African American Breast Cancer Epidemiology and Risk (AMBER) Consortium Study

Imputation Report

April 28, 2014

Contents

I. Summary and recommendations for dbGaP users

Genotype imputation is the process of inferring unobserved genotypes in a study sample based on the haplotypes observed in a more densely genotyped reference sample[1,2]. The University of Washington Genetics Coordinating Center (GCC) used IMPUTE2 software[3] to perform genotype imputation in the African American Breast Cancer Epidemiology and Risk (AMBER) Consortium study. 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. For more background and a detailed description of genotype quality control (QC) on this project, please see the report available through dbGaP (database of Genotypes and Phenotypes; phs000669).

II. Study data

a. Samples

The AMBER Consortium (http://www.theamberproject.org/) is a multi-study research project aimed at detecting the causes of early-onset and aggressive breast cancer in African American women. For this genetic component of the project, a total of 7,187 study samples were put into genotyping production at the Center for Inherited Disease Research (CIDR) during the fall of 2013. After the GCC's standardized QC procedures[4], genotypes were available for 6,860 unique study participants.

We selected samples for imputation using the set of recommended quality filters generated during genotype data cleaning. First we excluded samples with a missing call rate (MCR) > 2%, resulting in the exclusion of two samples. Second the GCC also recommends excluding chromosome anomalies from further analysis, where the anomaly results in unreliable genotype calling. Thus for imputation, we excluded entire sample by chromosome combinations where the whole chromosome was recommended for filtering due to an anomaly. For anomalies affecting only partial chromosomes, we initially did not exclude any genotype data. Rather, after imputation we zeroed out imputation segments affected by a partial chromosome anomaly. Ultimately, for most chromosomes a total of 6,858 study participants were imputed.

Figure 1 shows a principal component analysis (PCA) of all unduplicated study samples and select HapMap[5] samples. All study participants self-identified as African American. In terms of genetic ancestry, the participants appear to fall on a continuum between European and African ancestry as represented by HapMap reference populations. 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 using the same worldwide reference panel.

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

b. SNPs

This study was genotyped on a custom version of the Illumina HumanExome array, with a total of 405,555 SNPs: 246,519 from the standard Human Exome and 159,036 custom. The array was designed to human genome build 37. For the purposes of imputation, study SNPs were selected using GCC recommended 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 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 release 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 variant 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.

As seen in Table 1, many observed study SNPs are not in the imputation target (i.e. are type 3) and thus not ultimately used for imputation. This is likely due to the exome and custom content on the array, many of which may not have been observed in the 1000 Genomes project or were observed but not at the requisite frequency (discussed further in section III below). The count of type 3 SNPs is the difference between the “Study SNPs” and “Imputation basis” columns: 369,407-216,256=153,151 “type 3” SNPs, or ~41% of the study SNPs.

c. Data formatting

The study genotype data were initially accessed from the binary PLINK[6] file available in the dbGaP release, “Ambrosone_BreastCancer_TOP_subject_level,” with genotypes expressed in TOP alleles. When extracting data from this PLINK dataset, we (1) subset out by chromosome; (2) extracted only study SNPs passing the quality filter; and (3) specified a list of SNPs that required a strand flip to align with the “+” strand (discussed further in section IV). In addition, for chromosomes where one or more samples had a whole chromosomal anomaly recommended for filtering, a list of exclusion samples was passed to PLINK such that those anomaly-affected samples would not be included in that chromosomes' imputation (two samples on chromosome 2 and one sample on each of the following chromosomes: 7,13,18, and 21).

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

```
plink --bfile Ambrosone_BreastCancer_TOP_subject_level \
--extract snp.qualfilter.txt --flip fliplist.txt \
--keep sampkeep.txt --chr # --make-bed \
-–out Ambrosone_chr#
```

d. Pre-phasing

Historically, phasing and imputation were 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[7]. 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.

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. 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 run SHAPEIT2[8] pre-phasing software, which explicitly uses parent-offspring relationships. Other types of relationships (e.g., full and half sibships) also improve phasing, in that SHAPEIT2 can recognize long stretches of shared chromosome segments even when the exact familial relationships are not "known" by the software (B. Howie, OXSTATGEN electronic mailing list, Jan. 6, 2014).

We used a pre-phasing approach in this imputation, due to the advantages enumerated above. While there were no expected relatives in the study, genotype data cleaning revealed several relative pairs: 30 parent-offspring, 45 full sibling, and 71 second degree pairs. However, because of the second-degree relationships, pedigrees could not be fully specified. Thus we did not provide any explicitly annotated relationships when running SHAPEIT2 (i.e., in the input PLINK files, all parent identifiers were left as 'NA,' or missing). Rather, we relied on SHAPEIT2's ability to detect relatedness when not explicitly specified by the user.

The input files for SHAPEIT2 were the filtered, chromosome-specific PLINK files described above (see section II-c). In turn, SHAPEIT2 output best-guess haplotypes which were then fed directly into the IMPUTE2 imputation. SHAPEIT2 jobs were run multi-threaded across 12 compute cores. Runtimes ranged from 1 to 10 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 "#." The "`-S 200`" flag sets the number of conditioning states; "`-T 12`" is the flag for multi-threading across six cores.

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

III. Reference panel

Larger reference panels have been shown to increase imputation accuracy[2,9,10]. Previously, haplotypes from Phases 2[11] and 3[5] 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 populations[13]." 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 populations[14]. The Project has categorized each of these populations into four continental groupings: African (AFR), Americas (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 AMBER study participants, we used a worldwide reference panel of all 1,092 samples from the phase I integrated variant set, which is based on both low coverage whole genome and deep coverage exome sequence data.

We downloaded these reference panel data from a recent release on the IMPUTE2 website (see Web Resources). These data files were created by IMPUTE2 authors from the variant call format (VCF) files available from the Project. The authors used a two-step phasing process in which they first phased separately available SNP array data on these same samples to produce a haplotype "scaffold." Second, they used SHAPEIT2 to phase the sequencing data, using the SNP array scaffold from the previous step. This two-step phasing algorithm using SHAPEIT2 has been shown to yield higher quality imputation compared to the previous version of the IMPUTE2 reference data[15]. (Note currently only the autosomes are available in this new format; thus we used the previous version for the X chromosome only.)

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[10]. 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 38 million variants). Very low MAF variants are both harder to impute and, even if imputed error-free, it is unlikely most studies will be sufficiently powered to detect an association in downstream analyses. There is also the concern that a variant observed only once may reflect a sequencing error rather than a true variant. Therefore, we restricted imputation to variants with at least two copies of the minor allele in either the AFR or EUR 1000 Genomes continental groups. We included all three variant types (SNPs, indels, and SVs) in this imputation, based on findings[14] 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[16]. The initial study dataset contained TOP alleles, an Illumina naming method unrelated to "+" or "-" strand orientation[17] (also see Web Resources, Illumina 2006). Because all 1000 Genomes reference panel data are expected to be "+" strand relative to the reference, we first needed to convert TOP alleles to the "+" strand. Typically Illumina annotation contains the information necessary to perform this conversion. However, because this study used a custom array, some of the usual Illumina fields were missing from the array annotation (specifically the "RefStrand" field, which indicates whether the design alleles are on the "+" or "-" strand).

The GCC took the following steps to obtain allele mappings for the Ambrosone custom array. First, for the variants from the standard HumanExome array (n=246,519), the Illumina annotation file "HumanExome-12v1_A.csv" was used to obtain the mappings. Second, for the custom variants (n=159,036), a two-tiered bioinformatics approach was used. Initially UCSC BLAT searches (see http://genome.ucsc.edu/cgi-bin/hgBlat) were performed on the design sequences to determine the correspondence between design alleles and "+"/"-" strand. Where BLAT searches failed to yield a suitable match (n=201 variants), the UCSC Table Browser ("snp138" table, see http://genome.ucsc.edu/cgi-bin/hgTables) was used to determine "+" strand alleles. Plus strand mappings were ultimately obtained for all but 13 SNPs (strand ambiguous) and 28 insertion/deletion variants.

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. A total of 30 SNPs were flipped in this IMPUTE2 check, or < 0.008% of all input SNPs, suggesting that our initial attempt to orient to all "+" strand alleles was largely successful.

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).

a. Imputation segments

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

Ultimately we divided 23 chromosomes into 545 total segments, ranging from 7 segments on chromosomes 21 and 22 to 47 segments on chromosome 2. Using this study’s custom array, we calculated a mean of 678 study SNPs (type 2 + type 3) per segment (range 53-3,834; interquartile range 297-846). This suggested that imputation basis SNPs (type 2) would be adequately represented in our imputation segments. Imputing from an exome-based array will necessarily yield a lower density of observed SNPs per imputation segment compared to denser arrays, especially when those denser arrays have selected SNPs to tag regions genome-wide (i.e., not based on gene locations). However, our segmentation scheme yielded no segments without any observed SNPs and only 15 segments with less than 100 observed SNPs, which is typically our desired minimum threshold when checking study SNP density on proposed imputation segments. Furthermore, the IMPUTE2 authors recommend imputing with 5

MB or shorter segments, suggesting that combining segments to increase observed SNP density could have had negative consequences for imputation quality.

b. IMPUTE2 settings

The IMPUTE2 algorithm uses a “k_hap” value to specify which number of reference haplotypes should be used to impute each study sample. The implementation of this parameter is one of the ways imputation with a worldwide reference panel is made computationally feasible: i.e., the full set of reference samples is available to the imputation software, but it “chooses” a subset of reference haplotypes to impute each study sample based on perceived genetic similarity (for details, see Howie, Marchini, and Stephens, 2011[10]). The default k_hap value is 500; however, higher values are recommended when imputing into admixed populations such as African Americans. Thus for this project we set k_hap to 1,000.

Another IMPUTE2 default we altered was buffer size. 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.

c. Running IMPUTE2

An example of the command line syntax used to run IMPUTE2 on a chromosome 22 is shown below. Note the inclusion of the “`-os 0 2`” option, which specifies that only variants 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. The `-h` and `-l` flags refer to the 1000 Genomes reference panel files (the haplotypes and variant legend files, respectively).

```
impute2 -use_prephased_g -m genetic_map_chr22_combined_b37.txt \
-h ALL.chr22.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.haplotypes.gz \
-l ALL.chr22.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.legend.gz \
-int 20000000 25000000 -buffer 500 -allow_large_regions \
-known_haps_g Ambrosone_chr22.haps.gz \
-k_hap 1000 \
-filt_rules_l ma.cnt.gte2.afr.eur<1 \
-o Ambrosone_chr22.set2.gprobs -os 0 2 -o_gz \
-i Ambrosone_chr22.set2.metrics –verbose
```

Imputation jobs were run in parallel on a compute cluster consisting of 16 compute nodes, each containing two Intel Xeon E5-2630Lv2 Six-Core processors (15 MB cache), 128 GB of memory, and 2 TB of local storage. Due to the input of pre-phased haplotypes, the compute time required to impute most segments was one to three hours. In total, the imputation took approximately two days 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 at two levels: (1) by chromosome, where “23” denotes chromosome X; and (2) by participant consent level. Note that participants were imputed together, independent of consent level; raw output files were split into consent groups after the imputation analysis was completed.

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 “SUBJECT_ID” 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 (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 chromosomal anomaly that was recommended for filtering. 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., -1 -1 -1). Note that where the whole chromosome was recommended for filtering due to a chromosome anomaly, that individual was excluded from imputation on the given chromosome (see section II-c). After imputation segments were processed for anomalies, they were combined into per-chromosome .gprobs file, via the Unix ‘cat’ command.

The first five columns in these output files correspond to SNP ID; rs ID; physical position; and the two alleles, where the first allele shown is designated “allele A” and the second is designated “allele B.” (Note that allele A is the reference allele from the 1000 Genomes Project and thus bears no consistent relationship to the minor allele in IMPUTE2 output).

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:variant type"* is used, where "variant type" is either S (SNP), D (deletion is alternate allele), or I (insertion is alternate allele). 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 B" (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[9] and MACH[18]. (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 variant 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). The average "info" scores at SNPs with MAF < 0.05 fall below 0.6, while the remaining variants (those with MAF > 0.05) have average "info" scores between 0.6 and 0.65. The "info" score profile is similar for indels and SVs (panel B). We also plotted these metrics by chromosome, shown in Figure 4. No chromosomes appear to be extreme outliers, although the X chromosome appears to have slightly lower quality compared to the autosomes.

The imputation quality (e.g., average "info" scores by MAF bin and by chromosome) is notably lower than in studies using arrays designed for genome-wide coverage. We suspect this is due to two aspects of the imputation basis: both SNP density and SNP selection. Regarding density, the custom version of the HumanExome array used for this project contains just over 400K SNPs compared to the 1M or 2.5M SNP arrays that form our more typical imputation basis. Regarding SNP selection, the standard content of the array for this project was designed to focus on exome variants and therefore is likely to have relatively poor coverage of intergenic and nongenic regions of the genome. Furthermore, the array's standard content SNPs were not selected for their tagging properties, although some of the custom content was chosen to tag regions of interest. Because many of the exonic SNPs are rare, they are not observed in 1000 Genomes (or at least not observed with >2 minor alleles in AFR or EUR) and are thus ultimately not part of the imputation basis (i.e. are "study only" - or type 3 - SNPs). In fact, there were only 216,256 type 2 SNPs in this imputation basis, which is a significant reduction from the initial set of ~369 K eligible SNPs (i.e. those passing the quality filter from genotype data cleaning and mapped to chromosomes 1-22 and non-pseudoautosomal chromosome X – see Table 1).

Overall, the lower imputation quality profile we are observing for this study is likely explained by the relatively low number of imputation basis SNPs combined with the fact that they are mostly not selected to be tagging other SNPs. However, to investigate this empirically we re-made the plot shown in Figure 3 - showing average info scores across the MAF spectrum - separately for two different subsets of imputed variants: (1) all exonic variants and (2) all variants within 60 kb of a custom variant (i.e. from the 159,036 variants added as custom content to the standard exome array). We selected these subsets of variants based on the expectation that given the imputation basis, imputation quality in these regions should be higher compared to the genome-wide averages. To identify exonic variants, we used the R package *rtracklayer*[19] to interface with the UCSC Browser database (http://genome.ucsc.edu/cgi-bin/hgTable) and retrieve exon coordinates for all the RefSeq genes (hg19 reference, “RefSeq” track, “refGene” table). We then used the R package *IRanges*[20] to determine which imputed variants fell within an exon, yielding 420,852 variants (of ~24.2M total). To create the second subset of variants, we first identified all imputation basis (type 2) SNPs that were from the custom portion of the array. Note these custom variants were chosen to tag selected regions of the genome and were not limited to exons. Next we isolated those imputed variants that fell within 60 kb of one or more type 2 custom variants (again using the *IRanges* package), yielding 12,115,056 variants – roughly half all imputed variants.

Figure 5 shows the distribution of info scores across the MAF spectrum for (A) all imputed variants (similarly to Figure 3A but not dichotomized by variant type); (B) all imputed variants in exons, and (c) all imputed variants within 60 kb of a custom type 2 variant. There is a notable improvement in imputation quality in the subset of exonic variants: for MAF > 0.01, mean info is 0.7-0.75, which is ~0.1 higher compared to the genome-wide averages. The improvement in quality in the imputed subset within 60 kb of a custom variant is more modest: ~0.05 compared to the genome-wide averages. In Figure 6 we also present the distribution of MAF (calculated across imputation study samples) of all imputation basis (type 2) variants. That 39.1% of imputation basis variants have an allele frequency of ≤ 5 % is also likely contributing to the reduced imputation quality. That is, imputing from arrays with more common SNPs is likely to facilitate matching of haplotypes between study and reference samples and ultimately yield higher imputation quality. In summary, we found that imputation quality was higher in regions where we would expect it to be based on the imputation basis (see Figure 5). In addition, we confirmed that a high proportion of imputation basis variants were at MAF < 5% in study samples, suggesting another reason why imputation quality is lower than on more standard (or more frequently imputed from) arrays.

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[16]. 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[16]. 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 84.7% of all imputed variants for downstream analyses, while more stringent thresholds of 0.5 and 0.8 would retain 48.7% and 16.0% of imputed variants, respectively. However, users should be aware that setting stringent quality thresholds has been shown to result in missing true positive associations[21]. We also calculated the percentage of variants passing these various info level filters in the two subsets of imputed variants considered above: all exonic and those within 60 kb of a custom variant. A comparison of these percentages is shown in Table 3.

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

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 216,256 masked SNPs (i.e. all type 2 SNPs).

Figure 7 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. First, there is a sharp decline in empirical dosage r^2 for low-frequency variants (MAF < 0.05). As MAF increases, however, average correlation values increase to ~0.8. There is a slight dip again at MAF > 0.40, where mean dosage r^2 is between 0.7 and 0.8. The differences between unfiltered (black points) and filtered (gray points) data series demonstrate the utility of filtering by the “info” quality metric, which is available for all imputed variants. This filtering improves the quality metrics profile for masked SNPs across the entire range of MAF bins and notably removes the dip in dosage r^2 at the very common variants (MAF > 0.40). Thirdly, Figure 7C 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[23]. 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 4 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 we have done here. The former way of calculating this metric often leads to higher mean concordance, especially when imputed genotypes are filtered on maximum probability. There are also advantages and disadvantages to both metrics: dosage r^2 and concordance. The advantages of dosage r^2 include (1) precedence in the literature for evaluating imputation accuracy[7,8,24] ; (2) less sensitivity to allele frequency than concordance; (3) similarity to information metrics commonly reported by imputation software (for a review, see Marchini and Howie, 2010); and (4) incorporation of imputation uncertainty by using expected allelic dosage rather than most likely genotype. However, one important downside is that r^2 has high variance at low MAF[25]. Concordance is advantageous as a widely used metric that is easily interpretable; however, it ignores imputation uncertainty and is very sensitive to allele frequency, as low MAF variants may yield high concordance purely by chance[23].

Lastly, when discussing imputation quality there can be several different meanings of “efficiency.” Figure 7A illustrates one definition: the percentage of imputed variants passing a given quality filter (“info” ≥ 0.8, e.g.). This metric is > 70% 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[1,2,12,16,26], including family-based analysis[27]. 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[28], an R package developed by the GCC; PLINK (with the --dosage option: http://pngu.mgh.harvard.edu/~purcell/plink/dosage.shtml); MACH2qtl/dat[18]; SNPTEST[29]; ProbABEL[30]; BIMBAM[31]; SNPMStat[32]; and the R package snpMatrix[33]. For a comparison of methods to account for genotype uncertainty in imputed data, see Zheng et al[34].

VII. Summary

We have performed genotype imputation in the AMBER Consortium project, 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 GCC at the University of Washington in Seattle, WA. This report was reviewed and approved by study investigators Kathryn Lunetta and Steve Haddad at Boston University and by representatives of the CIDR genotyping center.

VIII. References

1. Browning, S. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* **124**, 439-50 (2008).
2. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
3. Howie, B., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
4. 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. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
9. 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).
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. Delaneau, O., Marchini, J. & Consortium, G.P. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *(in review)* (2013).
16. de Bakker, P. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
17. 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).
18. 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).
19. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-2 (2009).
20. Pages, H., Aboyoun, P. & Lawrence, M. IRanges: Infrastructure for manipulating intervals on sequences. R package version 1.20.7.
21. 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).

22. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet* **125**, 163-71 (2009).
23. Lin, P. *et al.* A new statistic to evaluate imputation reliability. *PLoS One* **5**, e9697 (2010).
24. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
25. Evangelou, E. & Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379-89 (2013).
26. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet* **4**, e1000279 (2008).
27. Shi, M. *et al.* Using imputed genotypes for relative risk estimation in case-parent studies. *American journal of epidemiology* **173**, 553-9 (2011).
28. 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).
29. 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).
30. 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).
31. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).
32. 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).
33. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum Hered* **64**, 45-51 (2007).
34. 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_SHAPEIT2_9-12-13.html on Dec. 10th, 2013.

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/.

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	32,361	16,775	1,848,647
2	29,463	18,311	2,009,999
3	22,620	13,306	1,687,304
4	14,865	8,593	1,700,513
5	19,542	12,278	1,550,065
6	22,772	14,618	1,514,323
7	16,557	9,759	1,378,341
8	15,520	9,834	1,339,751
9	16,763	10,289	1,021,748
10	20,358	13,839	1,171,955
11	22,235	12,336	1,169,602
12	19,520	11,561	1,132,099
13	7,308	4,649	852,129
14	13,548	8,459	777,070
15	10,585	5,401	697,975
16	17,081	10,416	743,530
17	16,080	7,777	647,402
18	7,183	4,875	675,032
19	18,633	9,017	524,072
20	8,593	4,668	531,416
21	4,271	2,547	327,532
22	7,398	4,113	316,584
X	6,151	2,835	797,138
Totals	***369,407***	***216,256***	***24,414,227***

[†] Study SNPs passing pre-imputation filters (IMPUTE2 variants types 2 and 3).

[††] Study SNPs passing pre-imputation filters and overlapping with imputation target variants in 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 ≥ 2 copies of minor allele in either the AFR or EUR 1000 Genomes panel.

Table 2. An overview of the 1,092 samples in the 1000 Genomes Project phase 1 worldwide reference panel, which was used to impute all study participants. The Project assigned each population to one of four continental groupings: African (AFR), Americas (AMR), Asian (ASN), and European (EUR)[12]. 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 Americas 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. Percentage of imputed variants passing various info filter thresholds. The latter two rows demonstrate the improvement in imputation quality when considering subsets of variants expected to have higher quality based on imputation basis (i.e. exome + custom variants).

Set	Number of imputed variants	Percentage (%) passing info thresholds of:		
		0.3	0.5	0.8
All imputed	24,197,971	84.7	48.7	16.0
Exonic	420,852	88.1	62.2	27.8
Within 60 kb of a custom type 2 (imputation basis) SNP	12,115,056	87.9	56.8	22.3

Table 4. 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	83,910	0.988 (0.995)	0.498 (0.528)
≥ 0.05	132,346	0.917 (0.967)	0.798 (0.933)

XI. Figures

Figure 1. Principal component analysis of 70 HapMap samples and 6,829 unduplicated study samples. On the left, HapMap samples are color-coded by population: Utah residents (CEPH) with Northern and Western European ancestry (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); and Yoruba in Ibadan, Nigeria (YRI). On the right, study samples are color-coded according to self-designation as Hispanic or non-Hispanic. Note all study samples are self-identified African American. The percent variance explained by each of these first two components is noted on the axis labels. (Also Figure 10 from the genotype QC report.)

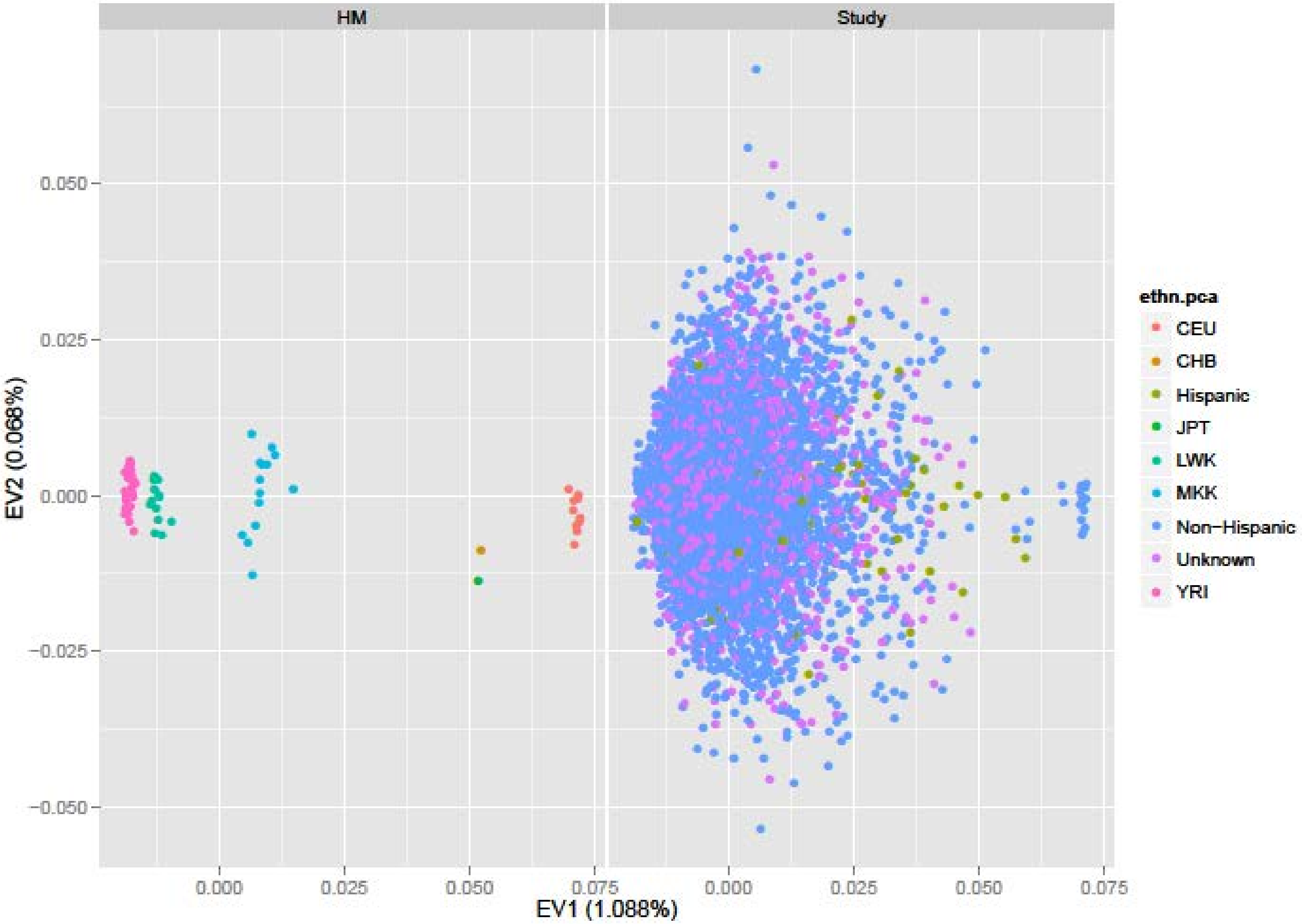


Figure 2. A schematic of variant types as defined in the IMPUTE2 imputation algorithm. Each individual is represented by a unique color in the horizontal bar(s), and alternate alleles at each variant are represented as *A* and *B*. Section (A) represents phased reference haplotypes, where two samples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

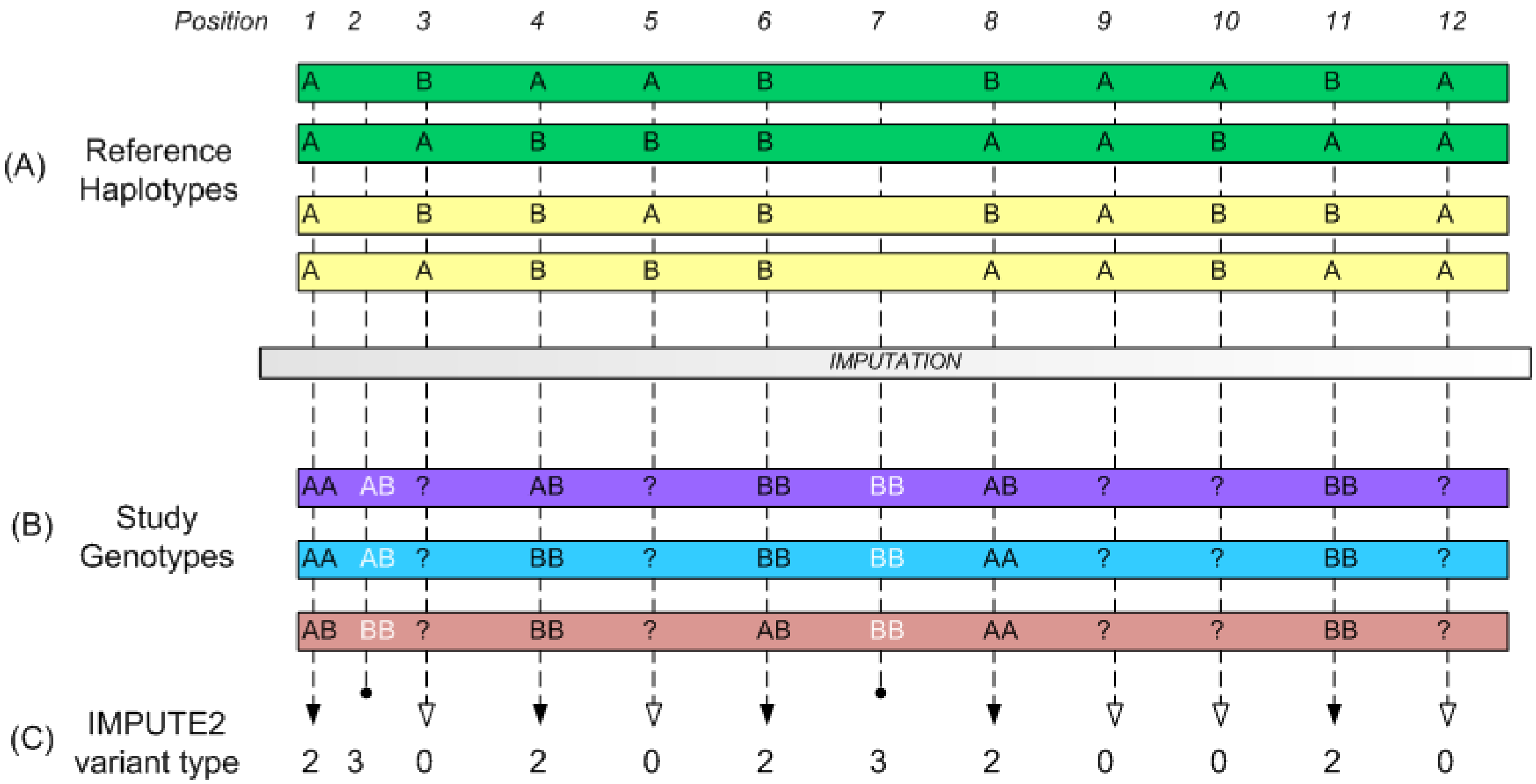
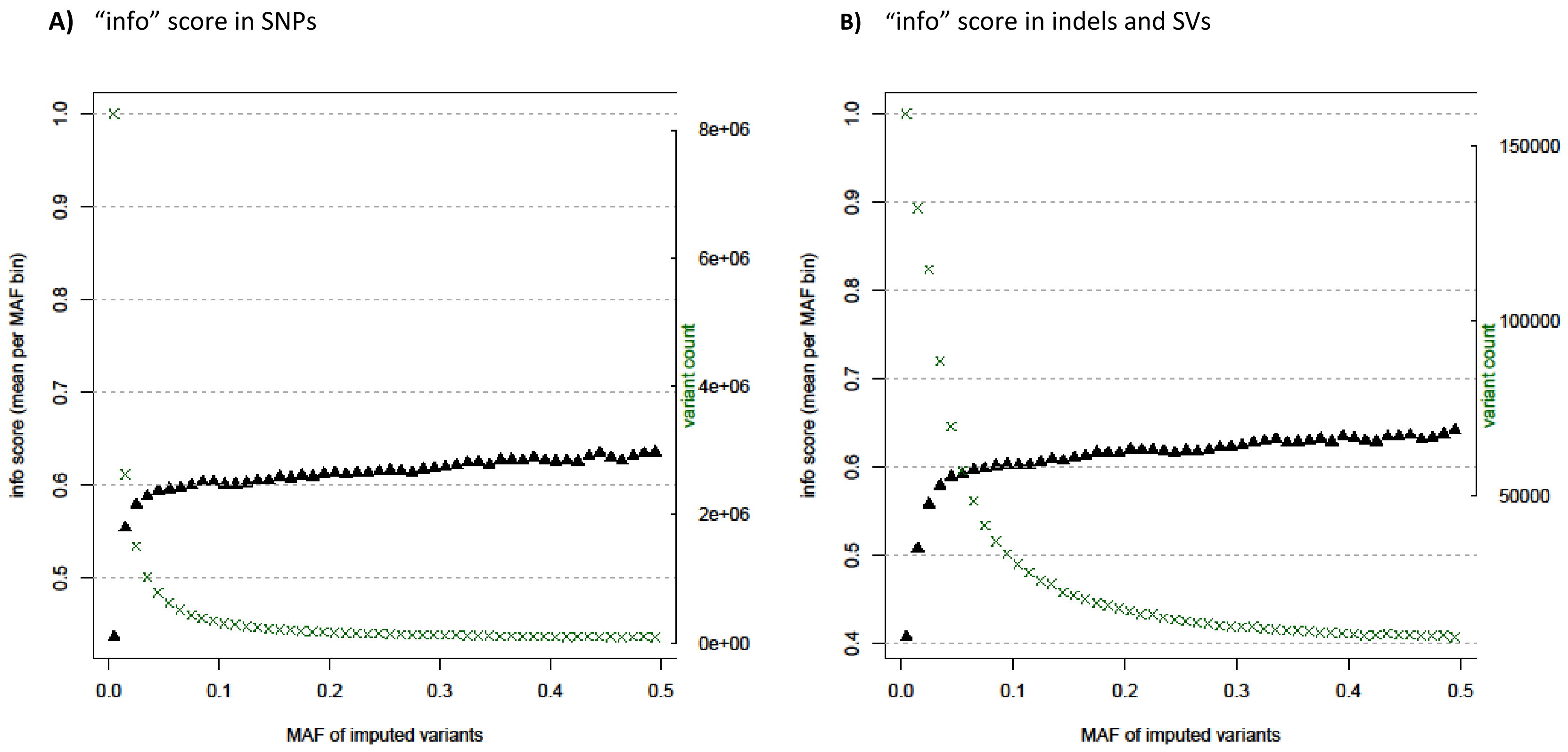


Figure 3. Summaries of quality metrics at all imputed variants. In each plot, imputed variants are binned by MAF (0.01 intervals) along the x-axis and then the mean “info” metric per bin is plotted on the y-axis, in panel (A) for SNPs and in panel (B) for indels and SVs. In each panel, the secondary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



A) “info” score in SNPs

B) “info” score in indels and SVs

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

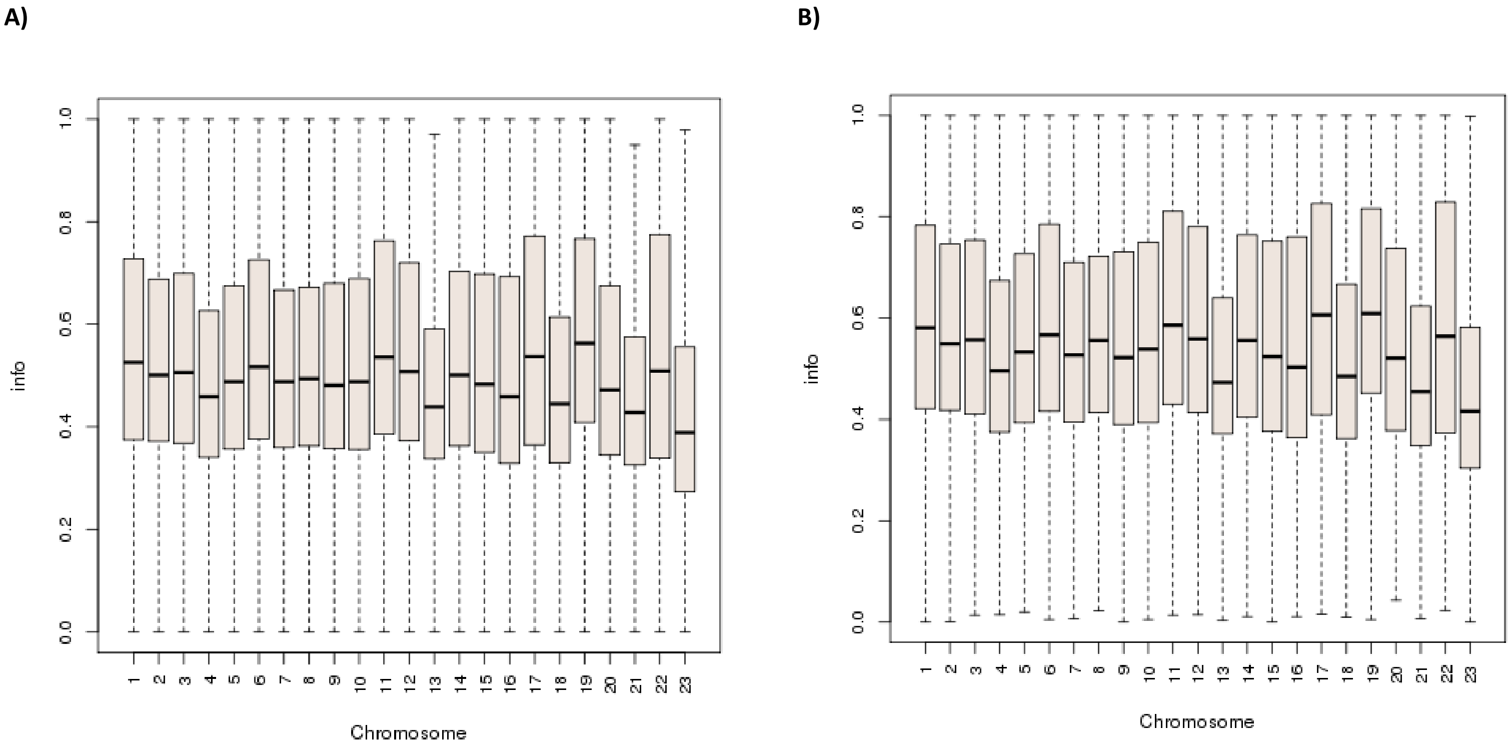


Figure 5. Comparing summaries of quality metrics at (A) all imputed variants, (B) all imputed variants in exons, and (c) all imputed variants within 60 kb of a custom type 2 variant. In each plot, imputed variants are binned by MAF (0.01 intervals) along the x-axis and then the mean "info" metric per bin is plotted on the y-axis; the secondary y-axes indicate the count of variants in each MAF bin.

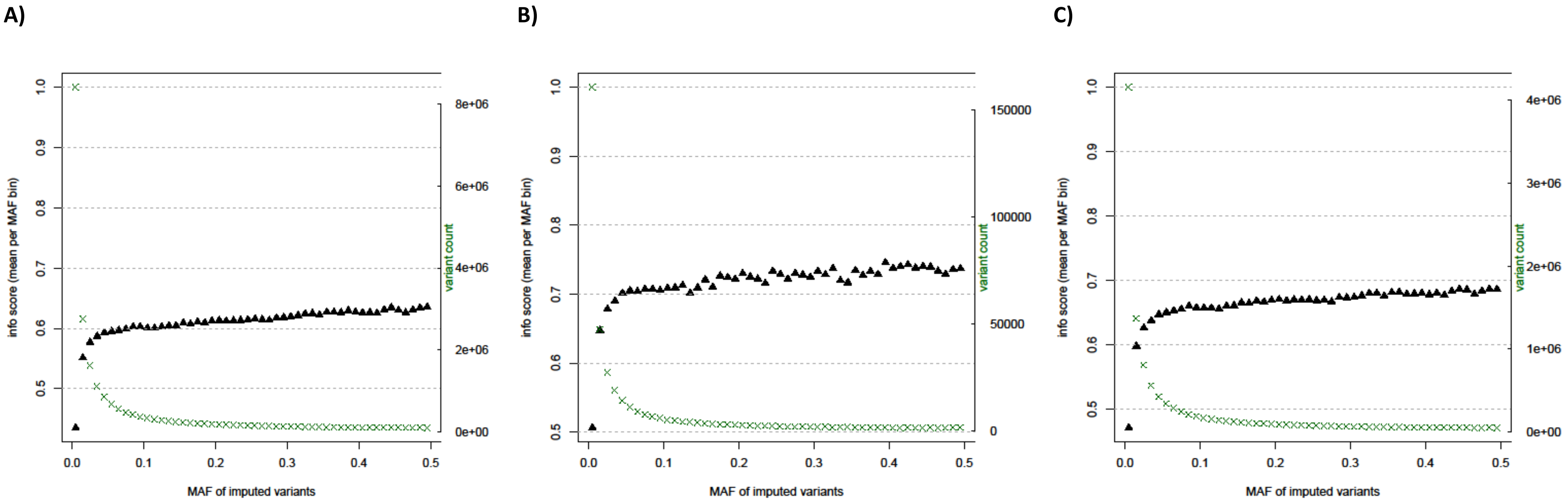


Figure 6. Histogram of MAF in all imputation basis (type 2) variants, where MAF was calculated in all imputed study samples (n=6,858).

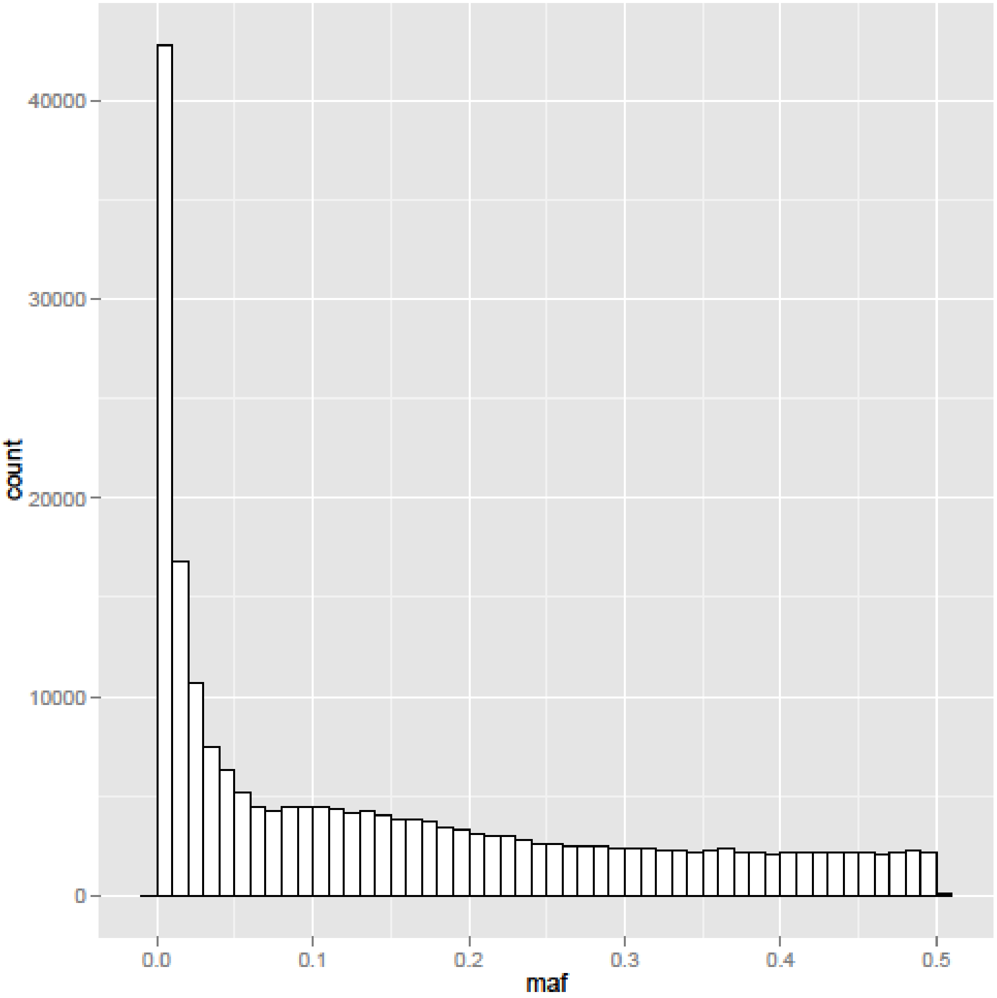
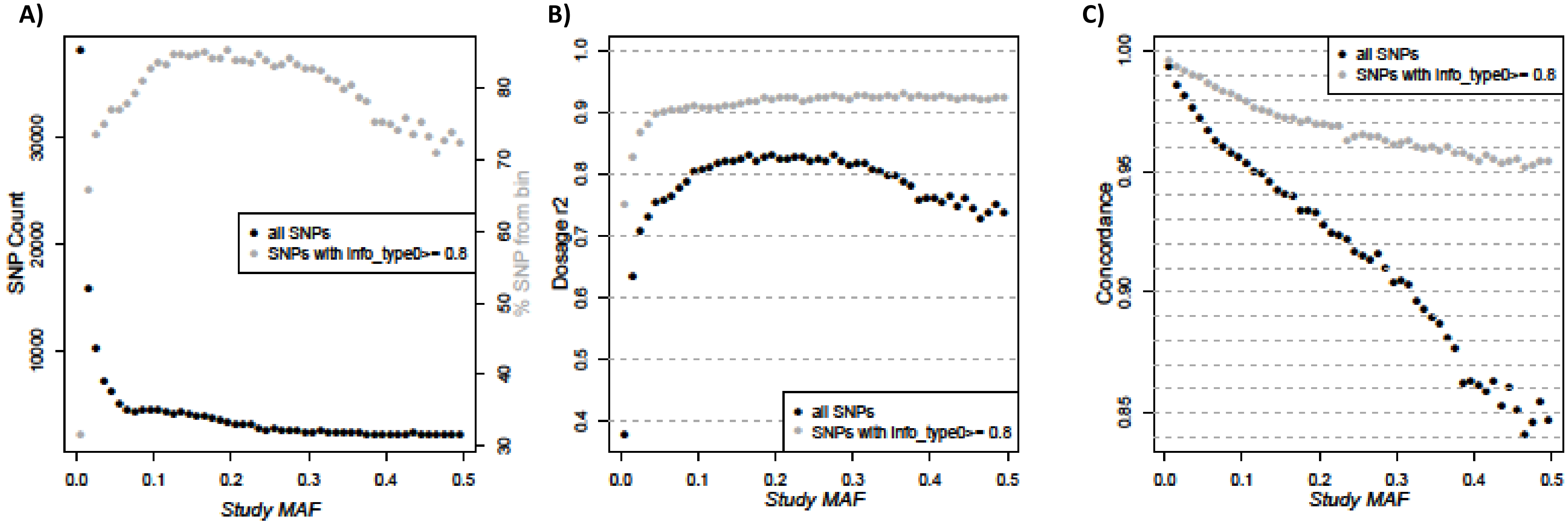


Figure 7. Quality metrics for all masked SNPs, grouped into MAF bins at 0.01 intervals. Panel (A) shows the number of SNPs per MAF bin and, on the secondary y-axis, the percentage of SNPs in the bin passing an “info” filter threshold of ≥ 0.8. Panel (B) plots the average empirical dosage r^2 metric per 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 (545 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. -1 -1 -1). The fields in this file are:
    - **subjectID**: participant level identifier assigned by the GCC, 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 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 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

b. **SNP flip list**.

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

c. **Sample-subject mapping.** The identifier used in the imputation output is the "SUBJECT_ID." A mapping of "SUBJECT_ID" to "scanID," which corresponds to one genotype scan, is provided in the file "subjectid2scanid.txt." Note the first two columns in this file are equivalent to the keep list "sampkeep.txt" used to extract imputation samples from the subject level PLINK file (see section II-c). The columns in this file are:

- **family:** family identifier
- **SUBJECT_ID**: participant level identifer, used in imputation output
- **scanID**: scan level identifier, corresponding to one genotyping instance

- **whole.chrom.anomaly**: for samples excluded from a given chromosome's imputation due to whole chromosome anomalies, this field contains the filtered chromosome(s) integer codes. Otherwise, NA
- **consent**: participant consent value, where 1=Cancer Research and Methods; 2=Breast cancer only, collaborative agreement.