Statistica Sinica **23** (2013), 1441-1461
doi:http://dx.doi.org/10.5705/ss.2012.141s

ACCURATE CONSTRUCTION OF LONG RANGE HAPLOTYPE IN UNRELATED INDIVIDUALS

Nicholas A Johnson[1], Stephanie J. London[2], Isabelle Romieu[3]
Wing H. Wong[1], Hua Tang[1]

[1]*Stanford University,* [2]*National Institute of Environmental Health Sciences, and* [3]*International Agency for Research on Cancer, France*

Abstract: Haplotype, or the sequence of alleles along a single chromosome, has important applications in phenotype-genotype association studies, as well as in population genetics analyses. Because haplotype cannot be experimentally assayed in diploid organisms in a high-throughput fashion, numerous statistical methods have been developed to reconstruct probable haplotype from genotype data. These methods focus primarily on accurate phasing of a short genomic region with a small number of markers, and the error rate increases rapidly for longer regions. Here we introduce a new phasing algorithm, `emphases`, which aims to improve long-range phasing accuracy. Using datasets from multiple populations, we found that `emphases` reduces long-range phasing errors by up to 50% compared to the current state-of-the-art methods. In addition to inferring the most likely haplotypes, `emphases` produces confidence measures, allowing downstream analyses to account for the uncertainties associated with some haplotypes. We anticipate that `emphases` offers a powerful tool for analyzing large-scale data generated in the genome-wide association studies (GWAS).

Key words and phrases: Expectation maximization, graphical model, haplotype, phasing.

1. Introduction

The term haplotype refers to the combination of alleles at multiple loci along a chromosome. In population genetics studies, haplotype data provide richer information regarding the shared genealogical history between two chromosomal segments than genotype at each marker separately. Thus, genome-wide pattern of haplotype similarity has been used to generate fine-scale delineation of population structure and relationship between individuals (Lawson et al. (2012)). Furthermore, haplotype data have been used to detect recombination hotspots as well as signature of recent positive selection; for example, the extended haplotype homozygosity test (EHH), one of the most widely used test for detecting select sweep, is based on comparing the observed length of a haplotype carrying a putatively beneficial allele to its expected length under neutral evolution

(Myers et al. (2005); Sabeti et al. (2002)). Despite the usefulness of haplotype information, most current genomic technologies generate genotype data, or unordered pairs of alleles. Although new sequencing techniques hold promise to directly assay haplotypes, these methods currently are not amenable for large-scale genetic studies (Yang, Chen, and Wong (2011); Kitzman et al. (2011); Fan et al. (2011)). While reconstructing haplotypes from genotypes is straightforward in some special settings (e.g., in the presence of relatives, in sperm, or for X chromosomes in males), statistical inference of haplotype from autosomal genotype data with no known relatives is challenging.

The work presented here is motivated by two recent studies. Using data in a cohort of $35,528$ Icelandic individuals genotyped by deCode, Kong et al. (2008) demonstrated that the accuracy of haplotypes inferred by the commonly used approaches (such as `FastPHASE` described below) are limited to a short range: the error rate rise to 30% for phasing a region of $6.4Mb$. These authors proposed a novel long-range phasing method (LRP) that is based on the rationale that, in a large sample, some individuals may be close relatives, such as second cousins. The LRP implements a heuristic algorithm that identifies "surrogate" parents, who share at least one allele identical-by-state (IBS) with the proband over an extremely long region (e.g., $1,000$ consecutive SNPs). The phasing of the proband are inferred by constructing the two *obligatory* haplotypes that must be shared by his/her surrogate parents. The LRP method was remarkably successful for the deCode sample considered in Kong et al.: for the $10Mb$ MHC region on chromosome 6, 87% of the individuals can be fully phased, and 7% of the remaining individuals can be phased for 90% of the heterozygous sites; overall, 93.7% of the heterozygous sites can be phased unambiguously. However, in many human populations, randomly sampled and unrelated individuals are on average much less closely related than they are in the Icelandic population. As a result, the LRP method does not apply because many individuals find no surrogate parents in the sample, unless a substantial fraction of the population is sampled. Nonetheless, the LRP method offers a powerful intuition, which we aim to capture in the current work.

A second study that motivated our research is Higasa et al. (2009), who genotyped 100 haploid Japanese genomes from complete hydatidiform moles (CHM) that arise when an empty egg with no nucleus is fertilized by a normal sperm. In their search for regions that harbor signature of recent positive selection, the authors found that accurate haplotype, derived from CHM or trio-phased chromosomes, can identify extended haplotype sharing that is obscured in statistically inferred haplotypes. These results underscore the needs to improve accuracy in existing statistical methods for haplotype inference.

In the next section, we provide a brief overview of existing phasing algorithms. Section 3 presents a new approach, `emphases` that is aimed to improve

long range haplotype construction using unrelated genotype data. The method is designed to analyze dense genotype data in large samples, such as those generated in GWAS. In assessing the performance of the proposed method (Section 4), we make use of a unique cohort of 492 parent-offspring trios that enable unambiguous determination of haplotypes at a majority of sites.

2. Background and Existing Phasing Algorithms

Numerous methods have been developed to infer haplotypes. The method of Clark (1990) begins by identifying a pool of unambiguous (homozygous) individuals, and phases the remaining individuals based on a parsimony heuristic that seeks to minimize the total number of distinct haplotypes in the sample. For a small number of linked markers, multinomial-based models fitted by the Expectation-Maximization (EM) algorithms can be quite effective (Excoffier and Slatkin (1995); Hawley and Kidd (1995); Long, Williams and Urbanek (1995)). The partition-ligation (PL-EM) algorithm of Niu et al. (2002) was proposed to accelerate computation and to keep the EM algorithm from becoming trapped in poor local modes. These methods perform reasonably well in identifying common haplotypes. However, the multinomial model is inappropriate for rare haplotypes, which is a serious weakness because for any fixed sample size of individuals, a majority of haplotypes become rare or unique as the number of markers increases – either by increasing marker density or by expanding the genomic region.

An important feature of the phasing problem, which was ignored by the early methods, is that all haplotypes are related through a genealogy; as such a *novel* haplotype that resembles a common haplotype is more plausible than one that resembles no other observed haplotype. This intuition motivated the coalescent-based Bayesian approach in the program, `PHASE` (Stephens, Smith and Donnelly (2001); Scheet and Stephens (2005)). Informally, the rationale that underlies `PHASE` can be understood using a simple example. The HapMap phase III project genotyped 50 parent-offspring trios from the Yoruba population in Nigeria (YRI); the trio relationship allows the accurate phasing of the 200 parental haplotypes. How could these haplotype templates be used to phase a new individual unrelated to any of the HapMap trios? `PHASE` seeks the most plausible configuration, in which the new haplotypes are derived from the HapMap haplotype templates through recombination and (rare) mutation events. To phase multiple individuals in the absence of an appropriately known haplotype templates, `PHASE` aims to jointly model all unobserved haplotypes. `PHASE` implements a Markov Chain-Monte Carlo (MCMC) algorithm which constructs a Markov chain with the stationary distribution corresponding to the desired posterior distribution of the haplotypes given genotypes. Each step of the MCMC algorithm samples a pair of haplotypes for an individual; the likelihood of these proposed haplotypes

is computed based on their similarities to the putative haplotypes of other individuals, which serve as haplotype templates and are treated as known without error. Novel haplotypes arise by recombinations and mutations in the observed haplotypes at rates that are determined by a population genetic model and coalescent. It is useful to observe that the haplotype templates idea in PHASE can be thought of as an elegant generalization of "surrogate" parents in LRP, but the mosaic model allows these surrogate parents to contribute haplotype segments of any length. In many subsequent studies that compares phasing methods , PHASE consistently outperforms other existing methods on both simulated and real data, Marchini et al. (2006b). However this algorithm is computationally intensive, preventing analysis of large-scale datasets such as those generated in genome-wide association studies (GWAS). In the rest of this paper, all discussion pertaining PHASE refers to the latest version, PHASE 2.1.1, unless otherwise specified.

Two software packages have been routinely used to analyze large-scale high-density SNP data: FastPHASE proposed by Scheet and Stephens (2006) and Beagle described in Browning and Browning (2009). Both programs employ Hidden Markov Models (HMM's), which can be computed efficiently via the EM algorithm and without lengthy MCMC runs. Like PHASE, FastPHASE also models each haplotypes as a mosaic of other haplotypes. However, whereas PHASE treats every haplotype as a hidden state, FastPHASE groups similar ones into smaller number of local haplotype clusters. The number of hidden states is fixed and constant, and is usually set to a small number (ten to twenty) in practice. Similarly, Beagle uses an HMM that represents local haplotype clusters. While FastPHASE uses a fixed number of clusters across the entire genomic region, Beagle allows the number of hidden states to vary depending on the empirical linkage disequilibrium (LD). Although the Beagle HMM typically has many more hidden states than that in FastPHASE, the computation in Beagle is not hampered because most hidden-state transition and emission probabilities are exactly zero. Owing to the much smaller number of hidden states, both FastPHASE and Beagle substantially improve computational efficiency. The assessment of estimation accuracy has focused on switching error, which is the number of errors between consecutive heterozygous sites. By these criteria, Beagle and FastPHASE achieve similiar accuracy as PHASE. We note that both of types of error measures local accuracy; phasing accuracy at longer distance has not been systematically investigated.

Several phasing algorithms have been developed recently with the primary goals of imputing untyped or missing variants. These methods include MaCH introduced by Li et al. (2010) and IMPUTE2 developed by Howie, Donnelly and Marchini (2009). Since the primary goal of the proposed method is phasing

haplotypes, we restrict our comparison to `FastPHASE` and `Beagle`. An excellent recent review surveys existing phasing methods can be found in Browning and Browning (2011), which compares the performance of a large number of methods.

3. Method

3.1. Model and notations

The basic building block of `emphases` is an HMM that resembles both `PHASE` and `FastPHASE`, in which the hidden state represents the haplotype and the observed state genotype. We reasoned that, as the number of individuals increases, so does the benefit of treating each haplotype as a distinct template (hidden state); at the same time, model parameters, rates of recombination and mutation, can be accurately estimated from data alone without the coalescent priors. Therefore, as in `PHASE`, the HMM in `emphases` considers every haplotype in the sample as a hidden state. Unlike `PHASE`, we eliminated the population genetic model, so that the model parameters can be estimated using an EM algorithm without the MCMC. The key innovations of `emphases` are the two computationally efficient optimization moves, which differ from the standard forward-backward and viterbi algorithms commonly employed in HMM.

We now describe the HMM that underlies `emphases`. Let $G_{jm} \in \{0, 1, 2\}$, ($j = 1, \ldots, N$ and $m = 1, \ldots, M$) be the genotype of individual j at marker m, and let $H = \{H_1, \ldots, H_{2N}\}$ represents the corresponding (unobserved) haplotypes. At each step, the two haplotypes of a single individual, j, are re-estimated; the current estimates of haplotypes in the remaining individuals are treated as templates. To emphasize the difference, the haplotypes to be re-estimated are denoted as B^s_m, where $s = 1, 2$, while the templates are denoted as $H^{(t)}_{-j}$, in which the superscript t indexes iteration and reminds us that the templates change in successive iterations. Since the two haplotypes (B^1, B^2) are excluded from the templates, the total number of templates is always $2N - 2$. As each haplotype in individual j is modeled as a mosaic of haplotypes in $H^{(t)}_{-j}$, it is convenient to introduce a hidden sequence, A_m, which indicates the haplotype index in $H^{(t)}_{-j}$ that acts as the template at marker m. Naturally, there should be two sequences, A^s_m, where $s = 1, 2$ for the two haplotypes; however, the two sequences are modeled as independent Markov Chains and hence the superscript is suppressed when there is no risk of confusion. It is also understood that A_m and B_m refer to the unphased individual, j, in each step.

Focusing on a single haplotype, the transition probabilities of the hidden states, A_m, are:

$$P_j(A^s_m = a \mid A^s_{m-1} = a', \rho) = \begin{cases} (1 - \rho_m) + \frac{\rho_m}{2N-2} & a = a', \\ \frac{\rho_m}{2N-2} & a \neq a'. \end{cases} \tag{3.1}$$

In other words, a jump in the hidden state occurs as a Markov jump process with rate ρ_m; when a jump occurs, the new state is sampled uniformly among all templates, including the original state, a'. Given the hidden state, the observed alleles are specified by

$$P_j(B_m = h \mid A_m = a, H_a = h', \theta_m) = 1_{(h \neq h')}\theta_m + 1_{(h=h')}(1 - \theta_m).$$

Thus, the alleles in B copy the corresponding haplotype template, but with a mismatch probability of θ that represents mutation events. If the genotype at B_m is missing, it can be simply imputed according to the template. Therefore this algorithm can be used for imputing missing genotypes. To initialize the HMM, let $\rho_1 = 1$ and $A_0 = 1$, forcing a jump at $m = 1$.

The objectives of both `PHASE` and `emphases` can be formulated as finding A and H such that the haplotype pairs in H are compatible with the observed genotype, G, while maximizing the log of a pseudo likelihood

$$S(H, \rho, \theta) = \sum_{j=1}^{N} \Big\{ \mathbf{1}_{[B^1 + B^2 = G]} \sum_{s=1}^{2} \log q_j(B^s; \rho, \theta, H_{-j}^{(t)}) \Big\}, \tag{3.2}$$

where

$$q_j(B; \rho, \theta, H_{-j}^{(t)}) = \sum_{A} \prod_{m=1}^{M} P_j(B_m \mid A_m, H_{-j}^{(t)}, \theta_m) P_j(A_m \mid A_{m-1}, \rho_m) \tag{3.3}$$

can be considered as a score associated with each haplotype. In the appendix, we describe the forward algorithm for computing this score, as well as an EM algorithm for estimating parameters ρ and θ. In `PHASE`, ρ_m is assumed to be proportional to the local recombination rate, and both ρ_m and θ_m have priors based on coalescent. Results in Scheet and Stephens (2006) show that, at moderate sample size, eliminating these priors has little effect on the haplotype inference but greatly simplifies the computation.

3.2. Haplotype optimization

In theory, we can obtain haplotype configurations for each individual by drawing haplotypes proportional to the score in (3.3), which is computed by the forward-backward equation (see the appendix). However, this approach is computationally expensive because, at each marker, the number of potential hidden states is on the order of N^2 for a diploid individual (each haplotype can take one of the $2N-2$ states). The computation quickly becomes intractable for moderate numbers of individuals. This is the primary reason that motivated `FastPHASE` to reduce the number of hidden states by forming much smaller number of haplotype clusters. Similarly, in the recently developed imputation method, `MaCH`,

Algorithm 1 `emphases` algorithm

Initialize H with haplotypes estimated using another fast algorithm, such as `FastPHASE` or `Beagle`.

```
Do 15 forward-backward iterations and update parameters, ρ and θ.
for iteration t = 1, ..., T, do
  for each individual j, do
    (B^1, B^2) ← (H^(t)_{2j-1}, H^(t)_{2j})
    Scan each site, accept / reject single site edits on (B^1, B^2) and recombination
    between (B^1, B^2) (Section 3.2);
    Update (H^(t)_{2j-1}, H^(t)_{2j}) ← (B^1, B^2)
  end for
  Update parameters, ρ and θ.
end for
for each individual j, do
  Compute confidence measures at each site (Section 3.3).
end for
```

the number of hidden states is capped by using only a subset of the sample as templates (the program recommends a cap of $200 - 300$.)

`emphases` overcomes the computational challenge in a different way. The algorithm takes as its input a set of initially phased haplotypes, which can be generated by any of the existing phasing algorithms, such as `FastPHASE`, `Beagle` or `MaCH`. These haplotypes are then improved via two types of *local* optimization moves. Each optimization step proposes a new phasing configuration, $\widetilde{B}^1$ and $\widetilde{B}^2$ for individual j, and evaluates the change in the score

$$\Delta(\widetilde{B}^s, B^s) = \sum_{s=1}^{2} \log q_j(\widetilde{B}^s; \rho, \theta, H^{(t)}_{-j}) - \sum_{s=1}^{2} \log q_j(B^s; \rho, \theta, H^{(t)}_{-j}). \tag{3.4}$$

If $\Delta(\widetilde{B}^s, B^s)$ exceeds a pre-defined threshold (set to $\epsilon = 0.3$ in the implementation), the proposal phasing configuration is accepted and the corresponding entries in the template pool, H, are updated with $(\widetilde{B}^1, \widetilde{B}^2)$. Both optimization steps allow re-use of computations, such that after a move is accepted or rejected at one marker, a proposed move can be evaluated at a nearby marker at little additional cost. Further, the score in (3.3) can be computed for the two haplotypes separately at all except a few sites, reducing the computation complexity to the order of NM when optimizing the haplotypes for a single individual. A sweep over all N individuals therefore takes N^2M time (as opposed to using standard forward-backward algorithm, which would take N^3M operations). Below we outline the idea of these optimization steps; computational details are described in the appendix.

The first optimization step is referred to as a *single site edit*, which attempts to switch phasing at a single marker in an individual. Focusing on marker m, the proposed haplotypes, $\widetilde{B}^s$, are allowed to differ from the current haplotypes at site m, provided that the two alleles are compatible with the observed genotype. The score in (3.3) is computed for all possible proposal haplotypes, and the haplotype that achieves the highest score is accepted if $\Delta(\widetilde{B}^s, B^s) > \epsilon$. Specifically, if the genotype G_{jm} is heterozygous, $\widetilde{B}^s_m$ reverses the phasing of B^s_m; if G_{jm} is missing, all possible genotype and phasing configurates are considered. In turn, each site is considered for single site edit, and multiple single site changes may occur sequentially. The haplotype template corresponding to the individual is updated when all sites are scanned. The single site edit is particularly useful to correct the mistakenly imputed and phased genotypes.

The second optimization step introduces a crossover (recombination) between the two current haplotypes at a specific site. At a given marker, $m^\star$, let

$$\widetilde{B}^1_m = \begin{cases} B^1_m & m \le m^\star, \\ B^2_m & m > m^\star. \end{cases} \tag{3.5}$$

The complementary haplotype is defined as $\widetilde{B}^2_m$. In turn, each marker is considered as the potential site of recombination; the site that achieved the greatest improvement is used to update H, as described above. Again, because the haplotypes are unchanged at all except one site, the computation can be achieved in NM operations. The crossover move is designed to correct the most common phasing mistake, a switch error.

In each iteration, the algorithm optimizes the haplotypes of each individual in turn using the single site edit and crossover moves, and updates the corresponding templates in H. Once all individuals have been updated, the model parameters, ρ and θ are re-estimated using the EM-algorithm. The process is repeated until covergence or for a pre-specified number of iterations.

3.3. Measuring haplotype confidence

Most phasing algorithms assign a likely configuration at each marker, yet the uncertainties associated with these estimates are not the same across markers. LD pattern varies tremendously across the genome, as a consequence of the variation in fine-scale recombination rate and the stochasticity through the genealogical process. In regions where LD is weak, phasing may be unreliable and impact subsequent analysis. Therefore, it is desirable to produce confidence measures in addition to the inferred haplotypes.

The two optimization steps described above offer two natural measures of phasing uncertainties. Given the haplotype estimated using `emphases`, we define

a single site phasing confidence for individual j at marker m, χ^{SS}_{jm}, as the maximal increase in (3.4) that can be achieved by a single site edit at m. When the genotype G_{jm} is homozygous, phasing is unambiguous and no confidence measure is produced. Likewise, we define the confidence of phasing between sites m and m' as the maximal increase in the score that can be achieved by a crossover move at a site between m and m'. Thus, let Δ^{CR}_{jk} be the the value of (3.4) achieved by a recombining the two current haplotypes at site k; the pairwise confidence between sites m and m' is

$$\chi^{CR}_{jmm'} := \max_{m \le k < m'} \Delta^{CR}_{ji}. \tag{3.6}$$

The two confidence measures are complementary. Negative values of large magnitude in both χ^{SS}_{jm} and $\chi^{SS}_{jmm'}$ indicate that the estimated phasing configuration in the region is substantially better than alternative configurations. On the other hand, if either or both of the measures are close to zero or positive, the estimated haplotypes are associated with high uncertainties in the sense an alternative phasing configuration is almost equally likely. In data analysis presented in the next section, we find that these confidence measures provide useful information in identifying regions of high uncertainties.

4. Numerical Examples and Results

Description of data. We use parent-offspring trio datasets to assess the performance of the proposed method. The parent-child family structure permits unambiguous determination of the haplotypes in the children at most markers, except triple heterozygous sites (i.e. both parents and the child are heterozygous) and sites where one or more members have missing genotypes; these two latter scenarios are collectively referred to as ambiguous sites henceforth. Neglecting the rare recombination events in the meiosis of the parents, the haplotypes in a child allows the reconstruction of the four parental haplotypes. Haplotypes thus constructed are treated as the gold standard; ambiguous sites are excluded from computing the error rate. The first dataset consists of 45 CEU (European Americans from Utah) trios and 50 YRI (Yoruba population from Nigeria) trios, genotyped on the Illumina HumanHap650Y arrays as part of the International HapMap Project, Phase III (The Interational HapMap Consortium (2007)). Specifically, we focus on chromosome 9, which includes $23,814$ SNPs. As a gold standard, we use the phased haplotype data generated by the HapMap consortium, excluding ambiguous sites. In all analyses, the CEU and YRI trios were analyzed separately. The second dataset includes 492 Mexican trios (MEX), genotyped as part of a GWAS of asthma. The recruitment and sample characteristics were described previously (Hancock et al., 2009). This dataset is useful for two reasons.

First, it allows us to assess the performance of various phasing methods in an genetically admixed population, a subject that has not been systematically studied. Second, as we illustrate below, the accuracy of haplotype inference depends on the sample size; therefore, as one of the largest trio-sample that has been genotyped to date (984 parents), analyzing this data provides a more realistic measure of the phasing accuracy expected in a GWAS study.

In applying all phasing algorithms, genotypes of the parents but not the children are used as input data; thus these analyses are comparable to analyzing unrelated individuals. `FastPHASE` was run with 20 hidden states, 10 restarts, and 35 EM iterations per restart; the number of hidden states doubles the default parameters and gives a lower error rate. Default setting was used for `Beagle`. For `emphases`, the output from either `FastPHASE` or `Beagle` were used as initial values, with 15 initial iterations of parameters update, followed by 8 iterations of haplotype optimization and parameter updates, as outlined in Algorithm 1.

Measurement of accuracy. Most previous studies comparing haplotype inference methods have focused on the rate of switching error, which is defined as the error probability between consecutive pairs of heterozygous sites. The switching error rate meaningfully measures the short-range phasing accuracy. However some analyses, such as the extended haplotype homozygosity (EHH) test used for detecting recent positive selection, depend on accurate phasing over a longer range, Sabeti et al. (2002). In these settings, the relevant quantity is the error rate between non-consecutive pairs of SNPs. As this error rate is expected to increase for markers further apart, we characterize the error rate as a function of distance. In what follows, SNPs are indexed according to their physical positions, $1, 2, \ldots, M$, and the distance between SNPs i and j is defined as $|i-j|$. This metric is sensible because none of the methods considered here, `FastPHASE`, `Beagle`, and `emphases`, incorporates either physical or recombination distance. The error rate at distance, d, is then computed over all pairs of markers that are d SNPs apart.

Phasing accuracy across populations. This first set of analyses aims to compare the phasing accuracies of `FastPHASE`, `Beagle`, and `emphases`, evaluating error rates between consecutive and non-consecutive SNPs, using trio data from the populations CEU, YRI and MEX. The sample sizes are 88, 100, and 100 individuals for the three populations, respectively. Table 1 compares the error rates between consecutive heterozygous sites. We observe that `Beagle` is slightly more accurate compared to `FastPHASE` on CEU and YRI and slighted less accurate on MEX; `emphases` reduces these errors by $10-40\%$, achieving an accuracy comparable to previous analysis using `PHASE` (Marchini et al. (2006a)). Although the terminal error rates of `emphases` are similar using either `Beagle` or `FastPHASE` output as initial values, we note that `emphases` does not perform well using a completely random phasing configuration, and therefore should not be used as a

Table 1. Switch error rates (%) on trio datasets representing three diverse populations. CEU=HapMap individuals of European ancestry from Utah; YRI: HapMap Yoruba individuals of African ancestry, from Nigeria; MEX=Mexican individuals from Mexico City. *Sample sizes refers to the number of unrelated individuals used for analysis; †`emphases` using `Beagle` output as initial phasing configuration; ‡`emphases` using `FastPHASE` output as initial phasing configuration. `FastPHASE` was run with 20 hidden states, 10 restarts, and 35 EM iterations per restart. For `Beagle`, default setting was used.

Population	N^*	`Beagle`	`Beagle`+`emphases`†	`FastPHASE`	`FastPHASE`+`emphases`‡
CEU	88	6.58	4.76	5.36	4.57
YRI	100	9.72	5.75	7.29	5.26
MEX	100	5.68	4.10	4.30	3.82

stand-alone method. For all methods, the switch error rates vary across the three populations analyzed, with the highest error in YRI and the lowest in MEX. We hypothesize that the error rate is highest in YRI because LD is lowest in the African population. The Mexican individuals derive substantial ancestry from the Native American ancestral populations (Johnson et al., 2011), which exhibit the highest LD because of historical population bottlenecks (Jakobsson et al., 2008). It is also interesting to observe that the improvement due to `emphases` is more substantial in YRI compared to CEU and MEX, and particularly compared to `FastPHASE`. These observations are consistent with the intuition that grouping haplotypes into a small number of clusters increases phasing error, and that the negative impact is especially strong for populations with greater haplotype diversity (YRI).

Figure 1 displays the error rates for non-consecutive pairs of heterozygous sites as a function of inter-marker distance. For clarity, this figure shows the error rates using `FastPHASE` and `emphases` with `FastPHASE` as initial values. The corresponding results, comparing `Beagle` and `emphases` with `Beagle` initialization are shown in Supplemental Figure 1 and are qualitatively similar. As expected, the error rates for all methods increase with inter-marker distance, with the limit approaching the random guess error rate of 50%. On the CEU trio data, the error rate incurred by `FastPHASE` (dotted red line) rises from 1.57% to 5.48% to 34.5% at distances of $1, 5$ and 50 markers, respectively, corresponding to $2,600, 17,900$, and $204,800$ bp. Applying `emphases` to the output of `FastPHASE`, phasing accuracy improves regardless of marker distance; the improvement of `emphases` is most prominent for moderately spaced markers (19.9% reduction at ~ 40 SNPs apart). Analysis of the HapMap YRI dataset yields qualitatively similar results; however, `emphases` achieves a greater error reduction on this set: 20% for consecutive SNP pairs, and $> 30\%$ for pairs 30 markers apart. Again, despite somewhat

different accuracy between `Beagle` and `FastPHASE` we find that `emphases` produces essentially identical error rates using the output of either program as initial values for all three populations and at all marker distances considered. Curiously, Figure 1 suggests that for short range, phasing error is highest in YRI, but at a distance of greater than 40 SNPs the phasing accuracy is the highest in YRI. We postulate that the reason lies in the genetic diversity and structure of the HapMap YRI sample. The YRI sample represents Yoruba individuals from Nigeria; previous studies have found that this population has greater genetic diversity, and the linkage disequilibrium (LD) between markers is generally weaker in the African populations than in non-African populations (The Interational HapMap Consortium (2007)). As a result, phasing is generally more challenging in an African population than in other populations, consistent with the highest phasing error rate in short ranges for all methods considered. For markers further apart, reduced LD leads to higher error rates in all three populations. However, whereas the parents in CEU and MEX datasets are "unrelated," cryptic relationships, such as uncle-niece and sib-pairs, have been reported in the HapMap YRI individuals (Pemberton et al. (2010)). Exploiting the shared haplotypes between the related individuals enables `emphases` to achieve a greater improvement over `Beagle` and `FastPHASE` in YRI sample than it does in CEU and MEX samples.

Phasing accuracy as a function of sample sizes. The second set of analyses aims to understand the impact of sample size. Subsets of 100, 400, and 800 individuals were randomly sampled from a total of 984 parents of the Mexican trios; these samples, as well as the complete sample were analyzed using `FastPHASE`, `Beagle` and `emphases`. The switch error rates shown in Table 2 suggest that the accuracy of both `Beagle` and `emphases` (with either `Beagle` or `FastPHASE` as initial values) improve with an increased sample size; in fact, the error rate of `emphases` is reduced by a factor of almost two when the sample size increases from 100 to the full set of 984. A greater sample size not only improves phasing accuracy at closely located SNPs, but also benefits long range phasing. Figure 2 plots pairwise phasing error rate as a function of SNP distance for sample sizes of 100 and 984, suggesting that, for both `Beagle` and `emphases`, an increased sample size can substantially reduce phasing error even for markers that are far away. For both small and large sample sizes, `emphases` achieves a lower error rate compared to `Beagle`, and the improvement is substantial for intermediate distance ($\sim$ 100 SNPs apart). A full description of the error rate for intermediate sample sizes can be found in Supplemental Figure 2, which also shows the error rates using `FastPHASE` and `emphases` with `FastPHASE` initialization. For all sample sizes and inter-marker distance considered, `emphases` consistently achieves the lowest error rates. In contrast to `Beagle` and `emphases`, whose performance greatly improves with an increased sample size, the error rates of `FastPHASE`

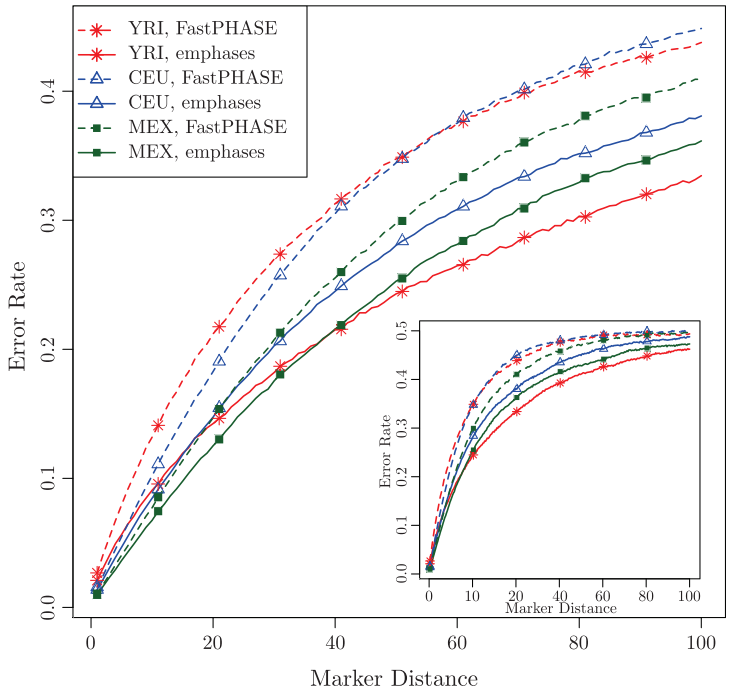


Figure 1. Phasing error rate as a function of inter-marker distance. Error rate are evaluated as the proportion of pairs of heterozygous sites; marker distance are measured by the numbers of markers between such pair of SNPs, but the intervening SNPs may not all be heterozygous. Other experimental settings are identical to that in Table 1.

changes little with sample size (Supplemental Figure 3). We attribute this pattern to the fixed haplotype clusters used. This observation is also consistent with the findings in Browning and Browning (2011) that `FastPHASE` can outperform `Beagle` for a small sample size, but the trend reverses as the sample size increases. Nonetheless, `emphases` achieves similar error rate using either `Beagle` or `FastPHASE` as initial values. In terms of computational burden, the CPU requirement for `emphases` increases quadratically with the sample size for fixed number of markers, and increases linearly with the number of markers for fixed sample size; the memory requirement for `emphases` is proportional to (markers $\times$ samples). On an Intel 3GHz processor, the computational times required for analyzing the complete MEX data (984 individuals and 23,814 markers) are 104.6, 1.15 and 8.85 hours for `FastPHASE`, `Beagle` and `emphases` (with eight optimization iterations), respectively. The memory required for `emphases` is approximately 1GB.

Phasing accuracy as a function of confidence measure. The confidence measures described in the previous section are based on the likelihood ratio of the cur-

Table 2. Switch error rate (%) for Mexican trio data (MEX). Subsets of individuals (N) were randomly sampled. Other settings are described in the caption of Table 1.

N	`Beagle`	`Beagle`+EM	`FastPHASE`	`FastPHASE`+EM
100	5.68	4.10	4.30	3.82
200	4.61	3.33	3.92	3.20
400	3.75	2.67	3.74	2.66
800	3.12	2.10	3.65	2.18
984	2.94	1.97	3.61	2.09

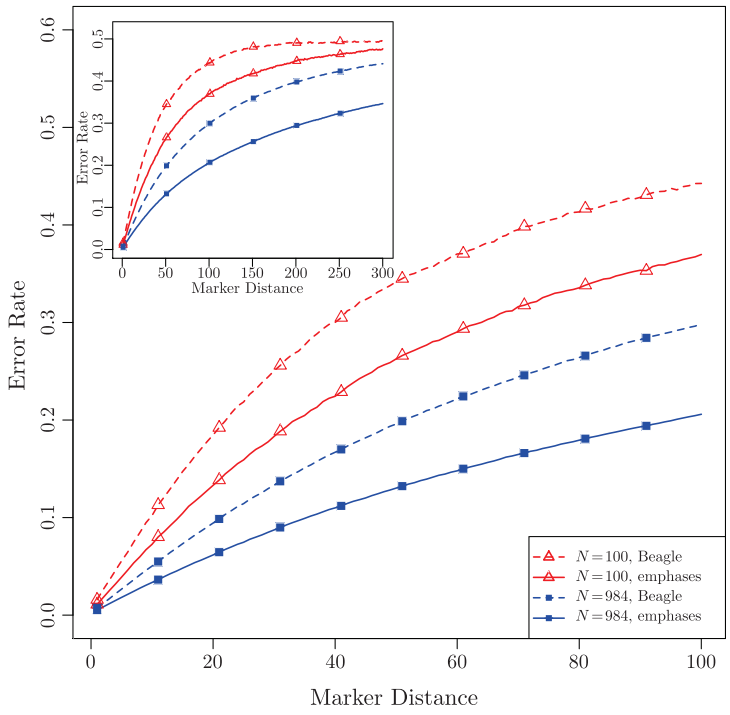
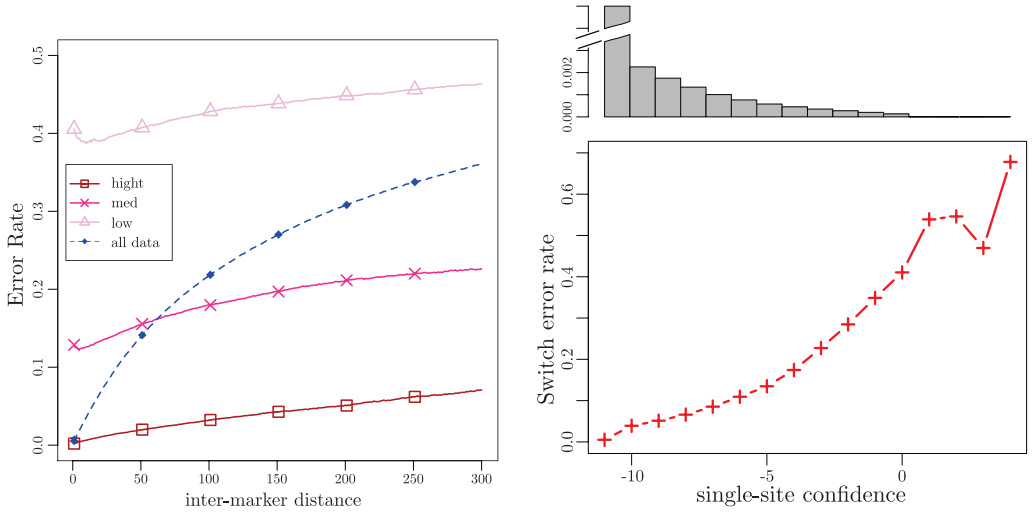


Figure 2. Phasing error rate as a function of SNP distance at various subsample-sizes of the Mexican panel. The `FastPHASE` error rate does not decrease appreciably with the sample size as both `Beagle` and `emphases` do. On the full panel of 984 parents the `emphases` switch error rate is 72% that of `Beagle` and 56% that of `FastPHASE` (regardless of which method is used as initialization).

rent phasing configuration to the next best alternative configuration that can be achieved by a single site edit (χ^{SS}_{jm}) or a single recombination between the two haplotypes ($\chi^{CR}_{jmm'}$). To assess the informativeness of the crossover confidence, we evaluated the empirical error rates for MEX data, stratifying pairs of markers by the crossover confidence, $\chi^{CR}_{jmm'}$: $(-\infty, -5]$, $(-5, -1]$ and $(-1, \infty)$,



(a) Cross-over confidence (b) Single-site confidence

Figure 3. Phasing error rate for the full MEX data using `emphases` with `Beagle` as initialization. (A): error rate (y-axis) between pairs of markers are stratified by cross-over confidence measure, and plotted against the inter-marker distance (x-axis). High conf: $\chi^{CR} \in (-\infty, -5]$; med conf: $\chi^{CR} \in (-5, -1]$; low conf. $\chi^{CR}_{jmm'} \in (-1, \infty)$. While the error rate increases with inter-marker distance in all strata, the empirical error rate differ appreciably depending on χ^{CR}, for any fixed inter-marker distance. (B): switch error (y-axis) between consecutive heterozygous sites increases as the maximum of the single-site confidence score at the two sites, $\max(\chi^{SS}_m, \chi^{SS}_{m'})$. The histogram indicates the distribution of single-site confidence scores, with $> 90\%$ of sites having a score less than 10 (broken bar).

corresponding to high, medium and low confidence, respectively. We found that the phasing error rate shows a stronger dependency on the confidence measure than on marker distance. In other words, the error rate is lower for distant markers with high confidence scores than for proximal markers with low confidence scores (Figure 3a). Of course, the number of high-confidence pairs diminishes as the distance grows: in the full MEX data, half of the SNP-pairs can be phased with high crossover confidence at a distance of 37 SNPs. Thus, the confidence scores identify haplotypes with high uncertainties, and should be down-weighted or excluded in subsequent haplotype-based analyses. Similarly, we examined the switch error rate between consecutive heterozygous sites as a function of single-site confidence measure. Figure 3(b) suggests that the probability of phasing error increases monotonically as the worse of the two single-site confidence scores increases, $\max(\chi^{SS}_m, \chi^{SS}_{m'})$. We note that both single-site and crossover confidence scores may exceed 0, because these scores depend on haplotype templates; thus

an update in one individual affects the scores in other individuals. However, this occurs rarely after the first few iterations: in the Mexican data analysis, less than 0.1% of the sites have $\chi^{SS} > 0$ after 8 iterations.

5. Discussion

The ability to accurately construct haplotypes from unphased genotype data plays essential roles in population and medical genetics research. Previous studies have largely focused on accuracy over short range, such as the error rate between consecutive heterozygous sites (switching error), yet many haplotype-based analyses assume accurate phasing over a long range (Sabeti et al. (2002)). Results on the data presented in Section 4 indicate that phasing error rate increases rapidly as a function of marker spacing, and underscores the need for methods that improve long range haplotype phasing in unrelated individuals. The LRP method proposed in Kong et al. (2008) has shown promising long range phasing performance in a large Icelandic cohort; however, this method relies on a substantial number of pairs of individuals in the sample sharing haplotypes spanning hundreds or thousands of markers. This requires that either the underlying population is inbred, or a large fraction of the population is included in the sample. We examined the allele sharing in 10 regions of 1000 SNPs in the HapMap CEU and YRI parents, as well as the 984 Mexican parents, and found no pairs of individual sharing a haplotype at this length. Therefore, it is unlikely that LRP can be directly applied to GWAS data collected in non-founder populations.

Here we propose a new computational approach, implemented in the program `emphases`, with the primary goal to improve the long range phasing accuracy. The model underlying `emphases` retains a key feature of PHASE: both adopt an HMM in which every haplotype in the sample is treated as a hidden state. There is a connection between this idea and the heuristic notion of "surrogate" parents that underlies LRP: the hidden state can be thought as the best surrogate parent, and the objective in `emphases`, which favors staying in a hidden state for long segments, achieves a similar effect as LRP that identifies surrogate parents based on long range allele sharing. In other words, the HMM in `emphases` and `PHASE` seek the most-likely surrogate parents in the sample. Seen this way, `PHASE` and `emphases` can be thought of as a generalization of LRP. By implementing new optimization algorithms and making use of an initial phasing configuration produced by existing phasing algorithms, `emphases` overcomes the computational challenge that limits the application of `PHASE` to large datasets. Examples presented in Section 4 demonstrates that, compared to the two widely used phasing algorithms, `FastPHASE` and `Beagle`, `emphases` improves phasing accuracy, both in terms of switching error rate between consecutive heterozygous sites and error rate between markers that are far apart. We attribute the improved performance

to the much larger number of hidden states used in `emphases` than in `FastPHASE` and `Beagle`, both of which reduce the number of hidden states by grouping haplotypes into clusters. An R package implementing `emphases`is available at `http://med.stanford.edu/tanglab/software/emphases.latest.zip`

Our analysis of the CEU, YRI and Mexican trio data provide some insights that may benefit future development and applications of phasing algorithm. First, while `emphases` improves long range phasing accuracy over `Beagle` and `emphases`, the error rate between distant markers can still be high. We propose two confidence scores to measure phasing uncertainties, which should be incorporated in subsequent analyses. Second, the analysis of Mexican data at varying sample size indicates that the performance of `emphases` improves with the increased sample size. This suggests that `emphases` has the potential to substantially improve phasing in large samples, such as those generated in GWAS. Furthermore, with the rapid accumulation of genotype and sequencing data from GWAS, further improvement in phasing accuracy maybe achieved by combining datasets representing closely related populations. Thus, computationally efficient algorithms that take advantage of these rich resources should be explored further.

Acknowledgement

The research was supported by NIGMS grant GM073059 to HT and by the Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, DHHS, to SJL. NAJ is supported by the Stanford Genome Training Program (T32 HG000044) and a Ric Weiland fellowship.

Appendix

Forward-backward algorithm and parameter estimation.

Here we derive the forward algorithm for computing the score in (3.3), $q_j(B^s;\rho,\theta,H^{(t)}_{-j})$. In what follows, we use shorthand notations $\mathcal{T}(a^* \to a)$ for transition probability from hidden state a^* to a, and $\mathcal{E}(B_m)$ to denote the emission probability of observing an allele given the haplotype template $(H^{(t)}_{-j})$, hidden state (A_m), and mismatch probability (θ_m). As explained in Section 3.2, the optimization steps compute the score of a proposed phasing configuration that assumes that phasing is known except at the candidate marker of single-site edit or the candidate marker of a cross-over modification. Therefore, the score $q_j(B;\rho,\theta,H^{(t)}_{-j})$ can be computed as the product of the two haplotypes separately. To compute $q_j(B^1;\rho,\theta,H^{(t)}_{-j})$, define the forward messages as $\overrightarrow{\nu}_m(k) = P(B^1_1,\ldots,B^1_m,A_m=k;\rho_m,\theta,H^{(t)}_{-j})$ following Rabiner (1989). The initiation condition is set as $\overrightarrow{\nu}_0(k) = 1/(2N-2)$ and $\rho_0 = 1$. In standard HMM, the induction

of $\overrightarrow{\nu}_{m+1}(a)$ for all a and m requires $O(N^2M)$ operations. In our setting, however, the transition probability specified in (3.1) means that the transition probability depends only on whether $a^* = a$, and not on the specific values of a and a^*. Therefore, $\overrightarrow{\nu}_m$ can be computed in $O(N)$ operation as

$$\begin{aligned}\overrightarrow{\nu}_{m+1}(a) &= P(B_1^1, \ldots, B_{m+1}^1, A_{m+1} = a) \\ &= \sum_{a^*=1}^{2N-2} P(B_1^1, \ldots, B_m^1, A_m = a^*)\mathcal{T}(a^* \to a)\mathcal{E}(B_{m+1}^1) \\ &= \mathcal{E}(B_{m+1}^1)\Big[\frac{\rho_m}{2N-2}\sum_{a^*=1}^{2N-2}\overrightarrow{\nu}_m(a^*) + (1-\rho_m)\overrightarrow{\nu}_m(a)\Big],\end{aligned}$$

where $\sum_{a^*=1}^{2N-2}\overrightarrow{\nu}_m(a^*)$ is a constant that only needs to be calculated once for each m. The backward messages is defined similarly as $\overleftarrow{\mu}_m(a) = P(B_m^1, \ldots, B_M^1, A_m = a)$, with $\overleftarrow{\mu}_M(a) = 1$ for all hidden states, a. Since the Markov chain is reversible, the backward messages can be computed in the same way as the forward messages by reversing the sequence (Lai, Xing, and Zhang (2008)). Together, the variables $\overrightarrow{\nu}$ and $\overleftarrow{\mu}$ allow us to compute the posterior, $P(A_m \mid B; H_{-j}^{(t)}, \rho_m, \theta_m)$.

The parameters ρ_m and θ_m are estimated using the EM algorithm. Given the hidden sequences of the two haplotypes in individual j, A^{js}, the EM estimates for ρ_m amounts to counting the number of jumps between markers m and $m+1$; likewise, θ_m is estimated by the fraction of mismatches at the sites. We note that, by Jensen's inequality, the parameter updates increase the score in (3.2) and optimize a minorizing Q-function, even though (3.2) is not a true likelihood.

Haplotype optimization.

We describe the single-site and the cross-over optimization steps, both require evaluating (3.3) for proposed haplotypes, $\widetilde{B}^s$. For single-site edits, as explained in Section 3.2, $\widetilde{B}^s$ differs from B^s at a single marker, m. Therefore, the score of $\widetilde{B}^s$ matches that of B^s in the interval of $[1, \ldots, m-1]$ and $[m+1, \ldots, M]$, and can be computed by stitching the scores corresponding to the two flanking segments, while summing over possible hidden states at site m :

$$q_{jm}(\widetilde{B}^s) = \sum_{a^*}\overrightarrow{\nu}_{m-1}(A_{m-1})\overleftarrow{\mu}_{m+1}(A_{m+1})\mathcal{T}(A_{m-1} \to a^*)\mathcal{T}(a^* \to A_{m+1})\mathcal{E}(\widetilde{B}_m^s).$$

This score is computed for single-site edited haplotypes, $(\widetilde{B}^1, \widetilde{B}^2)$, which are compatible to the observed genotype. The phasing configuration that achieves the maximum score is then compared with the current haplotypes to determine whether an edit is desirable.

The computation for the cross-over optimization is analogous. Consider the proposed haplotype formed by a cross-over at site m, $\widetilde{B}^1 = (B_1^1, \ldots, B_m^1,$

B^2_{m+1}, B^2_M). The pseudo likelihood can be computed by multiplying the forward and backward messages while summing across possible states at the seam, $\sum_{a^*} \overrightarrow{\nu}_m(a^*)\overleftarrow{\mu}_{m+1}(a^*)$. In implementing the cross-over moves, we scan each site sequentially for potential site of recombination, starting from one end of the chromosome. Rather than taking the first site at which a cross-over achieves higher score, we instead search for a local maximum before performing the recombination.

References

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Amer. J. Human Genetics* **84**, 210-223.

Browning, S. R. and B. L. Browning (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703-714.

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

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplo- type frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921-927.

Fan, H. C., Wang, J., Potanina, A. and Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51-57.

Hancock, D. B., Romieu, I., Shi, M., Sienra-Monge, J. J., Wu, H., Chiu, G. Y., Li, H., del Rio-Navarro, B. E., Willis-Owen, S. A., Weiss, S. T., Raby, B. A., Gao, H., Eng, C., Chapela, R., Burchard, E. G., Tang, H., Sullivan, P. F. and London, S. J. (2009). Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children. *PLoS Genet.* **5**, e1000623.

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

Higasa, K., Kukita, Y., Kato, K., Wake, N., Tahira, T. and Hayashi, K. (2009). Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its significance for the analyses of positively selected regions. *PLoS Genet* **5**, e1000468.

Howie, B. N., Donnelly, P. and Marchini, J. (2009). A exible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A. and Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003.

Johnson, N. A., Coram, M. A., Shriver, M. D., Romieu, I., Barsh, G. S., London, S. J. and Tang, H. (2011). Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet.* **7**, e1002410.

Kitzman, J. O., Mackenzie, A. P. Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E. and Shendure, J. (2011). Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59-63.

Kong, A. Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H. and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**, 1068-1075.

Lai, T. L., Xing, H. and Zhang, N. (2008). Stochastic segmentation models for arraybased comparative genomic hybridization data analysis. *Biostatistics* **9**, 290-307.

Lawson, D. J., Hellenthal, G., Myers, S. and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453.

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816-834.

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

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R. and Donnelly, P. (2006a). A comparison of phasing algorithms for trios and unrelated individuals. *Amer. J. Hum. Genet.* **78**, 437-450.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., Donnelly, P. and Consortium, I. H. (2006b). A comparison of phasing algorithms for trios and unrelated individuals. *Amer. J. Hum. Genet.* **78**, 437-450.

Myers, S., Bottolo L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324.

Niu, T., Qin, Z. S., Xu, X., Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Amer. J. Hum. Genet.* **70**, 157-169.

Pemberton, T. J., Wang, C., Li, J. Z., Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Amer. J. Hum. Genet.* **87**, 457-464.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pp. 257-286.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P. Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for largescale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Amer. J. Hum. Genet.* **78**, 629-644.

Stephens, M. and Scheet, P. (2005). Accounting for Decay of Linkage Disequi- librium in Haplotype Inference and Missing-Data Imputation. *Amer. J. Hum. Genet.* **76**, 449-462.

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

The Interational HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861.

Yang, H., Chen, X. and Wong, W. H. (2011). Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12-17.

Google Mountain View, 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA.

Department of Statistics, Stanford University, Stanford, CA 94305 and Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA.

E-mail: nickaj@gmail.com

Epidemiology Branch and Laboratory of Respiratory Biology, National Institute of Environmental Health Sciences, National Institutes of Health, U.S.

Department of Health and Human Services, Research Triangle Park, NC 27709 USA.

E-mail: london2@niehs.nih.gov

National Institute of Public Health, Cuernavaca, Mexico, and International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France.

E-mail: iromieu@gmail.com

Department of Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: whwong@stanford.edu

Department of Genetics, 300 Pasteur Drive, Stanford University, Stanford, CA 94305, USA.

E-mail: huatang@stanford.edu

(Received May 2012; accepted March 2013)