A Longitudinal Resource for Genetic Research in Behavioral and Health Sciences

Imputation Report

1000 Genomes Project reference panel (Phase 3)

November 2nd, 2016

Contents

I. Summary and recommendations for dbGaP users

Genotype imputation is the process of inferring unobserved genotypes in a study sample based on the haplotypes observed in a more densely genotyped reference sample[1,2]. The University of Washington Genetic Analysis Center (GAC) used IMPUTE2 software[3,4] to perform genotype imputation in the Wisconsin Longitudinal Study (WLS). This report provides a detailed account of data preparation and imputation; describes the imputation output, including file formats and quality metrics; and makes recommendations for downstream analyses.

Imputed results are provided as the probability of each of the three genotype states at each variant for every study participant. We recommend incorporating these imputed probabilities into any downstream analyses (e.g., as allelic dosages), rather than collapsing the probability information into the most likely genotype. Quality metrics are provided that can be used for filtering imputation results on a per-variant basis. For additional background information and a detailed description of genotype quality control (QC) on this project, please see the genotype QC report available through dbGaP (database of Genotypes and Phenotypes).

II. Study data

a. Samples

The WLS is a long-term study of a random sample of men and women who graduated from Wisconsin high schools in 1957 and their siblings. The addition of genetic data to WLS creates opportunities for genetic studies of aging, behavior, cognition, personality, mental health, health, disease, and mortality. A total of 9,606 study samples, including duplicates, were put into genotyping production at the Center for Inherited Disease Research at Johns Hopkins University. After the GAC's standardized QC procedures[5], including resolution of sample quality and identity, genotypes are available on dbGaP for 9,012 unique WLS study participants. (Note for 15 pairs of monozygotic twins, only one member of each pair is retained in the unique set of 9,012 participants.)

The GAC QC procedures yielded a set of recommended SNP and sample filters, as described in the genotype QC report. These recommended filters were then used to select study samples and SNPs for imputation. Typically we exclude from imputation samples with missing call rate (MCR) > 2%; however, no samples failed this criterion. Therefore, the imputation included all 9,012 unique genotyped study participants posted to dbGaP.

For samples with gross chromosome anomalies affecting an entire chromosome, those samples were removed from the affected chromosome's imputation but imputed for all non-affected (i.e., non-anomalous) chromosomes. As a result, one sample was excluded from chromosome 15 imputation, three from chromosome X imputation, and fifteen from imputation of the pseudo-autosomal regions PAR1 and PAR2. Note partial chromosome anomalies were addressed in post-imputation processing, described in section VI.

Figure 1 shows a principal component analysis (PCA) of 9,018 unique study participants with population reference samples from the International HapMap Project[6]. (Note the difference between 9,018 samples in the PCA and 9,012 in the imputation is due to six participants who withdrew consent.) This PCA was done to establish ancestry orientation of study samples and presented here to visualize macro patterns of genetic ancestry. In Figure 1, study samples are color-coded by self-identified race

group, a variable that is intentionally not included in the dbGaP posting per WLS Institutional Review Board (IRB) stipulation.

WLS participants are recruited from a relatively homogeneous population of predominately Northern and Western European ancestry. Therefore, as expected, the majority of study samples self-identify as "white" and cluster in the PCA space with HapMap reference populations of European descent. There are, however, WLS participants self-identifying as American Indian/Alaska Native, Asian, Black or African American, or as more than one race. For the most part, these samples cluster in the PCA space with the reference population(s) nearest to their self-identified race group. The IMPUTE2 algorithm discussed below recommends the use of a worldwide or "cosmopolitan" reference panel, irrespective of the genetic ancestry composition of study samples. Thus, we imputed all study samples together in one group, to the same worldwide reference panel.

The local subject level identifier ("SUBJECT_ID" in annotation files) was used as the individual identifier throughout, which can be mapped to the local sample-level identifier ("SAMPLE_ID" in annotation files) using the sample-subject mapping files provided in the Supplementary Files (section XII).

b. Variants

This project was genotyped on the Illumina OmniExpress array, humanomniexpress-24-v1-1 annotation version A, designed to human genome build 37/hg19. For the purposes of imputation, study variants were selected using GAC-recommended quality filters described in the genotype QC report. In addition to requiring variants to pass the quality filter ("quality.filter"=TRUE), we further restricted to variants with (1) known chromosome and position (i.e., exclude unmapped variants with chromosome or position = 0); (2) located on chromosomes 1-22, X, or XY (pseudo-autosomal); and (3) with unique positions, which involved removing where "redundant"=TRUE and/or "dup.pos.disc"=TRUE. A summary of initial input variants is shown in Table 1; a list of these variants is available in the Supplementary Files.

Observed genotypes, which have a maximum probability of 1, are included in the imputation output. Where an observed study variant had sporadic missing data, the missing genotypes were imputed by the pre-phasing software. Additionally, variants genotyped in the study but not used as imputation input (i.e., not passing the pre-imputation quality filters) may also appear in imputed results when available in the reference panel. This data formatting pipeline could result in discrepancies between observed genotypes posted in the primary genotype release and these imputed data. The variant annotation files accompanying this report can be used to differentiate between observed study variants used in the imputation input and the imputed variants. We refer to the former set of variants as the "imputation basis" and to the latter as the "imputation target." These terms are analogous to the IMPUTE2 definitions of "type 2" and "type 0" variants, respectively. (Note that "type 1" variants occur only when more than one reference panel is used with IMPUTE2.) Lastly, we refer to study variants that do not occur in the reference as "study only," or "type 3" in IMPUTE2; these are also retained in imputation output. See Figure 2 for a visual representation of these variant types.

c. Indel processing

The OmniExpress array contains one insertion/deletion variant, or "indel." Indels require extra processing to prepare the study dataset for imputation to a 1000 Genomes reference panel. Specifically,

alleles and — in some instances — base pair positions given in the Illumina manifest must be transformed for the imputation software to identify overlap between variants in the study data and variants in the reference panel. For example, Illumina describes indel alleles with a "-" for the deletion, e.g. "A/-," whereas the VCF (variant call format) convention used in 1000 Genomes lists the reference base directly upstream of the deletion, e.g. "GA/G." Harmonizing study dataset indel annotation with 1000 Genomes requires knowing the map position of indels. However, the one OmniExpress indel was unmapped in the Illumina manifest and thus not able to be harmonized. Unmapped variants are also excluded from the imputation basis, so the lack of aligned indel alleles and positions had no effect on the imputation.

d. Data formatting

The study genotype data were initially accessed from a binary PLINK[7] file with genotypes expressed in TOP alleles. PLINK data formatting proceeded in two steps. First, at the level of the genome-wide binary PLINK file, we used genomic strand information to identify and flip the strand of SNPs where the TOP alleles were not aligned to the plus ("+") strand of the human genome reference assembly (see section IV). In this first PLINK step we also updated monomorphic variants, where the PLINK file initially shows a "0" as one of the two alleles, to instead replace that "0" with the minor (i.e., unobserved) allele. This was done to allow the imputation software to find overlapping variants between the study and reference based on matching positions and alleles.

As a second PLINK formatting step, we divided the PLINK dataset into chromosome-specific binary files. In this step, we also (1) set haploid genotypes (male chromosome X) called as heterozygotes to missing; (2) extracted study samples, i.e., removed genotyping controls and any samples with a whole chromosome anomaly on the given chromosome; and (3) extracted only study variants selected for imputation. Below is an example of the command line syntax used to create the filtered binary files, for generic chromosome "#":

```
plink --bfile Herd_WLS_plusAlleles \
--extract snp.qualfilter.txt --keep sampkeep.txt \
--set-hh-missing --chr # --make-bed --–out Herd_WLS_chr#
```

e. Pre-phasing

When genotype imputation was first introduced, phasing and imputation were done jointly in a unified process. Since then, the alternative approach of "pre-phasing" has emerged as a practical way to maintain imputation accuracy while minimizing computation time, as available reference panels increase in number and in size[8]. Pre-phasing involves phasing the diploid study data prior to imputation and is amenable to most any pairing of phasing and imputation software. The computational arguments for pre-phasing are that (1) imputing into pre-phased haplotypes is much faster than imputing into unphased genotypes and (2) pre-phased data facilitates future updates to imputation, as improved reference panels become available. Although pre-phasing may introduce a small loss of accuracy, due to the lack of incorporating haplotype uncertainty information into the imputation step, the advantages appear to outweigh the disadvantages for most genome-wide imputations.

An additional advantage to pre-phasing in family studies is that it can incorporate family structure, while most imputation algorithms "ignore" relatedness due to both computational and programmatic constraints. However, most phasing software can incorporate family structure, such that pre-phasing enables use of family information during at least the phasing step if not the imputation step. With the SHAPEIT2[9] pre-phasing software used in this imputation project, parent-offspring (PO) relationships are used when explicitly specified. Furthermore, other non-specified relationships (full siblings and half-sib-like relationships, e.g.) also help to improve pre-phasing, as the software is able to detect long stretches of shared chromosomes despite not "knowing" up front about these other family structures.

We used a pre-phasing approach in this imputation, due to the advantages enumerated above. The WLS genotyped cohort comprises 2,263 families of two to four members each, the majority of which are full sibling pairs. However, no explicit family structure was present in the PLINK file provided to SHAPEIT2, for two reasons: (1) due to IRB stipulation, no pedigrees are being released for this study and (2) the lack of PO pairs among study samples means no explicit pedigree information would have been used by SHAPEIT2 anyway. Note SHAPEIT2 can still leverage non-specified relatedness to inform and improve pre-phasing, as discussed above.

We input the filtered, chromosome-specific PLINK files (see II-d) to SHAPEIT2, which returned the best guess haplotypes for each chromosome. These best guess haplotypes were then fed directly into the IMPUTE2 imputation. SHAPEIT2 jobs were run multi-threaded across 12 compute cores. Runtimes ranged from 4 to 35 hours, depending on the size of the chromosome. The two pseudo-autosomal regions of chromosome X (PAR1 and PAR2) were each pre-phased separately. Below is an example of the command line syntax used to run the SHAPEIT2 program on a generic chromosome "#":

```
shapeit2 -B Herd_WLS_chr# \
-M genetic_map_chr#_combined_b37.txt \
-O Herd_WLS _chr#.haps.gz Herd_WLS_chr#.sample.gz \
-S 200 –T 12 -L shapeit_chr#.log
```

III. Reference panel

Larger reference panels have been shown to increase imputation accuracy[2,4,10]. Initially, haplotypes from Phases 2[11] and 3[6] of the International HapMap Consortium served as the reference panel for many imputation analyses. Advancements in genome-wide resequencing technology have since yielded alternatives to these historically standard HapMap panels —namely the 1000 Genomes Project — enabling the imputation of many more and rarer variants[2,12].

The 1000 Genomes Project aims to "discover, genotype, and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations"[13]. The Project has proceeded in several stages. First, the pilot phase included low-coverage whole-genome sequencing in 179 samples, from which ~15M variants were detected. The pilot project was described in a 2010 manuscript[13]. Next, phase 1 was published in late 2012, which comprised ~39M variants (SNPs, indels, and structural variants) in 1,092 samples from 14 different populations worldwide[14]. The final phase, phase 3, was released in September 2014 and includes ~81M variants in 2,504 samples from 26 populations worldwide[15]. The Project has categorized each of the phase 3 populations into five

continental groupings or "super populations": African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). Sample counts by population and by continental panel are shown in Table 2. The variant calls in this final dataset are based on both low coverage whole genome and deep coverage exome sequence data. Additionally, phased haplotypes were generated using a "scaffold" of high quality SNP array data, a process which has been shown to yield higher quality imputation compared to phasing without a scaffold of array SNPs[16].

To impute the WLS study participants, we used a worldwide reference panel of all 2,504 phase 3 samples. We downloaded these reference panel data from the IMPUTE2 website (see Web Resources). These data files were created by IMPUTE2 authors from the VCF files available from the Project, using version 4 of the phase 3 release.

The IMPUTE2 method enables the computationally efficient use of all available reference panel samples, bypassing the problematic step of *a priori* choosing the mixture of haplotypes most representative of the study samples. Instead, when given a worldwide reference, IMPUTE2 will select an appropriate subset of the available reference haplotypes for each study haplotype in each genomic region[4]. While this approach eases the computational burden of using all reference samples, it still may not warrant the imputation of all available reference variants (i.e., approximately 81M variants). Very rare variants are both harder to impute and, even if imputed error-free, it is unlikely most studies will be sufficiently powered to detect an association in downstream analyses. There is also the concern that a variant observed only once may reflect a sequencing error rather than a true variant.

Here we restricted imputation to only variants with minor allele count (MAC) of at least four in any one of the five super populations (AMR, AFR, EAS EUR, or SAS). The EUR super population alone was not selected for this MAC-based filter given that not all WLS participants are of primarily European descent. To filter imputation target variants based on minor allele counts across multiple panels, we manually edited the .legend files available from the IMPUTE2 website. Specifically, we added an additional column "ma.cnt.gte4.allpanels" which contained a "1" for variants passing the filter (i.e., with at least four copies of the minor allele in at least one panel) and "0" otherwise. We included all three variant types (SNPs, indels, and structural variants) in this imputation, based on findings from the 1000 Genomes Project that imputation accuracy at indels and structural variants (SVs) can be comparable to that of SNPs[14].

IV. Strand alignment

Accurate imputation is dependent upon the study and reference panel allele calls being on the same physical strand of DNA relative to the human genome reference sequence. In practice, however, this crucial step is not always straightforward[17]. The initial study dataset contained TOP alleles, an Illumina naming method unrelated to "+" or "-" strand orientation[18] (also see Web Resources, Illumina 2006). Because all 1000 Genomes reference panel data are expected to be "+" strand relative to the reference, we first needed to convert TOP alleles to the "+" strand. Illumina annotation typically contains the information necessary to perform this conversion: the "RefStrand" column that indicates whether the design alleles are on the "+" or "-" strand. Here we used the array manifest, "humanomniexpress-24-v1-1-a_Anno.csv," to identify which SNPs to flip to convert a TOP allele PLINK dataset to plus ("+") strand alleles.

To ensure strand consistency, IMPUTE2 automatically addresses strand alignment at strand unambiguous SNPs (i.e. not A/T or C/G variants) by comparing allele labels. That is, where a strand unambiguous SNP in the study data is found to have different nucleotides compared to the reference panel, the strand is flipped in the study data. Notably, only four strand unambiguous SNPs were flipped by IMPUTE2 during the imputation. We did not, however, invoke the additional, optional strand alignment check "-align_by_maf." This option compares MAF between the reference and study samples at strand ambiguous SNPs (A/T or C/G) and, where necessary, flips the study data to make the minor alleles consistent. This method may be prone to erroneous strand flips at strand ambiguous SNPs with MAF close to 50%. Another disincentive for using the "align_by_maf" option is that allele frequencies are likely to differ between study and reference samples due to different genetic ancestry composition. Thus, we instead chose to rely on the SNP annotation to align strand-ambiguous SNPs to the plus strand, with the expectation that this approach would yield fewer strand misalignments compared to invoking the "align_by_maf" flag.

V. Imputation software and computing resources

Imputation analyses were performed using IMPUTE version 2.3.2, a freely available software program (see section IX, Web resources).

a. Imputation segments

We imputed chromosomes in segments due to (1) IMPUTE2 reports of improved accuracy over short genomic intervals and (2) our desire to expedite imputation by parallelizing jobs over a multi-core compute cluster. Segments were defined in an iterative process, following a series of recommendations set forth by IMPUTE2 authors. (Note that while IMPUTE2 provides recommendations for the segmentation method, it is up to the user to define the segments.) We first created 5 MB segments over the length of each chromosome. Next, at segments spanning a centromere we divided the segment at the centromere and merged each of the two partial segments into the segment either immediately up- or down-stream. This avoided segments spanning the centromere. Last, we checked segments for presence of observed study variants, in light of the logical recommendation from IMPUTE2 authors that each segment contain at least some imputation basis (i.e. type 2) variant.

Ultimately we divided 23 chromosomes into 547 total segments, ranging from 7 segments on chromosome 21 to 47 segments on chromosome 2. The two pseudo-autosomal regions PAR1 and PAR2 were each imputed as a separate segment. Using OmniExpress array variants passing the basic quality filter, we calculated a mean density of 1,271 study variants per segment (range 71 – 4,931), which we took as evidence that imputation basis variants would be adequately represented in our imputation segments. The segmentation scheme is available as a supplementary file.

b. IMPUTE2 settings

The IMPUTE2 algorithm uses a "k_hap" value to specify which number of reference haplotypes should be used to impute each study sample. The approach contributes to the computational feasibility of imputation with a worldwide reference panel: i.e., the full set of reference samples is available to the imputation software, but it "chooses" a subset of reference haplotypes to impute each study sample based on perceived genetic similarity (for details, see Howie, Marchini, and Stephens, 2011[4]). The

default k_hap value is 500; however, higher values are recommended when imputing into admixed populations. We and others have also found that increasing k_hap is unlikely to cause harm, except in the instance where maximizing k_hap (i.e., forcing the use of all available reference haplotypes) can yield a slight decrease in accuracy[19]. For this imputation we set k_hap to 2,000, ~40% of available reference haplotypes, which for past projects we have found to be a reasonable compromise between accuracy and computation time.

c. Running IMPUTE2

An example of the command line syntax used to run IMPUTE2 on chromosome 22 is shown below. The file specified by the `-known_haps_g` flag is the phased haplotypes output by SHAPEIT2. The `-h` and `-l` flags refer to the 1000 Genomes reference panel files (the haplotypes and variant legend files, respectively). The `-filt_results_l` imposes the MAC-based variant filter.

```
impute2 -use_prephased_g -m genetic_map_chr22_combined_b37.txt \
-h 1000GP_Phase3_chr22.hap.gz \
-l 1000GP_Phase3_chr22.legend.gz \
-int 20000001 25000000 -allow_large_regions \
-known_haps_g Herd_WLS_chr22.haps.gz \
-k_hap 2000 -filt_rules_l ma.cnt.gte4.allpanels<1 \
-o Herd_WLS_chr22.set2.gprobs -os 0 2 3 -o_gz \
-i Herd_WLS_chr22.set2.metrics –verbose
```

Imputation jobs were run in parallel on a compute cluster consisting of 12 compute nodes, each containing two Intel Xeon E5645 Six-Core processors (12 MB cache), 96 GB of memory, and 1.5 TB of local storage. The average runtime for each imputation segment was 8.4 hours (range < 1 – 19 hours).

VI. Imputation output

Imputation output files are divided by chromosome, where “23” denotes chromosome X. Each of the two pseudo-autosomal regions (X_PAR1 and X_PAR2) has a separate output file. For more information on the file formats described below, see Web Resources: “IMPUTE2 file format descriptions.” In addition, data dictionaries for each of these output file types are included in the imputation data release.

a. Sample files

Because the imputation output files do not contain sample identifiers, separate “.sample” files are provided to indicate the order of samples in the imputation results. That is, the order of samples in the “.sample” files matches the order of individuals in the imputation output. In these .sample files, ID_1 is the family ID and ID_2 is the individual ID_2. In this project, ID_1 and ID_2 are both equal to the local subject-level ID: “SUBJECT_ID” in annotation files.

b. Phased haplotypes

Phased haplotypes output by SHAPEIT2 are provided in gz-compressed “.haps” files. These haplotypes include all genotyped study variants selected for imputation. We additionally provide the phased haplotypes in VCF format, created with the SHAPEIT2 “-convert --output-vcf” flag. This alternate format

is provided to enable downstream use with the cloud-based imputation servers, such as the one hosted at the University of Michigan (https://imputationserver.sph.umich.edu) that accepts pre-phased VCF files as input. Imputation servers enable imputation to the Haplotype Reference Consortium[20].

c. Genotype probabilities

Imputation results are posted in chromosome-specific genotype probabilities files (".gprobs," also gz-compressed). Our first step in creating these files from the native IMPUTE2 output was to zero out any imputed genotypes in regions affected by gross chromosomal anomalies (see section 7 of the genotype QC report for details on anomaly detection). A sample's genotypes were zeroed out across the entire length of any imputation segment overlapping with or containing a gross chromosomal anomaly. Included in the supplementary files section of this report are (1) the chromosome and base pair coordinates of each imputation segment and (2) a list of all anomalous subject-segment combinations where imputed genotypes were set to missing (i.e., 0 0 0). Note that where the whole chromosome was recommended for filtering due to an anomaly, the sample was excluded from imputation on the given chromosome. After imputation segments were processed for anomalies, they were combined into per-chromosome .gprobs file, via the Unix "cat" command.

The first five columns in these output files correspond to SNP ID; rs ID; physical position; and the two alleles, where the first allele shown is designated "allele A" and the second is designated "allele B." Each subsequent set of three columns corresponds to the genotype probabilities of the three genotype classes (AA, AB, and BB) for a single individual. These genotype files contain three variant types as defined in the IMPUTE2 algorithm: type 0 (imputation target), type 2 (imputation basis), and type 3 (study only). The type for each line of the genotype probabilities files can be determined using the accompanying metrics files. Note there are no sample identifiers in the probabilities files, necessitating the use of auxiliary files to align imputed probabilities with sample information (see VI-a, above).

d. Quality metrics

Each genotype probabilities file is accompanied by a variant annotation and quality metrics file, with each row of a genotype file corresponding to a row in the variant annotation file. These metrics files were output by IMPUTE2 (the "-i" or info file); the only modifications we made were to (1) combine segmented files into one metrics file per chromosome and (2) delete the uninformative "snp_id" field. Columns in these files are defined below, based on IMPUTE2 online documentation (see Web Resources).

- **rs_id:** Variant identifier. For imputed variants in dbSNP and thus with a reference SNP (rsID) number, this identifier is constructed as "*rsID:position:ref:alt*", where "ref" and "alt" are the reference and alternate allele, respectively. For imputed variants not in dbSNP, this identifier is constructed as "*chr:position:ref:alt*." For type 2 (imputation basis) and type 3 (study only) variants, this field will be the identifier used in the input study dataset.
- **position:** Base pair position (GRCh37)

- **a0**: allele A and also the reference allele, following the convention in the 1000 Genomes Project reference panel files. Note type 3 variants do not have the same correspondence between allele A and reference.
- **a1**: allele B and also the alternate allele, following the convention in the 1000 Genomes Project reference panel files. Note type 3 variants do not have the same correspondence between allele B and alternate.
- **exp_freq_a1:** Expected frequency of allele B (a1) in the genotype probabilities output file.
- **info**: A statistical information metric, which is highly correlated with the squared correlation metrics output by BEAGLE[10] and MaCH[21]. (For a more in-depth comparison between these metrics, see the supplementary information in Marchini and Howie, 2010.) Values range from 0 to 1, where 1 means no uncertainty in the imputed genotypes. As noted in the IMPUTE2 online documentation, negative info scores can occur when the imputation is very uncertain, and -1 is assigned to the value when it cannot be calculated (i.e. is undefined). Note type 2 and type 3 variants will have info values of ~1. For type 0 variants, however, the info metric may be useful for filtering imputed results prior to downstream analyses, as discussed further in section VI-e.
- **certainty:** Average certainty of best-guess genotypes. This metric is also sometimes referred to as the "quality score" (QS) and is calculated as the average of the maximum probability across all samples for a given variant.
- **type**: Internal type assigned to each variant where type 0 denotes imputed variants (in 1000 Genomes but not study data); type 2 denotes imputation basis variants (observed in the study data and used to impute type 0); and type 3 denotes study variants not overlapping with imputation target variants. See Figure 2 for a schematic of these variant types.

Note: the following fields are defined only at type 2 variants, which are involved in leave-one-out masking experiments (see section VI-d).

- **concord_type0:** Concordance between observed and most likely imputed genotype
- **r2_type0:** Squared correlation between observed and imputed allelic dosage
- **info_type0**: Info quality metric for a type 2 variant treated as type 0 (i.e. when it was masked)

Figure 3 illustrates the relationship between MAF and imputation quality, with average info scores plotted for groups of variants binned by MAF (bin sizes of 0.01). Imputed SNPs (top row, panels A and B) are plotted separately from imputed indels and SVs (bottom row, panels C and D). While average info score for SNPs in the first MAF bin is 0.66, for common SNPs (MAF ≥ 0.05) average info scores increase to ~0.97. The average info scores for SVs and indels are slightly lower, leveling off at ~0.93 when MAF is ≥ 0.05.

We also plotted info scores by chromosome in Figure 4. The large interquartile ranges seen across all chromosomes are likely due to the high proportion of variants imputed to be very rare or monomorphic. This is not surprising given our liberal selection criteria for imputation target variants, which probably included many variants not seen in the predominately (though not exclusively) European study population. Imputation quality appears relatively consistent across the genome, though median values are decreased for chromosome X and the pseudo-autosomal regions PAR1 and PAR2. For

chromosome X, this is likely due to the decreased density of genotyped SNPs relative to the autosomes. The PAR regions may be negatively affected by their small size (PAR1 at 2.6 Mb and PAR2 at 0.3 Mb), which exposes a greater proportion of variants to "edge effects:" decreased imputation quality near the end of segments.

Downstream analyses of imputed results should take into account the uncertainty of imputed genotypes; however, there is no strong consensus on the best way to do this[17]. The GAC recommends a variant level filter, in which only variants with a quality metric (IMPUTE2 info or BEAGLE allelic r^2, e.g.) above a certain cutoff value are taken forward into downstream analyses. For example, there is precedent for including only variants with a quality metric of ≥ 0.3[17]. Other threshold values > 0.3 are also reasonable based on the user's desired balance between stringency and inclusivity. In this imputation, choosing a threshold of > 0.3 would retain 67.7% of all imputed variants for downstream analyses, while more stringent thresholds of 0.5 and 0.8 would retain 59.0% and 41.8% of imputed variants, respectively. (See above paragraph for a discussion of info score distribution among imputed variants.) However, users should be aware that setting stringent quality thresholds has been shown to result in missing true positive associations[22].

Another filtering approach is at the level of imputed genotypes. There is precedence for only analyzing genotypes imputed at a probability ≥ 0.9 and zeroing out all remaining genotypes[23]. However, genotype-level filtering does not make use of the full information at a given marker and therefore may be less desirable than the variant level filters described above.

e. Masked variant analysis

A common way to assess imputation quality, beyond the theoretical calculations of accuracy discussed above, is to intentionally "mask" a subset of the variants genotyped in the study sample (i.e. remove from the imputation basis), impute the masked variants as if they were unobserved, and then compare these imputed results to the observed genotypes. The comparison can be made to either (1) the most likely imputed genotype, yielding a somewhat coarse concordance measure and/or (2) the estimated allelic dosage, yielding a more granular correlation measure.

Consider imputed results represented as the probability of the AA, AB, and BB genotype. For the i^{th} sample and the j^{th} variant, the expected A allelic dosage is $E(d_{ij})$= 2*P(AA) + 1*P(AB) + 0*P(BB). The squared correlation between the expected allelic dosage $E(d_{ij})$ and the observed allelic dosage $O(d_{ij})$ over individuals can be calculated at each masked variant, assuming the observed genotype is the true genotype. This correlation metric is an empirical version of the imputation r^2 metrics of MaCH and BEAGLE, which are highly correlated with the IMPUTE2 info score.

This type of masked variant analysis is integrated into every IMPUTE2 imputation run: each study variant (type 2) is removed from imputation in a leave-one-out fashion, imputed (treated as type 0); and then compared to the imputation input. In the metrics files output by IMPUTE2, each type 2 variant includes results from the masked variant test, including concordance and correlation between imputed and observed results, as well as the info metric from treating the variant as type 0. Below we assess the quality metrics of the 686,143 masked SNPs in this imputation. These plots and summaries exclude variants where the empirical dosage r^2 ("r2_type0") is undefined, which occurs when the variant is either observed or imputed to be monomorphic. IMPUTE2 sets r2_type0 to "-1" in these instances;

these undefined r2_type0 observations are excluded from the plots and summaries so as not to artificially decrease the summary metrics (e.g., mean r2_type0).

Figure 5 summarizes the concordance and correlation metrics, with masked SNPs binned according to MAF in the observed study genotypes (0.01 intervals). The first panel (A) shows the number of SNPs per MAF bin and, and panel (B) shows the percentage of SNPs in the bin with "info_type0" ≥ 0.8. In panels C and D, each data point indicates the average value of all SNPs in that MAF bin for the metric indicated on the y-axis. The black data series include all masked SNPs while the gray data series excludes SNPs with "info_type0" < 0.8. The metric shown in panel (C) is the correlation between masked and imputed allelic dosages; the metric in panel (D) is the concordance: the fraction of identical genotypes between the most likely imputed and observed genotype.

Several salient points emerge from these graphs. First, there is a decline in empirical dosage r^2 for less common and rare variants. As MAF increases, however, average correlation values increase to ~ 0.95. Second, the differences between unfiltered (black points) and filtered (gray points) data series demonstrate the utility of filtering by the info quality metric, which is available for all imputed variants. This filtering improves the quality metrics profile for masked SNPs across the entire range of MAF bins. Thirdly, Figure 5D illustrates how overall concordance is heavily influenced by MAF, as for SNPs with MAF < 5% simply assigning imputed genotypes to the major homozygous state would yield > 90% concordance[24]. Thus, there is a bias of high concordance values at low MAF SNPs, where major homozygotes are likely to be imputed "correctly" just by chance. To alleviate this bias, in Table 3 we report average concordance and correlation values in two groups of masked SNPs: MAF < 0.05 and MAF ≥ 0.05.

We also use masked SNP results to assess strand consistency between study and reference panel data. As discussed in section IV, IMPUTE2 can detect and correct strand mismatches at strand-unambiguous SNPs (i.e., those with alleles other than A/T or C/G). Only four such flips were reported genome-wide, which suggests that our initial strand annotation was mostly correct. However, strand misalignments at strand ambiguous SNPs may go undetected. With masked results, we can compare concordance to quality (concord_type0 vs. info_type0) as masked variants with high quality but low concordance may indicate a strand flip. Figure 6 illustrates this check. Each point is a masked SNP: strand-ambiguous points are plotted as black asterisks and all other points as semi-transparent green circles. Points in the bottom right quadrant are potential strand issues, where concordance is low (concord_type0 < 0.5) but quality is relatively high (info_type0 > 0.5). As seen with the density curves in the outer margin of the plots, very few SNPs are in this potential problem quadrant (56 total) — only one of which is strand ambiguous. The presence of strand unambiguous SNPs suggests that other factors instead of strand misalignment may result in variants ending up with this type of masked SNP metric profile.

Users should note the following aspects of this and other masked SNP tests. While converting imputed probabilities to most likely genotypes is not recommended for association testing, it provides an easily interpretable quality metric for masked SNP tests. Furthermore, concordance can also be reported by averaging over all masked genotypes, rather than by calculating a concordance rate at each masked SNP and then taking the average of those per-SNP values as we have done here. The former way of calculating this metric often leads to higher mean concordance, especially when imputed genotypes are filtered on maximum probability. There are also advantages and disadvantages to both

metrics: dosage r^2 and concordance. The advantages of dosage r^2 include (1) precedence in the literature for evaluating imputation accuracy[8,9,25]; (2) less sensitivity to allele frequency than concordance; (3) similarity to information metrics commonly reported by imputation software (for a review, see Marchini and Howie, 2010); and (4) incorporation of imputation uncertainty by using expected allelic dosage rather than most likely genotype. However, one important downside is that r^2 has high variance at low MAF[26]. Concordance is advantageous as a widely used metric that is easily interpretable; however, it ignores imputation uncertainty and is very sensitive to allele frequency, as low MAF variants may yield high concordance purely by chance[24].

Lastly, when discussing imputation quality there can be several different meanings of "efficiency." Figure 5B illustrates one definition: the fraction of imputed variants passing a given quality filter (info ≥ 0.8, e.g.). This metric is quite high in most MAF bins > 0.1. An alternate meaning of imputation "efficiency" is the fraction of samples imputed above a given maximum probability threshold (probability ≥ 0.9, e.g.), calculated at each SNP. This metric is relevant if one were filtering imputed data at the genotype level rather than on a per-SNP level, as it equates to the percentage of samples whose data will be used at each SNP. However, given that genotype-level filtering is not recommended, the per-SNP efficiency metric, as described above, was not included here. Users can easily produce this metric by taking the imputed genotype data files; converting into most likely genotypes, using a probability threshold; and then calculating the percent missingness at each SNP.

f. Downstream analysis

Many references are available for users desiring further information on imputation methods, including recommendations and caveats for downstream analyses[1,2,13,17,26]. Prior to such analyses, users may need to filter imputed results and/or reformat the imputation output. IMPUTE2 is part of a suite of GWAS software that is useful in these post-imputation tasks. For example, QCTOOL may be used to filter imputed data by the IMPUTE2 info score. The data formatting program "fcGene" is another file conversion tool that is compatible with IMPUTE2 output (see Web Resources). Programs for performing association analyses with imputed genotype probabilities include GWASTools[27], an R package developed by the GAC; PLINK[7] (with the --dosage option: https://www.cog-genomics.org/plink2/assoc#dosage); MACH2qtl/dat[21]; SNPTEST[28]; ProbABEL[29]; BIMBAM[30]; SNPMStat[31]; and the R package snpMatrix[32]. For a comparison of methods to account for genotype uncertainty in imputed data, see Zheng et al.[33].

VII. Summary

We have performed genotype imputation in WLS, using a worldwide 1000 Genomes Project phase 3 reference panel and IMPUTE2 software. The imputed genotypes and accompanying marker annotation and quality metrics files are available through the authorized access portion of the dbGaP posting. These imputation analyses were performed and documented by Sarah Nelson under the leadership of Cathy Laurie and Bruce Weir, within the GAC at the University of Washington (UW) in Seattle, WA. This report was reviewed and approved by study investigators and by representatives of the CIDR genotyping center.

VIII. References

1. Browning, S. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* **124**, 439-50 (2008).
2. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
3. Howie, B., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
4. B. Howie, J.M., and M. Stephens. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
5. Laurie, C.C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* **34**, 591-602 (2010).
6. Altshuler, D. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
7. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
8. 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**, 955-9 (2012).
9. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
10. Browning, B. & Browning, S. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
11. Frazer, K. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
12. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
13. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
14. McVean, G. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
15. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
16. Delaneau, O., Marchini, J. & The Genomes Project, C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**(2014).
17. de Bakker, P. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
18. Nelson, S.C., Laurie, C.C., Doheny, K.F. & Mirel, D.B. Is 'forward' the same as 'plus'?…and other adventures in SNP allele nomenclature. *Trends in Genetics* **28**, 361-363 (2012).
19. Nelson, S.C. *et al.* Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Human molecular genetics* (2016).
20. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-83 (2016).
21. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).

22. Beecham, G.W., Martin, E.R., Gilbert, J.R., Haines, J.L. & Pericak-Vance, M.A. APOE is not associated with Alzheimer disease: a cautionary tale of genotype imputation. *Ann Hum Genet* **74**, 189-94 (2010).
23. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet* **125**, 163-71 (2009).
24. Lin, P. *et al.* A new statistic to evaluate imputation reliability. *PLoS One* **5**, e9697 (2010).
25. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
26. Evangelou, E. & Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379-89 (2013).
27. Gogarten, S.M. *et al.* GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* **28**, 3329-31 (2012).
28. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
29. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).
30. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).
31. Hu, Y.J., Lin, D.Y. & Zeng, D. A general framework for studying genetic effects and gene-environment interactions with missing data. *Biostatistics* **11**, 583-98 (2010).
32. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum Hered* **64**, 45-51 (2007).
33. Zheng, J., Li, Y., Abecasis, G.R. & Scheet, P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* **35**, 102-10 (2011).

IX. Web resources: data and software

The 1000 Genomes Project. "About the 1000 Genomes Project." Retrieved from http://www.1000genomes.org/about on March 7, 2011.

The 1000 Genomes Project. IMPUTE2 Haplotypes for Phase 3. Retrieved from https://mathgen.stats.ox.ac.uk/impute/1000GP%20Phase%203%20haplotypes%206%20October%202014.html on Oct. 31st, 2014.

The 1000 Genomes Project. Phase 3 final release. Available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ Delaneau O (Version 2.r790, c2011-2012) SHAPEIT: Segmented HAPlotype Estimation and Imputation Tool [software]. Available from https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.

Delaneau O (Version 2.r790, c2011-2012) SHAPEIT: Segmented HAPlotype Estimation and Imputation Tool [software]. Available from https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.

Genome-wide Association Study Software Suite : CHIAMO, GTOOL, IMPUTE, SNPTEST, HAPGEN, GENECLUSTER, BIA, HAPQUEST (c2007). Available from http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html.

Howie B and Marchini J (c2007-2012) IMPUTE version 2.3.2 [software]. Available from https://mathgen.stats.ox.ac.uk/impute/impute_v2.html.

Howie B and Marchini J (September 23, 2010). "Using IMPUTE2 for phasing of GWAS and subsequent imputation," a document distributed with IMPUTE2 example code. Available at http://mathgen.stats.ox.ac.uk/impute/prephasing_and_imputation_with_impute2.tgz.

Illumina, Inc. (2006). "TOP/BOT" Strand and "A/B" Allele [Technical Note]. Available from http://www.illumina.com/documents/products/technotes/technote_topbot.pdf

IMPUTE 2 background. Retrieved from https://mathgen.stats.ox.ac.uk/impute/impute_background.html, February 21, 2012.

IMPUTE2 file format descriptions. Retrieved from http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html , February 7, 2012.

Freeman C and Marchini J. (c2007-2011) GTOOL Software Package (Version 0.7.5) [software]. Available from http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html.

Purcell S. PLINK (Version 1.07, c2009) [software]. Available from http://pngu.mgh.harvard.edu/purcell/plink/.

Roshyra, NR. fcGENE [software]. Available from http://sourceforge.net/projects/fcgene/.

X. Tables

Table 1. Variant summary

Chromosome	Study SNPs†	Imputation basis††	Imputation Output*
1	57,054	56,147	2,555,978
2	55,547	55,120	2,764,508
3	45,400	44,978	2,303,766
4	38,835	38,558	2,329,049
5	40,564	39,436	2,093,036
6	46,310	45,551	2,063,587
7	36,520	35,970	1,895,362
8	35,649	35,366	1,815,748
9	31,619	31,334	1,413,498
10	37,612	37,281	1,609,087
11	35,240	34,813	1,599,948
12	34,220	33,762	1,543,617
13	26,713	26,554	1,156,309
14	22,437	22,278	1,054,213
15	20,820	20,593	959,652
16	21,930	21,543	1,041,767
17	19,464	19,109	909,944
18	20,933	20,741	909,580
19	14,284	13,940	739,828
20	17,802	17,545	714,371
21	9,837	9,766	449,672
22	10,048	9,831	444,797
X	16,171	15,541	1,129,961
PAR	424	386	52,071
Totals	***695,433***	***686,143***	***33,549,349***

† Study variants passing pre-imputation filters (IMPUTE2 variant types 2 and 3).
†† Study variants passing pre-imputation filters and overlapping with the imputation target (type 2).
*Imputation output is the sum of imputation target (type 0), imputation basis (type 2), and study only (type 3) variants. Reference panel variants have been filtered to only variants with MAC ≥ 4 in any one of the five 1000 Genomes super populations.

Table 2. An overview of the 2,504 samples in the 1000 Genomes Project phase 3 worldwide reference panel. The Project assigned each population to one of five continental groupings or "super populations": African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). This table is based on reference panel data downloaded from IMPUTE2 and the sample summary provided by the Project (see Web resources).

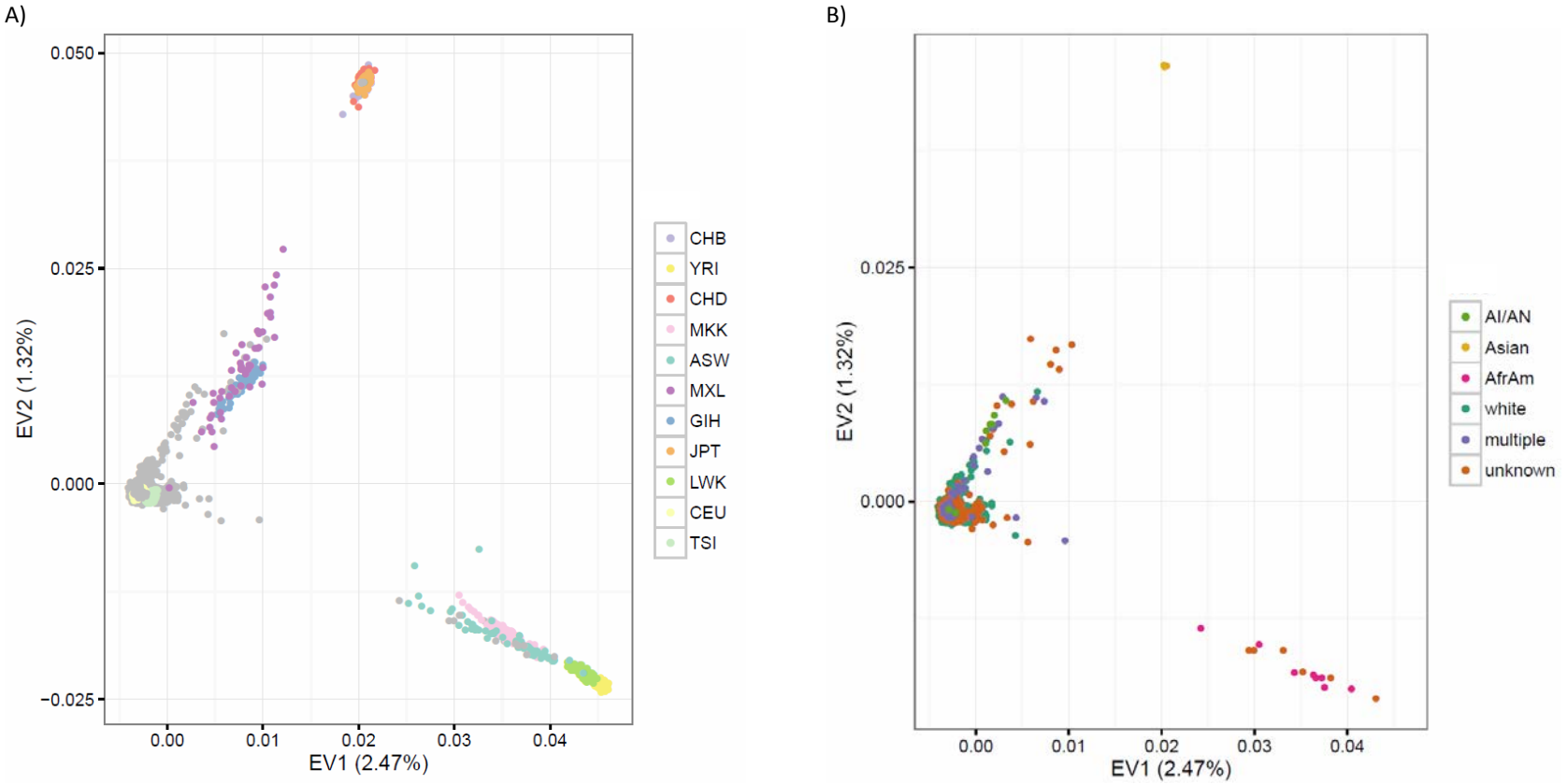
Full Population Name	**Abbreviation**	**Number of Samples**
African Caribbean in Barbados	ACB	96
African Ancestry in Southwest US	ASW	61
Esan in Nigeria	ESN	99
Gambian in Western Division, The Gambia	GWD	113
Luhya in Webuye, Kenya	LWK	99
Mende in Sierra Leone	MSL	85
Yoruba in Ibadan, Nigeria	YRI	108
Total African ancestry	*AFR*	*661*
Colombian in Medellin, Colombia	CLM	94
Mexican ancestry in Los Angeles, California	MXL	64
Peruvian in Lima, Peru	PEL	85
Puerto Rican in Puerto Rico	PUR	104
Total Americas ancestry	*AMR*	*347*
Chinese Dai in Xishuangbanna, China	CDX	93
Han Chinese in Beijing, China	CHB	103
Southern Han Chinese, China	CHS	105
Japanese in Tokyo, Japan	JPT	104
Kinh in Ho Chi Minh City, Vietnam	KHV	99
Total East Asian ancestry	*EAS*	*504*
Bengali in Bangladesh	BEB	86
Gujarati Indian in Houston, TX	GIH	103
Indian Telugu in the UK	ITU	102
Punjabi in Lahore, Pakistan	PJL	96
Sri Lankan Tamil in the UK	STU	102
Total South Asian ancestry	*SAS*	*489*
Utah residents with Northern and Western European ancestry	CEU	99
Finnish in Finland	FIN	99
British in England and Scotland	GBR	91
Iberian populations in Spain	IBS	107
Toscani in Italia	TSI	107
Total European ancestry	*EUR*	*503*

Table 3. Quality metrics for all masked SNPs dichotomized into groups of MAF < 0.05 vs. MAF ≥ 0.05. The third column shows the number of variants in each MAF group. Mean and median values are presented for overall genotype concordance and empirical dosage r^2 (in IMPUTE2 metrics files, labeled as "concord_type0" and "r2_type0," respectively). No info threshold has been applied here, such that all masked and imputed variants are included in these averages, except where r2_type0 was undefined (see report narrative).

MAF (in study samples)	**Number of variants**	**Mean (Median) Overall Concordance**	**Mean (Median) empirical dosage r^2**
< 0.05	110,905	0.998 (1.00)	0.833 (0.949)
≥ 0.05	575,238	0.982 (0.996)	0.957 (0.989)

XI. Figures

Figure 1. Principal component analysis of HapMap reference samples (panel A, with study samples in gray) and 9,018 unique study samples (panel B), using 84,740 autosomal SNPs selected based on linkage disequilibrium, MCR, and MAF. Color coding is either by reference population* (for HapMap samples) or by self-identified race (for study samples, where AI/AN=American Indian/Alaska Native; AfrAm=Black or African American; multiple=More Than One Race; unknown=Unknown or Not Reported). The percent variance explained by each of these first two components is noted on the axis labels. Also Figure 9 from the genotype QC report.



*CEU=Utah residents with Northern and Western European ancestry from the CEPH collection; CHB=Han Chinese in Beijing, China; GIH=Gujarati Indians in Houston, Texas; JPT=Japanese in Tokyo, Japan; LWK=Luhya in Webuye, Kenya; MXL=Mexican ancestry in Los Angeles, California; MKK=Maasai in Kinyawa, Kenya; TSI=Toscani in Italia; YRI=Yoruba in Ibadan, Nigeria

Figure 2. A schematic of variant types as defined in the IMPUTE2 imputation algorithm. Each individual is represented by a unique color in the horizontal bar(s), and alternate alleles at each variant are represented as *A* and *B*. Section (A) represents phased reference haplotypes, where two samples (4 phased chromosomes) are shown. Section (B) represents three study samples with genotype calls, as would be observed in GWAS array experiment. Section (C) identifies the variant type of each position shown. Type 2 variants have data in both the reference and the study samples: positions 1, 4, 6, 8, and 11. Type 0 variants have data in the reference but not in the study samples: positions 3, 5, 9-10, and 12. Thus, data at type 2 variants (imputation basis) are used to impute type 0 variants (imputation target) in the study samples. Type 3 variants are those in study samples but not in the reference; ultimately, these are extraneous to the imputation, which is why they are shown in white text. This figure is a based off of IMPUTE2 background documentation (see Web Resources).

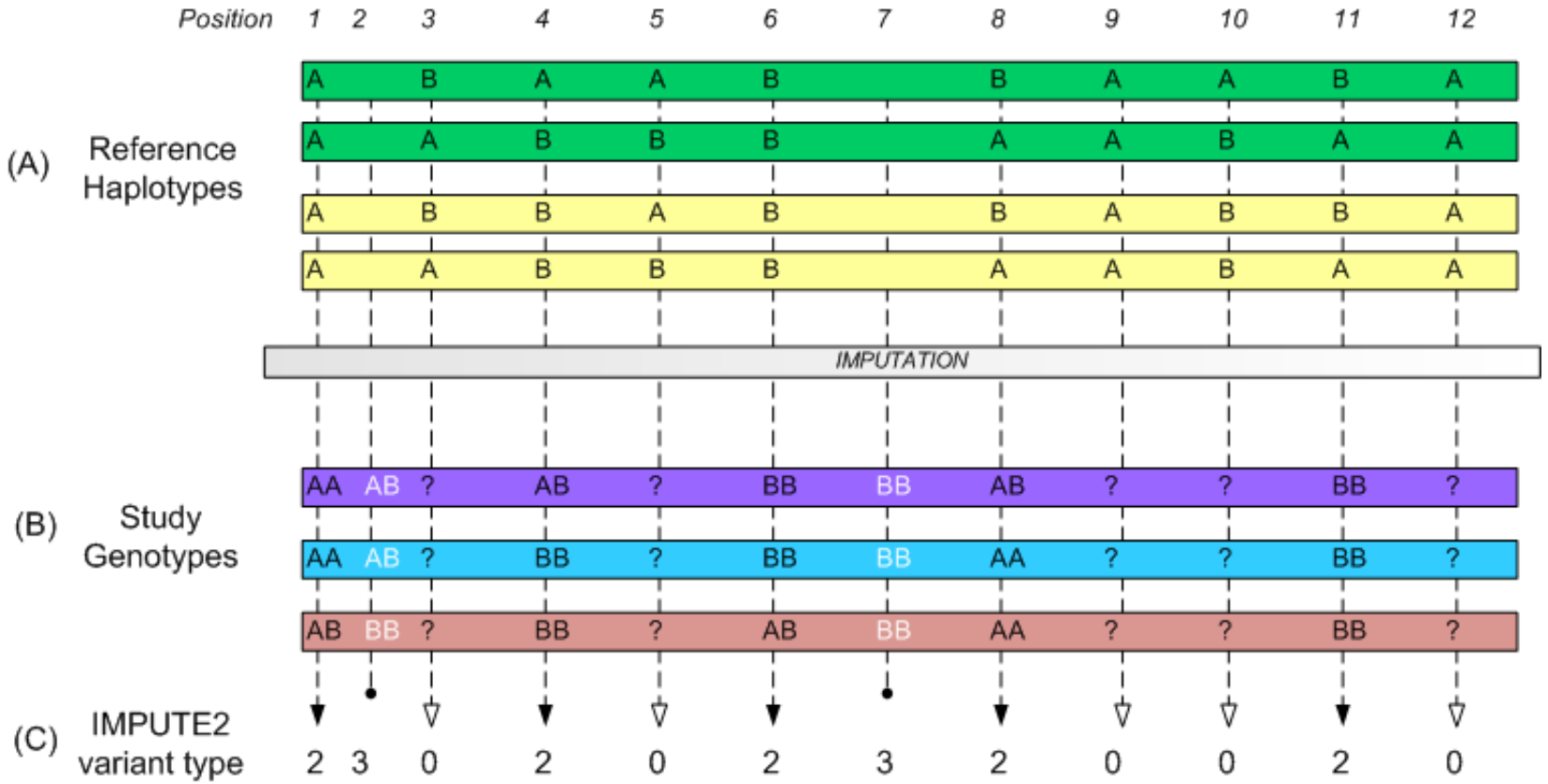
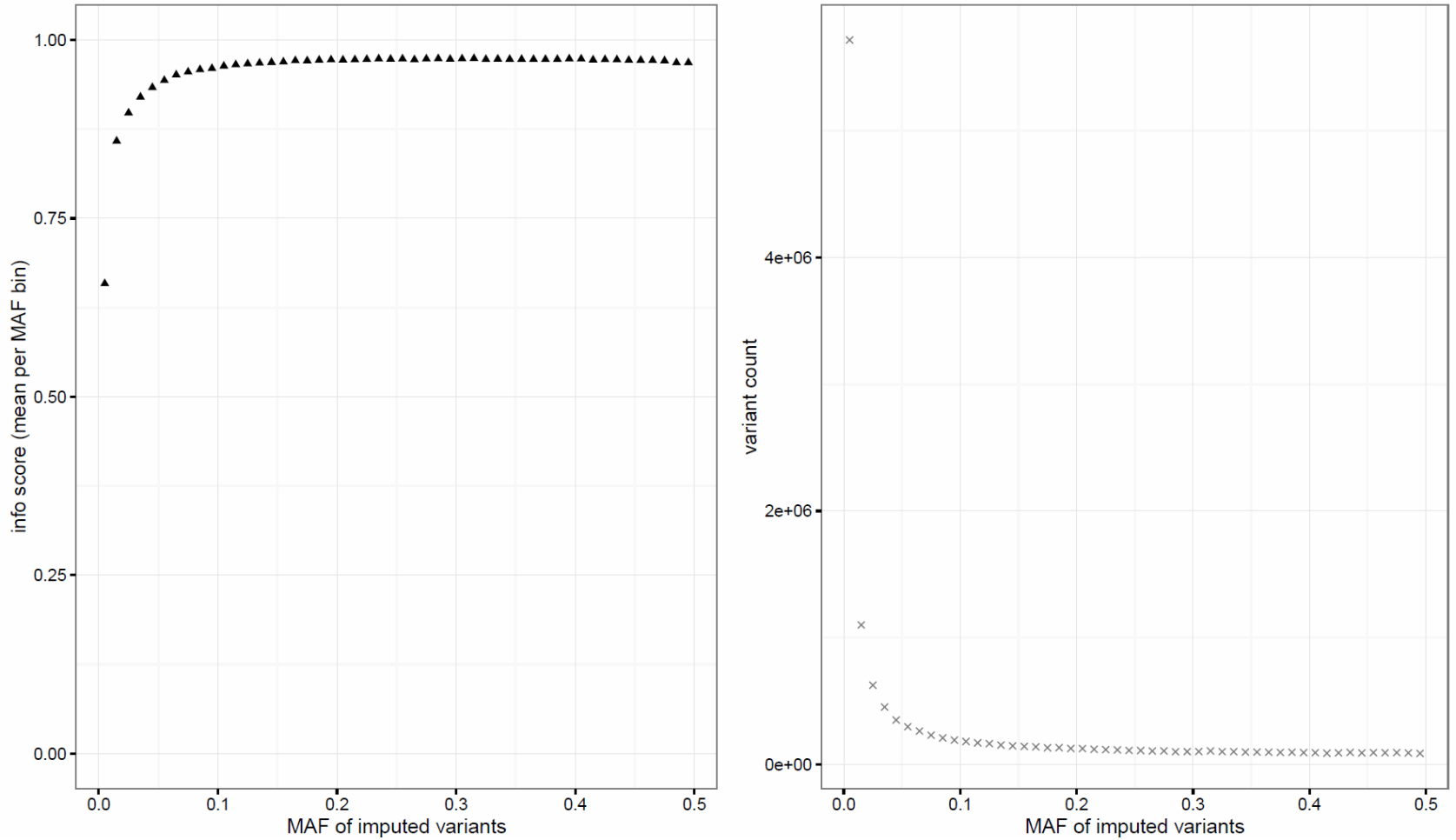


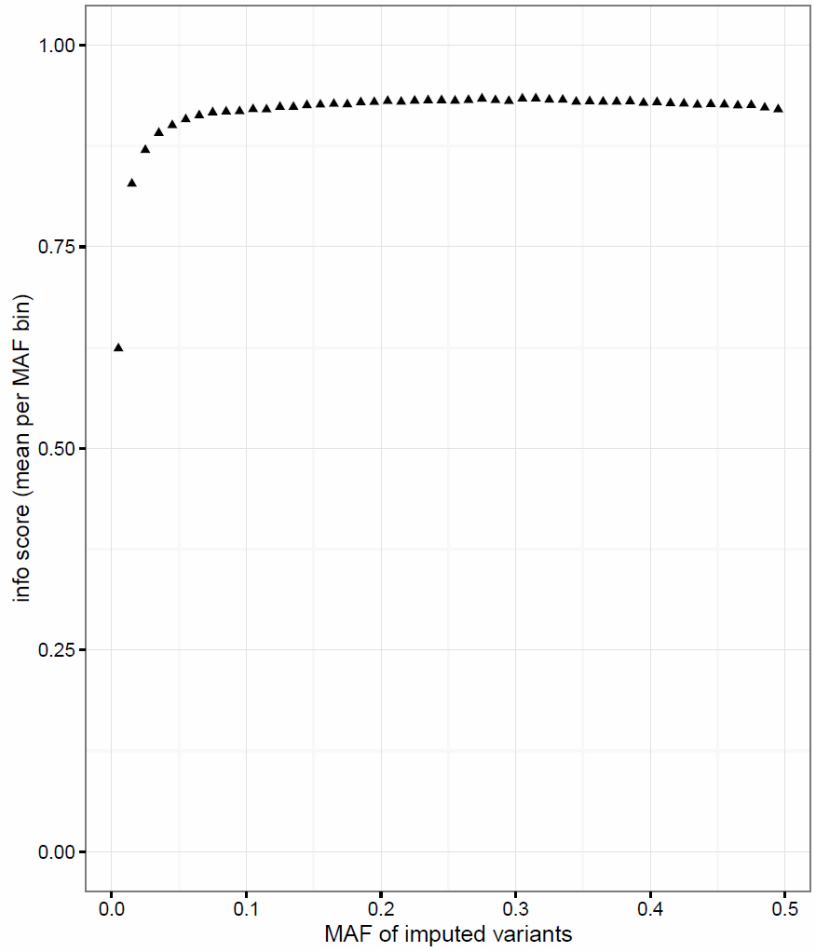
Figure 3. Summaries of quality metrics at all imputed variants. In each plot, imputed variants are binned by MAF (0.01 intervals). The top row is for SNPs (panels A and B) and the bottom row is for SVs and indels (panels C and D). Plots in the first column (A and C) have the mean info metric per bin plotted on the y-axis. Plots in the second column (B and D) indicate the count of variants in each MAF bin.

A) SNP metrics

B) SNP counts



C) SV/indel metrics



D) SV/indel counts

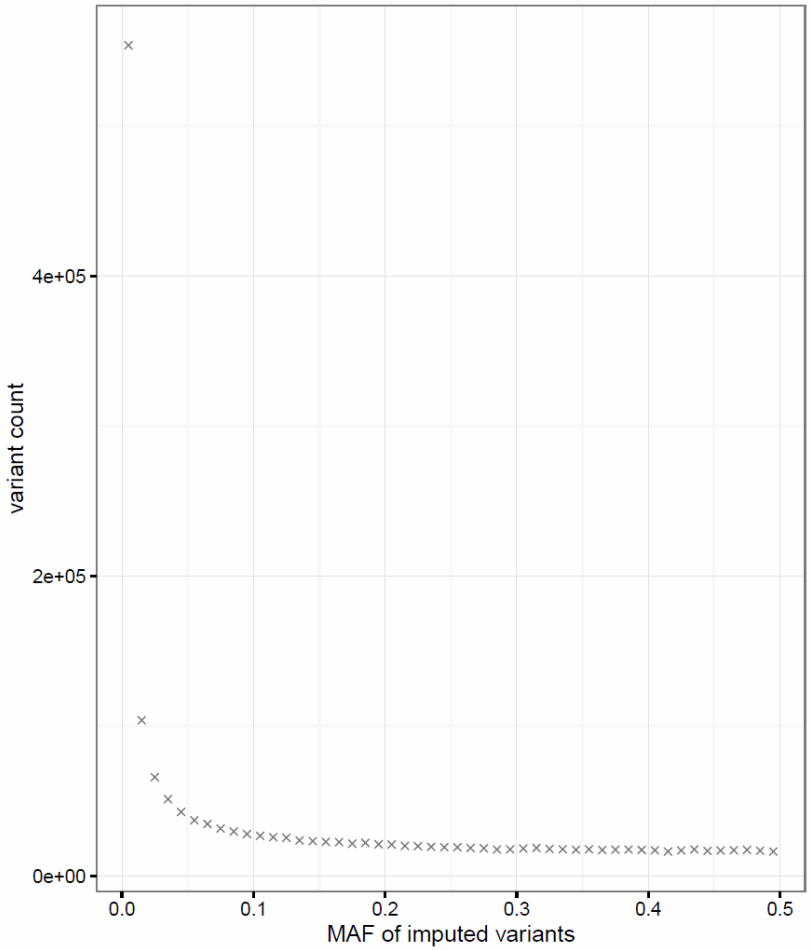
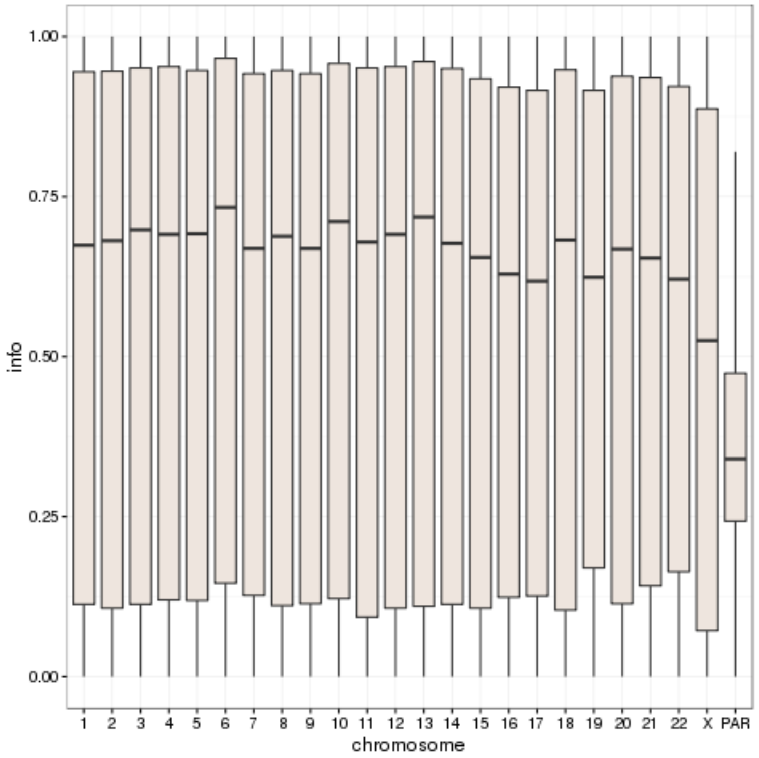


Figure 4. A comparison of the info score imputation quality metric, plotted by chromosome for all imputed SNPs (panel A) and indels and SVs (panel B). Outlier values are not displayed in these box plots. Pseudo-autosomal variants (in PAR1 and PAR2) have been combined and are plotted under the "PAR" label.

A) SNPs



B) SV/indels

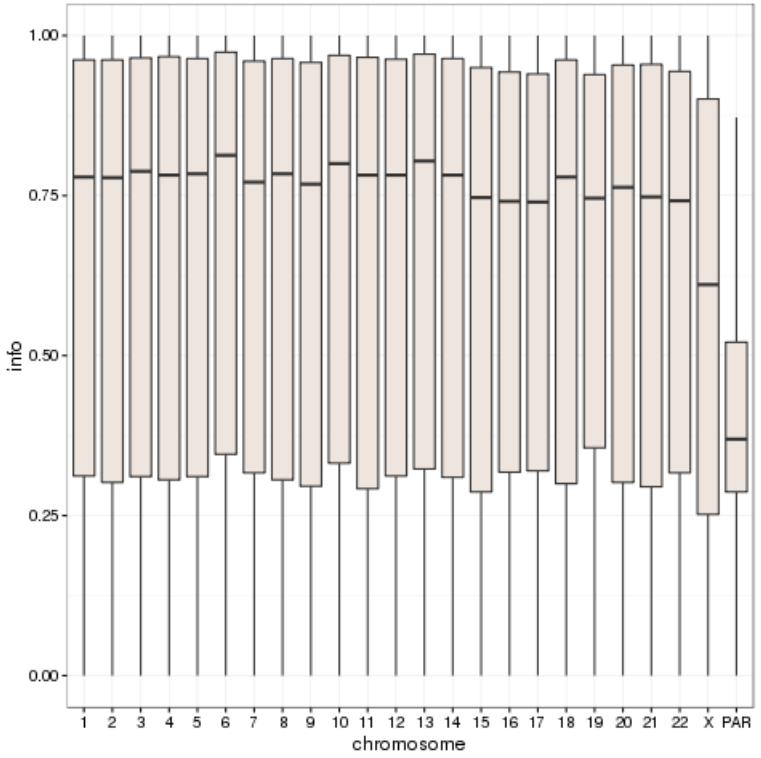


Figure 5. Quality metrics for all masked SNPs, grouped into MAF bins at 0.01 intervals. Panel (A) shows the number of SNPs per MAF bin, and panel (B) shows the fraction of SNPs in the bin passing an info filter threshold of ≥ 0.8. Panel (C) plots the average empirical dosage r^2 metric per MAF bin, and panel (D) is the concordance between the observed and the most likely imputed genotype at masked SNPs within each MAF bin. In this bottom row of plots, the black circles give the mean metric over all masked SNPs in the MAF bin while the gray triangles are the mean for masked SNPS passing an info filter of ≥ 0.8.

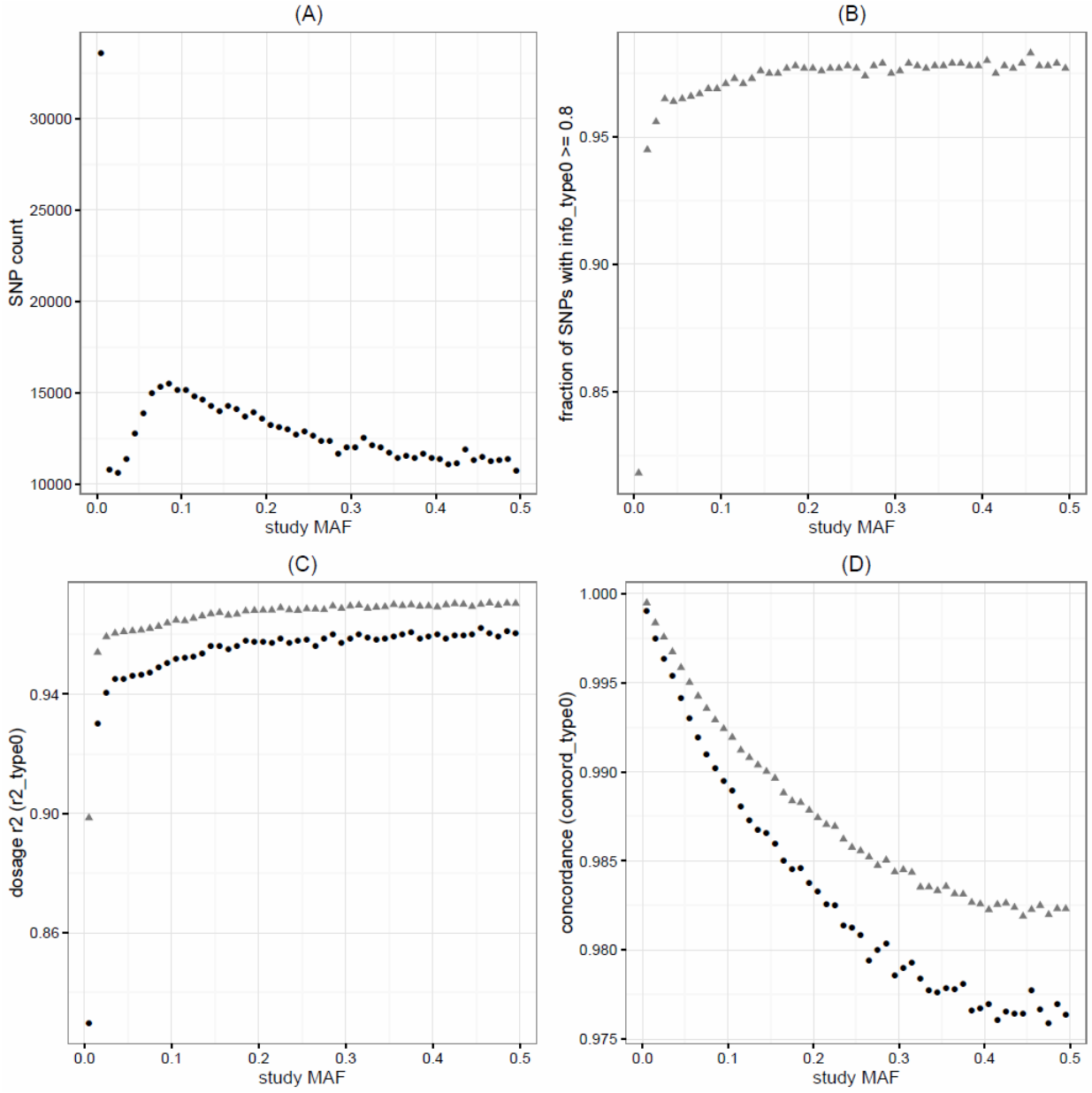
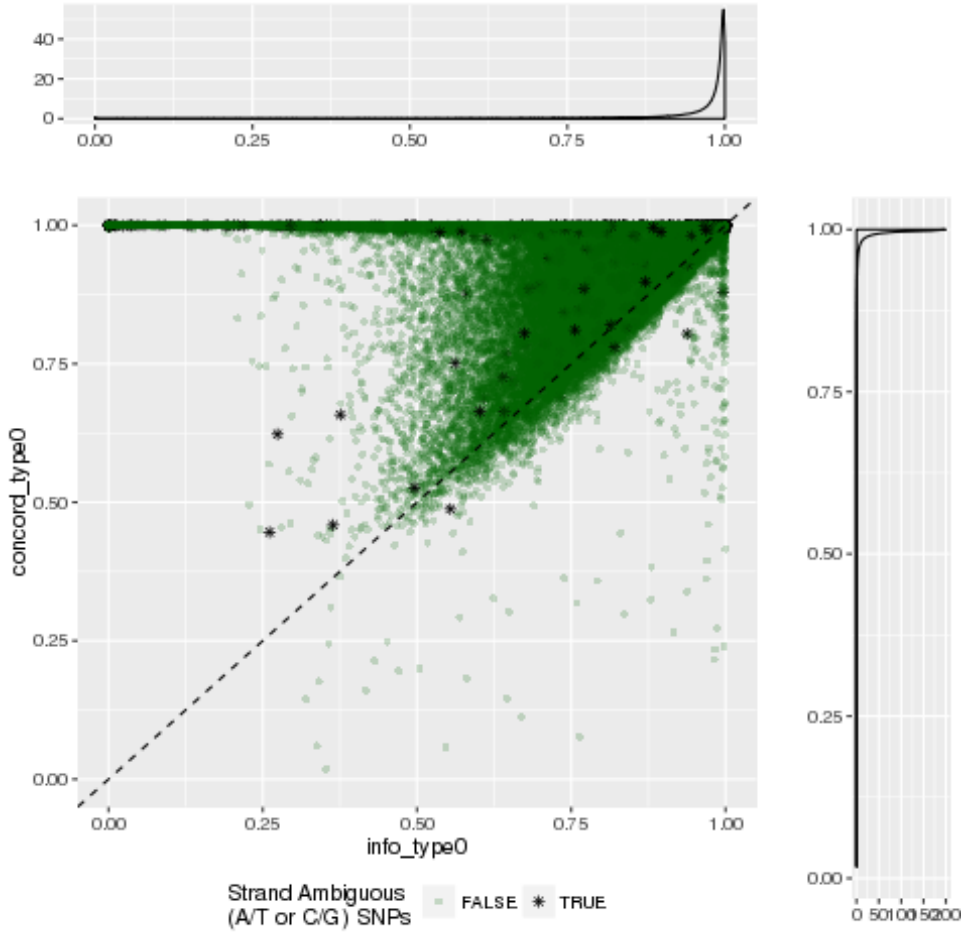


Figure 6. Strand consistency check using masked variant results. Masked concordance is plotted on the y-axis (concord_type0), and the info score from masking (info_type0) is plotted on the x-axis. Each point is a masked SNP, with strand ambiguous SNPs plotted as black asterisks. Density plots along the two axes illustrate the density of points in the scatterplot.



XII. Supplementary files

a. Chromosome anomalies. Genotypes in imputed segments of the genome harboring a gross chromosomal anomaly have been filtered out of the final genotype probabilities files. The following two supplementary files provide information related to this chromosomal anomaly filtering.

1. The file "imputation_segments.csv" is a list of the chromosome and base pair coordinates of each imputation segment (547 total). These coordinates were supplied to IMPUTE2 with the "-int" flag, to define imputation chunks. The fields in this file are:

- **chrom:** chromosome
- **segment:** imputation segment ID
- **mb.start:** start coordinate, in mega base pairs
- **mb.end:** end coordinate, in mega base pairs
- **bp.start:** start coordinate, in base pairs
- **bp.end:** end coordinate, in base pairs

2. The file "filtered_map.txt" is a list of subject-segment combinations where imputed genotypes were set to missing (i.e., 0 0 0). The fields in this file are:

- **SUBJECT_ID**: local subject-level identifier
- **chrom:** chromosome
- **segment:** imputation segment ID

b. Variant selection.

The file "snp.qualfilter.txt" is a list of genotyped variants passing GAC recommended quality filters from genotype cleaning process and also mapped to build 37. This list may be used to construct a keeplist for use with the PLINK `--extract` flag, to perform the initial sub setting of variants from the binary file (see II-d). The variant dimension in this file corresponds to the "Study variants" column of the Variant Summary in Table 1. The columns in these text files are:

- **rs.id**: variant name
- **chrom:** chromosome number, in build 37 mapping

c. Variant flip list.

The file "fliplist.txt" is a list of imputation basis variants requiring a strand flip to align with the plus strand of the human genome reference, based on Illumina annotation. For these variants, the Illumina TOP alleles in the initial input binary PLINK file were annotated as being on the "-" strand. For more information, see sections II-d and IV.

d. Sample-subject mapping. The identifier used in the imputation output is the local "SUBJECT_ID." A mapping of "SUBJECT_ID" to the local sample level identifier "SAMPLE_ID" is the file "sampsubj_mapping.txt." The columns in this file are:

- **SUBJECT_ID**: local (study investigator's) participant level identifer, used in imputation output
- **SAMPLE_ID**: local sample-level identifier

- **whole.chrom.anomaly**: for samples excluded from a given chromosome's imputation due to whole chromosome anomalies, this field contains the filtered chromosome(s) integer codes (where 23 = chromosome X). Otherwise, NA.