Supplementary materials

Revision 1.1
Author: 23andMe Research Team
Date: July 29th, 2026

1. Phenotype Summary Table

The descriptions of each phenotype, their respective numbers of cases and controls, and the counts of loci reaching genome-wide significance (5e-8) across genetic ancestry groups are provided in https://permalinks.23andme.com/pdf/supplemental-r8_g1-phenotype-table.pdf

2. Genotyping array platforms

The v2 platform was based on the Illumina HumanHap550 BeadChip and contained a total of about 560,000 variants, including about 25,000 custom variants selected by 23andMe. The v3 platform was based on the Illumina OmniExpress BeadChip and contains a total of about 950,000 variants and custom content to improve the overlap with our v2 array. The v4 platform was a fully custom array of about 570,000 variants and included a lower redundancy subset of v2 and v3 variants with additional coverage of lower-frequency coding variation. The v5 platform was based on the Illumina Global Screening Array, consisting of approximately 654,000 preselected variants and approximately 50,000 custom content variants. Samples that failed to reach 98.5% genotyping call rate were reanalyzed.

3. WGS data processing procedures

3.1. UKBB 200K WGS reference panel

The publicly available UKBB 200K WGS (late 2021 release) includes 200,031 whole genome samples sequenced to 32.5x coverage. These individuals are predominantly of white British ancestry, but with substantial South Asian and African minorities. Starting with the multi-sample VCFs[1], we applied additional QC and phasing with SHAPEIT5[2] to convert them into an imputation panel. Specifically, we retained alleles that passed the following filters:

- sites with > 6 alleles or > 10% missing were removed
- alleles were split with bcftools norm -m -any
- QDalt >= 6
- AAScore >= 0.35
- ABHet >= 0.175 && ABHom >= 0.900
- Not monomorphic
- allele count > 2 or VEP impact is one of LOW / MODERATE / HIGH

Following these filters, 297,158,813 mutations remained. Genotypes were then phased with SHAPEIT5 using an approach very similar to that outlined in Hofmeister et al. (2023)[2].

3.2. 23andMe WGS reference panel

We selected 14,403 participants from various internal and external WGS datasets. All samples were aligned and duplicate marked using one of two similar pipelines. The high coverage 1000 Genomes CRAM[3] (processed at NYGC) and recently received 23andMe WGS data (processed at the Broad Institute 2019 onwards) used the well-known functional equivalence pipeline[4] for

alignment and post-processing of reads. Older WGS data and the GTEx v8[5] data were processed using an internal pipeline including bwa mem 0.7.15-r1140 alignment, duplicate marking with samblaster v0.1.24, and no BQSR. Some of 23andMe's internal WGS efforts have been previously described[6,7].
The criteria for including participants in the panel were as follows:

- consented for research at time of sequencing
- depth (mean coverage) >= 10
- contamination < 0.05 (estimated by verifybamid[8])
- ratio of transitions/transversions is > 1.9 and <= 2.1
- r^2 with genotyping array >= 0.95 (for 23andMe research participants only)
- >= 3M variants called

Variants were called in each individual sample using DeepVariant v1.2.0[9] to produce GVCFs. The GVCFs were then jointcalled using GLnexus v1.4.1[10,11]. The following quality controls were applied to variants in order:

- sites with > 6 alleles or > 10% missing were removed
- alleles were split with bcftools norm -m -any
- ALT != "*" (overlapping upstream deletion) were removed
- (SNP & AQ > 44) | (Indel & AQ > 39) were retained (AQ = allele quality)
- sites with excessive heterozygosity (p < 1e-30) were removed
- sites with allele count = 1 or 2 were removed unless they had a VEP IMPACT of {LOW, MODERATE, HIGH}

Finally, variants were phased using SHAPEIT5. It's important to note that SHAPEIT5 imputed missing genotypes and produced a final panel without missingness. The final 23andMe reference panel included 14,403 samples and 85,075,659 variants.

4. Merging and further filtering of imputed dosages

We merged the 85,075,659 variants from the 23andMe imputation panel with the 297,158,813 variants from the UKBB WGS imputation panel. Out of a total of 311,492,457 variants, 240,750,442 were exclusive to one panel and 70,742,015 were present in both panels.

Potentially functional variants with VEP impact categorized as LOW / MODERATE / HIGH (N = 10,728,077) were set as PASS by default since these may be enriched for causal mutations. Subsequently, these functional variants were further filtered based on imputation r^2. In order to reduce the number of non-coding variants, we implemented a filter taking into account both minor allele frequency (MAF) and imputation r^2.This filter applied a more stringent threshold for rarer variants. Specifically, for the non-coding variants (N = 300,764,380) in each panel, we first determined whether the variant passed the following QC criteria in any one of the populations:

$log_{10}MAF + log_{10}r^2 > X$ where X = -4.699 for European; X = -4 for other genetic ancestry groups.

Using the above QC criteria, 80,438,411 and 159,490,516 variants were set as PASS in the 23andMe WGS and UKBB 200K WGS reference panels, respectively.
No further action was needed for variants that passed in only one panel. However, for variants that passed in both panels, we used a simple quality metric to determine which panel took precedence. Specifically, for each panel we computed the effective minor allele count and retained the panel with the higher count:

$$\sum_i \sum_j r^2_{ij} f_{ij} N_{ij}$$

where r^2, f, and N are the imputation quality, MAF and sample size respectively, in genotyping array i and ancestry j. The total number of variants before and after filtering are shown in Table 1.

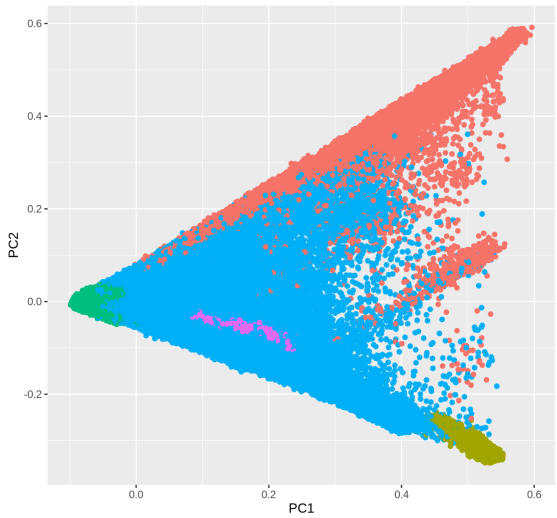
Table 1. *The overall count of variants in the imputation reference panel.*

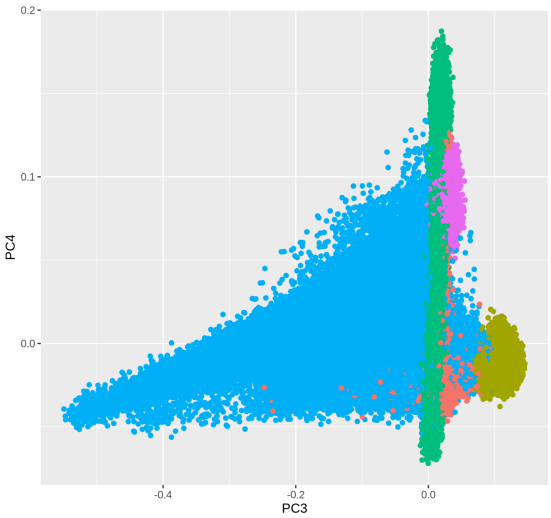
	23andMe WGS	**UKBB 200K WGS**	**Union**
Before filtering	85,075,659	297,158,813	311,492,457
After filtering	80,438,411	159,490,516	172,729,416
After precedence	43,766,815	128,962,601	172,729,416

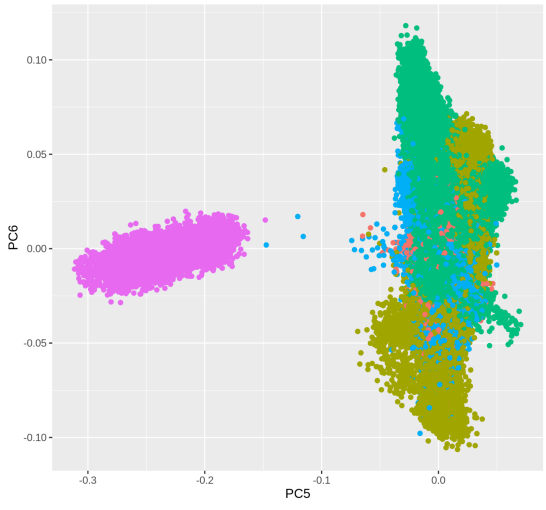
5. GWAS variant QC

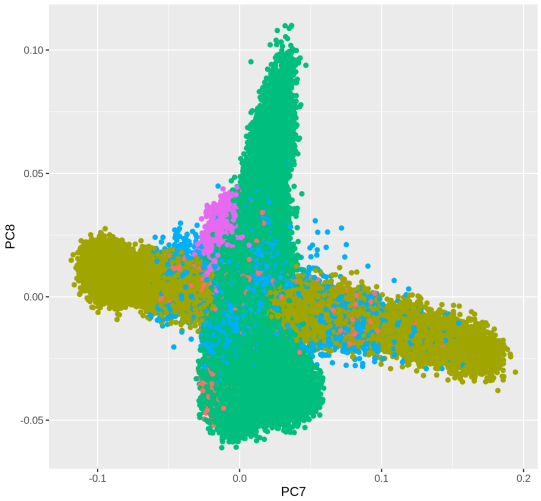
Low imputation quality variants (r2 < 0.5 averaged across all platforms or a minimum r2 < 0.3 within any platform) or those with evidence of batch effects (analysis of variance (ANOVA) F-test across 20 date buckets, P < 10−50) were excluded. Genotyped variants were excluded if present only on v2 arrays (due to small sample size), failed a Mendelian transmission test in trios (P < 10−20), failed a batch effect test (date bucket ANOVA P < 10−50) or had a call rate < 90%. For quantitative and ordinal traits, variants with MAF < 0.001 were also excluded. After QC, the final GWAS dataset included association results for approximately 117M variants for binary traits and 17M variants for quantitative or ordinal traits.

6. PCA plots for the ancestry groups in GWAS









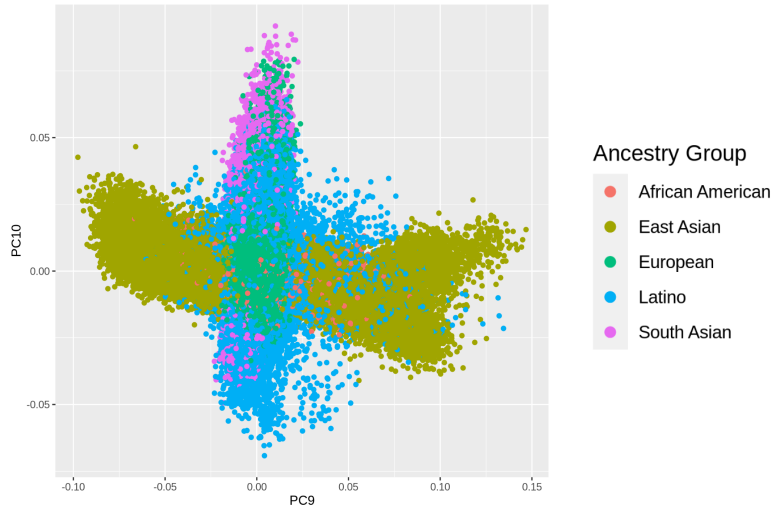


Figure 1. *PCA scatter plots from the overall PC1 to PC10, illustrating similarities and differences among research participants across African American, East Asian, European, Latino, and South Asian ancestry groups.*

7. References

1. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
2. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
3. 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).
4. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
5. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
6. O'Connell, J. *et al.* A population-specific reference panel for improved genotype imputation

in African Americans. *Commun. Biol.* **4**, 1–9 (2021).

7. Shi, J. *et al.* GWAS of cataract in Puerto Ricans identifies a novel large-effect variant in ITGA6. 2023.07.25.23293173 Preprint at https://doi.org/10.1101/2023.07.25.23293173 (2023).
8. Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30**, 185–194 (2020).
9. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
10. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. 343970 Preprint at https://doi.org/10.1101/343970 (2018).
11. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).