A Comparison of Phasing Algorithms for Trios and Unrelated Individuals

Jonathan Marchini,[1] David Cutler,[2] Nick Patterson,[3] Matthew Stephens,[4] Eleazar Eskin,[5] Eran Halperin,[6] Shin Lin,[2] Zhaohui S. Qin,[7] Heather M. Munro,[7] Gonçalo R. Abecasis,[7] and Peter Donnelly,[1] for the International HapMap Consortium

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom; [2]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore; [3]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA; [4]Department of Statistics, University of Washington, Seattle; [5]Computer Science Department, Hebrew University, Jerusalem; [6]The International Computer Science Institute, Berkeley; and [7]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor

Knowledge of haplotype phase is valuable for many analysis methods in the study of disease, population, and evolutionary genetics. Considerable research effort has been devoted to the development of statistical and computational methods that infer haplotype phase from genotype data. Although a substantial number of such methods have been developed, they have focused principally on inference from unrelated individuals, and comparisons between methods have been rather limited. Here, we describe the extension of five leading algorithms for phase inference for handling father-mother-child trios. We performed a comprehensive assessment of the methods applied to both trios and to unrelated individuals, with a focus on genomic-scale problems, using both simulated data and data from the HapMap project. The most accurate algorithm was PHASE (v2.1). For this method, the percentages of genotypes whose phase was incorrectly inferred were 0.12%, 0.05%, and 0.16% for trios from simulated data, HapMap Centre d'Etude du Polymorphisme Humain (CEPH) trios, and HapMap Yoruban trios, respectively, and 5.2% and 5.9% for unrelated individuals in simulated data and the HapMap CEPH data, respectively. The other methods considered in this work had comparable but slightly worse error rates. The error rates for trios are similar to the levels of genotyping error and missing data expected. We thus conclude that all the methods considered will provide highly accurate estimates of haplotypes when applied to trio data sets. Running times differ substantially between methods. Although it is one of the slowest methods, PHASE (v2.1) was used to infer haplotypes for the 1 million–SNP HapMap data set. Finally, we evaluated methods of estimating the value of r^2 between a pair of SNPs and concluded that all methods estimated r^2 well when the estimated value was ≥ 0.8.

The size and scale of genetic-variation data sets for both disease and population studies have increased enormously. A large number of SNPs have been identified (current databases show 9 million of the posited 10–13 million common SNPs in the human genome [International HapMap Consortium 2005]); genotyping technology has advanced at a dramatic pace, so that 500,000 SNP assays can be undertaken in a single experiment; and patterns of correlations among SNPs (linkage disequilibrium [LD]) have been catalogued in multiple populations, yielding efficient marker panels for genomewide investigations (see the International HapMap Project Web site). These genetic advances coincide with recognition of the need for large case-control samples to robustly identify genetic variants for complex traits. As a result, genomewide association studies are now being undertaken, and much effort is being made to develop efficient statistical techniques for analyzing the resulting data, to uncover the location of disease genes. In addition, the advances allow much more detailed analysis of candidate genes identified by more traditional linkage-analysis methods.

Many methods of mapping disease genes assume that haplotypes from case and control individuals are available in the region of interest. Such approaches have been successful in localizing many monogenic disorders (Lazzeroni 2001), and there is increasing evidence, of both a practical and theoretical nature, that the use of haplotypes can be more powerful than individual markers in the search for more-complex traits (Puffenberger et al. 1994; Akey et al. 2001; Hugot et al. 2001; Rioux et al. 2001). Similarly, haplotypes are required for many population-genetics analyses, including some methods for inferring selection (Sabeti et al. 2002), and for studying recombination (Fearnhead and Donnelly 2001; Myers and Griffiths 2003) and historical migration (Beerli and Felsenstein 2001; De Iorio and Griffiths 2004).

It is possible to determine haplotypes by use of experimental techniques, but such approaches are considerably more expensive and time-consuming than modern high-throughput genotyping. The statistical determination of haplotype phase from genotype data is thus potentially very valuable if the estimation can be done accurately. This problem has received an increasing amount of attention over recent years, and several computational and

Received September 2, 2005; accepted for publication December 29, 2005; electronically published January 26, 2006.

Address for correspondence and reprints: Dr. Jonathan Marchini, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom. E-mail: marchini@stats.ox.ac.uk

Am. J. Hum. Genet. 2006;78:437–450. 0002-9297/2006/7803-0012$15.00

statistical approaches have been developed in the literature (see Salem et al. [2005] for a recent literature review). Existing methods include parsimony approaches (Clark 1990; Gusfield 2000, 2001), maximum-likelihood methods (Excoffier and Slakin 1995; Hawley and Kidd 1995; Long et al. 1995; Fallin and Schork 2000; Qin et al. 2002), Bayesian approaches based on conjugate priors (Lin et al. 2002, 2004*b*; Niu et al. 2002) and on priors from population genetics (Stephens et al. 2001; Stephens and Donnelly 2003; Stephens and Scheet 2005), and (im)perfect phylogeny approaches (Eskin et al. 2003; Gusfield 2003). Up to now, no comprehensive comparison of many of these approaches has been conducted.

The forthcoming era of genomewide studies presents two new challenges to the endeavor of haplotype-phase inference. First, the size of data sets that experimenters will want to phase is about to increase dramatically, in terms of both numbers of loci and numbers of individuals. For example, we might expect data sets consisting of 500,000 SNPs genotyped in 2,000 individuals in some genomewide studies. Second, to date, most approaches have focused on inferring haplotypes from samples of unrelated individuals, but estimation of haplotypes from samples of related individuals is likely to become important. When inferring haplotypes within families, substantially more information is available than for samples of unrelated individuals. For example, consider the situation in which a father-mother-child trio has been genotyped at a given SNP locus. With no missing data, phase can be determined precisely, unless all three individuals are heterozygous at the locus in question. Of loci with a minor-allele frequency of 20%, for example, just 5.1% will be phase unknown in trios, but this rises to 32% in unrelated individuals. With missing data, other combinations of genotypes can also fail to uniquely determine phase.

In this study, we describe the extension of several existing algorithms for dealing with trio data. We then describe a comprehensive evaluation of the performance of these algorithms for both trios and unrelated individuals. The evaluation uses both simulated and real data sets of a larger size (in terms of numbers of SNPs) than has been previously been considered. We draw the encouraging conclusion that all methods provide a very good level of accuracy on trio data sets. Overall, the PHASE (v2.1) algorithm provided the most accurate estimation on all the data sets considered. For this method, the percentages of genotypes whose phase was incorrectly inferred were 0.12%, 0.05%, and 0.16% for trios from simulated data, HapMap CEPH trios, and HapMap Yoruban trios, respectively, and 5.2% and 5.9% for unrelated individuals in simulated data and the HapMap CEPH data, respectively. The other methods considered in this study had comparable but slightly worse error rates. The error rates for trios are comparable to expected levels of genotyping error and missing data and highlight the level of accuracy that the best phasing algorithms can provide on a useful scale. We also observed substantial variation in the speed of the algorithms we considered. Although it is one of the slowest methods, PHASE (v2.1) was used to infer haplotypes for the 1 million–SNP HapMap data set (International HapMap Consortium 2005). In addition, the data sets used in this comparison will be made available, to form a benchmark set to aid the future development and assessment of phasing algorithms. Finally, we evaluated methods of estimating the value of r^2 between a pair of SNPs. The most accurate method for estimating r^2 was to first use PHASE to infer the haplotypes across the region and then to estimate r^2 between the pair of SNPs as if the haplotypes were known. All methods estimated r^2 well when the estimated value was $\geqslant 0.8$.

Material and Methods

In this section, we describe the algorithms implemented in this study. Since most of these algorithms have been described elsewhere, we give only a brief overview of each method, together with some details concerning how each method was extended to cope with father-mother-child trios. Following a description of our notation and the assumptions made by each method, there is one subsection for each new method. Individuals who contributed to the development of the trio version of each method are shown in parentheses as part of the subsection title. In each subsection, expressed opinions are those of the contributing authors of that subsection and not of the combined set of authors as a group. We conclude with a concise overview that relates the different methods according to the assumptions they make about the most-plausible haplotype reconstructions.

Notation and Assumptions

We consider m linked SNPs on a chromosomal region of n trio families, where each trio consists of a mother, a father, and one offspring. We use the following notation throughout. Let $G = (G_1, \ldots, G_n)$ denote all the observed genotypes, in which $G_i = (GM_i, GF_i, GC_i)$ denotes the ith trio. GF_i, GM_i, and GC_i denote the observed genotype data for the father, mother, and child, respectively, and each are vectors of length m—that is, $GF_i = (GF_{i1}, \ldots, GF_{im})$, with $GF_{ik} = 0$, 1, or 2 representing homozygous wild-type, heterozygous, or homozygous mutant genotypes, respectively, at SNP marker k. Similarly, let $H = (H_1, H_2, \ldots, H_n)$ denote the unobserved haplotype configurations compatible with G, in which $H_i = (HM_i, HF_i)$, where $HM_i = (HM_{i1}, HM_{i2})$ and $HF_i = (HF_{i1}, HF_{i2})$ denote the haplotype pairs of the mother and father, respectively. We use the notation $HF_{i1} \oplus HF_{i2} = GF_i$ to indicate that the two haplotypes are compatible with the genotype GF_i. Also, we let $\Theta = (\theta_1, \ldots, \theta_s)$ be a vector of unknown population haplotype frequencies of the s possible haplotypes that are consistent with the sample.

All of the following algorithms make the assumption that

all the parents are sampled independently from the population and that no recombination occurs in the transmission of haplotypes from the parents to children.

PHASE (M.S. and J.M.)

The PHASE algorithm (Stephens et al. 2001; Stephens and Donnelly 2003; Stephens and Scheet 2005) is a Bayesian approach to haplotype inference that uses ideas from population genetics—in particular, coalescent-based models—to improve accuracy of haplotype estimates for unrelated individuals sampled from a population. The algorithm attempts to capture the fact that, over short genomic regions, sampled chromosomes tend to cluster together into groups of similar haplotypes. With the explicit incorporation of recombination in the most recent version of the algorithm (Stephens and Scheet 2005), this clustering of haplotypes may change as one moves along a chromosome. The method uses a flexible model for the decay of LD with distance that can handle both "blocklike" and "nonblocklike" patterns of LD.

We extended the algorithm described by Stephens and Scheet (2005) to allow for data from trios (two parents and one offspring). We treat the parents as a random sample from the population and aim to estimate their haplotypes, taking into account both the genotypes of the parents and the genotype of the child. More specifically, we aim to sample from the distribution $\Pr(HF,HM|GF,GM,GC)$ (compared with sampling from $\Pr(HF,HM|GF,GM)$, as shown in the work by Stephens and Scheet [2005]). To do this, we use a Markov chain–Monte Carlo (MCMC) algorithm very similar to that of Stephens and Scheet (2005), but, instead of updating one individual at a time, we update pairs of parents simultaneously. Note that the observed genotypes may include missing data at some loci, in which case the inferred haplotype pairs will include estimates of the unobserved alleles. When updating the parents in trio i, this involves computing, for each possible pair of haplotype combinations ($HF_i = \{hf,hf'\}; HM_i = \{hm,hm'\}$) in the two parents, the probability

$$\Pr\left(HF_i = \{hf,hf'\},HM_i = \{hm,hm'\}|GF_i,GM_i,GC_i,HF_{-i},HM_{-i},\rho\right) \propto \alpha_i\beta_i\gamma_i ,$$

where

$$\alpha_i = (2 - \delta_{hfhf'})\pi(hf|HF_{-i},HM_{-i},\rho,\mu)\pi(hf'|HF_{-i},HM_{-i},\rho,\mu) ,$$

$$\beta_i = (2 - \delta_{hmhm'})\pi(hm|HF_{-i},HM_{-i},\rho,\mu)\pi(hm'|HF_{-i},HM_{-i},\rho,\mu) ,$$

and

$$\gamma_i = \Pr[GC_i|HF_i = (hf,hf'),HM_i = (hm,hm')] ,$$

and where $\delta_{hh'}$ is 1 if $h = h'$ and is 0 otherwise; HF_{-i} and HM_{-i} are the sets HF and HM with HF_i and HM_i removed, respectively; π is a modification of the conditional distribution of Fearnhead and Donnelly (2001); ρ is an estimate of the population-scaled recombination rate, which is allowed to vary along the region being considered; and μ is a parameter that controls the mutation rate (see Stephens and Scheet [2005] for more details). The probability $\Pr[GC_i|HF_i = (hf,hf'),HM_i = (hm,hm')]$ is calculated assuming no recombination from parents to offspring and is therefore trivial to compute. We also assume no genotyping error. As a result, this probability is typically equal to 0 for a large number of parental diplotype configurations consistent with the parental genotypes, so the children's genotype data substantially reduces the number of diplotype configurations that must be considered. As in the work of Stephens and Scheet (2005), we use Partition Ligation (Niu et al. 2002) to further reduce the number of diplotype configurations considered when estimating haplotypes over many markers. This approach is not the most efficient, but it involved few changes to the existing algorithm.

wphase *(N.P.)*

The model underlying *wphase* was developed on the basis of ideas proposed by Fearnhead and Donnelly (2001) that introduced a simple approximate model for haplotypes sampled from a population. The algorithm differs from the PHASE algorithm above in three ways:

1. PHASE uses MCMC to sample configurations, whereas *wphase* performs a discrete hill climb. *wphase* computes a pseudolikelihood function or score for a putative haplotype reconstruction, H, of the form

$$S(H) = \prod_{i=1}^{n} \alpha_i\beta_i\gamma_i ,$$

   where α_i, β_i, and γ_i are defined as in the description of PHASE above. The method attempts to maximize the score by iteratively applying a set of "moves" that make small changes to the reconstruction.
2. PHASE and *wphase* differ in the precise form of the conditional distributions, π, used to calculate the factors α_i and β_i. As explained above, PHASE uses a modification of the conditional distribution of Fearnhead and Donnelly (2001), whereas *wphase* uses the conditional distributions introduced by Li and Stephens (2003).
3. PHASE internally re-estimates a *variable* recombination rate across the region, whereas *wphase* uses an externally input *constant* recombination rate across the region. Specifically, *wphase* uses $\rho = 0.05$ and $\theta = 0.02$.

In our opinion, the second and third differences are more important than the first. Although use of an MCMC offers some theoretical advantages, particularly the possibility of inference with use of multiple imputation of haplotypes, this is rarely used in practice (see David Clayton's SNPHAP algorithm for a notable exception [Clayton Web site]). If only one haplotype reconstruction is to be used (e.g., in HapMap), then maximizing a pseudolikelihood function is likely to produce a good solution. Testing in simulation has shown that *wphase* nearly always returns a score that is as good as or better than the value of the true haplotypes. This suggests that the quality of the reconstruction can be improved only by refining the score, not by altering the details of the hill climb. The difference in the form of the conditional distributions described above may lead to improved reconstructions (Stephens and Scheet 2005).

In the special case of the resolution of singleton SNPs that occur in the same individual, the conditional distributions used with PHASE will result in a more plausible solution than those used with *wphase*. The effect this difference has for nonsingleton SNPs remains unclear.

In addition, internally estimating a variable recombination rate is important, and its absence is a major weakness of the current version of *wphase*. True recombination rates vary greatly across the genome (McVean et al. 2004; Myers et al. 2005) and between various simulated regions in our test set.

Initial comparisons with PHASE version 1 (Stephens et al. 2001) at the time of development showed *wphase* to have very similar performance but not enough improvement to make it important to publish quickly. Since then, *wphase* has hardly improved, the main change being support for trio data, but PHASE underwent a major revision, with significant performance enhancements (Stephens and Donnelly 2003; Stephens and Scheet 2005).

HAP2 (S.L., A.C., and D.C.)

Haplotype and missing data inference was performed with HAP2, the details of which have been published elsewhere (Lin et al. 2004*b*). In short, HAP2 takes a Bayesian approach to haplotype reconstruction, set forth by Stephens et al. (2001), of dynamically updating an individual's haplotypes to resemble other haplotypes in the sample at each iteration in an MCMC scheme. The differences between this algorithm and the PHASE algorithm described above are as follows.

1. The conditional distributions, π, used at each iteration to sample the reconstruction of each individual are a special case of those used in PHASE, in which recombination is not explicitly modeled and a parent-independent mutation model is assumed. Specifically, the probability of observing a new haplotype is given by a Hoppe urn model (Hoppe 1987) or, equivalently, a Dirichlet, rather than coalescent-based, prior distribution for the haplotypes. Stephens et al. (2001) point out that the mode of the posterior distribution of this model will be close to the maximum-likelihood estimate sought by the expectation-maximization (EM) algorithm.
2. Whole haplotypes are not taken into account during the calculation of the conditional distributions. In reconstruction of an individual's haplotypes only, data at sites that are ambiguous for that individual are used. This difference results in a large increase in the speed of the algorithm.
3. A variant-partition ligation method (Niu et al. 2002) is used for the piecemeal reconstruction of haplotypes. We set the boundaries of the atomistic units to coincide with those of high-LD blocks. These regions were defined to be contiguous sequences in which all pairwise $|D'|$ (Lewontin 1988) among segregating sites are >0.8. The two-locus haplotype frequencies needed for the calculation of these values were estimated by the Weir-Cockerham two-point EM algorithm (Weir 1996). In our program, LD blocks longer than six sites were split to make atomistic units computationally manageable. Also, orphaned segregating sites that were not linked with any high-LD blocks were absorbed into the adjacent block containing a site with the maximum r^2 to the orphan.

With nuclear-family data, our program reconstructs the haplotypes of parents with children's genotypes used to constrain the former's haplotype space. On a more technical note, whenever an individual's haplotypes cannot be reconstructed to be equivalent to other haplotypes found in the population sample, a parent-independent mutation model is assumed that gives equal weight to all plausible reconstructions; this situation is rarely encountered in practice, because of the atomistic units used in the algorithm.

The goal of our program was to create a tool that achieves highly accurate haplotype reconstruction but that could be used, with reasonable execution times, on enormous data sets. The ultimate intent was to use the haplotypes reconstructed in this manner as alleles in disease-association studies (Lin et al. 2004*a*).

HAP (E.H. and E.E.)

HAP was extended (Halperin and Eskin 2004) to allow it to cope with genotypes typed from father-mother-child trios. The HAP algorithm assumes that the ancestral history of the haplotypes can be described by a perfect phylogeny tree. A perfect phylogeny tree is a genealogical tree with no recombinations and no recurrent mutations. HAP considers all phase assignments that result in a set of haplotypes that are almost consistent with a perfect phylogeny. Each assignment, H, is then given a score, $S(H)$, that is the maximum likelihood of the solution, under the assumption that the haplotypes were randomly picked from the population. More specifically,

$$S(H) = \max_{\Theta} \prod_{i}^{n} \theta_{HM_{i1}} \theta_{HM_{i2}} \theta_{HF_{i1}} \theta_{HF_{i2}} ,$$

where $HF_{i1} \oplus HF_{i2} = GF_i$ and $HM_{i1} \oplus HM_{i2} = GM_i$. HAP then chooses the phase assignment with the highest score. To phase a long region, HAP applies the perfect phylogeny model in a sliding window to short overlapping regions. These overlapping predictions are combined using a dynamic programming-based tiling algorithm that chooses the optimal phase assignment for the long region that is most consistent with the overlapping predictions of phase in the short regions (see Halperin and Eskin [2004] for more details).

Within a short region, the extension of HAP to trios must take into account the fact that the haplotypes of the children are copies of the haplotypes of the parents. We assume there are no recombinations or mutations between the parents and the children in the trios. This allows us, first, to unambiguously resolve the phase of the trios in many of the positions. For the remaining positions, we use HAP to enumerate all possible phase assignments. This results in a set of haplotypes that are almost consistent with a perfect phylogeny. In that enumeration, we exclude the solutions that contradict Mendelian inheritance within a trio. For each such solution, we give the likelihood score, which is the probability to observe the parents' haplotypes in our sample. We pick the solution with maximum likelihood as a candidate solution. To further improve the solution, we use a local search algorithm. The local search algorithm starts from the solution given by HAP, and it re-

peatedly changes the phase of one of the trios to a different possible phase and checks whether the likelihood function has increased. If it has increased, we use the new solution as the candidate solution and repeat this procedure. If no local change can be applied to increase the likelihood, we stop and use the solution as a putative solution for this region. HAP has been successfully applied to several large genomic data sets, including a whole-genome survey of genetic variation (Hinds et al. 2005).

tripleM and PL-EM (Z.S.Q., T. Niu, and J. Liu)

The tripleM algorithm is a direct extension of the EM algorithm (Dempster et al. 1977) used in maximum-likelihood haplotype reconstruction for unrelated individuals (MacLean and Morton 1985; Excoffier and Slakin 1995; Hawley and Kidd 1995; Long et al. 1995; Chiano and Clayton 1998; Qin et al. 2002); "PL-EM" is the name given to the version of the algorithm for unrelated data.

Assuming that there is no recombination event in this chromosomal region during meiosis, we write down the probability of observing the genotype data in a single trio family:

$$P(G_i|\Theta) \propto \sum_{HF_{i1}\oplus HF_{i2}=GF_i} \sum_{HM_{i1}\oplus HM_{i2}=GM_i} \times\theta_{HF_{i1}}\theta_{HF_{i2}}\theta_{HM_{i1}}\theta_{HM_{i2}} I_{HF_i\oplus HM_i=GC_i} ,$$

where $I_{HF_i\oplus HM_i=GC_i}$ is the indicator function for the event that $\exists u \in HF_i$ and $v \in HM_i$, such that $u \oplus v = GC_i$.

Assuming, further, a complete independence of the n trio families, we have the joint probability of the data from all the families as the product of that of individual ones. In the E step of the $(t+1)$th iteration of the EM algorithm, we compute the Q function as $Q(\Theta|\Theta^{(t)}) = \sum_{g=1}^{s} E_{\Theta^{(t)}}(n_g|G)\log\theta_g$, where

$$E_{\Theta^{(t)}}(n_g|G) = \sum_{i=1}^{n} \frac{\theta_{a_i}^{(t)}\theta_{b_i}^{(t)}\theta_{c_i}^{(t)}\theta_{d_i}^{(t)} I_{\{g\in\{a_i,b_i,c_i,d_i\}\text{and}\{a_i,b_i\}\oplus\{c_i,d_i\}=GC_i\}}}{\sum_{a_i'\oplus b_i'=GF_i}\sum_{c_i'\oplus d_i'=GM_i}\theta_{a_i'}^{(t)}\theta_{b_i'}^{(t)}\theta_{c_i'}^{(t)}\theta_{d_i'}^{(t)} I_{\{\{a_i',b_i'\}\oplus\{c_i',d_i'\}=GC_i\}}} .$$

In the M-step, the frequency vector is updated by maximizing the Q function, which gives rise to

$$\theta_g^{(t+1)} = \frac{E_{\Theta^{(t)}}(n_g|G)}{4n} .$$

For k linked SNPs, the total number of all possible distinct haplotypes is 2^k. The regular EM algorithm is unable to handle such a large number of SNPs, and computational techniques are required to allow this method to be applied to large regions. Partition-ligation (Niu et al. 2002) can be applied to solve this problem. At the beginning, the SNPs are divided into disjoint pieces, typically no more than eight SNPs in each piece. The above EM-based algorithm is then applied to all the trio families, to infer haplotype frequencies in each subset of markers. Since phasing on these subsets of markers is performed independent of one another, these steps can be performed in parallel, to speed up the process. Subsequently, adjacent pieces are ligated using the same EM algorithm. To keep the computation cost in check, only nonrare haplotypes are retained in each EM step. Essentially, tripleM is a direct extension of the PL-EM algorithm for haplotype reconstruction, seen in the work of Qin et al. (2002), and this approach has been used to construct haplotype phase for general pedigrees in the work of Zhang et al. (2005).

Summary of Methods

The descriptions of the above algorithms indicate that there are strong similarities among the models and assumptions they use (see table 1 for a summary of the properties of the five methods). We have also found it useful to consider differences among the methods from a formal point of view, in terms of the probability model on which they are based. We find it useful to think of each of the models from a Bayesian point of view, even though this may not be how all of the methods were developed and subsequently described. Within this framework, we wish to make inferences about the unknown haplotype reconstruction, H, and the population allele frequencies of the haplotypes, Θ, conditional on a set of observed genotype data G—that is, we wish to infer the posterior distribution $p(\Theta,H|G)$. By use of the Bayes rule, this can be written as

$$\Pr(\Theta,H|G) \propto \Pr(G|H)\Pr(H|\Theta)\Pr(\Theta) ,$$

and each method can be described in terms of the three factors on the right side of this expression. All five of the methods considered here use essentially the same expression for the first two factors. The first factor, $\Pr(G|H)$, models how consistent the haplotype configuration H is with the observed genotype data G. So, for trio data,

$$\Pr(G|H) = \prod_{i=1}^{n} \Pr[GC_i|HF_i = (hf,hf'),HM_i = (hm,hm')] ,$$

where $\Pr[GC_i|HF_i = (hf,hf'),HM_i = (hm,hm')]$ is computed under the assumption of no recombination between parents and child.

The second factor models the probability distribution of the haplotype reconstruction, H, given the population allele frequencies, Θ. All of the methods make the assumption of random mating in the population, to derive the following probability model:

$$\Pr(H|\Theta) = \prod_{i}^{n} \theta_{HM_{i1}}\theta_{HM_{i2}}\theta_{HF_{i1}}\theta_{HF_{i2}} .$$

Earlier, we saw that the key idea behind the PHASE algorithm is that, over short genomic regions, sampled chromosomes tend to cluster together into groups of similar haplotypes. This "clustering property" is encapsulated through the specification of a prior distribution on the population haplotype frequencies, Θ. PHASE and *wphase* use a prior that approximates the coalescent with recombination that puts more weight on distributions in which clusters of similar haplotypes tend to have nonzero frequency. Unfortunately, it is not possible to write down the form of this prior distribution directly, since PHASE and *wphase* directly specify the conditional dis-

Table 1

Properties of Haplotype-Inference Algorithms Used in the Present Study

Algorithm	Inference	Clustering Property	Recombination Model	Partition Ligation	Output
PHASE	MCMC	Approximate coalescent model	Estimated variable rates	Fixed chunk size	Best guess/sample/estimates of uncertainty
wphase	Maximum pseudo-likelihood	Approximate coalescent model	Fixed constant rate	Fixed chunk size	Best guess
HAP2	MCMC	None	None	LD-based variable chunk size	Best guess/sample/estimates of uncertainty
PL-EM/tripleM	Maximum likelihood (via EM)	None	None	Fixed chunk size	Best guess
HAP	Constrained maximum likelihood	Perfect phylogeny constraints	None	Overlapping chunks	Best guess

Table 2

Details of Simulated Data Sets Used in the Assessment of the Algorithms

Data Set	Details
ST1	100 data sets of 30 trios simulated with constant recombination rate across the region, constant population size, and random mating. Each of the 100 data sets consisted of 1 Mb of sequence.
ST2	Same as ST1, but with the addition of a variable recombination rate across the region.
ST3	Same as ST2, except a model of demography consistent with white Americans was used.
ST4	Same as ST3, with 2% missing data (missing at random).
SU1	100 data sets of 90 unrelated individuals simulated with constant recombination rate across the region, constant population size, and random mating. Each of the 100 data sets consisted of 1 Mb of sequence.
SU2	Same as SU1, but with the addition of a variable recombination rate across the region.
SU3	Same as SU2, except a model of demography consistent with white Americans was used.
SU4	Same as SU3, with 2% missing data (missing at random).
SU-100 kb	Since some studies may be concerned only with the performance of phasing algorithms on lengths of sequence shorter than 1 Mb, we simulated a set of data sets identical to set SU3, except that the sequences were only 100 kb in length. Each of these 100-kb data sets was created by subsampling a set of 1,180 simulated haplotypes. The remaining 1,000 haplotypes were used to estimate the "true" population haplotype frequencies. This allowed a comparison of each method's ability to predict the haplotype frequencies in a small region of interest.

Table 3

Details of the Real Data Sets Used in the Assessment of the Algorithms

Data Set	Details
RT-CEU	100 data sets consisting of 30 HapMap CEU trios across 1 Mb of sequence. For each data set, we created 30 new data sets, each with a different trio altered so that the transmission status of the alleles in one of the parents is switched. By switching only one trio at a time to create a new data set, the majority of the genotypes are unaltered, and a minimum amount of new missing data is introduced. In each region, the error rates for the different algorithms were calculated using only the phase estimates in the altered trios.
RT-YRI	Same as RT-CEU, except 30 HapMap YRI trios were used.
RU	We used HapMap CEU sample to create artificial data sets of unrelated individuals by simply removing the children from each of the trios. Since the phase of a large number of heterozygous genotypes will be known from the trios, we can use these phase-known sites to assess the performance of the algorithms for unrelated data. One hundred 1-Mb regions were selected at random from the CEU sample and processed in this way.

tributions needed to provide inference. (This does not guarantee that PHASE will converge to a proper probability distribution, but it is not thought to be a problem in practice [Stephens and Donnelly 2003].) A prior distribution that can be written down explicitly is the Dirichlet prior on haplotype frequencies,

$$\Pr(\Theta) = \frac{\Gamma(\sum_{j=1}^{s} \lambda_j)}{\prod_{j=1}^{s} \Gamma(\lambda_j)} \prod_{j=1}^{s} \theta_j^{\lambda_j - 1},$$

and this distribution does not encourage clustering of haplotypes in any way. Since HAP2 does not use all of the available data, it is not strictly correct to say that the method uses this prior. It has been suggested that, if HAP2 *did* use all of the available data, then the method would produce reconstructions very similar to those produced by a method that attempts to maximize the likelihood, such as the PL-EM/tripleM method. Differences in the partition-ligation schemes used by HAP2 and PL-EM/tripleM will also contribute to differences in their performance. A related approach, called "SNPHAP" (Clayton Web site), is based on the same model that underlies PL-EM but uses different computational tricks to deal with long regions. Thus, we would expect that this method would produce very similar results to those of PL-EM. Finally, the constraints on haplotype reconstructions in HAP can be thought of in terms of a prior distribution that encourages clusterings of haplotypes, although it would be difficult to write this down explicitly.

Results

Data Sets

To provide a comprehensive comparison of the algorithms, we constructed the following large sets of simulated and real data sets.

Simulated data.—We simulated haplotypes, using a coalescent model that incorporates variation in recombination rates and demographic events. The parameters of the model were chosen to match aspects of data from a sample of white Americans. Precise details of the parameters used and how they were estimated can be found in the work of Schaffner et al. (2005). Ascertainment of SNPs was modeled by simulating two extra groups of eight haplotypes. For each marker, two pairs of haplotypes were chosen randomly from each group of eight (independently, from marker to marker), and the marker was considered "ascertained" if either pair was heterozygous. Markers were then thinned to obtain the required 1 SNP per 5 kb density that was used throughout the present study. The details of the simulated data sets are given in table 2. Before the actual performance tests, two sets of simulated data, together with the answers, were provided to all those involved in writing and extending the algorithms described above, to facilitate algorithm development.

Real data.—We also used publicly available data from the HapMap project to compare the different algorithms. The HapMap data consists of genotypes of 30 trios from a population with European ancestry (denoted "CEU"), 30 trios from a population with African ancestry (denoted "YRI"), and 45 unrelated individuals from each of the Japanese and Chinese populations (denoted "JPT" and "CHB," respectively). For both CEU and YRI samples, we randomly selected 100 1-Mb regions with ~1 SNP per 5 kb. The form of genotype data on trios is such that the transmission status of many alleles can be identified unambiguously. Thus, the genotypes of other plausible offspring can be created by switching the transmission status of the alleles in the parents' genotypes. This process is illustrated in figure 1. A summary of the real data sets used is given in table 3. It is worth noting that, in total, the data sets created in this way represent 6,100 Mb of genetic data consisting of ~1.22 million SNPs. As such, it was not possible to apply all of the algorithms to the real data sets because of limitations on the computational resources available to the authors at the time of the study.

Criteria

We used six different criteria to assess the performance of the algorithms.

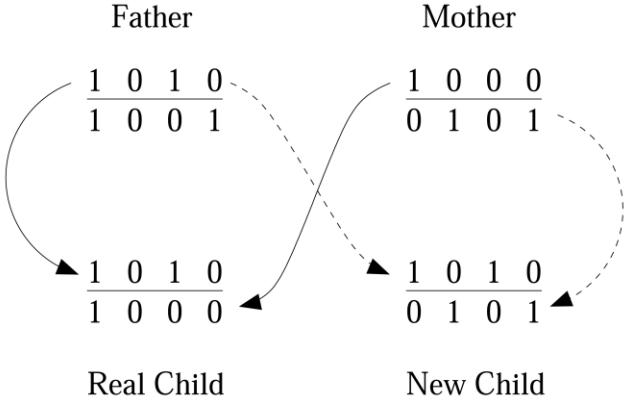

Figure 1 The method of constructing new data sets with artificially induced ambiguous sites from real trio data. The example in the figure consists of a father-mother-child trio at four SNPs. The genotypes at all sites are such that the haplotypes of each individual can be inferred exactly. A new "alternative universe" child can be created by swapping the transmission status of the haplotypes in one of the parents. In this example, both children inherit the "1010" haplotype from the father but inherit different haplotypes from the mother; the real child inherits the "1000" haplotype, and the new child inherits the "0101" haplotype. When the trio consisting of the father, the mother, and the new child is considered, we see that the transmission status of the fourth SNP is now not known unambiguously if we consider just the genotypes at the site. The performance of phasing algorithms can be assessed for these data sets by their ability to reconstruct the correct phase at these sites.

Switch error.—Switch error is the percentage of possible switches in haplotype orientation, used to recover the correct phase in an individual or trio (Lin et al. 2004*b*).

Incorrect genotype percentage (IGP).—We counted the number of genotypes (ambiguous heterozygotes and missing genotypes) that had their phase incorrectly inferred and expressed them as a percentage of the total number of genotypes. To calculate this measure, we first aligned the estimated haplotypes with the true haplotypes, to minimize the number of sites at which there were phase differences. For the trio data, this alignment is fixed by the known transmission status of alleles at nonambiguous sites. For the real data sets in which the truth for the missing data was not known, we removed such sites from consideration in both the numerator and the denominator. We believe the utility of this measure lies in its comparison with levels of genotyping error and missing data.

Incorrect haplotype (IHP).—IHP is the percentage of *ambiguous* individuals whose haplotype estimates are not completely correct (Stephens et al. 2001). It is worth noting that, as the length of the considered region increases, all methods will find it harder to correctly infer entire haplotypes. Thus, this measure will increase with genetic distance and eventually reach 100%, once the region becomes long enough.

Missing error.—Missing error is the percentage of incorrectly inferred missing data. To calculate this measure, we first aligned the estimated haplotypes with the true haplotypes, to minimize the number of sites at which there were phase differences. This alignment ignored the sites at which there was missing data. We then compared the estimated and true haplotypes at the sites of missing data and counted the number of incorrectly imputed alleles and then expressed this as a percentage of the total number of missing data.

χ^2 distance.—For SU 100-kb data sets, we also used the estimated haplotypes produced by each method to define a vector of haplotype frequencies $\{q_1, \ldots, q_k\}$, and we compared these with the population frequencies $\{p_1, \ldots, p_k\}$, using the χ^2 difference

$$\sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i} .$$

Two of the methods (PHASE and HAP) also produced explicit estimates of the population haplotype frequencies; these were also compared with the population frequencies by use of the same measure.

Running time.—For each of the methods, we recorded the running time for a subset of the simulated data sets. Because of limitations on the amount of available computing resources and the portability of some code, it was not possible to run all the methods on the same computer, so we also report some details of the computing resources used by each of the authors.

The switch error, IGP, IHP, and missing error were calculated by summing the number of errors or switches across all data sets and individuals and dividing by the total number of possible errors or switches across all data sets and individuals. Some of the real data sets have missing data; thus, the true haplotypes are not known completely, and it is not possible to calculate the switch error, IGP, or IHP measures. To deal with this problem, we calculated the error measures in a given individual or trio, using only sites for which there is no missing data.

Performance

The performance of the methods on the simulated and real data sets are shown in tables 4–7. When interpreting these results, it should be kept in mind that these results

Table 4

Error Rates for the Methods Applied to the Data Sets for Simulated Trio and Unrelated Individuals

Error Measure and Recombination Rate	Error Rate (%)				
	PHASE	*wphase*	HAP	HAP2	tripleM
Switch error:					
ST1	***.74***	.98	2.14	2.58	3.02
ST2	***.22***	.22	1.51	5.97	2.87
ST3	***1.36***	2.23	2.4	2.95	3.81
ST4	***1.48***	2.34	2.62	3.17	4.12
SU1	***2.4***	3.7	6.5	6.9	9.0
SU2	***2.2***	3.7	9.8	15.1	13.1
SU3	***4.8***	6.2	7.1	8.2	11.0
SU4	***5.3***	6.9	7.8	9.2	11.4
SU-100 kb	***4.3***	5.3	5.6	5.7	8.3
IGP:					
ST1	***.05***	.08	.17	.23	.24
ST2	***.02***	.02	.11	.43	.20
ST3	***.12***	.20	.21	.27	.33
ST4	***.12***	.19	.20	.29	.34
SU1	***2.5***	3.5	7.9	7.1	5.8
SU2	***2.4***	4.3	9.5	11.0	8.0
SU3	***5.1***	5.8	8.5	8.6	8.2
SU4	***5.2***	5.8	8.4	8.7	8.0
SU-100 kb	***1.5***	1.8	1.9	2.0	2.3
IHP:					
ST1	***5.5***	6.5	12.8	17.2	18.6
ST2	***1.9***	1.9	11.4	36.2	21.2
ST3	***10.4***	14.2	17.0	20.8	24.8
ST4	***10.3***	14.7	17.8	21.3	25.0
SU1	***35.5***	48.0	88.6	73.5	61.1
SU2	***40.4***	52.1	97.1	99.0	83.4
SU3	***59.1***	66.4	90.1	85.1	81.4
SU4	***60.8***	68.0	90.6	87.1	81.5
SU-100 kb	***17.2***	19.4	21.8	22.2	24.7
Missing error:					
ST4	***1.46***	1.89	4.36	5.26	3.38
SU4	***7.3***	9.0	11.6	15.0	19.4

Note.—The results for the best-performing method in each row are highlighted in bold italics.

are specific to the density of SNPs and sample size of the data sets used. Several clear patterns are evident in these tables.

Overall, the performance of all the data sets is very good. For the best method, we observed percentages of genotypes that had their phase incorrectly inferred: 0.12% for trios and 5.2% for unrelated individuals on simulated data sets, 0.05% and 5.9% on HapMap CEPH trios and unrelated individuals, respectively, and 0.16% on HapMap Yoruban trios (table 4). These results clearly show the difference in error rates between the use of trio and unrelated samples (compare ST and SU data sets in table 4). The error rates for the trio data sets are comparable to expected levels of genotyping error and missing data and highlight the level of accuracy that the best-phasing algorithms can provide on a useful scale.

For the trio data sets, the PHASE algorithm consistently provided the best performance (compare methods for ST data sets of table 4). Of the other methods, the *wphase* algorithm is the next best and is followed by HAP, HAP2, and tripleM, in that order. The only exception is that tripleM sometimes has a better performance than HAP2 (i.e., for ST2 data set, regardless of error measure).

For the data sets of unrelated individuals, the PHASE algorithm consistently provided the best performance, followed by *wphase* (compare methods for SU data sets of table 4). Of the other methods, PL-EM seems to perform better than both HAP and HAP2 in terms of IGP and IHP but less well in terms of switch error. This suggests that the haplotypes that PL-EM infers incorrectly require a relatively large number of switches to be made correct.

As expected, the performance is better for trio data sets than for unrelated individuals. Another useful summary of the performance of the algorithms that highlights the differences between the use of trio and unrelated data is the rate of switch errors per unit of physical distance. For the real-data-set comparisons shown in table 6, the results of the PHASE algorithm correspond to an average of one (trio) switch error every 8 Mb and every 3.6 Mb for the RT-CEU and RT-YRI data sets, respectively. For the RU data sets, we observed an average of one switch error every 333 kb of sequence. As mentioned above, these figures are relevant only to the SNP density and sample size of the data sets analyzed.

The performance of PHASE is improved in the scenarios in which recombination occurs in hotspots (ST2 and SU2 in table 4), compared with the scenarios that have constant recombination rates (ST1 and SU1 in table 4). This pattern does not hold, in general, for the other methods.

The error rates for simulated data depend on the demographic models assumed, because there is a difference in performance of the data sets simulated using a model of demography that is based on real data (ST3 and SU3 in table 4) and those simulated using a model that assumes constant population size and random mating (ST2 and SU2 in table 4).

The error rates for the data simulated with "CEU-like" demography are higher than real CEU data sets (compare ST4 and SU4 data sets in table 4 with RT-CEU and RU-CEU in table 6). It is difficult to specify the exact reason for this, but potential explanations include differences in the amount and pattern of missing data, differences in the levels of recombination, and differences in the real and simulated demographic events.

There is a large variation in the running times of the different methods (see table 7). For the simulated trio data sets, the fastest algorithm was tripleM, at 1.5 s. The algorithms HAP2, HAP, *wphase,* and PHASE took 12, 15, 4,480, and 8,840 times as long, respectively. For simulated unrelated data sets, HAP was the fastest algorithm, at 35.1 s. The algorithms HAP2, PL-EM, PHASE, and *wphase* took 3.6, 7.5, 1,114, and 12,205 times as long, respectively. Even so, PHASE was successfully applied to infer haplotypes from phase I of the HapMap project (1 million SNPs genotyped in two sets of 30 trios and a set of 89 unrelated individuals [International HapMap Consortium 2005]). (See J.M.'s Web site for online material with details of the haplotype estimation for phase I of the HapMap project.)

Estimation of r^2

In a given study, it is often of interest to consider the pattern of (pairwise) LD across a region for which genotype data has been collected. Estimates of LD are useful for visualization of the LD structure in a region or for purposes of defining a set of tagging SNPs for use in association studies (Johnson et al. 2001; Carlson et al. 2004). A commonly used measure is the squared correlation coefficient (r^2) within haplotypes; it cannot be calculated directly from genotype data. We evaluated the following methods of estimating r^2 between a pair of SNPs within a given region.

1. First, estimate haplotypes with the algorithms considered in the present study (PHASE, *wphase,* HAP, HAP2, and tripleM/PL-EM) and then estimate r^2 between each pair of SNPs, as if these were the true haplotypes.
2. Use genotypes for pairs of markers to estimate r^2 with the EM algorithm (pairwise) (Weir 1996).
3. Use the genotype correlation (GC).

We applied these methods to the simulated trio and unrelated data sets with (ST4 and SU4) and without (ST3 and SU3) missing data. For each of the 100 regions within each of these data sets, we first calculated the mean squared error of the true and estimated r^2, aver-

Table 5

Average χ^2 Distances of the True versus Estimated Population Haplotype Frequencies for Each Algorithm, Applied to the 100-kb Simulated Data Sets of Unrelated Individuals

Population Frequency	Comparison of χ^2 Distances by Algorithm					
	PHASE	*wphase*	HAP	HAP2	PL-EM	Sample Haplotypes
All frequencies	.70 (***.44***)	.77	.69 (.76)	.67	.83	.50
Frequencies >5%	.030 (***.027***)	.034	.034 (.07)	.034	.066	.028

Note.—The estimated haplotypes produced by each method were used to construct estimates of the population frequencies. In addition, HAP and PHASE provided explicit estimates of the population haplotype frequencies; the χ^2 distances for these approaches are given in parentheses. The χ^2 distances were calculated by summing over all population frequencies and by summing over all population frequencies >5%. The final column shows the χ^2 distance between the true population frequencies and the estimates produced by the true sample haplotypes. The results for the best-performing method in each row are highlighted in bold italics.

aged these values across the 100 data sets, and took the square root to give a root-mean-square-error (RMSE) measure.

The results are given in table 8 and show that all the methods do well at estimating r^2. The methods that estimate haplotypes do better than the methods that use only pairs of markers or use the GC. The most accurate estimates were obtained using PHASE to estimate haplotypes.

To gain a sense of the actual difference in estimates produced by the different methods, we chose a typical data set from each of the trio and unrelated data sets and plotted the true and estimated r^2 for the PHASE, pairwise, and GC methods (fig. 2). The figure shows that all methods have a tendency to produce errors on low values of r^2, but that high values of r^2 (>0.8) are estimated well. The figure also shows that the GC method is much less accurate than the PHASE and pairwise methods.

Benchmarks

To date, no comprehensive comparison has been performed between existing phasing algorithms. When comparisons have been performed, they have often involved small data sets of limited relevance. It is our intention that the data sets used in the present study form the basis of a benchmark set of data made freely available for the further development and open assessment of methods. Instructions for obtaining these data sets can be found at the authors' Web site.

Discussion

Inference of haplotype phase continues to be an important problem. With the advent of genomic-scale data sets, the size of the inference task has grown well beyond that on which many methods were developed and originally compared. The motivation for the present study was the HapMap project, in the first phase of which 1 million SNPs were genotyped in two sets of 30 trios and one set of 89 unrelated individuals. We extended some of the best current phasing algorithms to deal with trio data and undertook a comprehensive performance assessment of the algorithms for large simulated and real data sets.

The results of the comparison are encouraging. All of the algorithms produce comparable error rates. The most accurate algorithm was PHASE (v2.1). For this method, the percentages of genotypes whose phase was incorrectly inferred were 0.12%, 0.05%, and 0.16% for trios from simulated data, HapMap CEPH trios, and HapMap Yoruban trios, respectively, and 5.2% and 5.9% for unrelated individuals in simulated data and HapMap CEPH data, respectively.

When these results are interpreted, it is important to

Table 6

Error Rates for Methods Applied to the Real Data Sets

Error Measure and Sample	Error Rate (%) of Algorithm Applied to Real Data Sets				
	PHASE	*wphase*	HAP	HAP2	tripleM/PL-EM
Switch error:					
RT-CEU	***.53***	…	3.30	1.81	…
RT-YRI	***2.16***	…	7.34	…	…
RU	***5.43***	…	6.92	8.21	…
IGP:					
RT-CEU	***.05***	…	.28	.15	…
RT-YRI	***.16***	…	.49	…	…
RU	***5.84***	…	7.13	7.42	…
IHP:					
RT-CEU	***6.20***	…	20.07	17.51	…
RT-YRI	***15.7***	…	42.02	…	…
RU	***82.6***	…	91.9	90.8	…

Note.—Data sets are based on the HapMap data. Not all methods were run on these data sets, because of restrictions on the computational resources available to the authors. The results for the best-performing method in each row are highlighted in bold italics. See table 2 for description of data sets RT and RU.

Table 7

Findings for Each Method on the ST4 and ST3 Sets of Simulated Data

Algorithm	Mean Running Time by Data Set		Processor Details
	ST4	SU4	
wphase	1 h 52 min	119 h	Intel Xeon (2.8 GHz)
HAP	22.3 s	***35.1 s***	Intel Xeon (3.06 GHz)
HAP2	18.4 s	2 min 6 s	AMD Opteron 248 (2.2 GHz)
PHASE	3 h 32 min	10 h 52 min	AMD Opteron 246 (2.0 GHz)
tripleM/PL-EM	***1.5 s***	4 min 22 s	Intel Pentium (2.4 GHz)

Note.—The fastest performing method in each column is highlighted in bold italics.

remember that these error rates were produced on data sets with the particular average SNP density (1 SNP per 5 kb) and number of individuals used by HapMap and that care should be taken when trying to extrapolate these error rates to data sets with different numbers of individuals and different densities of SNPs. Generally speaking, the practical experience of all the authors involved in this study and previous simulation results (Stephens et al. 2001) lead us to believe that error rates will decrease with increased SNP density and increased sample sizes. We also have no evidence to suggest that the relative performance of the methods will change. For the data sets considered in the present study, the error rates for the trio data sets are comparable to expected levels of genotyping error and missing data and highlight the level of accuracy that the best phasing algorithms can provide on a useful scale.

The models underlying the methods studied here involve various assumptions. These assumptions will invariably be false for real data sets, and it is of interest to assess the extent to which performance changes with departures from these assumptions. For example, all the methods explicitly assume that parents of the trio data sets or the individuals in the unrelated data sets were sampled independent of the population. This may not be true in disease studies in which the trios may have been chosen because the child is affected or when a large proportion of the unrelated individuals are cases. Such sampling schemes will tend to lead to a departure from the explicit Hardy-Weinberg equilibrium (HWE) assumption of all the methods. For disease models in which risk increases with the number of risk alleles, such biased sampling will tend to increase the amount of homozygosity in the sample around the disease loci, which tends to reduce the number of ambiguous genotypes. Other disease models can be conjectured that would decrease homozygosity, but analyses focused on this point have suggested that departures from the HWE assumption are not a great cause for concern (Stephens et al. 2001). In addition, during the HapMap Project, it became clear that there was some unexpected relatedness between individuals in some of the analysis panels (International HapMap Consortium 2005), but our analysis shows that the results of all algorithms are still good. One could study extreme departures from the assumptions made by the approaches studied here (Niu et al. 2002), but we feel the most-informative measures of performance for many applications will be the behavior of the methods on the large real data sets we studied.

We anticipate several forthcoming challenges for haplotype-inference methods. One is to deal with inference in pedigrees that are more complex than trios (Abecasis and Wigginton 2005). Another, post-HapMap and other genomic resources, is to incorporate information about haplotypes known to be present in a population—and their frequency—in the inference of haplotypes from newly sequenced or genotyped individuals from the same or

Table 8

Accuracy of r^2 Estimation

Data Set and Algorithm	RMSE	
	With Missing Data	Without Missing Data
Trios:		
PHASE	.003	.002
wphase	.004	.003
HAP	.007	.004
HAP2	.007	.004
tripleM	.004	.005
Pairwise	.011	.009
GC	.032	.030
Unrelated individuals:		
PHASE	.011	.011
wphase	.015	.014
HAP	.022	.022
HAP2	.022	.020
PL-EM	.025	.029
Pairwise	.019	.018
GC	.025	.023

Note.—For each data set, we calculated the mean squared error of the true and estimated r^2, averaged these values across the 100 data sets, and took the square root to give an RMSE. RMSE is based on PHASE estimated haplotypes (PHASE, wphase, HAP, HAP2, tripleM/PL-EM), pairwise EM algorithm (pairwise), and GC. Results are based on the simulated data sets of trios with and without missing data (ST4 and ST3) and unrelated individuals with and without missing data (SU4 and SU3).

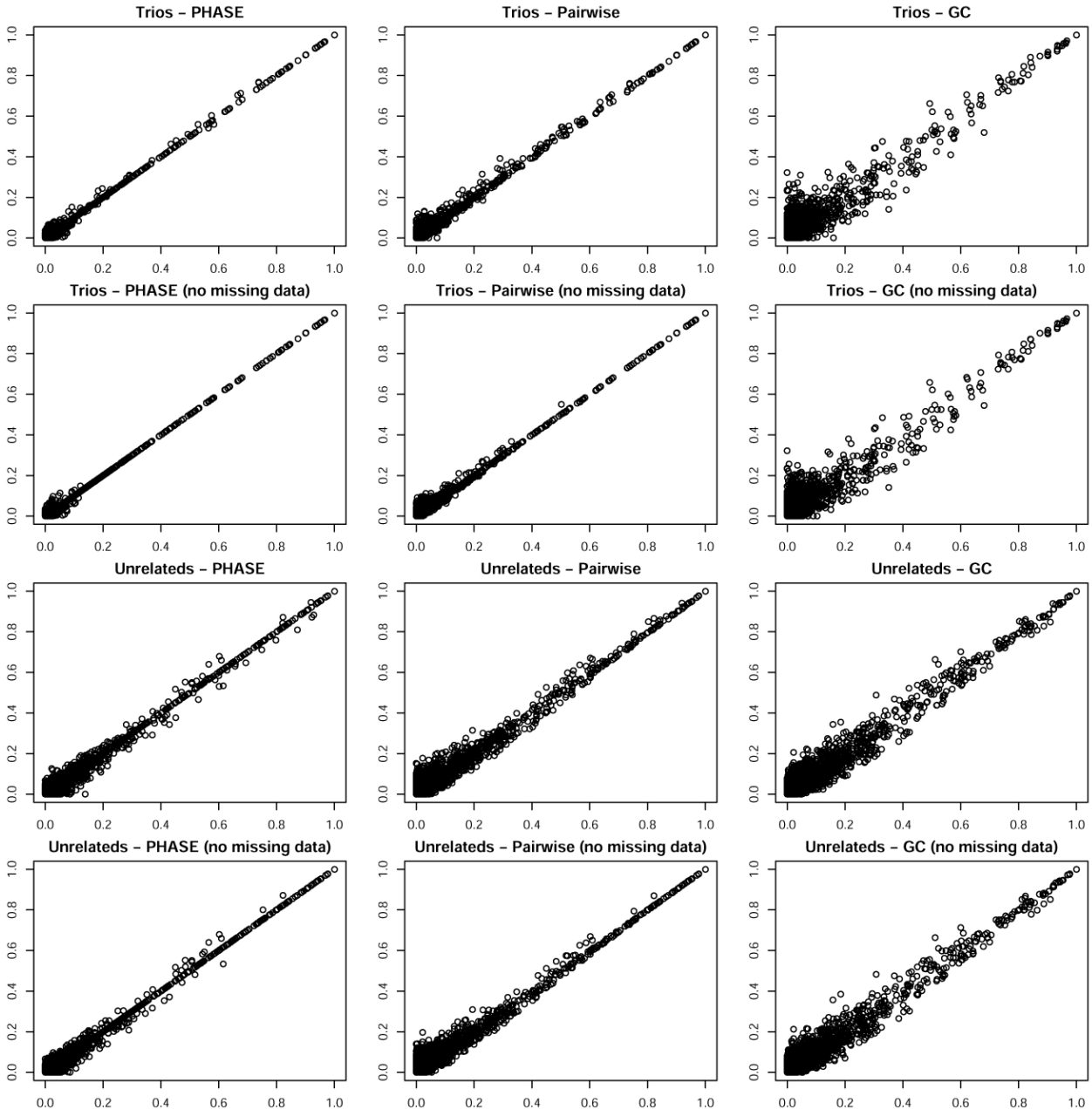

Figure 2 True (*X*-axis) and estimated (*Y*-axis) r^2 for the PHASE (*left column*), pairwise (*center column*), and GC methods (*right column*) and with (*rows 1 and 3*) and without (*rows 2 and 4*) missing data. Rows 1 and 2 show the differences for a trio data set, whereas rows 3 and 4 show the differences for the unrelated individuals data set. The data set was chosen at random from the 50 data sets analyzed.

a closely related population. A number of the methods described above lend themselves to this setting, and work in this direction is under way.

A different type of question, which is also unresolved, is whether and, if so, how best to use inferred haplotypes in downstream analyses. All of the methods considered produce a most likely set of haplotypes for their respective models, but some (the ones based on MCMC) can naturally produce a sample of plausible haplotype reconstructions that encapsulate the uncertainty in the estimates (see table 1). The question is whether it is important to use these estimates of uncertainty in downstream analyses.

We saw above that, for estimation of r^2, considerable improvements in accuracy result from inferring haplotype phase (by use of PHASE) and then estimating r^2 from the inferred haplotypes. At least in some settings, improved estimation of recombination rates and of his-

torical recombination events can also result from first estimating haplotypes and then treating these as known (International HapMap Consortium 2005). We did not look at estimating r^2 by integrating out over the uncertainty in the haplotypes, and this might perform even better than the two-stage procedure we did use.

In contrast, several studies have suggested that simply "plugging in" haplotype estimates to analysis methods can be suboptimal—namely, the studies of Morris et al. (2004) in the context of fine mapping and Kraft et al. (2005) for estimation of haplotype relative risks. Both studies used maximum likelihood for phase estimation. We saw above that this is one of the worst-performing methods considered here, effectively because this method does not give more weight to solutions in which haplotypes cluster together. In addition, both studies considered situations—20 SNPs in 1 Mb for 100 cases/controls in the work of Morris et al. (2004) and 4 SNPs for 200 cases/controls in that of Kraft et al. (2005)—in which there remained considerable uncertainty over estimated haplotypes. The density of SNPs in future studies will vary—depending on available resources, the genotyping platform used to assay the data, and the way in which the assayed SNPs have been chosen—and will likely lie somewhere between the density of the HapMap samples and the sparse simulated data sets considered by Morris et al. (2004) and Kraft et al. (2005). Knowledge of the haplotypes from the HapMap project should allow us to make much more accurate estimates of haplotypes, whatever the density of the markers in the future projects. Thus, uncertainty in haplotypes will be much less than it would have been if the HapMap data were not available. We suggest that the jury remains out on this question, pending studies that use the best phase-estimation methods on realistic-sized data sets and studies that take the HapMap data into account, for which accurate phase estimation is more likely.

In the specific context of disease-association studies, there remains an open question about how best to combine information across markers. Doing so could but need not necessarily use haplotype information. Chapman et al. (2003) have shown that, in a particular framework, the cost, in terms of additional parameters, of including haplotypes in the analysis, rather than simply using multilocus genotypes, outweighs the benefits for *detecting* a disease variant. In contrast, Lin et al. (2004*a*) show that haplotype information has an important role in detecting rare variants. Different issues arise in localization, and Zollner and Pritchard (2005) have shown that haplotypes can be valuable in this context.

Acknowledgments

We are grateful to Steve Schaffner for help and advice in using a sophisticated coalescent-based simulator, which allowed us to generate haplotype data with complex demographics. J.M. was supported by the Wellcome Trust. P.D. was supported by the Wellcome Trust, the National Institutes of Health (NIH), The SNP Consortium, the Wolfson Foundation, the Nuffield Trust, and the Engineering and Physical Sciences Research Council. M.S. is supported by NIH grant 1RO1HG/LM02585-01. N.P. is a recipient of a K-01 NIH career-transition award. G.R.A. is supported by NIH National Human Genome Research Institute grant HG02651. E.E. is supported by the California Institute for Telecommunications and Information Technology, Calit2. Computational resources for HAP were provided by Calit2 and National Biomedical Computational Resource grant P41 RR08605 (National Center for Research Resources, NIH).

Web Resources

The URLs for data presented herein are as follows:

Authors' Web site, http://www.stats.ox.ac.uk/~marchini/phaseoff.html
Clayton Web site, http://www-gene.cimr.cam.ac.uk/clayton/software/ (for the SNPHAP algorithm)
HAP, http://research.calit2.net/hap/
International HapMap Project, http://www.hapmap.org/
J.M.'s Web site, http://www.stats.ox.ac.uk/~marchini/HapMap.Phasing.pdf (for details of how haplotypes were inferred for the PHASE v.1 HapMap)
PHASE, http://www.stat.washington.edu/stephens/software.html
PL-EM, http://www.people.fas.harvard.edu/~junliu/plem/click.html
tripleM, http://www.sph.umich.edu/csg/qin/tripleM/

References

Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767

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

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci USA 98:4563–4568

Carlson C, 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

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

Chiano M, Clayton D (1998) Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 62:55–60

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

De Iorio M, Griffiths R (2004) Importance sampling on coalescent histories. II. Subdivided population models. Adv Appl Probab 36: 434–454

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via EM algorithm. J R Stat Soc B 39:1–38

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

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

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estima-

tion for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959
Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318
Gusfield D (2000) A practical algorithm for optimal inference of haplotypes from diploid populations. Proc Int Conf Intell Syst Mol Biol 8:183–189
——— (2001) Inference of haplotypes from samples of diploid populations: complexity and algorithms. J Comput Biol 8:305–323
——— (2003) Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. Paper presented at the Proceedings of the 6th Annual International Conference on Computational Biology, Washington, DC
Halperin E, Eskin E (2004) Haplotype reconstruction from genotype data using imperfect phylogeny. Bioinformatics 20:1842–1849
Hawley M, Kidd K (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86: 409–411
Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
Hoppe F (1987) The sampling theory of neutral alleles and an urn model in population genetics. J Math Biol 25:123–159
Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease Nature 411:599–603
International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237
Kraft P, Cox D, Paynter R, Hunter D, De Vivo I (2005) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261–272
Lazzeroni L (2001) A chronology of fine-scale gene mapping by linkage disequilibrium. Stat Methods Med Res 10:57–76
Lewontin R (1988) On measures of gametic disequilibrium. Genetics 120:849–852
Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233
Lin S, Chakravarti A, Cutler D (2004*a*) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36:1181–1188
——— (2004*b*) Haplotype and missing data inference in nuclear families. Genome Res 14:1624–1632
Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129–1137
Long J, Williams R, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810
MacLean C, Morton N (1985) Estimation of myriad haplotype frequencies. Genet Epidemiol 2:263–272
McVean G, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584
Morris AP, Whittaker JC, Balding DJ (2004) Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide–polymorphism genotype data. Am J Hum Genet 74:945–953
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324
Myers S, Griffiths R (2003) Bounds on the minimum number of recombination events in a sample history. Genetics 163:375–394
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
Puffenberger E, Kauffman E, Bolk S, Matise T, Washington S, Angrist M, Weissenbach J, Garver KL, Mascari M, Ladda R, Slaugenhaupt SA, Chakravarti A (1994) Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3:1217–1225
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
Rioux J, Daly M, Silverberg M, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29: 223–228
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
Salem M, Wessel J, Schork J (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2:39–66
Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583
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, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, Sunderland, MA
Zhang K, Sun F, Zhao H (2005) HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics 21:90–103
Zollner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092