PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Freeze 5 Genome-Wide Genotypes

Brett Vanderwerff[1,2,*], Lars G. Fritsche[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: brettva@umich.edu

1 Changes From Freeze 4

- Available genotyped cohort size increased by 10,270 participants
- Genetic data for the newly added participants were collected on our new customized Illumina Global Screening Array. See section 7.5 for recommendations for analyzing these data with the legacy CoreExome arrays

2 Overview of Genotyped Cohort

Freeze 5 contains genome-wide genotypes for a total of 70,439 participants. These participants come from the Michigan Genomics Initiative (MGI, n=59,828) study and the following MGI partner studies: Michigan Predictive Activity and Clinical Trajectories (MIPACT, n=6,065), Metabolism Endocrinology & Diabetes (MEND, n=2,793), Mental Health BioBank (MHB2, n=666), Michigan and You – Partnering to Advance Research Together (n=605), Biobank to Illuminate the Genomic Basis of Pediatric Disease (BIGBiRD, n=306), PROviding Mental health Precision Treatment (n=154), Immune Precision in Solid Organ Transplantation (n=16), and Michigan Neurological Disorders Precision Health Objective (n=6).

Among participants in Freeze 5, the genotype-inferred sex was 37,358 (≈ 53%) females and 33,081 males. The median age, as calculated from date of birth in electronic health record as of January 1st 2022 or time of death, was 60 years (median of 62 years for males and 57 for females). 173 participants were under 18 years of age (**Figure 1**). The self-reported race of participants as recorded during a medical office visit consisted of Caucasian (n=60,598), African American (n=4,561), Unknown (n=2,946), Asian (n=1,910), American Indian or Alaska Native (n=355), Native Hawaiian and Other Pacific Islander (n=69) (**Figure 2**). The inferred majority genetic ancestry of the participants was primarily

European (n=61,113) with smaller numbers of African (n=4,450), Western Asian (n=1,885), Eastern Asian (n=1,426), Central/South Asian (n=963), and Native American (n=602) descent (**Figure 2**).

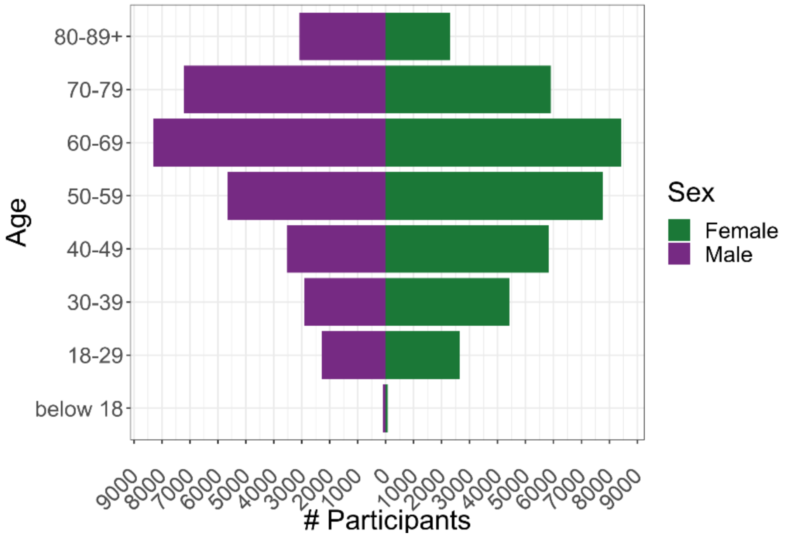

Figure 1: Age and sex distribution. The distribution of genotype-inferred sex and age as calculated as of January 1st 2022 for living participants or as of deceased date for non-living participants.

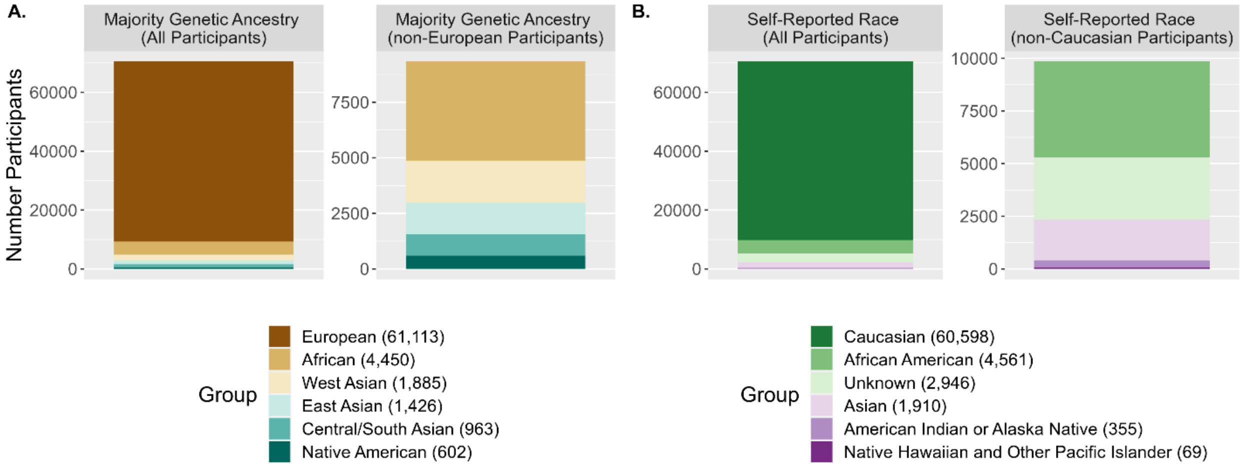

Figure 2: Genotype-inferred majority ancestry and self-reported race. (A.) Majority ancestry as inferred for MGI participants using the ADMIXTURE software with Human Genome Diversity Panel genotypes and continental population labels used as reference. **(B.)** Race as self-reported by MGI participants during a medical office visit.

3 Overview of Genetic Data

We offer genotypes experimentally determined at ≈ 570K sites for 60,176 participants by one of three versions of a customized Illumina Infinium CoreExome genotyping array and at 682,590 sites for 10,263 participants by a customized Illumina Infinium Global Screening Array (GSA). Following genotype imputation using the Trans-Omics for Precision Medicine (TOPMed) panel, Freeze 5 contains 307,883,040 variants. 285,866,195 of these variants are single nucleotide variants (SNVs) and 22,016,845 of these variants are short insertion deletions (indels). 46,873,824 SNVs and 3,589,605 indels (50,463,429 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 ≈ 80% (40,588,573) of sites had MAF ≤ 1% (**Figure 3**).

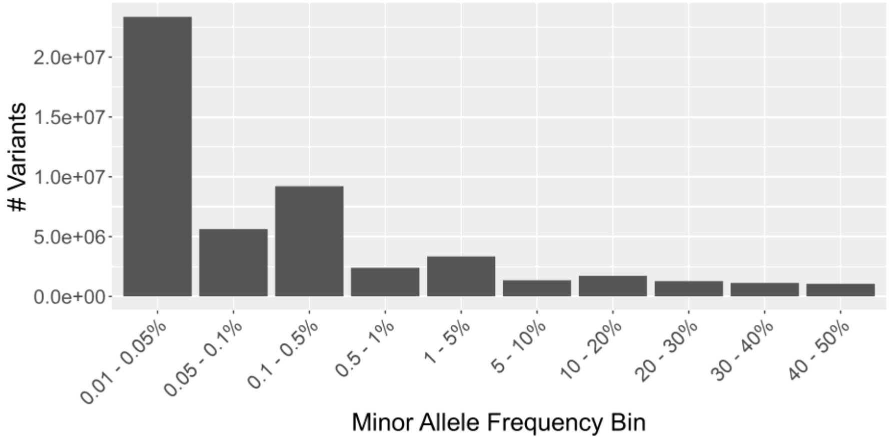

Figure 3: Frequency of imputed variants. Frequency distribution of variants imputed in MGI from the TOPMed reference panel. Only variants that pass the standard post-imputation filter (Rsq ≥ 0.3 and minor allele frequency ≥ 0.01%) are plotted.

The available genotype data sets in Freeze 5 are described in **Table 1**. Most analyses should use the standard release of Freeze 5 with genotypes imputed from TOPMed, filtered by post-imputation Rsq and MAF. A release of these data unfiltered for Rsq or MAF is also available. Raw data of directly assayed genotypes are also available for each array, where we flagged variants that failed QC filters. We also offer data sets generated by inner join of all CoreExome array versions or by inner join of both the CoreExome array versions and the GSA 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	570,506	19,826
CoreExome v1.1	574,490	37,921
CoreExome v1.3	573,648	2,429
CoreExomes v1.0-v1.3 merged*	498,710	60,176
GSA v1.3*	682,590	10,263
GSA v1.3 + CoreExomes v1.0-v1.3 merged	159,750	70,439
TOPMed imputed unfiltered	307,883,040	70,439
TOPMed imputed filtered†	50,463,429	70,439

Table 1: Genotype data available with Freeze 5. The total number of variants associated with the intermediate and imputed data sets available with the release of Freeze 5. †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. Custom probes corresponding to ≈ 60K variants were added 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 5.

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 using directly assayed genotypes that passed QC filters. We phased participants assayed on the CoreExome arrays jointly by running Eagle (v 2.4.1) in non-reference mode. In a separate batch we phased participants assayed on the GSA using the TOPMed Imputation Server pipeline (v1.6.6), which runs Eagle v2.4 in reference mode with a panel of 194,512 haplotypes from diverse samples[5,6].

5.3 Genotype Imputation

We imputed unobserved genotypes into the phased haplotypes of directly assayed genotypes based on a large reference panel of whole genome sequences. We imputed participants assayed on the CoreExome array using minimac4 v1.0.2 within v1.5.7 of the TOPMed Imputation Server[7]. Due to server limits on sample size, we performed imputation in 3 chunks of ≈ 20K samples/chunk. We then merged the separate imputations and updated the MAF, Rsq, and EmpRsq fields as the average of the values from each imputation chunk. In a separate batch we genotype imputed participants assayed on the GSA array using minimac4 v1.0.2 within v1.6.6 of the TOPMed Imputation Server. The workflow we used to merge imputed data generated for participants assayed on the CoreExome or GSA array is described in **Figure 4**.

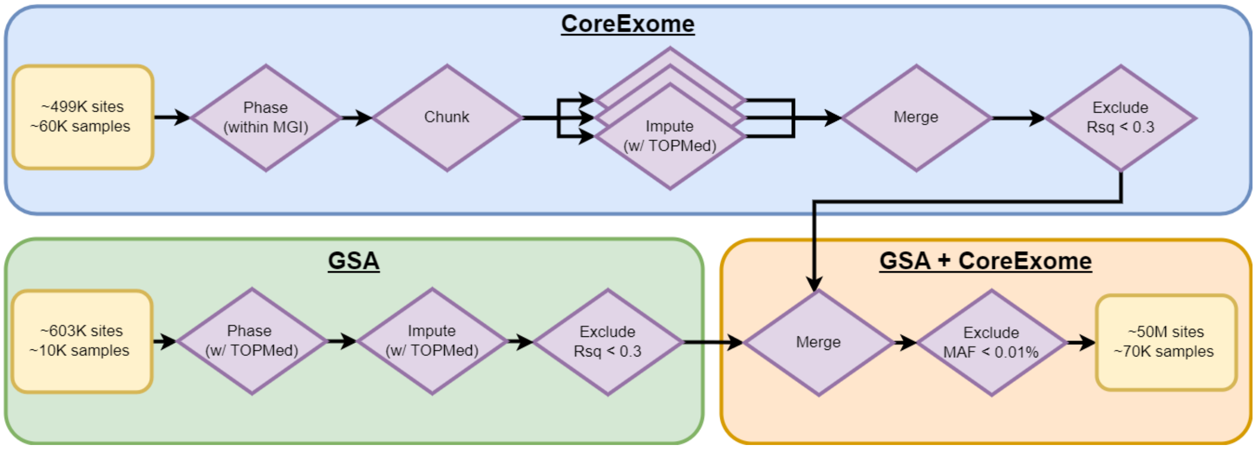

Figure 4: Phasing and imputation workflow for GSA and CoreExome arrays. We process genotypes from participants assayed on either the CoreExome or Global Screening Array (GSA) through phasing and genotype imputation in separate batches. We then merge the imputations based on each array to generate a single high-quality dataset that contains imputed genotypes from all participants included in Freeze 5. 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[8]. We merged genotypes from ≈ 160K QC filtered sites measured across all MGI participants with those of a reference panel of Human Genome Diversity Project genotypes[9]. 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. Genetic ancestry inferred by this method was summarized 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 did not match the available self-reported gender information of the participant 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 (**Figure 5**). 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)[10–12].

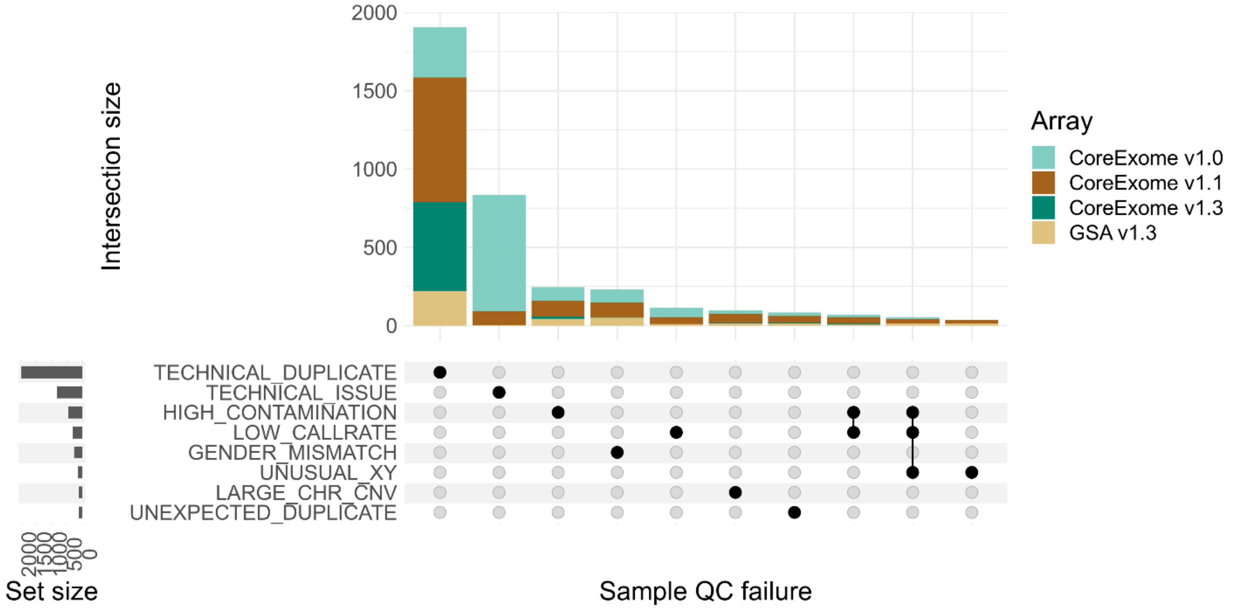

Figure 5: Sample QC outcomes. UpSet plot of the number of samples that fail various sample QC measures. TECHNICAL_DUPLICATE, sample with same ID and similar genotypes as other sample; TECHNICAL ISSUE, excluded DNA extraction batch; HIGH_CONTAMINATION, estimated contamination > 2.5 %; LOW_CALLRATE, sample call-rate < 99%; GENDER_MISMATCH, reported gender different from genotype-inferred sex; UNUSUAL_XY, unusual XY composition; LARGE_CHR_CNV, chromosomal call-rate drop > 5 %; UNEXPECTED_DUPLICATE, sample pair w/ different IDs & similar genotypes. The first 10 largest intersections are plotted.

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)[13]. 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 99%.

To merge data of the 3 CoreExome arrays, we first flagged and excluded ≈ 2.7K variants with $p < 1e\text{-}4$ in any pairwise comparisons of allele frequency between CoreExome array version 1.0, 1.1, or 1.3 with Fisher’s exact test. We then merged data across the CoreExome arrays by inner join and flagged and excluded 57 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.

For a subset of high-quality variants assayed on the GSA (HWE $p \geq 1e\text{-}4$, GenTrain score ≥ 0.15, Cluster Separation score ≥ 0.3, and call rate ≥ 99%), we observed a large difference between the alternate allele frequency in the European unrelated sample of MGI and the deeply sequenced genomes from 1000 Genomes Project samples (**Figure 6**)[14]. Thus, we flagged and excluded 2,223 variants assayed on the GSA where the alternate allele frequency difference between these data sets was larger than +/- 10%. We did not apply this variant QC step to the CoreExome array as the alternate allele frequency for only 139 variants was larger than +/- 10% that of 1000 Genomes Project samples, which seems more consistent with sampling variation.

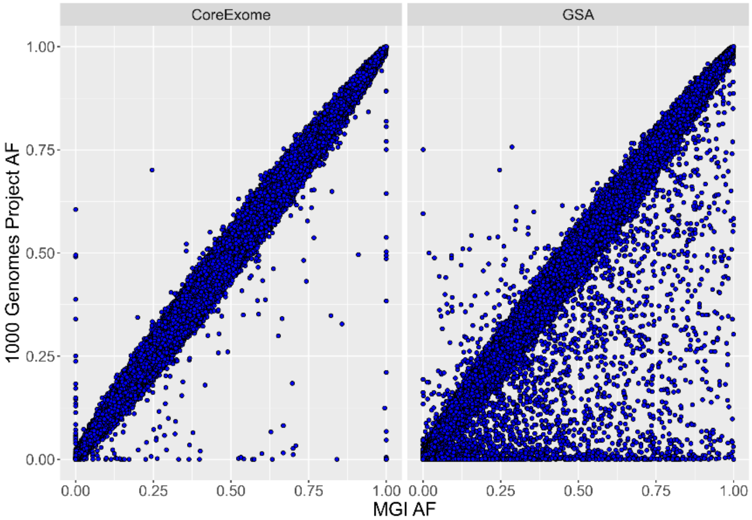

Figure 6: Allele frequency of MGI and 1000 Genomes Project samples. Alternate allele frequency (AF) among high-quality autosomal sites with Hardy-Weinberg equilibrium test $p \geq 1e\text{-}4$, call-rate ≥ 99%, Cluster Sep. score ≥ 0.3, and GenTrain score ≥ 0.15 assayed in the European unrelated sample of MGI with the CoreExome array **(left)** or GSA **(right)** compared to the AF observed in the high coverage whole genome sequence data of European unrelated 1000 Genomes Project samples.

We generated a high-quality genotype dataset that included all MGI participants in Freeze 5 by combining variants from the merged CoreExome arrays and the GSA after excluding 211 variants with p < 1e-3 when evaluating allele frequency between the array types with Fisher's exact test. An overview of the workflow we used to apply variant QC to the GSA and CoreExome arrays is described in **Figure 7** and numbers of sites excluded per variant QC criteria applied to each array is described in **Figure 8.**

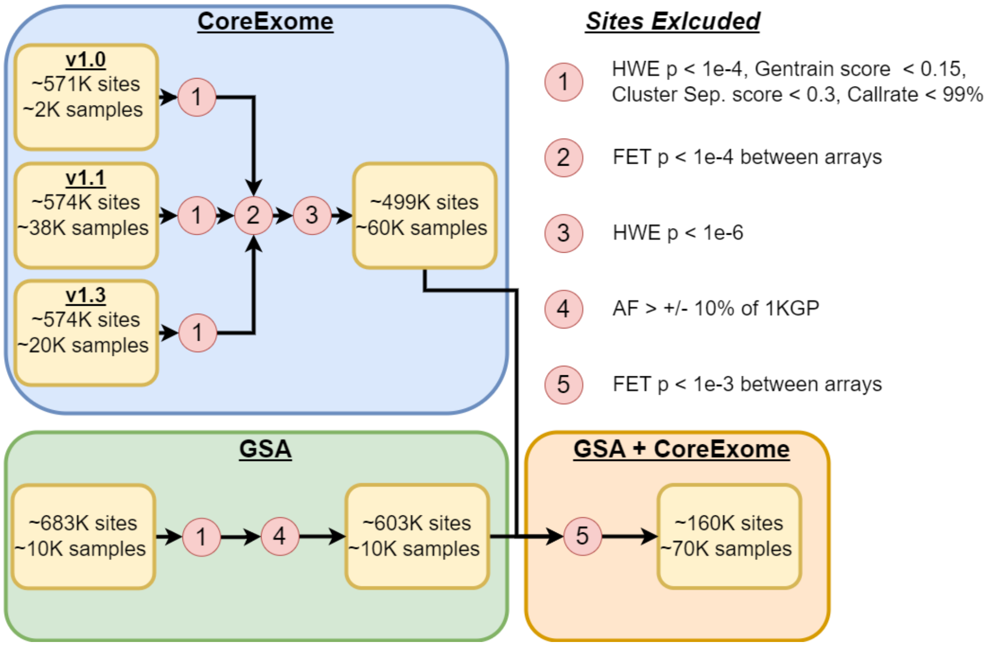

Figure 7: Workflow for variant QC. We apply variant QC to generate several sets of high-quality genotype data from the CoreExome arrays, the Global Screening Array (GSA), or merges of the GSA and CoreExome arrays. The red numbered circles represent steps taken to exclude sites based on variant QC. HWE, Hardy-Weinberg Equilibrium exact test; FET, Fisher's exact test; AF, alternate allele frequency; 1KGP, 1000 Genomes Project.

6.3 Genotype Imputation QC

We first filtered each of the CoreExome- and GSA-based imputations to exclude poorly imputed sites with Rsq < 0.3. We then merged these Rsq filtered data and recalculated MAF as the weighted average of the MAF from the CoreExome- and GSA-based imputations before applying a second filter to exclude very rare variants with MAF < .01%.

7 Quality Evaluation

7.1 Genotype Quality

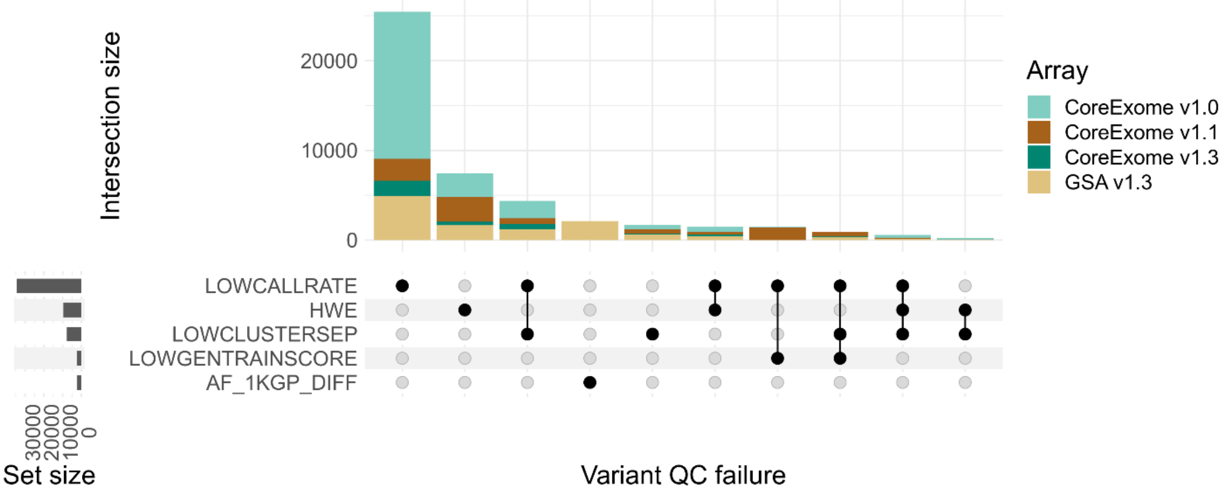

Figure 8: Variant QC outcomes. UpSet plot of the number of well-mapping sites that fail variant QC measures in each array. LOWCALLRATE, call rate < 99%; HWE, Hardy-Weinberg equilibrium test p < 1e-4 before array merge; LOWCLUSTERSEP, Cluster Sep. score < 0.3; LOWGENTRAINSCORE, GenTrain score < 0.15; AF_1KGP_DIFF, > +/- 10% difference in alternate allele frequency compared to 1000 Genomes Project samples. The first 10 largest intersections are plotted.

To determine the accuracy of directly assayed genotypes we measured genotype concordance for each array using 126, 339, 281, and 85 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 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	126	99.81	99.94	99.69	99.90
CoreExome v1.1	339	99.86	99.95	99.84	99.95
CoreExome v1.3	281	99.85	99.96	99.76	99.95
GSA v1.3	85	99.79	99.92	99.69	99.91

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

We evaluated phasing quality by switch error rate (SWE), a metric that describes the total number of haplotype switches that occur over the total number of heterozygous sites where haplotype switches are possible[15]. To obtain known maternal and paternal haplotypes, we used pedigree information inferred with KING (v2.1.3) to phase 77 parent-parent-offspring “trios” assayed on the CoreExome and 12 trios assayed on the GSA using Beagle v4.0[16]. We then removed the parents of each trio from the rest of the cohort before phasing the remaining samples with Eagle as described in section 5.2. We calculated SWE by counting haplotype switches that occurred at heterozygous sites between trio children phased with Eagle or their Beagle pedigree phased counterparts[17]. 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 higher in participants genotyped on the CoreExome array (**Figure 9**). The SWE evaluation was limited to trio children of majority European ancestry.

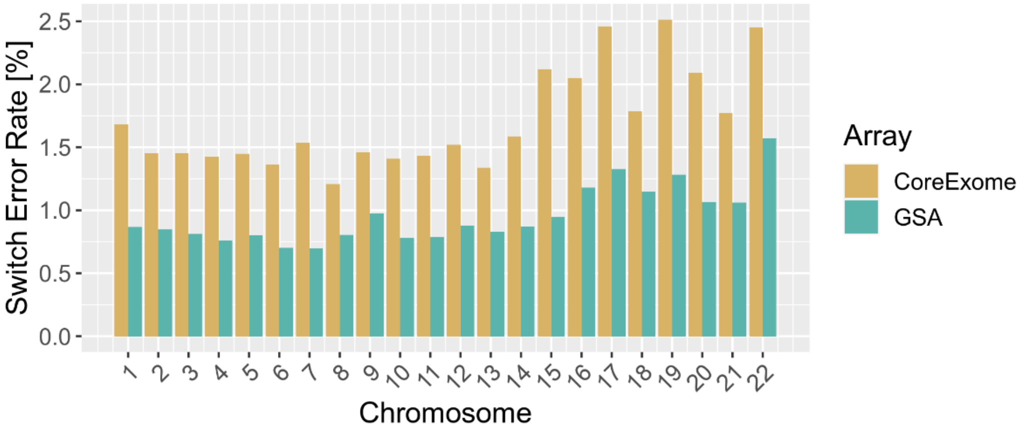

Figure 9: Phasing quality. Evaluation of phasing quality in children of parent-parent-offspring trios in the MGI cohort by switch error rate (SWE). SWE is summarized across participants assayed on either the CoreExome array or GSA. SWE across all autosomes was determined by evaluating the total number of haplotype switches that occurred over the total number of heterozygous sites where haplotype switches were possible.

7.3 Genotype Imputation Quality

We used the ”Rsq” and ”EmpRsq” metrics produced by the genotype imputation software Minimac4 to evaluate imputation quality[7]. The Rsq metric estimates imputation accuracy 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[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 CoreExomes or GSA as input for genotype imputation both Rsq and EmpRsq improved with increasing MAF and mean EmpRsq was > .9 when evaluating sites with MAF > 1% (**Figure 10**).

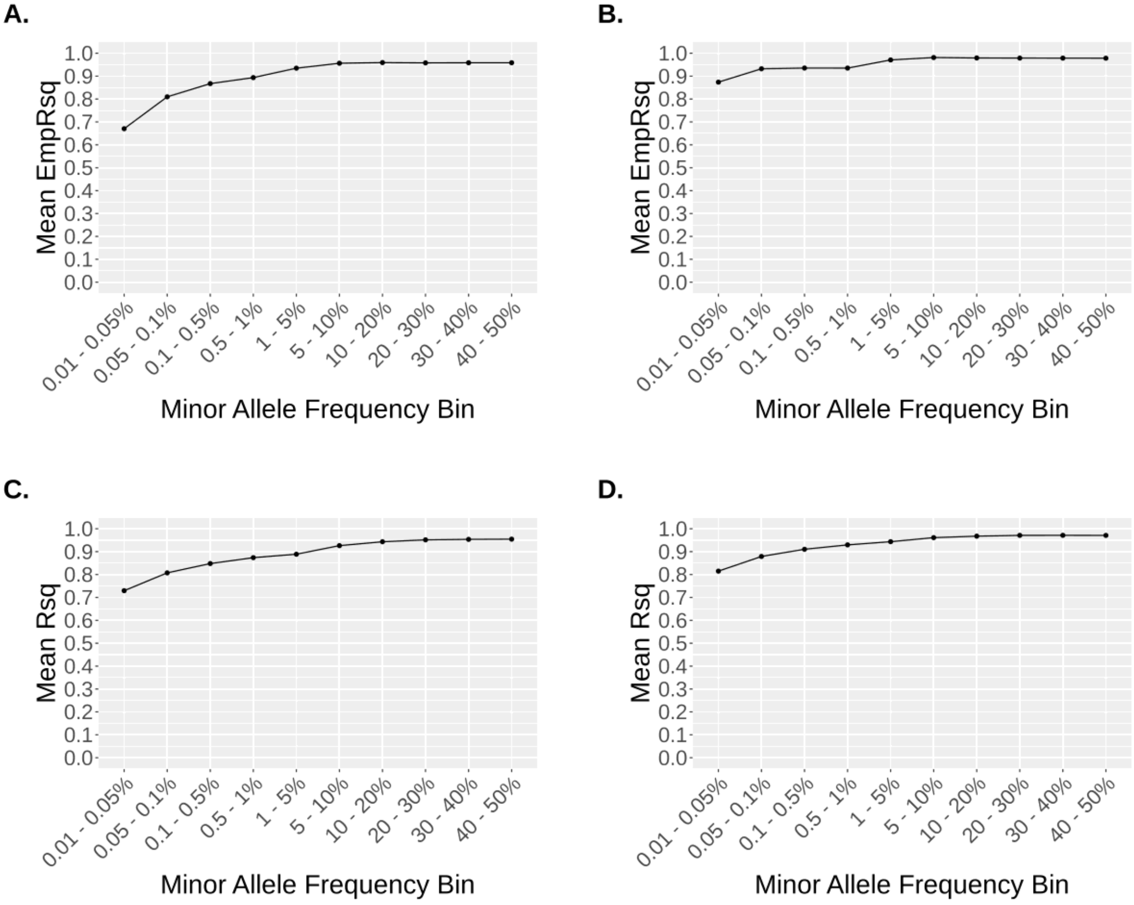

Figure 10: Imputation quality. Summary of imputation quality metrics when genotypes from all participants assayed on the CoreExome array or GSA were used as input for genotype imputation. The Pearson correlation coefficient of known and imputed genotypes (EmpRsq) for **(A.)** the CoreExome array and **(B.)** the GSA at 145,727 sites that were directly assayed on each array and imputed from the TOPMed panel with Rsq ≥ 0.3 and MAF ≥ 0.01%. The estimated correlation between imputed and expected genotypes (Rsq) for 45,330,463 sites that were imputed from the TOPMed panel with Rsq ≥ 0.3 and MAF ≥ 0.01%. when using the **(C.)** CoreExome array and **(D.)** the GSA array genotypes as input.

7.4 Principal Components

We calculated the first 20 principal components (PCs) for samples of all participants included in Freeze 5. 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 65,008 participants unrelated to the 3rd degree or closer and computed PCs using these samples with FlashPCA2 v2.0[19]. We then projected the remaining 5,431 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 European global ancestry fraction > 0.9 (50,102 unrelated & 3,974 related, **Figure 11**). We offer to compute study-specific PCs at the request of investigators.

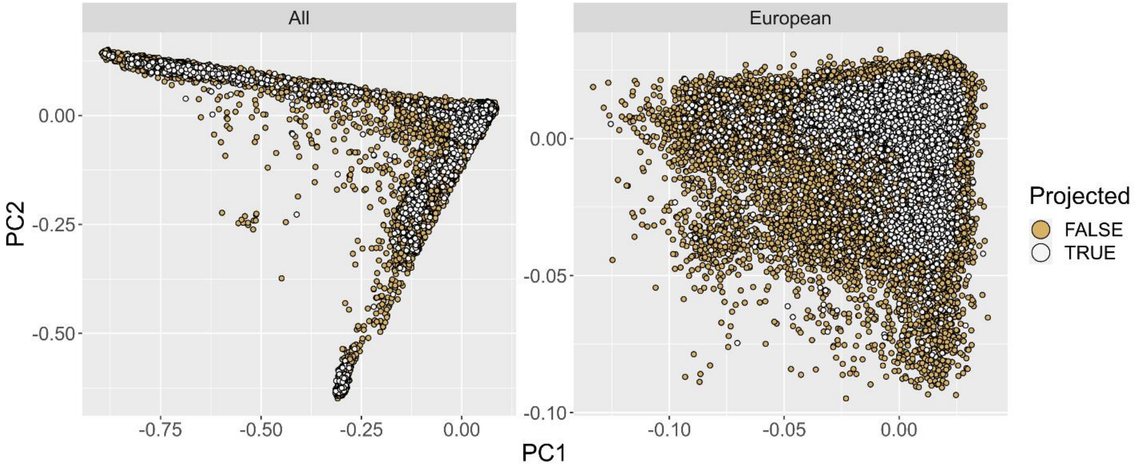

Figure 11: Principal component analysis. Principal components (PC) 1 and 2 computed from samples of **(left)** all unrelated participants included in Freeze 5 and **(right)** participants with inferred European global ancestry fraction > .9. The samples of related participants are projected into the PC space generated from samples of unrelated participants.

7.5 Empirical Comparison of Meta-analysis and Joint-analysis

Two options for analyzing data from participants genotyped on only 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 5 to determine how substitutable these approaches are.

We converted International Classification of Diseases (ICD)-9 and ICD-10 codes collected for 53,795 European ancestry participants included in Freeze 5 to phecode phenotypes using the R PheWAS package (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	Number Cases	Case:Control Ratio
Celiac disease	Digestive	459	1:71
Disorders of bilirubin excretion	Endocrine/metabolic	730	1:65
Primary hypercoagulable state	Hematopoietic	819	1:52
Hypoglycemia	Endocrine/metabolic	1,247	1:25
Type 1 diabetes	Endocrine/metabolic	2,306	1:15
Breast cancer	Neoplasms	3,354	1:13
Cancer of prostate	Neoplasms	3,374	1:5
Iron deficiency anemias	Hematopoietic	5,004	1:7
Atrial fibrillation	Circulatory system	5,860	1:4
Asthma	Respiratory	9,726	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 MAF > .01% and Rsq > 0.3 in both datasets and included covariates for age as of January 1st 2022 for living participants or as of deceased date for non-living participants, recruiting study, genotype-inferred sex, and the first 10 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 and we additionally included a covariate for genotyping array.

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 > .98 among sites with MAF > 1% for p-values and MAF ≥ .1% for betas (**Figure 12**). The median -log10(p-value) among sites with MAF ≤ 1% was 1.61 for the joint-analysis and 1.53 for the meta-analysis. The median -log10(p-value) among sites with MAF > 1% was 1.59 for both the meta-analysis and joint-analysis.

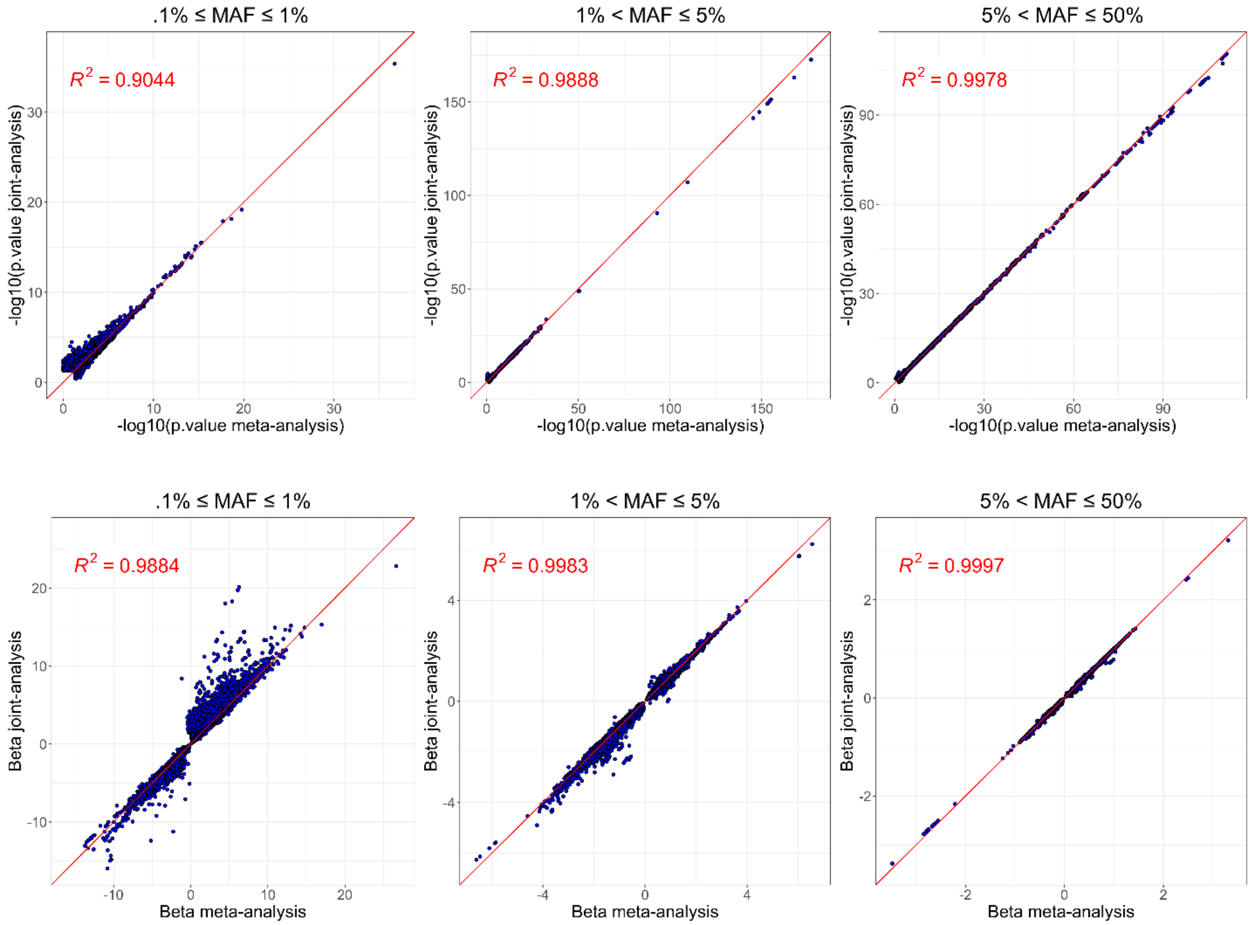

Figure 12: Comparison of p-values and betas from meta- and joint-analysis. Results for -log10(p-values) **(top row)** and betas **(bottom row)**. Points plotted are any hit with a p-value < .05 in either the joint-analysis or meta-analysis across any of the 10 phecode phenotypes evaluated. R^2 is the square of the Pearson correlation coefficient between meta- and joint-analysis. MAF, minor allele frequency.

For each phenotype, we inspected quantile-quantile (QQ) plots that compare p-values from a null uniform [0,1] distribution to p-values observed in either the meta- or the joint-analysis. We show a representative pair of QQ-plots from GWAS from the phecode phenotype asthma (**Figure 13A-B**) which were virtually identical and follow the null closely in the range of moderately significant p-values. For each meta- and joint-analysis we calculated the genomic inflation factor (λ) from sites with MAF > 1%[25]. Median λ was .99 for both meta- and joint-analysis, which is close to the expected value of 1 for a well-controlled GWAS with limited polygenic signal[26]. R^2 of λ across meta- and joint-analysis was > .73 (**Figure 13C**).

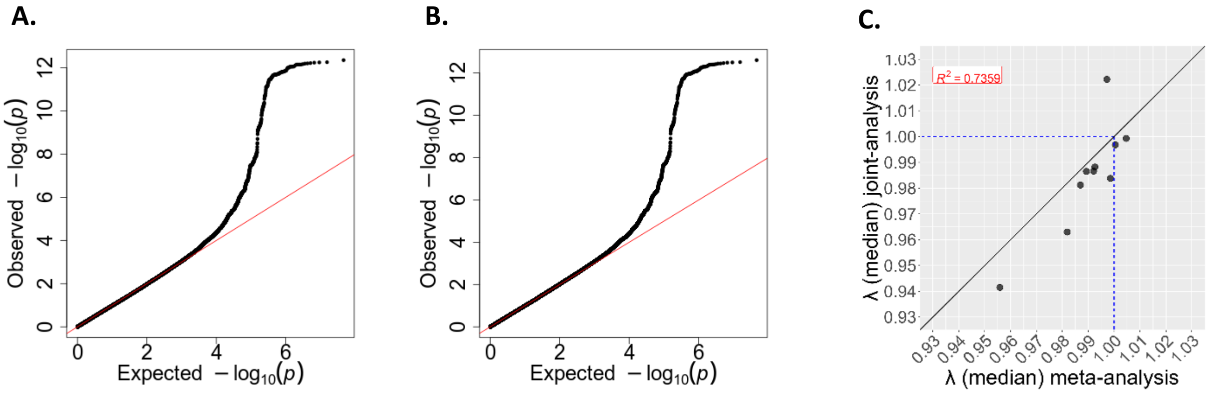

Figure 13: Observed vs. expected signal in meta- and joint-analysis. Quantile-quantile plots comparing p-values from a null uniform [0,1] distribution to p-values from **(A.)** meta-analysis and **(B.)** joint-analysis for the phecode phenotype asthma. **(C.)** genomic inflation factor (λ) computed from sites with minor allele frequency > 1% for each of 10 phecode phenotypes evaluated by each meta- and joint-analysis. R^2 is the square of the Pearson correlation coefficient of λ from meta- and joint-analysis.

Taken together, these results suggest that for variants with MAF > 1%, jointly analyzing participants genotyped on either the GSA and CoreExome array performs near identical to meta-analyzing separate GWAS performed on each of the arrays. For variants with MAF ≤ 1%, the meta-analysis is more conservative than the joint analysis. 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.

8 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.
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. TOPMed Imputation Server. https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about.

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

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

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

11. 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).

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

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

14. 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).

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. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).

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

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

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. PheWeb. https://pheweb.org/MGI-freeze3/.

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).

25. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

26. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807–812 (2011).