UK Biobank imputation using a Genomics England reference panel

Version 1.0 (9 March 2022)

Authors : Sinan Shi, Simone Rubinacci, Loukas Moutsianas, Alex Stuckey, Anna Need Sile Hu, The Genomics England Research Consortium, Mark Caulfield, Simon Myers, Jonathan Marchini

Contents

1 Introduction

This document describes an imputation reference panel developed using high coverage (30x) sequencing data from the Genomics England (GEL) 100,000 Genomes project, and the imputation of the UK Biobank (UKB) dataset using this GEL reference panel. This imputed dataset has been developed for use by any researcher with appropriate UKB access approval.

A reference panel refers to a set of haplotypes at a dense set of SNPs, indels and structural variants, which can be used to impute genotypes into study samples 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[1,2]. This increases the power of the association study, facilitates meta-analysis and the ability to fine-map causal variants.

The UKB dataset was originally imputed using a combined Haplotype Reference Consortium (HRC) and UK10K reference panel that was developed from relatively low-coverage sequencing datasets[3]. The GEL reference panel benefited from high coverage sequencing and more closely matched ancestry to the UKB participants, showing a significant improvement in imputation accuracy to its predecessor. The resulting imputed UKB autosomal data has 342 million SNPs and short indels, over 4 times more variants than the HRC+UK10K imputed data.

2 GEL phasing and imputation reference panel

2.1 Genomics England 100,000 Genome Project data

The Genomics England 100,000 Genomes Project was launched in 2013, focusing on rare diseases and cancer [4]. Over 120,000 genomes have been sequenced. It comprises genomes from 73,700 rare diseases patients (disorders affecting ≤1 in 2000 persons)[4] and their close relatives, and 46,539 genomes from cancer patients.

The GEL reference panel is built on the aggregated dataset (aggV2), comprising 78,195 samples from both rare disease and cancer germline genomes. Samples are sequenced with 150bp paired-end reads on the IlluminaHiSeq X and processed with the Illumina North Star Version 4 Whole Genome Sequenced Workflow (iSAAC Aligner v03.16.02.19 and Starling small variant caller v2.4.7), and aligned to the GRCh38 human reference genome. The individual gVCF files are aggregated into multi-sample VCF files using Illumina gVCF genotyper and normalised with vt v0.57721. The sample

level quality control has been carried out by Genomics England and details can be found at https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046780 The aggregated multi-sample VCF dataset (aggV2) comprises over 722 million SNPs and short indels (<=50bp). Multi-allelic variants were decomposed into biallelic variants.

2.2 Relatedness and Ancestry

As the GEL individuals have been mainly recruited from hospitals in England[6], the population structure of GEL resembles that of the UK Biobank[3]. **Table 1** provides a breakdown of self-reported ancestry across the dataset. The large number of White British/Irish and relatively large South Asian sample size help to boost phasing and imputation accuracy for these populations[1].

	Self-reported Ethnicity	Number of Samples
White or White British	White British	49,641 (63.48%)
	White Irish	1,048 (1.34%)
	Other White	4,100 (5.24%)
Asian or Asian British	Pakistani	2,885 (3.69%)
	Indian	1,751 (2.24%)
	Bangladeshi	647 (0.82%)
	Chinese	209 (0.27)
	Other Asian	1,180 (1.5%)
Black or Black British	African	991 (1.2%)
	Caribbean	652 (0.83%)
	Other Black	217 (0.27%)
Other	Other ethnic groups	1,151 (1.47%)
Mixed	Mixed	1,446 (1.85%)
Unknown	Unknown	12,277 (15.7%)

Table 1 : Breakdown of self-reported ethnicities.

The sample relatedness in the reference panel is high. According to the self-reported data, only 27,346 samples (34.97%) are unrelated. 11,584 (14.81%), 32,679 (41.79%), and 6,586 (8.43%) samples can find 2, 3 and >3 family members in the dataset. Among the related samples, 17,871 (22.85%) are marked as proband, 15,908 (20.34%) as mother to the proband, 12,409 (15.8%) as father to the proband, 3,149 (4.03%) as siblings to the proband, and 1,512 (1.93%) as other relatedness, such as grandparents or cousin to the proband. The high relatedness was leveraged to apply a powerful Mendel error filter for data quality control, and for accurate phasing of rare variants directly through transmission.

To identify parent-child relationship for phasing we combined information from self-reported relatedness, IBD (identity by descent)[5,6] and Mendel errors. Firstly, 30,000 autosomal variants that meet the following criteria are randomly selected for the analysis. (1) Pass the mean genotype quality and depth filter; (2) pass allele balance filter; (3) missingness < 1%; (4) inbreeding coefficient > -0.1; (5) LD-pruned r^2 < 0.1 with window size of 500Kb; (6) Hardy Weinberg equilibrium test p-value > 0.01; (7) intersect with 1000 Genome phase 3 data; (8) excluding high LD sites identified in Price et al., 2008 study.[10] We then carried out the following procedure on the selected variants. We selected samples with pairwise IBD0 < 0.1 and IBD1 > 0.7 as potential parent/child pairs. For all potential parent/child pairs matching the self-reported relationships, we calculated the Mendel errors, separating duo (parent-child) and trio (mother-father-child) families. The Mendel error cut-off are Q3+1.5IQR, and Q3+4.5IQR for trios and duos in order to identify and remove uniparental-disomy and isodisomy cases. Furthermore, we marked samples as unrelated when it was inconsistent with the self-reported age, i.e. parent should be at least 14 years older than the child. Through this procedure, we identified 12,816 (16.39%) samples are in a duo families and 35,106 (44.9%) in a trio families. As such, 30,273 (38.71%) samples were treated as unrelated for phasing.

2.3 Genotype calling and site filtering

The GEL variants are called individually. A small number of genotyping errors in individuals may cause many false positive sites. In addition to the sample level QC carried out by Genomics England, we applied further site level quality control based on the aggregated VCFs,

- **Genotype quality (GQ) + depth (DP)** : Individual genotypes with either GQ < 15 or DP <10 were marked as missing.
- **Missingness**: remove sites that have missing rate higher than 5%, including the missing genotypes flagged by GQ + DP filter.
- **Allele balance (ABhet):** allele depths (AD) for REF and ALT are expected not to have a huge discrepancy for each heterozygous individual genotypes. We first obtained the allele balance for each genotype, i.e. AD_REF/(AD_REF + AD_ALT). We then counted the number of sites where 0.25 < ABhet < 0.75 and marked them as pass. Sites with less than 75% pass rate were removed.
- **Mendel**: No more than 3 Mendelian errors among all duo and trio families for sites with allele frequency < 0.001 and 7 Mendelian errors for sites with allele frequency >= 0.001
- **Hardy-Weinberg equilibrium (HWE):** Sites where the Hardy-Weinberg Equilibrium (HWE) p-value < 10^{-5} in self-reported White British samples were removed.

- **gnomAD allele frequency (gnomAD)**: We removed sites that showed discrepancy in allele frequency between GEL and gnomAD. To do this we used a Fisher's exact test of the allele frequency difference and a p-value threshold of 10^{-10}.
- **Unrelated singletons**: we removed singletons that did not occur in related families.
- **Additional filters** : we chose a set of more lenient filters for those relatively common sites (AF>0.001) found in the external datasets (TOPMed, HRC, 1000 Genomes)[7–9]. For these sites we used missingness threshold of 25%, a Mendel error threshold of 250 per site and gnomAD allele frequency filter p-values of 10^{-20}. All other filters on GQ, DP, ABHet and HWE were kept as above.

The break-down with the sites removed by each filter are shown in the **Table 2**. The final reference panel has 342,560,554 autosomal variants. The overall Ts/Tv ratio increased from 1.1 to 1.8 after filtering.

	Number of SNPs left after the applying filter (removed %)	Number of Indels/SV left after applying the filter (removed %)	Total number of variants after applying the filter (removed %)
Raw	630,967,910	91,374,497	722,342,407
+ GQ/DP + missingness	428,701,462 (-32%)	55,702,335 (-39%)	484,403,797 (-32%)
+ABhet	411,285,423 (-3%)	42,963,226 (-14%)	454,248,649 (-4%)
+Mendel errors	410,854,761 (-0.07%)	41,905,560 (-1%)	452,760,321 (-0.2%)
+HWE	410,764,722 (-0.01%)	41,868,797 (-0.04%)	452,633,515 (-0.01%)
+gnomaAD	410,628,878 (-0.02%)	41,815,306 (-0.05%)	452,444,184 (-0.02%)
+Singleton	309,825,243 (-16%)	31,639,011 (-11%)	341,464,254 (-15%)
+Additional filters	310,844,262 (+0.16%)	31,716,292(+0.08%)	342,560,554(+0.15%)

Table 2 : Variant filtering. The table shows the effect of each filter applied sequentially from top to bottom. The number of variants (SNPs, Indels/SVs and Total variants) and the percentage removed is shown in each row.

2.4 Comparison of GEL, TOPMed and HRC datasets

The GEL reference panel consists of 342 million autosomal variants, among which 32 million (9.26%) are INDELs with the average length of 4.54 and the maximum length of 50. We compared the GEL reference panel to the widely used TOPMed[7] and HRC[8] reference panels. The GEL panel has 8 times and 1.1 times more variants than the HRC v2 and TOPMed r2 panels respectively. **Figure 2** compares the three datasets overall and in different frequency bins. Limited by low coverage sequencing technology, HRC has very few rare variants with the allele frequency lower than 10^{-4}. The number of rare variants captured are more comparable between TOPMed and GEL, as both

used high coverage sequencing technology. Despite the different sequencing depths and sample sizes, all three panels captured a similar set of relatively common variants ($AF > 10^{-4}$), with less than 4% unique to each panel (**Figure 2** c-d), whereas around half of the rarer variants ($AF \leq 10^{-4}$) from GEL and TOPMed cannot be found in each other (**Figure 2** a-b).

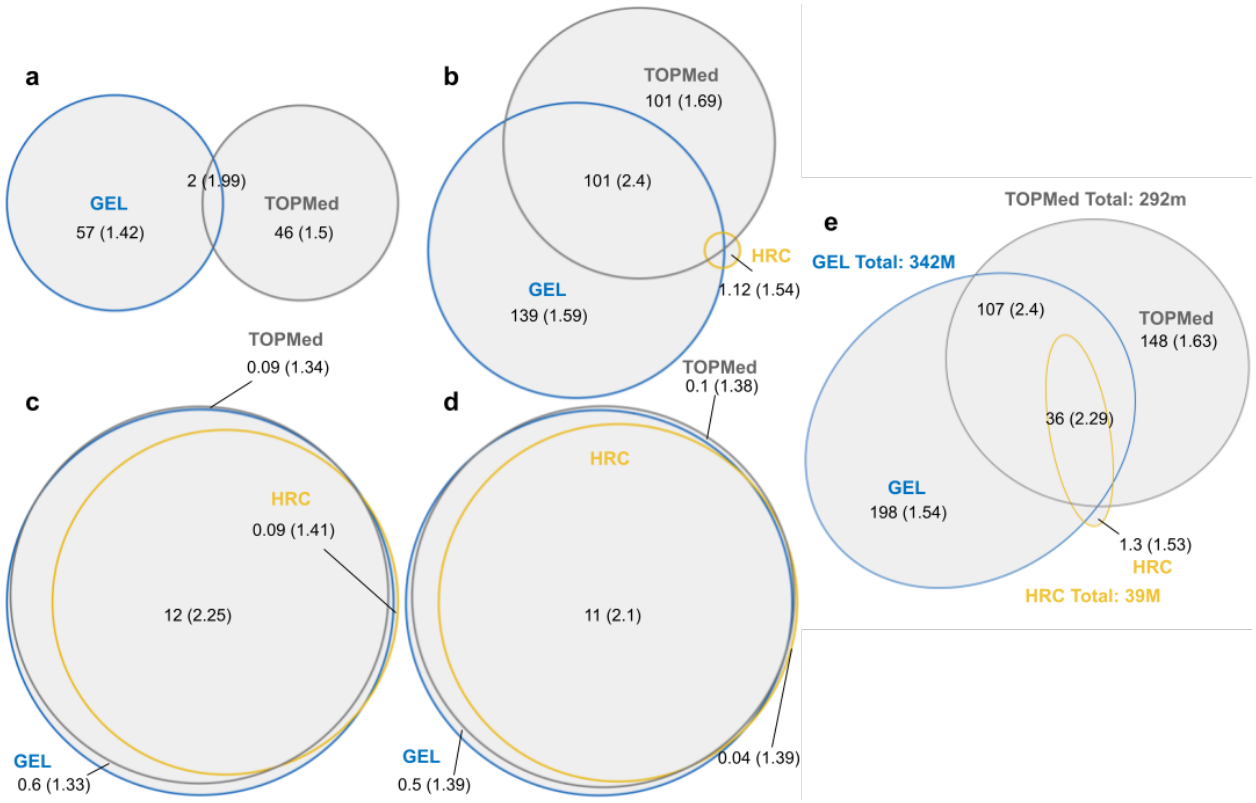

Figure 1: **Venn diagram comparing variants from GEL, HRC and TOPMed reference panel.** Allele frequency for variants existing in more than one reference panel is assigned with the highest allele frequency among all the panels. The Venn diagrams show variants with (a) $AF < 10^{-5}$, (b) $10^{-5} \leq AF < 10^{-4}$, (c) $10^{-4} \leq AF < 10^{-2}$, (d) $10^{-2} \leq AF < 1$, and (e) all variants. The numbers show the variants count of each region (in millions of variants) and followed by Ts/Tv ratio of these variants.

2.5 Reference panel haplotype phasing

Haplotype phasing was carried out using SHAPEIT4.2.2 [10]. We used a multi-stage strategy that leveraged the relatedness within the GEL dataset as much as possible.

In the first stage we used the makeScaffold https://github.com/odelaneau/makeScaffold to determine the phase of as many genotypes in each duo and trio as possible. The vast majority of genotypes can be phased using this process, with a small number of genotypes whose phase is ambiguous due to heterozygosity or missingness patterns. These genotypes were phased using SHAPEIT4.2.2.

To phase the unrelated samples, we first created a phased scaffold of common variants, and then the remaining variants were phased onto this scaffold. To create the scaffold we phased common variants with the minor allele frequency ≤ 0.01, using the phased related samples as the reference panel.

Phasing of the remaining rarer variants were then phased onto the scaffold in chunks containing around 300,000 sites with 30,000 sites on each side as buffer. The phased duo/trio data was used as a reference panel in this step. The chunks and are merged and concatenated using bcftools[11]. Concatenation of the phased chunks is possible as each set of variants have been phased onto the scaffold. The phasing step was computationally intensive and took about 6500 CPU days to accomplish.

In the initial phasing of the dataset, the sites identified using the additional filters (**Table 1**) were not included, and were subsequently phased into the full reference panel in a final step.

2.6 Phasing accuracy

Phasing accuracy could affect the quality of imputation and other downstream applications[1]. We phased the parents of mother-father-child trios from the 1000 Genomes Project[12] using reference panels from HRC and GEL and assessed the phase accuracy from that derived from Mendelian inheritance in each trio. We measured the phasing accuracy using switch error rate, which is the ratio of the number of possible switches required to obtain the true haplotype phase and the inferred one and the number of heterozygotes minus 1, where the phases are inferred. The phasing experiment was carried out on 589 trio families from diverse ethnic backgrounds.

Figure 3 shows the switch error of each sample phased from GEL and HRC panels. The GEL phased haplotypes obtained a lower switch error rate than HRC phased haplotypes for all samples with CEU (Northern European from Utah), African, South Asian and East Asian ancestry. The mean GEL phased haplotype switch error rate is 0.18%, 0.33%, 0.31% and 0.73% for European, African, South Asian and East Asian samples, comparing to 0.22%, 0.43%, 0.31%, 1.07% using HRC reference panel. The

population structure of the reference panel is a key factor in determining the phasing performance. Due to the absence of the South American samples in GEL, HRC outperforms GEL when phasing Peruvians, Mexicans, and Puerto Ricans.

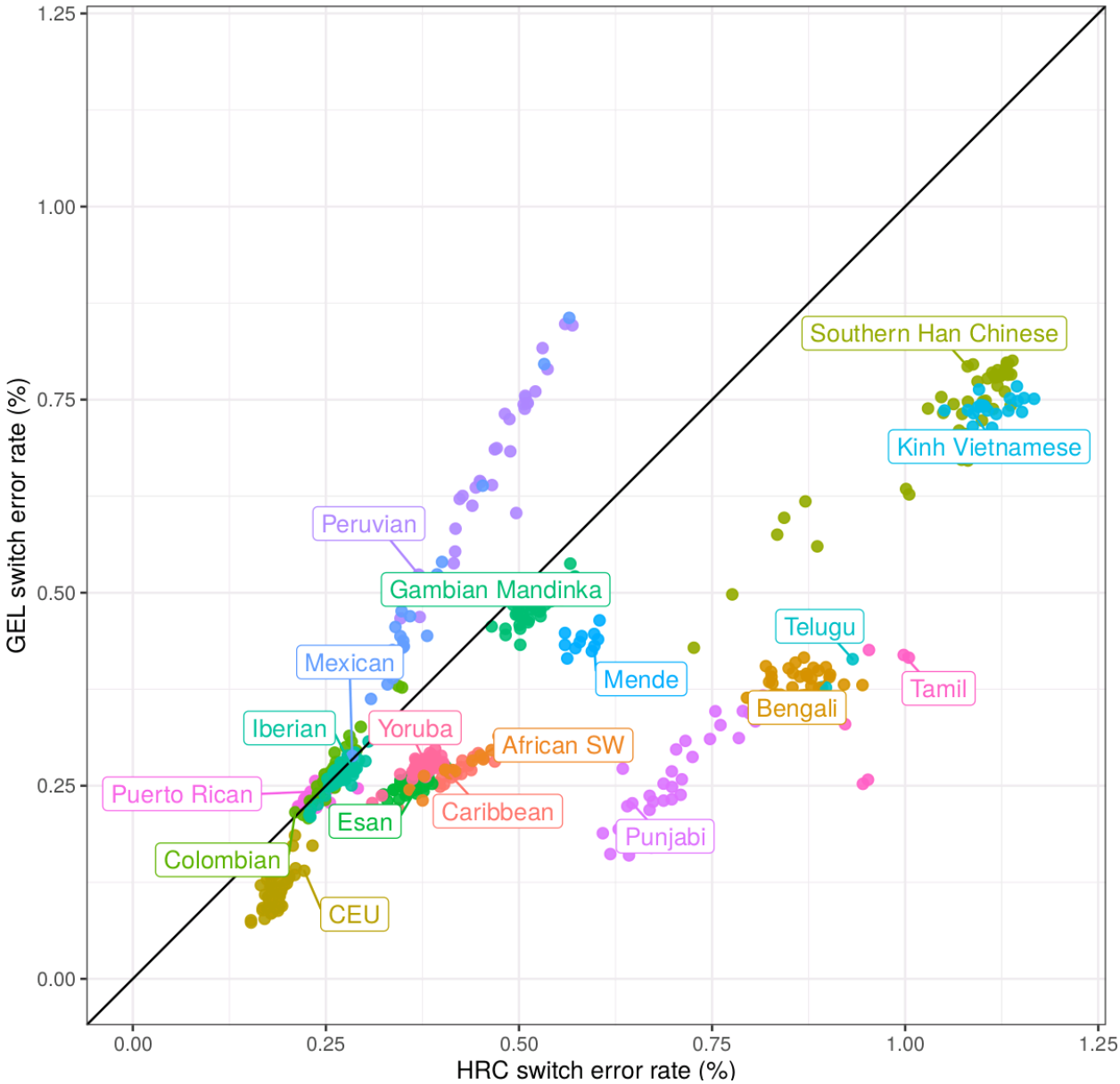

Figure 2: 1000 Genome high coverage sequencing trio data phasing switch error rate comparing haplotypes inferred by trio transmission and computationally inferred using reference panels.

2.7 Imputation accuracy

We assessed the utility of the GEL reference panel for imputation, compared to the TOPMed and HRC reference panels. We used high coverage, whole-genome sequencing data from 1000 genome samples[12]. SNP positions in the UK Biobank Axiom array[3] were used to create a pseudo-SNP array dataset, masking genotypes in 1000 Genome sequencing samples, except the sites existing in the Axiom array. We performed Hardy Weinberg equilibrium tests (HWE) within each of the 26 1000 Genome populations, filtering out sites which failed any of these tests with the p-value $< 10^{-10}$, resulting 716,473 bi-allelic SNPs across the whole autosome. The pseudo-SNP array dataset was then phased one chromosome at a time using SHAPEIT4[10].

The HRC reference panel was lifted over from GRCh37 to GRCh38 using GATK Picard LiftoverVCF[13]. The resulting GRCh38 HRC reference has 39,115,765 autosomal variants. TOPMed imputation is carried out using the TOPMed imputation server with TOPMed r2 reference panel[7]. We used IMPUTE5 [14] to impute from the GEL and HRC panels. We converted the reference panels to the IMP5 format to facilitate efficient imputation[14].

The imputed genotypes from the GEL, TOPMed and HRC reference panels were then compared to the sequencing data as the ground truth, stratified by allele frequency. Squared correlation r^2 between the imputed allele dosages and the 1000 Genome sequencing data were calculated, stratified by gnomAD (v3.3.1) minor allele frequency. As we focus on showing the overall performance of the reference panel across different allele frequencies, only variants that are in common with the gnomAD variants are taken into account. As a result, the number of variants measured may differ across reference panels. We also stratified the imputation results from the 2405 1000G samples into 6 groups : 661 African (AFR), 347 American (AMR) , 504 Eastern Asian (EAS), 489 South Asian (SAS), 313 non-Finnish European (NFE) samples and 91 British (GBR) samples.

Figure 4 shows the imputation results. The GEL imputation accuracy outperforms HRC panel in all allele frequency bins for all ethnicities. The GEL panel out-performs the TOPMed panel in GBR and SAS samples, especially for rarer variants. This is likely due to the GEL panel having a larger number of individuals with British and South Asian ancestry than the TOPMed panel. At MAF<10^{-5}, the GEL imputation performance (r2) for GBR samples is 0.6, compared to 0.3 and 0.29 using TOPMed and HRC, respectively. At MAF<2×10^{-4}, the r^2 are 0.75, 0.64, and 0.48 for GEL, TOPMed and HRC, respectively. These results suggest that the GEL reference panel is a well matched reference panel to impute the UK Biobank samples, of which over 80% are White British or Irish, and for which South Asian is the second largest ethnic group. The

TOPMed panel out-performs GEL and HRC in African, American and East Asian samples.

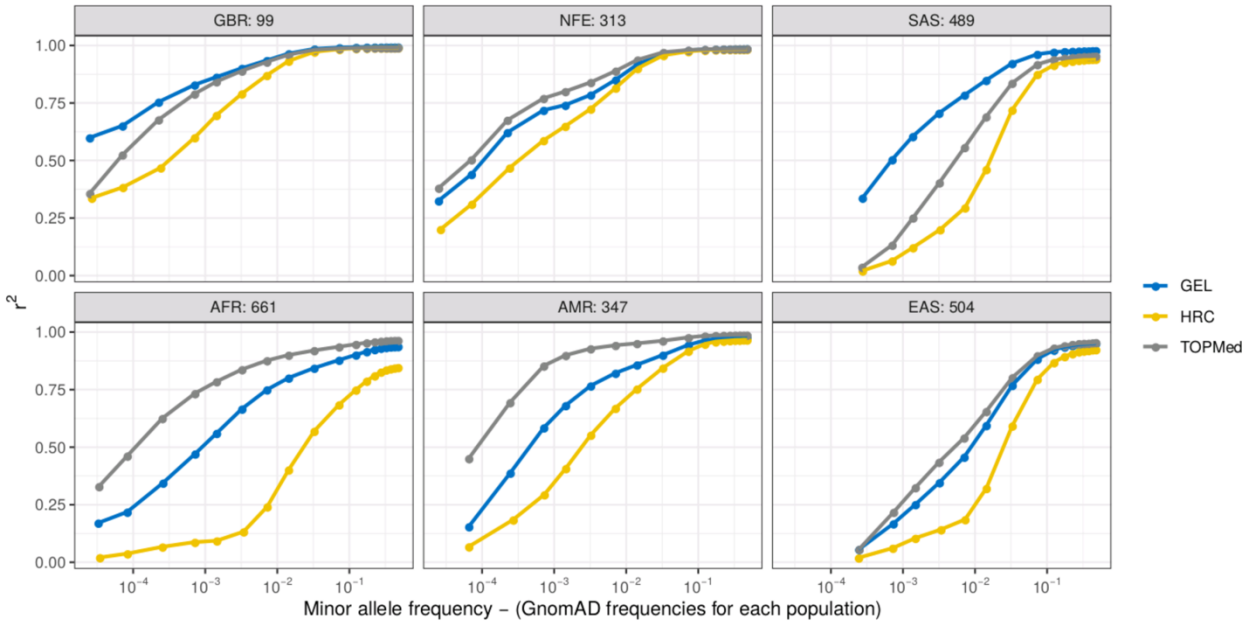

Figure 3: **Comparison of imputation performance using different reference panels.** The x-axis shows non-reference allele frequency on a log scale, focusing in on rarer variants. The y-axis is imputation performance (r^2). The performance of the reference panels HRC (yellow), TOPMed (grey), GEL (blue) are shown as lines in each plot. The variants are stratified by GnomAD allele frequency (v3.3.1) of their corresponding population.

3 UK Biobank imputation

3.1 UK Biobank SNP array data quality control and phasing

The UK Biobank SNP array data consists of 784,256 autosomal variants. We removed the set of 113,515 sites identified by the previous centralized UK Biobank as failing quality control[3], but also removed an additional set of 39,165 sites failing a test of Hardy-Weinberg equilibrium in 409,703 White British samples, with the p-value threshold of 10^{-10}. The SNP array data was also lifted over from GRCh37 to GRCh38 using GATK Picard LiftOver tool[13]. The alleles with mismatching strand were flipped wherever possible. 495 variants are removed due to the incompatibility between the two reference genomes, resulting into the final SNP array data with the size of 631,081 autosomal variants to be phased and imputed.

Haplotypes estimation of the SNP array data is a prerequisite for imputation. Phasing is carried out one chromosome at a time using SHAPEIT4.2.2 without a reference panel. We ran SHAPEIT4 using its default 15 MCMC iterations and 30 threads. The runtime varies from 2 hours to 30 hours for each chromosome.

3.2 UK Biobank imputation using GEL reference panel

The autosomal imputation by GEL reference panel was carried out using IMPUTE5 (v1.1.4). The SNP array data was divided into 408 consecutive and overlapping chunks across the genome and each chunk was further divided into 24 sample batches with each batch contains 20,349 samples. IMPUTE5 was run on each of the 9,792 subsets using a single thread and default settings, at a speed less than 4 minutes per genome, totalling about 1200 CPU days to impute all UK Biobank samples. Sample batches were merged using QCTOOL and the non-overlapping chunks were concatenated using cat-bgen[17].

3.3 Imputation quality

IMPUTE information (INFO) assesses the genotype imputation uncertainty ranging from 0 to 1. A high INFO implies higher imputation quality, with a value of 1 indicating no uncertainty, and 0 complete uncertainty about the genotype imputation. If an imputed variant on N samples has an INFO scored at α, it implies that the statistical power of association tests are approximately equivalent to αN perfectly observed genotype data[1]. There is no single correct answer for which threshold to use. To perform GWAS on the UK Biobank data with 500,000 samples, it's typical to use the variants with INFO higher 0.3, equivalent to >150,000 perfectly observed samples.

We compared GEL imputed UK Biobank INFO scores to those from the existing HRC+UK10K imputed dataset[3]. The proportion of GEL imputed variants passing the INFO threshold of 0.3 are 8%, 78% and 98% for MAF<=0.0001%, 0.0001%<MAF <=0.001%, and 0.001% < MAF<=0.01%, compared to 4%, 54%, and 78% for the HRCUK10K imputed data **(Figure 4)**. Among the 65 million variants imputed by both GEL and HRCUK10K panels, 87% achieved better INFO score using the GEL panel (**Figure 6**).

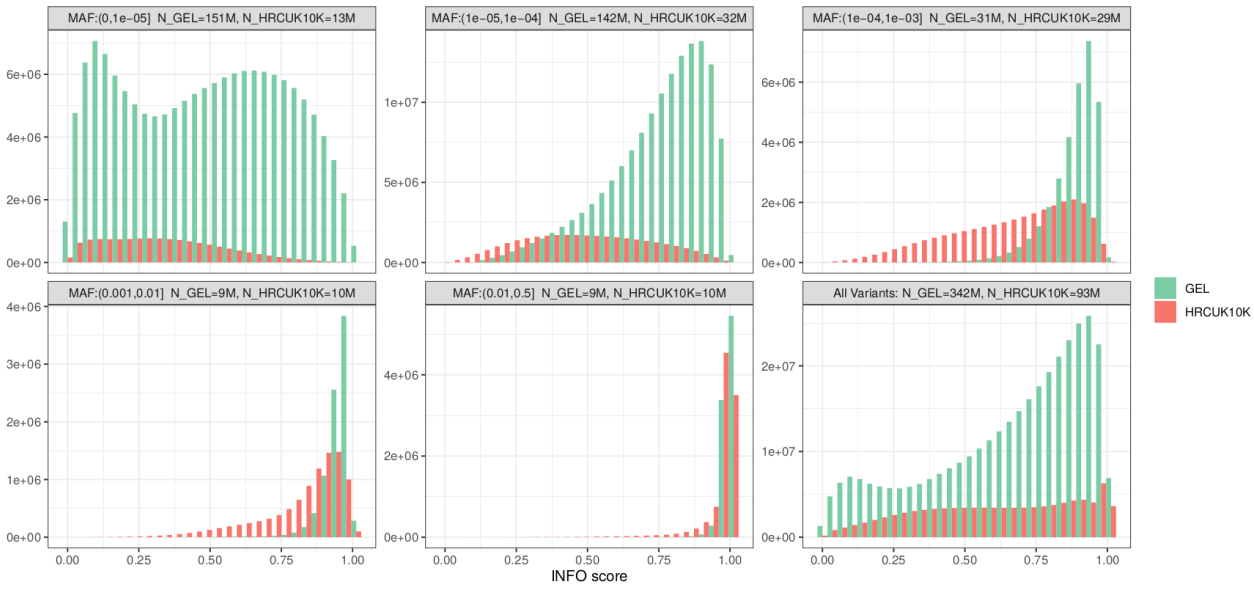

Figure 4: Imputation INFO score histogram comparison between GEL and HRCUK10K imputed UKB data. Each panel shows the distribution of INFO scores for GEL and HRCUK10K imputed variants in different MAF bins. The total number of variants in each bin is provided in the panel legend.

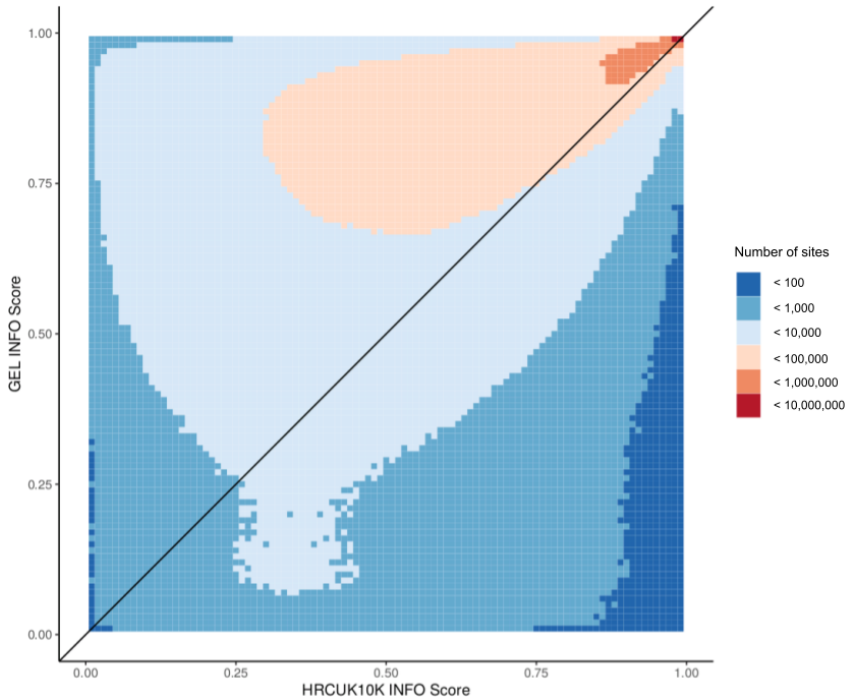

Figure 5: Comparison of INFO scores at sites in both GEL and HRC+UK10K reference panels. A heatmap of the scatter plot of UKB INFO scores from the 65 millions sites in both GEL and UK10K panels.

3.4 BGEN files

The imputed UK Biobank data for all autosomes, consisting 342,573,817 variants in 488,315 samples, are stored in 22 **8-bit zstd compressed BGEN files**[15]. Variants are aligned to the GRCh38 reference genome. Each chromosome is stored in a separate BGEN file with the file size ranging from 19Gb to 103Gb, and totalling of 1.2Tb. All variants are assigned with a unique ID, either using the rsid from dbSNP build 155 (https://ftp.ncbi.nih.gov/snp/latest_release/) wherever possible, or in the form of chr:pos_ref_alt, when an rsid is not available for the variant.

Imputed genotypes are stored as genotype posterior probabilities, consisting three 8-bit floating number, representing the probability of the genotype being homozygote reference, heterozygote or homozygote alternate, respectively. Despite having more than 3 times the number of variants, the GEL imputed UKB BGEN files are nearly half of the size to its predecessor, HRCUK10K imputed BGEN files (2.2Tb). This is mainly due to the storage method, of which the genotype dosages are stored in 8-bit instead of 16-bit. The compression of the rare variants is highly efficient using the BGEN format.

4 Funding and Data Access

The UK Biobank data was obtained via approved application number 48031. Access to the GEL dataset was via the GEL Research Environment. The UK Biobank dataset was uploaded to the GEL Research Environment and all analysis was carried out there. Genomics England and the 100,000 Genomes Project was funded by the National Institute for Health Research, the Wellcome Trust, the Medical Research Council, Cancer Research UK, the Department of Health and Social Care and NHS England. The research formed part of the work carried out by Genomics England Clinical Interpretation Partnerships (GeCIP) Population Genomics domain. Sinan Shi was funded via a Wellcome Trust Collaborator award to Jonathan Marchini, Simon Myers and Garrett Hellenthal (200186/Z/15/Z). We thank the participants of the 100,000 Genomes Project and the UK Biobank Project who made this study possible.

5 References

1. Marchini, J. & Howie, B. Genotype Imputation for Genome-Wide Association Studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

3. Bycroft, C. *et al.* The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **562**, 203–209 (2018).

4. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

5. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

6. Browning, S. R. & Browning, B. L. High-Resolution Detection of Identity by Descent in Unrelated Individuals. *Am. J. Hum. Genet.* **86**, 526–539 (2010).

7. Taliun, D. *et al.* Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

8. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

9. Auton, A. *et al.* A Global Reference for Human Genetic Variation. *Nature* **526**, 68–74 (2015).

10. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

11. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

12. Byrska-Bishop, M. *et al.* High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. *bioRxiv* 2021.02.06.430068 (2021) doi:10.1101/2021.02.06.430068.

13. DePristo, M. A. *et al.* A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data. *Nat. Genet.* **43**, 491–498 (2011).

14. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype Imputation Using the Positional Burrows Wheeler Transform. *PLOS Genet.* **16**, e1009049 (2020).

15. Band, G. & Marchini, J. BGEN: A Binary File Format for Imputed Genotype and Haplotype Data. 308296 (2018) doi:10.1101/308296.