

Epidemiology of Biomarkers of Risk and Progression in Late Onset Alzheimer Disease

Imputation Report - 1000 Genomes Project reference panel

November 21, 2012

Contents

- I. [Summary and recommendations for dbGaP users](#)
- II. [Study data](#)
 - a. Samples
 - b. SNPs
 - c. Data formatting
 - d. Pre-phasing
- III. [Reference panel](#)
- IV. [Strand alignment](#)
- V. [Imputation software and computing resources](#)
- VI. [Imputation output](#)
 - a. Phased output
 - b. Genotype probabilities
 - c. Quality metrics
 - d. Masked SNP analysis
 - e. Downstream analysis
- VII. [Summary](#)
- VIII. [References](#)
- IX. [Web resources](#)
- X. [Tables](#)
- XI. [Figures](#)
- XII. [Supplementary files](#)

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 Genetics Coordinating Center (GCC) used IMPUTE2 software³ to perform genotype imputation in this genome-wide association study of late onset Alzheimer disease in Caribbean Hispanics (Richard Mayeux, PI). Imputed results are provided as the probability of each of the three genotype states at each SNP, 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-SNP basis. For a detailed description of genotype quality control (QC) on this project, please see the genotype data cleaning report available through dbGaP (database of Genotypes and Phenotypes; study accession number phs000496).

II. Study data

a. Samples

The aim of this genome-wide association study (GWAS) is to identify new and confirm previously found genetic loci for late onset Alzheimer disease (LOAD). Participants are Dominican and Puerto Rican individuals from three different studies: a cohort, a family-based, and a case-control study. The dataset comprises Caribbean Hispanics living in Manhattan, NY, NY and residents of the Dominican Republic and Puerto Rico, as well as both families and unrelated individuals.

A total of 3,239 samples were successfully genotyped at the Center for Inherited Disease Research (CIDR). After the GCC's standardized QC procedures⁴, genotypes were available for 3,139 unique study participants. Each of these study samples had an overall missing call rate < 2% and was therefore included in this imputation. Within these 3,139 participants was one pair of monozygotic twins, of which only one was retained for imputation such that 3,138 samples were imputed. Initially, no data were removed based on gross chromosomal anomalies identified during data cleaning. Rather, chromosomal segments containing anomalies were filtered out of imputed results after imputation was completed, as discussed later in this report.

Figure 1 shows a principal component analysis (PCA) of study samples with select HapMap⁵ control populations, which illustrates that the genetic ancestry of most study participants seems to be mainly along the European-African transition line, with the addition of Native American ancestry in a subset of participants. 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. This approach also aligned with our desire to impute the maximum number of participants, in that no study samples were excluded from imputation based on principal components values.

The local subject level identifier (“SUBJID” in annotation files) was used as the individual identifier throughout, which can be mapped to the local sample scan identifier (“SAMPID,” corresponding to one genotype scan) using the sample-subject mapping file provided in the Supplementary Files (section XII).

b. SNPs

This LOAD GWAS was genotyped on the Illumina HumanOmni1-Quad_v1 SNP array, designed to human genome build 37. For the purposes of imputation, study SNPs were selected using CC recommended SNP quality filters described in the 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 in the same manner as the completely unobserved SNPs and should be treated with the same caveats. 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 GWAS release and these imputed data. The SNP annotation files accompanying this report can be used to differentiate between observed study SNPs used in the imputation input and the imputed SNPs. We refer to the former set of SNPs as the “imputation basis” and to the latter as “imputation target” SNPs. These terms are analogous to the IMPUTE2 definitions of “type 2” and “type 0” SNPs, respectively. (Note that “type 1” SNPs occur only when more than one reference panel is used with IMPUTE2.) Lastly, we refer to study SNPs that do not occur in the reference as “study only” SNPs, or “type 3” in IMPUTE2. See Figure 2 for a visual representation of these SNP types.

c. Data formatting

The study genotype data were initially accessed from the binary PLINK⁶ file available in the dbGaP release, “CIDR_Mayeux_Top_subject_level_24Oct2012,” with genotypes expressed in TOP alleles. The Illumina annotation file, which included genomic strand information, was thus used to identify the SNPs requiring a strand flip to convert the TOP allele to the “+” strand of the human genome reference assembly (see 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 SNPs passing the quality filter; (4) updated parental identifiers to conform with relatedness constraints of pre-phasing software (discussed in the following section); and (5) specified a list of SNPs that required a strand flip to align with the “+” strand, based on Illumina annotation. Below is an example of the command line syntax used to create the filtered binary files, for generic chromosome “#”:

```

plink --bfile CIDR_Mayeux_Top_subject_level_24Oct2012 \
--extract snp.qualfilter.txt --flip fliplist.txt \
--keep sampkeep.txt --update-parents update-parents.txt \
--set-hh-missing --chr # --make-bed \
--out Mayeux_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 size⁷. 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 GWAS.

An additional advantage to pre-phasing in family studies is that it can incorporate family structure, while most imputation algorithms “ignore” relatedness due to computational and programmatic constraints. To our knowledge, only BEAGLE software⁸ has the capability to utilize sample relatedness during the imputation step (when imputing into duos and trios), but this process is very computationally intensive and currently not practical for large family studies. 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.

We utilized the pre-phasing approach in this imputation, due to the advantages enumerated above. This GWAS contained over 200 families, ranging in size from 2 to 23 members. Pre-phasing enabled much of this family information to be used when determining haplotype phase in these samples. Following “best practices” guidelines in the IMPUTE2 documentation (see Web resources), we first phased the study data with the SHAPEIT⁹ haplotype estimation tool. We used a pre-release version of SHAPEIT version 2, which is “more accurate than version 1, especially in large datasets” (personal communication, Jonathan Marchini, May 9, 2012). The SHAPEIT authors kindly made SHAPEIT2 available to us in pre-release form, pending publication of their new and improved methods.

In its current implementation, SHAPEIT2 supports phasing of trios, duos, and unrelated samples. However, related individuals can only occur in one trio or duo. Therefore, in order to carry out pre-phasing via SHAPEIT, first we pruned familial relationships at individuals (n=19) occurring in more than one trio or duo. For the most part, this involved keeping only one offspring as related in multi-sibling families where at least one parent was genotyped,

while making the remaining siblings unrelated. In each of these situations, we prioritized keeping the proband as the unrelated offspring. The familial pruning was achieved using the “--update-parents” flag when extracting data from the PLINK binary dataset (see previous section), setting parental ID values to “0” at selected family members.

With these familial adjustments made, we then input the filtered, chromosome-specific PLINK files (see II-c) into SHAPEIT2 and received the best guess haplotypes as output. Haplotype phase was determined across all autosomes, for 5 trios, 46 duos, and 3,031 unrelateds. The phasing type for each participant (duo or unrelated) is included in the supplementary files section of this report, including whether they were pruned for relatedness. SHAPEIT2 ignores all sample relatedness when phasing chromosome X, thus all 3,138 samples were phased as unrelated on this sex chromosome.

Each per-chromosome SHAPEIT2 job was distributed across multiple computing cores, via the “--thread” argument. Runtimes ranged from 6 to 42 hours, 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 --input-bed Mayeux_chr#.bed \
Mayeux_chr#.bim Mayeux_chr#.fam\
--input-map genetic_map_chr#_combined_b37.txt --states-phase 200 \
--output-max Mayeux_chr#.haps.gz Mayeux_chr#.sample.gz \
--thread 6 --output-log shapeit_chr#.log
```

III. Reference panel

Larger reference panels have been shown to increase imputation accuracy^{2,8,10}. Previously, haplotypes from Phases 2¹¹ and 3⁵ 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^{2,12}.

The 1000 Genomes Project aims to “discover, genotype, and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations¹³.” In October 2011, the Project released the first version of the phase I integrated variant set, containing SNPs, insertion/deletions (indels), and structural variants (SVs) in 1,092 samples from 14 different populations¹⁴. The Project has categorized each of these populations into four continental groupings: African (AFR), American (AMR), Asian (ASN), and European (EUR). To impute these LOAD GWAS study 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 region¹⁰. While this approach eases the computational burden of using all reference samples, it still may not warrant the imputation of all available reference SNPs (i.e. approximately 38 million variants). Very low MAF SNPs 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 SNPs in downstream analyses. Therefore, we restricted imputation to SNPs with at least four copies of the minor allele in any one of the four 1000 Genomes continental groups: AFR, AMR, ASN, or EUR. We also excluded indels and SVs (approximately 1.5M variants), due to the current lack of information regarding how accurately such variants can be imputed.

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 straightforward¹⁵. The initial study dataset contained TOP alleles, an Illumina naming method unrelated to “+” or “-” strand orientation¹⁶ (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.

As further assurance of 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. 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, 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 base pair 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 currently 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 HumanOmni1-Quad_v1 array SNPs, we calculated an average density of 2,032 GWAS SNPs per segment (range 51-11,830; interquartile range 1,574-2,304). 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 haplotypes 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 CC 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 Mayeux_chr22.haps.gz \
-filt_rules_l ma.cnt.gte4.allpanels<1 sv.indel>0 \
-o Mayeux_chr22.set1.gprobs -os 0 2 -o_gz \
-i Mayeux_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. Due to the input of pre-phased haplotypes, the compute time required to impute most segments was approximately one hour. In total, the imputation took approximately one week of calendar time, after accounting for pre-phasing and the degree of parallelization enabled by our compute cluster.

VI. Imputation output

Imputation output files are divided by chromosome, where “23” denotes chromosome X. All study participants are consented for general 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.”

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). The SHAPEIT2 phasing type (trio, duo, or unrelated) is available in Supplementary Files accompanying this report. Note the family structures used for phasing may differ from the complete pedigree information, where familial pruning was required to conform to SHAPEIT2 data input requirements. 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 (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. 0.33 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.

The first five columns in these output files correspond to SNP ID; rs ID; physical position; and the two SNP 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 SNP types as defined in the IMPUTE2 algorithm: type 0 (imputation target) and type 2 (imputation basis). The SNP 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 SNP annotation and quality metrics file, with each row of a genotype file corresponding to a row in the SNP 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:** SNP 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 SNPs 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 BEAGLE⁸ and MACH¹⁷. (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 SNPs will have “info” values of ~1. For type 0 SNPs, however, the “info” metric is 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 SNP.
- **type:** Internal type assigned to each SNP where type 0 denotes imputed SNPs (in 1000 Genomes but not study data) and type 2 denotes imputation basis SNPs (observed in the study data and used to impute type 0). Note type 3 SNPs have been excluded with the IMPUTE2 option “-os 0 2.” See Figure 2 for a schematic of these SNP types.

Note: the following fields are defined only at type 2 SNPs, 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 SNP treated as type 0 (i.e. when it was masked)

Figure 3 includes distributions of the “info” and “certainty” metrics for all imputed SNPs (panels A and B, respectively). In Figure 3C, average “info” scores are plotted in SNPs grouped by imputed MAF (bin sizes of 0.01), demonstrating the relationship between MAF and imputation quality. 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 > 0.9. 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, indicating that imputation quality at X chromosome SNPs is comparable to autosomal SNPs.

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 this¹⁵. The CC recommends a SNP level filter, in which only SNPs 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 SNPs with a quality metric of ≥ 0.3 ¹⁵. 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 94.8% of all imputed SNPs for downstream analyses, while more stringent thresholds of 0.5 and 0.8 would retain 89.9% and 70.7% of imputed SNPs, respectively. Others have also suggested using different quality metric threshold values for different MAF bins of imputed SNPs¹⁸.

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 genotypes¹⁹. 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 $E(d_{ij}) = 2 * P(\text{AA}) + 1 * P(\text{AB}) +$

$0 \cdot P(BB)$. 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 924,700 masked SNPs (i.e. all type 2 SNPs).

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 includes all masked SNPs while the gray data series excludes 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 decline in empirical dosage r^2 for low-frequency variants (MAF < 0.05). As MAF increases, however, average correlation values level off to > 0.9 . 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 $> 90\%$ concordance²⁰. 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, via the concordance measures. 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.

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, here we do not include the per-SNP efficiency metric described above. 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^{1,2,12,15,21}. Programs for performing association analyses with imputed genotype probabilities include PLINK (with the `--dosage` option: <http://pngu.mgh.harvard.edu/~purcell/plink/dosage.shtml>), MACH2qtl/daf¹⁷, SNPTEST²², ProbABEL²³, BIMBAM²⁴, SNPStat²⁵, and the R package snpMatrix²⁶. For a comparison of methods to account for genotype uncertainty in imputed data, see Zheng et al²⁷. IMPUTE2 is part of a suite of GWAS software that is useful in post-imputation data filtering and formatting tasks (see Web Resources, “Genome-wide Association Study Software Suite”). For example, QCTOOL may be used to filter imputed data by the IMPUTE2 “info” score as recommended in section VI-c.

VII. Summary

We have performed genotype imputation in this GWAS of late onset Alzheimer disease in Caribbean Hispanics, using a worldwide 1000 Genomes Project 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 Genetics Coordinating Center at the University of Washington (UW) in Seattle, WA. This report was reviewed and approved by study investigators Richard Mayeux, Joseph Lee, and Amanda Sena, at Colombia University.

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. Laurie, C.C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* **34**, 591-602 (2010).
5. Altshuler, D. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
6. 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).
7. 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).
8. 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).
9. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
10. B. Howie, J.M., and M. Stephens. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
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. de Bakker, P. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
16. 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).
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. Liu, E.Y. et al. Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. *Genet Epidemiol* **36**, 107-17 (2012).
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. 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).

23. 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).
24. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).
25. 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).
26. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum Hered* **64**, 45-51 (2007).
27. 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 v v1.532, c2011) 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 (c2007-2011) IMPUTE version 2.2.2 [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/>

X. Tables

Table 1. SNP summary

Chromosome	Study SNPs [†]	Imputation basis ^{††}	Imputation Output
1	77,733	76,550	1,639,361
2	73,005	72,267	1,781,822
3	59,339	58,740	1,501,569
4	55,270	54,642	1,517,997
5	53,787	53,235	1,378,896
6	69,969	68,840	1,348,512
7	49,129	48,431	1,228,557
8	49,132	48,602	1,188,831
9	43,282	42,746	911,106
10	49,431	48,825	1,040,951
11	46,231	45,663	1,038,281
12	44,950	44,421	1,006,836
13	32,613	32,334	756,741
14	28,302	28,003	689,733
15	27,773	27,420	618,025
16	29,696	29,183	664,752
17	26,255	25,811	572,210
18	26,313	26,065	597,557
19	20,755	20,238	469,034
20	25,667	25,210	469,329
21	13,231	13,059	289,475
22	13,419	13,235	284,543
X	21,464	21,180	637,930
Totals	936,746	924,700	21,632,048

[†] Study SNPs passing pre-imputation filters (IMPUTE2 SNP 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) SNPs. Type 0 SNPs have been restricted to those with 4 or more copies of the minor allele in AFR, AMR, ASN, or EUR 1000G samples.

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 all study 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
<i>Total African ancestry</i>	<i>AFR</i>	<i>246</i>
Colombian in Medellin, Colombia	CLM	60
Mexican Ancestry in Los Angeles, CA	MXL	66
Puerto Rican in Puerto Rico	PUR	55
<i>Total American ancestry</i>	<i>AMR</i>	<i>181</i>
Han Chinese in Beijing, China	CHB	97
Han Chinese South, China	CHS	100
Japanese in Tokyo, Japan	JPT	89
<i>Total Asian ancestry</i>	<i>ASN</i>	<i>286</i>
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
<i>Total European ancestry</i>	<i>EUR</i>	<i>379</i>

Table 3. Quality metrics for all masked SNPs, dichotomized into groups of $MAF < 0.1$ vs. $MAF \geq 0.1$. 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	126,536	0.991 (0.995)	0.822 (0.900)
≥ 0.05	798,164	0.971 (0.986)	0.925 (0.964)

XI. Figures

Figure 1. Principal component analysis of 3,138 study participants with 1,203 HapMap controls. HapMap samples are color-coded by population: African ancestry in Southwest USA (ASW); Utah residents (CEPH) with Northern and Western European ancestry (CEU); Han Chinese in Beijing, China (CHB); Chinese in Metropolitan Denver, Colorado (CHD); Gujarati Indians in Houston, Texas (GIH); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Mexican ancestry in Los Angeles, California (MEX); Maasai in Kinyawa, Kenya (MKK); Tuscans in Italy (TSI); Yoruba in Ibadan, Nigeria (YRI). The percent variance explained by each of these first two components is noted on the axis labels. (Also Figure 9c from the genotype QC report.)

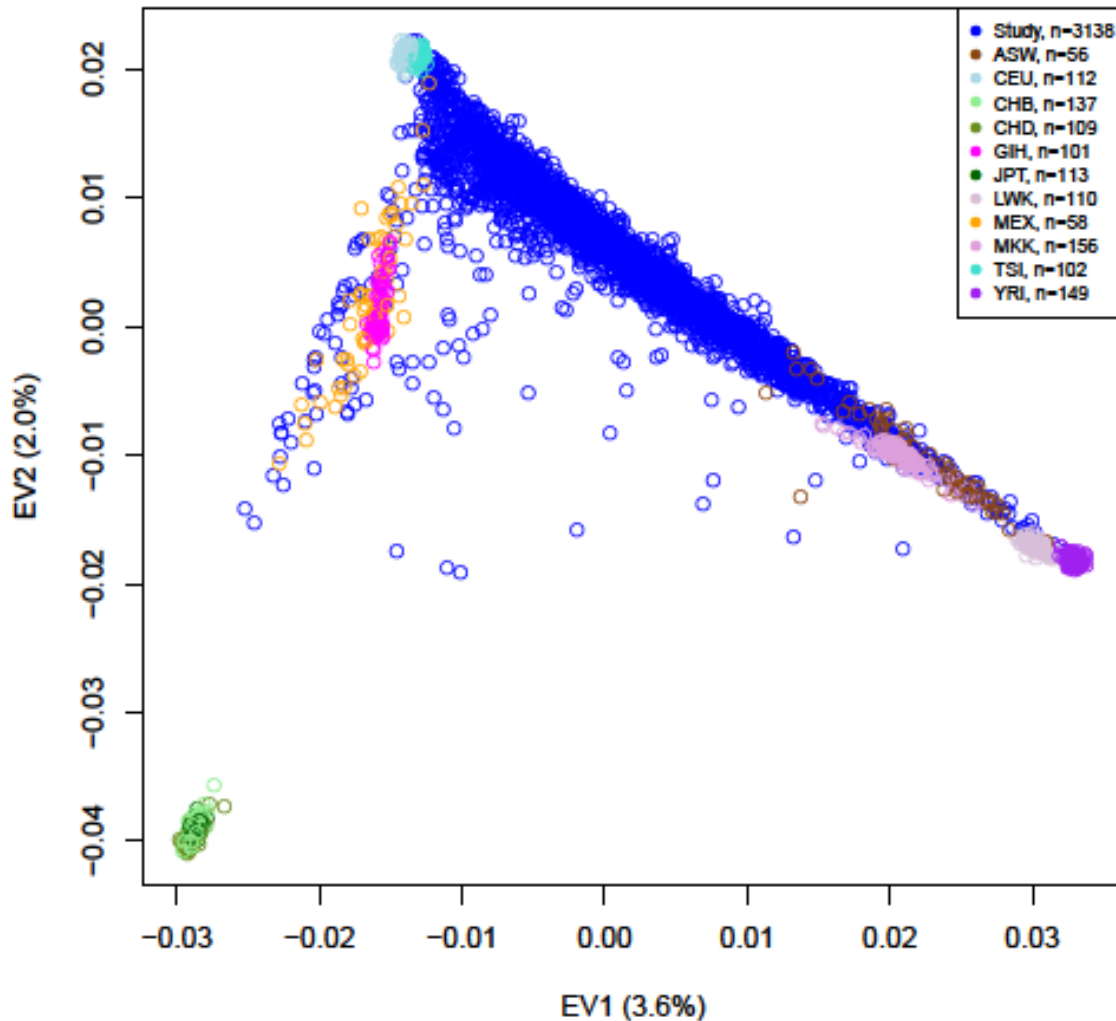


Figure 2. A schematic of SNP types as defined in the IMPUTE2 imputation algorithm. Each individual is represented by a unique color in the horizontal bar(s), and alternate alleles at each SNP are represented as A and B. Section (A) represents phased reference haplotypes, where two samples (4 phased chromosomes)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

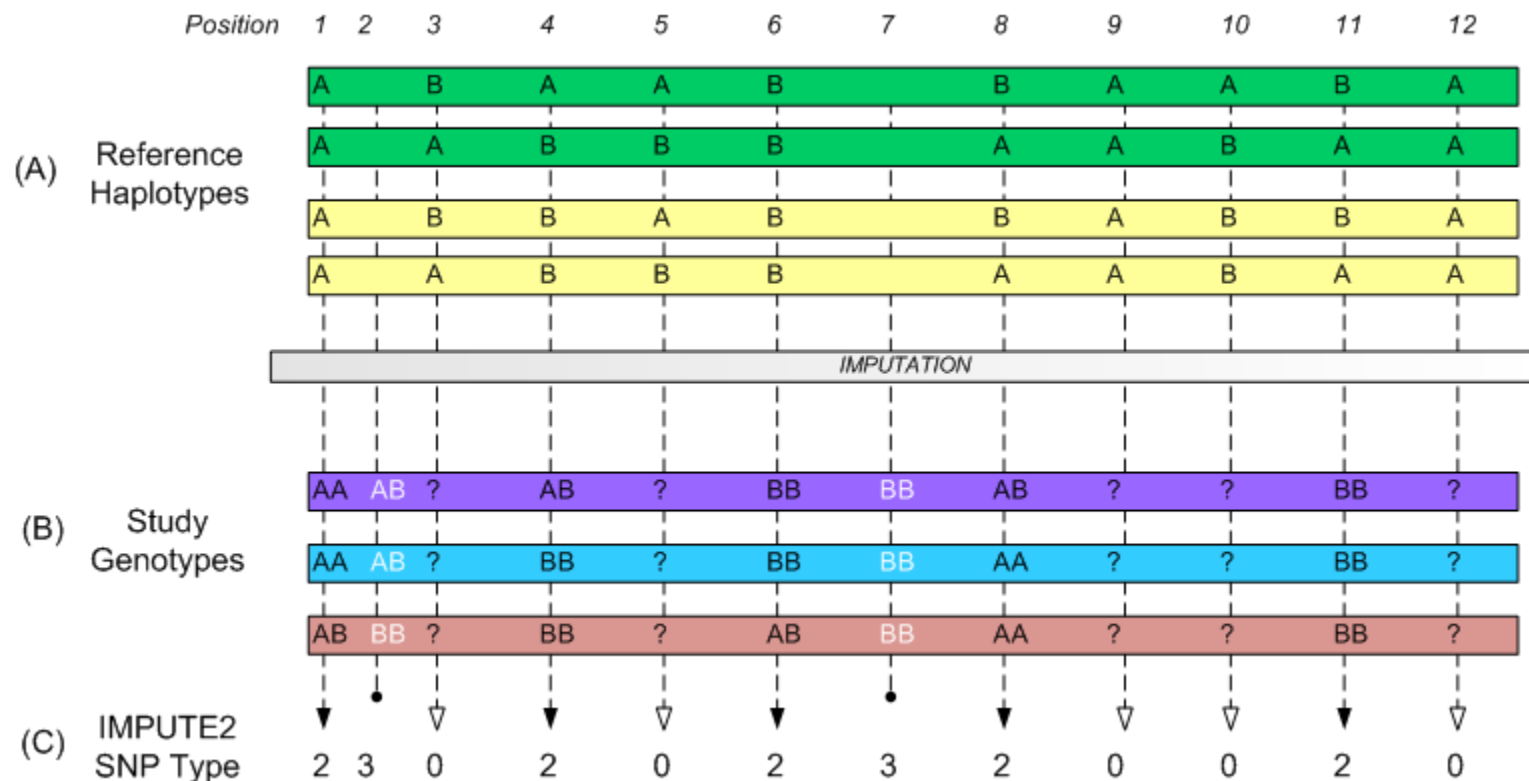
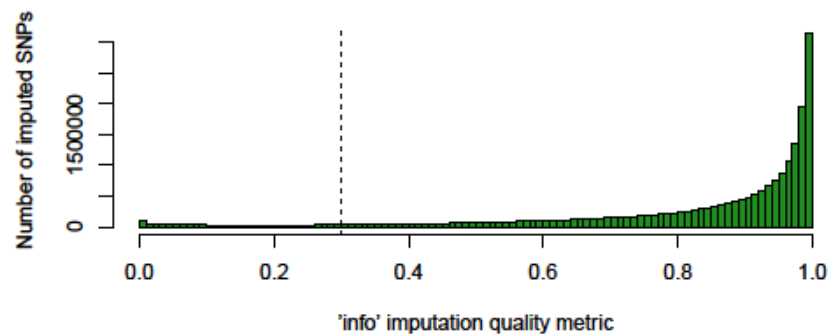
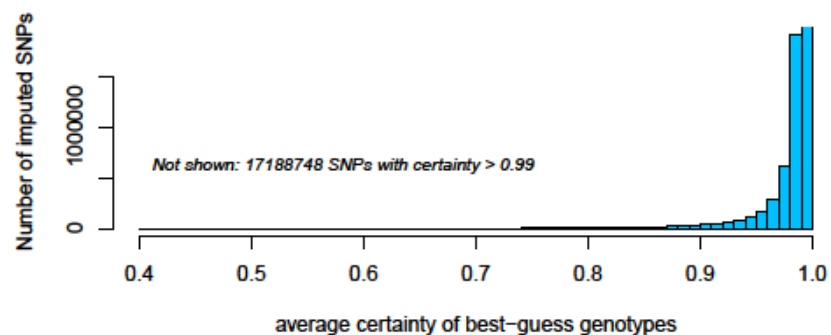


Figure 3. Summaries of quality metrics at all imputed SNPs. Panel A shows the distribution of the “info” quality metric, with a dashed line indicating a potential 0.3 threshold value. Panel B is the distribution of “certainty,” the average certainty of best-guess genotypes. Panel C summarizes the relationship between the “info” score and MAF. The secondary axis indicates the count of SNPs in each MAF bin (0.01 intervals).

A)



B)



C)

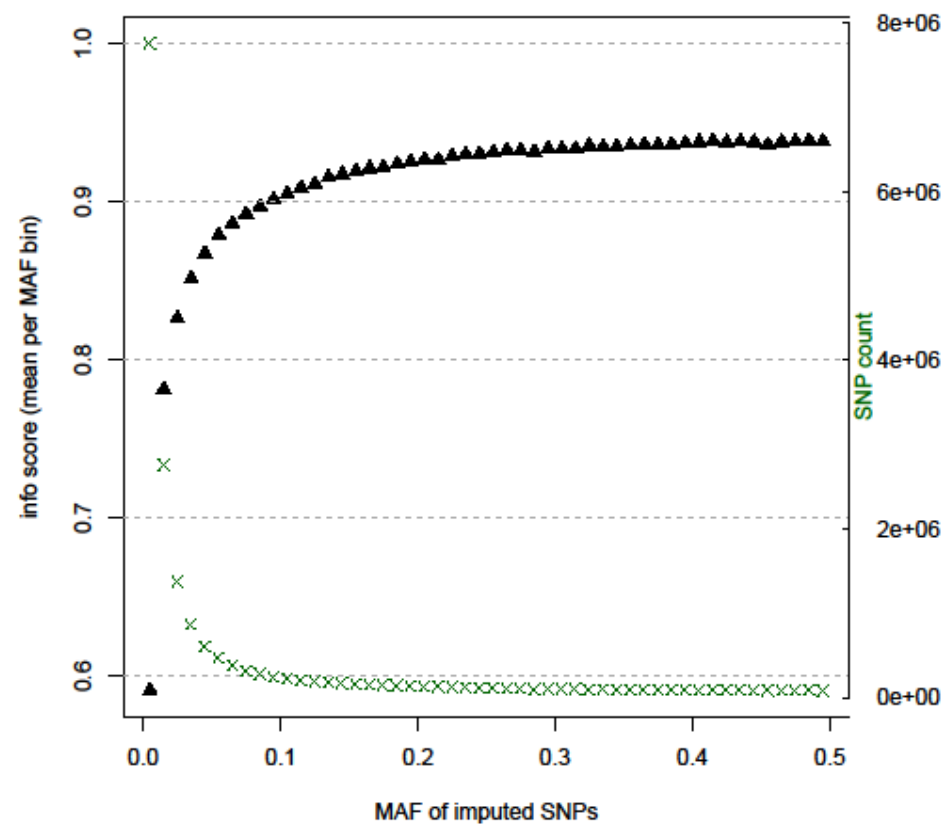
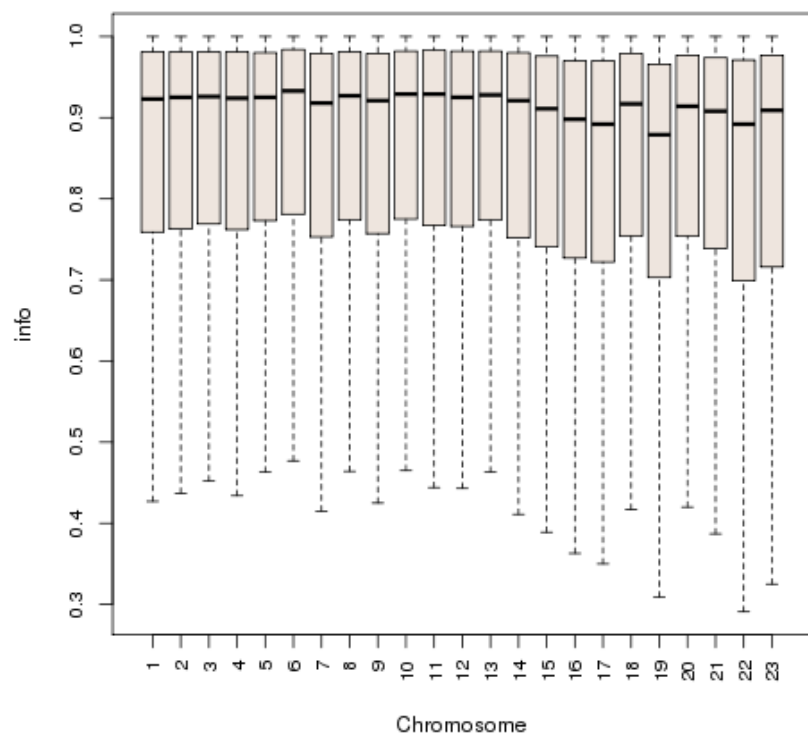


Figure 4. A comparison of imputation quality metrics by chromosome for all imputed SNPs, “info” in panel A and “certainty” in panel B. Outlier values are not displayed in these box plots. On the x-axis, “23” denotes the X chromosome.

A)



B)

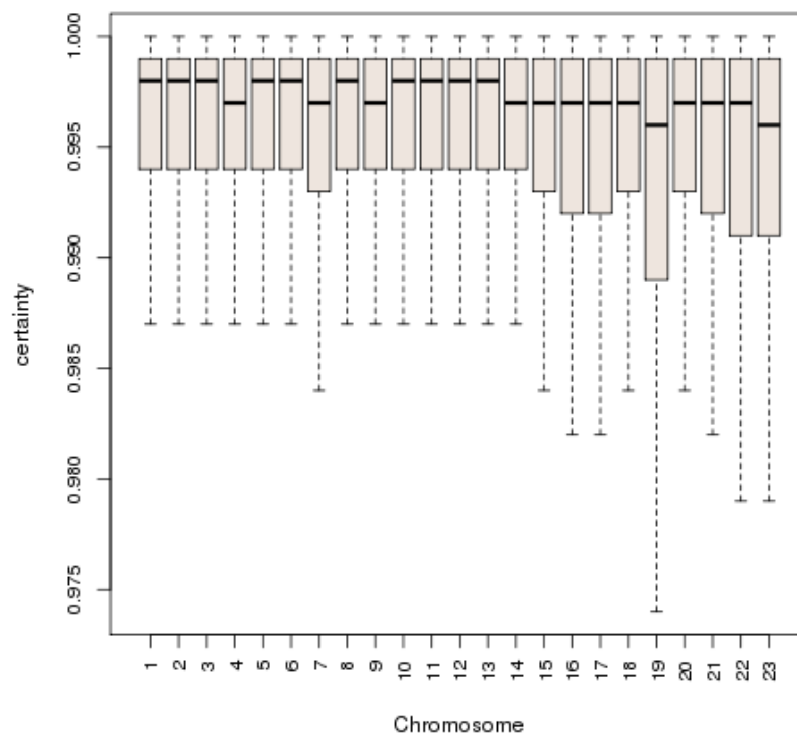
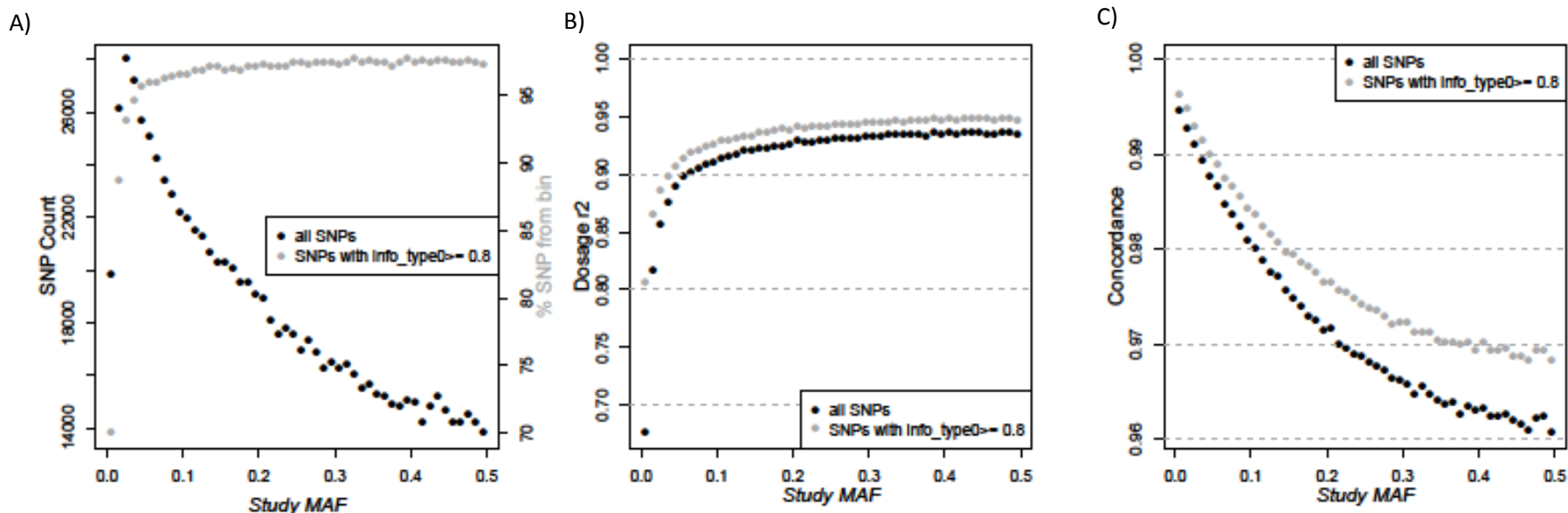


Figure 5. Quality metrics for all masked SNPs, grouped into MAF bins at 0.01 intervals. Panel (A) shows the number of SNPs per MAF bin and, on the secondary y-axis, the percentage of SNPs in the bin passing an “info” filter threshold of ≥ 0.8 . Panel (B) plots the average empirical dosage r^2 metric 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 SNPs passing CC recommended quality filters from genotype cleaning process and also mapped to build 37. This list may be used to construct a keelist for use with the PLINK `--extract` flag, to perform the initial sub setting of SNPs from the binary file (see II-c). The SNP dimension in this file corresponds to the “Study SNPs” column of the SNP Summary 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 “subjectID.” A mapping of “subjectID” to “scanID,” which corresponds to one genotype scan, is provided in the file ***subjectid2scanid.txt***. The columns in this file are:

- **family:** family identifier
- **SUBJID:** local (study investigator’s) participant level identifier18 , used in imputation output
- **subjectID:** participant level identifier assigned by the GCC
- **SAMPID:** local (study investigator’s) sample level identifier
- **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.

- **pruned.relationship:** Logical vector set to TRUE where one or both of the participant's parental IDs were set to "0" in order to carry out the pre-phasing step; otherwise, FALSE.