

1 Evaluating the quality of the 1000 Genomes

2 Project data

3 **Saurabh Belsare¹, Michal Sakin-Levy², Yulia Mostovoy², Steffen Durinck³, Subhra Chaudhry³, Ming Xiao⁴, Andrew**
4 **S. Peterson³, Pui-Yan Kwok^{1,2,5}, Somasekar Seshagiri³ and Jeffrey D. Wall^{1,6,*}**

5 ¹Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, 94143, USA, ²Department of Dermatology,
6 University of California, San Francisco, San Francisco, CA, 94143, USA, ³Department of Molecular Biology, Genentech Inc., 1 DNA Way,
7 South San Francisco, CA, 94080, USA, ⁴School of Biomedical Science, Engineering, and Health Systems, Drexel University, Philadelphia,
8 PA, 19104, USA, ⁵Cardiovascular Research Institute, San Francisco, San Francisco, CA, 94143, USA, ⁶Department of Epidemiology and
9 Biostatistics, University of California, San Francisco, San Francisco, CA, 94143, USA

10 **ABSTRACT** Data from the 1000 Genomes project is quite often used as a reference for human genomic
11 analysis. However, its accuracy needs to be assessed to understand the quality of predictions made using this
12 reference. We present here an assessment of the genotype, phasing, and imputation accuracy data in the 1000
13 Genomes project. We compare the phased haplotype calls from the 1000 Genomes project to experimentally
14 phased haplotypes for 28 of the same individuals sequenced using the 10X Genomics platform. We observe
15 that phasing and imputation for rare variants are unreliable, which likely reflects the limited sample size of
16 the 1000 Genomes project data. Further, it appears that using a population specific reference panel does not
17 improve the accuracy of imputation over using the entire 1000 Genomes data set as a reference panel. We
18 also note that the error rates and trends depend on the choice of definition of error, and hence any error
19 reporting needs to take these definitions into account.

20 INTRODUCTION

21 The 1000 Genomes Project (1KGP) was designed to provide a comprehensive description of human genetic variation
22 through sequencing multiple individuals¹⁻³. Specifically, the 1KGP provides a list of variants and haplotypes that can be
23 used for evolutionary, functional and biomedical studies of human genetics. Over the three phases of the 1KGP, a total of
24 2504 individuals across 26 populations were sequenced. These populations were classified into 5 major continental

*Correspondence: jeff.wall@ucsf.edu

25 groups: Africa (AFR), America (AMR), Europe (EUR), East Asia (EAS), and South Asia (SAS). The 1KGP data was generated
26 using a combination of multiple sequencing approaches, including low coverage whole genome sequencing with mean
27 depth of 7.4X, deep exome sequencing with a mean depth of 65.7X, and dense microarray genotyping. In addition, a subset
28 of individuals (427) including mother-father-child trios and parent-child duos were deep sequenced using the Complete
29 Genomics platform at a high coverage mean depth of 47X. The project involved characterization of biallelic and
30 multiallelic SNPs, indels, and structural variants.

31 Given the low depth of (sequencing) coverage for most 1KGP samples, it is unclear how accurate the imputed haplotypes
32 are, especially for rare variants. We quantify this accuracy directly by comparing imputed genotypes and haplotypes
33 based on low-coverage whole-genome sequence data from the 1KGP with highly accurate, experimentally determined
34 haplotypes from 28 of the same samples. Additional motivation for our study is given below.

35 **Phasing** It is important to understand phase information in analyzing human genomic data. Phasing involves resolving
36 haplotypes for sites across individual whole genome sequences. The term '*diploomics*'⁴ has been coined to describe
37 "*scientific investigations that leverage phase information in order to understand how molecular and clinical phenotypes are*
38 *influenced by unique diplotypes*". The diplotype shows effects in function and disease related phenotypes. Multiple
39 phenomena like allele-specific expression, compound heterozygosity, inferring human demographic history, and
40 resolving structural variants requires an understanding of the phase of available genomic data. Phased haplotypes are
41 also required as an intermediate step for genotype imputation.

42 Phasing methods can be categorized into methods which use information from multiple individuals and those which rely
43 on information from a single individual⁵. The former are primarily computational methods, while the latter are mostly
44 experimental approaches. Some computational approaches use information from existing population genomic databases
45 and can be used for phasing multiple individuals. These, however, may be unable to correctly phase rare and private
46 variants, which are not represented in the reference database used. On the other hand, some methods use information
47 from parents or closely related individuals. These have the advantage of being able to use Identical-By-Descent (IBD)
48 information, and allow long range phasing, but require sequencing of more individuals, which adds to the cost. A few
49 methods which use these approaches are: PHASE⁶, fastPHASE⁷, BEAGLE⁸⁻⁹, SHAPEIT¹⁰⁻¹¹, EAGLE¹²⁻¹³ and IMPUTE v2¹⁴.

50 Experimental phasing methods, on the other hand, often involve separation of entire chromosomes followed by
51 sequencing of short segments, which can then be computationally reconstructed to generate entire haplotypes. These
52 methods do not need information from individuals other than the one being sequenced. These methods involve

53 genotyping being performed separately from phasing. These methods fall into two broad categories, namely dense and
54 sparse methods¹⁴. Dense methods resolve haplotypes in small blocks in great detail, where all variants in a specific region
55 are phased. However, they do not inform the phase relationship between the haplotype blocks. These involve diluting
56 high molecular weight DNA fragments such that fragments from at most one haplotype are present in each unit. Sparse
57 methods can resolve phase relationships across large distances, but may not inform on the phase of each variant in a
58 chromosome. In these methods, a low number of whole chromosomes is compartmentalized such that only one of each
59 pair of haplotypes is present in each compartment. These compartmentalizations are followed by sequencing to generate
60 the haplotypes.

61 In this work, we use phased haplotypes generated using the 10X Genomics method which uses linked-read sequencing¹⁵.
62 1 nanogram of high molecular weight genomic DNA is distributed across 100,000 droplets. This DNA is barcoded and
63 amplified using polymerase. This tagged DNA is released from the droplets and undergoes library preparation. These
64 libraries are processed via Illumina short-read sequencing. A computational algorithm is then used to construct phased
65 haplotypes based on the barcodes.

66 **Imputation** Imputation involves the prediction of genotypes not directly assayed in a sample of individuals.
67 Experimentally sequencing genomes to a high coverage is an expensive process. Low coverage sequencing or arrays can
68 be used as low-cost methods for sequencing. However, these methods may lead to uncertainty in estimated genotypes
69 (low coverage sequencing) or missing genotype values for untyped sites (arrays). Imputation can be used to obtain
70 genotype data for missing positions using reference data and known data at a subset of positions in individuals which
71 need to be imputed. Imputation is used to boost the power of GWAS studies¹⁶, fine mapping a particular region of a
72 chromosome¹⁷, or performing meta-analysis¹⁸, which involves combining reference data from multiple reference panels.

73 Imputation uses a reference panel of known haplotypes with alleles known at a high density of haplotyped positions. A
74 study/inference panel genotyped at a sparse set of positions is used for sequences which need to be imputed. Performing
75 imputation involves two basic steps:

- 76 • Phasing genotypes at genotyped positions in the study/inference panel
- 77 • Haplotypes from the inference panel which match those in the reference panel at the positions in the study panel
78 are assumed to match in all other positions

79 Various imputation algorithms perform these steps sequentially and iteratively or simultaneously.

80 Factors affecting the quality of the phasing and imputation are (1) size of reference panel (2) density of SNPs in reference
81 panel (3) accuracy of called genotypes in the reference panel (4) degree of relatedness between sequences in reference
82 panel and study sequences (5) ethnicity of the study individuals in comparison with the available reference data and (6)
83 allele frequency of the site being phased or imputed⁵.

84 Multiple methods have been developed for genotype imputation¹⁹. fastPHASE⁷, MACH²⁰⁻²¹, BEAGLE^{8,22-23}, and *IMPUTE* v2¹⁴
85 are some widely used methods for imputation.

86 An analysis of the imputation accuracy for the HapMap project has been performed about a decade ago²⁴, but no similar
87 detailed analysis exists for assessing the phasing and imputation of the 1000 Genomes project, particularly comparing the
88 database against experimentally phased sequences. We present here a detailed assessment of the quality of phasing and
89 imputation for the 1000 Genomes database, particularly as a function of minor allele frequency and inter-SNP distances
90 for biallelic SNPs.

91

92 MATERIAL AND METHODS

93 Input Data

94 Processed VCFs were downloaded from the 1000 Genomes website. This data is available for each chromosome
95 separately. To obtain agreement with the experimental data, 1000 Genomes VCFs corresponding to the GRCh38 assembly
96 were downloaded. Experimental data was sequenced using the 10X Genomics platform for 28 individuals: 5 GM, 18 HG,
97 and 5 NA. The GM and NA individuals were originally part of the HapMap project while the HG are from the 1000
98 Genomes project. Thirteen of these individuals were processed at UCSF and sequenced at Novogene, while the remaining
99 individuals were processed and sequenced at Genentech. The populations from which each of the individuals come (as
100 listed in the Coriell Catalog) are:

- 101 • South Asia (SAS):
 - 102 ○ Gujarati Indians in Houston, Texas, USA (HapMap) [GIH] - GM21125*, NA20900, NA20902
 - 103 ○ Punjabi in Lahore, Pakistan [PJL] - HG03491, HG03619
 - 104 ○ Sri Lankan Tamil in the UK [STU] - HG03679, HG03752, HG03838*

- 105 ○ Indian Telugu in the UK [ITU] - HG03968
- 106 ○ Bengali in Bangladesh [BEB] - HG04153, HG04155
- 107 ● East Asia (EAS):
 - 108 ○ Han Chinese in Beijing, China (HapMap) [CHB] - GM18552*, NA18570, NA18571
 - 109 ○ Chinese Dai in Xishuangbanna, China [CDX] - HG00851*, HG01802, HG01804
 - 110 ○ Kinh in Ho Chi Minh City, Vietnam [KHV] - HG02064, HG02067
 - 111 ○ Japanese in Tokyo, Japan (HapMap) [JPT] - NA19068*
- 112 ● Africa (AFR):
 - 113 ○ Luhya in Webuye, Kenya (HapMap) [LWK] - GM19440*
 - 114 ○ Gambian in Western Division, The Gambia [GWD] - HG02623*
 - 115 ○ Esan from Nigeria [ESN] - HG03115*
- 116 ● Europe (EUR):
 - 117 ○ Toscani in Italia (Tuscans in Italy) (HapMap) [TSI] - GM20587*
 - 118 ○ British from England and Scotland, UK [GBR] - HG00250*
 - 119 ○ Finnish in Finland [FIN] - HG00353*
- 120 ● America (AMR):
 - 121 ○ Mexican Ancestry in Los Angeles, California, USA (HapMap) [MXL] - GM19789*
 - 122 ○ Peruvian in Lima, Peru [PEL] - HG01971*

123 Asterisks next to sample IDs refer to samples processed at UCSF.

124 **Preprocessing 1000 Genomes Data**

125 The 1000 Genomes data was separated into individual and chromosome specific VCFs using *vcftools*²⁵. Further, the
126 variants were filtered for biallelic SNPs, phased, filtered for PASS, and indels were removed. The experimentally phased
127 data also had a very small fraction of unphased SNPs, which were removed by filtering with *vcftools*. The analysis was
128 performed only for autosomes.

129 **Phasing Analysis**

130 The alternate (ALT) allele frequencies of all the SNPs of interest were obtained from the 1000 Genomes data and
131 converted to minor allele frequencies to be able to analyze switch error as a function of minor allele frequencies. The
132 filtered SNPs from the experimental data were split into phase sets, based on phase set information available in the
133 experimental VCF files. Switch error was calculated between the experimental and 1000 Genomes data for each phase set
134 in each chromosome of each individual from the experimental dataset. Switch error is defined as *percentage of possible*
135 *switches in haplotype orientation used to recover the correct phase in an individual*²⁶ or *proportion of heterozygous positions*
136 *whose phase is wrongly inferred relative to the previous heterozygous position*²⁷. *vcftools* returns the switch error as well as
137 all positions of switches occurring along the chromosome.

138 **Switch Error as a Function of Minor Allele Frequency** ALT allele frequencies were accessed for each of the switch
139 positions from the data and were converted to minor allele frequencies. Distribution of switch positions as a function of
140 minor allele frequency was plotted for each chromosome in each individual.

141 **Switch Error as a Function of Inter SNP Distance** Positions of each SNP were accessed from the data. The number of
142 intermediate switches were counted for all pair of SNPs, not only consecutive SNPs. If the number of switches between
143 two SNPs were odd, a switch error was counted. This was used to calculate the distribution of switch errors as a function
144 of inter-SNP distance.

145

146 **Imputation Analysis**

147 The entire imputation analysis is performed for each chromosome for each individual.

148 **Generate Recombination Map** *IMPUTE v2*¹³ makes available recombination maps for each chromosome using the 1000
149 Genomes data for the GRCh37 assembly. A recombination map was obtained for each chromosome for GRCh38 by lifting
150 over the GRCh37 maps using the *liftover* software. ~8k positions (0.2%) were removed from the lifted over
151 recombination map because *liftover* resulted in them being in the incorrect order.

152 **Generate Reference Panel** A reference haplotype panel was generated for all individuals from the 1000 Genomes data by
153 subsetting it to the specific population of interest. 1000 Genomes data for the individuals which were experimentally
154 sequenced was not included in the reference panel. *vcftools* was used to filter out the individuals of interest from the 1000

155 Genomes data. *bcftools* was used to convert the VCF data to haps-sample-legend format. An alternate approach was also
156 used, where the entire 1000 Genomes data was used to generate a reference haplotype panel.

157 **Generate Study Panel** A study panel was generated for the experimentally sequenced individuals selected. The study
158 panel is assumed to be genotyped at positions corresponding to the Illumina **InfiniumOmni2.5-8** array. Array positions
159 were lifted over from GRCh37 to GRCh38 using *liftover*. 1000 Genomes haplotypes (since 1000 Genomes data is
160 prephased, the study panel is also in the form of haplotypes rather than genotypes) for those positions for those
161 individuals were selected to create the study panel using *vcftools*. Filtered VCF files were converted to the haps-sample
162 format using *bcftools*.

163 **Run Imputation** Missing positions are imputed using *IMPUTE v2*. Imputation was performed in 5Mb windows. The
164 genotype output by imputation was converted to VCF format using *bcftools*. VCFs produced over all windows were
165 combined using *vcf-concat*. *IMPUTE v2* generally phases the typed genotyped sites in study panel. This is followed by
166 imputation which is performed by assuming that haplotypes in the study panel that match the haplotypes in the reference
167 panel at the typed sites also match in the untyped sites. *IMPUTE v2* then performs an iterative process performing
168 multiple Monte-Carlo steps alternating phasing and imputation. For this analysis, however, as haplotypes from the 1000
169 Genomes project were directly used to generate the study panel, the phasing step was not performed.

170 **Filter Positions** For one part of the analysis, i.e. estimating errors in the positions represented in the experimentally
171 phased VCFs (henceforth called experimental SNPs), the positions from those VCFs were filtered from the imputed data
172 using *vcftools*. Experimental genotypes from the experimental VCFs were obtained for each individual of interest using
173 *vcftools*. SNPs with duplicate entries in either the imputed or experimental data were removed. Continent-specific allele
174 frequencies were obtained for the experimental SNPs from the 1000 Genomes data using *vcftools*, to be able to analyze
175 switch error as a function of Minor Allele Frequencies. For the other part of the analysis, i.e. estimating errors for all
176 positions in the 1000 Genomes data, the allele fractions were similarly obtained for all of the SNPs.

177 **Imputation Error** Imputation error was computed as fraction of genotypes being incorrectly identified. Imputation error
178 was computed for both, the SNPs in the experimental data and all the SNPs in 1000 Genomes data. Error is computed as a
179 function of minor allele frequency. The continent-specific minor allele frequencies were used for analyzing the imputation
180 error.

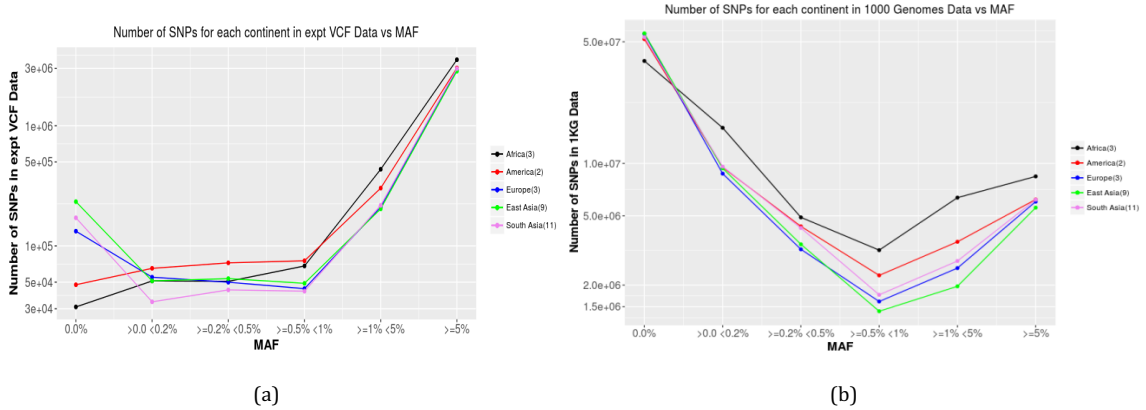
181 For all analysis where error rate is computed as a function of the continent-specific minor allele frequency (genotyping
182 error and imputation error; Figs. 1,2,7,8), the minor allele frequencies are binned as $MAF=0.0\%$, $0.0\% < MAF < 0.2\%$,
183 $0.2\% \leq MAF < 0.5\%$, $0.5\% \leq MAF < 1\%$, $1\% \leq MAF < 5\%$, $MAF \geq 5\%$. For the analysis where all 1000 Genomes minor
184 allele frequencies are used (phasing error and imputation error comparing use of multiple reference panels; Figs. 3, 4, 9),
185 the minor allele frequencies are binned into only five bins, i.e. there is no $MAF=0.0\%$ bin. Rest of the bins are the same as
186 for the continent-specific MAF bins.

187 **Experimental Methods**

188 **Samples processing:** HMW Genomic DNA was extracted and converted into 10x sequencing libraries according to the
189 10X Genomics (Pleasanton, CA, USA) Chromium Genome User Guide and as published previously²⁸. Briefly, GEMS were
190 made with 1.25ng HMW template gDNA, Master-mix Genome Gel Beads and partitioning oil on the microfluidic Genome
191 Chip. Isothermal incubation of the GEMs (for 3 h at 30°C; for 10 min at 65°C; stored at 4°C) produced barcoded fragments
192 ranging from a few to several hundred base pairs. After dissolution of the Genome Gel Bead in the GEM Illumina Read 1
193 sequencing primer, 16bp 10x barcode and 6bp random primer are released. The GEMs were then broken and the pooled
194 fractions were recovered. Silane and Solid Phase Reversible Immobilization (SPRI) beads were used to purify and size
195 select the fragments for library preparation. Library prep was performed according to the manufacturer's instructions
196 described in the Chromium Genome User Guide Rev C. Libraries were made using 10x Genomics adapters. The final
197 libraries contain the P5 and P7 primers used in Illumina bridge amplification. The barcoded libraries were then quantified
198 by qPCR (KAPA Biosystems Library Quantification Kit for Illumina platforms). Sequencing was done using Illumina HiSeq
199 4000 with 2×150 paired-end reads. Raw reads were processed, aligned to the reference genome, and had SNPs called and
200 phased using 10X Genomics' Long Ranger software (version 2.1.1 or 2.1.6) with the "wgs" pipeline with default settings.

201 **RESULTS**

202 The 1000 Genomes project chromosome-specific VCFs for the GRCh38 assembly contain between 6.4M (chr1) to 1.1M
203 (chr22) variants over all the 2504 individuals. After filtering for biallelic SNPs, phased, filtered for PASS, removing indels,
204 we are left with 6.15M (chr1) to 1.05M (chr22) variants. The experimentally phased data from the 10X Genomics platform
205 has different numbers of called variants for each sequenced individual. For chromosome 1, the number of called variants
206 varies from 414K to 494K across the 28 individuals, while, for chromosome 22, the number of called SNPs varies from
207 104K to 120K. After performing a similar filtering for the experimental data, the number of biallelic PASS phased SNPs
208 ranges between 298K and 357K for chromosome 1 and 64K and 75K for chromosome 22.

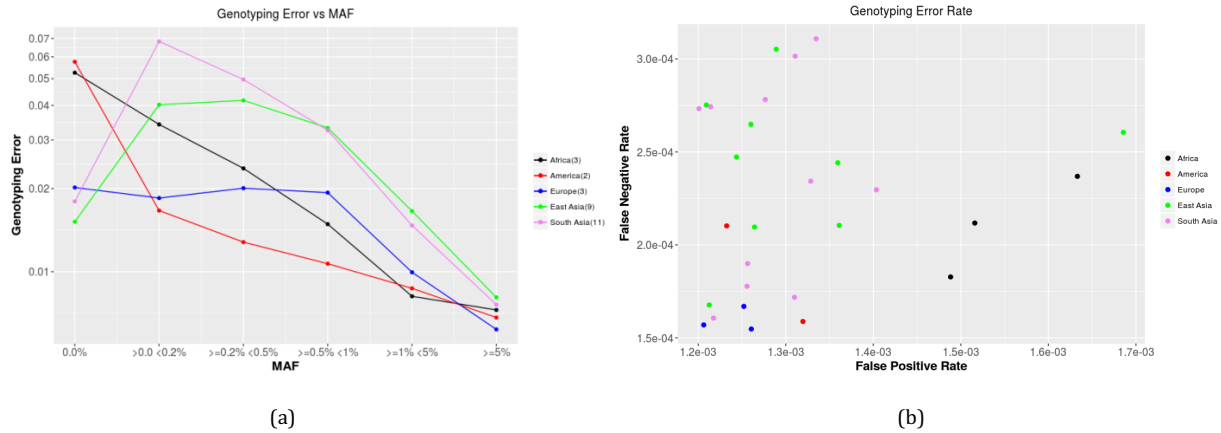


209
 210
 211 **Figure 1** Distribution of SNPs as a function of continent-specific minor allele frequencies (a) only experimental SNPs (b) all 1000
 212 Genomes SNPs

213 The SNPs from the experimentally phased VCFs (Fig. 1a), averaged over continent groups show that the vast majority of
 214 SNPs in this selection have high continent-specific MAF values (> 5%). Comparing across continents for the continent
 215 invariant SNPs, the African and American individuals have an order of magnitude less continent invariant SNPs than the
 216 European, East Asian and South Asian individuals. However, if we look at all the SNPs in the 1000 Genomes Data (filtered
 217 for biallelic PASS phased SNPs) as a function of continent-specific MAF, the distribution we observe has a very different
 218 trend. There is a significant over-representation of the very low continent-specific MAF SNPs (< 0.1%), $\sim 5 * 10^7$, as
 219 compared to all the subsequent higher MAF SNPs, which all range $< 1 * 10^7$.

220 These discrepancies between the numbers in the 1000 Genomes data and in the experimentally phased data, as well as
 221 the differing trends as a function of MAF occur because the 1000 Genomes data includes a SNP if even one individual in
 222 the 2504 individuals has a variant (heterozygous or homozygous-alternate) at that position while the experimental data
 223 includes a SNP only if that particular individual has a variant (heterozygous or homozygous-alternate) at that position.
 224 This results in a much larger number of overall SNPs being present in the 1000 Genomes data as compared to the
 225 experimental and also the majority of the 1000 Genomes SNPs having extremely low MAF, as those would occur only in
 226 one or a few individuals.

227 **Genotyping Error**



228

229

230

231

232

Figure 2 Genotyping error (a) in the experimental VCF positions as a function of continent-specific minor allele frequency averaged over all chromosomes over all individuals in each continent (b) false positive vs false negative rates (defined in text) for all 1000 Genomes SNPs

233

234

235

236

237

238

239

240

Genotyping error is computed comparing the 1000 Genomes genotypes with the experimental genotypes. The experimental genotypes for all SNPs not present in the experimental VCF for each individual are assumed to be homozygous reference. Mismatched genotypes are counted as errors. Figure 2a looks at the errors (fraction of genotypes which are incorrect) for the experimental VCF positions as a function of the continent-specific minor allele frequencies. There is higher error at the population invariant sites (MAF=0.0%) in the African and American populations than the European, East Asian and South Asian populations. This correlates with a lower total number of population invariant SNPs in those continents (Fig. 1a). For non-invariant SNPs, we observe, as expected, a decreasing error rate with increasing minor allele frequency, to a <1.5% error genotyping error rate for the SNPs with minor allele frequencies > 1%.

241

242

243

244

245

246

247

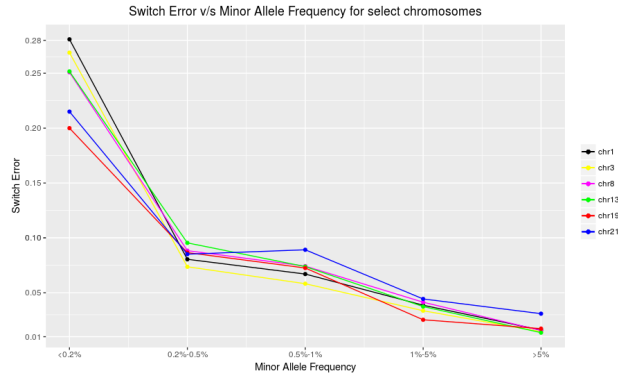
248

249

Comparing false positive (sites non-homozygous reference in 1000 Genomes data and homozygous reference in the experimental data) vs false negative (sites homozygous reference in 1000 genomes data and non-homozygous reference in the experimental data) error rates for all 1000 Genomes sites (Fig. 1b), we see that the genotyping for the European and American individuals is very accurate, with both low false positive and false negative rates. The East Asian and South Asian populations both have mostly low false positive rates, but show a wide range (factor of 2) of false negative rates, while showing only a ~15% variation in the false positive rates for most individuals. In contrast, the African individuals mostly have relatively low false negative rates, but have among the highest false positive rates. This indicates that the sequencing in the 1000 Genomes project has over called non-homozygous reference variants in African individuals compared to the rest, and over called SNPs as homozygous reference in some of the East and South Asian individuals.

250

Phasing

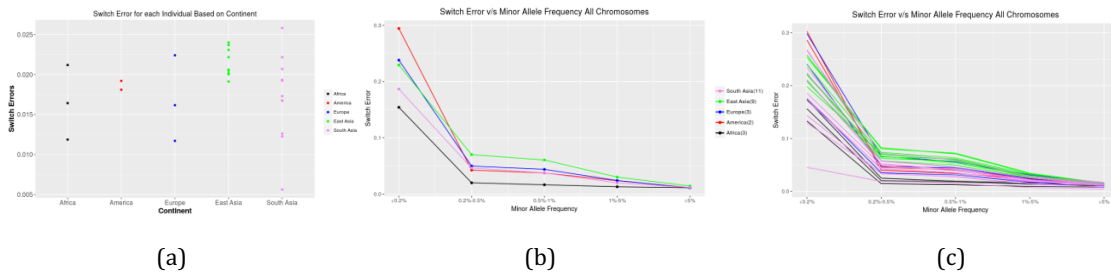


251

252 **Figure 3** Switch error as a function of Minor Allele Frequencies for different individual chromosomes. Chromosome 21 shows higher
 253 switch error for large MAF values

254 Phasing errors are all analyzed for overall 1000 Genomes minor allele frequencies, not continent specific MAFs.

255 Comparing the switch error across individual chromosomes (Fig. 3), we observe that the switch error ranges between 25
 256 - 30% for the rare MAF (< 0.1%) SNPs, falling to < 5% for SNPs with MAFs 1 - 5%. The majority of SNPs, which fall in the
 257 MAF > 5% category, have an error < 2.5%. However, a comparatively higher switch error at larger MAF values (> 5%) is
 258 observed for chromosome 21. This plot (Fig. 3) shows only a subset of chromosomes a single individual (GM18552), but
 259 this trend is observed for all other chromosomes and individuals studied.



260

261

(a)

(b)

(c)

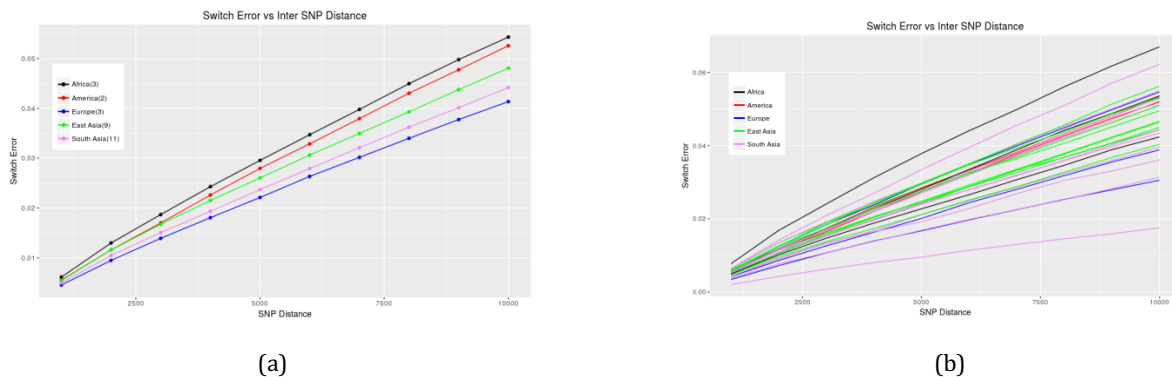
262 **Figure 4** Switch error (a) Total switch error (number of switches in experimental SNPs/total number of experimental SNPs) for each
 263 individual (b) Switch error as a function of Minor Allele Frequencies averaged over all individuals in each continent. (c) Switch error as a
 264 function of Minor Allele Frequencies for all individuals colored by continent.

265 Figure 4a shows the total switch error for each of the individuals. The total switch errors for all the individuals studied go
 266 up to ~ 2.5%. The switch errors for the East Asian individuals are grouped together, while those for the South Asian
 267 individuals show greater variability. This is in line with the general observation that South Asian populations have an
 268 overall greater heterogeneity than do East Asian populations [J. Wall, Unpublished data]

269 Analyzing the switch error as a function of minor allele frequency averaged over all chromosomes of all individuals of a
270 population (Fig. 4b), we observe low switch error, < 5%, for low minor allele frequencies (MAF) (1 – 5%). For rare SNPs
271 with MAF (0.2 – 1%), the switch error is ~ 5 – 10%. For extremely rare minor allele SNPs, i.e. MAF < 0.2%, the error is
272 much higher, i.e. 15 – 35%. For all higher MAF values (> 5%), the error is < 2.5%. The average error rate for the
273 individuals from the African populations is almost the same over the range of MAF values > 0.1%.

274 As observed in Figure 4c, the differences in the error rates between individuals decrease with increasing minor allele
275 frequency. Individuals from South Asia show a larger variation in error as a function of MAF as compared to individuals
276 from East Asia. The individuals from the African populations have the lowest switch error over the range of MAF values.
277 Individual NA20900, an individual from the Gujarati Indians in Houston (GIH) population has the lowest switch error as a
278 function of minor allele frequency for the low MAF SNPs.

279



280

281

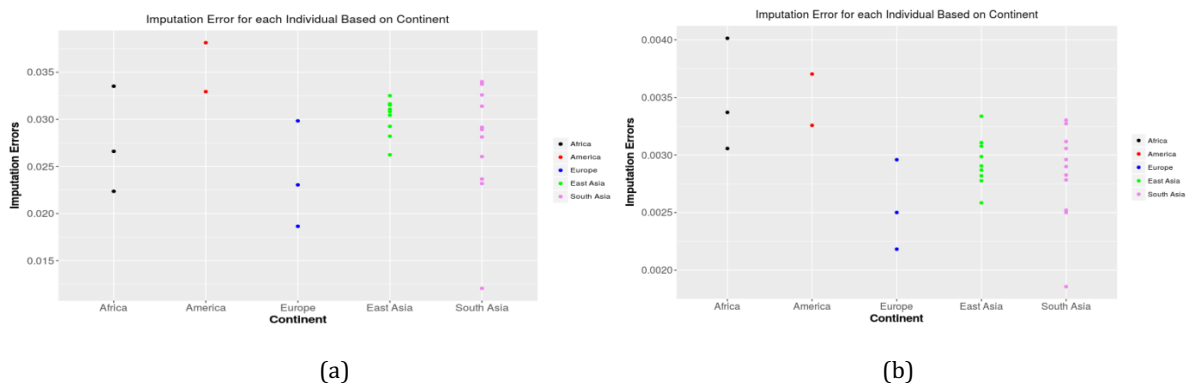
282 **Figure 5** Switch error as a function of inter-SNP distance (a) Switch error as a function of inter-SNP distances averaged over individuals
283 in each continent. (b) Switch error as a function of inter-SNP distances for all individuals colored by continent.

284 We also analyzed phasing error as a function of the distances between SNPs (Fig. 5). The phasing error increases as a
285 function of the inter-SNP distance, i.e. SNPs which are further apart are more likely to be out of phase with each other. The
286 within population trends are the same as for switch error vs MAF, where the individuals from South Asia show a larger
287 spread as compared to the individuals from East Asia. Individual NA20900 shows the lowest error rate, same as for the
288 comparison of error vs MAF (Fig. 4c).

289 Comparing the switch error as a function of MAF vs. the switch error as a function of inter-SNP distance, we see that the
290 individuals from the African populations show distinctly opposite trends. For low MAF SNPs, the error is the lowest
291 averaging over the African individuals, while across the range of inter-SNP distances, the average over the African

292 individuals was the highest error. The reason this occurs can be understood from the fact that there are a higher number
293 of low MAF SNPs in the African individuals in the experimental data (Fig. 1a), as well as an overall higher number of SNPs
294 in those individuals, leading to a higher SNP density for these individuals. In addition, there is less linkage disequilibrium
295 (LD) in the individuals from the African populations, which would make it harder to phase them accurately²⁹⁻³⁰. Hence,
296 pairs of SNPs are more likely to be out of phase with each other, leading to higher switch error as a function of inter-SNP
297 distance.

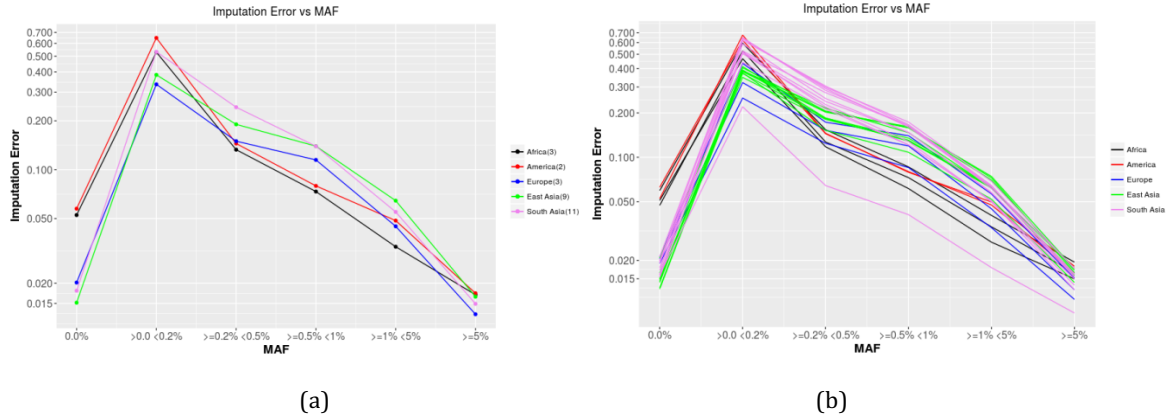
298 Imputation



299
300 (a) (b)
301 **Figure 6** Total imputation error (a) Total imputation error in experimental SNPs (number of incorrect genotypes in all experimental
302 SNPs/total number of experimental SNPs) for each individual (b) Total imputation error in all 1KG SNPs (number of incorrect genotypes
303 in all 1KG SNPs/total number of 1KG SNPs) for each individual

304 Imputation error is computed as the fraction of SNPs with incorrectly imputed genotypes. However, depending on the
305 subset of SNPs under consideration, the error can be computed in two different ways, (1) fraction of experimental SNPs
306 incorrectly imputed and (2) fraction of all 1KG SNPs incorrectly imputed. In the case of the second definition of error, the
307 experimental calls for all the positions not in the experimental VCFs are set to homozygous-reference.

308 Figure 6a shows the total imputation error in the experimental SNPs while Figure 6b shows the total imputation error in
309 the 1KG SNPs for each of the individuals. The total imputation errors in the experimental SNPs for all the individuals
310 studied go up to $\sim 4\%$. For this subset of SNPs, the two American individuals have the among the highest imputation
311 errors. The imputation errors for the East Asian individuals are grouped together, while those for the South Asian
312 individuals show greater variability. This agrees with our observations for the switch error (Fig. 4a). In the 1KG SNPs, on
313 the other hand, since we are looking at a much larger set of SNPs, most of which are homozygous-reference in any given
314 individual, we see a much smaller error $< \sim 1\%$.



315

316

317 **Figure 7** Imputation accuracy experimental VCF positions (a) Imputation error in the experimental SNPs as a function of Minor Allele

318 Frequencies averaged over individuals in each continent. (b) Imputation error in the experimental SNPs as a function of Minor Allele

319 Frequencies for all individuals colored by continent.

320 **Imputation error in experimental SNPs** Figure 7a shows a wide range of error rates as function of the continent-

321 specific minor allele frequency. The continent invariant positions (MAF=0.0%) are imputed almost as accurately as the

322 high MAF (>5% in 3 populations, and >1% in two populations) SNPs. In these positions, we make the same observation as

323 we did for the original genotyping in the 1000 genomes reference data (Fig. 2a), i.e. the errors in the European, East Asian

324 and South Asian individuals for these continent invariant positions are lower than those for the American and African

325 individuals. For the very rare SNPs, i.e. MAF < 0.2%, the error is as high as ~ 60%. These extremely high error rates are

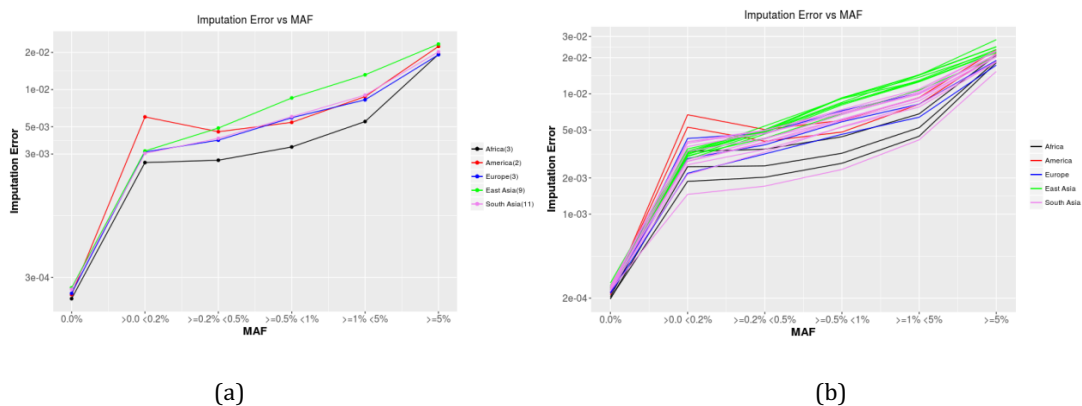
326 only observed in the American individuals and a few of the South Asian individuals. For the rest of the individuals, the

327 error rates are < 50%. In the mid-range of MAF values, i.e. 0.2% to 1%, the errors range between 10 – 20%. The SNPs with

328 higher MAF values are fairly accurate, with errors < 2% for common SNPs (MAF > 5%). This can also be seen looking at all

329 the individuals separately (Fig. 7b). The South Asian (Gujarati in Houston, Texas) individual NA20900 still shows the

330 lowest error rate as a function of MAF for imputation, just as it does for the switch error (Fig. 4c).



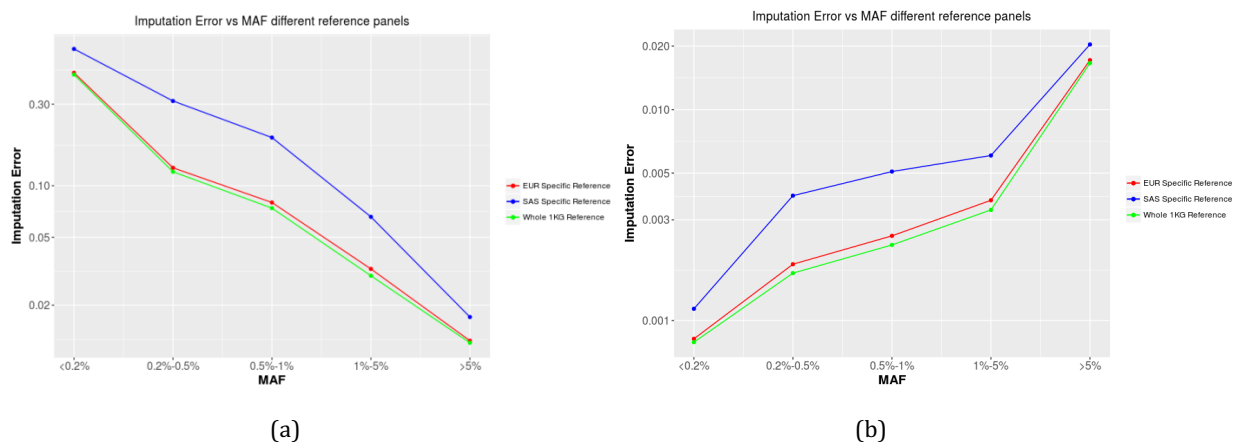
331

332

333 **Figure 8** Imputation accuracy all 1KG SNPs (a) Imputation error in all the 1000 Genomes positions as a function of Minor Allele
 334 Frequencies averaged over individuals in each continent. (b) Imputation error in all the 1000 Genomes positions as a function of Minor
 335 Allele Frequencies for all individuals colored by continent.

336 **Imputation error in all 1KG SNPs** Computing the error using all the 1KG SNPs, we see a different trend for the errors as
 337 a function of minor allele frequency (Figs. 8a, 8b). The invariant sites have very low errors $\sim 10^{-4}$. For the variant sites, the
 338 errors increase as a function of minor allele frequency, as opposed to decreasing as they do in the experimental only SNPs.
 339 The reason this happens is that contrasting the number of experimental SNPs (Fig. 1a) with the numbers of all 1KG
 340 SNPs (Fig. 1b), while the number of low MAF SNPs is 1-2 orders of magnitude less than the number of SNPs with MAF >
 341 5% in the experimental data, the number of very low MAF SNPs is 2-10 times greater than the number of SNPs with MAF
 342 > 5% in the whole 1000 Genomes data. The vast majority of the very low MAF SNPs in the whole 1000 Genomes data are
 343 homozygous-reference, since those SNPs show variation in only one or very few 1000 Genomes individuals. Hence,
 344 imputation predictions get most of those positions correct in most of the individuals. As a result, the fraction of those very
 345 rare SNPs which are predicted incorrectly is much lower when considering all the 1000 Genomes SNPs as compared to
 346 only considering the experimental SNPs, where most of the SNPs are high MAF SNPs.

347 Consistent with the observations for the experimental only SNPs, at very rare SNPs (MAF < 0.2%), the American
 348 individuals still have the highest error rate. The individuals from the South Asian populations still show a greater spread
 349 than those from the East Asian populations. Individual NA20900 still shows the lowest error rate as with previous
 350 observations.



351
 352 (a) (b)
 353 **Figure 9** Imputation error as a function of Minor Allele Frequencies for European individuals comparing the European reference panel
 354 v/s the entire 1KG reference panel (a) experimental SNPs (b) All 1000 Genomes SNPs

355 **Comparison of reference panels** Here, we compare the imputation errors resulting from using different reference
356 panels for imputation. A continent-specific reference panel for the individual of interest, a reference panel which includes
357 all of the 1000 Genomes individuals, and a continent-specific reference panel for a different continent from the one from
358 which the individuals are, are chosen. The minor allele frequencies used here are for all the overall 1000 Genomes minor
359 allele frequencies, instead of a continent-specific minor allele frequency, since we want to understand the impact of the
360 choice of reference panel, and continent-specific MAFs would not align with the whole reference or the reference from
361 another continent. In this case, we look at the imputation error in the 3 European individuals when imputation is carried
362 out with the European reference, the South Asian reference, and the whole 1000 Genomes reference.

363 The observed result for experimental only SNPs (Fig. 8a) when comparing reference panels for the European individuals
364 is very similar when looking at all 1000 Genomes SNPs (Fig. 8b). The imputation accuracy when using the entire 1000
365 Genomes data as a reference panel gives a slightly better accuracy than using just a European specific reference panel.
366 The error while using an incorrect reference panel, however, is up to a factor of 2 greater than the error when using the
367 appropriate reference, or when using the whole 1000 Genomes reference panel. The trend of error as a function of MAF is,
368 again, the opposite of what was observed when looking at only the experimental SNPs.

369 **DISCUSSION**

370 The 1000 Genomes Project data have been widely used as a reference for estimating continent-specific allele frequencies,
371 and as a reference panel for phasing and imputation studies. Since the project's design involved low-coverage (~7X)
372 sequencing for most of the samples, it was unknown *a priori* how accurate the 1KGP's genotype and haplotype calls were,
373 especially for rare variants. This accuracy obviously directly impacts the usefulness of the 1KGP data. With the advent of
374 inexpensive, commercial platforms for experimentally phasing whole genomes, it is possible to directly quantify the
375 genotype and haplotype error rates of the 1KGP data.

376
377 Our comparison of 28 experimentally phased genomes with the 1KGP data found that the latter is highly accurate for
378 common and low-frequency variants (i.e., $MAF \geq 0.01$). As expected, accuracy declined with decreasing MAF, with rare
379 variants ($MAF < 0.01$) not reliably imputed onto haplotypes. Surprisingly though, the genotype calls were reasonably
380 accurate even for rare variants. This observation may not generalize to other low-coverage sequencing studies due to the
381 complicated and labor-intensive protocol used for variant calling in the 1KGP. We conclude that the 1KGP data is best
382 used as a reference panel for imputing variants with $MAF \geq 0.01$ into populations closely related to the 1KGP groups, and
383 is probably of limited utility for imputation in rare variant association studies. Larger subsequent imputation panels, such

384 as the one generated by the Haplotype Reference Consortium (HRC)³¹, are likely much more useful for imputing rare
385 variants, at least in well-studied European populations. However, even this large reference panel may be of limited
386 usefulness for imputation into other human groups. While our results suggest that using a region-specific reference panel
387 (for the correct region) for imputation is only slightly worse than using a worldwide panel, the choice of an incorrect
388 regional panel makes the imputation considerably worse. So, large European-based haplotype reference panels will be of
389 limited utility for imputing variants into East Asian, South Asian, or African-American genomes, while imputation studies
390 involving understudied groups such as Middle Easterners, Melanesians or Khoisan are likely to have error rates
391 substantially higher than what was observed in our study. This is a consequence of the fact that most rare variants are
392 region-specific; imputation only works when the variant being imputed shows up often enough in the reference panel. In
393 summary, while the 1KGP and HRC provide valuable genomic resources that can augment the power of GWAS in groups
394 with European ancestry, additional large-scale genome sequencing of diverse human populations will be necessary to
395 obtain comparable benefits of imputation in genetic association studies of non-European groups.

396

397 Finally, we note that the absolute error rate varied by an order of magnitude, depending on the specific definitions of
398 error that were used. This highlights the importance of definitional clarity in studies that evaluate the accuracy of
399 genomic resources.

400 **Conflicts of Interest Declaration**

401 Genentech authors hold shares in Roche. The other authors declare no conflicts of interest.

402 **Acknowledgments**

403 JDW was supported in part by NIH grant R01 GM115433. SB was supported by Genentech research grant CA0095684.

404

405 **LITERATURE CITED**

- 406 1. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale
407 sequencing. *Nature* 467, 1061–1073.

- 408 2. The 1000 Genomes Project Consortium (2012) An integrated map of human genetic variation from 1092 human
409 genomes. *Nature* 491, 56–65.
- 410 3. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526, 68–
411 74.
- 412 4. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011) The importance of phase information
413 for human genomics. *Nature Reviews Genetics* 12, 215–223
- 414 5. Browning, S. and Browning, B. (2011). Haplotype phasing: existing methods and new developments. *Nature*
415 *Reviews Genetics* 12, 703–714
- 416 6. Stephens, M. and Scheet, P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and
417 missing-data imputation. *Am. J. Hum. Genet.* 76, 449–462
- 418 7. Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data:
419 applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644
- 420 8. Browning, S. and Browning, B. (2007). Rapid and accurate haplotype phasing and missing-data inference for
421 whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097
- 422 9. Browning, B. and Browning S. (2009). A unified approach to genotype imputation and haplotype-phase
423 inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223
- 424 10. Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype
425 inference. *BMC Bioinformatics* 9
- 426 11. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012) A linear complexity phasing method for thousands of genomes.
427 *Nature Methods* 9, 179–181
- 428 12. Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef Y. A., Finucane, H. K. Schoenherr, S., Forer, L.,
429 McCarthy, S., Abecasis, G. R., et al., (2016) Reference-based phasing using the haplotype reference consortium
430 panel. *Nature Genetics* 48, 1443–1448
- 431 13. Loh, P.-R., Palamara, P. F., and Price, A. L. (2016) Fast and accurate long-range phasing in a UK biobank cohort.
432 *Nature Genetics* 48, 811–816
- 433 14. Howie, B. N., Donnelly, P., and Marchini, J. (2009) A flexible and accurate genotype imputation method for the
434 next generation of genome-wide association studies. *PLOS Genetics* 5
- 435 15. Synder, M. W., Adey, A., Kitzman, J. O., and Shendure, J. (2015) Haplotype-resolved genome sequencing:
436 experimental methods and applications. *Nature Reviews Genetics* 16, 344–358

- 437 16. Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou,
438 S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.*, (2016) Haplotyping germline and cancer genomes with high-
439 throughput linked-read sequencing. *Nature Biotechnology* *34*, 303–311
- 440 17. Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009) Designing genome-wide association studies: sample
441 size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* *5*
- 442 18. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007) A new multipoint methods for genome-
443 wide association studies by imputation of genotypes. *Nature Genetics* *37*, 906–913
- 444 19. Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P.,
445 Andersen, G., *et al.* (2008) Meta-analysis of genome wide association data and large-scale replication identifies
446 additional susceptibility loci for type 2 diabetes. *Nature Genetics* *40*, 638–645
- 447 20. Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nature Reviews*
448 *Genetics* *11*, 499–511
- 449 21. Li, Y., Willer, C. J., Sanna, S., and Abecasis, G. R. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*
450 *10*, 387–406
- 451 22. Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010) Mach: using sequence and genotype data to
452 estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* *34*, 816–834
- 453 23. Browning, S. (2006). Multilocus association mapping using variable-length markov chains. *Am. J. Hum. Genet.* *78*,
454 903–913
- 455 24. Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg N. A., and Scheet, P. (2009) Genotype-
456 imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* *84*, 235–250
- 457 25. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G.
458 T., Sherry, S. T., *et al.* (2011). The variant call format and vcftools. *Bioinformatics* *27*, 2156–2158.
- 459 26. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E. and Lin, S., Qin, Zhaohui, S. Q., Munro, H.
460 M. Abecasis, G. R., *et al.*, (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am. J.*
461 *Hum. Genet.* *78*, 437–450
- 462 27. Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from
463 population genotype data. *Am. J. Hum. Genet.* *73*, 1162–1169
- 464 28. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., Jaffe, D. B. (2017) Direct determination of diploid genome
465 sequences. *Genome Research* *27*(5), 757-767

- 466 29. Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., Di Rienzo, A. (2001) Gene conversion and
467 different population histories may explain the contrast between polymorphism and linkage disequilibrium
468 levels. *Am. J. Hum. Genet.* 69, 831–843
- 469 30. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A.,
470 Faggart, M., et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229
- 471 31. McCarthy, S., Das, S., Kretschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C.,
472 Danecsek, P., Sharp, K., et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nature*
473 *Genetics* 48(10), 1279-1283