EPOCH imputation | Technical report and summary

Abstract

As part of the EPOCH study protocol, blood samples were collected from 336 multi-ethnic study participants for subsequent genetic analyses. DNA samples were genotyped using the Infinium Omni2.5-8 BeadChip. This report summarizes the procedures implemented for genotype imputation of untyped autosomal loci.

Imputation has been successfully completed for a subset of genetic variants and EPOCH participants that met quality control standards outlined in this document. The data set used as the basis for the completed imputation included all but three participants with high rates of missing genotypes. Further discussion is warranted to determine whether the data should be re-imputed with the exclusion of any other participants. Issues that may warrant exclusion of additional participants from re-imputation include unexpectedly low or high rates of heterozygosity, discrepancies between self-reported and genetic sex, or evidence of being an outlier with respect to genetic ancestry.

Introduction

Genotype imputation refers to the statistical prediction of genotypes that were not directly measured. Genotype imputation has been shown to significantly improve statistical power for genome-wide association analyses.

Imputation has been made possible by the development of robust genomic reference panels, most notably the 1000 Genomes project and the International HapMap Project. Such reference panels have been developed by performing whole genome sequencing of large, ancestry-representative samples. Imputation algorithms predict genotypes at untyped loci using the genotypes observed at observed loci and the patterns of sequence variation observed in an appropriate reference panel.

Imputation algorithms estimate the probability of each possible allelic and then output the expected number of minor alleles at a given locus within an individual. For directly genotyped loci, the minor allele count is confined to integer values. Because the expected number of minor alleles is derived from a probability distribution, imputed genotypes exist on a continuous scale which indicates the most likely genotype at a given locus.

The example below illustrates observed and imputed genotypes at six loci within an individual. Loci 1, 3, and 5 were directly genotyped by microarray. Genotypes at loci 2, 4, and 6 were imputed. In this example, the individual is most likely homozygous major at locus 2, heterozygous at locus 6, and homozygous minor at locus 4.

Figure 1. Example of imputed genotype data

	Locus	**1**	**2**	**3**	**4**	**5**	**6**
Minor allele count (N)	**Observed**	2	-	0	-	1	-
	Imputed	2	0.10	0	1.89	1	1.10

Raw genotype data must be thoroughly quality controlled prior to imputation. The quality of the data input for imputation directly impacts the accuracy of imputed genotypes. Imputation accuracy is a significant determinant of statistical power for downstream genome-wide association analyses.

Data quality assessment

A series of quality control statistics were produced to assess the suitability of each variant and individual for imputation. Following each quality control step, variants and individuals with potential issues warranting exclusion from imputation were flagged for further review.

Variant validity

Variants present in the EPOCH GWAS data were compared to variants represented on the most up-to-date version of the Omni 2.5 Beadchip. Variants that were represented in both the EPOCH data and the current Omni 2.5 array were then compared to assess whether the mapping of each variant had changed. Variants were retained for imputation if they were present in both sources and if they mapped to the same chromosome and position in each.

The raw EPOCH GWAS data set contained 2,399,785 variants, spanning the 22 autosomes, the sex chromosomes, and mitochondrial DNA. 97.8% of variants represented in the EPOCH GWAS data were represented on the most up-to-date version of the Omni 2.5 Beadchip and were mapped to the same chromosome and position in each source.

Table 1. Variants removed during variant validation

Reason for removal	***Variants removed (N)***
Absent from the Omni 2.5-8 v1.4 Beadchip	51,714
Mapped to a different chromosome in the Omni 2.5-8 v1.4 Beadchip	1,156
TOTAL	52,870

Genotype missingness

Background

Missingness is an important indicator of the quality of genotypes measured by microarray. Call rate is the complement of missingness, representing the percentage of non-missing genotypes per individual or per variant. For example, a missingness rate of 5% corresponds to a 95% call rate.

The missingness rate per variant is calculated as the percentage of sampled individuals for whom no genotype could be measured. A high individual call rate indicates that the microarray was able to successfully measure genotypes from a given individual's DNA sample. A low individual call rate may indicate that the sample was compromised in some way prior to genotyping.

The missingness rate per individual is calculated as the percentage of microarray variants for which no genotype could be measured. A high variant call rate indicates that the microarray was able to capture the genotypes of study individuals with a high degree of success. A low variant call rate indicates that the microarray was unable to consistently measure the genotype for a given variant.

Variant and individual missingness can be non-random in nature. For example, heterozygous genotypes are sometimes more difficult to capture than homozygous genotypes. Non-random genotype missingness therefore presents a significant risk for confounding in genome-wide association analyses. Filtering variants and individuals to minimize missingness helps ensure the validity of results obtained in downstream analyses.

EPOCH

Genotype missingness per variant was evaluated within a subset of individuals with less than 5% missing genotypes (call rate>95%).

Missingness per individual was calculated within the same subset of individuals with less than 5% missing genotypes. The rate of missing genotypes per individual was calculated across all variants, and separately across common variants only (minor allele frequency greater than or equal to 5% (MAF>=5%).

Individual heterozygosity and sex discrepancy

Background

Heterozygosity refers to the proportion of an individual's genotypes that are heterozygous. For a biallelic variant, a heterozygous genotype includes one copy of the common allele and one copy of the minor allele. Unexpectedly high or low rates of heterozygosity can result from sample contamination or consanguinity, which may compromise the accuracy of imputation.

Sex discrepancy refers to a mismatch between the reported sex of an individual and their sex as predicted by X chromosome heterozygosity. Discrepancies between reported sex and imputed sex can indicate sample mix-ups.

Sample problems such as heterozygosity and sex discrepancy can reduce imputation accuracy. Samples with heterozygosity or sex discrepancy issues must be excluded from imputation to vent the loss of statistical power in downstream association analyses of imputed genotypes.

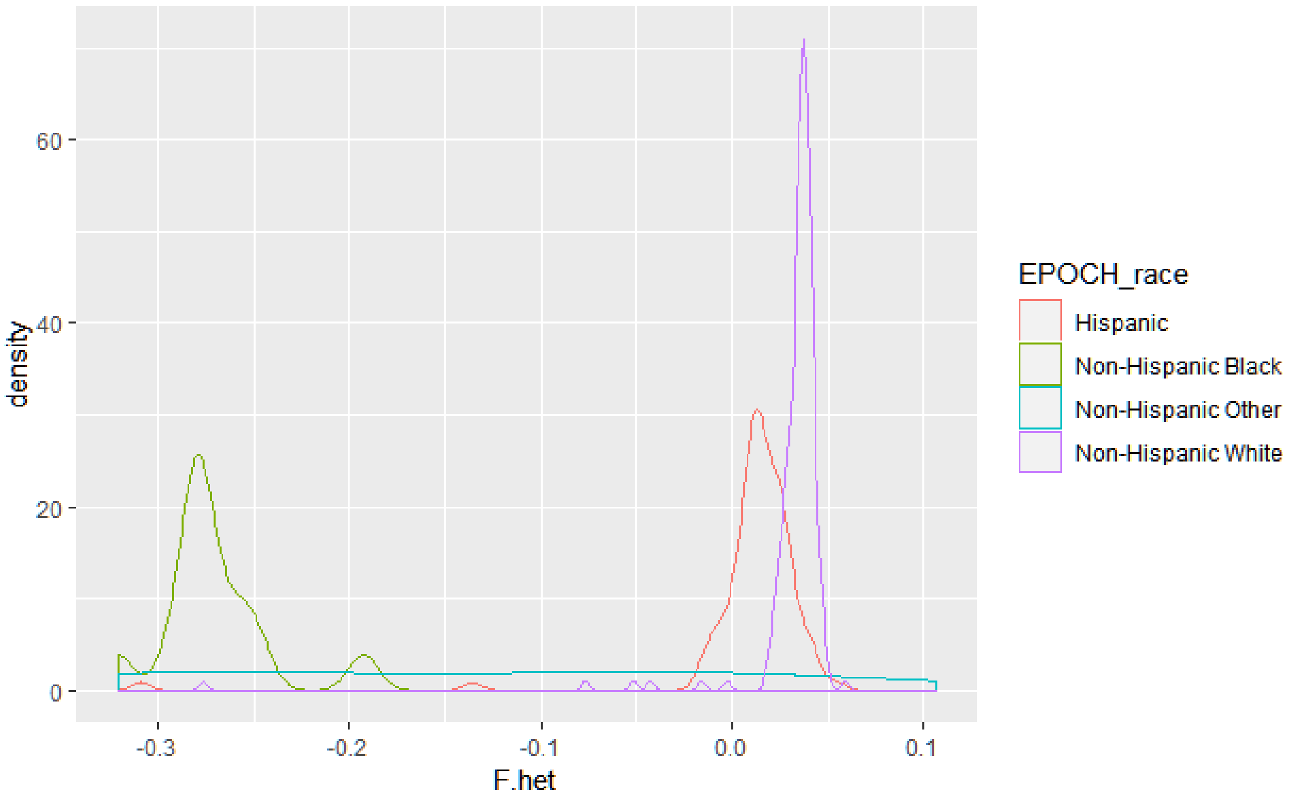
EPOCH

Heterozygosity and sex mismatch statistics were calculated for individuals with less than 5% missing genotypes, using variants with less than 2% missing genotypes. Imputed sex as predicted by X chromosome heterozygosity was compared to participant sex as reported in EPOCH clinical data.

Due to the multi-ethnic nature of the EPOCH cohort and small sample sizes of subgroups, preliminary findings regarding heterozygosity problems and sex discrepancies in the EPOCH sample were inconclusive. Heterozygosity and sex discrepancy statistics, as calculated by Plink 1.9, depend on the expected minor allele frequency of each variant. Unless external allele frequencies are provided, the expected minor allele frequency is calculated based on the alleles observed in the sample. This method has proved insufficient for quality control of EPOCH’s minority participants, as the allele frequencies found in non-Hispanic white populations are often dramatically different than those found in minority populations.

Heterozygosity and sex discrepancy statistics for the EPOCH cohort are currently being re-calculated with reference to the minor allele frequencies observed in the 1000 Genomes project. By using the allele frequencies found in the 1000 Genomes, the calculations will no longer be biased by ethnicity. Updated results will be provided as soon as possible. Unbiased results will then be used to definitively conclude whether any EPOCH participants should be excluded from genotype imputation due to heterozygosity or sex discrepancy issues.

Figure 2. Distribution of the F statistic for heterozygosity within each ethnic subgroup



Principal component analysis (PCA) of genetic ancestry

Background

Principal component analysis is performed as part of pre-imputation quality control to identify and remove participants with genetic ancestry that differs dramatically from the overall sample. Inclusion of PCA outliers in genotype imputation can lead to lower imputation accuracy.

The top ten principal components for genetic ancestry are commonly used as covariates in genome wide association analyses. Inclusion of ancestry principal components helps to minimize the potential for confounding by ethnicity (i.e., population stratification).

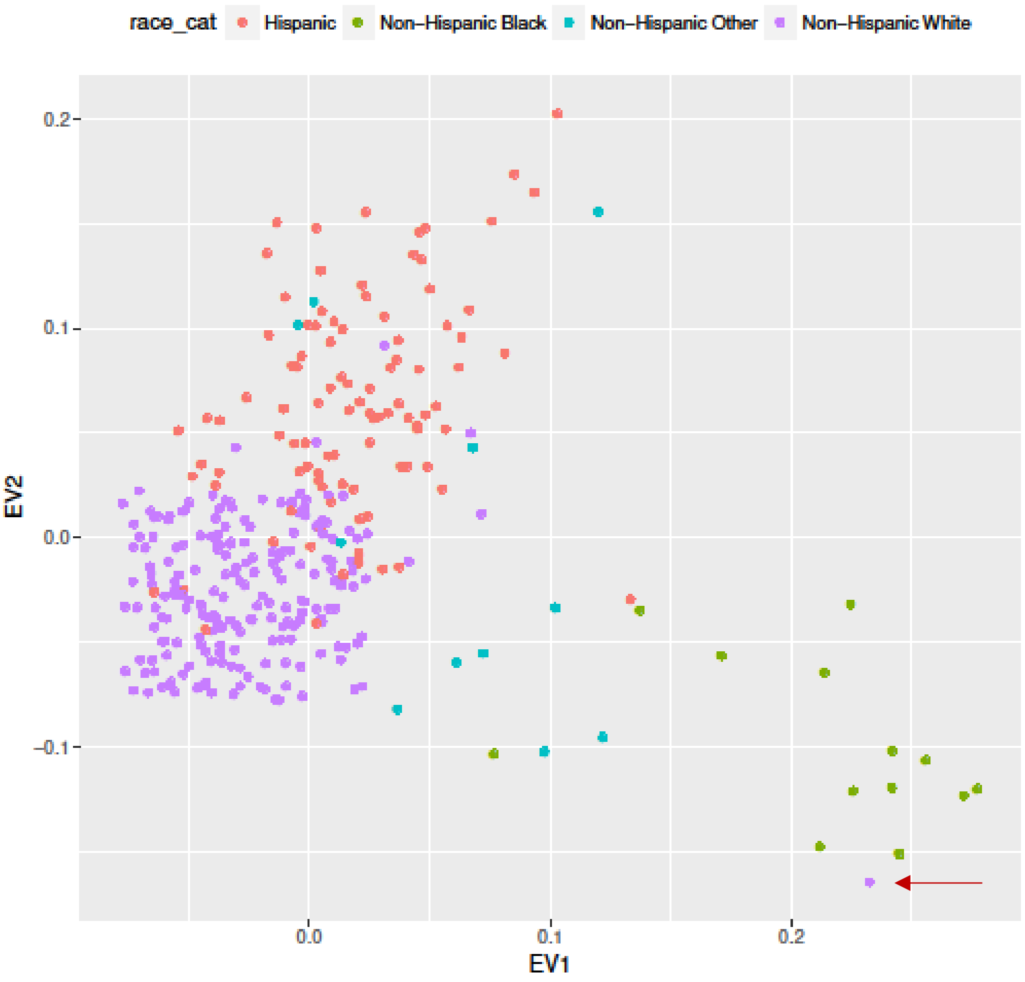
EPOCH

Principal component analysis of genetic ancestry was conducted to identify ancestry outliers and to produce covariates for use downstream genome-wide association analyses. In preparation for PCA, variants that passed the quality control standards outlined above underwent linkage disequilibrium (LD) pruning to identify a subset of pairwise-independent variants with a maximum pairwise correlation of 0.2.

Following LD pruning, Plink 1.9 was used to calculate the top ten principal components for genetic ancestry within the EPOCH cohort. Principal components were calculated across the full EPOCH cohort, and separately for the subgroup of non-Hispanic white participants to facilitate subsequent sensitivity analyses.

Calculated principal components were evaluated for the presence of participants with ancestry that differed dramatically from the remainder of the sample within the same ethnic subgroup.

Figure 3. First and second principal components of genetic ancestry as calculated From common variants (MAF>5%)



The PCA plot represented in Figure 2 (above) suggests the presence of a potential outlier within the non-Hispanic white subgroup of the EPOCH cohort. While the majority of the non-Hispanic white participants cluster together at the left of the plot, this participant lies at the far right within a cluster of non-Hispanic black participants. It is possible that this participant is an admixed individual who self-identified as non-Hispanic white when asked to indicate his or her categorical race/ethnicity. Further discussion is warranted to decide whether imputation should be performed again with the exclusion of this participant.

Quality control

Following the generation of quality control statistics and plots, quality control standards were implemented to produce a quality-controlled dataset suitable for imputation. Strand alignment was performed and inclusion/exclusion thresholds were applied to retain variants and individuals with high call rates and reliable genotypes.

Strand alignment

Background

Each human chromosome contains two strands of DNA, the ‘forward’ strand and the ‘reverse’ strand. The DNA sequence contained on the reverse strand is a mirror image of the sequence encoded on the forward strand. The allele present on each homologous chromosome can be expressed in terms of the nucleotide present on the forward strand, or its complementary base on the reverse strand. For the purposes of genome-wide association analyses, either representation of genotype may be used. By contrast, genotype imputation requires knowledge of the DNA sequence as read from the forward strand.

Within the DNA sequence on a given strand are segments that are commonly inherited as a unit. These co-inherited regions of the genome, known as haplotype blocks, allow us to infer the allele present at an unobserved locus based on the alleles observed at nearby loci.

Microarrays commonly capture some genotypes on the forward strand and others on the reverse strand. The strand orientation of each variant in the raw genotype data is dependent on the microarray platform and the specific version number of the array used.

Strand alignment is most accurately accomplished with the use of an array-specific ‘strand file.’ Closely resembling a manifest file, a strand file contains information about each variant in the microarray, including its reference and alternative alleles, its chromosome and position, and the strand on which the alleles are reported. With this information, a simple algorithm can be applied to ‘flip’ the strand of variants whose genotypes are represented on the reverse strand. In practice, this equates to substituting alleles represented on the reverse strand for their complementary bases.

Figure 4. Nucleotides as represented on the reverse (-) strand and the forward (+) strand

Raw		Aligned	
Allele	**Strand**	**Allele**	**Strand**
A	-	T	+
T	-	A	+
C	-	G	+
G	-	C	+

Flipping variants originally represented on the reverse strand produces an aligned dataset in which all genotypes are reported in terms of the alleles present on the forward strand of DNA within each homologous chromosome. Alignment of all genotypes to the forward strand enables haplotype inference, which in turn permits accurate imputation of unmeasured genotypes.

EPOCH

EPOCH genotyping was done using the Illumina Omni 2.5 microarray. There was initially some difficulty retrospectively identifying which version number of the Omni 2.5 was used. With the help of the Genomics Core, Illumina Tech Support was able to identify the EPOCH project and confirm that the Omni2.5-8-v1.1 was used for genotyping.

Genotypes produced by Illumina microarrays can be exported from Illumina’s Genome Studio software in three strand formats: TOP, ILMN, and SOURCE. The strand format used to export EPOCH’s raw genotypes from Genome Studio was unknown. In order to identify the strand format computationally, the reference and alternate alleles given in the raw *.bim file were compared to the alleles represented in the three possible output formats. This process confirmed that the genotypes were exported using the TOP strand format.

Strand alignment was performed using resources curated by the Wellcome Trust Institute, as recommended by the Blood Cell Consortium (BCX) and by other major consortia. Having confirmed the genotyping platform used, the version number of the Omni 2.5, and the strand export format, raw EPOCH genotypes were aligned using a strand file curated by the Wellcome Trust Institute that corresponds uniquely to the Omni2.5-8-v1.1 with genotypes represented in the TOP strand format. Strand alignment was executed using a script developed at the Wellcome Trust Institute and published for free use. Resources curated by the Wellcome Trust Institute were leveraged at the recommendation of the analysis plans developed for the Blood Cell Consortium (BCX) and other major consortia. Documentation regarding the resources used for strand alignment of the EPOCH GWAS data can be found at http://www.well.ox.ac.uk/~wrayner/strand/index.html.

Inclusion thresholding

Genotypes measured in the EPOCH data were filtered to retain only the variants that are still represented on the most recent version of Illumina's Omni 2.5 microarray (Omni2.5-8-v1.4).

Strict variant- and individual-level call rate thresholds were applied to generate a high-quality subset of data to be imputed. Individuals with less than 5% missing genotypes were retained in the sample to be imputed. Variants with less than 2% missing genotypes across individuals were retained to inform the imputation.

421,735 variants were removed due to missing genotypes in more than 2% of EPOCH participants. After thresholding, 1,925,180 variants remained in the data set for imputation.

Three participants were removed due to more than 5% missing genotypes genome-wide. 333 participants remained in the data set for imputation, including 91 cases who were exposed to gestational diabetes mellitus (GDM) and 242 controls who were not. The total genotyping rate in samples retained for imputation was 99.7%.

Table 2. Participants removed due to genotype call rate < 95% (> 5% missing genotypes)

PID	*Sex*	*Race*	*GDM*	*Frequency missing genotypes (%)*
20365	Male	Non-Hispanic White	No	20.7%
20483	Female	Hispanic	Yes	13.3%
20531	Female	Non-Hispanic White	No	82.4%

Michigan Imputation Server

Quality control

Prior to initiating imputation, the Michigan Imputation Server automatically performs its own rigorous quality control procedures. Uploaded data that fails any of the server's quality control standards is automatically rejected for imputation. An outline of the Michigan Imputation Server's quality control pipeline can be found at https://imputationserver.readthedocs.io/en/latest/pipeline/.

Imputation parameters

Many algorithms and software exist for the purposes of genotype imputation. The Michigan Imputation Server implements the Minimac3 method of imputation.

The server requires the user to upload quality controlled genotype data and specify parameters for the imputation algorithm. The user must select a reference panel, a phasing method, and a representative population for quality benchmarking.

For the imputation of EPOCH autosomal genotypes, the 1000 Genomes Phase 3 (Version 5) served as the reference panel. Eagle v2.3 was selected as the phasing method. Given the multi-ethnic nature of the EPOCH study cohort, the 'Mixed' population was selected for use in the server's quality control procedures.

Imputation results

Imputation was successfully completed for a data set including all EPOCH participants with genotyping call rates greater than 95% and all variants meeting inclusion thresholds described in this document. Further discussion is warranted to determine whether data should be re-imputed with the exclusion of any participants with evidence of outlier PCA values, unexpectedly low or high rates of heterozygosity, or sex discrepancies.