Health and Retirement Study, Combined Phases 1-3

Imputation Report - 1000 Genomes Project reference panel

September 27, 2013

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 panel[1,2]. The University of Washington Genetics Coordinating Center (GCC) used IMPUTE2 software[3] to perform genotype imputation in the Health and Retirement Study (HRS). 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, rather than taking the most likely imputed genotype. Quality metrics are provided that can be used for filtering imputation results on a per-variant basis. For a detailed description of genotype quality control (QC) on this project, please see the report available through the database of Genotypes and Phenotypes (dbGaP, study accession number phs000428).

II. Study data

a. Samples

The HRS (http://hrsonline.isr.umich.edu/) is a large, longitudinal study of Americans over age 50 aimed at monitoring health, social, economic, and numerous other factors related to aging and retirement. A random subset of the ~26,000 total participants was selected to participate in enhanced face to face interviews and biological specimen collection between 2006 and 2010. This imputation includes respondents who consented to the saliva collection in 2006 (Phase 1), 2008 (Phase 2), or 2010 (Phase 3). The Phase 1 and 2 participants were genotyped and imputed together previously (see dbGaP accession numbers phg000207 and phg000264). The Phase 3 participants were genotyped in early 2013, on the same array as Phases 1-2. Each of the genotyping datasets (Phases 1-2, 3) underwent the GCC standardized QC procedures[4] for genotype data cleaning. The three phases were ultimately combined, yielding a total of 15,620 unique HRS participants: 12,505 from the initial Phase 1 and 2 genotyping and 3,115 from Phase 3.

All HRS participants were imputed together, including the previously genotyped and imputed Phases 1-2 participants. Samples were selected for imputation from the combined dataset using the set of recommended quality filters generated during data cleaning. In brief, samples of questionable identity or poor quality were excluded from imputation, including 53 Phases 1-2 samples with missing call rate > 2%. We also excluded sample-chromosome combinations where a large chromosomal anomaly was recommended for filtering. This resulted in the exclusion of nine samples from chromosome X imputation.

Figure 1 shows a principal component analysis (PCA) of all HRS study participants, who have a wide range of ancestries and self-identified ethnicities. The IMPUTE2 algorithm discussed below recommends the use of a worldwide 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 ("SUBJID" in annotation files) was used as the individual identifier throughout, which can be mapped to the local and GCC-assigned scan level identifiers ("SAMPID" and "scanID", respectively) using the sample-subject mapping file provided in the Supplementary Files of this report (section XII).

b. SNPs

Both HRS genotype datasets were genotyped on the Illumina HumanOmni2.5 array, but different versions: Phases 1-2 on HumanOmni2.5-4v1 and Phase 3 on HumanOmni2.5-8v1. Both versions of the array are designed to human genome build 37. SNPs retained in the combined dataset were those common to both versions of the array and with matching rsID, chromosome, and position in both array annotations. SNPs discordant between control samples run on both array versions were excluded. For the purposes of imputation, study SNPs were selected using the combined quality filter described in the Phase 3 genotyping QC report.

A summary of initial input SNPs is shown in Table 1; a list of these SNPs is available in the Supplementary Files. Observed genotypes (which have a probability of 1) are included in the imputation output. Where an observed study SNP had sporadic missing data, the missing genotypes were imputed by the pre-phasing software. Additionally, SNPs 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 dbGaP genotype releases and these imputed data. The variant annotation files accompanying this report can be used to differentiate between observed study SNPs 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. See Figure 2 for a visual representation of these variant types.

c. Data formatting

The study genotype data were initially accessed from the combined, binary PLINK[4] file available in the dbGaP release, "HRS_phase123_TOP," with genotypes expressed in TOP alleles. The Illumina annotation files, which included genomic strand information, were used to identify the variants requiring a strand flip to convert the TOP allele to the "+" strand of the human genome reference assembly (discussed further in section IV).

When extracting data from this PLINK dataset, we (1) subset out by chromosome; (2) set haploid genotypes (male chromosome X) called as heterozygotes to missing; (3) extracted only study variants passing the quality filter; (4) updated parental identifiers to reflect parent-offspring relationships identified during data cleaning; and (5) specified a list of variants that required a strand flip to align with the "+" strand, based on Illumina annotation. Furthermore, when formatting the chromosome X PLINK file, the nine samples with a large chromosomal anomaly recommended for filtering were removed with the PLINK "`--remove`" flag.

Below is an example of the command line syntax used to create the filtered binary files, for generic chromosome "#":

```
plink --bfile HRS_phase123_TOP \
--extract snp.qualfilter.txt --flip fliplist.txt \
--keep sampkeep.txt --set-hh-missing --chr # --make-bed \
--update-parents update_parents.txt -–out HRS2_chr#
```

d. Pre-phasing

Historically, phasing and imputation have been done jointly in a unified process. More recently, the alternative approach of "pre-phasing" has been suggested as a way to maintain imputation accuracy while minimizing computation time, as available reference panels increase in number and in size[5]. 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 imputation.

For some studies, an additional advantage to pre-phasing is that it can incorporate family structure, while most imputation algorithms "ignore" relatedness, due to 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. Although there were no expected relatives among HRS study participants, genotype data cleaning revealed several first and second-degree relationships. While pre-phasing does not utilize full siblings and half-sib-like relationships, parent-offspring pairs are used. Thus we updated parent-offspring relationships in the initial PLINK dataset, by using the "`--update-parents`" flag (see previous section). The family structure information is included in the Supplementary Files section of this report. For each participant, we annotate whether they were phased as part of a duo, trio, or as unrelated.

We phased the study data with the SHAPEIT2[6] haplotype estimation tool, inputting the filtered, chromosome-specific PLINK files (see II-c) and receiving the best guess haplotypes as output. 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 18 hours to 6 days, depending on the size of the chromosome. Below is an example of the command line syntax used to run the SHAPEIT2 program on a generic chromosome "#":

```
shapeit2 -B HRS2_chr# \
-M genetic_map_chr#_combined_b37.txt \
-O HRS2_chr#.haps.gz HRS2_chr#.sample.gz \
-S 200 –T 12 -L shapeit_chr#.log
```

III. Reference panel

Larger reference panels have been shown to increase imputation accuracy[2,7,8]. Previously, haplotypes from Phases 2[9] and 3[10] 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, enabling the imputation of many more and rarer variants[2,11].

The 1000 Genomes Project aims to "discover, genotype, and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations[12]." In October 2011, the Project released the first version of the phase I integrated variant set, containing single nucleotide polymorphisms (SNPs), insertion/deletions (indels), and structural variants (SVs) in 1,092 samples from 14 different populations[13]. The Project has categorized each of these populations into four continental groupings: African (AFR), American (AMR), Asian (ASN), and European (EUR). Sample counts from each of the 14 populations and four continental panels are show in Table 2. To impute these HRS participants, we used a worldwide reference panel of all 1,092 samples from the phase I integrated variant set (v3, released March 2012), which is based on both low coverage whole genome and deep coverage exome sequence data. We downloaded these reference panel data, for 22 autosomes and the non-pseudo autosomal portions of chromosome X, from the IMPUTE2 website (see Web Resources), which had been created from the variant call format (VCF) files available from the Project.

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[8]. 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 39 million variants). Very low frequency 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 at these loci in downstream analyses. Therefore, we restricted imputation to variants with at least four copies of the minor allele in any of the four 1000 Genomes continental groups (AFR, AMR, ASN, or EUR). We included all three variant types (SNPs, indels, and SVs) in this imputation, based on findings[13] from the 1000 Genomes Project that imputation accuracy at indels and SVs can be comparable to that of SNPs.

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 (“reference”). In practice, however, this crucial step is not always straightforward[14]. The initial study dataset contained TOP alleles, an Illumina naming method unrelated to “+” or “-“ strand orientation[15] (also see Web Resources, Illumina 2006). Because all 1000 Genomes reference panel data are expected to be “+” strand relative to the reference, we initially used Illumina annotation to identify and flip all the SNPs where the TOP allele was not on the “+” strand.

We initially observed a discrepancy in strand annotation for 463 SNPs when comparing the two versions of the array annotation (HumanOmni2.5-4v1 and HumanOmni2.5-8v1). In an attempt to resolve these discrepancies, we performed a BLAT[16] search of the TOP genomic sequence (“TopGenomicSequence” in Illumina annotation) for each of the 463 SNPs. According to the BLAT results, the HumanOmni2.5-8v1 strand annotation appeared to be correct for the majority (~80%) of SNPs. Thus we used the HumanOmni2.5-8v1 array version to construct the fliplist. We also excluded the subset of 463 SNPs with discrepant strand information that were also strand ambiguous (A/T or C/G alleles), as subsequent imputation steps would not be able to identify a strand misalignment at these SNPs. Only 32 SNPs were excluded on this basis (i.e. both strand ambiguous and with discrepant strand information in the two array annotations).

As further assurance of strand consistency, IMPUTE2 automatically addresses strand alignment at strand unambiguous SNPs (i.e. not A/T or C/G) 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. 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 ethnic composition. Thus, we instead chose to rely on the SNP annotation alone to align strand-ambiguous SNPs to the + 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, a freely available software program (see section IX, Web resources). 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. We first created 5 MB segments over the length of each chromosome from the first to last position appearing in the reference panel (i.e. starting at the first imputation target rather than position=1). Secondly, segments either overlapping the centromere or at the terminal ends of chromosomes were then merged into the segment immediately upstream. We then checked each segment for the presence of type 0 SNPs, as it is not logical to impute over an interval with no imputation target SNPs. These checks led to additional merging of centromere-adjacent segments on chromosomes 1, 3, 9, 16, and X. Ultimately we divided 23 chromosomes into 552 total segments, ranging from 6 segments on chromosomes 21 to 47 segments on chromosome 2. (Note that while IMPUTE2 provides recommendations for the segmentation method, it is up to the user to implement these criteria and actually define the segments.)

Lastly, we assessed our segmentation scheme in light of the recommendation from IMPUTE2 authors that each segment contain at least some observed GWAS (i.e. type 2) SNPs. Using Illumina HumanOmni2.5-4v1 array SNPs, we calculated an average density of 4,415 GWAS SNPs per segment (range 4-11,610; interquartile range 3,652-4,993). We took this as evidence that GWAS SNPs would be adequately represented in our proposed segments.

By default, IMPUTE2 flanks imputation segments with a 250 kb buffer, where type 2 SNPs are used to estimate haplotype structure but ultimately discarded from the imputation output. We chose to double the buffer size to 500 kb, which is closer to the 1 MB buffer size the GCC has previously used with BEAGLE imputation software. An example of the command line syntax used to run IMPUTE2 on the first 5 MB segment of chromosome 22 is shown below. Note the inclusion of the “-os 0 2” option, which specifies that only SNPs of types 0 and 2 should be written to imputation output files (i.e. removes type 3 “study only” SNPs from output). The file specified by the “-known_haps_g” flag is the phased haplotypes output by SHAPEIT2.

```
impute2 -use_prephased_g -m genetic_map_chr22_combined_b37.txt \
-h ALL_1000G_phase1integrated_v3_chr22_impute.hap.gz \
-l ALL_1000G_phase1integrated_v3_chr22_impute.legend.gz \
-int 16000001 2.1e+07 -buffer 500 -allow_large_regions \
-known_haps_g HRS2_chr22.haps.gz \
-filt_rules_l ma.cnt.gte4.allpanels < 1 \
-o HRS2_chr22.set1.gprobs -os 0 2 -o_gz \
-i HRS2_chr22.set1.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. Despite using pre-phased haplotypes, the large sample size caused imputation jobs to be both long running (four hours on average, but ranging from one to eight hours) and high memory (up to 32 GB, averaging ~15 GB). Due to the memory usage, each 12-core node could only accommodate three to four jobs at once. With this decreased parallelization (i.e. running three to four jobs instead of 12 jobs on each 12-core node), the elapsed calendar time for imputation was approximately three weeks.

VI. Imputation output

Imputation output files are divided by chromosome, where “23” denotes chromosome X. All study participants are consented for non-profit research use, thus alleviating the need to further divide the output by consent level. 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. Phased output

Results from the SHAPEIT2 “pre-phasing” step are posted as gz-compressed “.haps” and “.sample” files, both in IMPUTE2 input format. There are two identifiers in these files: ID_1, which corresponds to the PLINK family ID, and ID_2, corresponding to the PLINK individual-level ID. Note that the individual-level ID is the local subject ID (the field labeled “SUBJID” in annotation files). Regardless of the user’s desire for phased input haplotypes, the “.sample” files will likely be necessary for any downstream analyses, as sample identifiers are not included in the imputation output. The order of samples in the “.sample” files is the order of individuals in the imputation output files described below.

b. 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 raw IMPUTE2 output was to zero out any imputed genotypes in regions affected by gross chromosomal anomalies. 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. 0.33 0.33 0.33, or equal probabilities of each of the three genotype classes).   After imputation segments were processed for anomalies, they were combined into per-chromosome .gprobs file, via the Unix ‘cat’ command. Note that where the whole chromosome was recommended for filtering due to a gross chromosome anomaly, that individual was excluded from imputation on the given chromosome (see section II-c above).

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 two variant types as defined in the IMPUTE2 algorithm: type 0 (imputation target) and type 2 (imputation basis). 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).

c. 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 somewhat redundant "snp_id" field. Columns in these files are defined below, based on IMPUTE2 online documentation (see Web Resources).

- **rs_id:** variant identifier. For variants in dbSNP, the reference SNP (rs) number. Otherwise, the naming convention "*chr#-position"* is used. Note that where a single position is identified differently in the study and reference data (possible for type 2 variants only), this field reflects the identifier from the study dataset rather than from the reference.
- **position:** Base pair position (GRCh37)
- **exp_freq_a1:** Expected frequency of "allele A" (equivalent to "allele 1") in the genotype probabilities output file
- **info**: A statistical information metric, which is highly correlated with the squared correlation metrics output by BEAGLE[7] and MACH[17]. (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 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 SNP where type 0 denotes imputed variants (in 1000 Genomes but not study data) and type 2 denotes imputation basis variants (observed in the study data and used to impute type 0). Note type 3 variants have been excluded with the IMPUTE2 option "`-os 0 2`." 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). We have plotted imputed SNPs (panel A) separately from indels and SVs (panel B). While average "info" scores at SNPs with MAF < 0.10 fall below 0.9, the remaining SNPs (those with MAF > 0.10) have average "info" scores around 0.95. The average "info" scores for SVs and indels are slightly lower, leveling off at 0.9 when MAF is 0.10 or more. We also plotted these metrics by chromosome, to assess quality in the slightly more complicated X chromosome imputation. As seen in Figure 4, the X chromosome does not appear to be an outlier for either SNPs or indels/SVs, indicating that imputation quality at X chromosome variants is comparable to the autosomes.

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[14]. The GCC 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[14]. 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 98.0% of all imputed variants for downstream analyses, while more stringent thresholds of 0.5 and 0.8 would retain 93.2% and 76.7% of imputed variants, respectively. However, users should be aware that setting stringent quality thresholds has been shown to result in missing true positive associations[18].

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[19]. However, genotype-level filtering does not make use of the full information at a given marker and therefore may be less desirable than the SNP level filters described above.

d. Masked SNP analysis

A common way to assess imputation quality, beyond the theoretical calculations of accuracy discussed above, is to intentionally "mask" a subset of the SNPs genotyped in the study sample (i.e. remove from the imputation basis), impute the masked SNPs 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} SNP, 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 SNP, 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 SNP analysis is integrated into every IMPUTE2 imputation run: each study SNP (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 SNP includes results from the masked SNP test, including concordance and correlation between imputed and observed results, as well as the “info” metric from treating the SNP as type 0. Below we assess the quality metrics of all SNPs masked in this imputation, a total of 1,945,761 masked SNPs (i.e. all type 2 SNPs). Note that while the Omni2.5M array contains some indels, none were type 2 variants (i.e. not in the imputation basis) and thus were not included in the masking.

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, on the secondary y-axis, the percentage of SNPs in the bin with “info_type0” ≥ 0.8. In panels B and C, 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 exclude SNPs with “info_type0” < 0.8. The metric shown in panel (B) is the correlation between masked and imputed allelic dosages; the metric in panel (C) is the concordance: the fraction of identical genotypes between the most likely imputed and observed.

Several salient points emerge from these graphs. Firstly, there is a sharp decline in empirical dosage r^2 for lower frequency variants (MAF < 0.05). As MAF increases, however, average correlation values level off to > 0.95. Secondly, 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 SNPs. This filtering improves the quality metrics profile for masked SNPs across the entire range of MAF bins. Thirdly, Figure 5C 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[20]. 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.

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 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.

Lastly, when discussing imputation quality there can be several different meanings of "efficiency." Figure 5A illustrates one definition: the percentage of imputed SNPs 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 percentage 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.

e. Downstream analysis

Many references are available for users desiring further information on imputation methods, including recommendations and caveats for downstream analyses[1,2,11,14,21]. 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 as recommended in section VI-c. 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[22], an R package developed by the GCC; PLINK (with the --dosage option: http://pngu.mgh.harvard.edu/~purcell/plink/dosage.shtml); MACH2qtl/dat[17]; SNPTEST[23]; ProbABEL[24]; BIMBAM[25]; SNPMStat[26]; and the R package snpMatrix[27]. For a comparison of methods to account for genotype uncertainty in imputed data, see Zheng et al[28].

VII. Summary

We have performed genotype imputation in the Phases 1-3 of the Health and Retirement Study, using a worldwide 1000 Genomes Project reference panel and IMPUTE2 software. The imputed genotypes and accompanying variant annotation and quality metrics files are available through the authorized access portion of the dbGaP posting. These imputation analyses were performed by Wenying Zheng and documented by Sarah Nelson, under the leadership of Cathy Laurie and Bruce Weir, within the GCC at the University of Washington (UW) in Seattle, WA. This report was reviewed and approved by study investigators David Weir, Sharon Kardia, and Jennifer Smith, at the University of Michigan.

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. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
5. 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).
6. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
7. 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).
8. B. Howie, J.M., and M. Stephens. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
9. Frazer, K. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
10. Altshuler, D. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
11. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
12. Durbin, R.M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
13. McVean, G. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
14. de Bakker, P. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
15. 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).
16. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
17. 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).
18. 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).
19. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet* **125**, 163-71 (2009).
20. Lin, P. et al. A new statistic to evaluate imputation reliability. *PLoS One* **5**, e9697 (2010).
21. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet* **4**, e1000279 (2008).
22. 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).

23. 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).
24. 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).
25. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).
26. 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).
27. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum Hered* **64**, 45-51 (2007).
28. 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. Retrieved from http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html on April 19, 2012.

The 1000 Genomes Project. Phase1 integrated release version3 [released April 2012]. Available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/

Delaneau O (Version 2.r644, c2011-2012) SHAPEIT: Segmented HAPlotype Estimation and Imputation Tool [software]. Available from http://www.shapeit.fr/.

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 2.3.0 [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/.

The Regents of the University of Michigan (2013). Health and Retirement Study. Retrieved from http://hrsonline.isr.umich.edu/, September 27, 2013.

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	165,995	151,463	1,748,658
2	176,221	161,581	1,898,462
3	148,905	136,430	1,600,113
4	137,831	126,400	1,620,107
5	132,230	121,114	1,469,430
6	131,969	120,889	1,441,584
7	115,887	106,310	1,310,531
8	113,938	105,033	1,260,536
9	92,620	85,819	968,540
10	107,690	98,947	1,108,716
11	104,503	95,783	1,105,001
12	101,513	92,816	1,074,574
13	75,348	69,215	809,808
14	69,313	63,743	736,419
15	65,636	60,415	659,228
16	68,313	63,113	702,378
17	58,711	53,845	611,389
18	62,412	57,768	638,001
19	40,465	37,009	502,780
20	51,066	47,472	499,158
21	28,685	26,369	309,838
22	28,964	26,964	303,166
X	40,169	37,263	692,472
Totals	***2,118,384***	***1,945,761***	***23,070,889***

† Study SNPs passing pre-imputation filters (IMPUTE2 variants types 2 and 3).
†† Study SNPs passing pre-imputation filters and overlapping with the reference panel (type 2).
Imputation output is the sum of imputation basis (type 2) and imputation target (type 0) variants. Type 0 variants have been restricted to those with at least four copies of the minor allele in any of the four 1000 Genomes continental ancestry panels.

Table 2. An overview of the 1,092 samples in the 1000 Genomes Project worldwide reference panel (phase I integrated variant set v3, March 2012), which was used to impute HRS participants. Each population was assigned to one of four continental groupings: African (AFR), American (AMR), Asian (ASN), and European (EUR). All haplotypes in the phased reference panel are for unrelated, founder individuals only. 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 Ancestry in Southwest US	ASW	61
Luhya in Webuye, Kenya	LWK	97
Yoruba in Ibadan, Nigeria	YRI	88
Total African ancestry	***AFR***	***246***
Colombian in Medellin, Colombia	CLM	60
Mexican Ancestry in Los Angeles, CA	MXL	66
Puerto Rican in Puerto Rico	PUR	55
Total American ancestry	***AMR***	***181***
Han Chinese in Beijing, China	CHB	97
Han Chinese South, China	CHS	100
Japanese in Tokyo, Japan	JPT	89
Total Asian ancestry	***ASN***	***286***
Utah residents (CEPH) with Northern and Western European ancestry	CEU	85
Toscani in Italia	TSI	98
British in England and Scotland	GBR	89
Finnish in Finland	FIN	93
Iberian populations in Spain	IBS	14
Total European ancestry	***EUR***	***379***

Table 3. Quality metrics for all masked SNPs, dichotomized into groups of MAF < 0.05 vs. MAF ≥ 0.05. The second column shows the number of SNPs 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 SNPs in each MAF category are included in these averages.

MAF **(in study samples)**	**Number of SNPs**	**Mean (Median)** **Overall Concordance**	**Mean (Median)** **empirical dosage r^2**
< 0.05	774,796	0.995 (0.998)	0.817 (0.901)
≥ 0.05	1,170,965	0.983 (0.994)	0.953 (0.983)

XI. Figures

Figure 1. Principal component analysis (PCA) of unique, unrelated HRS samples from the combined dataset, Phases 1-3. Color-coding is according to self-identified race; plotting symbol denotes self-identified ethnicity (Mexican American, Other Hispanic or not Hispanic). The axis labels indicate the percentage of variance explained by each eigenvector. Also Figure 16 from the Phase 3 genotype QC report.

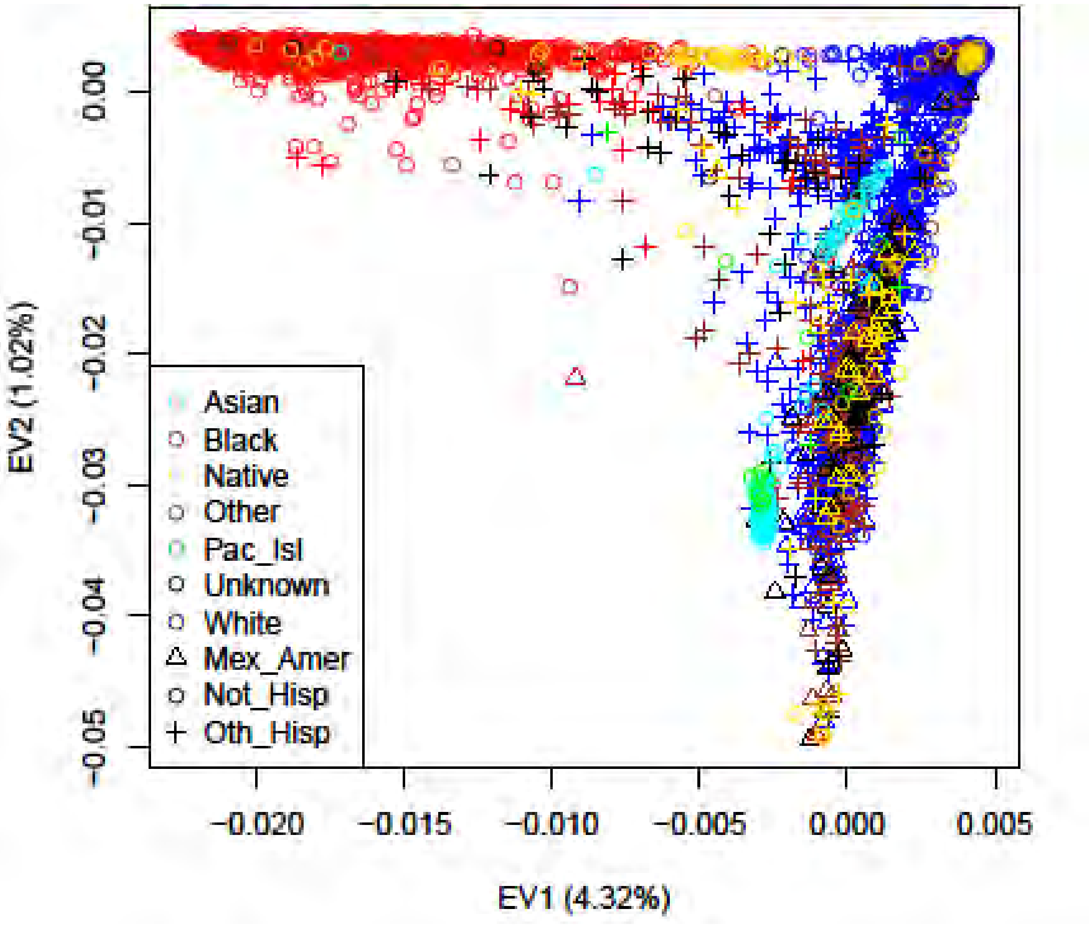


Figure 2. A schematic of variant types as defined in the IMPUTE2 imputation algorithm. Each individual is represented by a unique color in the horizontal bar(s), and alternate alleles at each variant are represented as *A* and *B*. Section (A) represents phased reference haplotypes, where two samples (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).

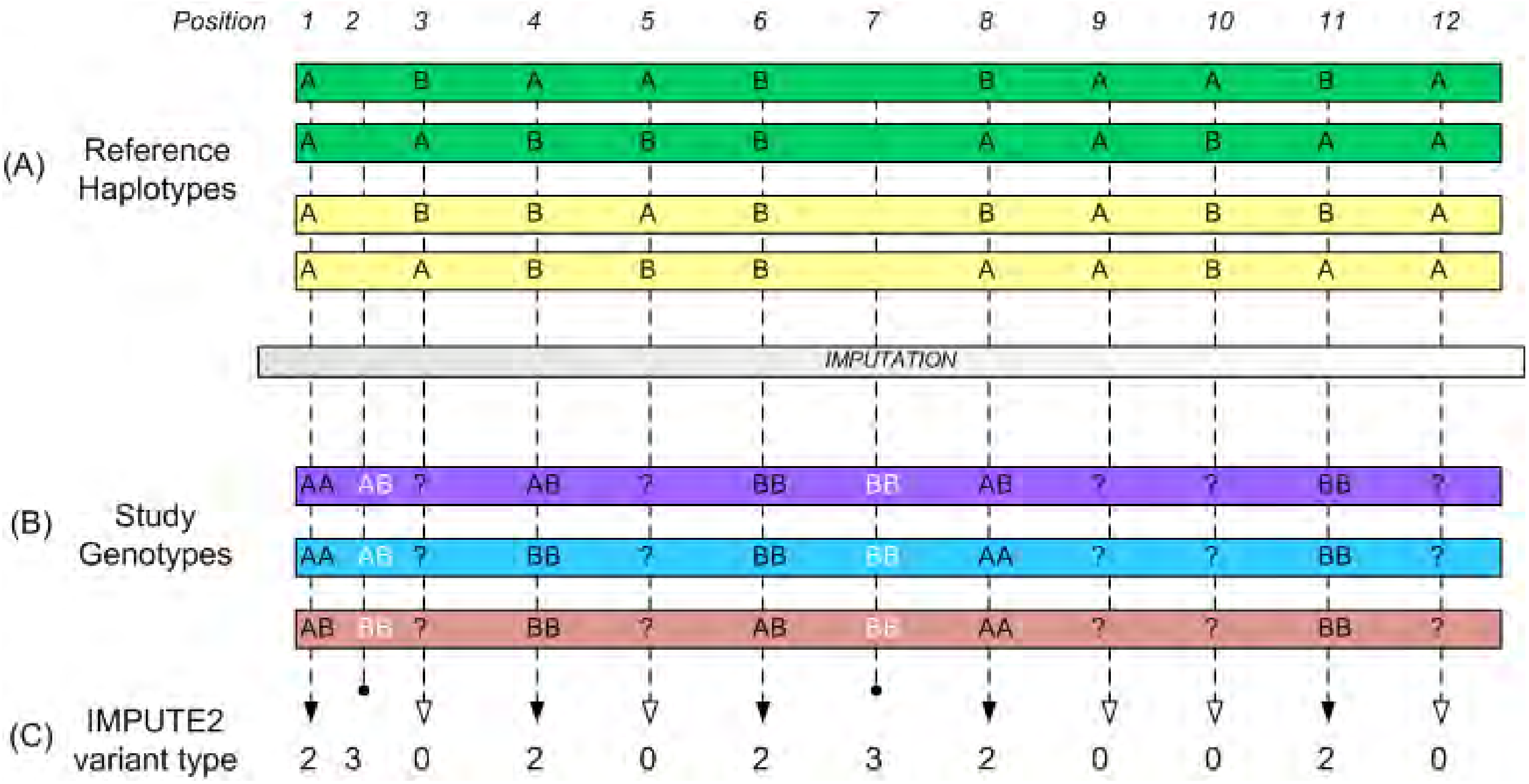


Figure 3. Summaries of quality metrics at all imputed variants: SNPs, SVs, and indels. In each plot, imputed variants are binned by MAF (0.01 intervals) along the x-axis and then average "info" score per bin is plotted on the y-axis. Panel (A) is for SNPs and panel (B) for indels and SVs. The secondary y-axes indicates the count of variants in each MAF bin.

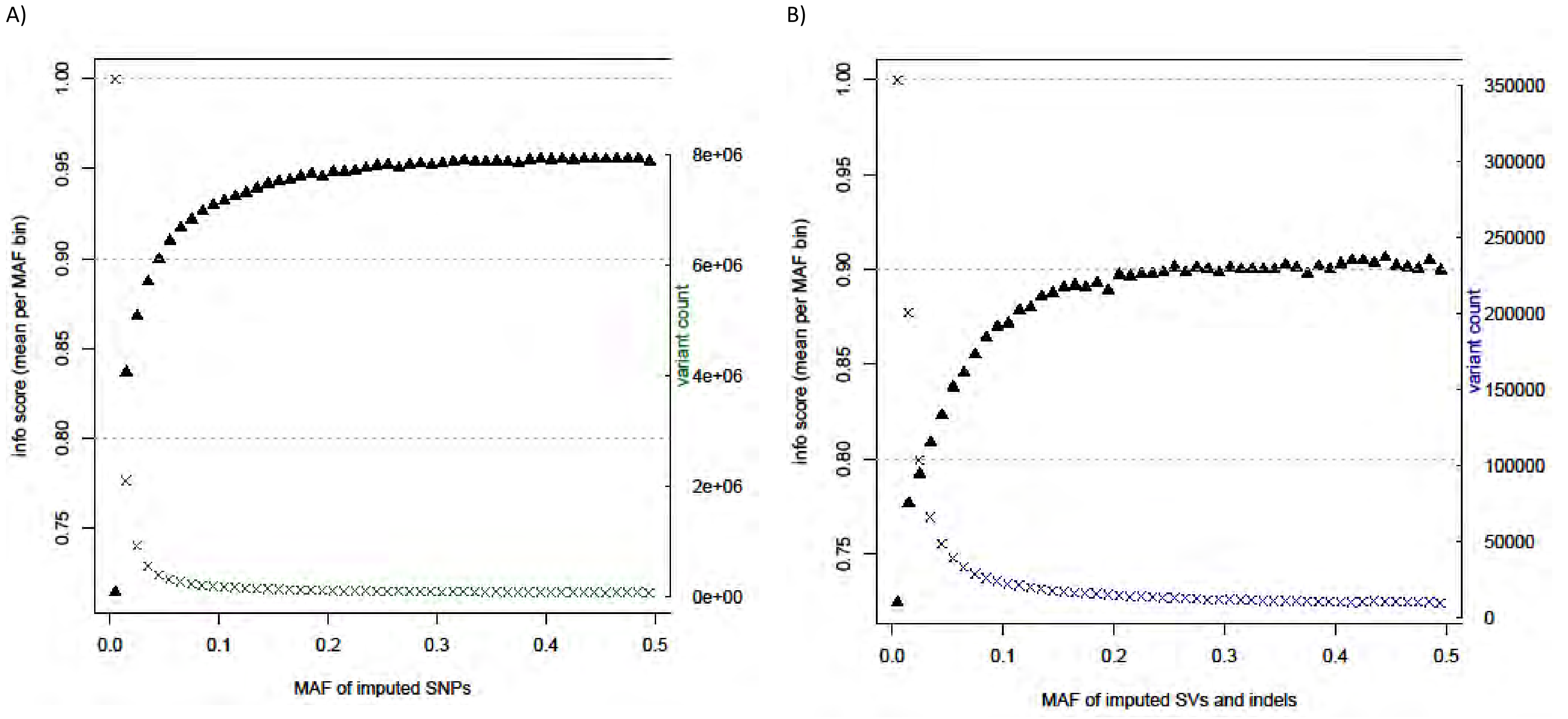
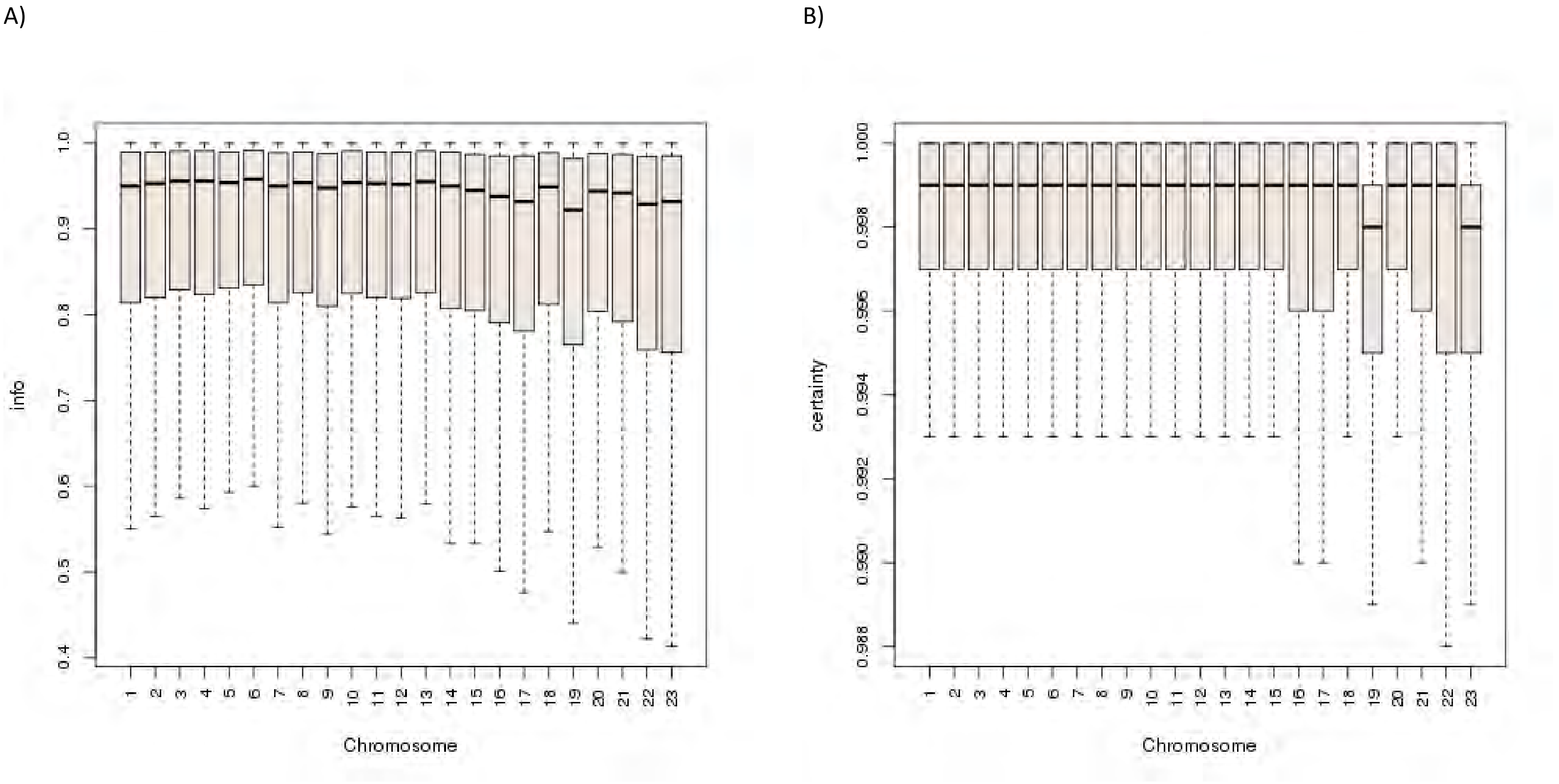


Figure 4. A comparison of imputation quality metrics by chromosome for all imputed SNPs, indels, and SVs: “info” in panel (A) and “certainty” in panel (B) for SNPs, “info” in panel (C) and “certainty” in panel (D) for indels and SVs. Outlier values are not displayed in these box plots. On the x-axis, “23” denotes the X chromosome.



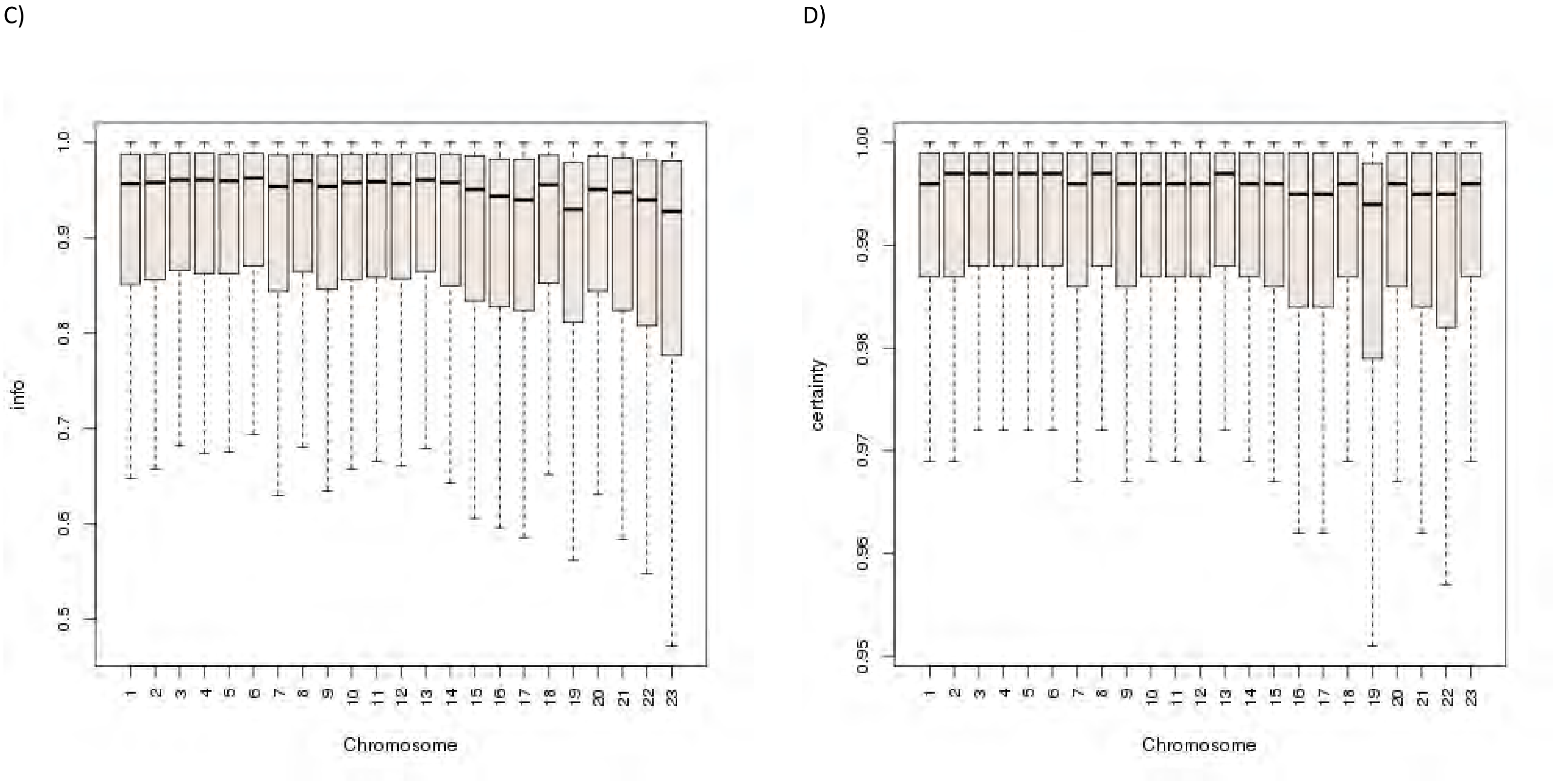
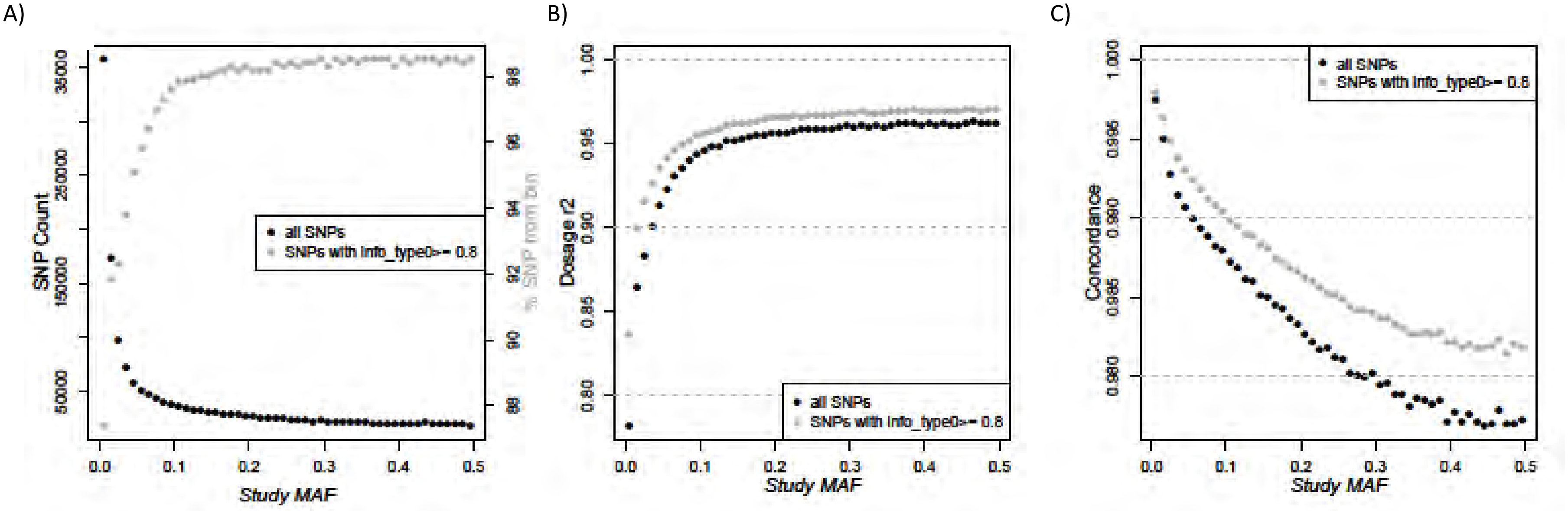

C)
D)
info
1.0
0.9
0.8
0.7
0.6
0.5
certainty
1.00
0.99
0.98
0.97
0.96
0.95
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome

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, on the secondary y-axis, the percentage of SNPs in the bin passing an “info” filter threshold of ≥ 0.8. Panel (B) plots the average empirical dosage r^2 metric per MAF bin, both before and after filtering on the “info” score (black and gray data series, respectively). Similarly, panel (C) is the concordance between the observed and the most likely imputed genotype at masked SNPs within each MAF bin, with and without the “info” filter.



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 (552 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
2. The file "filtered_map.txt" is a list of subject-segment combinations where imputed genotypes were set to missing (i.e. 0.33 0.33 0.33, or equal probabilities of each of the three genotype classes). The fields in this file are:
    - **subjectID**: participant level identifier assigned by the CC, used in imputation output
    - **chrom:** chromosome
    - **segment:** imputation segment ID

b. **SNP selection.**

The file "snp.qualfilter.txt" is a list of genotyped variants passing GCC 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-c). The SNP dimension in this file corresponds to the "Study SNPs" column in Table 1. The columns in these text files are:

- **rs.id**: refSNP identifier in build 37.
- **chrom:** chromosome number, in build 37 mapping.

c. **Sample-subject mapping.** The identifier used in the imputation output is the "SUBJID." A mapping of "SUBJID" to "SAMPID," which corresponds to one genotype scan, is provided in the file "subjectid2scanid.txt." The columns in this file are:

- **SUBJID**: local (study investigator's) participant level identifer, used in imputation output
- **SAMPID**: local (study investigator's) sample level identifier, corresponding to one genotype scan
- **scanID**: scan-level identifier assigned by the GCC
- **sex**: male (M) or female (F)
- **phasing.type:** Sample type in the SHAPEIT phasing analysis. Possible values are: Unr, DuoC, DuoM, DuoF, TrioC, TrioM and TrioF, which stand for unrelated, duo child, duo mother, duo father, trio child, trio mother, and trio father, respectively.