PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Data Freeze 3 Technical Notes v1.1

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

1 Overview

Data Freezes are regular releases of phased and imputed genetic data sets derived from samples of patients in the Michigan Genomics Initiative (MGI) cohort. This document provides a brief description of the properties of Data Freeze 3 released by the MGI 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).

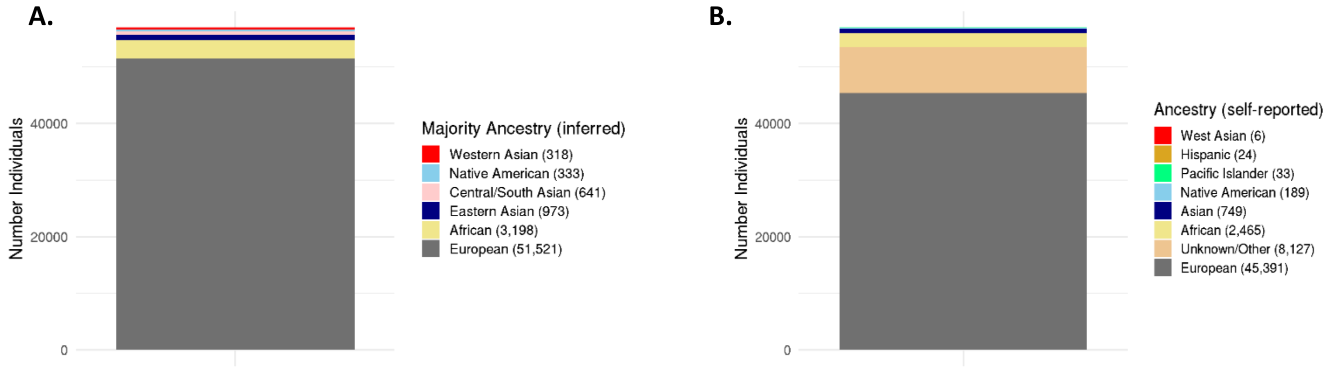

Figure 1. *Ancestry for each sample in the MGI cohort.* ***(A.)*** *Majority ancestry inferred from genetic data.* ***(B.)*** *Self-reported ancestry by study participants.*

Intermediate and Imputed Data Sets

	# Samples	*# Variants*
Genotyped Array 1.0	20,023	603,583
Genotyped Array 1.1	37,088	607,778
Merged Genotyped	56,984	502,256
Phased	56,984	502,255
Phased & Imputed unfiltered	56,984	40,457,219
Phased & Imputed filtered*	56,984	32,477,751

**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.*

After genotype imputation, the data set contains 502,255 genotyped variants and 39,954,964 imputed variants for a total of 40,457,219 variants (Table 1). Among that total, 30,029,291 variants (74%) are rare with a minor allele frequency (MAF) < 0.5% (Figure 2). Applying standard filters to remove poorly imputed variants (Rsq < 0.3) and very rare variants (MAF < 0.01%), generated a high-quality data set containing 32,477,751 variants.

The different types of data that are available with the release of Data Freeze 3 are described in Table 1. The Phased & Imputed filtered data set has 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 including appropriately flagged low-quality variants. A set of data produced by merging data from both array versions is available. These merged data have undergone both sample- and variant-level QC but are not phased. All datasets 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.

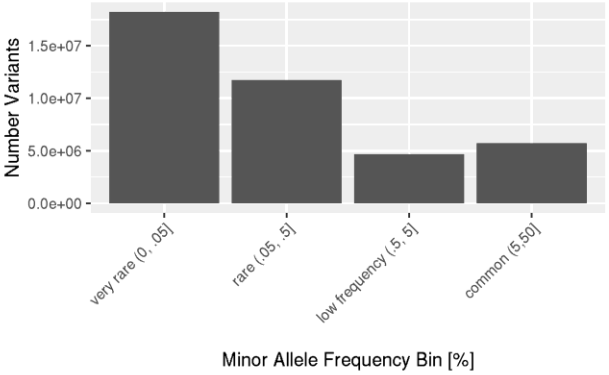

***Figure 2.** Number of variants in the phased and imputed data set (unfiltered) binned by minor allele frequency.*

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 and UM_HUNT_Biobank_v1-1_20006200_A1/Array 1.1. Both arrays were designed with the same backbones containing probes corresponding to ~570,000 total variants. This included ~ 240,000 tag single nucleotide and ~280,000 exonic variants. Custom probes corresponding to ~60,000 variants were incorporated into each array to detect candidate variants from GWASs, 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 detection of LoF variants, 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 - 1,152 samples.

To improve genotyping accuracy, all accumulated batches of samples 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 as per the GenomeStudio Genotyping Module protocol[1]. Where automatic clustering performed poorly, manual review and curation of cluster definitions was performed[2]. 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[3].

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*), data corresponding to variants that were represented on both arrays were combined. These combined data were used as input for phasing.

3.3 Phasing

Phasing was performed on the merged genotype data 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)[4]. Genetic map coordinates were inferred by using a reference genetic map of hg19 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 set by estimating genotypes that were not directly assayed on the arrays. 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)[5]. Each prepared chunk of the data set was then imputed against the Haplotype Reference Consortium r1.1 reference panel of 64,940 haplotypes[6]. 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.

3.5 Ancestry Inference

The majority ancestry of individuals corresponding to each sample was first inferred by performing principal component analysis using PLINK (v1.9)[7]. Principal component calculations were based off a reference genotype panel of Human Genome Diversity Project (HGDP) samples[8]. 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. To provide a more granular level of ancestry information, samples that were not inferred as European by projection PCA were analyzed with the software ADMIXTURE[9]. 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 at sites of sample and data production 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 gonosomal 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 analysis was performed with in-house developed R scripts. Pairwise relatedness between samples was

Samples Excluded by QC in Each Array

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

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

estimated using the relationship inference software KING (v2.1.3)[10]. Contamination between samples was estimated by the contamination detection software VICES[11]. PLINK was used to determine sample level call-rates.

4.2 Variant QC

To determine genotyping array probe specificity, probes were mapped to the Human Genome Reference Consortium Human Build 37 (GRCh37) and the revised Cambridge Reference Sequence of human mitochondrial DNA (rCRS) using the sequence alignment tool BLAT (v. 351)[12]. Variants where corresponding array probe(s) did not uniquely and perfectly map to the chromosome sequences of GRCh37 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[2]. Cluster Separation and GenTrain scores range from 0 to 1, with lower scores suggesting poor cluster separation and lower cluster quality[2]. Variants with a GenTrain score < 0.15 and/or a Cluster Separation score < 0.3 were excluded from the final data set.

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 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

Variants Excluded by QC in Each Array

			# Failing Variants	
Exclusion Flag	*Description*	*Array 1.0*	*Array 1.1*	*Both Arrays*
LOW_GENTRAIN	GenTrain score < 0.15	27	1,643	11
LOW_CLUSTER_SEP	Cluster Separation score < 0.3	1,583	718	948
LOW_CALLRATE	Call-rate < 99%	15,631	1,720	2,981
HWE_ARRAY	HWE test p < 10^{-4} within array	2,240	1,678	1,260
BATCH_EFFECT	Fisher’s exact test p < 10^{-3} between arrays	0	0	1,766
MONOMORPHIC	Minor allele frequency of 0	0	0	39,915
HWE_MERGED	HWE test p < 10^{-6} after array merge	0	0	33
	Total Variants:	***18,122***	***4,531***	***45,969***

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. The number of variants that failed on both arrays are also shown.*

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 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. Concordance was measured first before the application of variant-level QC and again after removing those variants that failed QC. Removing variants that failed QC led to increased genotype call concordance on both arrays (Table 4).

Array-based Genotype Concordance

	Pairs of Duplicates	*Pre-Variant QC Concordance*	*Post-Variant QC Concordance*
Array 1.0	153	99.74 %	99.91%
Array 1.1	304	99.91 %	99.94%

Table 4. *Concordance of genotype calls that were made for identical samples that were genotyped twice on the same array. Genotype 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)[13]. 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[14]. 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[13]. 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).

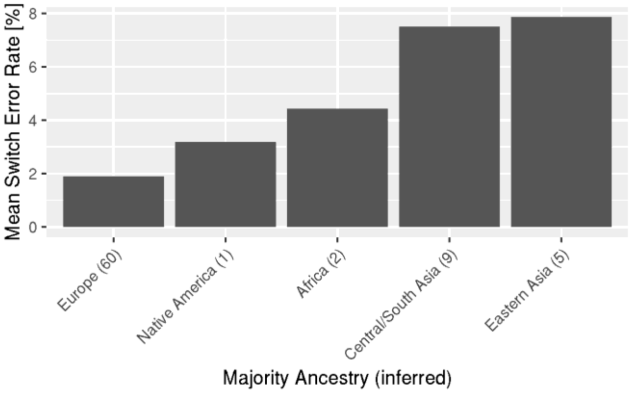

Figure 3. *Evaluation of phasing performance by switch error rate (SWE). SWE is summarized across several inferred majority ancestry groups.*

5.3 Imputation Evaluation

Imputation quality was measured by the metrics produced by the imputation software Minimac4. The metrics summarize imputation quality by estimating the correlation between imputed and expected genotypes at both all imputed sites (Rsq metric) and those sites both genotyped and imputed (Leave-one-out Rsq metric). Both quality control metrics improved with increasing MAF (Figure 4).

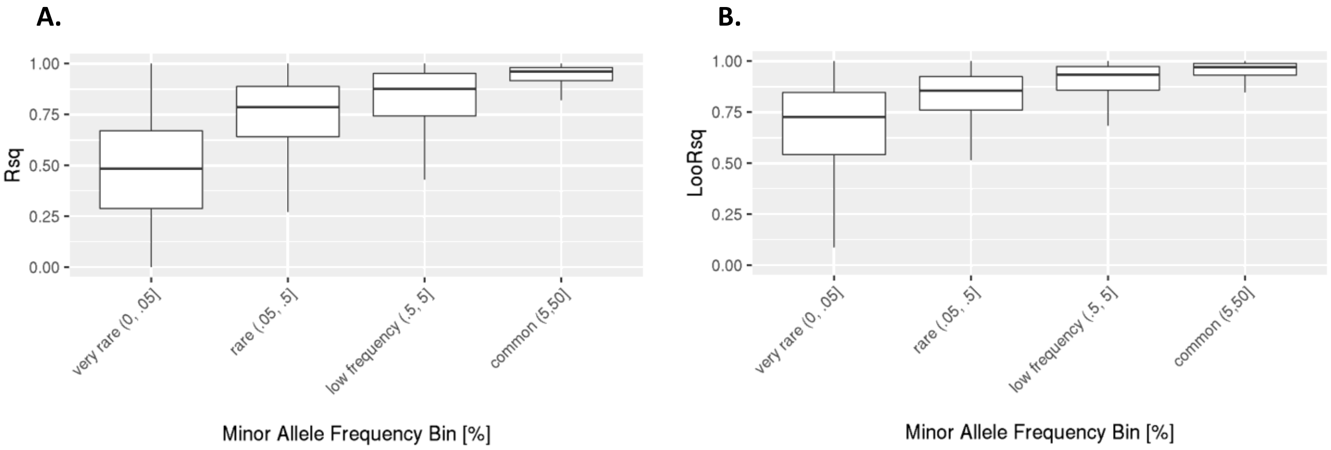

Figure 4. *Summary of imputation quality estimates for the phased and imputed data set (unfiltered) provided by the imputation software Minimac4. The estimated correlation between imputed and expected genotypes over several MAF bins at* ***(A.)*** *all imputed sites by Rsq and* ***(B.)*** *sites that were both typed and imputed by Leave-one-out Rsq (LooRsq).*

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[15]. The related samples were then projected onto the principal components of the unrelated samples. Using the same approach described above, 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).

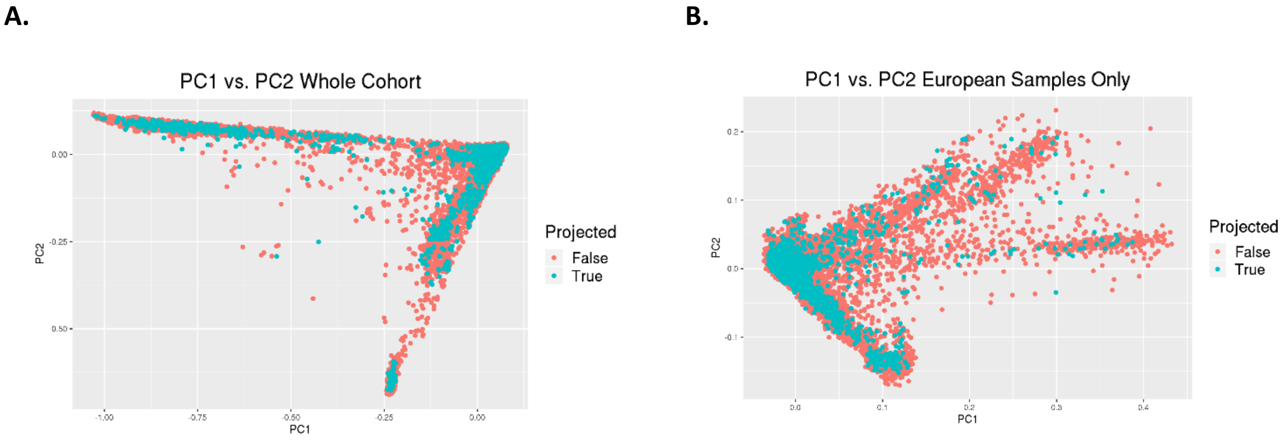

Figure 5. *Plots of the first and second principal components for* ***(A.)*** *all samples in the MGI cohort and* ***(B.)*** *those samples with majority European ancestry. For both cohorts, samples inferred to be related were projected onto the principal components of unrelated samples.*

6 References

1. GenomeStudio Documentation. https://support.illumina.com/array/array_software/genomestudio/documentation.html.
2. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
3. 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).
4. 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).
5. 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).
6. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
7. 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).
8. Stanford University. https://www.hagsc.org/hgdp/.
9. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
10. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
11. Zajac, G. J. M. *et al.* Estimation of DNA contamination and its sources in genotyped samples. *Genet. Epidemiol.* **43**, 980–995 (2019).
12. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

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

14. 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).

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