Michigan Genomics Initiative Data Freeze 3 Technical Notes v1.2

July 15, 2020

Brett Vanderwerff[1,2*], Lars Fritsche[1,2], Anita Pandit[1,2], Matthew Zawistowski[1,2], Michael Boehnke[1,2], & Sebastian Zöllner[1,2,3]

Author Affiliations:

[1]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

[2]Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

[3]Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA

* To whom correspondence regarding data preparation should be addressed: brettva@umich.edu

Major Changes from v1.1

- Freeze 3 data that were imputed using the TOPMed reference panel are now available, providing an increased number of well-imputed variants for analysis. Sections 2, 4.2, and 5.3 provide new information for these data.
- Figure 2 now reports the number of variants present in the imputed data sets *after* the application of standard post-imputation filters to exclude poorly imputed variants (Rsq < 0.3) or very rare variants (MAF < 0.01%).
- Table 4 now summarizes non-reference-homozygote genotype call concordance.

1 Overview

Data Freezes are regular releases of phased and imputed genetic data sets derived from samples of participants in the Michigan Genomics Initiative (MGI) cohort. This document provides a brief description of the properties of Data Freeze 3, released in March 2020. This data description is followed by an overview of the data generation methods, sample- and variant-level quality control (QC) strategies, and data quality evaluation.

2 Data Description

The Freeze 3 data sets contain phased and imputed haplotypes of 56,984 individuals of predominantly European (51,521) majority ancestry along with smaller numbers of predominantly African (3,198), East Asian (973), Central/South Asian (641), Native American (333), and Western Asian (318) descent (Figure 1).

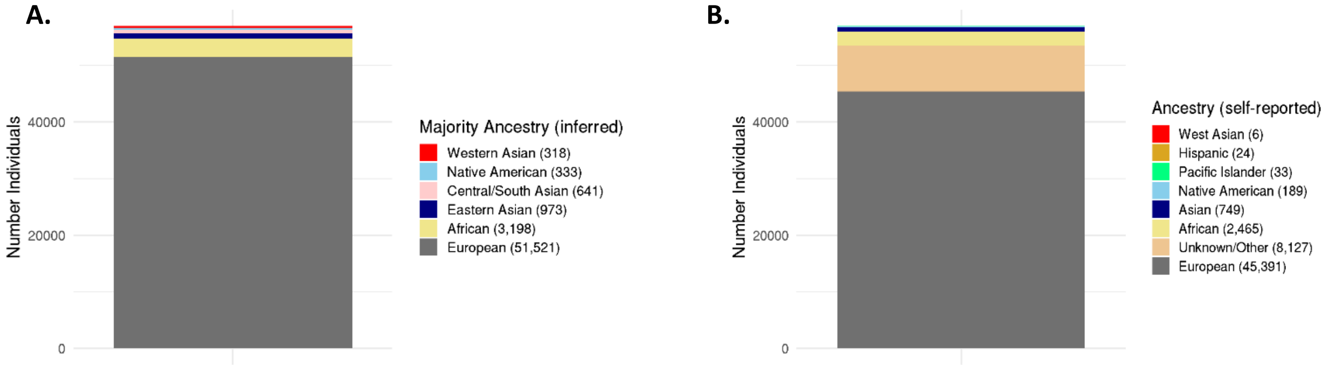

***Figure 1.** Ancestry for samples in the Freeze 3 MGI cohort. **(A.)** Majority ancestry inferred from genetic data. **(B.)** Self-reported ancestry by study participants.*

Freeze 3 data are available mapped to the coordinates of the Human Genome Reference Consortium Human Build 37 (GRCh37) or 38 (GRCh38). These versions were imputed with the Haplotype Reference Consortium r1.1 (HRC) or the Trans-Omics for Precision Medicine r2 (TOPMed) genotype reference panels, respectively[1,2].

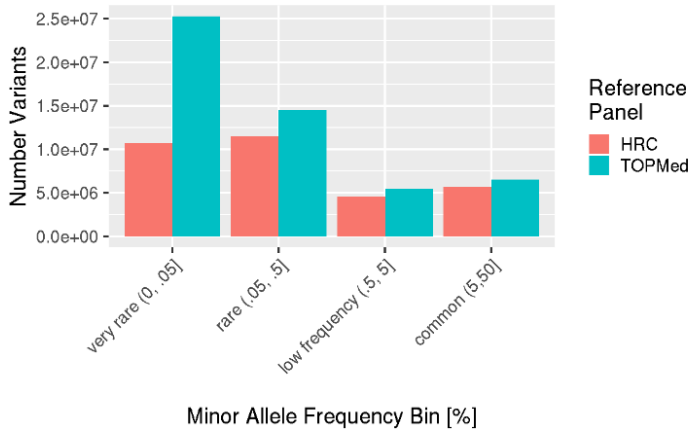

Figure 2. *Number of high-quality variants (Rsq > 0.3, MAF > 0.01%) in the data sets imputed with the HRC or TOPMed reference panels binned by minor allele frequency.*

After imputation with the HRC panel, Freeze 3 contains 502,255 genotyped and 39,954,964 imputed only variants for a total of 40,457,219. All variants imputed using the HRC reference panel are single nucleotide variants (SNVs). Applying standard post-imputation filters to remove poorly imputed variants (Rsq < 0.3) and very rare variants (minor allele frequency (MAF) < 0.01%), resulted in a high-quality data set containing 32,477,751 variants. After imputation with the TOPMed panel, Freeze 3 contains 501,607 genotyped and 307,016,210 imputed only variants for a total of 307,517,817 variants. 21,981,323 and 285,536,494 of these variants are indels and SNVs, respectively. 3,692,031 indels and 48,165,288 SNVs (51,857,319 variants total) pass the standard post-imputation Rsq and MAF filter. Among variants imputed with the TOPMed panel, 49% have MAF < 0.05%; among variants imputed with the HRC panel, 33% have MAF < 0.05% (Figure 2).

The different types of data that are available with the release of Data Freeze 3 are described in Table 1. The imputed data sets where standard post-imputation Rsq and MAF filters have been applied have the highest quality imputed and genotyped variants. Intermediate files that include more variant calls are also available. The rawest form of data available are genotype calls for each array after sample-level QC; these data sets include appropriately flagged low-quality variants. A data set produced by merging data from both array versions is also available. These merged data have undergone both sample- and variant-level QC but are not phased. All data sets are provided in VCF format.

To access these data, please apply through our ticketing system (submit a "Custom Data Request" in JIRA): https://doctrjira.med.umich.edu/. You will need to submit an IRB application through IRBMED to access these data, which you can apply for in eResearch Regulatory Management: https://its.umich.edu/academics-research/research/eresearch. For further assistance, please contact the Research Scientific Facilitators at phdatahelp@umich.edu, who can guide you through the data request process.

Intermediate and Imputed Data Sets

	# Samples	*# Variants*	
		GRCh37	*GRCh38*
Genotyped Array 1.0	20,023	603,583	603,583
Genotyped Array 1.1	37,088	607,778	607,778
Merged Genotyped	56,984	502,256	501,608
Phased	56,984	502,255	501,607
HRC Imputed (unfiltered)	56,984	40,457,219	-
HRC Imputed (filtered*)	56,984	32,477,751	-
TOPMed Imputed (unfiltered)	56,984	-	307,517,817
TOPMed Imputed (filtered*)	56,984	-	51,857,319

**Variants with R^2< 0.3 AND/OR MAF < 0.01% excluded*

Table 1*. The total number of samples and variants associated with the imputed and intermediate data sets available with Data Freeze 3.*

3 Data Production

3.1 Genotype Calling

All samples were genotyped by the University of Michigan Advanced Genomics Core on one of two custom array versions based on the Illumina Infinium CoreExome-24 bead array platform:UM_HUNT_Biobank_11788091_A1/Array 1.0 or UM_HUNT_Biobank_v1-1_20006200_A1/Array 1.1. Both arrays were designed with the same backbones containing probes corresponding to ~570,000 variants: ~240,000 tag single nucleotide variants and ~280,000 exonic variants. Custom probes corresponding to ~60,000 variants were incorporated into each array to detect candidate variants from GWAS, nonsense and missense variants, ancestry informative markers, and Neanderthal variants. This custom content included probes corresponding to ~30,000 predicted Loss-of-Function (LoF) variants. LoF variants require *de-novo* genotyping by two probe-based design. Due to a design flaw, ~21,000 predicted LoF variants in the custom content were paired with only a single probe during the array design. As these single probes are not optimal for LoF variant detection, LoF variants associated with a single probe design were flagged as “experimental” and excluded from the data set before phasing and imputation. Samples were genotyped on a rolling basis in batches of approximately 576 to 1,152 samples.

To improve genotyping accuracy, all accumulated sample batches processed on each genotyping array were combined for array-wise genotype calling at the time of Data Freeze creation. Raw Intensity Data files produced from array scanning were imported into GenomeStudio 2.0 running the Genotyping Module v2.0.4 and the GenTrain clustering algorithm v3.0. Automatic clustering of variants was performed following the GenomeStudio Genotyping Module protocol[3]. Where automatic clustering performed poorly, manual review and curation of cluster definitions was performed[4]. Data were then exported from GenomeStudio and used as input for the rare variant caller ZCall (v3.4) to recover rare variants that may have been misclustered by the automatic clustering process[5].

3.2 Merging Data Across Genotyping Arrays

MGI samples have been genotyped across multiple array versions. These array versions had identical design backbones but were synthesized in different batches. After removing variants where genotype data significantly differed across arrays (see *Section 4.2, Variant QC*), variants that represented the intersection of both arrays were used as input for phasing.

3.3 Phasing

Phasing was performed on the merged genotype data across all participants to estimate haplotypes. The data set was first divided into 23 separate files containing genotype data for chromosomes 1-22 or the non-pseudoautosomal (PAR) regions of chromosome X. Each of these files were independently phased using the software EAGLE (v 2.4.1)[6]. Genetic map coordinates were inferred by using a reference genetic map of GRCh37 or GRCh38 that was available with the distribution of Eagle. The entire MGI cohort was phased together without the use of a reference panel (“within-cohort” phasing).

3.4 Imputation

Imputation was used to expand the size of the phased data sets by estimating genotypes that were not directly assayed on the arrays. We offer two data sets, one imputed with the HRC reference panel mapped to GRCh37, the other imputed with the TOPMed reference panel mapped to GRCh38.

3.4.1 Imputation with the HRC Reference Panel

The HRC reference panel consists of 64,940 predominantly European haplotypes and 40,457,219 genetic variants[1]. In preparation for imputation, data files corresponding to chromosomes 1-22 or the non-PAR regions of chromosome X were divided into chunks using the automated chunking feature of the imputation software Minimac4 (v1.0.0)[7]. Minimac4 was set to output data in hard genotype, estimated alternate allele dosage, and estimated haploid alternate allele dosage formats. After the imputation process, chunked data were merged to the chromosome-level.

3.4.2 Imputation with the TOPMed Reference Panel

The TOPMed reference panel includes haplotypes from 194,512 ancestrally diverse samples and 308,107,085 genetic variants[2]. Imputation was performed using the TOPMed Imputation Server pipeline (v1.2.7) at https://imputation.biodatacatalyst.nhlbi.nih.gov/. Due to a limit on sample size by the pipeline, the full MGI cohort was divided at random into 3 evenly sized sub-cohorts and each sub-cohort was imputed separately. After imputation was complete, genotypes from the sub-cohorts were merged with Bcftools (v1.9)[8].

3.5 Ancestry Inference

The majority ancestry of MGI participants corresponding to each sample was first inferred by performing principal component analysis using PLINK (v1.9)[9]. Principal component calculations were based off a reference genotype panel of Human Genome Diversity Project (HGDP) samples[10]. MGI samples were projected onto the space created by the first two principal components of the HGDP samples. MGI samples were inferred to be of European ancestry if they fell within a circle drawn around European HGDP samples. The circle was defined by a radius that was 1/8 the distance between the European HGDP sample centroid and the centroid formed between European, East Asian, and African HGDP samples. A similar approach was taken to infer the ancestry of East Asian and African samples. The majority ancestries of samples that did not fall into the areas defined by these circles were inferred with the software ADMIXTURE[11]. MGI samples were merged with a reference panel of HGDP samples. Merged data were analyzed by running ADMIXTURE in supervised mode using the number of HGDP super-populations (K=6) as a template. Ancestry inferred by this method was summarized to the largest ancestry fraction reported by ADMIXTURE.

4 Data Quality Control

4.1 Sample QC

Sample-level QC was performed on a rolling basis as batches of samples were genotyped. This approach allowed prompt response and issue remediation if needed. A sample was flagged per batch and excluded from the Data Freeze if any of the following issues were raised during sample QC: (1) patient had withdrawn from the study, (2) genotype-inferred sex did not match the self-reported gender of the patient or self-reported gender was missing, (3) sample had an atypical sex chromosomal aberration (e.g. Klinefelter syndrome), (4) sample shared a kinship coefficient > .45 with another sample with a different ID, (5) sample-level call-rate was below 99%, (6) sample was a technical duplicate or twin of another sample with a higher call-rate either within the same array or across arrays, (7) estimated contamination level exceeded 2.5%, (8) call-rate on any individual chromosome was five-fold lower than that of all other chromosomes, or (9) sample was processed in a DNA extraction batch that was flagged for technical issues (Table 2). Sample QC

Samples Excluded by QC in Each Array

	#Failing Samples	
Description	*Array 1.0*	*Array 1.1*
Excluded DNA extraction batches	746	73
Duplicated sample with higher call-rate in same array	128	58
Sample pair w/ Identical IDs & discordant genotypes	16	2
Sample pair w/ different IDs & similar genotypes	10	494
Unusual XY composition, e.g. Turner syndrome	32	65
No gender information available	1	94
Reported gender different from genotype inferred sex	121	91
Estimated contamination > 2.5 %	118	144
Chromosomal call-rate drop > 5 %	15	39
Sample call-rate < 99%	97	115
Duplicated sample with higher call-rate in another array	79	35
Total Samples:	***1,363***	***1,210***

Table 2. *Numbers of samples excluded from Data Freeze 3 based on various QC outcomes. Sample exclusion counts are distributed among the arrays on which samples were processed.*

analysis was performed with in-house developed R scripts. Pairwise relatedness between samples was estimated using the relationship inference software KING (v2.1.3)[12]. Contamination between samples was estimated by the contamination detection software VICES[13]. PLINK was used to determine sample level call-rates.

4.2 Variant QC

To determine genotyping array probe specificity, probes were mapped to the sequences of GRCh37 or GRCh38 and the revised Cambridge Reference Sequence of human mitochondrial DNA (rCRS) using the sequence alignment tool BLAT (v. 351)[14]. Variants where corresponding array probe(s) did not uniquely and perfectly map to the chromosome sequences of the GRCh37, GRCh38, or the rCRS reference were excluded from analysis.

Several quality control flags were assigned to the remaining variants that were represented on both arrays (Table 3). "GenTrain" and "Cluster Separation" scores are internal QC metrics from the GenomeStudio Genotyping Module that measure the overall quality of clusters produced by the GenTrain algorithm[4]. Cluster Separation and GenTrain scores range from 0 to 1, with lower scores suggesting poor cluster separation and lower cluster quality[4]. Variants with a GenTrain score < 0.15 and/or a Cluster Separation score < 0.3 were excluded from the final data set.

Variants Excluded by QC in Each Array

	# Failing Variants GRCh37			*# Failing Variants GRCh38*		
Exclusion Flag	*Array 1.0*	*Array 1.1*	*Both Arrays*	*Array 1.0*	*Array 1.1*	*Both Arrays*
GenTrain score < 0.15	27	1,643	11	27	1,600	11
Cluster Separation score < 0.3	1,583	718	948	1,587	712	950
Call-rate < 99%	15,631	1,720	2,981	15,641	1,725	2,968
HWE test p < 10^{-4} within array	2,240	1,678	1,260	2,188	1,608	1,256
FET p < 10^{-3} between arrays	0	0	1,766	0	0	1,769
Minor allele frequency of 0	0	0	39,915	0	0	39,820
HWE test p < 10^{-6} after array merge	0	0	33	0	0	49
Total Variants:	***18,122***	***4,531***	***45,969***	***18,100***	***4,452***	***45,869***

Table 3. *QC outcomes for variants that were represented on both arrays. Depicted are the numbers of variants that failed either uniquely on Array 1.0 or uniquely on Array 1.1 for each exclusion flag. The number of variants that failed on both arrays are also shown.*

Deviation from Hardy-Weinberg equilibrium (HWE) for each variant was first tested at the array level in a sub-population of the complete MGI cohort that contained only individuals with recent European ancestry that were unrelated to the second degree (KING). HWE was rejected if an exact test produced a p-value < 10^{-4}.

To detect array-specific batch effects, Fisher's exact test (FET) was performed on variants that were represented on both arrays and passed QC. Variants that were associated with a p-value < 10^{-3} were assumed to differ between arrays due to batch-effects introduced during the genotyping process. Variants with a p-value below this threshold were pruned from the data set before merging genotype data across both arrays. After merging arrays, deviation from HWE was again tested in a subset of individuals with recent European ancestry that were unrelated to the second degree (PLINK, KING). Variants with a p-value < 10^{-6} were removed from the merged data set. Additionally, variants with a MAF of 0 across all individuals in the merged data set (monomorphic variants) were removed.

5 Data Quality Evaluation

5.1 Genotype Concordance

Pairs of samples that were genotyped more than once on each array version (technical duplicate samples) allowed for the assessment of genotype call concordance on each array. 153 and 304 pairs of technical duplicate samples were genotyped on UM_HUNT_Biobank_11788091_A1/Array 1.0 and UM_HUNT_Biobank_v1-1_20006200_A1/Array 1.1, respectively. Genotype call concordance rate between samples was determined by evaluating: (# concordant calls / # total calls) x 100. This calculation was performed both at all sites and only those sites where at least one sample of the duplicate pair had a non-reference-homozygote call. Concordance was measured before application of variant-level QC and after removing variants that failed QC. Removing variants that failed QC led to increased genotype call concordance on both arrays (Table 4).

Array-based Genotype Concordance

		Pre-Variant QC Concordance [%]		*Post-Variant QC Concordance [%]*	
	Pairs of Duplicates	*All*	*NRH*	*All*	*NRH*
Array 1.0	153	99.74	99.58	99.91	99.84
Array 1.1	304	99.91	99.87	99.94	99.92

Table 4. *Concordance of genotype calls that were made for identical samples that were genotyped twice on the same array. Genotype concordance was evaluated at both all genotyped sites and only those sites where at least one sample had a non-reference-homozygote (NRH) call. Concordance was measured both before and after the application of variant-level QC.*

5.2 Phasing Evaluation

Phasing quality was evaluated by switch error rate (SWE)[15]. To develop a "gold standard" phased reference sample, 77 parent-parent-child trios were first identified in the full MGI cohort with KING. The trios were phased using pedigree information with Beagle v4.0[16]. The parents of each trio were then removed from the full MGI cohort before phasing the remaining samples with Eagle as described in *Section 3.3, Phasing*. Children from the trios that were phased with Eagle were then compared to their "gold standard" pedigree phased counterparts. SWE across all autosomes was determined by evaluating the total number of strand switches that occurred over the total number of heterozygous sites where strand switches were possible[15]. Sites with Mendelian errors and those sites heterozygous in all trio members were not considered in the SWE calculation. SWE varied among different populations of inferred majority ancestry ranging from 1.9% in Europeans to 7.9% in East Asians (Figure 3).

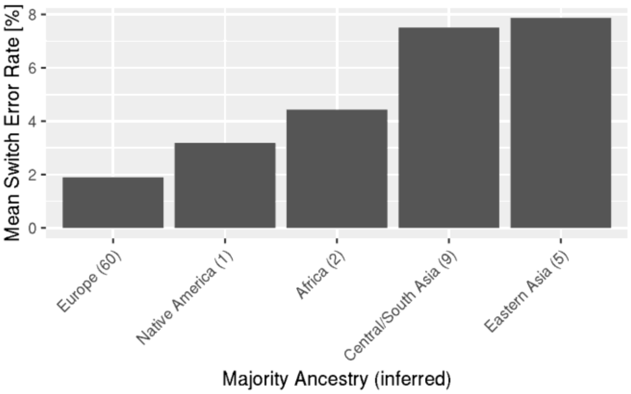

Figure 3. *Evaluation of phasing performance by switch error rate (SWE). SWE is summarized across several inferred majority ancestry groups.*

5.3 Imputation Evaluation

Imputation quality was measured by the Rsq and EmpRsq metrics produced by the imputation software Minimac4. The Rsq metric estimates imputation quality at imputed sites by the formula Var(HDS) / (p(1-p)) where HDS is the estimated haploid alternate allele dosage and p is the mean of HDS. EmpRsq is an imputation quality metric available at all sites that were both genotyped and imputed. It is defined as the Pearson correlation coefficient of known and imputed genotypes as if the known genotypes were masked. For the data set that was imputed with the TOPMed reference panel, Rsq and the MAF of each variant were estimated for the merged data set by taking the mean of each of these values across 3 separately imputed sub-cohorts (see *Section 3.4.2, Imputation with the TOPMed Reference Panel*). Both imputation quality metrics improved with increasing MAF when using either the HRC or TOPMed panel. Direct comparisons of imputation quality between Freeze 3 data imputed using the HRC or TOPMed reference panels are possible at sites that are imputed by both panels. Imputation of Freeze 3 using the TOPMed panel resulted in a relative increase in mean Rsq and EmpRsq compared to HRC-based imputation across all MAF bins except for the lowest bin ([0,0.01] %, Figure 4A-B). Rsq at all sites that were imputed by either the HRC or TOPMed panels were also evaluated. Rsq of SNVs and indels that were imputed using the TOPMed panel are comparable, no indels were imputed when using the HRC panel as reference (Figure 4C-D).

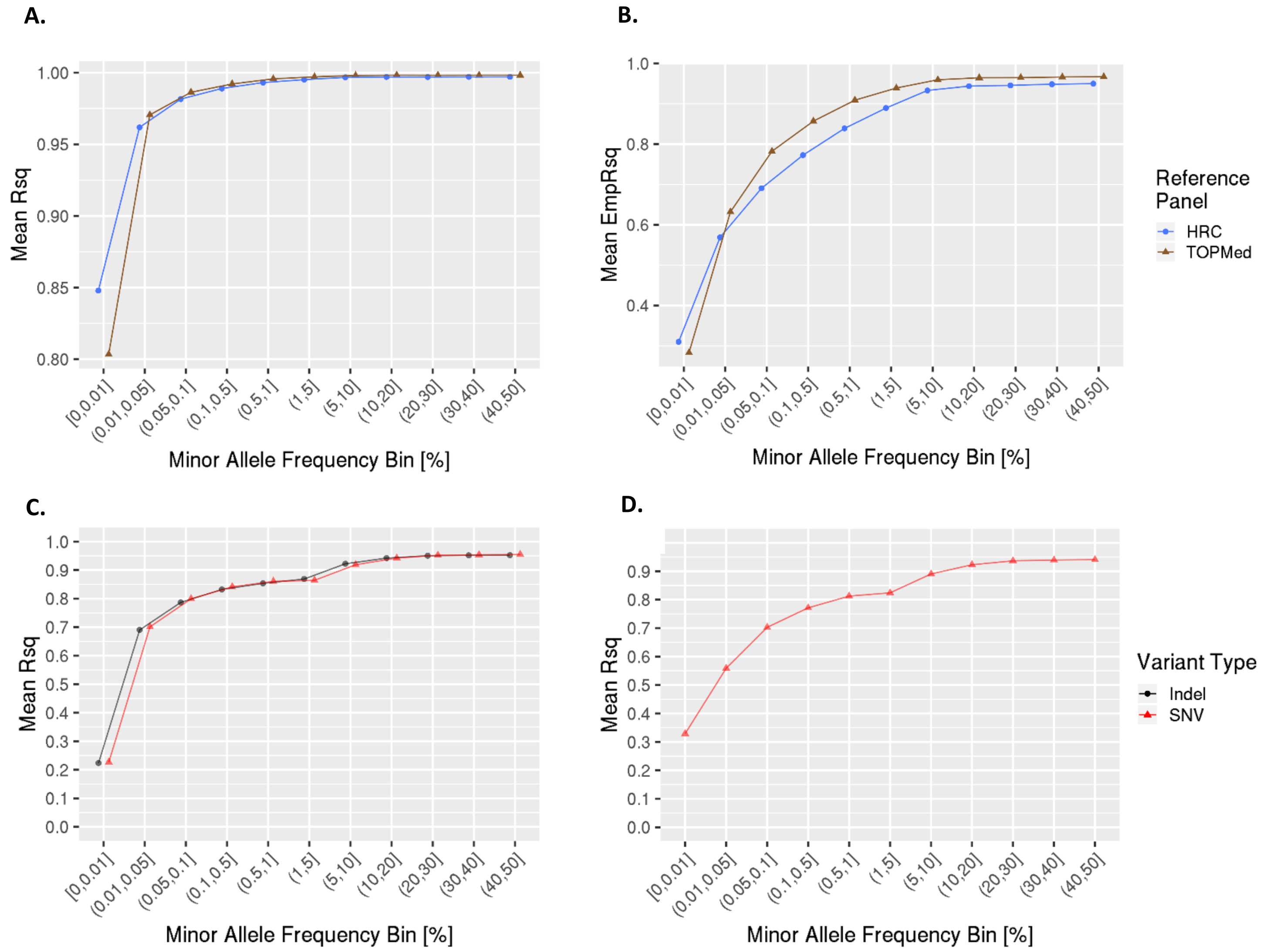

Figure 4. *Summary of imputation quality metrics for the data sets imputed with the HRC or TOPMed reference panels. 404,279 sites that were genotyped on the arrays and imputed across both reference panels were used to evaluate:* ***(A.)*** *the estimated correlation between imputed and expected genotypes (Rsq) and* ***(B.)*** *the Pearson correlation coefficient of known and imputed genotypes (EmpRsq). Rsq is summarized for all single nucleotide variants (SNVs) or indels that were imputed by using either the* ***(C.)*** *TOPMed reference panel (285,509,108 SNVs, 21,981,323 indels) or* ***(D.)*** *HRC reference panel (40,359,612 SNVs).*

5.4 Principal Component Calculation

The first 10 principal components for all samples in the cohort were calculated from quality-controlled genotype data. The data were first pruned to remove all variants with a MAF < 1%. Additionally, pairs of variants with a squared correlation > 0.5 within a walking window of 500 variants and a step size of 5 were thinned (PLINK). Variants in the major histocompatibility complex region were also removed. Relationship inferences were made to identify all individuals that were related to the second degree (KING). 8,342 inferred related samples were separated from the remaining 48,642 unrelated samples. Principal components were computed from the unrelated samples using FlashPCA2 v2.0[17]. The related samples were then projected onto the principal components of the unrelated samples. Using the same approach that was applied to the full MGI cohort, a second set of principal components were generated for only those samples with inferred majority European ancestry (45,293 unrelated & 6,228 related samples, Figure 5).

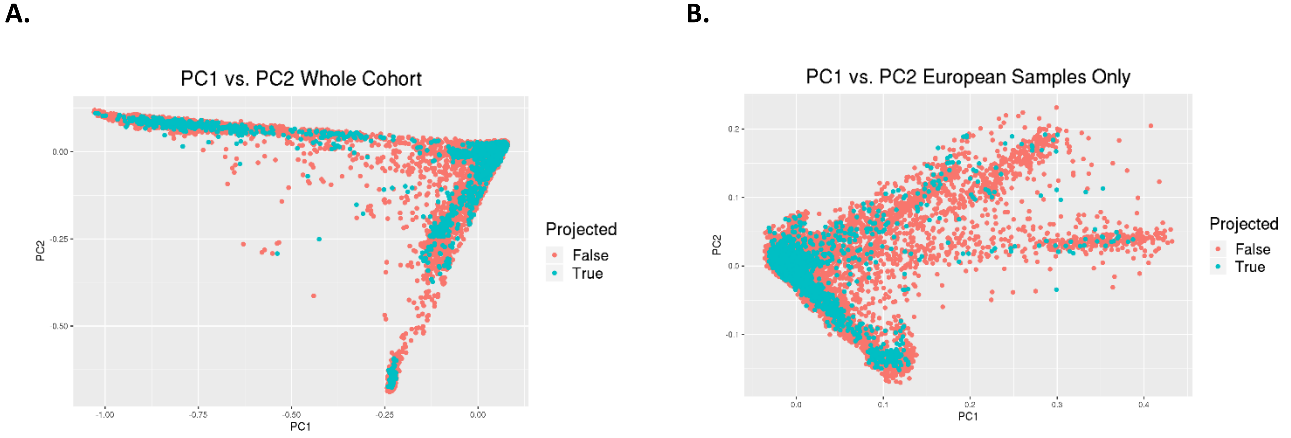

Figure 5. *Plots of the first and second principal components for* ***(A.)*** *all samples in the MGI cohort and* ***(B.)*** *those samples with inferred majority European ancestry. For both cohorts, samples inferred to be related were projected onto the principal components of unrelated samples.*

6 References

1. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

2. TOPMed Imputation Server. https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about.

3. GenomeStudio Documentation.

   https://support.illumina.com/array/array_software/genomestudio/documentation.html.

4. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).

5. Goldstein, J. I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinforma. Oxf. Engl.* **28**, 2543–2545 (2012).

6. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).

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**, 955–959 (2012).

8. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

9. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

10. Stanford University. https://www.hagsc.org/hgdp/.

11. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

12. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

13. Zajac, G. J. M. *et al.* Estimation of DNA contamination and its sources in genotyped samples. *Genet. Epidemiol.* **43**, 980–995 (2019).

14. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

15. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, (2018).

16. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

17. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).