PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Freeze 6 Genome-Wide Genotypes v1.1

Brett Vanderwerff[1*], Matthew Zawistowski[1], Lars G. Fritsche[1], Emily Bertucci-Richter[1], Snehal Patil[1,2], Michael Boehnke[1], Xiang Zhou[1], and Sebastian Zöllner[1,3]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, 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 Changes From Freeze 5

- Available genotyped cohort size increased by 10,306 participants.
- Phasing of all MGI samples performed with TOPMed as reference. In Freeze 5, samples assayed on the CoreExome array were phased without a reference panel.

2 Overview of Genotyped Cohort

This data release contains genome-wide genotypes for a total of 80,529 participants included as a part of Freeze 6. The majority of participants are recruited through the Michigan Genomics Initiative (MGI) Anesthesiology Collection Effort (ACE) while awaiting inpatient surgical procedures. In addition to MGI-ACE, Freeze 6 contains some participants recruited through various strategies by the following “MGI partner studies”: Michigan Predictive Activity and Clinical Trajectories (MIPACT), Metabolism Endocrinology & Diabetes (MEND), Mental Health BioBank (MHB2), Michigan and You – Partnering to Advance Research Together (MYPART), Biobank to Illuminate the Genomic Basis of Pediatric Disease (BIGBiRD), PROviding Mental health Precision Treatment (PROMPT), Immune Precision in Solid Organ Transplantation (ImPrec), Michigan Neurological Disorders Precision Health Objective (MIND-PRO), Michigan eArly disease Progression cohort in COPD (MGI-MAP-COPD), Integration of Immune Phenotypes in Autoimmune Skin Disease (PerMIPA), Inflammatory Bowel Disease Databank (IBD-Biobank), and MGI-Dysplasia-Associated Arterial Disease Precision Health Network (MGI-DAAD). Some participants may be enrolled in more than one of the above studies.

Among participants in Freeze 6, the genotype-inferred sex was ≈ 54%(43,350 participants) female and ≈ 46% (37,179 participants) male. The median age of participants, as calculated from date of birth in the electronic health record as of January 1st 2023 or time of death, was 60 years (median of 63 years for males and 57 for females) (**Figure 1**). Data from an additional 179 participants recruited through the BIGBiRD study that were under 18 years of age are available at request.

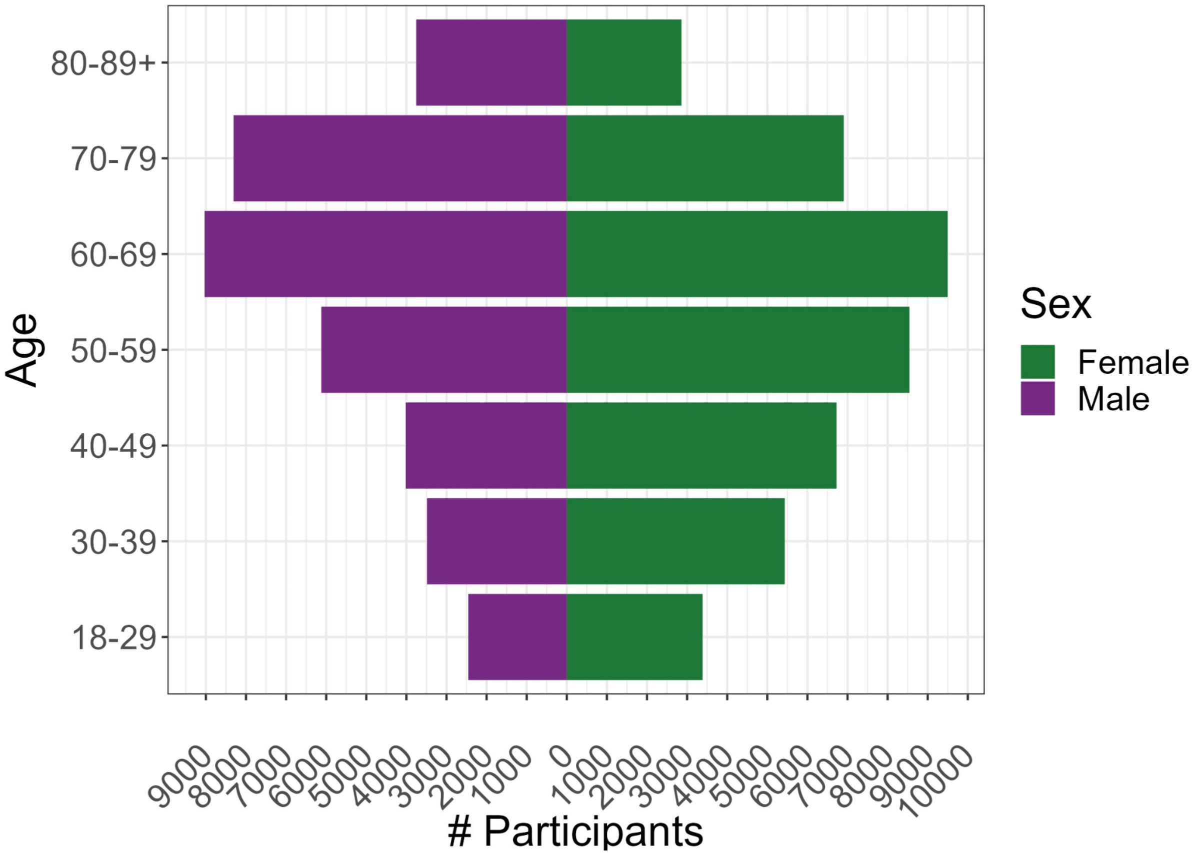


Figure 1: Age and genotype-inferred sex distribution. The distribution of genotype-inferred sex and age as calculated as of January 1st 2023 for living participants or as of deceased date for non-living participants.

The self-reported races of participants as recorded during a medical office visit consisted of Caucasian (n=69,120), African American (n=5,128), Asian (n=2,474), Other (n=1,862), Unknown (n=754), American Indian or Alaska Native (n=438), or Native Hawaiian and Other Pacific Islander (n=78). 334 participants refused to report a race and 341 had a missing value for race. The inferred majority genetic ancestry of the participants was primarily European (n=69,589) with smaller numbers of African (n=4,993), Western Asian (n=2,260), Eastern Asian (n=1,784), Central/South Asian (n=1,181), and Native American (n=722) descent (**Figure 2**).

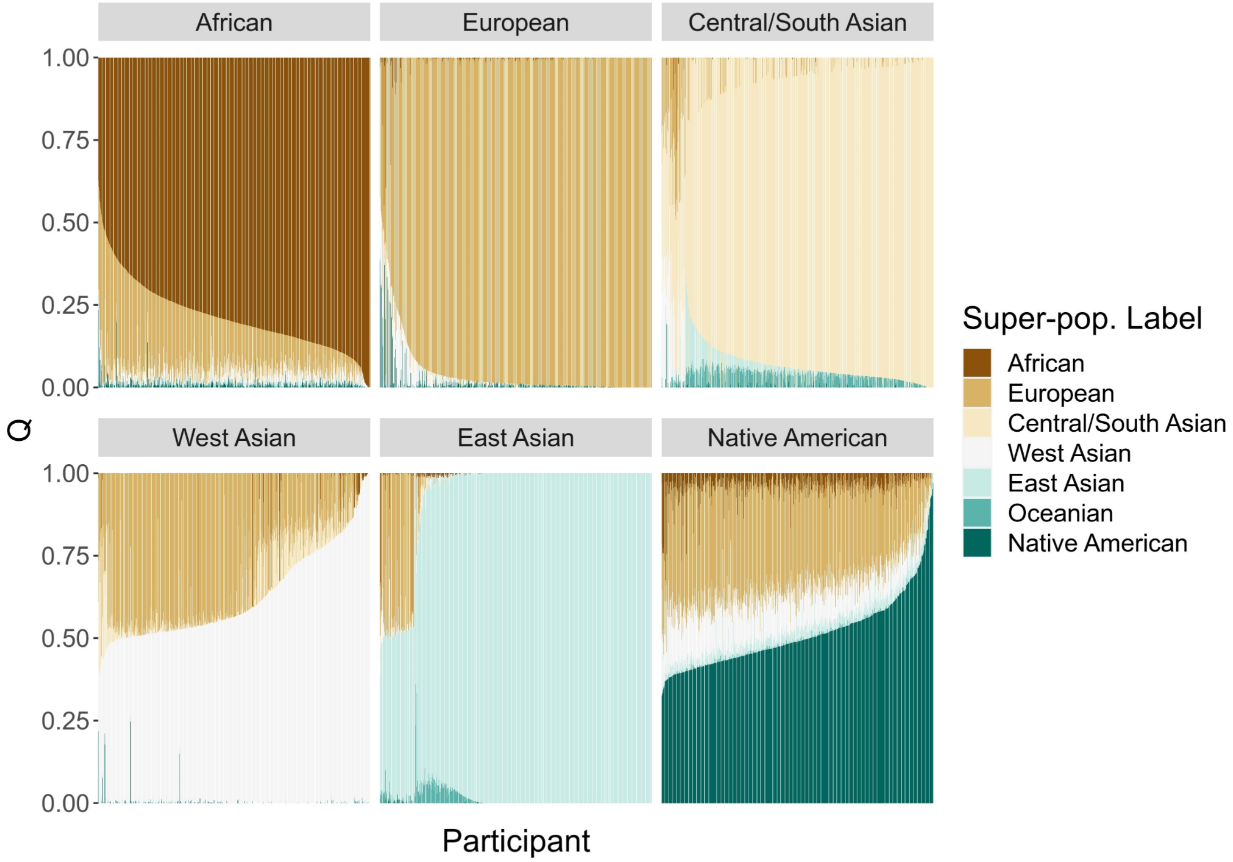


Figure 2: Genetic admixture of MGI participants. We inferred the genetic ancestry of MGI participants using the ADMIXTURE software with Human Genome Diversity Panel genotypes and super-population labels as reference. We defined the majority ancestry for each participant as the continental population 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.

3 Overview of Genetic Data

We offer genotypes experimentally determined at ≈ 570K sites for 60,715 participants by one of three versions of a customized Illumina Infinium CoreExome genotyping array and at 682,590 sites for 19,814 participants by a customized Illumina Infinium Global Screening Array (GSA). Following genotype imputation using the Trans-Omics for Precision Medicine (TOPMed) panel, Freeze 6 contains 307,726,597 variants. 285,723,035 of these variants are single nucleotide variants (SNVs) and 22,003,562 of these variants are short insertion deletions (indels). 48,739,064 SNVs and 3,739,062 indels (52,478,126 variants total) passed the standard post-imputation filters, which removed poorly imputed variants with Rsq < 0.3 and very rare variants with minor allele frequency (MAF) < 0.01%. In this filtered data-set ~81% (42,587,217) of sites had MAF ≤ 1% (**Figure 3**).

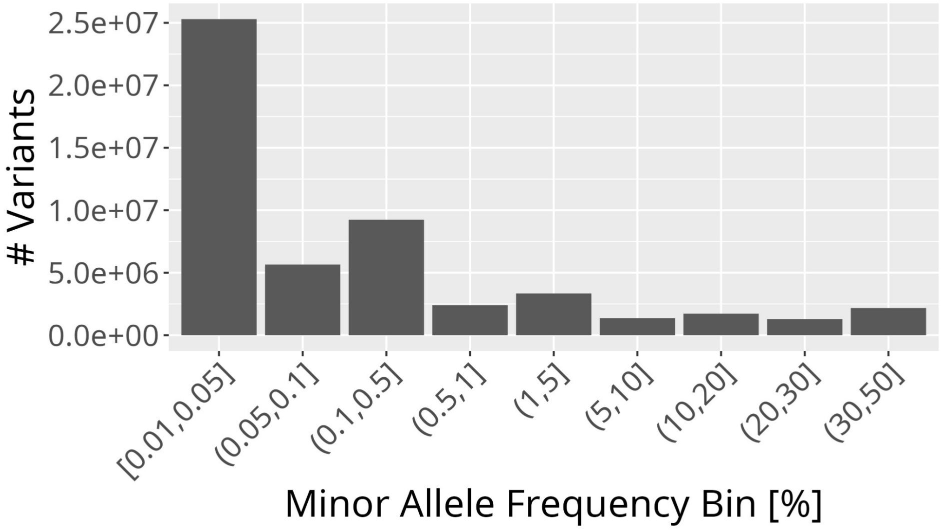


Figure 3: Counts of imputed variants. Counts of variants imputed in MGI from the TOPMed reference panel. Only variants that pass the standard post-imputation filter (Rsq ≥ 0.3 and minor allele frequency ≥ 0.01%) are plotted.

The available genotype data sets in Freeze 6 are described in **Table 1**. The standard analytic release of Freeze 6 with genotypes imputed from TOPMed and filtered by post-imputation Rsq and MAF is the most appropriate dataset for most association analyses. We also provide less processed datasets: (1) Imputed genotypes unfiltered for Rsq or MAF, (2) raw data of directly assayed genotypes for each array, where we flagged variants that failed QC filters, (3) a data set generated by merging all arrays versions after sample- and variant-level QC. All data sets are provided in VCF format and all genetic positions are in coordinates of human genome build GRCh38.

Data Set	**# Variants**	**# Participants**
CoreExome v1.0	556,441	19,834
CoreExome v1.1	566,704	37,782
CoreExome v1.3	570,520	3,099
CoreExomes v1.0-v1.3 merged*	508,367	60,715
GSA v1.3*	619,235	19,814
GSA v1.3 + CoreExomes v1.0-v1.3 merged	168,183	80,529
TOPMed imputed unfiltered	307,726,597	80,529
TOPMed imputed filtered†	52,478,126	80,529

Table 1: Genotype data available with Freeze 6. The total number of variants associated with the intermediate and imputed data sets available with the release of Freeze 6. Variant counts given for the non-imputed genotype data are counts of high quality variants that pass our quality control measures. †Variants with Rsq < 0.3 or MAF < 0.01% excluded; * versions available in both phased and unphased formats. TOPMed, Trans-Omics for Precision Medicine reference panel.

4 Data Access

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.

5 Data Production

5.1 Directly Assayed Genotypes

We genotyped biosamples collected from either blood or saliva at the University of Michigan Advanced Genomics Core (AGC) on either a customized version of the Illumina Infinium GSA-24 v1.3 or CoreExome-24 v1.0, v1.1, or v1.3.

The GSA contains fixed content corresponding to ≈ 654K variants, ≈ 85K of which are exonic. ≈ 514K variants provide genome-wide coverage and ≈ 119K represent curated clinical research variants, including variants with known disease associations, pharmacogenomics variants, and tag SNPs for HLA alleles. The remaining ≈ 10K fixed content variants were included for QC purposes, including variants for sample identification and ancestry inference. We customized our GSA by incorporating probes targeting ≈ 38K predicted Loss-of-Function (LoF) variants that were observed at least twice in individuals of the NHLBI TOPMed program, the source of reference haplotypes used to genotype impute MGI participants.[1]

The CoreExome v1.0, v1.1, and v1.3 are 3 different synthesis batches of the same array design / backbone and contain fixed content corresponding to ≈ 570K variants: ≈ 240K tag single nucleotide variants and ≈ 280K exonic variants. We added custom probes corresponding to ≈ 60K variants to each CoreExome array to detect candidate variants from genome-wide association studies (GWAS), nonsense and missense variants, ancestry informative markers, and Neanderthal variants. This custom content included probes corresponding to ≈ 30K predicted LoF variants. LoF variants require de-novo genotyping by two probe-based design. Due to a design flaw, ≈ 21K predicted LoF variants in the custom content are assayed with only a single probe. 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 QC-filtered data available with Freeze 6.

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.[2]

We performed two rounds of genotyping for most MGI samples. For sample-level QC we first called sample genotypes per automatic clustering of each sample batch processed by the AGC. At the time of Freeze creation we called genotypes in 4 separate batches for each the GSA, CoreExome v1.0, v1.1, or v1.3 arrays by automatic and joint clustering of all samples that were processed to date on each array and that passed sample QC filters, consequently a higher quality of genotype calls can be expected.

Where array-based automatic clustering performed poorly, we manually reviewed and curated cluster definitions.[3] 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.[4] Due to limited sample size, we did neither manually review cluster definitions nor perform the associated zCall work for the CoreExome v1.3 array.

5.2 Statistical Phasing

We inferred haplotypes for each participant by statistical phasing using directly assayed genotypes that passed QC filters as input for the software Eagle v2.4.[5] We tested two separate phasing approaches, one approach phasing MGI samples on a local machine without the use of a reference panel (“within-cohort” phasing) and one approach using the TOPMed Imputation Server pipeline (v1.7.3), which uses a panel of 194,512 haplotypes from whole genome sequences of diverse samples as reference.[6] We phased the final analytic MGI dataset by using the TOPMed panel as reference.

5.3 Genotype Imputation

We imputed unobserved genotypes into the phased haplotypes of MGI participants using the TOPMed reference panel available from the TOPMed Imputation Server. We implemented a sample-wise chunking approach due to server limits on sample size uploads. For samples assayed on the CoreExome array we performed genotype imputation in 3 chunks of ≈ 20K samples/chunk. In a separate batch we genotype imputed participants assayed on the GSA in a single ≈ 20K sample chunk. Directly genotyped sites with MAF < ~.5% were excluded from the imputation due to a QC step that was performed by the TOPMed Imputation Server.

Following imputation, we merged all 4 sample chunks using the Michigan Imputation Server post-processing tool “hds-util”.[7] hds-util computes estimated imputation quality (Rsq), allele frequency, and MAF in the merged data from the estimated haploid alternate allele dosage gotten across all chunks. hds-util does not compute the empirical imputation quality (ER2), but reports ER2 from each sample chunk in the merged data (delimited by a comma).We annotated the ID column of the VCF with rsids and the INFO column of the VCF with functional annotations available from dbSNP build 156 using SnpSift 5.1d[8,9]. The workflow we used to merge imputed data generated for participants assayed on the GSA or CoreExome array is described in **Figure 4**.

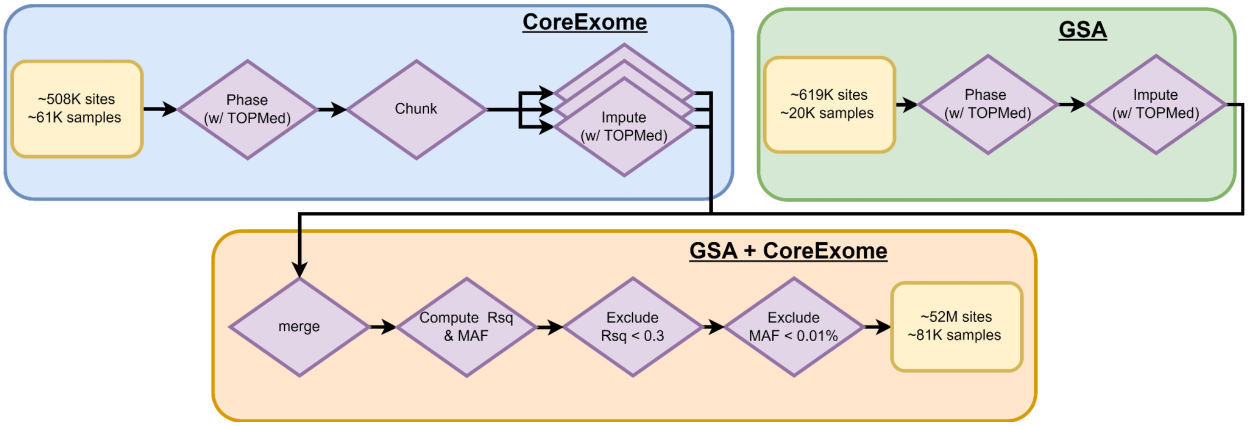


Figure 4: Phasing and imputation workflow for GSA and CoreExome arrays. We process genotypes from participants assayed on either the CoreExome or Global Screening Array (GSA) through phasing and genotype imputation in separate batches. We then merge the imputations based on each array to generate a single high-quality dataset that contains imputed genotypes from all participants included in Freeze 6. TOPMed, Trans-Omics for Precision Medicine reference panel; MAF, minor allele frequency; Rsq, estimated imputation quality.

5.4 Genetic Ancestry Inference

For the purposes of cohort description, we inferred the majority genetic ancestry of MGI participants by using the software ADMIXTURE.[10] We merged genotypes from ≈168K QC filtered sites measured across all MGI participants with those of a reference panel of Human Genome Diversity Project genotypes.[11] These merged data were analyzed by running ADMIXTURE in supervised mode using the number of Human Genome Diversity Project continental populations (K=7) as a template. We summarized genetic ancestry inferred by this method to the largest Q value (global ancestry fraction) reported by ADMIXTURE.

6 Quality Control

6.1 Sample QC

We performed sample-level QC on a rolling basis as batches of samples were genotyped. A sample was flagged per batch and excluded from the Freeze if any of the following issues were raised during sample QC: (1) participant had withdrawn from the study, (2) genotype-inferred sex had an unexplained mismatch with self-reported gender information or self-reported gender was missing, (3) sample had an atypical sex chromosomal aberration (e.g. Klinefelter syndrome), (4) sample had same genotypes but different ID of another sample, (5) sample-level call rate was below 99%, (6) sample was a duplicate of another sample with a higher call rate, (7) estimated contamination level exceeded 2.5%, (8) call rate on any individual chromosome was ≤ 95%, or (9) sample was processed in a DNA extraction batch that was flagged for technical issues. The numbers of samples excluded per sample QC criteria are described in **Figure S1**. 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).[12–14]

6.2 Variant QC

To determine genotype array probe specificity, we mapped probes to the sequences of GRCh38 and the revised Cambridge Reference Sequence of human mitochondrial DNA (rCRS) using the sequence alignment tool BLAT (v.351).[15] We excluded variants where the corresponding array probe(s) did not uniquely and perfectly map to the chromosome sequences of GRCh38, or the rCRS reference.

For variants assayed on each of the CoreExome and GSA arrays, we assigned quality control flags and excluded sites if (1) Hardy-Weinberg Equilibrium exact test (HWE) $p < 1e\text{-}4$ in a sub-population of MGI participants with majority European genetic ancestry that were inferred to be unrelated to the 2nd degree by KING (v2.1.3), (2) GenomeStudio “GenTrain” score < 0.15, (3) GenomeStudio “Cluster Separation” score < 0.3, or (4) call rate was less than 98%. For variants assayed on the CoreExome array we additionally flagged and excluded sites that had a call rate between 98% and 99% in at least two of the three array synthesis batches.

For a subset of high-quality variants on the CoreExome and GSA, we observed a large difference between the alternate allele frequency in the European unrelated sample of MGI and the deeply sequenced genomes from ancestry matched 1000 Genomes Project samples (**Figure 5**).[16] Thus, we flagged and excluded variants assayed on either the GSA or CoreExome arrays where the alternate allele frequency deviated from that observed in the 1000 Genomes Project samples by +/- 10%.

To merge data of the 3 CoreExome arrays, we first flagged and excluded ≈ 2.5K variants with $p < 1e\text{-}3$ in any pairwise comparisons of allele frequency between CoreExome array versions 1.0, 1.1, or 1.3 with Fisher’s exact test. We then merged data across the CoreExome arrays by inner join. Following the merge we flagged and excluded 43 variants with HWE $p < 1e\text{-}6$ in a subset of individuals with majority European ancestry that were inferred to be unrelated to the 2nd degree.

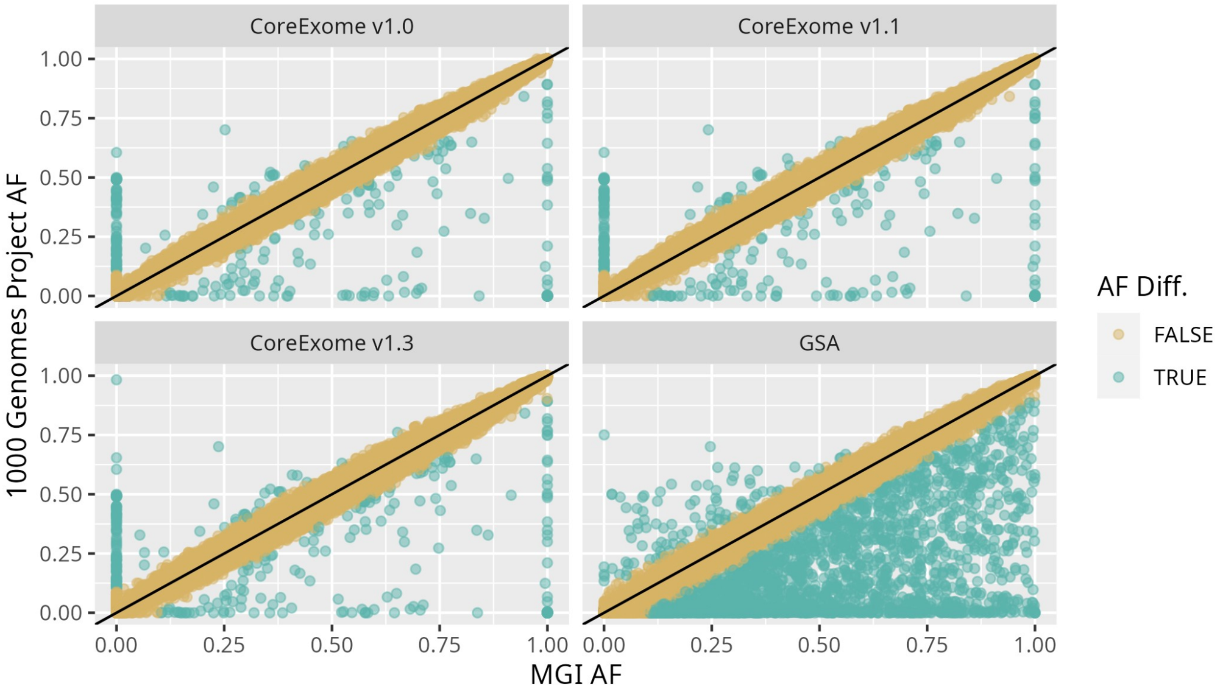


Figure 5: Allele frequency of MGI and 1000 Genomes Project samples. Alternate allele frequency (AF) among high-quality sites assayed in the European unrelated sample of MGI with the CoreExome arrays or GSA compared to the AF observed in the high coverage whole genome sequence data of European unrelated 1000 Genomes Project samples. Points filled with green color indicate that the AF difference between these data sets was larger than +/-10%. Diagonal line indicates y=x.

We generated a high-quality genotype dataset that included all MGI participants in Freeze 6 by combining variants from the merged CoreExome arrays and the GSA after excluding 116 variants with p < 1e-4 when evaluating allele frequency between the array types with Fisher’s exact test. Following the merge we flagged and excluded 32 variants with HWE p < 1e-6 in a subset of individuals with majority European ancestry that were inferred to be unrelated to the 2nd degree. An overview of the workflow we used to apply variant QC to the GSA and CoreExome arrays is described in **Figure 6** and numbers of sites excluded per variant QC criteria applied to each array are described in **Figure S2.**

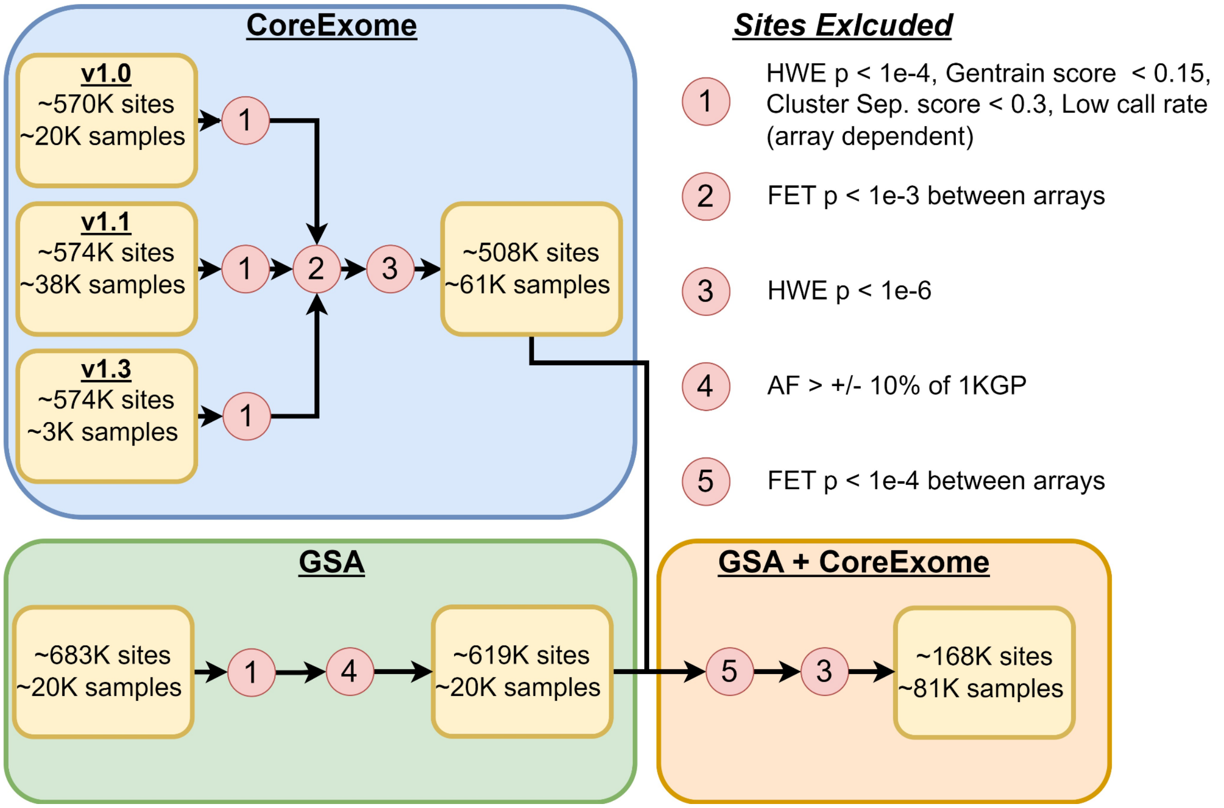


Figure 6: Workflow for genotype array variant QC. We apply variant QC to generate several sets of high-quality genotype data from the CoreExome arrays, the Global Screening Array (GSA), or merges of the GSA and CoreExome arrays. The red numbered circles represent steps taken to exclude sites based on variant QC. HWE, Hardy-Weinberg Equilibrium exact test; FET, Fisher's exact test; AF, alternate allele frequency; 1KGP, 1000 Genomes Project.

6.3 Genotype Imputation QC

Following merging the separately imputed sample chunks with hds-util, we applied a filter to exclude poorly imputed sites with Rsq < 0.3 and very rare variants with MAF < .01%.

7 Quality Evaluation

7.1 Genotype Quality

To determine the accuracy of directly assayed genotypes we measured genotype concordance for each array using 139, 319, 318, and 86 samples that were genotyped twice on the CoreExome v1.0, CoreExome v1.1, CoreExome v1.3 arrays or GSA v1.3, respectively. We considered genotypes concordant if they matched perfectly between duplicate samples. We evaluated concordance across a set of all genotypes (overall concordance) and a set of only those genotypes where at least one sample of the duplicate pair had a non-reference-homozygote call (non-reference concordance). We measured concordance before and after removing variants that failed QC. For all arrays, removing variants that failed QC increased genotype call concordance (**Table 2**).

		Overall Concordance		**Non-Reference Concordance**	
Array	**# Pairs**	**Pre-QC**	**Post-QC**	**Pre-QC**	**Post-QC**
CoreExome v1.0	139	99.8	99.90	99.68	99.85
CoreExome v1.1	319	99.87	99.93	99.84	99.92
CoreExome v1.3	318	99.85	99.91	99.76	99.90
GSA v1.3	86	99.79	99.88	99.68	99.84

Table 2: 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.

7.2 Phasing Quality

To evaluate an optimal phasing approach we compared the inferred phase quality from phasing MGI samples with and without a reference panel. We selected 5,000 participants with majority European genetic ancestry that were assayed on either the CoreExome or the GSA (10,000 samples total). We selected an additional 3,969 participants with majority African genetic ancestry that were assayed on the CoreExome array (we did not select African samples assayed on the GSA due to low sample size). For each selected dataset we phased chromosome 2 in two separate runs, once phasing without the use of a reference panel and once phasing with the TOPMed panel as reference. To estimate phasing quality we compared imputation quality measured as the square of the Pearson correlation coefficient (EmpRsq, see below) between known and imputed genotypes at sites with MAF ≥ .01%, expecting that higher phase quality will translate to higher EmpRsq.

Imputation quality was higher when phasing MGI with TOPMed as reference compared to phasing MGI without the use of a reference panel. We observed the largest quality differences in Africans genotyped on the CoreExome array among sites with MAF < 1%, in Europeans genotyped on the CoreExome array among sites with MAF < .5%, and in Europeans genotyped on the GSA among sites with MAF < .1% . When phasing with the TOPMed reference panel, we observed more notable imputation quality gains among African ancestry participants assayed on the CoreExome array relative to European ancestry participants assayed on the CoreExome array, particularly among common variants with MAF >1% **(Figure 7)**. Based on these results we phased the final analytical set of MGI data using the TOPMed panel as reference.

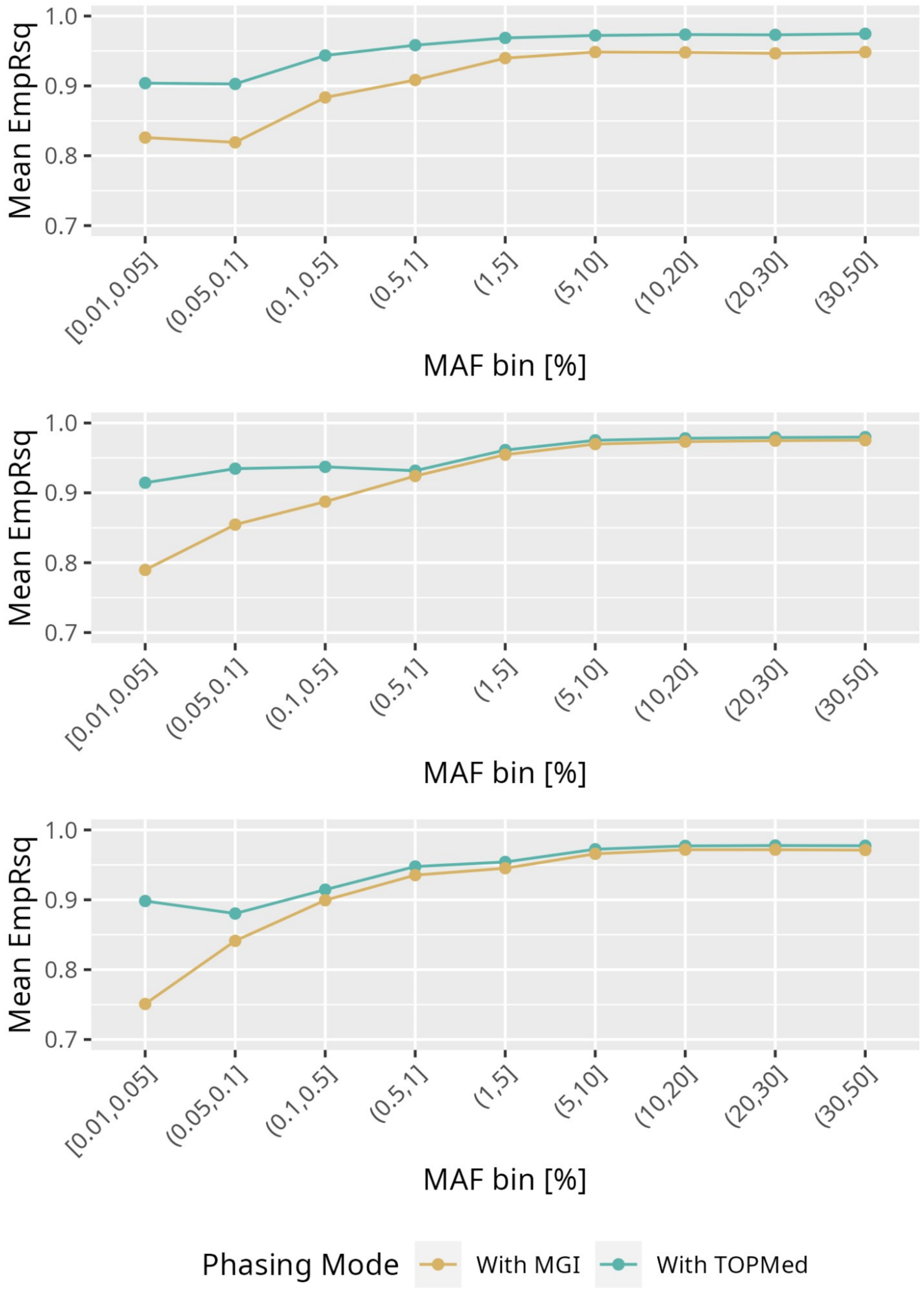


Figure 7: Estimation of phasing quality. Estimation of phasing quality from phasing MGI within cohort or using the Trans-Omics for Precision Medicine (TOPMed) reference panel. We used the Pearson correlation coefficient of known and imputed genotypes (EmpRsq) as proxy for phasing quality among **(top panel)** participants with majority African genetic ancestry genotyped on the CoreExome array, **(middle panel)** participants with majority European genetic ancestry genotyped on the CoreExome array, and **(bottom panel)** participants with majority European genetic ancestry genotyped on the Global Screening Array.

7.3 Genotype Imputation Quality

We used the ”Rsq” and ”EmpRsq” metrics produced by the genotype imputation software Minimac4 to evaluate imputation quality.[17] The Rsq metric estimates imputation accuracy at all imputed sites by the formula:

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

where $\hat{p}$ is the estimated 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.[18] The EmpRsq metric measures imputation quality at all sites that were both genotyped and imputed. It is defined as the square of the Pearson correlation coefficient of known and imputed genotypes as if the known genotypes were masked. When using either the GSA or CoreExome arrays as input for genotype imputation, mean Rsq improved with increasing MAF and mean EmpRsq was > .9 for all MAF bins evaluated. Mean EmpRsq was generally slightly higher for sites assayed on the GSA (**Figure 8**).

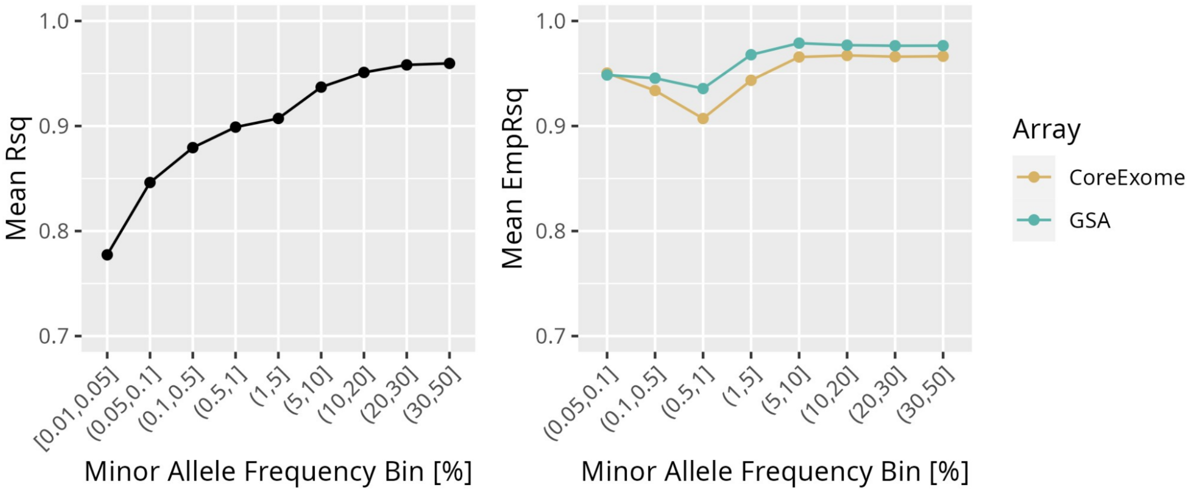


Figure 8: Imputation quality. (Left panel) the estimated correlation between imputed and expected genotypes (Rsq) for 52,478,126 sites that were imputed from the Trans-Omics for Precision Medicine reference panel with Rsq ≥ 0.3 and MAF ≥ 0.01%. Here Rsq is calculated from the estimated haploid alternate allele dosage using hds-util following the merge of separately imputed sample chunks from the GSA and CoreExome arrays. **(Right panel)** the Pearson correlation coefficient of known and imputed genotypes (EmpRsq) for the CoreExome array and the GSA at 147,469 sites that intersected both arrays and were imputed from the TOPMed panel with Rsq ≥ 0.3 and MAF ≥ 0.01%.

7.4 Principal Components

We calculated the first 20 genetic principal components (PCs) for samples of all participants included in Freeze 6. We pruned data to remove all variants with a MAF < 1% before thinning pairs of variants with a squared correlation > 0.5 within a walking window of 500 variants and a step size of 5 (PLINK). We used KING (v2.2.7) to identify participants unrelated to the 3rd degree or closer and computed PCs using these samples with FlashPCA2 v2.0.[19] We then projected the remaining samples from related participants onto the PC space generated from samples of unrelated participants. Using the same approach that was applied to samples of all participants, we generated a second set of PCs for only those samples from participants with inferred majority European genetic ancestry (**Figure 9**). We offer to compute study-specific PCs at the request of investigators.

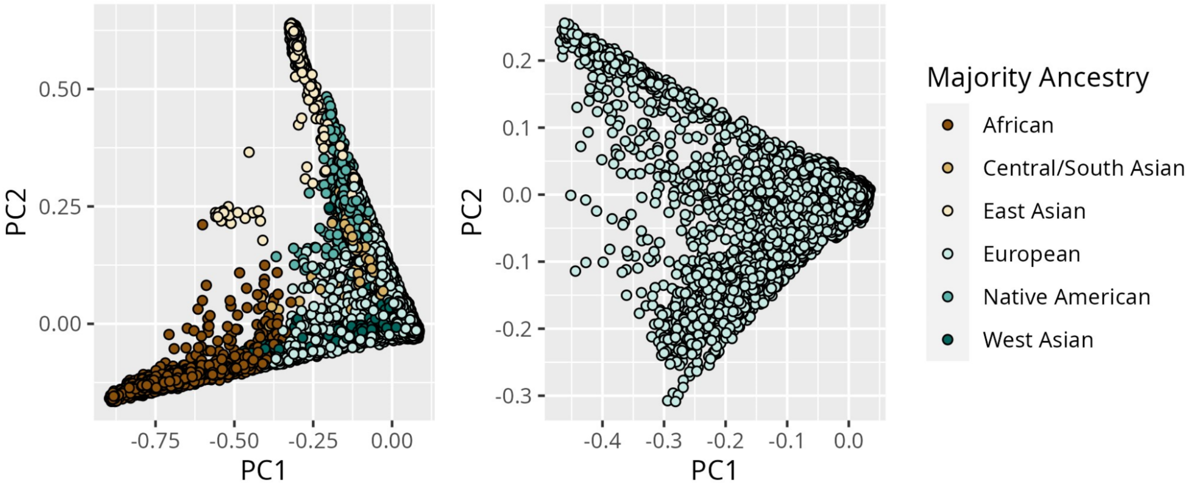


Figure 9: Principal component analysis. Principal components (PC) 1 and 2 computed from samples of **(left)** all participants included in Freeze 6 and **(right)** participants with inferred majority European ancestry. Each point is colored according to the majority genetic ancestry inferred using the ADMIXTURE software with Human Genome Diversity Panel genotypes as reference.

7.5 Empirical Comparison of Meta- and Joint-analysis

Two options for analyzing data from participants genotyped on either the CoreExome arrays or GSA together are 1.) meta-analysis of summary statistics from analyses run on imputed genotype data from each array family or 2.) joint-analysis of imputed genotype data pooled across all arrays. While meta-analysis is expected to perform similarly to joint-analysis[20], it requires more computational steps, thus we sought to empirically evaluate meta- and joint-analysis approaches in Freeze 6 to determine how substitutable these approaches are.

We converted International Classification of Diseases (ICD)-9 and ICD-10 codes collected for 69,257 European ancestry participants included in Freeze 6 to phecode phenotypes using the R PheWAS package with min.code.count = 1 (v0.99.5-5)[21], of these we selected 10 phecode phenotypes (**Table 3**) where we observed genome-wide significant ($p < 5e-8$) signals in a previous MGI Freeze.[22]

Phenotype	**Category**	**# Cases**	**Case Control Ratio**
Celiac disease	Digestive	725	1:57
Primary hypercoagulable state	Hematopoietic	1,273	1:43
Disorders of bilirubin excretion	Endocrine/metabolic	1,377	1:44
Hypoglycemia	Endocrine/metabolic	2,128	1:19
Type 1 diabetes	Endocrine/metabolic	3,500	1:13
Cancer of prostate	Neoplasms	4,304	1:5
Breast cancer	Neoplasms	4,543	1:12
Atrial fibrillation	Circulatory system	7,949	1:4
Iron deficiency anemias	Hematopoietic	8,161	1:5
Asthma	Respiratory	13,752	1:3

Table 3. Phenotypes evaluated by meta- and joint-analysis. The names and categories in addition to the number of cases and the case:control ratio for 10 phecode phenotypes evaluated by meta- and joint-analysis.

To perform meta-analysis, we first ran GWAS using SAIGE (v 44.6.4) on imputed genotype data collected from participants assayed on the CoreExome array or GSA separately.[23] We evaluated variants with Rsq ≥ 0.3 and MAF ≥ 0.01% in both datasets and included covariates for age as of January 1st 2023 for living participants or as of deceased date for non-living participants, genotyping array version, genotype-inferred sex, and the first 4 European genetic PCs. For each phecode phenotype, we then meta-analyzed the pair of summary statistics generated from SAIGE by running METAL in inverse variance weighted mode.[24]

To perform joint-analysis, we ran GWAS as described above with the exception that we provided imputed genotype data pooled from participants assayed on either the CoreExome array or GSA as input for SAIGE.

We compared meta-analysis and joint-analysis p-values and betas at all sites with p-value < .05 in either the meta- or the joint-analysis. -log10(p-value) and beta concordance between each approach increased with MAF and had high concordance with R^2 > .99 among sites with MAF > 1% for both p-values and betas (**Figure 10**). The median -log10(p-value) among sites with MAF ≤ 1% was 1.60 for the joint-analysis and 1.52 for the meta-analysis. The median -log10(p-value) among sites with MAF > 1% was 1.61 for the joint-analysis and 1.60 for the meta-analysis.

Taken together, these results suggest that for variants with MAF > 1%, jointly analyzing participants genotyped on either the GSA and CoreExome array performs nearly identical to meta-analyzing separate GWAS performed on each of the array families. For variants with MAF ≤ 1%, the meta-analysis and joint-analysis results are more discordant. On this basis we recommend that users testing variants with MAF > 1% in GWAS use joint-analysis and to consider meta-analysis when evaluating rarer variants.

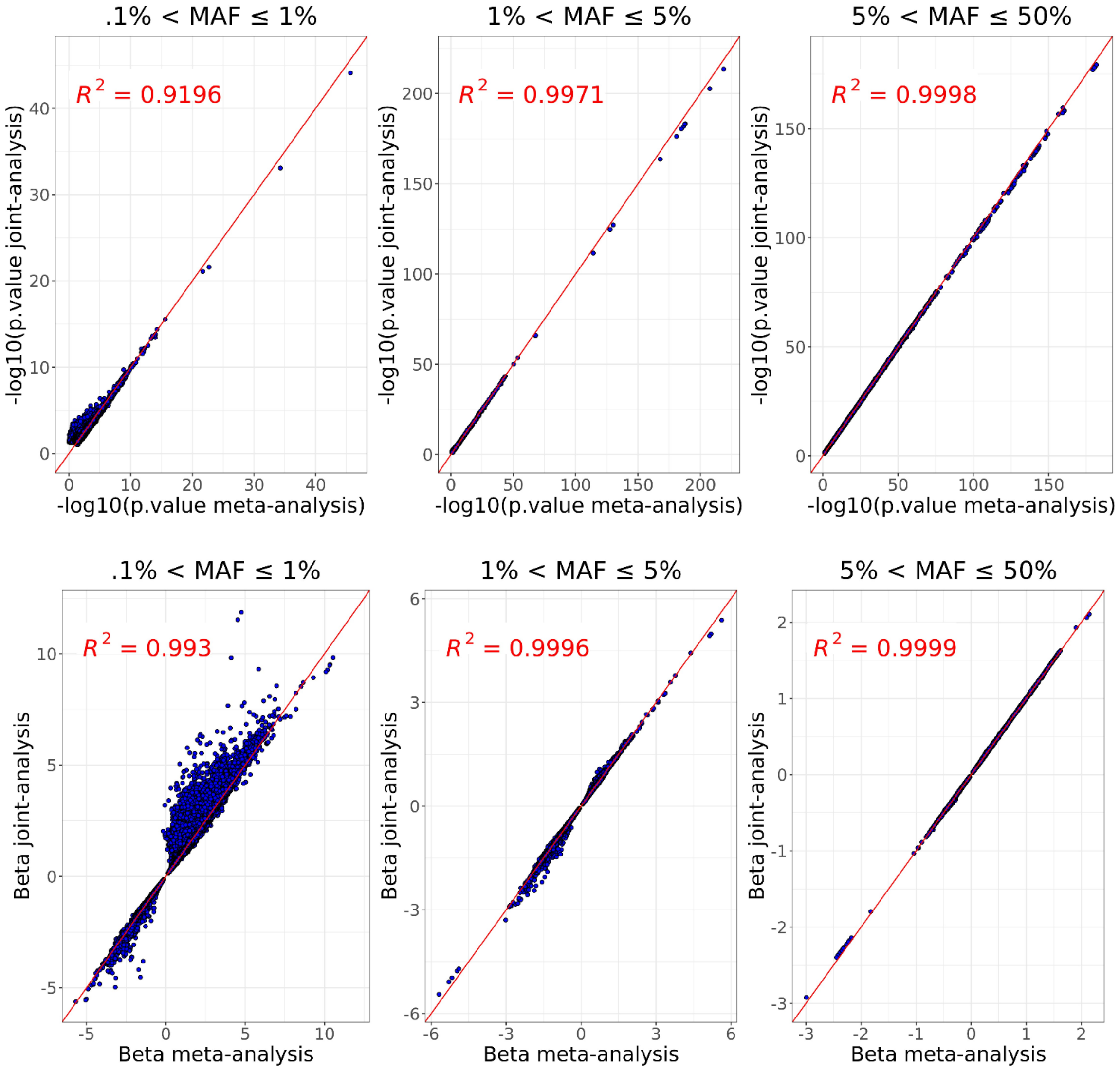
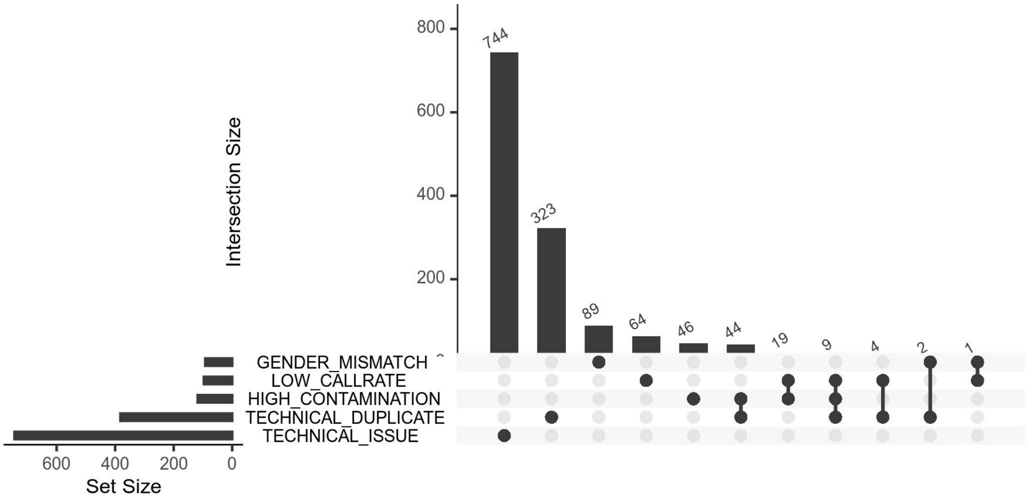


Figure 10. Comparison of p-values and betas from meta- and joint-analysis. Results for -log10(p-values) **(top row)** and betas **(bottom row)**. Points plotted are any hit with a p-value < .05 in either the joint-analysis or meta-analysis across any of the 10 phecode phenotypes evaluated. R^2 is the square of the Pearson correlation coefficient between meta- and joint-analysis. MAF, minor allele frequency.

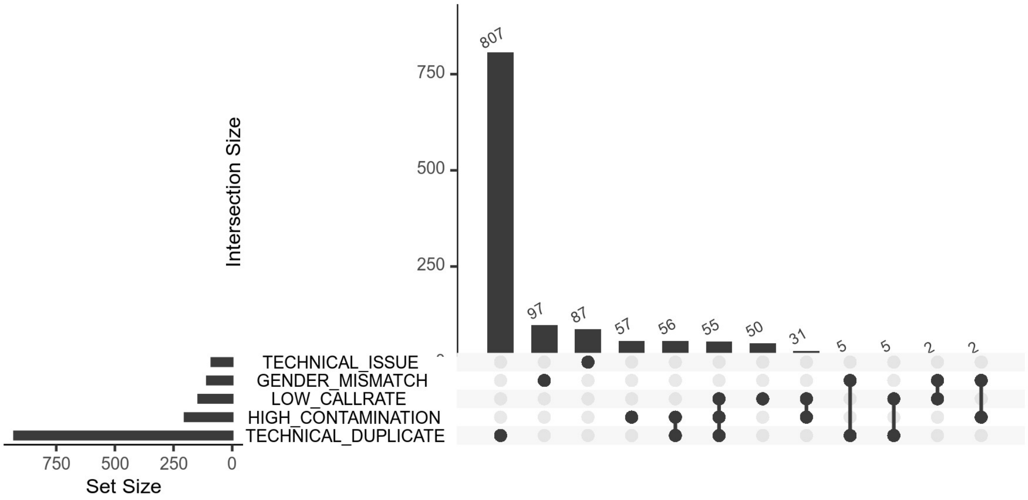
8 Supplementary Information

The following shows numbers of samples excluded per sample QC criteria described in section 6.1.

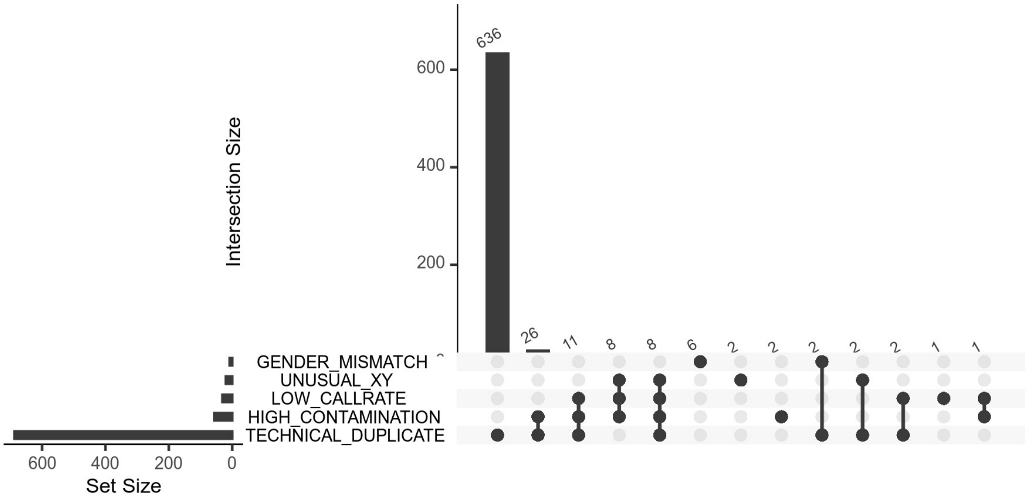
A.)



B.)



C.)



D.)

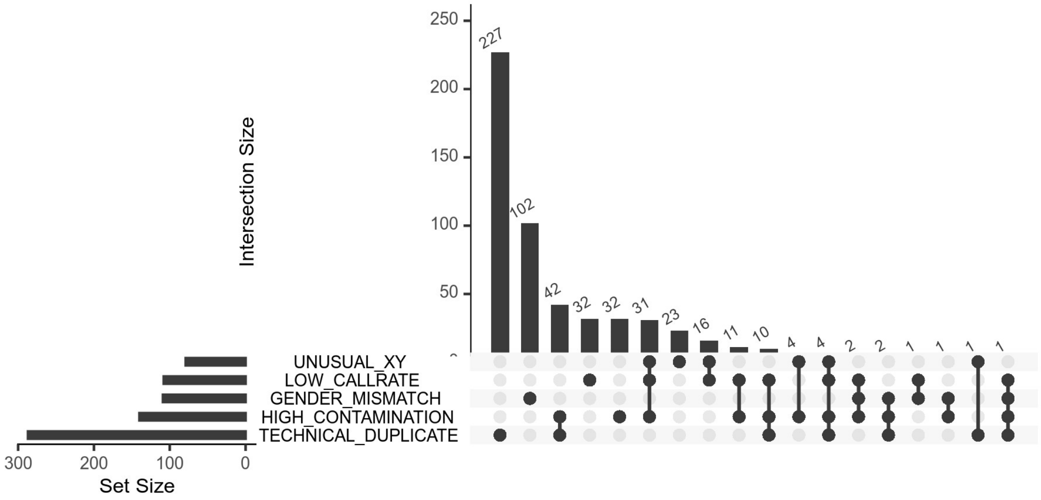
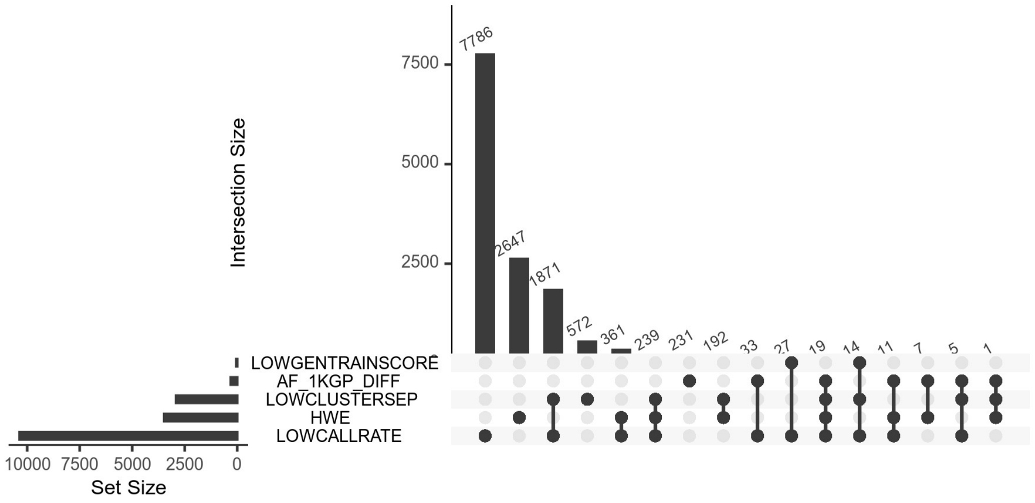


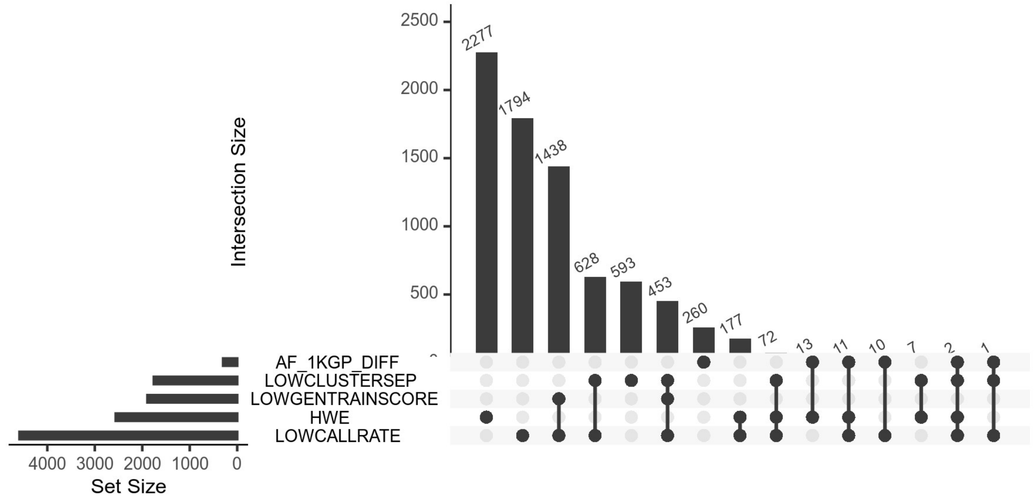
Figure S1: Sample QC outcomes. UpSet plot of the number of samples that fail various sample QC measures for the **A.)** CoreExome v1.0, **B.)**, CoreExome v1.1, **C.)**, CoreExome v1.3, or **D.)** Global Screening Array v1.3. TECHNICAL_DUPLICATE, sample with same ID and similar genotypes as other sample; TECHNICAL ISSUE, excluded DNA extraction batch; HIGH_CONTAMINATION, estimated contamination > 2.5 %; LOW_CALLRATE, sample call rate < 99%; GENDER_MISMATCH, reported gender different from genotype-inferred sex; UNUSUAL_XY, unusual XY composition. Only the first five largest sets are plotted.

The following shows numbers of variants excluded per variant QC criteria described in section 6.2.

A.)



B.)



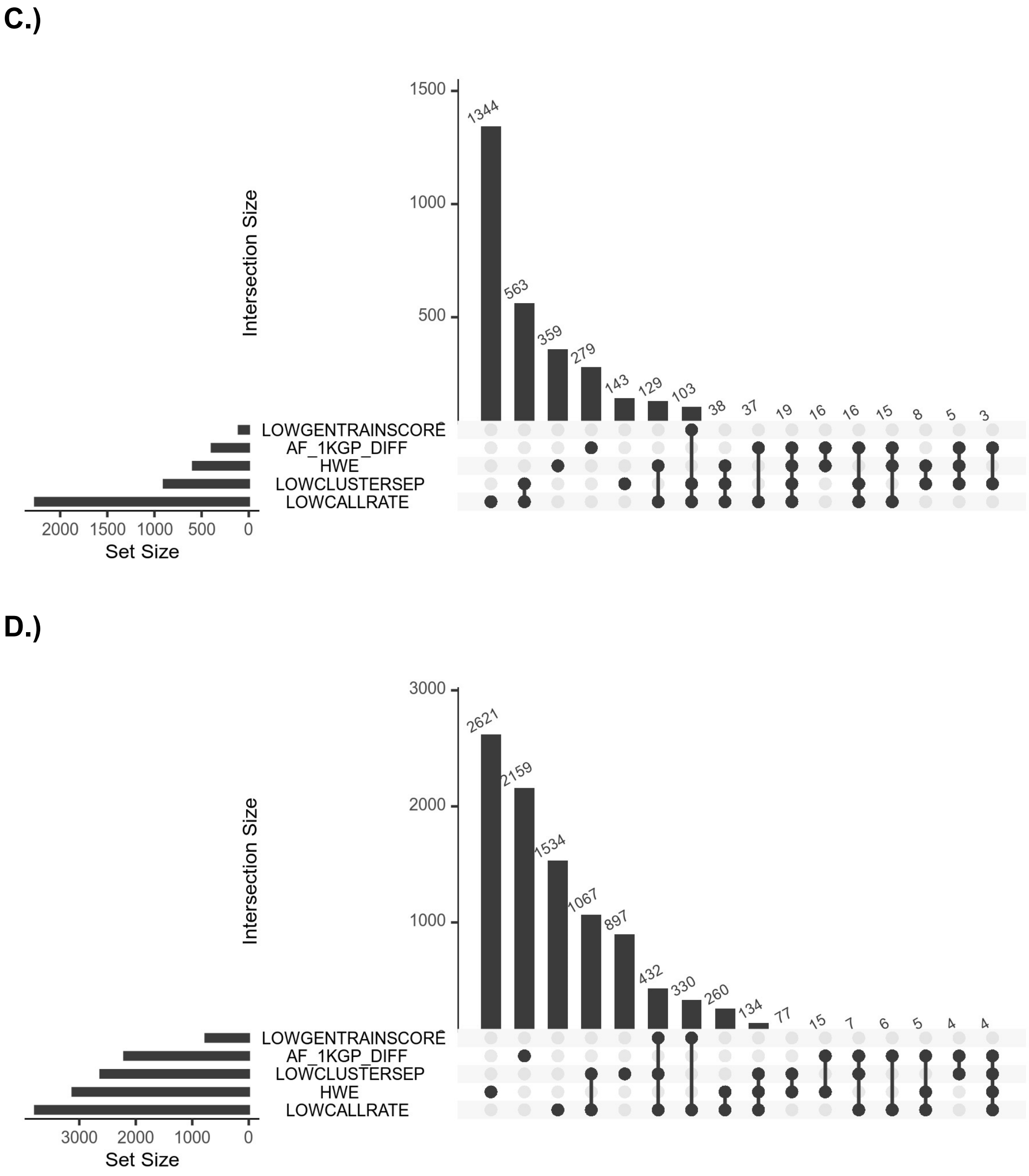


Figure S2: Variant QC outcomes. UpSet plot of the number of well-mapping sites that fail variant QC measures in the **A.)** CoreExome v1.0, **B.)**, CoreExome v1.1, **C.)**, CoreExome v1.3, or **D.)** Global Screening Array v1.3. LOWCALLRATE, call rate < 98% for all arrays or call rate between 98 and 99% in two of three CoreExome array versions; HWE, Hardy-Weinberg equilibrium test p < 1e-4 before array merge; LOWCLUSTERSEP, Cluster Sep. score < 0.3; LOWGENTRAINSCORE, GenTrain score < 0.15; AF_1KGP_DIFF, greater than +/- 10% difference in alternate allele frequency compared to 1000 Genomes Project samples. Only the first five largest sets are plotted.

9 References

1. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
2. GenomeStudio Documentation. https://support.illumina.com/array/array_software/genomestudio/documentation.html (2020).
3. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat Protoc* **9**, 2643–2662 (2014).
4. Goldstein, J. I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012).
5. 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).
6. TOPMed Imputation Server. https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about.
7. statgen/hds-util. https://github.com/statgen/hds-util (2023).
8. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
9. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9**, 677–679 (1999).
10. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
11. Human Genome Diversity Project. https://www.hagsc.org/hgdp/ (2020).
12. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

13. 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).

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

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

16. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

17. 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).

18. Minimac3 Info File - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/Minimac3_Info_File (2021).

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

20. Lin, D. Y. & Zeng, D. Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data. *Genet Epidemiol* **34**, 10.1002/gepi.20435 (2010).

21. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

22. MGI PheWeb. https://pheweb.org/MGI-freeze3/ (2022).

23. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).

24. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).