Published in final edited form as:
Hum Genet. 2008 December ; 124(5): 439–450. doi:10.1007/s00439-008-0568-7.

Missing data imputation and haplotype phase inference for genome-wide association studies

Sharon R. Browning
Department of Statistics, The University of Auckland

Abstract

Imputation of missing data and the use of haplotype-based association tests can improve the power of genome-wide association studies (GWAS). In this article, I review methods for haplotype inference and missing data imputation, and discuss their application to GWAS. I discuss common features of the best algorithms for haplotype phase inference and missing data imputation in large-scale data sets, as well as some important differences between classes of methods, and highlight the methods that provide the highest accuracy and fastest computational performance.

Keywords

genotype imputation; HapMap; GWAS

Introduction

Genome-wide association studies (GWAS) scan the entire genome for variants that are associated with a trait or disease of interest. These studies are proving to be successful in finding susceptibility loci underlying complex diseases (Altshuler and Daly 2007; Lango and Weedon 2008).

To find disease-associated variants using a case-control genome-wide association design, at least several thousand cases and several thousand controls are typically needed for adequate power (Altshuler and Daly 2007; Wang et al. 2005), while several hundred thousand or more SNPs are needed to cover the human genome adequately (Balding 2006). Thus, GWAS are considerably larger in scale than candidate gene association studies. Consequently, GWAS demand new methods of analysis that are computationally efficient and that make good use of the available data.

One way to improve the power of GWAS is to infer haplotype phase and use a haplotype-based method for association testing, in addition to applying single-marker association testing methods (Browning and Browning 2007a). A haplotype is a sequence of alleles that are on the same physical chromosome (i.e. that are inherited from the same parent). Since the observed genotypes are unordered pairs of alleles, haplotype phase must be inferred. Statistical methods estimate haplotype phase using linkage disequilibrium (LD), which is correlation between neighboring variants. Due to LD, haplotypes can be correlated with other variants within a region. Thus, testing haplotypes can enable one to detect associations with ungenotyped variants.

Address for correspondence: Sharon Browning, Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, e-mail: s.browning@auckland.ac.nz, fax: +649 3737018, phone: +649 3737599 x88745.

Another way to improve the power of GWAS is to use missing data imputation to infer genotypes for known, but ungenotyped, variants. These variants can then be tested for association with the trait (e.g. Scott et al. 2007; The Wellcome Trust Case Control Consortium 2007). At present the imputed variants are usually SNPs genotyped in the HapMap project (The International HapMap Consortium 2007). A third approach to improving power is to combine results across multiple studies, imputing genotypes when SNPs have been genotyped in some, but not all studies (e.g. Lettre et al. 2008; Zeggini et al. 2008).

I begin this review with an overview of the use of haplotype phase inference and missing data imputation in GWAS. This overview includes a discussion of the use of the HapMap for genotype imputation; a discussion of the applications of haplotype phase inference and of missing data imputation in GWAS; and a discussion of the relative merits of haplotype-based association testing versus single-marker association testing of imputed markers. I then give attention to the statistical methodology underlying the major GWAS-applicable methods for haplotype phase inference and missing data imputation. This includes a brief introduction to hidden Markov models and the Expectation-Maximization (EM) algorithm, a discussion of metrics for assessing the quality of results from haplotype inference and imputation methods, and a brief description of the models and statistical methodology underlying the major methods. I summarize results from published comparisons of methods, and discuss the relative merits of the methods.

The HapMap and imputation

A reference panel consists of a number of individuals genotyped at all markers of interest. The reference genotypes or haplotypes can be used to model patterns of variation, thus aiding the imputation of missing data (particularly missing genotypes, but also haplotype phase) in the remaining individuals.

The HapMap data are very well suited for this purpose. Phase II of the HapMap project (The International HapMap Consortium 2007) includes over 3.1 million single nucleotide polymorphisms (SNPs) genotyped on four panels of individuals. These panels are: 30 trios (two parents with one child) of individuals with northern and western European ancestry from the USA (the CEU panel), 30 trios of Yoruban individuals from Ibadan, Nigeria, 45 unrelated individuals from Tokyo, Japan and 45 unrelated Han Chinese individuals from Beijing.

One limitation of the HapMap (and of other resources, such as data from previous GWAS studies) is that it only covers a limited number of ethnicities. To date, most GWAS have used samples from populations with mostly northern and western European ancestry, so that the HapMap CEU panel is a close match. When a reference panel from one ethnicity is used to impute variation in a sample taking from another ethnicity the quality of imputation will be reduced somewhat, although using a pooled reference panel using all available ethnicities can give acceptable results (Chambers et al. 2008). In a balanced study design, in which missing data patterns in cases and controls are roughly the same, and in which cases and controls are drawn from a single population, a mismatch between ethnicities in the reference panel and the genotyped sample is not likely to result increased rates of false positive results. However in an unbalanced situation one might expect an increase in false positive results in a similar manner to that seen in the presence of population stratification (Campbell et al. 2005).

A second limitation of HapMap Phase II data is that the sample size within each ethnicity is quite low. A collection of 30 trios will provide 120 haplotypes. Thus, ability to estimate the haplotypic background of low frequency alleles (particularly those with population frequency < 2%) is extremely limited.

Applications of haplotype phase inference to GWAS

In the context of GWAS, the main application of haplotype phase inference is to enable the use of haplotype-based association methods. To reduce the computational burden for GWAS data, one can base the analysis on a single best estimate of haplotype phase for each individual (Browning and Browning 2007a; Scheet et al. 2007). It is also possible to obtain posterior probabilities of multiple possible haplotype configurations, and to sum over these in the downstream analysis. Haplotypic association analysis of GWAS have been successful in finding associated loci that are not genome-wide significant using single-marker tests (Browning and Browning 2008; Raelson et al. 2007).

An additional application of haplotype phase inference is to phase the reference panel. A number of imputation methods require the reference data to consist of phased haplotypes, although some will accept unphased genotypes. Reference panels comprised of trios (or other closely related individuals) have an advantage over unrelated panels of similar size in that haplotype phase can be inferred much more accurately in trios by using the rules of Mendelian inheritance at each genetic marker as well as the linkage disequilibrium correlation across markers (Marchini et al. 2006).

Applications of imputation to GWAS

Missing data imputation has several applications in the GWAS context. Firstly, one can use imputation to fill in the small proportion of genotypes that fail to pass quality control. Secondly, one can impute genotypes at markers that have not been typed in the study, by using a reference panel. Thirdly, one can use imputation to combine results from two or more studies that have been genotyped on differing sets of markers, again with the help of a reference panel. A fourth application, which is not considered in detail here, occurs in family-based GWAS, when a family member is not available for genotyping (Dudbridge 2008). The first three applications of imputation will be considered in more detail in the following subsections. Table 1 shows an example of each type of data.

Imputing missing genotypes

Table 1A shows an example of missing genotypes. Standard quality control procedures, applied before analysis, remove any individual or marker that has a high proportion of missing data. Thus, each individual and each marker will have a low proportion of missing data. No reference panel is needed to impute the missing genotypes, as the almost complete data from other individuals and the high marker density usually provide sufficient information to impute with high accuracy. Greater than 98% accuracy of imputed genotypes can be achieved in studies with 3000+ individuals genotyped at the density of the Affymetrix 500K array (Browning and Browning 2007b).

Early developers of methods for haplotype phase inference recognized the potential application to filling in missing genotypes (Hawley and Kidd 1995). Thus, this application is not new. However, in the GWAS context it has increasing importance, for two reasons. Firstly, the markers in a GWAS are sufficiently dense and the samples sizes are sufficiently large to allow accurate imputation of missing data. Secondly, due to the high marker density, multilocus association methods can be quite powerful, and multilocus methods typically require complete genotype data (and often haplotype phase as well). When applying a multilocus association method to multiple markers it is very inefficient to discard an individual just because that individual is missing a genotype at one of the markers. A high proportion of individuals will have data missing at one or more of the markers, so discarding these individuals will reduce the sample size substantially. Thus, imputation of missing genotype data is important for

maintaining power of association studies for single marker methods, and even more so for many multilocus methods.

Imputing genotypes at ungenotyped markers using a reference panel

Table 1B shows an example of markers that are not genotyped in the sample but that are genotyped in a reference panel (SNPs 2 and 5). The reference panel is essential if the genotypes at the ungenotyped markers are to be inferred in the sample. Note that the samples may have some missing genotypes at genotyped markers, and there may be some missing genotypes in the reference panel.

This type of imputation is closely related to the concept of tagging. In tagging, one chooses a subset of SNPs (the “tag” SNPs) from a larger set, such that every SNP in the larger set is highly correlated with one or more tag SNPs (Carlson et al. 2004; Johnson et al. 2001), or with a haplotype of tag SNPs (Pe'er et al. 2006). Thus testing tag SNPs for association with a trait should be almost as powerful as testing the larger set of SNPs. When the tagging concept was in early development, it was realised that tag SNPs could also be used to impute the variants that the tag SNPs were tagging, for the purposes of “finer mapping” (Chapman et al. 2003). This idea was extended for application to GWAS by several authors (Marchini et al. 2007; Nicolae 2006; Servin and Stephens 2007). Prior to the development of the HapMap resource, one needed to conduct a pilot study of all known polymorphism in a region, from which one chose tag SNPs. The tag SNPs could then be used as proxies for other known polymorphisms. One can now use HapMap data to select tag SNPs instead of conducting a pilot study, and once genotyping has been performed on one's own sample, one can combine the sample data with HapMap data and impute the remainder of the HapMap SNPs for the sample (The International HapMap Consortium 2007). As well as gaining statistical power by using imputation, one can reduce the cost of GWAS by using smaller (less expensive) arrays, such as the Illumina 300K rather than Illumina 550K array. However, a slightly larger sample size is needed to achieve comparable power when using a smaller array, as imputation is typically less accurate than genotyping (Anderson et al. 2008). In addition to imputation of HapMap SNPs, other types of variation may be imputed, such as the classical HLA alleles (Leslie et al. 2008; Listgarten et al. 2008).

The strategy of imputing HapMap SNPs has been adopted in several GWAS (e.g. Chambers et al. 2008; Scott et al. 2007; The Wellcome Trust Case Control Consortium 2007; Willer et al. 2008; Zeggini et al. 2008). This strategy has been successful in finding associations that would not have been found using only the original genotypes. For example, Zeggini et al. imputed 2.20 million HapMap SNPs in three collections of type 2 diabetes cases and controls (Zeggini et al. 2008). Two of the collections had been genotyped on the Affymetrix 500K GeneChip, while the third had been genotyped on the Illumina 317K chip. This imputation resulted in two significant results that would not have been found using only the original genotypes. One of these was a known association with *PPARG*, while the second was a novel association with CDC123-CAMK1D, which has been confirmed through genotyping in replication samples.

Combining studies from different genotyping platforms

It has become apparent that even large GWAS with several thousand cases and controls are underpowered to find disease susceptibility variants for many common diseases. Increased success can come from combining data across multiple studies to increase sample sizes and thus increase power. A significant challenge in combining results from such studies lies in combining results across studies that have genotyped different marker sets. One approach, used by a group of type 2 diabetes studies (Diabetes Genetics Initiative 2007; Scott et al. 2007; Zeggini et al. 2007), is to simply investigate the top results from each individual study,

performing additional genotyping in the samples from the other studies to attempt to replicate these results. A more powerful approach is to combine the data from all studies. This allows detection of associations that are not among the top hits in any one study, but that show a trend in each component study. In order to do so, the problem of assessing markers genotyped in some component studies but not in others must be addressed, which can be achieved through imputation (Pe'er et al. 2006). Several groups have recently taken this approach, combining data from studies that used different genotyping platforms by using HapMap-based imputation, and have found novel associations (Barrett et al. 2008; Lettre et al. 2008; Willer et al. 2008; Zeggini et al. 2008).

In general, accurate cross-platform imputation requires the use of a reference panel of individuals who have been genotyped on the majority of markers from both platforms. Table 1C shows two samples genotyped on different genotyping platforms, along with a reference panel. Some markers (such as SNP 4) may be genotyped on both platforms. The reference panel is essential unless the degree of overlap between the two genotyping platforms is very high. The HapMap data can serve as the reference panel, or one can make use of panels of control individuals who have been genotyped on several different marker sets, such as the 1958 British Birth Cohort (http://www.b58cgene.sgul.ac.uk), which has been genotyped on the Affymetrix 500K and Illumina 550K platforms.

Haplotypic tests versus testing imputed markers

Multilocus association tests, including haplotype-based association tests, seek to detect association between disease status and variants that have not been directly genotyped. While imputation seeks to test association between known, but ungenotyped, variants and disease status, haplotypic tests seek to test association between observed haplotypic backgrounds and disease status. Every genetic variant originally occurs on a particular haplotypic background, which is modified over time through recombination and mutation. Thus, the observed current haplotypic backgrounds can serve as proxies for such known or unknown genetic variation, particularly for low frequency (i.e. recent) variants. Haplotypic tests may be able to detect association with variants that are not included in any suitable reference panel, which increases their relative usefulness. Also, haplotypic tests may be able to find sets of *cis* interacting variants on the same haplotype background within a gene which will not be found using single marker tests of genotyped or imputed variants (Schaid 2004).

On the other hand, multilocus tests need to be applied very carefully. If haplotypes are tested indiscriminately, there will be a large increase in the amount of variation that is tested, with the need for correspondingly large multiple-testing correction, which reduces power. Sliding window approaches that test every haplotype of a fixed number of markers are particularly poor in this respect. If the window size is too high, too many tests will be applied and any association will be split over multiple haplotypes and will be undetectable. If the window size is too small, information is lost, reducing power. The optimal window size varies from one region to another, and even with optimal window size, it may be possible to cluster the haplotypes further to improve power.

The localized haplotype clustering testing method (Browning and Browning 2007a; Browning 2006) is quite parsimonious in selecting clusters of haplotypes to test for association, avoiding the problems inherent in sliding window approaches. The method was able to identify four novel associations in the Wellcome Trust data set, three of which have strong support from independent studies (Browning and Browning 2008; Zeggini et al. 2008). In contrast, imputation of HapMap SNPs did not result in novel findings in these data. It is worth noting that application of Beagle involved approximately 1.5 million haplotypic tests per disease, which is fewer than the number of tests applied when imputing HapMap variation. Thus, this

result suggests that at present haplotypic analysis can have an edge over analysis of imputed markers.

With the development of reference panels that cover a greater proportion of actual variation, the balance is likely to shift towards imputation. The 1000 Genomes Project (http://www.1000genomes.org/) will sequence the genomes of at least 1000 individuals from around the world, providing a publicly available reference panel that catalogues almost all SNPs and structural variants that have frequencies of 1% and higher.

Overview of methods for imputation and haplotype phase inference

Most methods for haplotype phase inference can also be used to perform imputation. In addition, there are imputation methods that are independent of haplotype phase inference. Although there are a very large number of methods for haplotype phase inference and/or missing data imputation, most of these are too computationally intensive for application to GWAS data. The focus here is on those methods that are most applicable to GWAS analysis. Before describing the methods themselves, some background information on Hidden Markov Models and the Expectation-Maximization (EM) algorithm, which are used by most of the methods, are presented and metrics for assessing accuracy of haplotype phase inference and imputation are discussed.

Hidden Markov models and the EM algorithm

Hidden Markov models (HMMs) are a natural choice of approach for inference of haplotype phase and missing genotypes. In an HMM, an underlying hidden (i.e. unobserved) state generates the observed data (see Rabiner 1989). In the context of haplotype phase and missing genotype inference, the observed data are the observed unphased genotypes (which may include errors and missing data), while the hidden state represents the haplotype phase and the true genotypes.

A Markov model is applied to the hidden states along the chromosome. Markov models have a very simple probabilistic structure that results in a relatively parsimonious model and facilitates computation. The observed data at a marker depend only on the hidden state at that marker (the hidden state is said to “emit” the observed data).

Computation on HMMs is achieved using numerical tricks that exploit the conditional independence structure of the model, and computation times generally increase linearly with the number of markers, and quadratically (or less) with the number of states at each marker. The specialized algorithms that are used are the Viterbi algorithm to find the most likely hidden state paths (i.e. phased haplotypes), the Baum forward-backward algorithm to compute posterior probabilities of hidden states (i.e. probabilities of haplotypes given the genotype data), and the Baum-Welch algorithm to fit model parameters by maximizing the likelihood. Details of these algorithms are given in a tutorial by Rabiner (1989).

The Baum-Welch algorithm is an EM algorithm (Rabiner 1989). EM algorithms iteratively update model parameters to maximize the model likelihood. In the context of haplotype phase and missing data imputation, the iteration usually proceeds as follows. First, one takes an estimate of haplotype phase and missing data values, which can be an arbitrary guess at the first iteration. Using the estimated full data (with haplotype phase and all genotypes), one estimates the other parameters of the model, such as recombination fractions. Then, using the fitted model and original observed genotype data, one re-estimates the haplotype phase and missing data values. These new estimates of haplotype phase and missing data values become the estimates used to initiate the next iteration of the algorithm. Typically, convergence (of measures of accuracy such as switch error or imputation error rates, described below) is

achieved in 10-100 iterations of the algorithm (e.g. Browning and Browning 2007b; Hawley and Kidd 1995; Scheet and Stephens 2006). An EM algorithm can get caught in a local maximum of the likelihood surface. A workaround to the problem of local maxima is to run the algorithm multiple times and use the solution with the highest final likelihood value; however, this does increase computation time.

EM algorithms are used in a frequentist framework, while Bayesian models are typically fit using Markov chain Monte Carlo (MCMC) algorithms. MCMC algorithms attempt to explore the entire model space, rather than simply find a maximum, and generally require tens of thousands of iterations. Thus, they are very computationally intensive, and are not suitable for routine analysis of large genome-wide data sets.

Metrics for assessing results

Two primary metrics are used for assessing accuracy of haplotype phase inference and imputation for large-scale data sets. The switch error rate (Lin et al. 2002; Stephens and Donnelly 2003) is the proportion of successive pairs of heterozygote markers in an individual that are phased incorrectly with respect to each other. This error rate can only be assessed when the true haplotypes are known, for example in simulated data, or when nuclear family data are available. The imputation error rate is the proportion of missing data genotypes that are correctly imputed. This error rate can be assessed in real data by masking (setting to missing) a small proportion of the genotypes, and attempting to impute them. Methods for haplotype phasing that achieve low switch error rates also tend to achieve low imputation error rates, because low switch error rates indicate accurate haplotype phasing, which in turn leads to high-quality imputation.

It is possible to tune the output from haplotype phasing and imputation methods to attempt to minimize these two measures of accuracy. While it is natural to use the “best” haplotypes as output – that is, the haplotypes with the highest likelihood values, which are the output of the Viterbi algorithm (see section on HMMs above) – these are not necessarily the haplotypes that will minimize switch and imputation error. For example, fastPHASE (Scheet and Stephens 2006) attempts to minimize switch error by moving through the heterozygous sites in an individual's genotypes, phasing each heterozygous site relative to the previous one according to the most frequent phasing seen in the sampled haplotype pairs. Imputation error can be minimized by setting the imputed genotype to be the one that maximizes the genotype posterior probability (this is not necessarily the same as maximizing the posterior probability for the entire haplotype pair).

In real data, for which the correct phase and missing genotype values are unknown, one can assess the imputation or phase accuracy by using the variability of the sampled haplotypes or genotypes. These samples can be obtained either from multiple iterations of a single long EM run (Li et al. 2007) or from the final iterations of multiple EM runs (Scheet and Stephens 2006). If the imputed genotypes or phases of successive heterozygote genotypes are almost identical over multiple samples, one can have high confidence in the quality of the imputation or phasing, whereas if the variability is high, the accuracy is likely to be low.

Methods for haplotype phase inference

In this section, I describe the methods for haplotype phase inference that are applicable to GWAS, and the statistical models on which many such methods are based.

Methods based on the Li and Stephens framework

A number of methods suitable for haplotype phase inference on large-scale data sets are based on variants of the "product of approximate conditionals" (PAC) models described in Li and Stephens (2003). I will refer to this family of models as the Li and Stephens framework. In these models, a subset of haplotypes is selected as a reference set, and each reference haplotype represents a (hidden) state of the HMM at each marker. The true haplotypes underlying the observed genotype data are assumed to be imperfect mosaics of the reference haplotypes. Points of change from one reference haplotype to another allow for historical recombination. The observed alleles may differ from the alleles on the underlying true haplotypes to allow for historical mutation and genotype error. As part of the model fitting process, parameters such as historical recombination rates between adjacent markers, and mutation rates may be estimated.

FastPHASE (Scheet and Stephens 2006) uses the Li and Stephens framework, with a fixed number of haplotype clusters in place of reference haplotypes. I will refer to this model as the Scheet and Stephens model. The model parameters, including definitions of the haplotype clusters and recombination and mutation rates are fit using an EM algorithm. By default, fastPHASE v1.2 chooses the optimal number of clusters from the range 5, 10 and 15 using cross-validation. For data sets with large numbers of individuals, the use of a larger number of haplotype clusters (e.g. 20 or 30 clusters rather than the default 5-15) improves imputation and phasing accuracy (Eronen et al. 2006). However, computation time increases quadratically with the number of haplotype clusters used.

Mach (Li et al. 2006; Li et al. 2007) is also based on the Li and Stephens framework. During each EM iteration of the model fitting, the current estimates of haplotype phase are used as the reference haplotypes. One at a time, an individual is removed from the set of reference haplotypes and is updated. The updated pair of haplotypes for the individual is sampled from the posterior probability distribution which is based on the current reference haplotypes. The recombination and mutation rates are estimated at the end of each iteration. In order to reduce the computational burden, one can restrict the number of states. In this case, a random subset of estimated haplotypes is used as the reference pool for each update of an individual. As the computational burden increases quadratically with the number of reference haplotypes, the ability to limit the number of states in this way is essential for data sets with large numbers of individuals. Impute (Marchini et al. 2007) is also based on the Li and Stephens framework, however Impute does not infer haplotype phase, so it is considered later, under methods for missing data imputation.

PHASE (Stephens and Scheet 2005; Stephens et al. 2001) has been considered a gold-standard in the field of haplotype phase inference, as it has achieved excellent results on small data sets (Marchini et al. 2006). PHASE version 2 (Stephens and Scheet 2005) takes a Bayesian approach and uses MCMC to fit parameters of a model based on the Li and Stephen framework. Because it uses MCMC, it involves long computation times. In addition, the current implementation of PHASE (v 2.1) cannot handle more than approximately 100 markers at once, so it must be applied to sliding windows of markers for larger data sets.

Beagle

Beagle (Browning and Browning 2007b) is based on a model that locally clusters haplotypes (Browning 2006). I will refer to this model as the Browning model in this review. In the Browning model, the observed haplotypes are grouped into clusters at each marker position, based on similarity of the haplotypes at markers in the local vicinity. As one moves along the model from one marker to the next, cluster membership tends to stay stable, with some changes due to historical recombination or mutation events.

The Browning model is an HMM, and EM-style updating is used to fit the model in Beagle. There are no explicit parameters such as recombination fractions in the Browning model. Instead, the model is represented by the clusters and by the possible transitions between them, along with the observed frequencies of those transitions. Important differences between the Browning model and the Li and Stephens framework (as implemented in Mach and Impute) are highlighted in Figure 1. Firstly, the number of states at each marker can vary. This allows Beagle to model differing levels of complexity at differing locations, while minimizing the computational burden. Secondly, in the Browning model a hidden state (the localized haplotype cluster) only emits a single type of allele (i.e. haplotypes with different observed alleles at a position cannot be in the same localized haplotype cluster). Thus, mutation is not explicitly modelled, although the states of the model will include any observed mutations. In addition, each state at one marker can transition to at most k states at the next marker, where k is the number of observed alleles for the marker (e.g. $k = 2$ for SNPs). These differences reduce the number of possible paths through the model for a given multilocus genotype, thus reducing the computational burden for each iteration of the EM algorithm. With the Browning model, there is no need to estimate parameters such as mutation rates and recombination rates explicitly, which appears to have the effect of reducing the number of iterations required (relative to Mach and fastPHASE which do estimate such parameters). A final difference is that many haplotype configurations are assigned a probability of zero by the Browning model. For example, the haplotype 111 has probability zero in Figure 1. This difference is necessary to allow the model to be so parsimonious, but means that the haplotype model must be constructed from all sampled individuals, rather than from a subset acting as a reference panel. If an individual's genotypes are not used in the model-building process, it is possible to encounter the situation in which there is no haplotype configuration in the model that is consistent with the individual's genotype. In summary, the Browning model is a much more parsimonious model than the Li and Stephens framework. Thus, there are many fewer parameters to estimate in the Browning model, which results in much faster computation times.

The localized haplotype clustering in the Browning model used by Beagle is conceptually similar to the clusters of the Scheet and Stephens model. One major difference between the clusters in the Browning model and the clusters in the Scheet and Stephens model is that the latter uses a fixed number of clusters, while the former allows the number of clusters to vary from one position to another. Another important difference is that clusters in Beagle are based on the current estimates of the haplotypes rather than on underlying ancestral haplotypes. For example, two haplotypes with different alleles at the current marker position will not be in the same cluster at this position with Beagle, whereas they might be in the same cluster with fastPHASE, if the haplotypes are otherwise very similar at nearby markers.

Other methods for haplotype phase inference

The EM algorithm can also be used directly for haplotype phase and missing data estimation (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). In this case, there is no mediating model, but the frequencies of the haplotypes are estimated directly. The methods can only be applied to small numbers of markers at once, because the haplotype frequencies become too low to be estimated with any accuracy when more than a handful of markers are considered. For large-scale data sets, then, haplotype phasing involves sliding a window along the chromosome, estimating haplotype phases within each window and piecing the fragments together over the whole chromosome. Due to the limitation on the number of markers that can be considered at once, and on the lack of a model to account for historical recombination and mutation, it is unlikely that direct EM algorithms will be able to achieve the accuracy of good model-based methods.

Another successful class of large-scale haplotype phase inference methods is based on piecing together observed haplotype segments, such as in HaploRec (Eronen et al. 2006). HaploRec divides reference haplotypes into fragments. All fragments with observed frequency greater than some threshold are placed in a dictionary. The probability of a haplotype is defined to be the product of corresponding fragment frequencies from the dictionary. As there are multiple ways to construct a given haplotype from the dictionary of fragments, the result is averaged over all such "segmentations". An EM algorithm is used to successively refine the estimated phased haplotypes. HaploRec does not impute missing data as part of the haplotype phasing process, in contrast to the other methods described above. The dictionary model of Ayers et al. (2007) is very similar, however Ayers et al. use MCMC to fit their model. This enables imputation of missing data, but makes the algorithm too slow for application to large data sets.

Consideration of phylogeny can be used in haplotype reconstruction, as in HAP (Halperin and Eskin 2004). This approach assumes no recurrent mutation, and no historical recombination within the window of markers, when building an ancestral tree for the haplotypes underlying the observed genotype data. This is inherently a block-based or window-based approach, which has some disadvantages, as mentioned above for the direct EM approaches.

There are a great many papers describing other methods for haplotype phase inference. Typically, these methods show excellent accuracy on small data sets (a handful of markers and fewer than 100 individuals) but have not been shown to have good accuracy for large data sets, such as those found in GWAS. In addition, most of these methods are computationally costly, and would be extremely difficult to apply to on a genome-wide scale.

Methods for missing data imputation

Except for HaploRec, all of the above haplotype phasing methods impute missing data as part of the process of inferring haplotype phase. In addition, Mach and fastPHASE have options to only impute missing data (i.e. not infer haplotype phase), which reduces computational time somewhat. Mach and fastPHASE have options to specify that the phase of some individuals is known – this is useful when including reference data that has been accurately phased with the use of data on related individuals, such as HapMap trios (The International HapMap Consortium 2005). Beagle version 3.0 also allows for use of a phase known reference panel (B.L. Browning and S.R. Browning, unpublished data; software available on request).

Two imputation-specific methods that are based on the Li and Stephens framework (Li and Stephens 2003) are Impute (Marchini et al. 2007) and Bim-Bam (Servin and Stephens 2007). Impute uses a panel of phased reference haplotypes to build a model, and requires user-specification of recombination rates and mutation rates. Thus, it avoids the need for an iterative model-building approach, but it may be sensitive to misspecification of model parameters, and does not utilize information contained in the other individuals on whom imputation will be performed. Bim-Bam uses fastPHASE to perform the imputation, but adds new methodology for using the imputed values in association testing. Missing data are imputed multiple times, with the imputed values being used in a Bayesian regression approach to test for association. It is beyond the scope of this review to discuss specific statistical techniques for using imputed genotypes in association testing, so we do not consider Bim-Bam in further detail here.

Rather than using all markers (on a chromosome, or within a large window) as potential predictors of genotype via phased haplotypes, several approaches use small sets of genotyped markers (usually tag SNPs). The predictors of genotype can be regression equations based on tag SNPs (Chapman et al. 2003), specific two or three marker haplotypes (Pe'er et al. 2006), or weighted averages of haplotype indicators (Lin et al. 2008; Nicolae 2006; Zaitlen et al. 2007).

In addition to specialized genetics-based approaches, one can use standard statistical techniques for imputing missing data, such as linear regression with variable selection, regression trees, or k-nearest neighbor methods. Yu and Schaid (2007) reviewed a number of these methods, and compared them to fastPHASE on masked HapMap data. They found that fastPHASE provided better results (gave lower imputation error rates) than any of the non-genetic methods.

Accounting for uncertainty in imputation

The degree of accuracy that can be achieved when estimating ungenotyped markers varies greatly depending on the extent of LD between the ungenotyped marker and nearby genotyped markers. Ungenotyped markers that are in low LD with nearby markers, or, equivalently, for which the estimated accuracy is low, are usually discarded rather than being carried forward into association analysis. For example, the study of Scott et al. (2007) imputed genotypes in 2335 individuals for 2.15 million HapMap SNPs with minor allele frequency > 1% in Caucasians that were not included on the Illumina 300K panel. In this study, 0.06 million imputed SNPs (3%) failed to have sufficiently high estimated accuracy, and were removed from the analysis. Nonetheless, there will be some uncertainty around the remaining estimated (imputed) missing genotypes that should not be ignored.

Lin et al. (2008) recommend using likelihood-based methods to integrate over uncertain haplotype phase and missing data values when imputing and testing ungenotyped variants. This avoids the potential loss of power inherent in the two-stage approach of imputing variants then using the imputed values (without accounting for uncertainty in these values) in the association tests. Most imputation methods provide posterior probabilities for imputed genotypes, which allows for accounting of uncertainty without taking the full likelihood approach. It is beyond the scope of this review to discuss the best ways to use imputed values in association tests, however various approaches have been described (Lin et al. 2008; Marchini et al. 2007; Nicolae 2006; Servin and Stephens 2007).

Whichever approach is used, tests based on ungenotyped variants are subject to the same problems as those on actual genotypes, such as effects of population stratification, and differential rates of missing data and genotyping error in cases and controls. Genotyping problems at a single marker can adversely affect imputation at multiple nearby imputation positions. Thus, replication of imputation-based results should always include actual genotyping of the implicated markers on a separate set of cases and controls.

Comparisons of methods

Both error rates and computing times need to be considered when assessing the performance of competing methods. The relative performance of the methods differs greatly as a function of sample sizes, marker densities and computing parameters such as the number of EM iterations. As a general rule, for both haplotype phase inference and missing data imputation, the larger the sample size, the more hidden states (e.g reference haplotypes in the Li and Stephens framework, or haplotype clusters in the fastPHASE and Beagle models) are needed in the model to achieve optimal performance. This is not surprising, as a larger sample size means a greater number of observed haplotypes, which can be better captured by greater model complexity.

Eronen et al. (2006) found that EM methods based on direct estimation of haplotype frequencies, such as PL-EM (Qin et al. 2002) are less accurate than model-based methods such as HaploRec, fastPHASE and PHASE on large data sets. Browning and Browning (2007b) showed that Beagle is faster and more accurate than HaploRec and fastPHASE on very large data sets (thousands of individuals and a density of at least 1 SNP per 3 kb). See Browning and

Browning (2007b) and Eronen et al. (2006) for comparisons of other methods on large sample sizes. Also see Marchini et al. (2006) for results on smaller sample sizes.

Although simulation results in Lin et al. (2008) suggest that the tag-based maximum likelihood approach is more powerful than a haplotype-based imputation approach using fastPHASE, the simulations included only five SNPs. With more markers, fastPHASE could obtain improved estimates of haplotype phase, and thus increase the accuracy of the imputation. Thus, these simulations are too limited to allow valid comparison of the relative performance of Lin et al.'s method and fastPHASE.

Conclusions

This paper has reviewed the best methods for haplotype phase inference and missing data imputation, and their application to GWAS. The best haplotyping methods in this context differ from those suggested in an earlier comparison (Marchini et al. 2006) with smaller sample sizes, because of the computational demands of whole genome data, and because methods that provide most accurate inference on data sets with small numbers of individuals do not necessarily provide the most accurate inference on larger data sets.

Missing data imputation is a new and exciting way to improve the power of GWAS. By means of a reference panel, one can impute ungenotyped SNPs and/or combine studies genotyped on different platforms. The development of larger reference panels, such as the 1000 Genomes Project, with more individuals and more variants, will make this approach increasingly useful. However, haplotype-based multilocus analysis (Browning and Browning 2007a) should not be neglected as a complement to single marker analysis.

For large sample sizes (>1000), Beagle has an advantage over other haplotype-phasing programs, in that its parsimonious modelling scales well to such large data sets while other methods have to be scaled back for computational reasons. For smaller sample sizes (100 individuals), for which computing times are not as significant, I have found that Mach provides excellent results, providing better accuracy than other methods such as fastPHASE and Beagle (S.R. Browning, unpublished results).

I have not directly compared the imputation-only methods with imputation using the haplotype-phasing methods. Some of the imputation methods do not actually output imputed genotypes, but only output the final results of testing for association (for example, Lin et al. 2008). These methods need to be compared in terms of power to detect association rather than accuracy of imputation. Of the available imputation methods, Impute and Mach have both been used for imputation in GWAS, and have yielded similar results (Barrett et al. 2008).

What makes for a good method for haplotype phasing and missing data inference for GWAS? Since the data sets are so large, they contain a lot of information. A careful balance must be maintained in the level of modelling that is applied. With low levels of modelling, such as in the direct EM methods, only a small number of markers can be considered simultaneously. This reduces the amount of information that can be extracted from the data. Careful selection of the markers that are used to provide information (as in Lin et al. 2008) helps, but may still be sub-optimal. Overly stringent modelling can also be disadvantageous with large data sets, as the data contain a lot of information that can be partially lost if an inadequate model is strongly imposed. Thus with large data sets, the data should be allowed to speak for themselves to quite an extent. Several of the more successful methods are very empirical. Mach and Impute use estimated or observed haplotypes directly as reference haplotypes, with other haplotypes being imperfect mosaics of these haplotypes. Beagle constructs a parsimonious model based on the estimated haplotypes, and allows only for certain mosaics of these haplotypes. To achieve the balance of over-modelling versus under-modelling, an extremely useful approach

is to recognize that haplotypes will tend to be locally similar to one another, with patterns of changes following those expected from the biological processes of recombination and mutation (Li and Stephens 2003). This principle underlies the models that are used by the most successful haplotype phasing algorithms, such as models based on the Li and Stephens framework, HaploRec's segmentation model and the Browning model.

Beyond modelling, the implementation of the method must be computationally efficient, and the method must have good convergence properties if it is iterative (as most of the methods are). HMMs facilitate efficient sampling of new haplotype estimates given the current model, and are thus extremely useful. The models based on the Li and Stephens framework and the Browing model are examples of HMMs. EM-style iterative methods, such as those used by Mach, fastPHASE, HaploRec and Beagle require much less computation than methods based on MCMC (such as PHASE). It is my opinion that MCMC is too slow for GWAS data sets, and that MCMC will not be able to provide useful solutions to the problem of haplotype phase estimation for large-scale data. Currently, EM algorithms are widely used, whereas other types of iterative maximization, such as variants of Newton's method (for example, Dudbridge 2008), are rare, however this may indicate fashion rather than inherent advantages of the EM approach.

Acknowledgments

The author thanks Brian Browning for helpful discussions and the anonymous reviewers for their comments. This work was supported by NIH grant 3R01GM075091-02S1.

Literature Cited

Altshuler D, Daly M. Guilt beyond a reasonable doubt. Nat Genet 2007;39:813–5. [PubMed: 17597768]

Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet 2008;83:112–9. [PubMed: 18589396]

Ayers KL, Sabatti C, Lange K. A dictionary model for haplotyping, genotype calling, and association testing. Genetic Epidemiology 2007;31:672–683. [PubMed: 17487885]

Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–91. [PubMed: 16983374]

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–62. [PubMed: 18587394]

Browning BL, Browning SR. Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. Genetic Epidemiology 2007a;31:365–375. [PubMed: 17326099]

Browning BL, Browning SR. Haplotypic analysis of Wellcome Trust Case Control Consortium data. Human Genetics 2008;123:273–280. [PubMed: 18224336]

Browning SR. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 2006;78:903–13. [PubMed: 16685642]

Browning SR. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 2008;178:2123–2132. [PubMed: 18430938]

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 2007b;81:1084–1097. [PubMed: 17924348]

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. Demonstrating stratification in a European American population. Nat Genet 2005;37:868–72. [PubMed: 16041375]

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74:106–20. [PubMed: 14681826]

Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 2008;40:716–8. [PubMed: 18454146]

Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. Human Heredity 2003;56:18–31. [PubMed: 14614235]

Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331–6. [PubMed: 17463246]

Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Human Heredity 2008;66:87–98. [PubMed: 18382088]

Eronen L, Geerts F, Toivonen H. HaploRec: efficient and accurate large-scale reconstruction of haplotypes. BMC Bioinformatics 2006;7:542. [PubMed: 17187677]

Excoffier L, Slatkin M. Maximum-Likelihood-Estimation of Molecular Haplotype Frequencies in a Diploid Population. Molecular Biology and Evolution 1995;12:921–927. [PubMed: 7476138]

Halperin E, Eskin E. Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics 2004;20:1842–9. [PubMed: 14988101]

Hawley ME, Kidd KK. Haplo - a Program Using the Em Algorithm to Estimate the Frequencies of Multisite Haplotypes. Journal of Heredity 1995;86:409–411. [PubMed: 7560877]

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233–7. [PubMed: 11586306]

Lango H, Weedon MN. What will whole genome searches for susceptibility genes for common complex disease offer to clinical practice? J Intern Med 2008;263:16–27. [PubMed: 18088250]

Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet 2008;82:48–56. [PubMed: 18179884]

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann HE, Mohlke KL, Hirschhorn JN. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 2008;40:584–91. [PubMed: 18391950]

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 2003;165:2213–2233. [PubMed: 14704198]

Li, Y.; Ding, J.; Abecasis, GR. Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference [abstract 2290]. Presented at the annual meeting of the American Society of Human Genetics; October 9-13, 2006; New Orleans, Louisiana. 2006. Available from http://www.ashg.org/genetics/ashg06s/index.shtml

Li, Y.; Willer, CJ.; Ding, J.; Scheet, P.; Abecasis, GR. In Silico Genotyping for Genome-Wide Association Studies [abstract 2071]. Presented at the annual meeting of the American Society of Human Genetics; October 23-27, 2007; San Diego, California. 2007. Available from http://www.ashg.org/genetics/ashg07s/index.shtml

Lin DY, Hu Y, Huang BE. Simple and efficient analysis of disease association with missing genotype data. Am J Hum Genet 2008;82:444–52. [PubMed: 18252224]

Lin S, Cutler DJ, Zwick ME, Chakravarti A. Haplotype inference in random population samples. Am J Hum Genet 2002;71:1129–37. [PubMed: 12386835]

Listgarten J, Brumme Z, Kadie C, Xiaojiang G, Walker B, Carrington M, Goulder P, Heckerman D. Statistical Resolution of Ambiguous HLA Typing Data. PLoS Computational Biology 2008;4:e1000016. [PubMed: 18392148]

Long JC, Williams RC, Urbanek M. An E-M Algorithm and Testing Strategy for Multiple-Locus Haplotypes. American Journal of Human Genetics 1995;56:799–810. [PubMed: 7887436]

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 2006;78:437–50. [PubMed: 16465620]

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 2007;39:906–913. [PubMed: 17572673]

Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol 2006;30:718–27. [PubMed: 16986160]

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 2006;38:663–7. [PubMed: 16715096]

Qin ZS, Niu T, Liu JS. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 2002;71:1242–7. [PubMed: 12452179]

Rabiner LR. A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 1989;77:257–286.

Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, Bradley WE, Croteau P, Nguyen-Huu Q, Segal J, Debrus S, Allard R, Rosenstiel P, Franke A, Jacobs G, Nikolaus S, Vidal JM, Szego P, Laplante N, Clark HF, Paulussen RJ, Hooper JW, Keith TP, Belouchi A, Schreiber S. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A 2007;104:14747–52. [PubMed: 17804789]

Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol 2004;27:348–64. [PubMed: 15543638]

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 2006;78:629–44. [PubMed: 16532393]

Scheet, P.; Stephens, M.; Abecasis, GR. Whole Genome Linkage Disequilibrium Association Mapping of Binary Traits [abstract 209]. Presented at the annual meeting of the American Society of Human Genetics; October 23-27, 2007; San Diego, California. 2007. Available from http://www.ashg.org/genetics/ashg07s/index.shtml

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007;316:1341–5. [PubMed: 17463248]

Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 2007;3:e114. [PubMed: 17676998]

Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73:1162–9. [PubMed: 14574645]

Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 2005;76:449–62. [PubMed: 15700229]

Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–89. [PubMed: 11254454]

The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–320. [PubMed: 16255080]

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61. [PubMed: 17943122]

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78. [PubMed: 17554300]

Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005;6:109–18. [PubMed: 15716907]

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008;40:161–9. [PubMed: 18193043]

Yu Z, Schaid DJ. Methods to impute missing genotypes for population data. Hum Genet 2007;122:495–504. [PubMed: 17851696]

Zaitlen N, Kang HM, Eskin E, Halperin E. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 2007;80:683–91. [PubMed: 17357074]

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638–45. [PubMed: 18372903]

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336–41. [PubMed: 17463249]

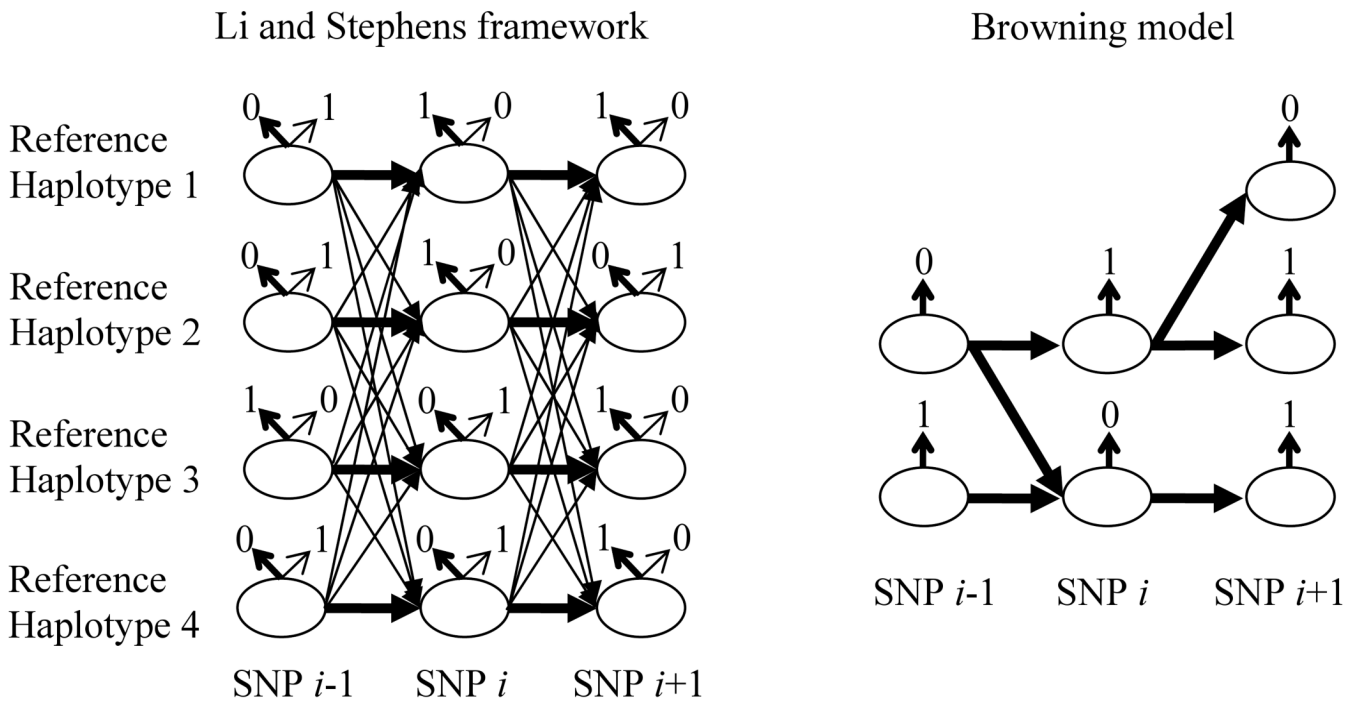

Figure 1.

Illustration highlighting major differences between models based on the Li and Stephens framework (2003) and the Browning model (Browning 2006). Excerpts of the models covering three markers (SNPs *i*-1, *i* and *i*+1) are shown. Ovals are hidden states of the models. For the Li and Stephens framework, these states are defined by the reference haplotypes, while for the Browning model the states are localized clusters of haplotypes. Note that the graphical representation of the Browning model is that given in Browning (2008), while earlier representations had states as edges rather than as nodes of the graph. The Browning model will tend to have fewer states at any given marker than will unconstrained models based on the Li and Stephens framework, and the number of states can vary from marker to marker for the Browning model but is fixed in the Li and Stephens framework. Arrows between states from one SNP to the next are transitions of the HMM. For the Li and Stephens framework, transitions with highest prior probability (those seen in the reference haplotypes) are shown with bold arrows, while thin arrows allow for historical recombination. For the Browning model, there are at most *k* transitions coming out of a state, where *k* is the number of alleles at the next marker (i.e. 2 for SNPs), which helps to keep the model parsimonious. Arrows coming out of the top of the states are possible emissions of the HMM, which are the observed alleles. For the Li and Stephens framework, emissions with highest prior probability (the alleles on the reference haplotypes) are shown with bold arrows, while thin arrows represent mutations to other alleles. The reference haplotypes here are 011, 010, 101 and 001. For the Browning model, there is only one possible emission from each state, which helps to keep the model parsimonious. The models shown are illustrative only. The actual form of the Browning model will vary depending on the alleles of the reference haplotypes outside this window of markers.

Table 1

Examples of three types of data for imputation

A. Missing genotypes						
Individual	**SNP 1**	**SNP 2**	**SNP 3**	**SNP 4**	**SNP 5**	**SNP 6**
Sample 1	AC	TT	AA	CC	AT	GG
Sample 2	CC	--	GG	GG	AA	GG
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sample N	AA	TC	AG	--	AA	TT
B. Ungenotyped markers and a reference panel						
Individual	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Sample 1	AC	--	AA	CC	--	TG
Sample 2	CC	--	GG	GG	--	--
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sample N	AA	--	AG	CC	--	TG
Reference 1	AC	TT	AG	CC	AT	GG
Reference 2	CC	CC	GG	--	TT	TG
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Reference M	AA	TC	AG	CG	AA	TT
C. Two studies genotyped on different platforms and a reference panel						
Individual	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Study 1 Sample 1	AC	--	AA	CC	--	TG
Study 1 Sample 2	CC	--	GG	GG	--	--
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Study 1 Sample N_1	AA	--	AG	CC	--	TG
Study 2 Sample 1	--	TC	--	GG	TT	--
Study 2 Sample 2	--	TT	--	GG	AT	--
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Study 2 Sample N_2	--	CC	--	CG	AA	--
Reference 1	AC	TT	AG	CC	AT	GG
Reference 2	CC	CC	GG	--	TT	TG
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Hum Genet. Author manuscript; available in PMC 2009 December 1.

A. Missing genotypes

Individual	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Reference M	AA	TC	AG	CG	AA	TT

Missing data is denoted --.