STUDY DESIGNS

Haplotype phasing: existing methods and new developments

Sharon R. Browning and Brian L. Browning‡*

Abstract | Determination of haplotype phase is becoming increasingly important as we enter the era of large-scale sequencing because many of its applications, such as imputing low-frequency variants and characterizing the relationship between genetic variation and disease susceptibility, are particularly relevant to sequence data. Haplotype phase can be generated through laboratory-based experimental methods, or it can be estimated using computational approaches. We assess the haplotype phasing methods that are available, focusing in particular on statistical methods, and we discuss the practical aspects of their application. We also describe recent developments that may transform this field, particularly the use of identity-by-descent for computational phasing.

Imputation
In the context of this article, this is the estimation of missing genotype values by using the genotypes at nearby SNPs and the haplotype frequencies seen in other individuals.

Calling genotypes
Estimating genotype values from raw data. Genotyping technology provides information about the underlying genotype, typically in the form of signal intensities or read counts of the two alleles. Statistical techniques are used to resolve this information into genotype calls. Typically, information across individuals is used, and correlation across SNPs (that is, haplotype phase) is also helpful.

**Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA.
‡Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington 98195, USA.
e-mails: sguy@uw.edu; browning@uw.edu*

doi:10.1038/nrg3054

With recent technological advances, enormous amounts of genotype data are being generated, both from increasingly comprehensive and inexpensive genome-wide SNP microarrays and from ever more affordable whole-genome and whole-exome sequencing tools. However, the vast amount of information in these data is best exploited through phased haplotypes, which identify the alleles that are co-located on the same chromosome. Because sequence and SNP array data generally take the form of unphased genotypes, it is not directly observed which of the two parental chromosomes, or haplotypes, a particular allele falls on. Fortunately, new advances in computational and laboratory methods promise improved determination of haplotype phase.

Methods for haplotype phasing have developed in response to improvements in technology that have changed the scale of genetic data. At first, genetic studies would typically assay only a single variant, and hence haplotype phase was irrelevant. As candidate gene sequencing became more accessible in the late 1980s, methods were developed for computational and experimental phasing of short regions containing a small number of genotyped polymorphisms. With the advent of genome-wide SNP microarrays and genome-wide association studies in around 2005, new computational methods were developed to handle whole-chromosome data efficiently. Laboratory-based methods for experimental phasing of whole-genome sequence data have also recently been developed.

The importance of haplotype phase information is increasing as we move into the era of large-scale sequencing. Applications of haplotype phase include understanding the interplay of genetic variation and disease[1], imputation of untyped genetic variation[2–4], calling genotypes in microarray and sequence data[5–10], detecting genotype error[11], inferring human demographic history[12], inferring points of recombination[13], detecting recurrent mutation[13] and signatures of selection[14] and modelling *cis*-regulation of gene expression[15].

In this Review, we cover the historical and recent developments in methods for computational phasing of genotypes from population data sets and family data sets, and in experimental methods for phasing single individuals. This Review mostly focuses on computational methods, both because the authors' expertise is in this area and because experimental methods are not yet cost-effective for large-scale use. We examine the strengths and weaknesses of the best currently existing methods and consider a few examples of their use. Finally, we discuss recent developments and current challenges in phasing methodology.

Computational haplotype phasing

Computational methods pool information across individuals to estimate haplotype phase from genotype data. Unrelated individuals can be phased by considering sets of common haplotypes that can explain the observed genotype data. The number of unrelated individuals present in a sample is a crucial factor in determining how well the phase can be estimated: the more individuals, the better the estimation. Related individuals, by contrast, can be phased by considering haplotypes that are shared identical-by-descent between individuals within families. This within-family information on identity-by-descent (IBD)

Unphased genotypes	Possible phasing A		Possible phasing B		Possible phasing C		Possible phasing D	
A/C	A	C	A	C	A	C	A	C
G/T	G	T	G	T	T	G	T	G
A/T	A	T	T	A	A	T	T	A
Population haplotype frequency	55%	0%	15%	5%	2%	3%	0%	20%
Population frequency of unordered haplotype pair	0%		2 × (15% × 5%) = 1.5%		2 × (2% × 3%) = 0.12%		0%	
Posterior probability of unordered haplotype pair	0%		1.5% / (1.5% + 0.12%) = 93%		0.12% / (1.5% + 0.12%) = 7%		0%	

Figure 1 | **Statistical phasing of unrelated individuals using haplotype frequencies.** Consider one individual with a heterozygous genotype at each of three SNPs in a region. There are four possible haplotype configurations that are consistent with the genotype data (possible phasing patterns A–D). Suppose that haplotype frequencies are available from other individuals in the population at these sites (provided below each phasing pattern). These frequencies may have been estimated from population data without additional modelling (with the *a priori* assumption that all haplotype frequency configurations are equally likely) or from a model that accounts for the biological processes of recombination and mutation (such as the Li and Stephens model[30]). The population frequency of a haplotype pair is obtained using the Hardy–Weinberg principle (independence of the two haplotypes within an individual); the factor of two in the frequency of the haplotype pairs accounts for both possible assignments of maternal and paternal origin to the two haplotypes. The posterior probabilities of the phased data are obtained from the population frequencies of the possible haplotype pairs. In this example, the posterior probability of phasing B (93%) is much greater than that of phasing C (7%).

Identical-by-descent
Two haplotypes are identical-by-descent if they are identical copies of a haplotype inherited from a common ancestor.

Cryptic relatedness
The undocumented existence of relatives within a sample.

Posterior distribution
Probabilities that account for the prior information and the information in the data. For haplotype phase estimation, the posterior distribution accounts for all available information, including the genotypes and the estimated haplotype frequencies in the population.

Expectation maximization algorithm
(EM algorithm). An iterative approach for finding the values of the unobserved data (such as haplotype phase) that maximize the statistical likelihood of the observed incomplete data. Although the likelihood increases with each iteration, the approach is not guaranteed to find the global maximum.

Partition–ligation
A divide-and-conquer strategy that is designed to reduce the computational burden for phasing methods that do not scale well with increasing region size. A large region is divided up into smaller regions, and haplotype phase estimates from the smaller regions are used to limit the possibilities when phasing the large region.

is much more informative for phase estimation than the haplotype frequency information used to phase unrelated individuals; however, haplotype frequency information across families or from the population can also be used to fill in the gaps in haplotype phase that are not determined by IBD sharing within families. Also, with unrelated individuals, some cryptic relatedness will exist that can be exploited with an IBD-sharing approach. Thus, computational phasing of related individuals and computational phasing of unrelated individuals are not completely separate problems.

Computational cost is an important factor when considering which computational phasing method to use. Generally, there is a choice of algorithms and algorithm parameters, and the researcher must select a trade-off between haplotype phase accuracy and computational cost. For large data sets of unrelated individuals, a method is needed that scales well with both the number of markers and the number of individuals. For family data, a method is needed that can handle the maximum family size present in the data (many methods scale exponentially with family size) and that also scales well with number of markers.

Computational phasing in unrelated individuals

Statistical approaches to phasing unrelated individuals rely on the modelling of haplotype frequencies. In situations in which several haplotype configurations are possible for an individual's genotypes, one can estimate, through statistical modelling of the data, the probability of any given haplotype configuration (FIG. 1) and either pick the most likely configuration or output a set of configurations that are sampled from the posterior distribution. Other computational approaches, such as parsimony[16,17] and long-range phasing[13] (discussed below), are rule-based methods; they do not directly model haplotype frequencies but are based on the assumption that the most likely haplotype configurations are those that are seen in other individuals.

Our description focuses on those methods that are most widely used or most historically important. We present the methods in approximate chronological order. There are many other computational phasing methods in use that are described elsewhere[18].

Clark's algorithm. Clark's algorithm[19] was the first published method for haplotype phase inference for three or more markers in unrelated individuals. The method is based on using unambiguous haplotypes (from individuals with at most one heterozygous marker) and parsimony (finding solutions that use the least number of unique haplotypes). The algorithm is only suitable for tightly linked polymorphisms. When polymorphisms are not tightly linked, there may be several reasonable haplotype phase assignments that correspond to an individual's genotype, and the method does not provide a means of choosing between such assignments. Clark anticipated the next significant advance in phasing methodology by observing that the expectation maximization algorithm (EM algorithm) could be used to phase small numbers of polymorphisms that are not tightly linked[19].

EM algorithm. An early application of the EM algorithm[20] to the haplotype phasing problem[21–23] involved treating all possible haplotype configurations as equally likely *a priori*. This phasing method is typically referred to as 'the EM algorithm', even though many other statistical phasing methods also use an EM approach as a part of their algorithms. The basic EM algorithm works well for a small number of genetic polymorphisms (up to around ten), but it quickly encounters computational constraints as the number of markers increases. The partition–ligation extension of the EM algorithm[24] increases the number of polymorphisms that can be handled computationally. However, for larger numbers of markers, the EM method is computationally expensive and loses accuracy by using a suboptimal model for haplotype frequencies. More accurate phasing can be obtained by better *a priori* modelling of probabilities of haplotype configurations, as is done by the coalescent-based and hidden Markov model (HMM) methods described below[25]. Many software implementations of the EM algorithm exist, including Arlequin[26] and PL-EM[24].

The EM algorithm is useful when a small number of polymorphisms in a short gene or haplotype block are to be studied. Clark's algorithm can also be used for this purpose, as can PHASE (see below), which would, in most cases, be a better choice. One application in this setting is haplotypic association testing. For example, Clark's algorithm was used to phase 13 tightly linked SNPs in the β2 adrenergic receptor (*ADRB2*) gene and found that a haplotype pair was significantly associated with bronchodilator response to a β-agonist in asthmatics, whereas individual SNPs were not[27]. A second application is determining whether a polymorphism that is seen in multiple populations has a single origin or whether it has independently arisen multiple times. This question can be answered by investigating whether the polymorphism occurs on a single haplotype background (single origin) or multiple haplotypes (multiple origins). For example, by using the EM algorithm on five SNPs, it was determined that the methylenetetrahydrofolate reductase 677T polymorphism is associated with a common haplotype in individuals from European, Asian and African populations[28]. This finding indicates that the polymorphism may have occurred on a haplotype that had a selective advantage.

Coalescent-based methods and hidden Markov models. Approximate coalescent models[29] were a breakthrough for modelling population haplotype frequencies[30–32]. These models recognize that new haplotypes are derived from old haplotypes by the processes of mutation and recombination. Because mutation and recombination events are rare over small genomic distances, haplotypes tend to look similar to each other. Thus, for example, if the haplotypes 1100 (where 0 and 1 represent two possible alleles at each of the four polymorphic sites) and 0001 are seen in the sample, it may also be likely that the haplotypes 1101 (formed by recombination) or 0011 (formed by mutation) are seen, but it is less likely that the haplotype 1111 (formed by a recombination and a mutation) will be seen. This approach forms the basis of many population-based statistical phasing methods, including PHASE[25,33], fastPHASE[34], MACH[4] and IMPUTE2 (REF. 35). In each case, the approximate coalescent gives rise to an HMM, and its parameters are estimated with the use of iterative algorithms, such as the stochastic EM algorithm[36,37]. The methods are described briefly below; additional details and comparisons of the methods are given in BOX 1.

PHASE[25,33] was for some time considered a gold standard for accuracy among population-based haplotyping algorithms[38]. It is still useful for small genomic regions, but it is very slow compared to newer algorithms. PHASE is suitable for moderately small numbers of markers (up to 100) and small sample sizes (up to several hundred individuals). For large genomic regions, other methods should be used, such as those described below. Available software includes PHASE itself and a faster implementation, SHAPE-IT[39].

FastPHASE[34] was an important milestone, because this algorithm made it possible to phase genome-wide SNP array data. For small numbers of individuals (up to 100), it is only a little less accurate than PHASE[34]. The speed of fastPHASE is partly achieved using a parsimonious clustering of haplotypes. For small sample sizes, this clustering captures almost all of the information. However, for larger sample sizes, computational feasibility is maintained at the cost of loss of information, leading to less accurate haplotypes than can be achieved with some of the more recent methods[40].

BEAGLE[40] is based on an HMM that does not explicitly model recombination and mutation, although these aspects are implicitly captured. The model clusters haplotypes at each locus, and the clustering adapts to the amount of information available so that the number of clusters increases globally with sample size and locally with increasing linkage disequilibrium (LD). Relative to fastPHASE, BEAGLE is an order of magnitude faster and is more accurate for medium and large sample sizes (>1,000 individuals) but is less accurate for small sample sizes (100 individuals)[40]. BEAGLE is not well suited for very small numbers of markers in a region (fewer than 100).

MACH[4] and IMPUTE2 (REF. 35) are new additions to the set of available statistical phasing methods. Both methods have primarily been used for the imputation of untyped variants but can also be used for haplotype phase inference and are based on the same approximate coalescent model[30]. These methods can handle larger data sets than PHASE can while giving greater accuracy for large sample sizes than fastPHASE can. In FIG. 2, we compare the performance of BEAGLE with that of MACH and IMPUTE2, as the haplotype phasing performance of MACH and IMPUTE2 has not previously been examined in detail. Using parameters suggested in the documentation for each program (FIG. 2a,b), MACH has the highest accuracy for the smaller sample sizes, and BEAGLE has the highest accuracy for larger sample sizes. There was more than an order of magnitude of difference in computing times between the method with the fastest computing time (BEAGLE) and the method with the slowest computing time (MACH). The accuracy of all programs can be improved at the cost of increased computing time (FIG. 2c,d). For MACH and IMPUTE2, increasing the model complexity by increasing the number of HMM states allows the methods to make better use of the information in the data and thus obtain more accurate results, although the program will take longer to run. For BEAGLE, accuracy is improved by combining the results from multiple runs.

One application of haplotype phase on a genome-wide scale is investigation of population structure. Auton *et al.*[41] used BEAGLE to phase almost 4,000 individuals from four continental regions at over 400,000 SNPs genome-wide. They used the phased haplotypes to compare patterns of haplotype diversity between populations.

Hidden Markov model
(HMM). A mathematically elegant and computationally tractable class of models in which the observed data are generated by an unobserved Markov process. A Markov process is a probabilistic process in which the distribution of future states (for example, states that are further along the chromosome) depends only on the current state and not on previous states.

Haplotype block
A short genomic region within which inter-marker linkage disequilibrium is strong.

Approximate coalescent
The coalescent is a model for the process by which the ancestry of alleles converges when looking back in time. An approximate coalescent is a model that generates patterns of genetic variation that are similar to patterns generated by the coalescent but that is computationally simpler.

Linkage disequilibrium
(LD). Non-independence (correlation) between genetic variants at the population level. In general, LD decreases with genomic distance and is not present between variants on different chromosomes.

Effective population size
The size of a population of randomly mating individuals that would show the same amount of genetic drift as is found in the actual population. The effective population size is usually smaller than the actual population size.

Box 1 | Approximate coalescent and HMM methods for computational phasing of unrelated individuals

A hidden Markov model (HMM) has underlying hidden states that are not directly observed. In haplotype phase inference, these states in some way represent the underlying true haplotypes. Transition probabilities determine the ways in which the hidden states can change from one chromosomal position to another, and emission probabilities link unobserved states to the observed data.

In the Li and Stephens[30] framework used by MACH[4] and IMPUTE2 (REF. 35), the hidden states are 'template haplotypes' — that is, haplotypes that have already been estimated within the sample. During each iteration of the estimation procedure, each individual's haplotypes are estimated using template haplotypes that were previously estimated in other individuals. Over successive iterations, the haplotype estimates improve, converging towards an optimal solution. MACH uses a random subset of sample haplotypes as templates, whereas IMPUTE2 uses a subset of haplotypes that are selected to be similar to the haplotypes of the individual currently being estimated. The IMPUTE2 strategy appears to permit greater improvement in accuracy as sample size increases and model complexity (the number of states) is held constant (FIG. 2). As part of the estimation procedure, MACH also estimates the transition probabilities for the hidden states (essentially the recombination rates) and the emission probabilities (representing the mutation rates). By contrast, IMPUTE2 takes as input the effective population size and recombination rates and uses these to derive transition rates and emission probabilities. This difference may account for some of the difference in computing times between the two methods (FIG. 2).

The model used by PHASE (version 2.1) is quite similar to the Li and Stephens framework but adds an additional set of parameters: the coalescent times between a given haplotype and the underlying template haplotype. All haplotypes — other than those of the individual being re-estimated at each step — are used as hidden states (templates), unlike MACH and IMPUTE2, which use only a subset of haplotypes as templates. This is one factor underlying the difference in computation times. Another factor is that PHASE uses Markov chain Monte Carlo methods to explore the space of all possible solutions, whereas MACH, IMPUTE2, BEAGLE and fastPHASE use stochastic expectation maximization (EM) to converge towards the most probable solutions.

BEAGLE[40] forms an HMM by locally clustering the haplotypes at each marker position along a chromosome. The haplotypes are locally clustered in such a way that haplotypes in the same cluster tend to have similar probabilities for alleles at downstream markers. The haplotype clusters are the hidden states. At each iteration of the algorithm, new haplotype estimates are sampled from the current state of the HMM conditional on the genotype data, and these haplotype estimates are used to build a new HMM. The model is parsimonious in several ways. First, the clustering of haplotypes keeps the number of underlying hidden states low. Second, the model only considers a small subset of all possible transitions between states at one position and states at the next position (whereas the Li and Stephens framework allows for all possible transitions). The transitions considered are those implied by the haplotype estimates used to build the current model. These differences between BEAGLE and the Li and Stephens framework are described in more detail elsewhere[46].

The fastPHASE[34] method also locally clusters haplotypes; however, the way the clustering is performed is different to that of BEAGLE. The BEAGLE approach allows different positions to have different numbers of clusters (hidden states), whereas fastPHASE uses the same number of clusters at each position. For small sample sizes, the optimal number of clusters can be determined and used, but for large sample sizes, the optimal number of clusters would be larger than is computationally feasible. FastPHASE is similar to PHASE, MACH and IMPUTE2 (but different from BEAGLE) in allowing for all possible transitions between states from one position to the next.

For example, they found that Japan has a lower diversity than Taiwan, and that south east Europe has a lower diversity than south west Europe. Haplotype patterns can also be used to detect signatures of selection. Sabeti *et al.*[42] used HapMap data[43] that had been phased with PHASE to look for unusually long haplotypes, which are a signature of positive selection. They found hundreds of strong candidates across the genome. Another important application of genome-wide phasing is to pre-phase data before performing imputation. Although pre-phasing data prior to imputation is not necessary for some imputation programs, it can substantially speed up the imputation process, but it also incurs a small loss in accuracy. The main imputation programs (including BEAGLE, MACH and IMPUTE2) are also phasing programs and are typically used for the pre-phasing step, if it is required. The largest public reference panels used for imputation (namely, the HapMap[43,44] and 1000 Genomes[8] projects) are available in phased versions. Haplotype association testing can also be performed on a genome-wide scale using phased haplotypes[45–47].

Making use of identity-by-descent. A recent development in computational phasing of haplotypes is the use of IBD information. Even in a sample of 'unrelated' individuals, distant relationships give rise to segments of IBD, which can be used for phasing, as described in more detail in the next section. The IBD that is useful in this context is that which is due to a recent shared ancestor, such as within the past 20 generations, which leads to detectable long segments of IBD[48,49]. A rule-based version of this approach was pioneered by Kong *et al.*[13] in their long-range phasing algorithm, which was applied to the Icelandic population. In this study, IBD tracts were identified by searching for long genomic segments (≥10 Mb) for which two individuals shared an allele at all markers in the segment. The IBD-based approach worked particularly well in that setting, because Iceland is a small, largely isolated population and because a high proportion of the existing population (over 10%) has been genotyped. Because the genotyped Icelandic sample is large relative to the population, for most individuals at most loci, it is possible to find multiple other individuals

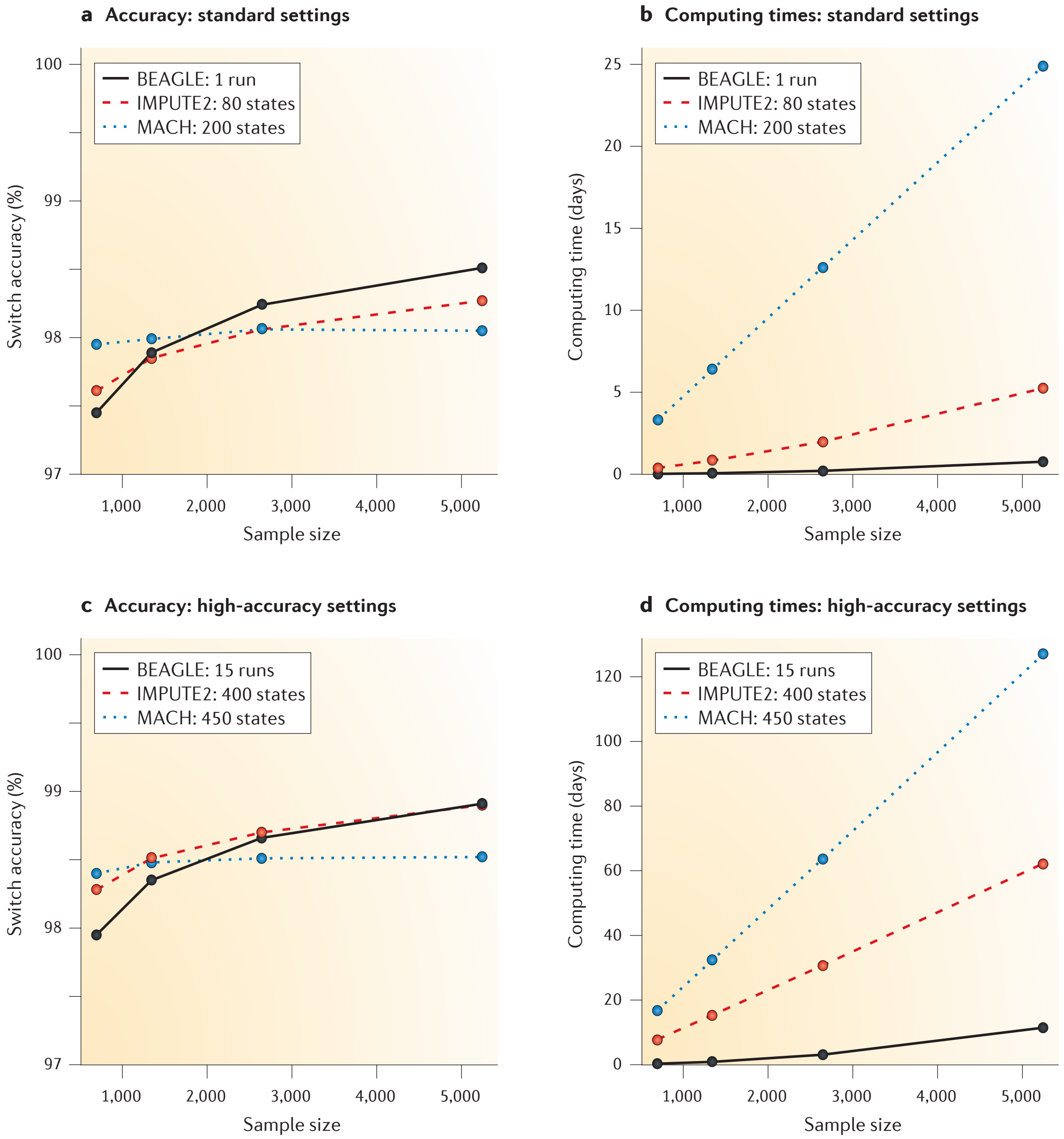

Figure 2 | **Comparison of recent statistical haplotype phasing methods.** We compared phasing accuracy and computation time for BEAGLE 3.3.1 (REF. 40), IMPUTE 2.1.2 (REF. 35) and MACH 1.0.16 (REF. 4). The sample was comprised of up to 5,200 controls from the Wellcome Trust Case Control Consortium 2 (REFS 84,85) and 44 offspring from the HapMap3 (REF. 44) CEU (Utah residents with northern and western European ancestry from the CEPH collection) trios that had been genotyped on Illumina Human1M SNP arrays. We evaluated accuracy for markers on chromosome 20 (21,166 markers after quality control filters). Phasing accuracy was measured in the HapMap trio offspring using markers that have their phase determined by parental genotypes. Accuracy is represented by switch error rate (BOX 2). BEAGLE was run with default settings with the low-memory option (use of the low-memory option does not affect accuracy but reduces memory usage at the cost of a 30–60% increase in computing time). To obtain results in a reasonable amount of time for MACH and to follow recommended practice for IMPUTE2, the data for MACH and IMPUTE2 were split into 11 chunks of 5.1 Mb and one 6.3 Mb chunk, leaving a 500 kb overlap for adjacent chunks. The two haplotypes for each individual were aligned across chunks using the phase of heterozygous genotypes near the centre of the overlap region, and the chunks were merged to yield chromosome-wide phasing. Computing times are for the whole chromosome and are obtained for MACH and IMPUTE2 by adding computing times for each chunk. **a**,**b** | This comparison used parameter settings that are based on the current documentation for each program. Parameter settings for IMPUTE2 followed parameters in a prototype phasing script downloaded from the IMPUTE2 website: '-phase -include_buffer_in_output -stage_one -k 80 -iter 30 -burnin 10 -Ne 11500'. MACH options were '--round 50 --states 200 --phase', as suggested in the MACH documentation. **c**,**d** | As in panels **a** and **b**, but there is increased model complexity or run time for each method to obtain improved accuracy. BEAGLE was run 15 times, and the results were combined by phasing successive heterozygotes using a majority vote from the 15 runs. MACH was run with 450 states (compared to 200 for the standard settings), and IMPUTE was run with 400 states (compared to 80 states for the standard settings).

SNP index	Unphased genotypes		Shared haplotype	IBD-phased genotypes				Possible phasing A				Possible phasing B			
	Individual 1	Individual 2		Individual 1		Individual 2		Individual 1		Individual 2		Individual 1		Individual 2	
1	A/C	A/C	?	?	?	?	?	A	C	A	C	C	A	C	A
2	C/T	C/C	C	C	T	C	C	C	T	C	C	C	T	C	C
3	T/T	T/G	T	T	T	T	G	T	T	T	G	T	T	T	G
4	G/G	A/G	G	G	G	G	A	G	G	G	A	G	G	G	A
5	C/C	C/C	C	C	C	C	C	C	C	C	C	C	C	C	C
	Population frequency of haplotype (second instance of shared haplotype in parentheses)							5%	6%	(5%)	0.1%	0.2%	0.3%	(0.2%)	3%
	Population frequency of ordered trio of haplotypes							5% × 6% × 0.1% = 3.0 × 10^{-4}%				0.2% × 0.3% × 3% = 1.8 × 10^{-5}%			
	Posterior probability of phasing (normalized population frequency of trio of haplotypes)							3 × 10^{-4}% / (3 × 10^{-4}% + 1.8 × 10^{-5}%) = 94%				1.8 × 10^{-5}% / (3 × 10^{-4}% + 1.8 × 10^{-5}%) = 6%			

Figure 3 | **Use of identity-by-descent to determine haplotype phase.** First, we discuss how to determine phase using identity-by-descent (IBD) alone (main columns 1–4).When two individuals are known to be identical-by-descent (for example, if they are a parent–offspring pair), the individuals share an allele at each marker, and this allele is determined by the genotype data when one or both individuals are homozygous. In this example, the two individuals with unphased genotypes shown in main column 2 are identical-by-descent. SNP 1 is heterozygous in both individuals and thus cannot be phased using IBD but may be able to be phased using population haplotype frequencies (see below). SNP 2 is homozygous in individual 2, and so the shared haplotype must have the C allele. Analogously, SNPs 3 and 4 are homozygous in individual 1, so the shared alleles are T and G, respectively. SNP 5 is homozygous in both individuals, so phasing is trivial. The inferred shared haplotype is shaded green. Use of IBD phasing alone gives the phasing shown in the IBD-phased haplotype columns, in which the phasing of SNP 1 is unknown. Second, we discuss how to determine phase using IBD and haplotype frequencies. Consider the same two identical-by-descent individuals as above. The phase is determined by IBD at SNPs 2–5 (main column 3) but is not determined at SNP 1, which is heterozygous in both individuals. Only haplotype phasings that satisfy the IBD-phasing constraints need be considered. Here the two identical-by-descent individuals are phased jointly, so the joint phase at SNP 1 must be consistent with the IBD, and the identical-by-descent haplotype is only included once in the probability of the haplotype configuration. The inferred identical-by-descent haplotype is shaded in main columns 5 and 6. Haplotype phasing pattern A is much more probable (94%) than phasing pattern B (6%).

who share a haplotype that is identical-by-descent and that can be used for phasing. Consequently, it was possible to phase approximately 90–95% of heterozygous markers in the Icelandic sample. Direct application of Kong *et al.*'s rule-based approach to large outbred populations is not currently practical, as it would require at least 1% of the population to be genotyped. It is likely that the applicability of IBD-based phasing can be extended to additional populations by using more sensitive methods for detecting IBD and combining IBD-based phasing with population haplotype frequency models. Software is available for long-range phasing using IBD[50,51]. These programs are suitable for phasing large pedigrees or samples from small populations in which all individuals are closely related.

The long-range haplotypes that were generated by Kong *et al.* have been used for several interesting applications. Kong *et al.*[52] used genealogy and the inferred haplotypes to determine the parental origin of alleles and to test for association with disease. They found several parental-origin-specific associations. Holm *et al.*[53] used the inferred haplotypes for accurate imputation of a putative rare causal variant in other individuals to obtain a stronger association signal. Kong *et al.*[13] also showed that the haplotypes can be used to study fine-scale recombination and to study the inheritance of recurrent mutations.

Computational phasing in related individuals

In related individuals for whom pedigrees are available, Mendelian constraints (or, more generally, IBD constraints) provide information for determining the haplotype phase at many sites. For example, a parent–offspring pair must share one allele that is identical-by-descent at every position. The identical-by-descent alleles at different sites on the same chromosome will be on a single haplotype in the child and on a single haplotype in the parent, assuming that recombination has not occurred between the sites in the transmission of the chromosome to the child. FIGURE 3 gives an example of the use of IBD to determine haplotype phase. More generally, if two individuals have IBD across a region on a chromosome, they must share one allele that is identical-by-descent at every position in the region, and the identical-by-descent allele will usually be on a single shared haplotype in both individuals. If one or both individuals have a homozygous genotype at a site within the region of IBD, the allele in the homozygous genotype must be the shared allele, so that the identical-by-descent allele is known, and the site is phased relative to all other sites in the region for which the identical-by-descent allele is known. Thus, for diallelic markers, such as SNPs, haplotype phase is only unknown at positions at which both individuals are heterozygous or not identical-by-descent or at positions at which one individual has a missing genotype.

Further information on haplotype phase is obtained when more than two relatives are considered simultaneously. For example, at diallelic markers in parent–offspring trios (mother–father–child), the only positions at which phase is not determined are those at which all three individuals are heterozygous (a small proportion of sites) or sites at which one or more of the individuals has a missing genotype. Larger families contain even more information on haplotype phase, although this is not trivial to extract. Linkage programs, such as GENEHUNTER[54], can extract this information, although they assume that sites are in linkage equilibrium (that is, not in LD). When sites are in LD, linkage programs that assume linkage equilibrium may falsely infer IBD in situations in which it is not present; this is a problem for pedigrees that have many ungenotyped individuals, and it leads to incorrect phasing[55]. Moreover, because these methods assume that markers are in linkage equilibrium, they cannot use information from population haplotype frequencies.

As an example of family-based phasing, sequence data on a nuclear family (two parents and two children) were analysed, and inheritance patterns and hence haplotype phase were inferred from this analysis. Phase information was used to look for genes in which the affected children had compound heterozygosity for dysfunctional variants, and this, in turn, enabled determination of the genes that were responsible for two rare syndromes affecting the children[56].

Compound heterozygosity
The presence of two deleterious variants located in the same gene but on different chromosome copies of an individual. It is possible to distinguish between compound heterozygosity and the occurrence of two variants on the same chromosome copy by determining the haplotype phase.

Long-range phasing in families. The Kong *et al.*[13] approach to IBD detection and phasing in unrelated individuals can also be applied to data from related individuals, and this approach can be used with sites in LD. However, diallelic markers can only be phased when one of the related individuals is homozygous or when one of the related individuals can be phased from other IBD relationships so that the allele on the shared haplotype can be determined. The use of IBD to phase related individuals provides essentially perfect phasing (barring genotype error and recent mutation) over long chromosomal regions at sites that can be phased with the IBD information alone (that is, not at sites at which the identical-by-descent individuals are all heterozygous). Fortunately, population haplotype frequency information is also available to estimate the phase at those ambiguous sites. Haplotype frequency information can provide accurate phasing over short genomic regions and thus, in principle, can fill in the gaps to provide an overall phasing that is highly accurate (FIG. 3). The use of haplotype frequency information with IBD-based phasing is currently an active area of research.

Using population haplotype frequency information. Some methods use both IBD and population haplotype frequency information to phase related individuals, although the existing methods are limited in various ways. It is possible to use family information in conjunction with the EM algorithm to estimate haplotype phase[57,58]. This approach can only analyse small genomic regions. MERLIN is a linkage program that allows limited LD in the form of clusters of tightly linked markers[59]. This approach to LD modelling is not adequate for highly dense genotype data, such as those generated from current genome-wide SNP microarrays or sequencing. BEAGLE[3], SHAPE-IT[39] and modifications of other programs[38] use IBD and haplotype frequencies to phase parent–offspring trios. These methods work well for trios and parent–offspring pairs but are not easily extended to larger families with multiple offspring or multiple generations owing to intrinsic limitations in the algorithms.

Ignoring relationship. It is possible to phase related individuals as if they were unrelated, using only population haplotype frequency information. FIGURE 4 demonstrates that the haplotype phase of closely related individuals will be estimated more accurately than that of unrelated individuals, even when the relationship information is ignored or unknown[60]. This result is because the occurrence of the same extended identical-by-descent haplotype several times in the sample helps in its estimation. Drawbacks of this approach are that it is possible to have inconsistencies between the haplotype phases of closely related individuals (that is, the phasing may imply the unlikely occurrence of several closely spaced recombinations, or imputed missing

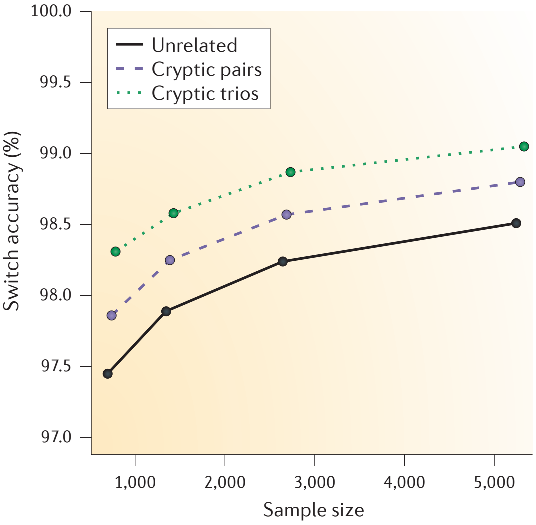

Figure 4 | **Accuracy of statistical phasing of cryptic relatives when relationship is not explicitly accounted for.** The same sets of individuals were phased as in FIG. 2, but either one parent of each HapMap CEU child ('cryptic pairs' results) or both parents ('cryptic trios' results) were added. Phasing was performed with BEAGLE assuming that all samples were unrelated. The 'unrelated' results are identical to those for BEAGLE in FIG. 2a and do not include any of the parents. It can be seen that adding relatives to the phase estimation greatly improves phase accuracy, even when the individuals are treated as unrelated. The phase accuracy would be substantially further improved by using the known relationships during the phase estimation.

D′

A measure of linkage disequilibrium (LD) between two markers. D′ takes values between 0 and 1. Absence of LD is indicated by 0, and 1 indicates maximum possible LD given the allele frequency of the markers.

genotypes may not be consistent with Mendelian rules), and the accuracy of the phase estimation will not be as high as it could be if the relationships were fully utilized. Nonetheless, this approach presents a simple solution that will provide acceptable accuracy for many applications.

Factors influencing phasing accuracy

A number of factors influence the achievable computational phasing accuracy. These include sample size, marker density, genotype accuracy, relatedness in the sample, ethnicity and allele frequency.

Sample size. Assuming that other factors are equal, the larger the sample size, the greater the haplotype phasing accuracy (FIG. 2). This is particularly the case when the statistical model can incorporate the large amount of information on population haplotype frequencies contained in larger data sets[40]. This applies to family data as well as to unrelated data[3] in situations in which haplotype frequency information is used to phase those sites that do not have phase determined by IBD. Thus, a simple and powerful strategy for improving haplotype phase accuracy is increasing the sample size through use of a reference panel of individuals from the same population and using a phasing method that can make effective use of the additional data.

Marker density. Whether increased marker density results in improved or reduced accuracy depends on the measure of accuracy being considered (BOX 2). On a per-marker basis, haplotype estimates are more accurate with denser data. However, on a regional basis and with an absolute measure of accuracy (totally correct haplotype over a region), having greater marker density results in more opportunities for error and thus lower accuracy.

Genotype accuracy. Genotype accuracy influences haplotype phase accuracy, because at least one of the two estimated haplotypes for an individual must be wrong whenever a genotype is mis-specified. When genotype data are noisy or incomplete, as is the case with low-coverage sequence data, one solution is to phase genotype likelihoods rather than called genotypes. Genotype likelihoods capture the uncertainty in the genotype data, and both EM-based and HMM-based phasing algorithms can be adapted to phase genotype likelihood data[5,6]. With genotype likelihood data, posterior genotype probabilities and haplotype phase are estimated simultaneously, which increases the accuracy of both tasks[6,8,61].

Degree of relatedness. Known relatedness, if used along with haplotype frequency (such as in parent–offspring trios), results in markedly superior haplotype phase estimation compared with using only unrelated individuals[38]. As we have shown above, even if closely related individuals are treated as unrelated individuals, their haplotypes will be more accurately estimated than those of unrelated individuals.

Sample ethnicity. African populations have more haplotype diversity and lower levels of LD compared with non-African populations, such as Europeans. Allele frequencies and density of polymorphisms are confounding factors when comparing accuracy across ethnicities, and the comparison will depend on the accuracy metric and also perhaps on the phasing algorithm used. Overall, there does not seem to be a clear pattern of differences in phasing accuracy between populations from different continents[3,38,62]. In the context of genotype imputation, including samples from closely related populations in the imputation reference panel can improve genotype imputation accuracy, particularly for low-frequency variants[63,64]. This suggests that when the

Box 2 | Metrics for comparing haplotype phasing methods

Metrics for computational phasing accuracy

Three primary metrics are used to measure computational haplotype phase accuracy: haplotype accuracy, imputation accuracy and switch error. It is generally sufficient to use one rather than all of the metrics when comparing algorithms, because the metrics tend to produce similar rankings. The haplotype accuracy and switch error metrics require the existence of gold-standard phased data. This gold standard may come from nuclear family data or from experimental phasing. When gold-standard data are available, switch accuracy is usually the most informative metric. The imputation accuracy metric is unique in that it can be applied to any data set without requiring the existence of gold-standard phased data. Thus, one can use this metric to make sure that the haplotype inference procedure is performing properly or to choose which program settings to use.

Haplotype accuracy. This measure relates to the proportion of haplotypes that are correctly inferred over the whole region of interest. This metric is typically only relevant for small numbers of markers, as the chance of correctly phasing a large region is very small, even for the best statistical phasing methods. This metric can be applied to simulated data for which the true haplotype phase is known, or it can be applied to real data for which Mendelian constraints from closely related individuals determine the true phase at most sites.

Imputation accuracy. Haplotype phasing algorithms generally impute sporadic missing data as part of the phasing algorithm. Some of the genotypes can be masked (that is, some genotypes can be set to 'missing data' status), allowing determination of the proportion of imputed alleles that are correctly imputed by the phasing algorithm. This metric can be applied to any data set, because it does not require knowledge of the true haplotype phase.

Switch error. When comparing an inferred haplotype phase to the true haplotype phase, it is possible to count how many switches (recombination events in the inferred phased haplotypes) are required to obtain the true haplotype phase. This comparison can be expressed as a rate: the number of switches required divided by the number of opportunities for switch error, which is the number of heterozygote markers in the individual's genotype minus 1 (the first heterozygote marker can be assigned an arbitrary phase).

Metrics for experimental phasing accuracy

The experimental phasing of an individual's genotypes is independent of statistical phasing using a reference panel (see 'Computational phasing in unrelated individuals' in the main text), provided that statistical phasing has not been used as part of the experimental phasing procedure. The statistical phase of pairs of heterozygous SNPs for which linkage disequilibrium is high (for example, $D' > 0.9$ or $D' = 1$) will be highly accurate and can be compared with the experimental phase to obtain a rate of concordance[69,70]. Similarly, if the individual has close relatives who have been genotyped, Mendelian or identity-by-descent constraints (see 'Computational phasing in related individuals' in the main text) can be used to determine accurately the phase of many SNPs, and the proportion of SNPs at which the experimental phase is discordant can be calculated[73,74]. In addition, for experimental phasing methods, the proportion of heterozygous SNPs at which the phase can be determined is an important factor, as this is typically much less than 100%. The proportion of SNPs at which the genotype is incorrect or missing also needs to be considered, as this can be lower than for methods generating unphased data.

Box 3 | Recent methods for whole-genome experimental phasing

These methods separate whole chromosomes (the first three methods below) or long haplotypes (the fourth method) using a variety of approaches. The separated chromosomes are either individually tagged or first combined into pools in such a way that most pools will contain at most one copy of each homologous chromosome or haplotype. The chromosomes are then sequenced or genotyped. All of these methods are at the proof-of-concept stage, so it is difficult to know which, if any, will develop into a widely used protocol.

Microdissection

Ma *et al.*[74] arrested cells in metaphase, then spread chromosomes and then microdissected them into subsets. Some subsets may contain two homologous chromosome copies, which cannot then be phased using that subset. They then genotyped each subset with a whole-genome genotyping array. Phased genotypes are available for chromosomes that have a single homologous copy in one of the subsets.

Fluorescence-activated cell sorting

Yang *et al.*[83] used fluorescence-activated cell sorting (FACS) to separate individual chromosomes; these were amplified and tagged before sequencing to enable reads to be mapped back to specific chromosome copies.

Microfluidics

Fan *et al.*[73] developed a specialized microfluidic device to capture a single metaphase cell and then partition the 46 chromosomes. Each chromosome was typed to determine its identity. Two pools were constructed that contained only one copy of each homologous chromosome. Each pool was genotyped separately with a whole-genome genotyping array to obtain phased genotypes.

Fosmids

Kitzman *et al.*[69] created a fosmid library of long haplotypes. These were separated into 100 pools. The probability that a given pool contains two non-homologous copies of the same chromosomal region is low. Barcode-labelled shotgun libraries were constructed from each pool. These were sequenced at low depth (2–3×). The barcodes were used to assemble the shotgun reads back into the haplotypes contained in the fosmid libraries. In addition, the individual was sequenced using standard next-generation sequencing with a higher depth (15×). The unphased genotypes were determined from this sequencing, and then the haplotypes from the earlier step were used to determine haplotype phase.

sample is small and no other individuals from the same population are available to use as a reference panel, the haplotype estimation accuracy can be improved by including samples from other populations, particularly from closely related populations, such as those from the same continent. For samples with admixed ancestry, such as African Americans, including samples from the ancestral populations may improve phasing accuracy[62].

Allele frequency. Rare variants are difficult to phase computationally because, to obtain high-confidence phase information, a variant must be seen several times within its haplotype context. In particular, computational approaches cannot phase mutations that have arisen *de novo* in an individual, unless data on the individual's offspring are available. For this important class of variants, experimental phasing methods are required.

Experimental phasing

Experimental phasing is expensive and labour-intensive. Nonetheless, when very accurate long-range haplotypes are required and close relatives are not available for IBD-based computational phasing, experimental phasing methods are available that can be applied during data generation. Also, sequencing technologies automatically produce some information on phase, and methods for using that information are beginning to be developed.

Whereas several experimental phasing methods provide complete phasing of whole chromosomes, other methods only provide phasing for long or short haplotype fragments. If only phasing for fragments is provided, computational methods must be applied to assemble overlapping fragments into larger haplotypes. This problem, known as the single-individual haplotype-reconstruction problem, is theoretically challenging and has received considerable attention[65]. In most cases, population data are not used, but several recent methods use both experimentally derived haplotype fragments and population information[66,67].

Whole-genome experimental phasing. The human reference sequence generated by the International Human Genome Sequencing Consortium was produced by first creating large-insert clones and then shotgun sequencing the clones[68]. The clone inserts are single haplotypes, resulting in haploid sequence. The same large-insert clone plus shotgun sequencing approach can be used directly to generate sequence data on phased haplotypes, although it is extremely expensive on a whole-genome scale. Recently, Kitzman *et al.*[69] combined this approach with next-generation sequencing to produce whole-genome sequence data that were mostly phased (BOX 3). The added cost of applying this approach, beyond the cost of the whole-genome sequencing, was approximately US$4,000 ($1,000 for labour and $3,000 for reagents) for the sequence of a single individual[69]. Suk *et al.*[70] used a similar approach and indicated a cost of under €6,000 ($8,600), including the cost of the whole-genome sequencing. These methods do not provide completely phased chromosomes, because the phased haplotype fragments must be pieced together, which can incur errors. Suk *et al.*[70] used ReFHap[71] to assemble the fragments, whereas Kitzman *et al.*[69] used a reimplementation of HapCUT[72]. Suk *et al.*[70] were able to phase 99% of SNPs, and the phased blocks had N50 length of 1 Mb (that is, 50% of resolved sequence is in a block of at least 1 Mb in length), whereas Kitzman *et al.*[69] had a lower coverage: 94% of SNPs phased into blocks with N50 length of 400 kb. A disadvantage of these approaches is that, although large chromosomal segments can be phased, the segments may not be accurately stitched together owing to missing phase information across regions of homozygosity exceeding fosmid size (40 kb)[70].

Other approaches to experimental phasing are based on various other means of separating homologous chromosomes or haplotype fragments before genotyping or sequencing. Some of these methods can phase whole chromosomes. A recent method that enables whole-genome phasing in an automatable approach is the use of a specialized microfluidics device to separate chromosomes from a single cell in metaphase[73]. The separated chromosomes can then be sequenced or SNP-genotyped (BOX 3).

One application of experimentally phased whole-genome sequence is population genetics analysis. Kitzman *et al.*[69] analysed the experimentally phased sequence of a Gujarati Indian individual and determined that the novel variants mostly fall on haplotypes that are

Reference panel
A collection of samples that are not of direct interest but that are included in an analysis for the purposes of increasing statistical power or accuracy for the samples of interest. Reference panels are commonly used for genotype imputation and can also be used for haplotype phasing.

Genotype likelihoods
Statistical likelihoods that encapsulate the relative evidence for each possible genotype call.

Fluorescence-activated cell sorting
(FACS). A type of flow cytometry in which individual particles (such as chromosomes) are separated and fluorescence intensities (from earlier staining) are measured.

not European-like. Another application is clinical interpretation of personal genomes. Suk *et al.*[70] found 171 genes with two or more potentially severe mutations in the genome of a German individual, of which 159 were experimentally phased. Of these, 86 were in *cis* and 73 were in *trans*. Configurations in *cis* leave one copy of the gene unchanged, which is likely to be less damaging. A further potential application of experimental phasing is determination of human leukocyte antigen (HLA) haplotypes[70,73] for donor–recipient matches in transplant medicine. More work is needed to assess whether the level of accuracy is sufficiently high for this application.

Whichever method is used, whole-genome experimental phasing is more expensive than generating unphased whole-genome data. The methods require an initial processing step, such as developing fosmid libraries or separating chromosomes. As such methods become more mainstream, it is likely that this initial processing will be automated, thereby saving time and reducing costs. Nonetheless, additional equipment and/or reagents are needed for this step. Furthermore, the sequencing or genotyping that follows the initial processing tends to involve some additional sequencing or genotyping beyond that required for generation of unphased data. For example, Kitzman *et al.*[69] and Suk *et al.*[70] generated unphased sequence data as well as the phased fosmid sequences to improve the quality of the final phased data, whereas Ma *et al.*[74] estimate that five to six genome-wide genotyping arrays are needed per sample on average for their chromosome microdissection method (BOX 3).

Phase information in sequence reads. Direct sequencing can also provide some information for phasing. Sanger sequencing produces reads that are quite long (>700-base read lengths are possible[75]). The use of paired-end sequencing also provides information for haplotype phasing. When a read encompasses two or more heterozygous genotypes of an individual, the phase of the heterozygote genotypes is determined, as each fragment from which a read or pair of reads is obtained is a single haplotype. Thus, if the fragments are long and sequence coverage is sufficiently high, a substantial amount of haplotype phase information can be obtained[76]. Sanger sequencing is too expensive for whole-genome sequencing on a large scale, but recently developed real-time, single-molecule sequencing methods are much cheaper, and these methods can yield sequence reads that exceed 1 kb[77,78]. Long read lengths may permit direct phasing from experimental sequence reads with sufficient sequence coverage.

Next-generation sequencing technology is considerably less expensive than Sanger sequencing, but the reads are shorter, providing less information for phasing[77]. Nonetheless, short reads — especially when they are paired-end reads — provide some information for phasing that can be incorporated into computational phasing[66,67]. Software for using phase information from next-generation sequence reads includes the Haplotype Improver software[67] and the 'read-backed phasing' algorithm that is incorporated into the Genome Analysis Tool Kit software[79].

Although whole-genome experimental phasing is likely to remain a niche application owing to its cost and complexity, the use of phase information from sequence reads is likely to become increasingly important. At present, phase information from sequence reads is not sufficient to determine haplotype phase fully, thus we expect to see the marriage of experimental and computational phasing as read information is incorporated into computational phasing methods.

Computational versus experimental phasing

There are several factors to consider when choosing whether to perform experimental phasing, computational phasing with related individuals or computational phasing with unrelated individuals. Computational phasing in unrelated individuals is the simplest and most inexpensive approach. Computational phasing in related individuals is straightforward if related samples are available and if the relationships are simple (in particular, parent–offspring pairs or trios are easily handled). The use of parent–offspring trios increases the genotyping or sequencing cost by threefold if the additional two individuals in each trio are not otherwise of interest. Experimental phasing increases the cost of data generation by two- to fivefold, requires high levels of technical expertise and may require investment in specialized equipment. Of the existing whole-genome experimental methods, the fosmid-based approaches[69,70] appear to be the least expensive (an approximately twofold increase in cost over standard sequencing), although the resulting haplotype phase has some gaps in coverage. Most types of phasing, excluding only the methods based on whole-chromosome separation, require more computing resources than those typically found in a desktop PC and at least a moderate level of bioinformatics expertise.

Computational phasing in unrelated individuals provides accurate phasing of common SNPs over small regions when the sample is large or when a large reference panel is used. This is adequate for a number of applications, including testing for haplotypic association, imputing ungenotyped common SNPs and comparing haplotypic diversity across populations. However, the accuracy is not high enough for some other applications, such as investigating compound heterozygosity, in which the variants of interest are at a low frequency.

Rare and low-frequency SNPs may be phased by genotyping or sequencing related individuals or by experimental phasing. Both approaches will provide a highly accurate phase at most SNPs, whether they are rare or common, and will provide phasing over long chromosomal regions or whole chromosomes. *De novo* mutations can be phased with experimental approaches or with data from offspring (if available). With most phasing methods, one can expect to have some positions at which the phase is unknown or incorrectly estimated. For example, the phase is not known with certainty at SNPs for which parent–offspring trios are all heterozygous (although the use of population data may allow these positions to be phased with a reasonably high level of confidence). As another example, fosmid-based experimental haplotyping results in long blocks

Barcode labelling
Tagging of each sample with a unique short sequence (barcode) before pooling samples. After sequencing, the sample corresponding to each read can be determined from the barcode.

Admixed ancestry
An individual has admixed ancestry if he or she has recent ancestors deriving from different continental populations.

Large-insert clones
Large haplotype fragments that are inserted into, for example, bacterial artificial chromosomes (BACs).

Shotgun sequencing
A sequencing method in which DNA is randomly sheared into small fragments before being sequenced.

Fosmid
A type of hybrid DNA molecule comprising bacterial DNA and a section of genomic DNA of ~40 kb in length.

Microfluidics
The manipulation of fluids on a very small scale. This approach can be used to separate individual chromosomes before sequencing for experimental phasing.

Metaphase
A stage of mitosis at which chromosomes are highly condensed, facilitating their separation for some experimental phasing methods.

Paired-end sequencing
Sequencing of haplotype fragments from each end. The two sequenced ends are typically separated by a gap.

of phased haplotypes, so the phase is unknown across block boundaries. In addition to the gaps in the phase, the phased haplotypes may be incorrect at some positions owing to underlying genotype or sequence errors.

Current challenges and future directions

The incoming flood of large-scale sequence data presents challenges for haplotype phasing. Computational phasing is difficult for low-frequency variants, and experimental phasing is currently too expensive for use on a large scale. Developments in statistical and experimental methods promise to meet these challenges. It remains to be seen whether whole-genome experimental phasing will end up being sufficiently inexpensive and automatable for common use, but the recently proposed methods suggest some promise. The use of phase information from short reads, together with statistical information from haplotype frequencies, is another area of development. Next-generation sequence read lengths are increasing in size as the technology develops, thus providing increased information about haplotype phase, although improved statistical methodology is needed to fully exploit this information.

To obtain improved computational phasing of data from unrelated individuals, there is a need for large panels consisting of thousands of individuals from different ethnicities. This will enable researchers with small sample sizes to borrow information on haplotype frequencies and will also enable the use of IBD-based haplotype phasing for improved accuracy. There is an additional need for large sets of individuals who can be used as reference panels for imputation. It would be advantageous if these panels were accurately phased to save computation time in the imputation analyses. The 1000 Genomes Project[8] is an important step towards achieving these goals; however, larger sets of individuals from each major population would provide further benefits. For some continental groups that have been the focus of existing studies, such as Europeans, large panels can be obtained by combining existing resources, but there is a practical need for large reference panels that have had careful quality control filtering and that have been accurately phased to be available as a single unified data set.

Virtually all of the existing methods for the statistical inference of the haplotype phase assume diallelic markers, although there are some exceptions[40,80,81]. There is a need to extend existing methods to incorporate multiallelic markers and to evaluate the accuracy of phasing methods when they are applied to data containing copy number variants.

As sequencing technology becomes less expensive and more ubiquitous, the computational challenges will become even more prominent. There is a need for even faster computational methods for haplotype phasing that are also highly accurate and able fully to exploit the information present in large samples. The use of IBD information that is latent in samples of 'unrelated' individuals shows promise for increasing haplotype phase accuracy in large samples. Recent work on improved resolution of IBD detection[48] should permit the extension of IBD phasing from founder populations with a high proportion of individuals genotyped[13] to outbred populations with a lower proportion of genotyped individuals.

Finally, computational haplotype phasing of related individuals that makes use of relationships (IBD constraints) and haplotype frequencies is a remaining challenge area. As the pendulum moves back from the common-disease, common-variant hypothesis — which has a focus on association studies in unrelated individuals — to a greater focus on rare variants that are most easily studied in family data[82], methods for the analysis of related individuals will be increasingly important.

1. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nature Rev. Genet.* **12**, 215–223 (2011).
2. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39**, 906–913 (2007).
3. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
4. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
5. Kang, H., Qin, Z. S., Niu, T. & Liu, J. S. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **74**, 495–510 (2004).
6. Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* **85**, 847–861 (2009).
7. Yu, Z., Garner, C., Ziogas, A., Anton-Culver, H. & Schaid, D. J. Genotype determination for polymorphisms in linkage disequilibrium. *BMC Bioinformatics* **10**, 63 (2009).
8. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
9. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **21**, 952–960 (2011).
10. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
11. Scheet, P. & Stephens, M. Linkage disequilibrium-based quality control for large-scale genetic studies. *PLoS Genet.* **4**, e1000147 (2008).
12. Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
13. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
This paper describes the use of an IBD-based phasing method called 'long-range phasing' in a large sample from the Icelandic population.
14. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
15. Tao, H., Cox, D. R. & Frazer, K. A. Allele-specific KRT1 expression is a complex trait. *PLoS Genet.* **2**, e93 (2006).
16. Gusfield, D. Haplotype inference by pure parsimony. *Lect. Notes Comp. Sci.* **2676**, 144–155 (2003).
17. Wang, L. & Xu, Y. Haplotype inference by maximum parsimony. *Bioinformatics* **19**, 1773–1780 (2003).
18. Weale, M. E. A survey of current software for haplotype phase inference. *Hum. Genomics* **1**, 141–144 (2004).
19. Clark, A. G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**, 111–122 (1990).
This paper describes the first computational phasing method for more than two markers.
20. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38 (1977).
21. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
This was one of the earliest papers describing the use of the EM algorithm for statistical phasing of unrelated individuals.
22. Hawley, M. E. & Kidd, K. K. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* **86**, 409–411 (1995).
23. Long, J. C., Williams, R. C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810 (1995).
24. Qin, Z. S., Niu, T. & Liu, J. S. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**, 1242–1247 (2002).
25. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
26. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).

27. Drysdale, C. M. *et al.* Complex promoter and coding region β 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
28. Rosenberg, N. *et al.* The frequent 5,10-methylenetetrahydrofolate reductase C677T polymorphism is associated with a common haplotype in whites, Japanese, and Africans. *Am. J. Hum. Genet.* **70**, 758–762 (2002).
29. McVean, G. A. & Cardin, N. J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B* **360**, 1387–1393 (2005).
30. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
This paper describes the approximate coalescent model used by the MACH and IMPUTE statistical phasing methods. The model is similar to that used by PHASE.
31. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Statist. Soc. B* **62**, 605–655 (2000).
32. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
33. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
This paper describes PHASE, which has been considered as a gold standard for computational phasing accuracy, although it is too computationally intensive to be applied to large data sets.
34. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
This paper describes fastPHASE, which was one of the first computational phasing methods suitable for genome-wide SNP data.
35. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
36. Celeux, G. & Diebolt, J. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statist. Quart.* **2**, 73–82 (1985).
37. Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A. & Golmard, J. L. A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. *Ann. Hum. Genet.* **68**, 165–177 (2004).
38. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
39. Delaneau, O., Coulonges, C. & Zagury, J. F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
40. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
This paper describes the BEAGLE method for statistical phasing in samples of unrelated individuals.
41. Auton, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19**, 795–803 (2009).
42. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
43. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
44. The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
45. Kenny, E. E. *et al.* Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. *Proc. Natl Acad. Sci. USA* **106**, 13886–13891 (2009).
46. Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124**, 439–450 (2008).
47. Tregouet, D. A. *et al.* Genome-wide haplotype association study identifies the *SLC22A3-LPAL2-LPA* gene cluster as a risk locus for coronary artery disease. *Nature Genet.* **41**, 283–285 (2009).
48. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
49. Browning, S. R. & Browning, B. L. High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* **86**, 526–539 (2010).
50. Hickey, J. M. *et al.* A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.* **43**, 12 (2011).
51. Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A. & Goddard, M. E. Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics* 24 Jun 2011 (doi:10.1534/genetics.111.128082).
52. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
53. Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genet.* **43**, 316–320 (2011).
54. Kruglyak, L., Daly, M. J., ReeveDaly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. H. Genet.* **58**, 1347–1363 (1996).
55. Schaid, D. J., McDonnell, S. K., Wang, L., Cunningham, J. M. & Thibodeau, S. N. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am. J. Hum. Genet.* **71**, 992–995 (2002).
56. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
57. Rohde, K. & Fuerst, R. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum. Mutat.* **17**, 289–295 (2001).
58. Zhang, K., Sun, F. & Zhao, H. HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics* **21**, 90–103 (2005).
59. Abecasis, G. R. & Wigginton, J. E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77**, 754–767 (2005).
60. Zhang, F. & Deng, H. W. Confounding from cryptic relatedness in haplotype-based association studies. *Genetica* **138**, 945–950 (2010).
61. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
62. Andres, A. M. *et al.* Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.* **31**, 659–671 (2007).
63. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84**, 235–250 (2009).
64. Jostins, L., Morley, K. I. & Barrett, J. C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19**, 662–666 (2011).
65. Geraci, F. A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem. *Bioinformatics* **26**, 2217–2225 (2010).
66. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics* **26**, i183–i190 (2010).
67. Long, Q., MacArthur, D., Ning, Z. & Tyler-Smith, C. HI: haplotype improver using paired-end short reads. *Bioinformatics* **25**, 2436–2437 (2009).
68. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
69. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotech.* **29**, 59–63 (2011).
This paper describes the use of an experimental phasing method that was applied to the sequence of an individual and the population-genetic inferences that were made using the phased haplotypes.
70. Suk, E.-K. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* 3 Aug 2011 (doi:10.1101/gr.125047.111).
71. Duitama, J., Huebsch, T., McEwen, G., Suk, E.-K. & Hoehe, M. R. in *Proc. 1st ACM Int. Conf. Bioinf. Comp. Biol.* 160–169 (Association for Computing Machinery, Niagara Falls, New York, 2010).
72. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
73. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotech.* **29**, 51–57 (2011).
74. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods* **7**, 299–301 (2010).
75. Hert, D. G., Fredlake, C. P. & Barron, A. E. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**, 4618–4626 (2008).
76. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
77. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
78. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
79. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
80. Su, S. Y. *et al.* Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics* **26**, 1437–1445 (2010).
81. Li, Z. *et al.* A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn). *Cell Res.* **19**, 519–523 (2009).
82. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Rev. Genet.* **11**, 415–425 (2010).
83. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl Acad. Sci. USA* **108**, 12–17 (2011).
84. The UK IBD Genetics Consortium & The Wellcome Trust Case Control Consortium 2. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nature Genet.* **41**, 1330–1334 (2009).
85. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

Acknowledgements

This study was supported by the US National Institutes of Health (NIH) awards R01HG005701 and R01HG004960. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Wellcome Trust.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Sharon R. Browning's homepage: http://faculty.washington.edu/sguy
Brian L. Browning's homepage: http://faculty.washington.edu/browning
Arlequin: http://cmpg.unibe.ch/software/arlequin3
BEAGLE: http://faculty.washington.edu/browning/beagle/beagle.html
fastPHASE: http://stephenslab.uchicago.edu/software.html
GENEHUNTER: http://linkage.rockefeller.edu/soft/gh
The Genome Analysis Toolkit: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
IMPUTE2: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
MACH: http://www.sph.umich.edu/csg/abecasis/MACH
MERLIN: http://www.sph.umich.edu/csg/abecasis/Merlin
***Nature Reviews Genetics* series on Study Designs:** http://www.nature.com/nrg/series/studydesigns/index.html
PHASE: http://stephenslab.uchicago.edu/software.html
PL-EM: http://www.people.fas.harvard.edu/~junliu/plem
'Read-backed phasing' algorithm: http://www.broadinstitute.org/gsa/wiki/index.php/Read-backed_phasing_algorithm
SHAPE-IT: http://www.griv.org/shapeit

ALL LINKS ARE ACTIVE IN THE ONLINE PDF