Michigan Genomics Initiative Data Freeze 4 Technical Notes

Brett Vanderwerff[1,2,*], Lars Fritsche[1,2], Anita Pandit[1,2], Snehal Patil[1,2,3], Matthew Zawistowski[1,2], Michael Boehnke[1,2], Xiang Zhou[1,2], and Sebastian Zöllner[1,2,4]

[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 Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[*] *To whom correspondence regarding data preparation should be addressed:* bretva@umich.edu

August 3, 2021

1 Changes From Data Freeze 3

- Increase of available genotyped cohort size by 3,231 participants.

2 Data Description

Data Freeze 4 is comprised of 60,215 genotyped participants of the Michigan Genomics Initiative (MGI). 28,251 ($\approx 47\%$) of participants are male and 31,964 ($\approx 53\%$) are female. The median age for males is 62 compared to 57 for females. The self-reported race of participants as recorded during a medical office visit is Caucasian (51,967), African American (3,859), Unknown (2,229), Asian (1,829), American Indian or Alaska Native (273), Native Hawaiian and Other Pacific Islander (58) (Figure 1). The inferred majority genetic ancestry of the included participants is primarily European (53,054) with smaller numbers of African (3,761), Eastern Asian (1,281), Central/South Asian (891), Western Asian (780), and Native American (448). Although we primarily describe the MGI cohort using majority ancestry labels, MGI participants exhibit a range of genetic admixture (Figure S1).

11,409 of participants in Data Freeze 4 are inferred to be related to at least one other participant in the cohort at 3$^{\text{rd}}$ degree or closer. 7,906 MGI participants have at least one inferred 2$^{\text{nd}}$ degree or closer relative (KING v2.2.7, Figure S2).

We assayed $\approx$ 570,000 genotypes for each participant via genome-wide genotyping array and imputed millions of additional genotypes with the Haplotype Reference Consortium r1.1 (HRC) or the Trans-Omics for Precision Medicine r2 (TOPMed) reference panels [1, 2].

After imputation with the HRC panel, Data Freeze 4 contains 40,494,480 variants mapped to build 37 of the human genome. 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 32,401,123 variants (Table 1, Table S1). In this high-quality data-set $\approx$ 32% (10,276,791) have MAF $\leq$ 0.05% (Figure 2, Table S2).

After imputation with the TOPMed panel, Data Freeze 4 contains 307,896,277 variants mapped to build 38 of the human genome. 285,879,432 and 22,016,845 of these variants are SNVs and indels,

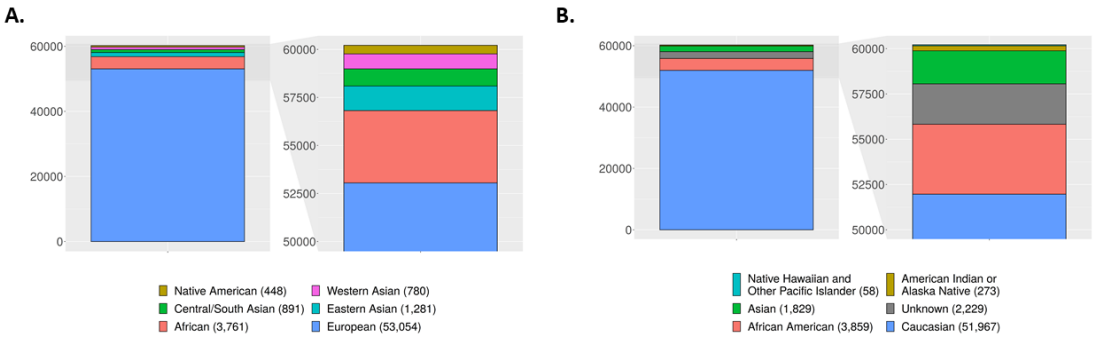

Figure 1: Genotype-inferred majority ancestry and self-reported race of MGI participants. (A.) Majority ancestry as inferred for MGI participants using the ADMIXTURE software with Human Genome Diversity Panel genotypes and continental population labels used as reference. **(B.)** Race as self-reported by MGI participants during a medical office visit. The left plot in each inset summarizes the full MGI cohort. The right plot in each inset is a zoom in view focusing on the non-European/non-Caucasian component of the full MGI cohort.

Data Set	# Samples	# Variants	
		GRCh37	GRCh38
CoreExome-24 v1.0	19,831	571,625	570,506
CoreExome-24 v1.1	37,953	575,621	574,490
CoreExome-24 v1.3	2,431	574,822	573,648
CoreExome arrays merged (unphased)	60,215	499,575	498,711
CoreExome arrays merged (phased)	60,215	499,574	498,710
HRC Imputed (unfiltered)	60,215	40,494,480	-
HRC Imputed (filtered*)	60,215	32,401,123	-
TOPMed Imputed (unfiltered)	60,215	-	307,896,277
TOPMed Imputed (filtered*)	60,215	-	52,740,810

Table 1: Intermediate and imputed data sets. The total number of variants associated with the intermediate and imputed data sets available with Data Freeze 4. Variant counts are given for versions of Data Freeze 4 mapped to the coordinates of GRCh37 or GRCh38. *, Variants with Rsq $<$ 0.3 or MAF $<$ 0.01% excluded.

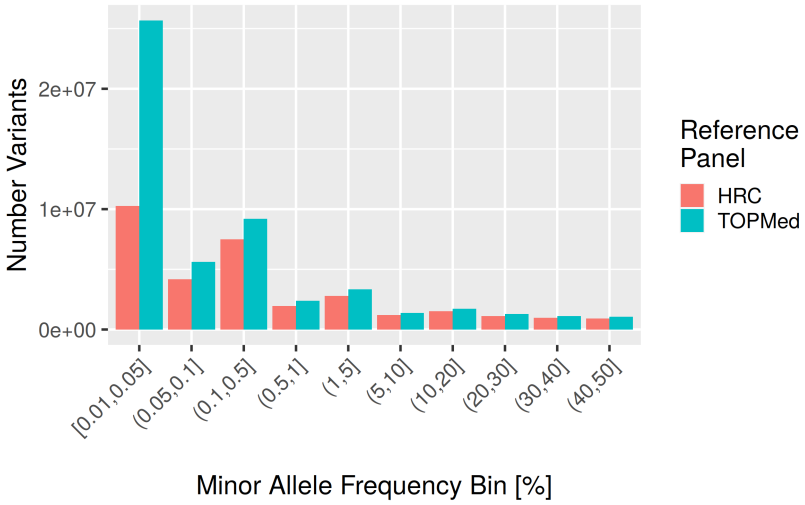

Figure 2: Distribution of variant frequency. The number of variants in the data sets imputed with HRC or TOPMed that fall into different minor allele frequency bins. Only variants that pass the standard post-imputation filter (Rsq ≥ 0.3 and MAF $\geq 0.01\%$) are plotted.

respectively. 48,986,377 SNVs and 3,754,433 indels (52,740,810 variants total) pass the standard post-imputation Rsq and MAF filter (Table 1, Table S1). $\approx$ 49% (25,678,109) of variants imputed with TOPMed that pass the standard post-imputation filter have MAF $\leq 0.05\%$ (Figure 2, Table S2).

The genotype data sets that are available with the release of Data Freeze 4 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 with genotype calls for each array after sample-level quality control (QC) (with appropriately flagged low-quality variants), a data set merging data from all array versions after sample- and variant-level QC (unphased), and the unfiltered set of HRC- or TOPMed imputed variants. 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.

3 Data Production

3.1 Genotype Calling

The University of Michigan Advanced Genomics Core (AGC) genotyped the samples of MGI participants on one of three custom array versions based on the Illumina Infinium CoreExome-24 v1.0, v1.1, or v1.3 bead array. All the arrays we used were designed with the same backbones containing probes corresponding to $\approx$ 570,000 variants: $\approx$ 240,000 tag single nucleotide variants and $\approx$ 280,000 exonic variants. We incorporated custom probes corresponding to $\approx$ 60,000 variants 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 $\approx$ 30,000 predicted Loss-of-Function (LoF) variants. LoF variants require de novo genotyping by two probe-based design. Due

to a flaw in our design, ≈ 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, we flagged LoF variants associated with a single probe design as "experimental" and excluded them from the data set before phasing and imputation. The AGC genotyped samples on a rolling basis in batches of ≈ 576 to 1,152 samples.

To produce genotype callsets, we imported raw Intensity Data files from array scanning into GenomeStudio 2.0 running the Genotyping Module v2.0.4 and the GenTrain clustering algorithm v3.0. To define the clusters that genotype calls are based on, we performed automatic clustering by following the GenomeStudio Genotyping Module protocol [3].

We performed two rounds of genotyping for most MGI samples. We first called sample genotypes per sample batch processed by the AGC. We used these preliminary callsets for sample-level QC (see subsection 4.1) and generated them by automatic clustering of only those samples belonging to each respective batch. We then called sample genotypes per array at the time of Data Freeze creation. These higher-quality callsets were generated by automatic clustering of all high-quality samples processed to date on each array.

Where array-based automatic clustering performed poorly, we performed manual review and curation of cluster definitions [4]. We used the rare variant caller ZCall (v3.4) to recover rare variants that may have been misclustered during the array-based automatic clustering process [5]. Due to limited sample size, we did not manually review cluster definitions or perform the associated ZCall work for the CoreExome v1.3 array.

3.2 Merging Data Across Genotyping Arrays

We first filtered to exclude variants where genotype data significantly differed across the 3 array versions we used to genotype MGI participants (see section subsection 4.2). We then merged these filtered data keeping only variants that intersect all array versions.

3.3 Phasing

We phased merged genotype data to estimate haplotypes. We divided the data set into 23 separate files containing genotype data for chromosomes 1-22 or the non-pseudoautosomal regions of chromosome X. We phased these files using the software Eagle (v 2.4.1) [6], using the reference genetic map of GRCh37 or GRCh38 distributed of Eagle. We phased the MGI cohort without the use of a reference panel ("within-cohort" phasing).

3.4 Imputation

We performed genotype imputation on the phased data sets to infer additional genotypes not directly assayed by array. 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.

The HRC reference panel consists of 64,940 predominantly European haplotypes and 40,405,505 genetic variants [7]. The TOPMed reference panel includes 194,512 haplotypes and 308,107,085 genetic variants from diverse samples [2].

We used the TOPMed Imputation Server pipeline (v1.5.7) at https://imputation.biodatacatalyst.nhlbi.nih.gov/ to impute MGI with the TOPMed reference panel. We used the Michigan

Imputation Server pipeline (v1.5.7) at https://imputationserver.sph.umich.edu/ to impute MGI with the HRC reference panel.

Due to server limits on sample size, we divided the full MGI cohort into 3 evenly sized sub-cohorts and imputed each sub-cohort separately. We reconstructed the full MGI cohort by merging imputed sub-cohorts with Bcftools (v1.9) [8].

3.5 Genetic Ancestry Inference

For the purpose of cohort description, we inferred the majority ancestry of MGI participants by using the software ADMIXTURE [9]. We merged genotypes of MGI samples with those of a reference panel of Human Genome Diversity Project (HGDP) samples [10]. These merged data were analyzed by running ADMIXTURE in supervised mode using the number of HGDP continental populations (K=7) as a template. Genetic ancestry inferred by this method was summarized to the largest Q value (ancestry fraction) reported by ADMIXTURE.

4 Data Quality Control

4.1 Sample QC

We performed sample-level QC on a rolling basis as batches of samples were genotyped. This approach allowed us to promptly detect and resolve issues 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) participant had withdrawn from the study, (2) genotype-inferred sex did not match the self-reported gender of the participant or self-reported gender was missing, (3) sample had an atypical sex chromosomal aberration (e.g. Klinefelter syndrome), (4) sample shared a kinship coefficient $\geq .45$ 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, (7) estimated contamination level exceeded 2.5%, (8) call-rate on any individual chromosome was $\leq 95\%$, or (9) sample was processed in a DNA extraction batch that was flagged for technical issues (Table 2). Our sample QC analysis was performed with in-house developed R and Python scripts. We estimated pairwise relatedness between samples with KING (v2.1.3), contamination between samples with VICES, and sample call-rates with PLINK (v1.9) [11, 12, 13].

4.2 Variant QC

To determine genotyping array probe specificity, we mapped probes 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]. We excluded variants where corresponding array probe(s) did not uniquely and perfectly map to the chromosome sequences of the GRCh37, GRCh38, or the rCRS reference.

We assigned quality control flags to the remaining variants that intersected all arrays. The number of well-mapping sites that fail in each array are provided in 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]. We excluded variants with a GenTrain score < 0.15 and/or a Cluster Separation score < 0.3 from the final data set.

Sample QC Flag	CoreExome-24 v1.0		CoreExome-24 v1.1		CoreExome-24 v1.3	
	#	%	#	%	#	%
UNUSUAL_XY	19	0.09	70	0.18	11	0.37
GENDER_MISMATCH	94	0.44	103	0.26	5	0.17
GENDER_MISSING	0	0	2	0.01	20	0.68
LOW_CALLRATE	97	0.46	143	0.37	18	0.61
LARGE_CHR_CNV	18	0.08	41	0.10	2	0.07
HIGH_CONTAM	118	0.56	201	0.51	38	1.29
TECH_ISSUE	739	3.49	80	0.20	1	0.03
TECH_DUP_ERROR	14	0.07	4	0.01	4	0.14
TECH_DUP	333	1.57	656	1.68	411	13.92
UNEXPECTED_DUP	22	0.10	61	0.16	28	0.95
Total Unique Failing Samples	1,394	6.58	1,195	3.06	477	16.16

Table 2: Sample QC outcomes. The total numbers and percentages of samples that fail various sample QC flags. UNUSUAL_XY, unusual XY composition (e.g. Turner syndrome); GENDER_MISMATCH, reported gender different from genotype-inferred sex; GENDER_MISSING, no gender information available; LOW_CALLRATE, sample call-rate < 99 %; LARGE_CHR_CNV, chromosomal call-rate drop > 5 %; HIGH_CONTAM, estimated contamination > 2.5 %; TECH_ISSUE, excluded DNA extraction batch; TECH_DUP_ERROR, sample pair w/ identical IDs & dissimilar genotypes; TECH_DUP, sample with same ID and similar genotypes as other sample; UNEXPECTED_DUP, sample pair w/ different IDs & similar genotypes.

Variant QC Flag	# Failing Variants GRCh37			# Failing Variants GRCh38		
	CE v1.0	CE v1.1	CE v1.3	CE v1.0	CE v1.1	CE v1.3
HWE	3,405	2,888	432	3,380	2,865	423
LOW_CALLRATE	17,922	4,157	1,767	17,871	4,100	1,743
LOW_CLUSTER_SEP	2,410	1,534	590	2,395	1,513	577
LOW_GENTRAIN	37	1,553	69	36	1,510	69
Total Unique Failing Variants	21,051	7,294	2,133	20,983	7,211	2,103

Table 3: Array-based variant QC. The number of well-mapping sites that fail variant QC flags in each array. Counts are given for versions of Data Freeze 4 mapped to the coordinates of GRCh37 or GRCh38. CE, CoreExome-24; HWE, Hardy-Weinberg equilibrium test p < 10^{-4} before array merge; LOW_CALLRATE, call-rate < 99%; LOW_CLUSTER_SEP, Cluster Sep. score < 0.3; LOW_GENTRAIN, GenTrain score < 0.15.

We tested variants for deviation from Hardy-Weinberg equilibrium (HWE) using a sub-population of MGI participants with majority European ancestry that were inferred to be unrelated to the 2nd degree by KING (v2.1.3). HWE was rejected if an exact test produced a p-value < 10^{-4}. For each CoreExome array, there were instances of variants failing QC by more than one of our array-based variant QC measures (Figure S3, Figure S4).

For each variant, we tested the hypothesis that genotype frequency does not differ between array versions with Fisher's exact test (FET). We assumed variants with a p-value < 10^{-3} differed in frequency due to batch-effects introduced during the genotyping process. ≈ 2,700 variants were excluded from each array based on FET results (Table S3).

After merging data across arrays, we tested again for deviation from HWE in a subset of individuals with majority European ancestry that were inferred to be unrelated to the 2nd degree. ≈ 60 variants with a p-value < 10^{-6} were excluded from each array (Table S3). We also excluded ≈ 36,000 monomorphic variants (MAF = 0) from each array (Table S3).

Array	# Pairs	All Sites		NRH Sites	
		Pre-QC	Post-QC	Pre-QC	Post-QC
CoreExome-24 v1.0	127	99.80	99.94	99.68	99.89
CoreExome-24 v1.1	318	99.90	99.95	99.88	99.93
CoreExome-24 v1.3	277	99.91	99.96	99.90	99.96

Table 4: Array-based genotype concordance. Concordance of genotype calls from samples 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 call. Concordance was measured both before and after the application of variant-level QC. Values are expressed as the percentage of concordant calls out of all compared calls. NRH, non-reference-homozygote.

5 Data Quality Evaluation

5.1 Genotype Concordance

We measured genotype concordance using 127, 318, and 277 pairs containing samples that were genotyped twice on the CoreExome-24 v1.0, CoreExome-24 v1.1, or CoreExome-24 v1.3 arrays, respectively. We considered genotypes concordant if they matched perfectly between samples. We evaluated concordance across a set of all genotypes and a set of only those genotypes where at least one sample of the duplicate pair had a non-reference-homozygote call. We measured concordance before and after removing variants that failed QC. For all arrays, removing variants that failed QC led to increased genotype call concordance (Table 4).

5.2 Phasing Quality

We evaluated phasing quality by switch error rate (SWE), a metric that describes the total number of strand switches that occur over the total number of heterozygous sites where strand switches are possible [15]. To obtain known maternal and paternal haplotypes, we used pedigree information inferred with KING (v2.1.3) to phase 118 trios using Beagle v4.0 [16]. We then removed the parents of each trio from the full MGI cohort before phasing the remaining samples with Eagle as described in subsection 3.3. We calculated SWE by counting switches that occurred at heterozygous sites between children phased with Eagle or their Beagle pedigree phased counterparts [15]. Sites with Mendelian errors and sites that were heterozygous in all trio members were excluded from our SWE calculation. SWE increases with decreasing chromosome length and is on average lower in European participants (Figure 3).

5.3 Imputation Quality

We used the "Rsq" and "EmpRsq" metrics produced by the imputation software Minimac4 to evaluate imputation quality. The Rsq metric estimates imputation quality at all imputed sites by the formula:

$$Rsq = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1-\hat{p})}$$

where $\hat{p}$ is the frequency of the alternate allele, D_i is the allele dosage for the i^{th} haplotype and n is the number of samples that are evaluated [17].

The EmpRsq metric measures imputation quality at all sites that were both genotyped and imputed.

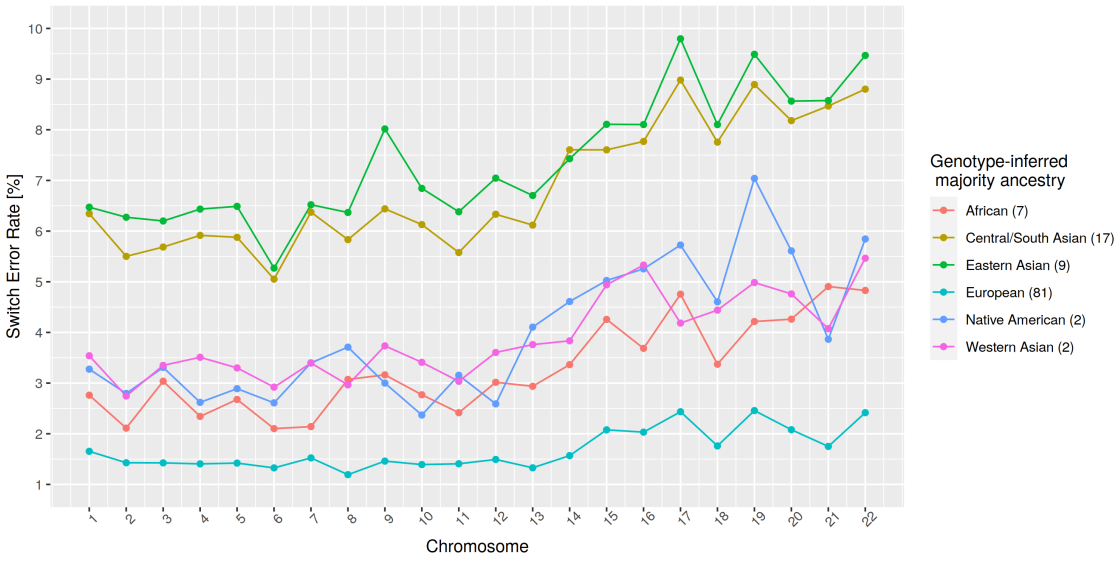

Figure 3: Phasing quality. Evaluation of phasing quality in trio children of the MGI cohort by switch error rate (SWE). SWE is summarized across several genotype-inferred majority ancestry groups. SWE across all autosomes was determined by evaluating the total number of strand switches that occurred over the total number of heterozygous sites where strand switches were possible. The value given in parentheses in the legend is the number of trio children used to estimate switch error rate.

It is defined as the Pearson correlation coefficient of known and imputed genotypes as if the known genotypes were masked. We estimated Rsq, EmpRsq, and the MAF of each variant by taking the mean of each of these values across 3 separately imputed chunks (see subsection 3.4).

Both Rsq and EmpRsq improved with increasing MAF when using either the HRC or TOPMed panel (Figure 4). Direct comparisons of imputation quality between data imputed using the HRC or TOPMed reference panels are possible at sites that are imputed by both panels. Imputation of Data Freeze 4 using the TOPMed panel resulted in a relative increase in mean Rsq and EmpRsq compared to HRC-based imputation across all MAF bins. We also evaluated Rsq at all sites that were imputed by either the HRC or TOPMed panels. Rsq of indels imputed using the TOPMed panel are comparable to Rsq of SNVs imputed with the TOPMed panel, no indels were imputed when using the HRC panel as reference.

5.4 Principal Components

We calculated the first 20 principal components for all samples in the cohort. The data were first pruned to remove all variants with a MAF $< 1\%$. Additionally, we thinned pairs of variants with a squared correlation > 0.5 within a walking window of 500 variants and a step size of 5 (PLINK). We also excluded variants in the major histocompatibility complex region (6:25000000-33500000). We used KING (v2.2.7) to identify 52,309 participants unrelated to the 2nd degree or closer and computed principal components using these samples with FlashPCA2 v2.0 [18]. The remaining 7,906 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 (46,301 unrelated & 6,753 related samples, Figure 5).

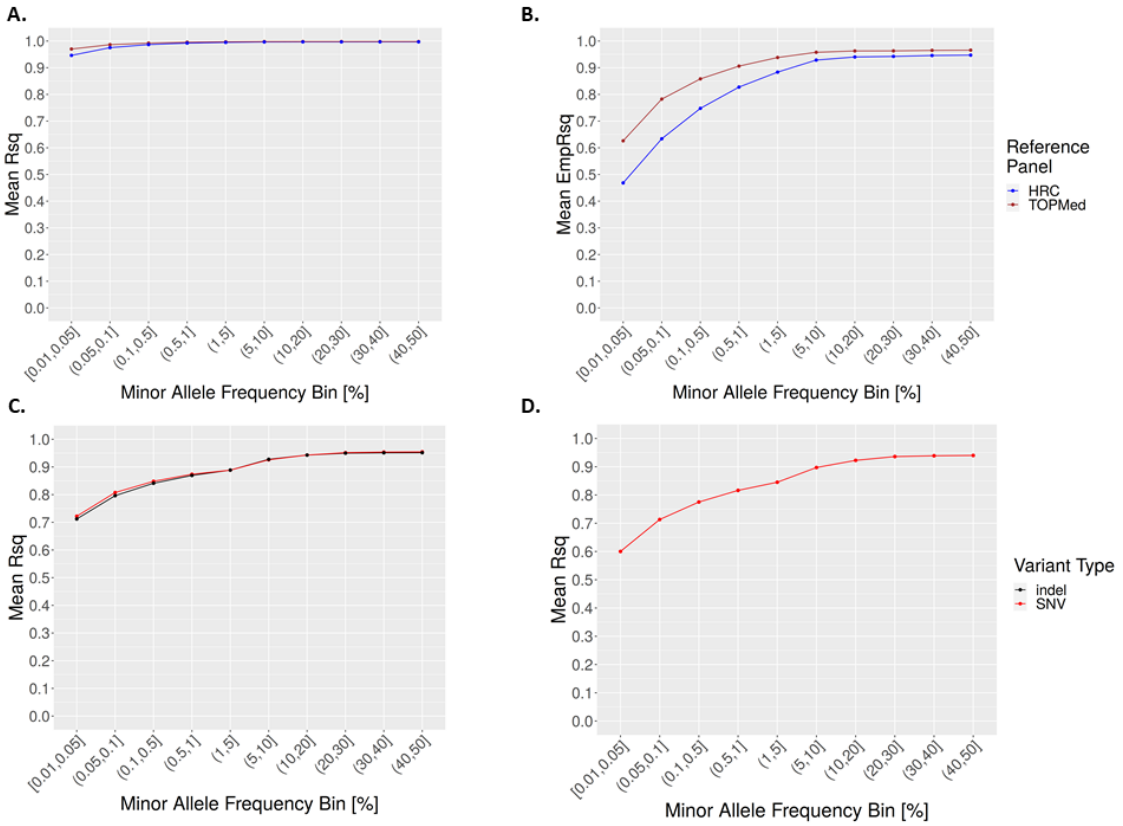

Figure 4: Imputation quality across frequency bins. Summary of imputation quality metrics for the data sets imputed with the HRC or TOPMed reference panels. 383,181 sites that were genotyped on the arrays and well-imputed (Rsq ≥ 0.3 and MAF $\geq 0.01\%$) across both reference panels were used to evaluate: **(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 well-imputed single nucleotide variants (SNVs) or indels that were imputed by using either the **(C.)** TOPMed reference panel (48,973,140 SNVs, 3,754,433 indels) or **(D.)** the HRC reference panel (32,308,248 SNVs). The given minor allele frequency bins are based on the frequency of variants as reported by the arrays.

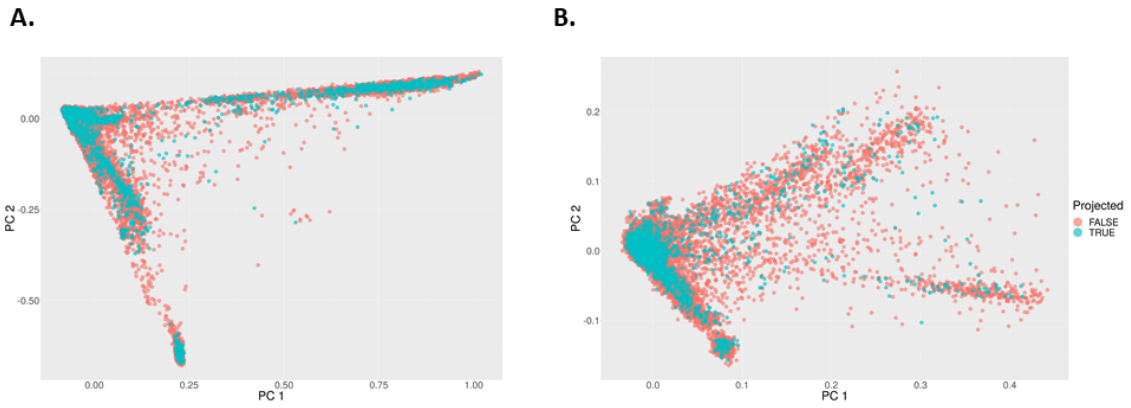

Figure 5: Principal component plots for full sample and European subsample. Plots of the first and second principal components for **(A.)** all 60,215 samples in the MGI cohort of Data Freeze 4 and **(B.)** 53,054 samples with inferred majority European ancestry by the software ADMIXTURE. For both cohorts, samples inferred to be related to the 2nd degree or closer were projected onto the principal components of unrelated samples.

References

[1] A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283 (2016).

[2] TOPMed Imputation Server. https://imputation.biodatacatalyst.nhlbi.nih.gov/#!

[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. *Nature Protocols* **9**, 2643–2662 (2014).

[5] Goldstein, J. I. *et al.* zCall: A rare variant caller for array-based genotyping: Genetics and population analysis. *Bioinformatics (Oxford, England)* **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. *Nature Genetics* **48**, 811–816 (2016).

[7] Michigan Imputation Server. https://imputationserver.sph.umich.edu/start.html#!pages/hrc-r1.1.

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

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

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

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

[12] Zajac, G. J. M. *et al.* Estimation of DNA contamination and its sources in genotyped samples. *Genetic Epidemiology* **43**, 980–995 (2019).

[13] Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575 (2007).

[14] Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research* **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 genetics* **14**, e1007308 (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. *American Journal of Human Genetics* **81**, 1084–1097 (2007).

[17] Minimac3 Info File - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/Minimac3_Info_File.

[18] Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: Principal component analysis of Biobank-scale genotype datasets. *Bioinformatics (Oxford, England)* **33**, 2776–2778 (2017).

6 Supplementary Tables and Figures

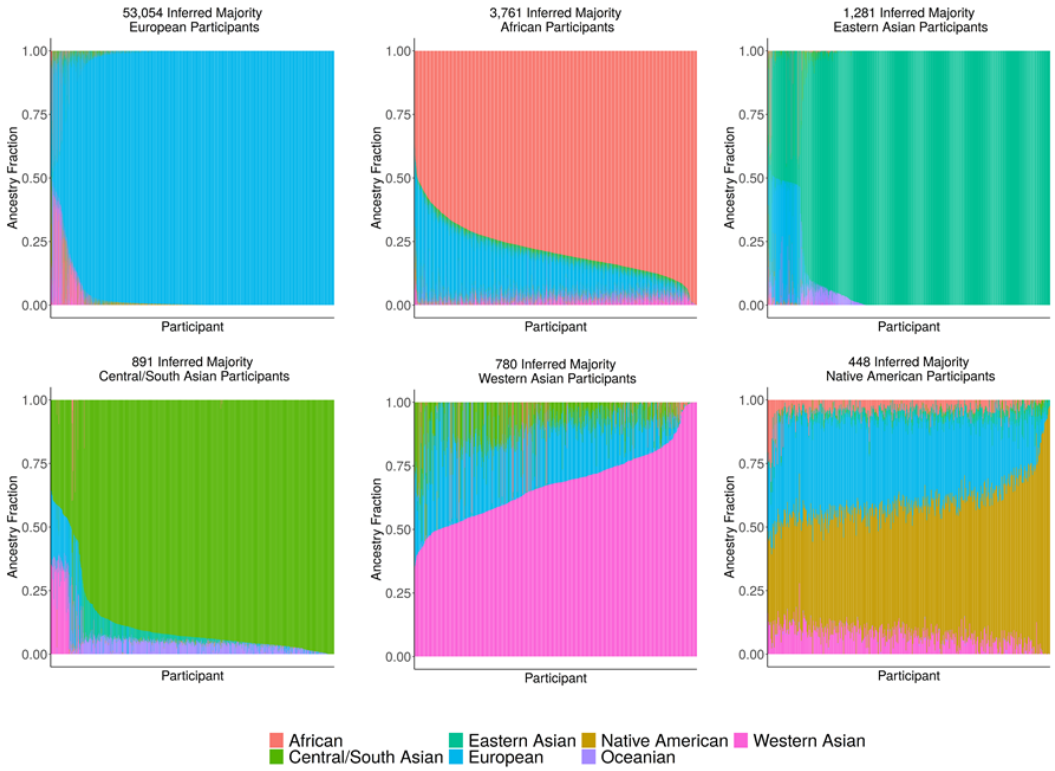

Figure S1: Genetic admixture of MGI participants. Genetic ancestry was inferred for MGI participants using the ADMIXTURE software with Human Genome Diversity Panel genotypes and continental population labels used as reference. Majority ancestry for each participant was defined as the continental population label with the largest reported Q value (ancestry fraction) from ADMIXTURE. Each inset is a stacked barplot of Q values for each participant belonging to the respective majority ancestry population.

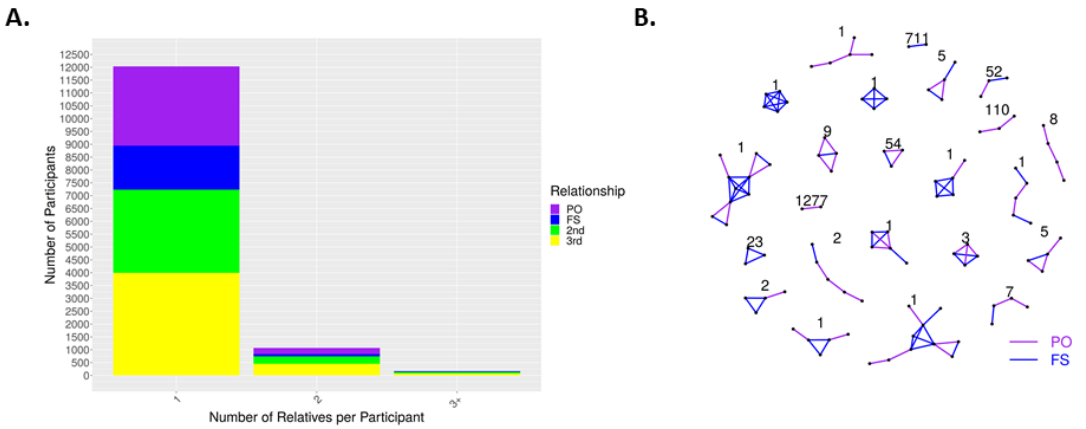

Figure S2: Inferred genetic relationships of MGI participants. **(A.)** The number of other MGI participants that each MGI participant is inferred to be related to. Relationships up to the 3rd degree are considered. **(B.)** Examples of unique family configurations in MGI that incorporate PO and FS relationships. PO, Parent-offspring; FS, full-sibling; 2nd (second-degree); 3rd (third-degree).

	# Variants (All)			# Well-imputed Variants		
Reference Panel	SNV	Indel	SNV & Indel	SNV	Indel	SNV & Indel
HRC	40,494,480	-	40,494,480	32,401,123	-	32,401,123
TOPMed	285,879,432	22,016,845	307,896,277	48,986,377	3,754,433	52,740,810

Table S1: Numbers of SNVs and indels. The number of single nucleotide variants (SNVs) and short insertion deletions (indels) in Data Freeze 4 after imputation with the TOPMed or HRC panel. Well-imputed variants are defined as sites with Rsq $\geq$ 0.3 and MAF $\geq$ 0.01%.

	Reference Panel	
MAF bin [%]	HRC	TOPMed
[0.01,0.05]	10,276,791	25,678,109
(0.05,0.1]	4,171,074	5,615,451
(0.1,0.5]	7,480,389	9,185,142
(0.5,1]	1,960,904	2,391,243
(1,5]	2,787,123	3,326,925
(5,10]	1,190,491	1,364,528
(10,20]	1,510,313	1,718,258
(20,30]	1,129,268	1,291,277
(30,40]	979,160	1,120,307
(40,50]	915,610	1,049,570

Table S2: Distribution of well-imputed variants by frequency. The number of well-imputed variants in Data Freeze 4 after imputation with the TOPMed or HRC panel. Well-imputed variants are defined as sites with Rsq $\geq$ 0.3 and MAF $\geq$ 0.01%.

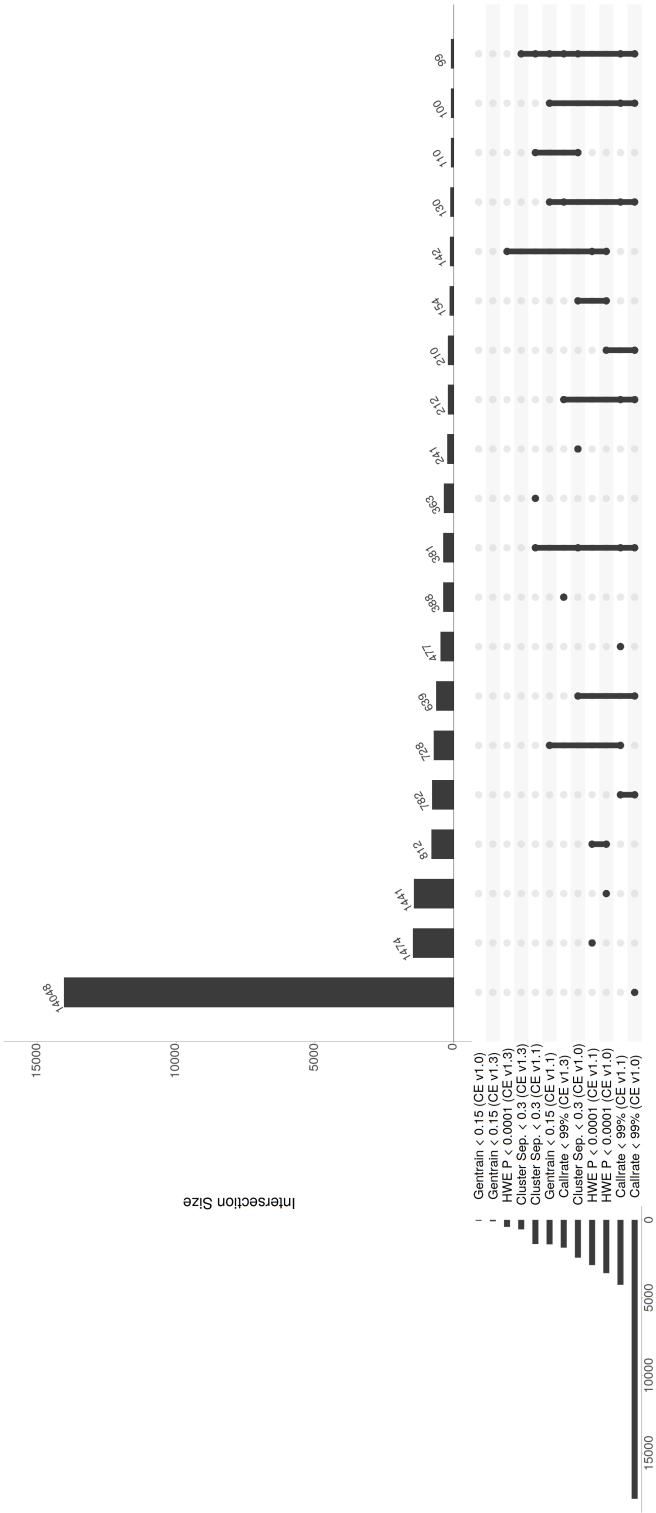

Figure S3: Array-based variant QC outcome intersections (GRCh37). Upset plot depicting the intersection of sets of variants that fail QC in the version of Data Freeze 4 mapped to the coordinates of GRCh37. Plot is sorted by intersection size with only the first 20 largest intersections shown. CE, CoreExome-24; Cluster Sep., Cluster Separation; HWE, Hardy–Weinberg equilibrium.

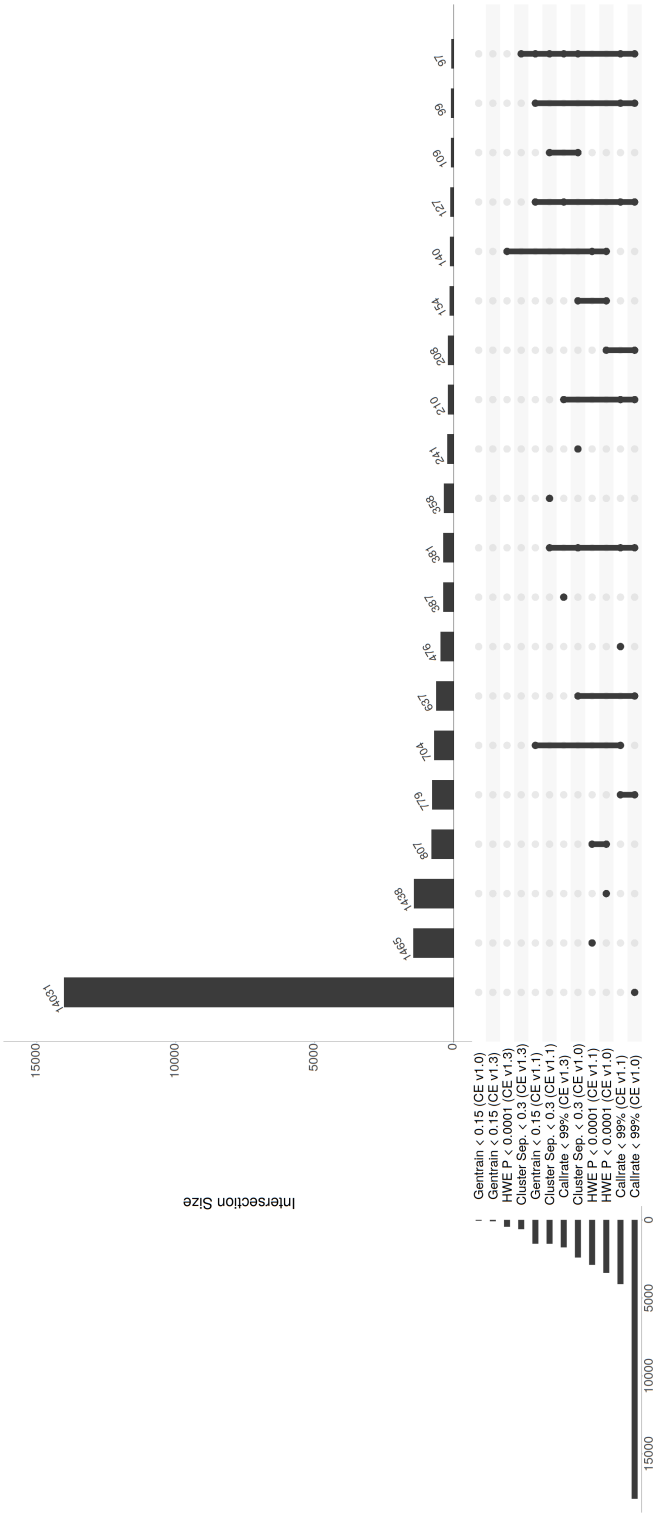

Figure S4: Array-based variant QC outcome intersections (GRCh38). Upset plot depicting the intersection of sets of variants that fail QC in the version of Data Freeze 4 mapped to the coordinates of GRCh38. Plot is sorted by intersection size with only the first 20 largest intersections shown. CE, CoreExome-24; Cluster Sep., Cluster Separation; HWE, Hardy–Weinberg equilibrium.

Variant QC Flag	# Failing Variants GRCh37			# Failing Variants GRCh38		
	CE v1.0	CE v1.1	CE v1.3	CE v1.0	CE v1.1	CE v1.3
HWE	3,405	2,888	432	3,380	2,865	423
LOW_CALLRATE	17,922	4,157	1,767	17,871	4,100	1,743
LOW_CLUSTER_SEP	2,410	1,534	590	2,395	1,513	577
LOW_GENTRAIN	37	1,553	69	36	1,510	69
FET	2,708	2,708	2,708	2,683	2,683	2,683
HWE_MERGED	58	58	58	57	57	57
MONOMORPHIC	36,680	36,680	36,680	36,588	36,588	36,588
Total Unique Failing Variants	60,497	46,740	41,579	60,311	46,539	41,431

Table S3: All variant QC. The number of well-mapping sites that fail any QC flag. Counts are given for versions of Data Freeze 4 mapped to the coordinates of GRCh37 or GRCh38. CE, CoreExome-24; HWE, Hardy-Weinberg equilibrium test $p < 10^{-4}$ before array merge; LOW_CALLRATE, call-rate < 99%; LOW_CLUSTER_SEP, Cluster Sep. score < 0.3; LOW_GENTRAIN, GenTrain score < 0.15; FET, Fisher's exact test p-value $< 10^{-3}$; HWE_MERGED, Hardy-Weinberg equilibrium test $p < 10^{-6}$ after array merge; MONOMORPHIC; minor allele frequency = 0 after array merge.