UK Biobank Phasing and Imputation Documentation

Version 1.2
13 November 2015

documentation author Jonathan Marchini
Department of Statistics, University of Oxford
on behalf of UK Biobank

Contributors to UK Biobank Phasing and Imputation

Jonathan Marchini (Statistics Dept, Oxford), Jared O'Connell (WTCHG, Oxford), Olivier Delaneau (University of Geneva), Kevin Sharp (Statistics Dept, Oxford), Warren Kretzschmar (WTCHG, Oxford), Gavin Band (WTCHG, Oxford), Shane McCarthy (WTSI, Hinxton), Desislava Petkova (WTCHG, Oxford), Claire Bycroft (WTCHG, Oxford), Colin Freeman (WTCHG, Oxford), Peter Donnelly (WTCHG, Oxford).

Table of Contents

Introduction

This document describes the analysis carried out to perform genotype imputation for the interim release of the UK Biobank (UKB) genotype data. It also provides advice about using the imputed data to carry out genome wide association studies (GWAS) or for extracting genotypes for use as covariates in other types of association study.

Genotype imputation[1,2] is the process of predicting genotypes that are not directly assayed in a sample of individuals. A reference panel of haplotypes at a dense set of SNPs, indels and structural variants, is used to impute genotypes into a study sample of individuals that have been genotyped at a subset of the SNPs. These '*in silico'* genotypes can then be used to boost the number of SNPs that can be tested for association. This increases the power of the study, the ability to resolve or fine-map the causal variants and facilitates meta-analysis. The result of the imputation process is a dataset with 73,355,667 SNPs, short indels and large structural variants in 152,249 individuals. See Box 1 of [1] for a quick visual overview of how genotype imputation works.

The process of imputation is divided into two steps (i) pre-phasing, and (ii) imputation. In the first step, the samples to be imputed are 'pre-phased' i.e a statistical method is applied to genotype data to infer the underlying haplotypes of each individual. In the second step, a different statistical method is used to combine the inferred haplotypes with a reference panel of haplotypes and impute the unobserved genotypes in each sample. The following two sections of this document describe how the pre-phasing and imputation was carried out on the ~150,000 samples.

Phasing and imputation can be a computationally intensive process. To avoid many different research groups having to carry this out independently, phasing and imputation was been carried out centrally.

Questions about using the imputed genotypes should be sent to the UKB Genetics mail list set up for this purpose. You can subscribe to the mail list here

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKB-GENETICS

Phasing

Filtering before phasing

To create an input data for the phasing we applied SNP QC filters as described in UK Biobank QC documention [3]. The samples were genotyped on two slightly different chips. Approximately 50,000 were genotyped as part of the UL BiLEVE study using a chip designed for that study (denoted UKBL), with the remaining samples (~100,000) genotyped on the UKB chip. Therefore, we applied different missingness filters on SNPs dependent upon chip. SNPs were removed based on the number of batches in which they are completely missing:

i. SNPs on both UKB chip and UKBL chip - remove them if they are missing in more than 3 batches (out of 33 batches)
ii. SNPs on the UKB chip and not the UKBL chip - remove them if they are missing in more than 2 batches (out of 22 batches)
iii. SNPs on the UKBL chip and not the UKB chip - remove them if they are missing in more than 1 batch (out of 11 batches)

1,037 sample outliers[3] were removed. Multi-allelic SNPs and SNPs with a minor allele frequency (MAF) < 1% were then removed from the dataset. These filters resulted in a dataset with 641,018 autosomal SNPs in 152,256 samples. Chromosome X phasing and imputation will be carried out at a later date.

Phasing method description

Phasing on the autosomes was carried out using a modified version of the SHAPEIT2[4] program modified to allow for very large sample sizes. This new method (which we refer to as SHAPEIT3) modifies SHAPEIT2's surrogate family approach to remove a quadratic complexity component of the algorithm[5]. In small sample sizes of a few thousand samples, this part of the algorithm, which involves calculating Hamming distances between current haplotypes estimates, contributes only a relatively small part to the computational cost. As sample sizes increase over 10,000 samples then this component becomes significant. The new algorithm uses a divisive clustering algorithm to identify clusters of haplotypes, and then calculates Hamming distances only between pairs of haplotypes within each cluster. Only haplotypes within each cluster are used as candidates for the surrogate family copying states in the HMM model. The resulting algorithm has complexity O(N log N) where N is the number of haplotypes in the dataset being phased. In practice, we have observed that the method exhibits scaling close to linear. This is a crucial feature of the method, especially for very large sample sizes, and a property not shared by other approaches[6,7]. The development of this approach is ongoing and there is substantial scope to make further improvements in speed and accuracy. A newer version is likely to offer an order of magnitude reduction in speed.

Validation of the phasing method

The accuracy of this new method was assessed by taking advantage of 72 mother-father-child trios that were identified in the UKB dataset[3]. This family information can be used to infer the phase of a large number of SNPs in the trio parents. These family inferred haplotypes were used as a truth set, as is common in the phasing literature[4]. The parents of each trio were removed from the dataset and then haplotypes were estimated across chromosome 20 in a single run of SHAPEIT3. This dataset consisted of 16,762 autosomal SNPs. The inferred haplotypes were then compared to the truth set using the switch error metric[4]. We obtained an exceptionally low switch error rate of 0.4% across the trio children reporting British ancestry. By adjusting parameters of the method we have observed switch error rates lower than 0.3%.

With switch error rates this low, long chunks of sequence of many megabases will be inferred correctly. Downstream imputation from such haplotypes will be highly accurate.

To assess the performance gain of phasing all 152,112 samples together, versus phasing in smaller subsets of samples two other test datasets of size 1,072 and 10,072 samples were created, also containing the trio children. The results are shown in full detail in **Table 1** and highlight the benefits of joint phasing of all the samples. These results clearly demonstrate the close to linear scaling of the SHAPEIT3 algorithm.

Sample size	Method	Switch Error (%)	Run time (hrs)	Run Time Scaling	Sample Size Scaling	Threads
1,072	SHAPEIT3	2.6	0.25	1	1	10
10,072	SHAPEIT3	1.3	2.5	10	9.4	10
152,112	SHAPEIT3	0.4	38.5	154	142	10

Table 1 : Phasing performance on UKB samples.

Whole genome phasing

Phasing was carried out in chunks of 5,000 SNPs, with an overlap of 250 SNPs between chunks. SHAPEIT3 was run on each chunk using 4 cores per job and S=200 copying states. As a part of the phasing process any remaining missing genotypes were imputed during the phasing. Chunks were ligated using a modified version of the hapfuse program.

Genotype imputation

Assessment of the UK Biobank Array for imputation

The UK Biobank Axiom array from Affymetrix was specifically designed to optimize imputation performance in GWAS studies [8]. An experiment was carried out to assess the imputation performance of the array, stratified by allele frequency, and to compare performance to some other commercially available arrays.

Performance was assessed using high-coverage, whole-genome sequence data made publicly available by Complete Genomics (CG).

Data from 10 samples from the European ancestry (CEU) population was used. All variant sites with a call rate below 90% were filtered out in order to only consider very reliable sites in the analysis. Only data from chromosome 20 was used.

To mimic a typical imputation analysis, a pseudo-GWAS dataset was constructed by extracting the CG SNP genotypes at all the sites included on a given array. All sites not on the array were then imputed using the UK10K reference panel [9]. Imputation was carried out using IMPUTE2 [10] which chooses a custom reference panel for each study individual in each 1 Mb segment of the genome. The k_{hap} parameter of IMPUTE2 was set to 1,000. All other parameters were set to default values. This experiment was repeated for 4 different genome-wide SNP arrays (a) Affymetrix UK Biobank Axiom array (b) Illumina Omni 2.5M array (c) Illumina Omni 1M Quad (d) Illumina Omni Express.

Variants were stratified into allele frequency bins and the squared correlation (R^2) was calculated between the allele dosages at variants in each bin with the masked CG genotypes. Since different arrays contain different numbers of variants it is important to make sure that imputation performance is measured at the same set of variants when comparing chips. To achieve this, both imputed and array variants were included in the R^2 analysis, so that the comparison measures the overall performance of each array. As a consequence, an array with more variants will gain an advantage, as it is reasonable to expect that directly genotyping a variant will yield more accurate genotypes than imputation. **Figure 1** shows the results of this analysis. The x-axis is non-reference allele frequency (%) on a log scale, which focuses in on rarer variants. The y-axis is imputation performance (R^2).

The salient points are

a. the UK Biobank chip (purple) outperforms the Illumina Omni 1M Quad (blue) and Illumina Omni Express (green), both which have comparable numbers of variants.
b. The UK Biobank chip performs almost as well as the Illumina 2.5M chip (red), which has ~3 times the number of SNPs. It is worth noting that the UKB chip and Illumina Omni 2.5M chip are very close in the 1-5% range. A likely consequence of the chip design process focusing in part on this frequency range [8].

The overall conclusion of this analysis is that the Affymetrix UKB array is a very good array from which to carry out genotype imputation. The caveat is that this analysis is focused on samples with European ancestry.

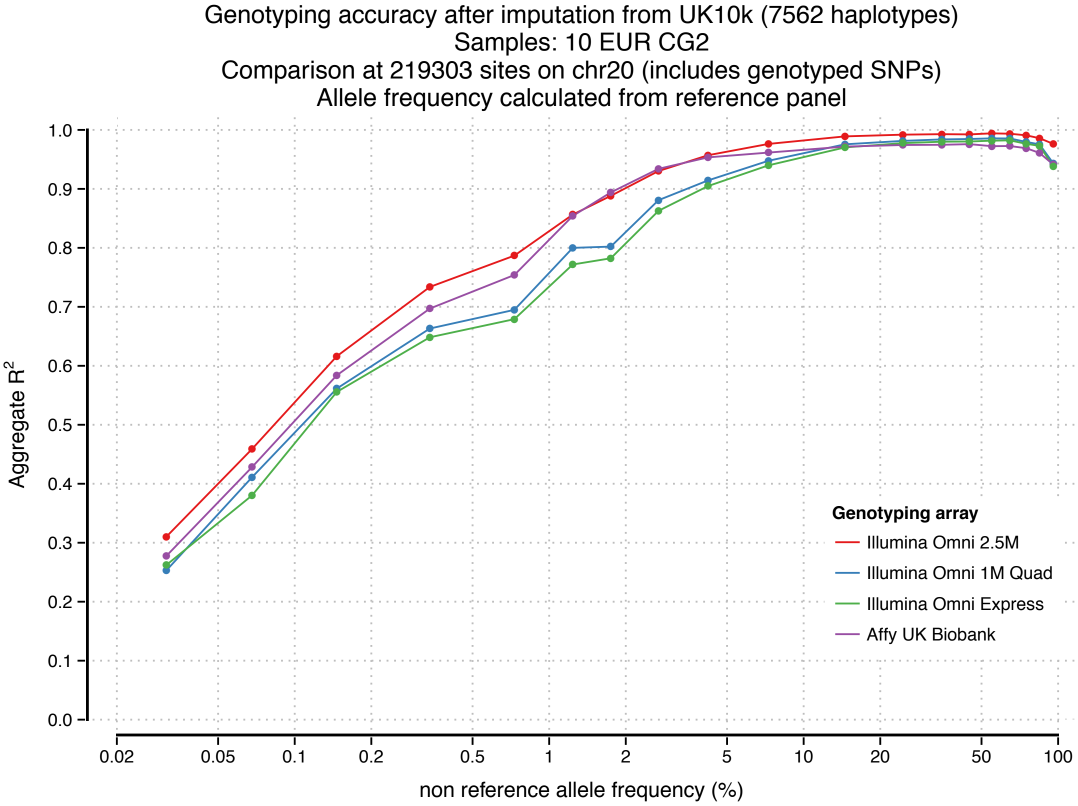

Figure 1 : Comparison of imputation performance of the UK Biobank Array and several other commercially available genotyping arrays.

Reference panel used for imputation

There are a number of factors that influence the accuracy of genotype imputation [1], but generally accuracy will increase as the number of haplotypes in the reference panel grows and if the ancestry of the sample haplotypes is a good match to the ancestry of the reference panel haplotypes. The UKB dataset consists of samples with a diverse range of ancestries, but with the majority of samples having British (or European) ancestry. For this reason it was desirable to use a reference panel with a large number of haplotypes with British and European ancestry, and also a diverse set of haplotypes from other world-wide populations. To achieve this the UK10K haplotype reference panel was merged together with the 1000 Genomes Phase 3 reference panel using the **–merge_ref_panels** option in the IMPUTE2 software (link).

Using this merged panel has been shown to produce a high-quality reference panel for imputation[9]. An advantage of this reference panel is that it includes SNPs, short indels and larger structural variants. The reference panel consists of 87,696,888 bi-allelic variants in 12,570 haplotypes.

Imputation method description

Imputation was carried out using the same algorithm as is implemented in the IMPUTE2 program. The current IMPUTE2 program is a very flexible tool for phasing and imputation that implements a general set of options. A new C++ program was written from scratch to focus exclusively on haploid imputation needed when samples have been pre-phased. This new version is both memory and computationally efficient compared to IMPUTE2. The method takes advantage of high correlations between inferred copying states in the HMM to reduce computation. We refer to this program as IMPUTE3.

Whole genome imputation

Imputation was carried out in chunks of 2Mb with a 250kb buffer region. A set of 2,000 haplotype copying states were used to impute each sample. Imputed variants in each non-overlapping part of each chunk were concatenated into per-chromosome files.

Information scores, minor allele frequencies and filtering

QCTOOL was used to calculate the minor allele frequency (MAF) and imputation information score of each imputed variant. The imputation information is a metric between 0 and 1. A value of 1 indicates that there is no uncertainty in the imputed genotypes whereas a value of 0 means that there is complete uncertainty about the genotypes. A value of α in a sample of *N* individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size of α*N.*

Many GWAS carried out to date have used filters on MAF and information score by applying a threshold on these metrics. There is no single correct threshold to use. However, as MAF decreases it is generally the case that imputation quality decreases. Previous studies have tended to use a filter on information between 0.3-0.5. Since these studies have typically consisted of hundreds or low thousands of samples an information of 0.3 corresponds to an effective sample size with limited power to detect associations. However, the UK Biobank dataset is considerably larger in size than most previous GWAS. An information measure of 0.3 in ~150,000 samples roughly corresponds to an effective sample size of ~45,000, which would be expected to yield very good power to detect association.

Some variants are imputed as monomorphic, or close to monomorphic i.e. no or almost no variation in the genotypes. Such sites were removed using QCTOOL using a filter on MAF of 0.001%. In addition, 7 samples were removed from the dataset due to these individuals having requested their data be removed from the study. The resulting dataset consists of 73,355,667 variants in 152,249 individuals.

The distribution of information scores at these 73,355,667 variants is shown in **Figure 2** (a). Plots stratified by MAF are also shown (b) MAF > 5% (c) 1%<=MAF<5% (d) 0.1%<=MAF<1% (e) 0.01%<=MAF<0.1% (f) 0.001%<=MAF<0.01%.

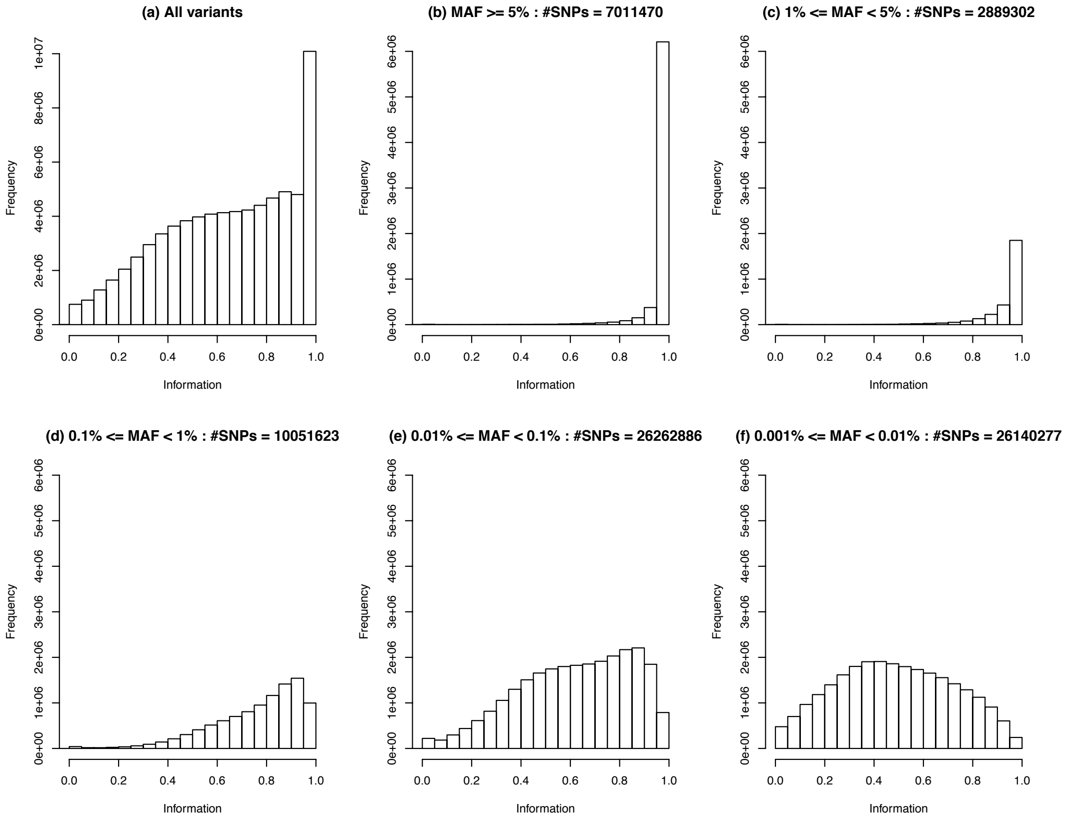

Figure 2 : Distribution of information scores at variants in the imputed dataset. The x-axis shows the information score on the scale 0 to 1.

Imputed genotype files

Let G_{ij} denote the genotype of the i^{th} sample at the j^{th} variant. The process of genotype imputation produces a probability distribution for each genotype i.e.

$p_{ij0} = P(G_{ij} = AA)$ $p_{ij1} = P(G_{ij} = AB)$ $p_{ij2} = P(G_{ij} = BB)$

where A and B are the two alleles at the variant. This probability triple (which sums to 1) is provided in the imputed genotype files for each imputed variants in all samples. SNP variants included in the phased dataset also occur in the imputed files in this format.

The imputed data is provided in a compressed binary BGEN file format. The BGEN file format is a binary version of the GEN file format.

The BGEN file format was chosen to provide good compression of the imputed data and ease of use for genetic association testing against traits and phenotypes. For example, programs commonly used such as SNPTEST and PLINK already read BGEN files, and QCTOOL can be used to filter, summarize, manipulate and convert the files to other formats.

The format stores one variant at a time (i.e. per row). As MAF decreases more compression is possible due to increased similarity between imputed genotypes across samples. The total size of the UKB Interim release dataset is 1.3Tb, with each chromosome file ranging in size from 20Gb to 109Gb. As the file format is binary the files are not viewable in normal text editors. Later in this document there is advice and guidance on working with these files.

The files are named as

chrNimpv1.bgen

where N is the number of the autosome (N = 1,….,22).

RS IDs were added into the BGEN files for as many variants as possible using available RS ID lists available from the UK10K website and the 1000 Genomes website.

RS IDs are useful, unique identifiers of SNPs and other variants and can be looked up in the dbSNP database. When researchers report associations of variants with diseases and traits they normally report the results using the RS ID.

Variant positions are reported in Genome Reference Consortium Human genome build 37 co-ordinates (GRChb37).

Sample files

In addition to the 22 autosomal BGEN files, there is file called impv1.sample

This file (refered to as the `sample file') is the part of the BGEN file format that stores information about each sample in the dataset. The format of this file is described on the GEN file format webpage.

The sample file has two header lines, followed by 1 line for each individual in the BGEN file. The order of the individuals in the sample file matches the order of the individuals in the BGEN file. **The order is important**. Programs that read bgen/sample pairs assume that the order matches between the files.

The sample file can be used to store information about each individual i.e. phenotypes and covariates. If phenotypes and covariates are added into the sample file then SNPTEST can be used to carry out association testing at each variant. Care should be taken in making sure that such information is correctly added to sample files. The format allows discrete and continuous phenotypes and covariates, as well as missing values (see file format webpage link above).

Differences between raw genotypes and imputed files

SNPs below 1% MAF were filtered out before the phasing step, however many of these SNPs will have been imputed. Therefore these SNPs will appear in the raw genotype files, and the imputed files, but may have different genotypes. As such, researchers should not be surprised if the results of analysis at these SNPs differ dependent upon which files are used.

An exemplar genome wide association study

A GWAS for the phenotype of height was carried out to assess the use of the UK Biobank genetic data as a resource for genetic association studies. There are already a substantial number of replicated associations [11]. The purpose of this analysis was not to report new associations, but rather to check that a reasonably standard GWAS pipeline produced valid results.

Sample filtering

Principal component analysis and the self-declared ethnicity were used to derive a "White British" subset of samples. In addition, samples were excluded if they had

(a) at least one related sample
(b) a genetically inferred gender that did not match the self-reported gender.
(c) ~500 extreme outliers [3].

These filters resulted in a dataset with 112,338 samples.

Taking account of the different arrays used

Some SNPs are only included on one of the UKBB or UKBL arrays. At such SNPs, missing genotypes will have been imputed as part of the phasing process, so that these SNPs will consist of a mixture of genotyped and imputed SNPs. This can lead to bias in association testing if there is some correlation between the phenotype and which array a sample was assayed on. The samples involved in the UKBL study were selected based on phenotypes associated with lung function[12], thus it may be possible for such associations to occur. There are at least 2 solutions to ameliorate any possible confounding due to array

a. carry out association tests conditioning on a binary indicator of array.
b. carry out separate tests of association in UKBB samples and UKBL samples and combine the results using meta-analysis.

Association testing

GWAS was performed at all variants using SNPTEST. An additive genetic model was fitted at each SNP, using gender, age, array and 10 principal components as covariates. That is, the example uses option (a) above.

The program option **–method expected** was used in the SNPTEST software, which converts the genotype probability triple to an expected genotype, d_{ij}, (often called the dosage), calculated as

$$d_{ij} = \sum_{k=0}^{2} kp_{ijk}$$

Results

The GWAS for height produced a substantial number of associated regions. These regions had a high correspondence to those genetic regions that have previously been replicated for height and described in the NHGRI GWAS Catalog [11]. The analysis suggested a significant number of novel loci could be identified. **Figure 3** shows a plot of the –log10 p-values for the height and BMI scans on chromosome 4.

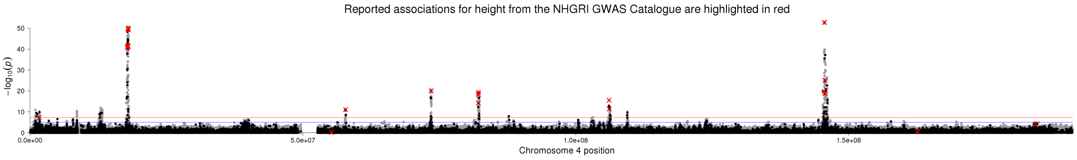

Figure 3 : Chromosome 4 GWAS for height. The x-axis shows physical position. The y-axis is –log10 p-value for each tested variant. Variants on the array are shown as black dots, imputed variants are shown as grey dots. Reported associations from the NHGRI GWAS Catalog are shown as red crosses. The blue and red horizontal lines are drawn at a –log10 p-value of 5 and 7.5 respectively.

File processing

We recommend that researchers use the QCTOOL program to handle the BGEN files. This program has options for extraction or removal of subsets of the data (SNPs and/or samples), and for file format conversion. See the QCTOOL examples page for information on command lines used to perform specific tasks.

The program SNPTEST can process BGEN files. It will automatically detect the BGEN file format if data files are named with the .bgen extension.

PLINK v1.9 can process BGEN files; at the time of writing BGEN files are specified using the --bgen option.

For further information on tools supporting the BGEN format, see the BGEN file format website.

References

1. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
2. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
3. The UK Biobank. *UK Biobank Genotyping QC documentation*. (2015).
4. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
5. O'Connell, J., Sharp, K., Delaneau, O. & Marchini, J. Haplotype estimation for biobank scale datasets. (2015) (submitted)
6. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40,** 1068–1075 (2008).
7. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91,** 238–251 (2012).
8. The UK Biobank Array Design Group. UK Biobank Axiom Array Content Summary. (2014).
9. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications* **6,** 8111 (2015).
10. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1,** 457–470 (2011).
11. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* **42,** D1001–6 (2014).
12. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3,** 769–781 (2015).