}^{l_i} \left(\pi^{ij} \sum_{(a,b):a \oplus b = g^{ij}} \theta_a \theta_b \right).$$

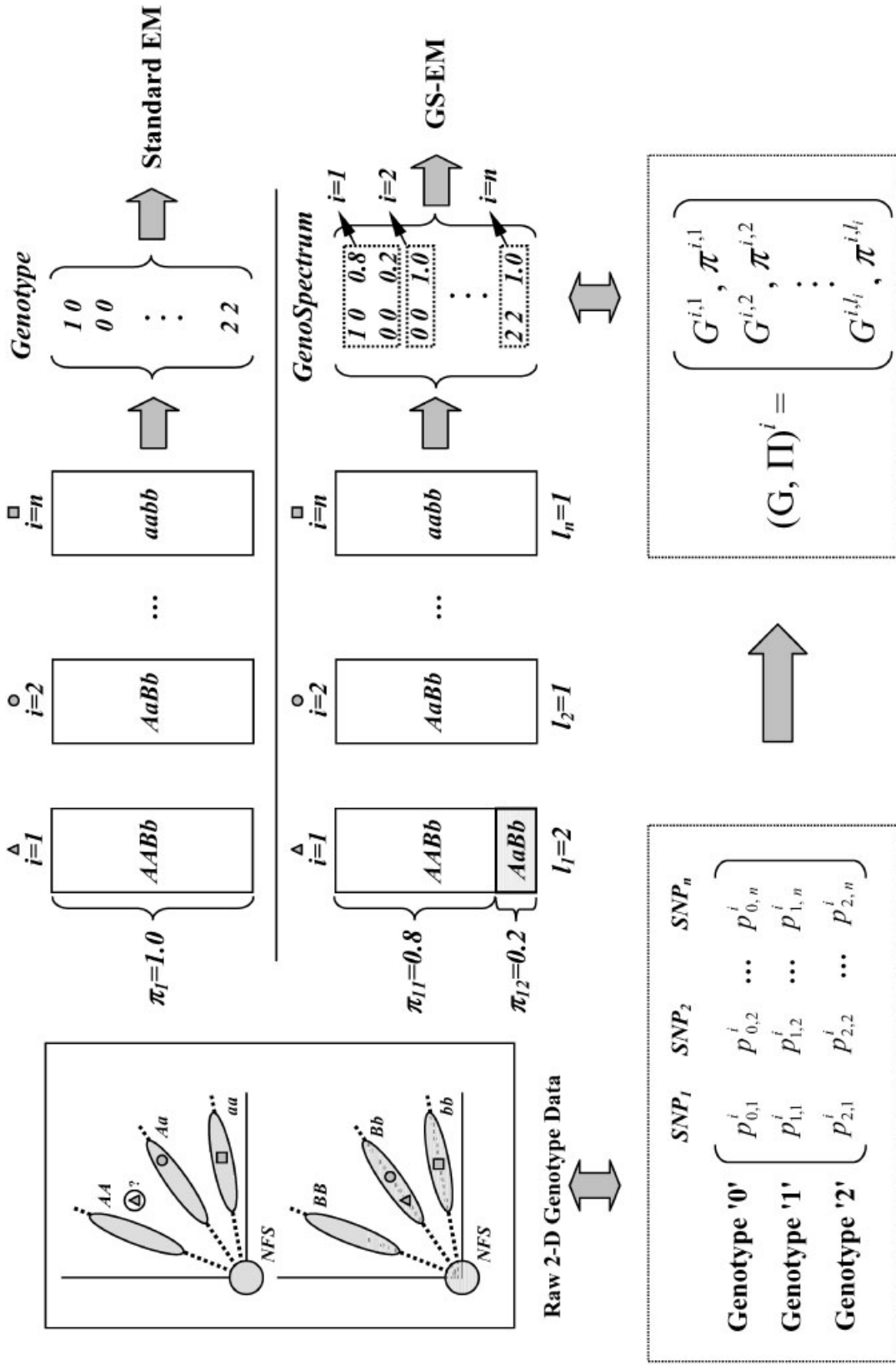


Fig. 4. Schematic depiction of genospectrum [Kang et al., 2004] for assigning probabilistic genotype scores. Haplotype of interest consists of $m=2$ SNPs, $snp1$, and $snp2$. Raw two-dimensional fluorescence intensity genotype data for these two SNPs are shown in upper left corner. AA , Aa , and aa denote homozygous major, heterozygous, and homozygous minor genotype clusters for $snp1$, and BB , Bb , and bb denote homozygous major, heterozygous, and homozygous minor genotype clusters for $snp2$, respectively. Solid triangles, circles, and squares represent datapoints for individuals 1, 2, and n , respectively. At top: Practice of conventional, deterministic genotype scoring, such that standard EM algorithm can be employed for haplotype inference. At bottom: Practice of probabilistic genotype scoring, giving rise to genospectrum (GS), such that GS-EM algorithm can be employed for haplotype inference.

The following EM iteration for Θ can then be obtained:

$$\begin{aligned}\theta_a^{(t+1)} &= \frac{E_{\Theta^{(t)}}(n_a | G, \Pi)}{2n} \\ &= \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^{l_i} \frac{\pi^{i,j} \theta_a^{(t)} \theta_{G^{i,j} \setminus a}^{(t)} \{1 + I(a = G^{i,j} \setminus a)\}}{\sum_{j=1}^{l_i} \pi^{i,j} \sum_{(u,v): u \oplus v = G^{i,j}} \theta_u^{(t)} \theta_v^{(t)}}\end{aligned}$$

where $G^{i,j} \setminus a$ denotes the complement haplotype such that $a \oplus [G^{i,j} \setminus a] = G^{i,j}$. Note that the standard EM algorithm is a special case of the GS-EM with $l_i = 1$ and $\pi^{i,j} = 1$. Because GS-EM uses more information from the raw genotype output than deterministic genotype scores, it is shown to make more accurate haplotype inferences, demonstrated through both simulated and empirical datasets [Kang et al., 2004]. It is notable that the strategy used by Kang et al. [2004] treated genotyping clustering and haplotyping phasing as two disjoint steps; however, these two steps can be modeled simultaneously, and haplotype information can actually provide crucial hidden LD information that can be used in reducing the chances of making errors in the genotype clustering step.

HAPLOTYPE INFERENCE METHODS FOR DNA POOLS

“DNA pooling” of individual samples was advocated to reduce the genotyping cost dramatically [Churchill et al., 1993; Amos et al., 2000; Daniels et al., 1998; Sham et al., 2002]. Inbar et al. [2002] developed an allele-specific, long-range PCR method for direct measurement of haplotype frequencies in DNA pools. However, this method is not feasible when neighboring SNPs have physical intervals > 20 kb. It should be noted that pooling complicates the configuration of haplotypes and therefore adds more ambiguities in estimating haplotype frequencies. Although creating two large DNA pools for cases and controls is very cost-efficient, such a procedure is extremely susceptible to errors not only in allele frequency estimation, but also in haplotype frequency estimation [Barratt et al., 2002; Yang et al., 2003]. As an alternative, the use of small DNA pools (e.g., a pool size of 2) [Hoh et al., 2003; Wang et al., 2003] was shown to be superior over the use of large pools. Consider pools of size K (i.e., comprising K individuals) each and a total of n such pools so that the total number of individuals is nK (It is noted that haplotype

inference without DNA pooling is just a special case when $K=1$). Then, in each pool, there are $2K$ haplotypes for K individuals. Two EM-based algorithms were developed to phase “pool genotype data.” Yang et al. [2003] developed an EM algorithm for obtaining MLEs and variance estimations of haplotype frequencies that can handle missing data in DNA pools. For m loci, there are 2^m possible distinct haplotypes, and we give each distinct haplotype a numerical label, such that all these 2^m possible haplotypes can be distinguished by their labels: $1, 2, 3, \dots, 2^m$. For the i th pool, let g_i denote the unphased “pool genotype,” which can be phased into α_i distinct haplotype configurations that are compatible with g_i . Note that each of the α_i distinct haplotype configurations can be represented by a $1 \times 2K$ vector, e.g., for the t th ($1 \leq t \leq \alpha_i$) distinct haplotype configuration of pool i , we can denote its configuration $L^{(i,t)}$ as $(l_1^{(i,t)}, l_2^{(i,t)}, \dots, l_{2K}^{(i,t)})$, such that $1 \leq l_1^{(i,t)} \leq l_2^{(i,t)} \leq \dots \leq l_{2K}^{(i,t)} \leq 2^m$, where $l_1^{(i,t)}, l_2^{(i,t)}, \dots, l_{2K}^{(i,t)}$ are simply the “label representation” of the $2K$ haplotypes for pool i in an ascending order of their labels. For example, for a haplotype with 3 SNPs ($m=3$), we label the 8 possible haplotype configurations (0 0 0), (0 0 1), (0 1 0), (0 1 1), (1 0 0), (1 0 1), (1 1 0), and (1 1 1) as 1, 2, ..., 8, respectively (0=major allele; 1=minor allele). For instance, if we form DNA pools of size 4 ($K=4$), then, and the 5th ($i=5$) pool with “pool genotype” $g_5=(1 0 0)$ (1=homozygous wild-type, 0=heterozygote, 2=homozygous mutant), a compatible haplotype configuration $L^{(i,t)}$ would be (1, 2, 2, 3, 3, 3, 3, 3). Suppose these $2K$ elements $l_1^{(i,t)}, l_2^{(i,t)}, \dots, l_{2K}^{(i,t)}$ of $L^{(i,t)}$ comprise $\beta_{(i,t)}$ distinct labels, $(dl_1^{(i,t)}, dl_2^{(i,t)}, \dots, dl_{\beta_{(i,t)}}^{(i,t)})$, such that $1 \leq dl_1^{(i,t)} < dl_2^{(i,t)} < \dots < dl_{\beta_{(i,t)}}^{(i,t)} \leq 2^m$ (in our example, $L^{(i,t)}$ is reduced to (1, 2, 3), such that $\beta_{(i,t)} = 3$), we then denote $c_1, c_2, \dots, c_{\beta_{(i,t)}}$ as the respective counts for labels $dl_1^{(i,t)}, dl_2^{(i,t)}, \dots, dl_{\beta_{(i,t)}}^{(i,t)}$ (in our example, c_1, c_2 , and c_3 would be 1, 2, and 5, respectively), and it is obvious that $\sum_{i=1}^{\beta_{(i,t)}} c_i = 2K$. Then, the likelihood function for the pooled genotype data is given by $L(\Theta) = P(G|\Theta) = \prod_{i=1}^n (\sum_{t=1}^{\alpha_i} w_t \prod_{k=1}^{2K} \theta_{tk})$, where $w_t = (2K)! / \prod_{j=1}^{\beta_{(i,t)}} (c_j)!$, and $\theta_{t1}, \theta_{t2}, \dots, \theta_{t2K}$ denote the haplotype frequencies for the respective haplotypes $l_1^{(i,t)}, l_2^{(i,t)}, \dots, l_{2K}^{(i,t)}$. According to Yang et al.

[2003], in the E-step, the probability of the j th haplotype configuration $L^{(i,j)}$ can be calculated as $p(L^{(i,j)}) = w_j \prod_{k=1}^{2K} \theta_{jk} / \sum_{t=1}^{z_i} w_t \prod_{k=1}^{2K} \theta_{tk}$. In the M-step, at the $(s + 1)$ th iteration, $\theta_h^{(s+1)} = \sum_{i=1}^n \sum_{t=1}^{z_i} c_{L^{(i,j)}}^h p(L^{(i,t)}) / 2nK$, where $c_{L^{(i,j)}}^h$ denotes the count of haplotype h in the j th haplotype configuration $L^{(i,j)}$ for pool i , and $p(L^{(i,t)})$ was estimated using $\Theta^{(s)}$. Yang et al. [2003] further showed that 1) $K=3$ or 4 appears to be optimal, and 2) $n > 30$ is needed for getting robust asymptotic variance estimates. Ito et al. [2003] also developed an EM algorithm called "LDpooled" that can phase pooled DNA samples. The disadvantage of "LDpooled" compared to Yang et al. [2003] is that "LDpooled" is incapable of handling missing data. It should be noted that pooling essentially ignored the issue of genotyping errors at the individual level, and the impact of individual genotyping error on the accuracy of haplotype frequency estimation based on DNA pools remains unexplored for the methods introduced by Yang et al. [2003] and Ito et al. [2003].

CONCLUSIONS

Haplotypes capture LD information in chromosomal regions descended from ancestral chromosomes. Such information is of considerable interest in population genetic and genetic epidemiological studies. Unveiling the human genome sequence is as revolutionary as Copernicus' revelation of an unfixed Earth, and has helped generate a genome-wide SNP map [Altshuler et al., 2000; Sachidanandam et al., 2001]. With widespread applications of new generations of genotyping techniques, especially high-density SNP arrays [Kennedy et al., 2003; Matsuzaki et al., 2004], the magical potential of the human genome will eventually be unlocked by linking haplotype information to biomarker and phenotypic data. It is foreseeable that the future development of statistical algorithms for addressing haplotype reconstruction will resort more and more to ideas based on combinatorial mathematics, graphical models, and machine learning. The performances of *in silico* methods utilizing these ideas will continue to improve in practical scenarios. Applications of various haplotype inference algorithms have already facilitated greatly the derivation of haplotype phase information from observed genotype data, which is an essential process for subsequent haplotype tagging SNP (htSNP) selection and htSNP-based association analyses. The

ultimate delineation of the human haplotype map using population- or family-based haplotype inference algorithms will become one of the stepping stones toward a complete understanding of genetic causes of complex diseases.

ACKNOWLEDGMENTS

I thank Prof. Jun S. Liu, Dr. Xin Lu, and Hosung Kang for insightful discussions.

REFERENCES

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30:97–101.
- Albagha OM, Tasker PN, McGuigan FE, Reid DM, Ralston SH. 2002. Linkage disequilibrium between polymorphisms in the human TNFRSF1B gene and their association with bone mass in perimenopausal women. *Hum Mol Genet* 11:2289–2295.
- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513–516.
- Amos CI, Frazier ML, Wang W. 2000. DNA pooling in mutation detection with reference to sequence analysis. *Am J Hum Genet* 66:1689–1692.
- Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. 2002. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 66:393–405.
- Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. *Nat Rev Genet* 5:251–261.
- Becker T, Knapp M. 2003. Comment on "The impact of genotyping error on haplotype reconstruction and frequency estimation." *Eur J Hum Genet* 11:637.
- Beyzade S, Zhang S, Wong YK, Day IN, Eriksson P, Ye S. 2003. Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction. *J Am Coll Cardiol* 41:2130–2137.
- Bixby RE, Wagner DK. 1988. An almost linear-time algorithm for graph realization. *Math Operations Res* 13:99–123.
- Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S. 2003. Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res* 13:2717–2724.
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74:106–120.
- Choi JH, Park HS, Oh HB, Lee JH, Suh YJ, Park CS, Shin HD. 2004. Leukotriene-related gene polymorphisms in ASA-intolerant asthma: an association with a haplotype of 5-lipoxygenase. *Hum Genet* 114:337–344.
- Chung RH, Gusfield D. 2003. Perfect phylogeny haplotyper: haplotype inferral using a tree model. *Bioinformatics* 19:780–781.
- Churchill GA, Giovannoni JJ, Tanksley SD. 1993. Pooled-sampling makes high-resolution mapping practical with DNA markers. *Proc Natl Acad Sci USA* 90:16–20.

- Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111–122.
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 63:595–612.
- Crandall KA, Templeton AR. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* 134:959–969.
- Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen MJ. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am J Hum Genet* 62:1189–1197.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc (Ser B)* 39:1–38.
- Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. 2001. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet* 28:361–364.
- Eriksson P, Deguchi H, Samnegard A, Lundman P, Boquist S, Tornvall P, Ericsson CG, Bergstrand L, Hansson LO, Ye S, Hamsten A. 2004. Human evidence that the cystatin C gene is implicated in focal progression of coronary artery disease. *Arterioscler Thromb Vasc Biol* 24:551–557.
- Ertekin-Taner N, Allen M, Fadale D, Scanlin L, Younkin L, Petersen RC, Graff-Radford N, Younkin SG. 2004. Genetic variants in a haplotype block spanning IDE are significantly associated with plasma Aβ₄₂ levels and risk for Alzheimer disease. *Hum Mutat* 23:334–342.
- Excoffier L, Slatkin M. 1995. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921–927.
- Ferrari SL, Ahn-Luong L, Garnero P, Humphries SE, Greenspan SL. 2003. Two promoter polymorphisms regulating interleukin-6 gene expression are associated with circulating levels of C-reactive protein and markers of bone resorption in postmenopausal women. *J Clin Endocrinol Metab* 88:255–259.
- Ferrari SL, Deutsch S, Choudhury U, Chevalley T, Bonjour JP, Dermizakis ET, Rizzoli R, Antonarakis SE. 2004. Polymorphisms in the low-density lipoprotein receptor-related protein 5 (LRP5) gene are associated with variation in vertebral bone mass, vertebral bone size, and stature in whites. *Am J Hum Genet* 74:866–875.
- Goldstein DR, Zhao H, Speed TP. 1997. The effects of genotyping errors and interference on estimation of genetic distance. *Hum Hered* 47:86–100.
- Gordon D, Heath SC, Ott J. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered* 49:65–70.
- Gusfield D. 2001. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J Comput Biol* 8: 305–323.
- Gusfield D. 2002. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In: *Proceedings of the 6th International Conference on Computational Biology, RECOMB'02*. p 166–175.
- Halperin E, Eskin E. 2004. Haplotype reconstruction from genotype data using Imperfect Phylogeny. *Bioinformatics* 20:1842–1849.
- Hoh J, Matsuda F, Peng X, Markovic D, Lathrop MG, Ott J. 2003. SNP haplotype tagging from DNA pools of two individuals. *BMC Bioinformatics* 4:14.
- Inbar E, Yakir B, Darvasi A. 2002. An efficient haplotyping method with DNA pools. *Nucleic Acids Res* 30:76.
- International HapMap Consortium. 2003. The International HapMap project. *Nature* 426:789–796.
- Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, Kamatani N. 2003. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am J Hum Genet* 72:384–398.
- Kang H, Qin ZS, Niu T, Liu JS. 2004. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *Am J Hum Genet* 74:495–510.
- Katzov H, Chalmers K, Palmgren J, Andreassen N, Johansson B, Cairns NJ, Gatz M, Wilcock GK, Love S, Pedersen NL, Brookes AJ, Blennow K, Kehoe PG, Prince JA. 2004. Genetic variants of ABCA1 modify Alzheimer disease risk and quantitative traits related to beta-amyloid metabolism. *Hum Mutat* 23:358–367.
- Kehoe PG, Katzov H, Feuk L, Bennet AM, Johansson B, Wiman B, de Faire U, Cairns NJ, Wilcock GK, Brookes AJ, Blennow K, Prince JA. 2003. Haplotypes extending across ACE are associated with Alzheimer's disease. *Hum Mol Genet* 12: 859–867.
- Kehoe PG, Katzov H, Andreassen N, Gatz M, Wilcock GK, Cairns NJ, Palmgren J, de Faire U, Brookes AJ, Pedersen NL, Blennow K, Prince JA. 2004. Common variants of ACE contribute to variable age-at-onset of Alzheimer's disease. *Hum Genet* 114:478–483.
- Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SP, Jones KW. 2003. Large-scale genotyping of complex DNA. *Nat Biotechnol* 21:1233–1237.
- Kirk KM, Cardon LR. 2002. The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet* 10:616–622.
- Li P, Wood T, Thompson JN. 2002. Diversity of mutations and distribution of single nucleotide polymorphic alleles in the human alpha-L-iduronidase (IDUA) gene. *Genet Med* 4: 420–426.
- Liu JS, Wu YN. 1999. Parameter expansion for data augmentation. *J Am Stat Assoc* 94:1264–1274.
- Liu JS, Sabatti C, Teng J, Keats BJ, Risch N. 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res* 11:1716–1724.
- Lu X, Niu T, Liu JS. 2003. Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. *Genome Res* 13:2112–2117.
- Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R. 2004. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 14:414–425.
- McDonald OG, Krynetski EY, Evans WE. 2002. Molecular haplotyping of genomic DNA for multiple single-nucleotide polymorphisms located kilobases apart using long-range polymerase chain reaction and intramolecular ligation. *Pharmacogenetics* 12:93–99.
- Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G. 1996. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res* 24:4841–4843.
- Mira MT, Alcais A, Nguyen VT, Moraes MO, Di Flumeri C, Vu HT, Mai CP, Nguyen TH, Nguyen NB, Pham XK, Sarno EN, Alter A, Montpetit A, Moraes ME, Moraes JR, Dore C, Gallant CJ, Lepage P, Verner A, Van De Vosse E, Hudson TJ, Abel L, Schurr

- E. 2004. Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature* 427:636–640.
- Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70:157–169.
- Nyholt DR. 2002. Genehunter: your “one-stop shop” for statistical genetic analysis? *Hum Hered* 53:2–7.
- Odeberg J, Holmberg K, Eriksson P, Uhlen M. 2002. Molecular haplotyping by pyrosequencing. *Biotechniques* 33:1104–1108.
- Qian D, Beckmann L. 2002. Minimum-recombinant haplotyping in pedigrees. *Am J Hum Genet* 70:1434–1445.
- Qin ZS, Niu T, Liu JS. 2002. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 71:1242–1247.
- Rieder MJ, Taylor SL, Clark AG, Nickerson DA. 1999. Sequence variation in the human angiotensin converting enzyme. *Nat Genet* 22:59–62.
- Rohde K, Fuerst R. 2001. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum Mutat* 17:289–295.
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
- Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. 2002. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet* 71:992–995.
- Sham P, Bader JS, Craig I, O’Donovan M, Owen M. 2002. DNA pooling: a tool for large-scale association studies. *Nat Rev Genet* 3:862–871.
- Sobacchi C, Vezzoni P, Reid DM, McGuigan FE, Frattini A, Mirolo M, Albhaga OM, Musio A, Villa A, Ralston SH. 2004. Association between a polymorphism affecting an AP1 binding site in the promoter of the TCIRG1 gene and bone mass in women. *Calcif Tissue Int* 74:35–41.
- 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.
- Spotila LD, Rodriguez H, Koch M, Tenenhouse HS, Tenenhouse A, Li H, Devoto M. 2003. Association analysis of bone mineral density and single nucleotide polymorphisms in two candidate genes on chromosome 1p36. *Calcif Tissue Int* 73:140–146.
- Steel MA. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *J Classif* 9:91–116.
- Stephens JC, Rogers J, Ruano G. 1990. Theoretical underpinning of the single-molecule-dilution (SMD) method of direct haplotype resolution. *Am J Hum Genet* 46:1149–1155.
- Stephens M, Donnelly P. 2000. Inference in molecular population genetics. *J R Stat Soc (Ser B)* 62:605–655.
- Stephens M, 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 NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989.
- Tregouet DA, Escolano S, Tired L, Mallet A, Golmard JL. 2004. A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. *Ann Hum Genet* 68:165–177.
- Tutte WT. 1960. An algorithm for determining whether a given binary matroid is graphic. *Proc Am Math Soc* 11:905–917.
- Wang S, Kidd KK, Zhao H. 2003. On the use of DNA pooling to estimate haplotype frequencies. *Genet Epidemiol* 24:74–82.
- White SN, Taylor KH, Abbey CA, Gill CA, Womack JE. 2003. Haplotype variation in bovine Toll-like receptor 4 and computational prediction of a positively selected ligand-binding domain. *Proc Natl Acad Sci USA* 100:10364–10369.
- Woolley AT, Guillemette C, Li Cheung C, Housman DE, Lieber CM. 2000. Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat Biotechnol* 18:760–763.
- Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N, Crocker N, Boyd PR, Zaykin DV, Purvis JJ. 2002. Effectiveness of computational methods in haplotype prediction. *Hum Genet* 110:148–156.
- Yang Y, Zhang J, Hoh J, Matsuda F, Xu P, Lathrop M, Ott J. 2003. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc Natl Acad Sci USA* 100:7225–7230.
- Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F. 2004. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res* 14:908–916.
- Zhang S, Pakstis AJ, Kidd KK, Zhao H. 2001. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet* 69:906–914.
- Zhong XB, Lizardi PM, Huang XH, Bray-Ward PL, Ward DC. 2001. Visualization of oligonucleotide probes and point mutations in interphase nuclei and DNA fibers using rolling circle DNA amplification. *Proc Natl Acad Sci USA* 98:3940–3945.
- Zhou G, Zhai Y, Dong X, Li Y, Zhang X, Zhang R, Li S, Li X, He F, Wei H, Chen X, Yao Z, Shen Y, Qiang B, He F. 2004. Variants in TNFRSF5 locus and association analysis with hepatitis B virus (HBV) infection. *Hum Mutat* 23:99–100.
- Zou G, Pan D, Zhao H. 2003. Genotyping error detection through tightly linked markers. *Genetics* 164:1161–1173.