BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 84–91
doi:10.1093/bioinformatics/bts632

Genetics and population analysis

Advance Access publication October 23, 2012

Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold

Androniki Menelaou[1] and Jonathan Marchini[1,2,*]

[1]Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom and [2]Wellcome Trust Centre for Human Genetics, Oxford, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Given the current costs of next-generation sequencing, large studies carry out low-coverage sequencing followed by application of methods that leverage linkage disequilibrium to infer genotypes. We propose a novel method that assumes study samples are sequenced at low coverage and genotyped on a genome-wide microarray, as in the 1000 Genomes Project (1KGP). We assume polymorphic sites have been detected from the sequencing data and that genotype likelihoods are available at these sites. We also assume that the microarray genotypes have been phased to construct a haplotype scaffold. We then phase each polymorphic site using an MCMC algorithm that iteratively updates the unobserved alleles based on the genotype likelihoods at that site and local haplotype information. We use a multivariate normal model to capture both allele frequency and linkage disequilibrium information around each site. When sequencing data are available from trios, Mendelian transmission constraints are easily accommodated into the updates. The method is highly parallelizable, as it analyses one position at a time.

Results: We illustrate the performance of the method compared with other methods using data from Phase 1 of the 1KGP in terms of genotype accuracy, phasing accuracy and downstream imputation performance. We show that the haplotype panel we infer in African samples, which was based on a trio-phased scaffold, increases downstream imputation accuracy for rare variants (R2 increases by >0.05 for minor allele frequency <1%), and this will translate into a boost in power to detect associations. These results highlight the value of incorporating microarray genotypes when calling variants from next-generation sequence data.

Availability: The method (called MVNcall) is implemented in a C++ program and is available from http://www.stats.ox.ac.uk/~marchini/#software.

Contact: marchini@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 19, 2012; revised on September 27, 2012; accepted on October 17, 2012

1 INTRODUCTION

Genome-wide association studies (GWAS) are known to be well powered for the detection of common variants that increase disease risk, and have uncovered many replicated associations in recent years. However, the effect sizes of many associated loci are relatively small and together explain only a small proportion of the heritability for many diseases and traits. It has been argued that rare variants with larger effects might harbour the remaining genetic variability (Manolio *et al.*, 2009). Next-generation sequencing allows whole-genome sequencing of a large sample of individuals so that a more detailed picture of human polymorphism can be obtained. Projects such as the 1KGP (The 1000 Genomes Project Consortium, 2012) are working to produce a catalogue of polymorphisms with low (minor allele frequency [MAF] $\in$ [0.005, 0.05]) and rare (MAF $<$ 0.005) allele frequencies.

The current analysis pipeline starts by *detecting* sites exhibiting evidence of polymorphism. Genotypes are then *called* at the polymorphic sites, and then haplotypes are estimated across all such sites, often as a by-product of calling the genotypes. The sets of haplotypes produced can be used as reference panels for imputation into GWAS (Marchini and Howie, 2010) and allow investigations of rare variant associations that go beyond previous imputation analyses from less complete reference panels such as HapMap.

Given the current costs for sequencing, it is still not viable to deeply sequence a large number of individuals, so given a fixed amount of resources, there is a trade-off between sample size and coverage. The consensus view (Li *et al.*, 2011; Kim *et al.*, 2010; The 1000 Genomes Project Consortium, 2012) favours low-coverage sequencing of large numbers of samples. For example, the 1KGP will be sequencing ~2500 individuals at ~4×, the UK10K project will sequence 4000 whole genomes at ~6× (http://www.uk10k.org/), and the Genome of the Netherlands is sequencing 750 individuals at ~12× (http://www.nlgenome.com/).

The current paradigm for detecting, genotyping and phasing polymorphic sites from low-coverage sequence data starts by mapping sequence reads to a reference genome. Mapped reads that overlap a given site (s) in a single individual (i) are then combined together to form *genotype likelihoods* (GLs). GLs are the probabilities of observing the reads given the underlying (unknown) genotype and can be written as $P(R_{is}|G_{is})$, where R_{is} and G_{is} denote the reads at site s and the true underlying genotype, respectively. Methods exist for calculating these GLs (Li *et al.*, 2009) and involve combining information about the bases observed in reads, while allowing for the possibility of sequencing errors via base and mapping quality scores. So, for example, if the set of reads only contains the A allele, then the GL $P(R_{is}|G_{is} = AA)$ will be larger than the other likelihoods, but other genotypes will still have non-zero likelihood. As sequence

*To whom correspondence should be addressed.

coverage increases, the likelihoods should become more 'peaked' around the true genotype. In low-coverage sequencing, there may be no reads spanning a site, in which case, the GLs will be identical across all possible genotypes.

Detecting polymorphic sites involves combining information across individuals at a site to infer whether there are at least two alleles observed across all individuals in the sample (Li *et al*., 2009). Once a site is detected, the site's genotypes can be called using a missing data likelihood that is optimised via an EM algorithm (Li *et al*., 2009), but this will only work well when coverage is high and the GLs contain very good information about the genotypes of all samples. When coverage is low, power to call genotypes can be gained by taking advantage of linkage disequilibrium (LD) between sites in close proximity. Methods that do this are extensions of phasing and imputation algorithms that pool information across samples and sites to infer multi-site genotypes and their underlying haplotypes (e.g IMPUTE2 [Howie *et al*., 2009], Beagle [Browning and Browning, 2009] and MACH [Li *et al*., 2010]).

In some studies, such as GWAS, the individuals being sequenced will also have been genotyped on a genome-wide single nucleotide polymorphism (SNP) chip. For example, in the 1KGP, individuals are both sequenced and genotyped on the Illumina OMNI2.5M chip. Some commercial vendors also return GWAS array data with sequencing results, so this design may become more prominent in the future. Up to now, chip genotypes have primarily been used to validate sequencing-based genotype calls, as the error rates of SNP chips are generally low (<0.5%) (O'Connell and Marchini, 2012).

In this article, we present a method that uses SNP chip genotypes to aid the process of genotype calling from sequencing reads. The first step of our method involves estimating haplotypes at the chip SNPs using an accurate phasing method (Delaneau *et al*., 2011) to form a *haplotype scaffold*. We then call genotypes at a polymorphic site using a model of the LD between the site and the surrounding sites in the haplotype scaffold. We use an Markov Chain Monte Carlo (MCMC) algorithm that iteratively updates the unobserved alleles at a site using the GLs and the local LD. We use a multivariate normal model to capture both allele frequency and LD information around each site that results in fast MCMC updates. When sequencing data are available from mother–father–child trios, Mendelian transmission constraints are easily accommodated into the updates. The method analyses one position at a time and results in a highly parallelizable method that can be applied to a large sample. We call our method MVNcall.

Our approach is similar to the QCALL approach (Le and Durbin, 2011), which also analyses one SNP at a time using a haplotype scaffold. QCALL builds approximate coalescent trees at each site and then infers phased genotypes by considering possible mutations on branches of the trees. This approach does not scale well as the number of sequences increases and is the reason why this method was not applied to the Phase 1 of the 1KGP. MVNcall is much faster and has enabled us to analyse the Phase 1 dataset.

In the sections that follow, we present in detail the underlying model developed for genotype calling and phasing. We apply our method and other methods to chromosome 20 of the Phase 1 of the 1KGP. We carry out an imputation analysis using the haplotype sets inferred by the different methods as reference panels, and we compare imputation accuracy.

2 METHODS

We assume that sequencing reads on M individuals have been used to detect a set of K polymorphic sites, with each site exhibiting two alleles. We use 0 and 1 to denote the reference and non-reference allele, respectively. For each site s we assume that the three GLs $P(R_{is}|G_{is}=0)$, $P(R_{is}|G_{is}=1)$ and $P(R_{is}|G_{is}=2)$ have been pre-calculated. Our method processes each of the K sites in turn. We estimate the phased genotypes at a single site (s) using all the GLs and the haplotype scaffold at D SNPs either side of s. We use H to denote the phased haplotype scaffold on the M individuals, where $H=\{h_{11},h_{12},\ldots,h_{M1},h_{M2}\}$ and h_{ij} denotes the jth haplotype of the ith individual and is a sequence of $2D$ alleles $h_{ij}=\{h_{ij1}\ldots,h_{ij2D}\}$ with $h_{ijl}\in\{0,1\}$. We use (h^s_{i1},h^s_{i2}) to denote the (unobserved) alleles of the ith individual at site s so that the (unobserved) genotype $G_{is}=h^s_{i1}+h^s_{i2}$. The purpose of our method is to infer these unobserved alleles in all M individuals so that we can construct the *complete* haplotypes of each individual at the $L=2D+1$ sites, which we denote as $h^*_{ij}=\{h_{ij1},\ldots,h_{ijD},h^s_{ij},h_{ij(D+1)},\ldots,h_{ij2D}\}$.

We have developed an MCMC algorithm to infer the unobserved alleles. Starting from an initial configuration of the alleles at site s, we cycle through the individuals in random order and update their two unobserved alleles. For the ith individual, the alleles are updated conditional on (a) the reads R_{is}, (b) the haplotypes of that individual h_{i1} and h_{i2}, and (c) the *complete* haplotypes of all individuals *except* individual i, which we denote H^*_{-i}.

More specifically, the conditional distribution we use can be written as

$$P((h^s_{i1},h^s_{i2})|R_{is},h_{i1},h_{i2},H^*_{-i})\propto P(R_{is}|G_{is})P(G_{is}|H^*_{-i},h_{i1},h_{i2}) \quad (1)$$

which makes the reasonable assumption that the reads R_{is} are conditionally independent of H^*_{-i}, h_{i1} and h_{i2} given G_{is}. Further, we specify that

$$P(G_{is}|H^*_{-i},h_{i1},h_{i2})=P(h^s_{i1}|h_{i1},H^*_{-i})P(h^s_{i2}|h_{i2},H^*_{-i}) \quad (2)$$

which makes another reasonable assumption that the two alleles (h^s_{i1},h^s_{i2}) are conditional independent given H^*_{-i} and their respective flanking haplotypes h_{i1} and h_{i2}. Figure 1 shows a pictorial representation of this update. This approach of constructing a Gibbs sampler by directly specifying conditional distributions is now standard in the field of haplotype estimation. See Appendix A in the article by Stephens and Donnelly (2003) for a discussion of why this approach is valid.

2.1 Modelling LD using a multivariate normal distribution

The only term of the model we need to specify is $P(h^s_{ij}|h_{ij},H^*_{-i})$, and this can be rewritten as

$$P(h^s_{ij}|h_{ij},H^*_{-i})\propto P(h^s_{ij},h_{ij}|H^*_{-i})=P(h^*_{ij}|H^*_{-i}) \quad (3)$$

where $P(h^*_{ij}|H^*_{-i})$ is the conditional distribution of a single haplotype h^*_{ij} given an observed set of other haplotypes H^*_{-i}. It is this distribution that models the LD between the alleles at site s and the sites in the haplotype scaffold. We use a multivariate normal model such that

$$P(h^*_{ij}|H^*_{-i})\sim N_L(\hat{\mu},\hat{\Sigma}). \quad (4)$$

We set the mean and covariance to be

$$\hat{\mu}=(1-\theta)\mu+\frac{\theta}{2}1_L \quad (5)$$

$$\hat{\Sigma}=\Sigma+\lambda I_L \quad (6)$$

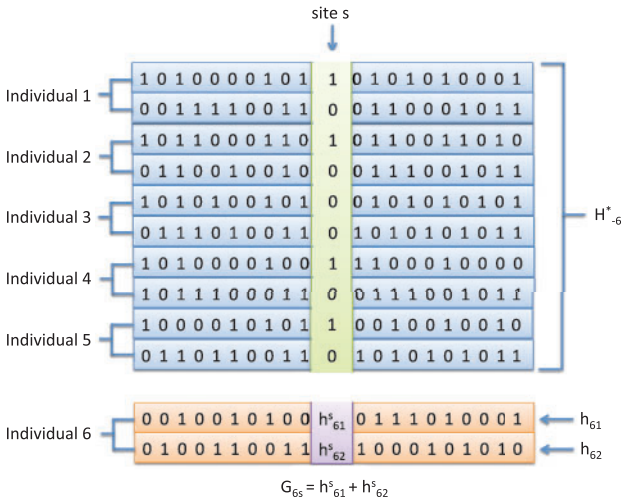

Fig. 1. A pictorial representation of a single update of our MCMC algorithm. In this example, there are six individuals with individual 6 being updated. The two unobserved alleles at the site s of individual 6 are denoted (h^*_{61}, h^*_{62}) (purple). These are flanked by the haplotypes (h_{61}, h_{62}) for individual 6 from the haplotype scaffold (orange). The set of complete haplotypes of the other five individuals (H^*_{-6}) are highlighted. These haplotypes combine the remaining haplotypes from the haplotype scaffold (blue) and the current estimates of the unobserved alleles in the five individuals (green). The genotype likelihoods for unobserved genotype $G_{6s} = h^*_{61} + h^*_{62}$ are shown below the plot. Our method updates (h^*_{61}, h^*_{62}) by sampling alleles using Eq. 1. We iterate this scheme over all individuals in random order. Updating all individuals once represents a single iteration

where 1_L is a L-vector of 1s, I_L is the L-dimensional identity matrix, and μ and Σ are the sample mean and covariance matrix of the complete haplotypes H^*_{-i} given by

$$\mu = \frac{1}{2M-2} \sum_{k=\{1,2\}} \sum_{m=1, m\neq i}^{M} h^*_{mk}, \tag{7}$$

$$\Sigma = \frac{1}{2M-2} \sum_{k=\{1,2\}} \sum_{m=1, m\neq i}^{M} (h^*_{mk} - \mu)(h^*_{mk} - \mu)^T. \tag{8}$$

The sample mean μ is the vector of allele frequencies at the L sites in H^*_{-i} and θ is a small constant defined as in the article by Wen and Stephens (2010). The sample covariance Σ captures LD information between the L sites, but because sites can be in perfect LD, we adjust the covariance by a small amount (λI_L) (Tychonoff, 1943). This ensures that the covariance matrix $\hat{\Sigma}$ is invertible, which is an operation we need to carry out when using this model. These adjustments essentially capture the idea that a given haplotype may contain unobserved variation not seen in H^*_{-i}.

Good computational efficiency of the MCMC updates can be achieved by taking advantage of the fact that the haplotype scaffold is fixed and not updated at each iteration. Our use of a multivariate normal model implies that the conditional distribution in Eq. 3 is a univariate normal distribution given by

$$P(h^s_{ij}|h_{ij}, H^*_{-i}) \sim N(\hat{\mu}^s_c, \hat{\Sigma}^s_c). \tag{9}$$

The mean $\hat{\mu}^s_c$ and variance $\hat{\Sigma}^s_c$ can be calculated using

$$\hat{\mu}^s_c = \hat{\mu}_s + \hat{\Sigma}_{s,-s} \hat{\Sigma}^{-1}_{-s,-s} (h_{ij} - \hat{\mu}_{-s}), \tag{10}$$

$$\hat{\Sigma}^s_c = \hat{\Sigma}_{s,s} - \hat{\Sigma}_{s,-s} \hat{\Sigma}^{-1}_{-s,-s} \hat{\Sigma}_{-s,s}, \tag{11}$$

where we use notation $\hat{\mu}_s$ to denote the element of $\hat{\mu}$ corresponding to the site s that is being phased and $\hat{\mu}_{-s}$ to denote all the other L–1 elements of $\hat{\mu}$ that correspond to sites in the haplotype scaffold flanking site s. Similarly, we use $\hat{\Sigma}_{s,s}$ to denote the marginal variance of site s, $\hat{\Sigma}_{-s,-s}$ to denote the covariance of the L–1 haplotype scaffold sites and $\hat{\Sigma}_{s,-s} = \hat{\Sigma}^T_{-s,s}$ to denote the vector of L–1 covariances between site s and the sites in the haplotype scaffold flanking site s. The key point here is that the inverse $\hat{\Sigma}^{-1}_{-s,-s}$ and mean vector μ_{-s} need only be calculated once at the beginning of the algorithm, and this results in efficient updates. Moreover, we set values in the inverse covariance matrix less than a threshold to 0, and we use a Yale representation of the resulting sparse matrix to speed up the computations (Golub and Van Loan, 1996).

Other approaches, such as IMPUTE (Marchini *et al.*, 2007) and MACH (Li *et al.*, 2010), use hidden Markov models to model a given haplotype as an imperfect mosaic of the set of other observed haplotypes and capture LD between sites. Our approach is simpler and results in fast updates. One notable feature of our method is that it uses a continuous distribution to model the discrete probability mass function that is $P(h^s_{ij}|h_{ij}, H^*_{-i})$. A similar model to ours has recently been shown to provide good performance for imputing unobserved allele frequencies and individual-level genotypes from summary or pooled data (Wen and Stephens, 2010).

2.2 MCMC algorithm as pseudo-code

Pseudo-code for our method is given as follows

(i) Initialize the alleles $(h^s_{i1}, h^s_{i2})\, i \in \{1, \ldots, M\}$ using a single-site expectation-maximization algorithm (see Supplementary Material)

(ii) Set $n = 1$; While $n \leq N$ do

  (a) Select a random ordering of individuals, denoted O.

  (b) For $i \in \{1, \ldots, M\}$ do

  (c) Sample new alleles $(h^s_{O_i 1}, h^s_{O_i 2})$ by sampling from the conditional distribution given in Eq. 1. When updating individual i we store the conditional probabilities

$$p^{(i,n)}_{(a_1,a_2)} = P^{(n)}((h^s_{i1} = a_1, h^s_{i2} = a_2)|R_{is}, h_{i1}, h_{i2}, H^{*n}_{-i}) \tag{12}$$

  where $a_1, a_2 \in \{0, 1\}$

(iii) We discard the first B iterations as burn-in and calculate the mean marginal posterior probabilities of each individual's phased alleles using a Rao–Blackwellised estimator as

$$P(h^s_{i1} = a_1, h^s_{i2} = a_2) = \frac{1}{N-B} \sum_{n=B+1}^{N} p^{(i,n)}_{(a_1,a_2)} \tag{13}$$

The final phased set of alleles is chosen to maximise these marginal posterior distributions.

(iv) Once all sites have been processed, the phased alleles at each site can then be placed into the haplotype scaffold at the appropriate location, resulting in a complete phased haplotype reference panel.

2.3 Parameter settings

MVNcall requires the user to set four parameters N, B, λ and D. We have investigated the sensitivity of the methods to variations in these parameters using the 1KGP dataset to derive a set of default parameters.

We carried out a convergence analysis of the MCMC algorithm to determine how many burn-in and sampling iterations produce stable results (see Supplementary Material). We varied both the number of

iterations $N \in \{0, 10, 50, 100, 200\}$ and burn-in iterations $B \in \{10, 50, 100\}$ to check convergence. Overall we observed that algorithm converges quickly. For the analyses that follow, we ran the method for $N = 100$ iterations with $B = 10$ burn-in iterations.

As we increase the value of λ, we increase the variance of the SNP sites being modelled, and thus 'flatten' the conditional distribution of a haplotype given the current conditioning set. This reduces the relative amount of LD information utilised when calling genotypes at the site of interest and increases the relative amount of information from the GLs. Using the data from the 1KGP, we tested different values of λ, ranging from (0.04, 0.08) with increments of 0.01, and we measured the genotype discordance on a subset of sites for which external data are available. We reached a minimum genotype discordance for $\lambda = 0.06$, and this value is used in all subsequent experiments.

We also ran our analysis of the 1KGP data using values of $D \in \{20, 30, 40, 50, 60, 70, 80\}$ and found that there was little added benefit beyond $D = 50$ (data not shown). This setting corresponds roughly to a window of all SNPs within 60 kb of the site of interest, and this forms our recommended setting for this parameter.

2.4 Method extensions

When using this method to call genotypes in large samples such as the 1KGP data, we found that the method's speed and accuracy could be improved by modifying the algorithm in the following three different ways.

2.4.1 Weighted surrogate family phasing Rather than update the ith individual's genotype given all of the haplotypes in H^*_{-i}, we only use a subset of haplotypes that most closely match the current haplotype estimates of the individual (Howie *et al*., 2009). This approach is called the "surrogate family" phasing approach, as it is a generalisation of the surrogate parent phasing approach of Kong *et al*. (2008). This method works well together with our use of a multivariate normal model to capture local haplotype structure. By excluding haplotypes with little relevant information to the individual being updated, the estimates of allele frequency and covariance better capture the relevant haplotype structure of the individual.

We investigated two different measures of haplotype similarity that are used to identify the set of closest haplotypes of each individual. The first one is the Hamming distance (HD) metric used in IMPUTE2 (Howie *et al*., 2009). The second metric is the perfect match distance (PMD), defined as the number of consecutive matches between two haplotypes, commencing from the site of interest and counting to the right and to the left of the position of interest. The PMD metric is not new in the literature. It has been used in haplotype-sharing methods for association studies under the assumption that individuals with similar phenotypes carry similar chromosomal regions around the risk locus (Beckmann, 2010; Li and Jiang, 2005). For both of these metrics, we only measure haplotype similarity using the scaffold sites. This has the advantage that the relevant calculations can be carried out once at the start of the algorithm and the set of haplotypes used by each individual is then fixed throughout the run of the MCMC method.

We have observed that the genotype discordance rates on the real datasets we analyse in this article are lower for PMD than HD (Supplementary Table S2). Breaking down the genotype discordance rate by genotype class, the largest difference between the genotype discordance rates are for the heterozygous genotypes. A possible advantage of the PMD over the HD measure is the fact that it accounts for the position of the site of interest in the window studied. As our model analyses one position at a time and not a window of SNPs simultaneously, PMD accommodates this information. The HD measure does not account for this, and it gives equal weight to mismatches irrespective of distance from the site of interest, which appears not to be ideal in our model.

In addition to choosing a subset of haplotypes for each individual, we found that using the value of the similarity metric to weight each of the haplotypes further improved performance (data not shown). We calculate the weights by normalizing the PMD of the haplotype by the total sum of PMD in the conditional set. The weights are used to calculate weighted sample means and covariances in Equations 7 and 8.

2.4.2 Model averaging We also investigated the influence of the number (k) of conditioning haplotypes. In our real data experiments, we found that genotype discordance for homozygous genotypes decreases as k increases. Whereas, for the heterozygous genotypes, the genotype discordance reaches a minimum when $k = 100$ and then increases as we increase the value of k (Supplementary Fig. S2). We subsequently found that adopting a model averaging approach improves performance across all genotype classes. We run the MCMC algorithm separately for the different values of k and at the end of the two runs we calculate the mean marginal posterior probabilities from the two runs from all non-burn-in iterations. Thus, we take the mean posterior probabilities from the two runs, giving equal weight to the two runs.

2.4.3 Incorporating trio information for genotype calling The model presented previously assumes that the individuals in the study are unrelated. In the case where the individuals are related, this information can be incorporated in the model to aid genotype calling and phasing. We have extended the method for the analysis of low-coverage sequence data on mother–father–child trios. The details of this method and a full comparison of its performance are included in the Supplementary Material.

2.5 Genotype accuracy comparisons

To assess the accuracy of the method in terms of both genotype calling and phasing, we applied the method on 1092 individuals on chromosome 20 from Phase 1 of the 1KGP. These samples came from 14 distinct population groups (see Supplementary Table S1). The average sequence coverage per individual is in the range 2–6×. The GLs were produced with SNPTools (http://www.hgsc.bcm.tmc.edu/cascade-tech-software_snp_tools-ti.hgsc). For the scaffold, we used phased genotype data collected on these individuals using the Illumina Omni2.5M microarray (OMNI2.5). The genotype data on these samples are a subset of a dataset of 2123 individuals genotyped on this chip, which consists of 327 mother–father–child trios, 42 parent–child duos and 1058 unrelated samples. Supplementary Table S1 lists how many trios, duos and unrelated samples there are in each population. The phasing software SHAPEIT (Delaneau *et al*., 2011) was used to phase the OMNI2.5 genotypes and create the haplotype scaffold of our method. This method used the trio and duo information to improve the phasing. On chromosome 20, this dataset consisted of 54 241 SNPs in the scaffold. We ran our method using $D = 50$ flanking sites in the scaffold each side of every site being called, and for two different values of conditioning haplotypes, one with 100 conditioning haplotypes and the second with all available haplotypes in the reference panel. We analysed samples in three continental groups plus the samples from admixed populations groups. The combinations are Asian and Admixed, European and Admixed and African and Admixed (see Supplementary Table S1). In the case of the Admixed individuals, the posterior genotypes from the African and Admixed run were used.

We make comparisons with three other methods that have been applied on the same dataset, Beagle, Thunder (MaCH) and SNPTools. Beagle, Thunder and our method provided results for all 1092 individuals, whereas SNPTools provided results for only 1041 individuals. Our comparisons are based on the set of 1041 individuals common to all four call sets. The genotype calls made by Thunder and SNPTools are obtained from the 1KGP website. We ran Beagle ourselves by dividing

the chromosome into 3-Mb chunks with 200-kb overlap and using the default parameters.

A subset of 601 individuals were present in HapMap3 (International HapMap 3 Consortium, 2010) and a subset of 591 individuals were also genotyped on an Affymetrix Axiom chip. We thus compared the genotype call sets from the four methods to the genotypes in these two datasets. In these comparisons, we excluded sites on the OMNI2.5 chip because our method assumes that these genotypes are fixed. There were 18 155 sites in the HapMap3 dataset that were *not* on the OMNI2.5 chip. There were 85 858 sites in the Axiom dataset that were *not* on the OMNI2.5 chip. The 591 HapMap3 individuals consisted of 160 with European ancestry, 167 with Asian ancestry, 162 with African ancestry and 102 with Admixed ancestry. The 601 Axiom individuals consisted of 167 with European ancestry, 162 with Asian ancestry, 165 with African ancestry and 107 with Admixed ancestry.

A subset of 170 Phase 1 samples were present in the HapMap3 dataset as trios and were phased in the HapMap3 dataset using this trio information. Of these samples, 112 were also present on the OMNI2.5 chip as trios and duos and the remaining 58 were present on OMNI2.5 as unrelated. Therefore, we can use the HapMap3 haplotypes to assess the accuracy of the haplotypes produced by our method in samples phased using related and unrelated samples. This comparison highlights the value of using a trio/duo-based scaffold in our method. We measured haplotype accuracy using the switch error rate (Delaneau *et al.*, 2011).

We also examined the effect of reducing the SNP density of the haplotype scaffold on the genotype accuracy. The details and results are described in full in the Supplementary Material and show that in populations with lower levels of LD, a denser scaffold is needed to preserve good genotype accuracy. We also used high-coverage sequencing data on a mother–father–child trio to assess genotype concordance and phasing accuracy using the extension of our method to allow for trio information. A full description of these experiments and the results are given in the Supplementary Material.

2.6 Imputation accuracy comparisons

We assessed the quality of the haplotype reference panels produced by MVNcall, Beagle, Thunder and SNPTools by using them as the basis for imputing genotypes into new samples. We used new samples that had been deeply sequenced (at 80× on average) by Complete Genomics (http://www.completegenomics.com/public-data/). Genotypes from such deeply sequenced samples will be accurate. After excluding the individuals who were present in Phase 1, 16 individuals came from three of the population groups of the 1KGP Phase 1 samples. More specifically, nine individuals had African ancestry, four individuals had European ancestry and three individuals had Mexican ancestry. For this analysis, we created four test datasets consisting only of genotypes at SNPs that are present on the Affymetrix 500K, Affymetrix 6.0, Illumina Human1M and OMNI2.5M chips. We then imputed genotypes at SNPs not on the various chips using the four different reference panels. We measured accuracy of the imputed genotypes using the Complete Genomics genotypes not present on the various chips. In the analysis based on chromosome 20, 324 907 non-Affymetrix 500K SNPs were imputed from 12 238 Affymetrix 500K chip SNPs, 314 138 non-Affymetrix 6.0 SNPs were imputed from 23 022 Affymetrix 6.0 chip SNPs, 311 238 non-Illumina Human1M SNPs were imputed from 25 939 Illumina Human1M chip SNPs and 289 837 non-OMNI2.5 SNPs were imputed from 47 317 OMNI2.5 SNPs. SNPs where the alleles were different in the Complete Genomics and 1KGP datasets were removed from the analysis. All the imputation experiments were performed with IMPUTEv2 (Howie *et al.*, 2011) using the default parameters and by splitting chromosome 20 into 5-Mb non-overlapping regions.

We focused the comparison in terms of imputation accuracy on SNPs with low and rare frequency variants, as such SNPs are a focus of much attention in current GWAS (McCarthy *et al.*, 2008). A standard measure of imputation accuracy is the Pearson R^2 correlation coefficient, which measures the correlation at each SNP between the true genotypes $G_{is}^{(true)}$ and the imputed dosages, defined as $G_{is}^{(dose)} = \sum_{g=0}^{2} gP(G_{is} = g) \in [0, 2]$. However, the R^2 metric behaves poorly in a setting where the sample size is small. This is because when the true genotypes of the individuals are the same for a given SNP, the R^2 is undefined. To avoid the problem of calculating per-SNP metrics for small sample sizes, we divided the imputed SNPs according to MAF, and then calculated an aggregate R^2 over all genotypes within each frequency bin. When used on small samples, this measure has good correlation with mean per-SNP R^2 in larger samples (The 1000 Genomes Project Consortium, 2012). The R^2 measure has direct relevance to the power of association studies, as the non-centrality parameter of the test for additive association is directly proportional to the R^2 between a marker locus and a causal locus (Chapman *et al.*, 2003).

3 RESULTS

3.1 Genotype accuracy comparison

Table 1 compares the percentage genotype discordance of the different methods on the HapMap3 dataset. The results are stratified by continental group and by genotype classes: homozygous reference genotypes (HomR), heterozygous genotypes (Het) and homozygous alternative genotypes (HomA). We also report the overall percent genotype discordance as well as the percent genotype discordance excluding the homozygous reference genotypes (NonRef).

In the African and Admixed samples, MVNcall has the lowest overall discordance rate (0.42%), which is mainly driven by the difference in the genotype discordance on the homozygous genotypes (0.19% for HomR and 0.57% for HomA), whereas the other methods have HomR and HomA discordances above

Table 1. Percent genotype discordance comparison of HapMap3 SNPs not in OMNI on chromosome 20

Method	Percent genotype discordance				
	HomR	Het	HomA	Overall	NonRef
African and admixed					
Beagle	0.23	0.95	0.63	0.51	0.83
MVNcall	**0.19**	0.78	**0.57**	**0.42**	0.70
Thunder	0.23	0.69	0.60	0.43	**0.66**
SNPTools	0.25	**0.66**	0.65	0.44	**0.66**
Asian					
Beagle	0.12	0.70	0.38	0.31	0.56
MVNcall	0.12	0.63	0.37	0.29	0.51
Thunder	**0.10**	0.56	**0.34**	**0.26**	0.46
SNPTools	0.12	**0.50**	0.36	**0.26**	**0.44**
European					
Beagle	0.11	0.66	0.40	0.30	0.56
MVNcall	0.11	0.53	0.40	0.27	0.48
Thunder	**0.09**	0.49	**0.35**	0.24	0.43
SNPTools	**0.09**	**0.44**	0.37	**0.23**	**0.41**

Note: Results are divided by population group and in the five columns we report the percent genotype discordance of: HomR: homozygous reference genotypes, Het: heterozygous genotypes, HomA: homozygous alternative genotypes, Overall: the overall genotype discordance and NonRef: the percent discordance on heterozygous and homozygous alternative genotypes. The lowest discordance rate in each column is indicated in bold.

0.23% and 0.60%, respectively. The NonRef discordance of MVNcall is 0.04% higher than that of Thunder and SNPTools owing to the higher Het discordance. MVNcall outperforms Beagle by 0.17% on the Het discordance and achieves 0.09% lower overall discordance and 0.13% lower NonRef discordance rates.

For the Asian and European groups, overall genotype discordance rates are lower for all methods, ranging from 0.23% to 0.30% for the European group and from 0.26% to 0.31% for the Asian group. Thunder achieves the lowest genotype discordance rates for the homozygous genotypes in these population groups, where SNPTools outperforms all other methods on calling heterozygous genotypes. MVNcall outperforms Beagle in both the European and Asian groups. Similar patterns hold when results are stratified into frequency bins MAF $\leq$ 3%, 3% $<$ MAF $\leq$ 5% and MAF $>$ 5% (Supplementary Tables S3–S5). Supplementary Tables S14–S19 report *P*-values for tests of equality between the overall discordance rates between all pairs of methods on each dataset.

Table 2 compares the percentage genotype discordance of the different methods on the Axiom dataset. The Axiom dataset has ~4.7 times the number of SNPs and many more rarer variants than the HapMap3 dataset, making this comparison more relevant. Of the 85 858 Axiom SNPs, there were 23 040, 38 890, 11 720 and 16 208 SNPs with MAF <2.5% in the African, Admixed, Asian and European groups, respectively. Comparing Tables 1 and 2, we see that overall discordance rates are higher in the Axiom dataset, driven by the larger number of rarer SNPs, which are harder to call. MVNcall outperforms other methods when calling homozygous genotypes on the African and Admixed groups, and has the lowest overall genotype discordance rate, tied with Thunder. Our method seems to have elevated discordance rates for the heterozygous genotypes at 1.36% compared with Thunder (1.17%) and SNPTools (1.23%) in the African and Admixed group. Beagle has the highest genotype discordance rates for heterozygous genotypes. In terms of overall discordance rate, MVNcall and Thunder have the lowest overall genotype discordance, followed by SNPTools and Beagle, in the African and Admixed group. In the Asian and European groups, the ranking of methods is SNPTools/Thunder, MVNcall and Beagle. In terms of the NonRef discordance rates, Thunder and SNPTools have the lowest discordance rates, followed by MVNcall and Beagle.

Table 2. Percent genotype discordance comparison on Axiom SNPs not in OMNI, divided by population group

	Percent genotype discordance				
	HomR	Het	HomA	Overall	NonRef
African and admixed					
Beagle	0.29	1.51	1.68	0.72	1.57
MVNcall	**0.25**	1.36	**1.65**	**0.66**	1.47
Thunder	0.30	**1.17**	1.67	**0.66**	**1.36**
SNPTools	0.30	1.23	1.72	0.68	1.41
Asian					
Beagle	0.14	2.08	1.32	0.54	1.75
MVNcall	0.14	2.06	**1.29**	0.54	1.73
Thunder	**0.13**	**1.90**	**1.29**	**0.51**	**1.64**
SNPTools	**0.13**	**1.90**	1.31	**0.51**	1.65
European					
Beagle	0.15	1.46	1.47	0.52	1.46
MVNcall	0.15	1.34	1.46	0.50	1.38
Thunder	**0.14**	1.27	**1.44**	**0.48**	1.33
SNPTools	**0.14**	**1.24**	**1.44**	**0.48**	**1.31**

Note: Results are divided by population group and in the five columns we report the percent genotype discordance of: HomR: homozygous reference genotypes, Het: heterozygous genotypes, HomA: homozygous alternative genotypes, Overall: the overall genotype discordance and NonRef: the percent discordance on heterozygous and homozygous alternative genotypes.

Similar patterns hold when results are stratified into frequency bins MAF $\leq$ 3%, 3% $<$ MAF $\leq$ 5% and MAF $>$ 5% (Supplementary Tables S6–S8). Supplementary Tables S9 and S10 show genotype discordance results for the Axiom and HapMap3 comparisons, respectively, stratified by both population and whether the individuals had scaffold genotypes phased in trios, duos or as unrelated samples. There is some evidence in these tables that genotypes called in samples that use a trio- or duo-phased scaffold are more accurate than in samples that use a scaffold phased with no family information. For example, overall discordance rates in 73 Yoruban samples that use a trio-phased scaffold and 75 Luhya samples that use an unrelated-phased scaffold are 0.42 and 0.63% in the HapMap3 dataset and 0.63% and 0.74% in the Axiom dataset, respectively. The numbers of individuals used to calculate error rates in some cells of these tables are not always large so it can be difficult to make firm conclusions.

The switch error rate of the haplotypes from the 112 Phase 1 samples called using a scaffold derived from trio and duos compared with the trio-phased HapMap3 haplotypes from the same samples was 0.89%. The switch error rate of the haplotypes from the 58 Phase 1 samples called using a scaffold derived from unrelated samples compared with the trio-phased HapMap3 haplotypes from the same samples was 2.65%. These results highlight the value of using trio and duo information when calling genotypes and estimating haplotypes from low-coverage sequence data.

3.2 Imputation accuracy comparison

Figure 2 and Supplementary Figures S3 and S4 show the results of our imputation accuracy comparisons when imputing from genotypes on the Affymetrix 500K chip (results from different chips are given in Supplementary Figs S5–S7). On the *x*-axis we plot the non-reference allele frequency (AF) of the imputed SNPs on the log scale and on the *y*-axis the aggregate R^2.

For the European group (Supplementary Fig. S3), SNPTools has the highest aggregate R^2 for rare SNPs compared with the other methods, which are close in performance across all frequencies. For the individuals with African ancestry, MVNcall clearly outperforms the other methods (Fig. 2), especially on SNPs with AF <0.05. The method increases aggregate R^2 by >0.05 for SNPs with AF <0.005 compared with the three other methods. This increase in R^2 will translate into an increase in power of association studies that interrogate rare variants. This pattern remains stable for all the three Affymetrix 6.0, Illumina Human1M and OMNI2.5M imputation scaffolds. For the

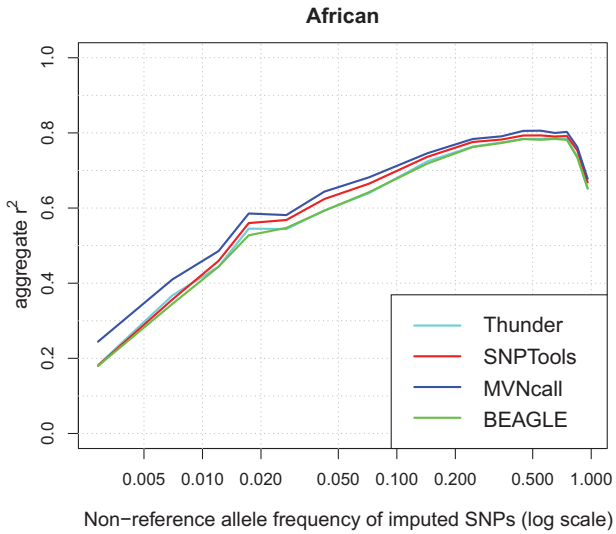

Fig. 2. Comparison of imputation accuracy in nine individuals with African ancestry using haplotype reference panels produced by Thunder (cyan), SNPTools (red), MVNcall (blue) and Beagle (green). We imputed 324 907 SNPs from 12 238 SNPs on the Affymetrix 500K chip on chromosome 20

Mexican individuals with admixed ancestry (Supplementary Fig. S4), we observe that SNPTools and MVNcall have better performance at rare variants than Thunder and Beagle.

4 DISCUSSION

Our proposed model for genotype calling and phasing assumes a study design where samples have been both sequenced and genotyped. The 1KGP uses exactly this design and this was a major motivation for developing such an approach. In addition, many cohorts of samples from GWAS have been genotyped using microarray chips, and it is likely that these samples might be sequenced in the future, providing a combined set of data that also fits in with our approach. In fact, some commercial vendors also return GWAS array data with sequencing results. The genotyped data from the microarrays can be fed into a phasing algorithm, and the resulting haplotypes serve as a haplotype scaffold in our model so that LD information between the scaffold sites and non-scaffold sites can be utilised. To capture this LD information, we use an approximate model that summarises the allele frequencies and the pairwise correlation between SNPs using a multivariate normal distribution. There is an increasing body of literature that the use of a normal distribution can capture enough information about correlations in alleles both between SNPs and between individuals to provide useful levels of inference (Nicholson *et al.*, 2002; Coop *et al.*, 2010; Wen and Stephens, 2010). Future extensions might utilise a normal approximation to efficiently call genotypes from low-coverage sequencing data together with a haplotype reference panel (Pasaniuc *et al.*, 2012) and without a haplotype scaffold. Our model would also be relatively easy to extend to other types of polymorphism such as indels and multi-allelic variants. Utilising phase informative reads when constructing haplotype panels from sequence data may also prove to be beneficial (He *et al.*, 2012).

When applied to data from Phase 1 of the 1KGP that include 1092 individuals from 14 distinct population backgrounds sequenced at 4× and genotyped on the OMNI2.5M chip, we find that our method outperforms Beagle in all population groups in terms of genotype accuracy. This is an interesting result because MVNcall analyses one site at a time rather than all sites jointly as in Beagle. It may be argued that jointly modelling all sites at once is a more appropriate strategy because this approach utilises LD information between all the sites being called. On the other hand, when the calling of genotypes is difficult, which can be the case when sequence coverage is low, it may be that the Beagle's joint approach suffers from convergence problems and the resulting haplotypes include errors due to this problem. The methods SNPTools and Thunder also model all sites jointly but perform slightly better than MVNcall on the European and Asian comparisons. It may be that the underlying models they use are much better than the model used in Beagle, or that the computational strategies that they use do a better job at avoiding convergence problems.

A major use of the haplotype panels produced by the 1KGP is imputation of genotypes into GWAS. It was thus natural that we assess the ability of MVNcall to produce haplotype panels in terms of downstream imputation performance. A key finding of this work is that the panel of haplotypes that we produce from the 1KGP African samples results in a clear boost in downstream imputation performance at rare variants, when compared with the haplotype panels produced by SNPTools, Thunder or Beagle. MVNcall results in a 0.05 increase in mean R^2 at rare variants below 1% frequency when compared with other methods. This is a substantial boost in accuracy, as studies of rare variants are in general less powerful. The R^2 metric used for our assessments has direct relevance to the power of such studies (Chapman *et al.*, 2003). In Admixed samples, MVNcall produces similar results to SNPTools, and both these methods have better performance than Thunder and Beagle at rare variants. We do not see the same increase in imputation performance when imputing European samples.

This may be because a substantial proportion (47%) of African samples in the reference panel have a haplotype scaffold that has been derived using trio phasing. We have shown that the haplotypes our method constructs when using a trio-phased scaffold have much lower switch error rate compared with the haplotypes from a scaffold phased without family information. It may be that it is this increase in haplotype accuracy that translates into a downstream effect on imputation performance. In contrast, only 4.4% of the European haplotypes in the reference panel were derived from a trio-phased scaffold. Almost 85% of the Mexican samples in the haplotype reference panel were derived using a trio-based scaffold, which might lead us to expect that imputation in Mexican samples should be better when using MVNcall reference haplotypes. The ancestry of Mexican individuals derives mostly from Europeans and Native Americans, with a small amount of African admixture. Imputing Mexican samples will work well when all three of these ancestries are well characterised in the reference panel. It may be the case that Mexican haplotype reference set does not provide a sufficient characterisation of Native American

haplotype diversity across the genome and this degrades imputation performance and drives the similarity between methods. Imputation of rare variants in Mexican samples is noticeably worse (Supplementary Fig. S4) than in the African and European samples (Fig. 2 and Supplementary Fig. S3).

Overall these results highlight the gains that can be made when using microarray genotypes to improve genotype calls and phasing using low-coverage sequence data. Future analysis of the 1000 Genomes Project data would benefit from fully incorporating the Illumina OMNI2.5M genotypes into the genotype calling and haplotype estimation process.

ACKNOWLEDGEMENTS

The authors are grateful to Bryan Howie for providing formatted datasets and scripts that facilitated the imputation accuracy comparisons. They thank Olivier Delaneau, Claire Churchhouse and Warren Kretzschmar for comments on the article.

FUNDING

A.M. acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC). J.M. was supported by United Kingdom Medical Research Council grant number G0801823.

Conflict of Interest: none declared.

REFERENCES

Beckmann,L. (2010) Haplotype sharing methods. In *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester.

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

Chapman,J. *et al.* (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.*, **56**, 18–31.

Coop,G. *et al.* (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Delaneau,O. *et al.* (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

Golub,G.H. and Van Loan,C.F. (1996) *Matrix computations*. 3rd edn. John Hopkins University Press, Baltimore, MD.

He,D. *et al.* (2012) Hap-seq: an optimal algorithm for haplotype phasing with imputation using sequencing data. *Lect. Notes Comput. Sci.*, **7262**, 64–78.

Howie,B. *et al.* (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, **1**, 457–470.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS. Genet.*, **5**, e1000529.

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Kim,S.Y. *et al.* (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.*, **34**, 479–491.

Kong,A. *et al.* (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.*, **40**, 1068–1075.

Le,S.Q. and Durbin,R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, **21**, 952–960.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,J. and Jiang,T. (2005) Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics*, **21**, 4384–4393.

Li,Y. *et al.* (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.

Li,Y. *et al.* (2010) Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.

Marchini,J. *et al.* (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Nicholson,G. *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. B*, **64**, 695–715.

O'Connell,J. and Marchini,J. (2012) Joint genotype calling with array and sequence data. *Genet. Epidemiol.*, **36**, 527–537.

Pasaniuc,B. *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.*, **44**, 631–635.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum Genet.*, **73**, 1162–1169.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Tychonoff,A.N. (1943) On the stability of inverse problems. *Doklady. Akademii. Nauk. SSSR*, **39**, 195–198.

Wen,X. and Stephens,M. (2010) Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.*, **4**, 1158–1182.