Genotyping Array Design and Data Quality Control in the Million Veteran Program

Haley Hunter-Zinck,[1,14] Yunling Shi,[1,14] Man Li,[1,2,14] Bryan R. Gorman,[1,3,14] Sun-Gou Ji,[1,4,14,15] Ning Sun,[5,6,14] Teresa Webster,[7] Andrew Liem,[1,3] Paul Hsieh,[1] Poornima Devineni,[1] Purushotham Karnam,[1] Xin Gong,[1] Lakshmi Radhakrishnan,[7] Jeanette Schmidt,[7] Themistocles L. Assimes,[8,9] Jie Huang,[1] Cuiping Pan,[8,9] Donald Humphries,[1] Mary Brophy,[1] Jennifer Moser,[10] Sumitra Muralidhar,[10] Grant D. Huang,[10] Ronald Przygodzki,[10] John Concato,[5,6,16] John M. Gaziano,[1,11] Joel Gelernter,[5,6] Christopher J. O'Donnell,[1] Elizabeth R. Hauser,[12,13] Hongyu Zhao,[5,6] Timothy J. O'Leary,[10] VA Million Veteran Program,[17] Philip S. Tsao,[8,9] and Saiju Pyarajan[1,11,*]

The Million Veteran Program (MVP), initiated by the Department of Veterans Affairs (VA), aims to collect biosamples with consent from at least one million veterans. Presently, blood samples have been collected from over 800,000 enrolled participants. The size and diversity of the MVP cohort, as well as the availability of extensive VA electronic health records, make it a promising resource for precision medicine. MVP is conducting array-based genotyping to provide a genome-wide scan of the entire cohort, in parallel with whole-genome sequencing, methylation, and other 'omics assays. Here, we present the design and performance of the MVP 1.0 custom Axiom array, which was designed and developed as a single assay to be used across the multi-ethnic MVP cohort. A unified genetic quality-control analysis was developed and conducted on an initial tranche of 485,856 individuals, leading to a high-quality dataset of 459,777 unique individuals. 668,418 genetic markers passed quality control and showed high-quality genotypes not only on common variants but also on rare variants. We confirmed that, with non-European individuals making up nearly 30%, MVP's substantial ancestral diversity surpasses that of other large biobanks. We also demonstrated the quality of the MVP dataset by replicating established genetic associations with height in European Americans and African Americans ancestries. This current dataset has been made available to approved MVP researchers for genome-wide association studies and other downstream analyses. Further data releases will be available for analysis as recruitment at the VA continues and the cohort expands both in size and diversity.

Introduction

The United States Department of Veterans Affairs (VA) initiated the Million Veteran Program (MVP) in 2011 to create a mega-biobank of at least one million samples with genetic data linked to nationally consolidated longitudinal clinical records.[1] The initial and continuing goal of MVP is to create a national resource for research to improve the health of United States veterans and, more generally, to contribute to our understanding of human health. MVP has currently collected samples from over 800,000 Veteran participants and expects to exceed a total of 1 million participants in the next 2 to 3 years.

Although MVP is similar in some respects to other large biobank projects, such as the UK Biobank; the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH); the China Kadoorie Biobank (CKB); and the DiscovEHR initiative,[2–4] it is unique in several ways. As one of the largest single biobanking efforts to date, MVP satisfies the need for larger genetic datasets while also benefiting from a very rich, nationally integrated longitudinal clinical database housed in the largest consolidated healthcare network in the United States. This feature allows for enhanced clinical phenotyping capabilities. The availability of additional self-reported health and lifestyle survey information augments clinical data from the Veterans Information Systems and Technology Architecture (VistA)—the VA's electronic health record (EHR).

Furthermore, with over 29% of participants self-reporting non-white ethnicity, MVP has substantial diversity in genetic ancestry and thus meets a pressing need for greater diversity in genome-wide association analyses so that researchers can discover novel associations,

[1] VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA 02130, USA; [2] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT 84132, USA; [3] Booz Allen Hamilton, McLean, VA 22102, USA; [4] Seven Bridges, Boston, MA 02129, USA; [5] VA Cooperative Studies Program, VA Connecticut Healthcare System, West Haven, CT 06516, USA; [6] Yale University School of Medicine, New Haven, CT 06510, USA; [7] Thermo Fisher Scientific, Santa Clara, CA 95054, USA; [8] VA Palo Alto Health Care System, Palo Alto, CA 94304, USA; [9] Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA; [10] Office of Research and Development, Veterans Health Administration, Washington DC 20571, USA; [11] Department of Medicine, Brigham and Women's Hospital and Harvard School of Medicine, Boston, MA 02115, USA; [12] Durham VA Health System, Durham, NC 27705, USA; [13] Department of Medicine, Duke University, Durham, NC 27617, USA

[14] These authors contributed equally to this work

[15] Present address: Sun-Gou Ji, Bridgebio Pharma, Palo Alto, CA

[16] Present address: Department of Medicine, Yale University, New Haven, CT and Center for Drug Evaluation and Research, Food and Drug Administration, Silver springs, MD

[17] Million Veteran Program membership is provided in the consortia section

*Correspondence: saiju.pyarajan@va.gov

https://doi.org/10.1016/j.ajhg.2020.03.004.

reduce false positives, and increase research equity.[5–8] As such, the MVP cohort provides an unprecedented opportunity for increasing the power of genome-wide association studies (GWASs) and will enable association discoveries regarding clinically important low-frequency and rare variants; such discoveries would only be possible in larger sample sizes. Reliable typing of these variants could provide explanations of missing heritability in complex or non-Mendelian diseases. However, the genetic diversity of MVP also poses challenges in genotype quality control.

In this report, we introduce the first installment of MVP genotype data consisting of 459,777 samples surveyed at 668,418 markers. In brief, we (1) describe the design of a research genotyping array with emphasis on clinically useful and/or rare variants applicable to multi-ethnic backgrounds; (2) describe the generation and quality control of genotyping data; (3) highlight some of the current MVP dataset's unique features, including exploratory analyses of genetic ancestry; and (4) replicate effect sizes of previously reported variants associated with height in European Americans and African Americans. Overall, we find that the MVP genetic dataset, linked to deep phenotypic data, is a high-quality and diverse resource for performing genetic analyses.

Material and Methods

Human Subjects and Data and Sample Collection

The VA Central IRB, as well as the local IRBs at the VA Boston Healthcare System and the VA Connecticut Healthcare System, approved this project. An overview of the recruitment strategies and protocols is given in a previous publication.[1] In brief, participants were recruited from approximately 60 VA healthcare facilities across the United States on a rolling basis. Informed consent was obtained from all participants. Participants consented to a blood draw and to have their DNA analyzed, as well as to linking their genetic information with their full clinical, survey, and other health data. Participants were also invited to answer two separate surveys about basic demographic information and lifestyle characteristics.

Blood drawn from consenting participants was shipped to the central biorepository in Boston, Massachusetts, where DNA was extracted and later shipped to two external vendors for genotyping on a custom Axiom array designed specifically for MVP (MVP 1.0). A description of the MVP 1.0 array design features is detailed in the Supplemental Information.

Thermo Fisher Scientific (formally Affymetrix) Axiom Genotyping Platform

The MVP 1.0 custom Axiom array is based on the Axiom Genotyping Platform. The Axiom genotyping platform utilizes a two-color, ligation-based assay using 30-mer oligonucleotide probes synthesized *in situ* onto a microarray substrate. Each single-nucleotide polymorphism (SNP) feature contains a unique oligomeric sequence complementary to the genomic sequence flanking the polymorphic site on either the forward or the reverse strand. Solution probes bearing attachment sites for one of two dyes, depending on the 3′ (SNP-site) base (A or T, versus C or G), are hybridized to the target complex, followed by ligation for specificity. Oligonucleotide sequences complementary to the forward or reverse strands are referred to as probesets. A marker (SNP or indel) can be interrogated by the probeset for the forward and/or reverse strand.

For additional details of the Axiom Genotyping Platform, see the Supplemental Materials and Methods.

Genotype Calling

We received unprocessed Axiom genotype data for 485,856 unique samples assayed by two vendors, referred to as vendor 1 and vendor 2, and performed genotype calling in batches grouped by vendor and sample processing date. By using data provided by the vendors and generated from our internal genotype calling process (see Supplemental Materials and Methods for details), we first analyzed the standard Axiom genotype quality metrics and compared these metrics between the two vendors.

After calling genotypes, we applied an advanced normalization procedure for mitigating plate-to-plate variation developed in collaboration with Thermo Fisher Scientific. The procedure was applied selectively on a per-batch basis to probesets exhibiting high plate-to-plate variance. After plate normalization, we applied standard marker quality-control procedures to clean and harmonize genotype calls across all the batches (Supplemental Materials and Methods), followed by advanced sample quality control (QC).

Advanced Sample QC

Sample Contamination

To detect and mitigate sample contamination, we assessed heterozygosity with PLINK, version 1.9, by calculating the F coefficient and quarantining samples with an F coefficient of less than −0.1. We assessed excess relatedness by using the relatedness inference software KING, version 2.0, and quarantined samples having a kinship coefficient of at least 0.1 with seven or more other samples within MVP. These samples had high dish QC (DQC) and low call rates and were outliers in comparison to the majority of samples in the MVP dataset (Figure S5D). Because a call rate below 98.5% correlated with excess sample heterozygosity or relatedness, we removed samples (15,436, or 3.00%) with call rates below this threshold.[9] All samples that were removed or quarantined from the current release of MVP data will be re-genotyped and included in the future data releases.

Sample Mislabeling

We identified samples and plates demonstrating potential mislabeling issues by analyzing genotype concordance between intentional duplicate samples that were sent blinded to the vendors as new samples for genotyping. Of the 25,867 intentional duplicate pairs, only 211 (0.82%) pairs were highly discordant (greater than 1% discordance). Samples on plates with discordant intentional duplicate pairs were quarantined for further analysis and re-genotyping. We also removed both samples and plates if the duplicate pair had a relatedness coefficient of less than 0.45. These precautions were taken out of concern about potential plate swaps and led to 9,975 samples' being quarantined.

Sample Misidentification

To discriminate between misidentified intentional duplicates (same samples intentionally genotyped twice), technical duplicates (controls repeatedly genotyped by vendors), and monozygotic twins, we calculated sample relatedness with the KING software, version 2.1.[10] Monozygotic twins were confirmed by cross-referencing EHR data. Pairs with birth dates differing by no more than one day and having unique participant identifiers and first names were considered verified monozygotic twin pairs. Unverified samples were quarantined as potentially mislabeled and will be re-genotyped.

Sex Check

To confirm sample gender, we extracted markers genotyped on the X chromosome while excluding the pseudoautosomal region, used the sex-check command from PLINK, and compared the expected F coefficient on the X chromosome to the gender recorded in the sample's EHR for all samples.[11] Participants whose reported gender differed from that inferred by PLINK were quarantined from subsequent analysis. We also removed remaining samples on plates with four or more gender mismatches to account for potential plate swaps. The threshold is relatively low because of the low percentage of females in our dataset.

Advanced Marker QC

Advanced Marker QC Pipeline

We implemented three main approaches to create the advanced marker QC pipeline: (1) exclude probeset calls from all batches for probesets that failed advanced QC tests; (2) exclude probeset calls in a given batch for which the probeset is not recommended; and (3) choose the best probeset per marker for markers interrogated by multiple probesets and exclude probeset calls from all batches for the "not-best" probesets. Details of each steps of the advanced marker QC are available in Supplemental Materials and Methods and in Figures S4, S6A, and S7A.

The advanced marker QC pipeline produced an inclusion list of probesets that met quality standards across the entire MVP dataset. For each batch, we included a probeset in the dataset if it met all three of the following criteria: (1) it was included in the inclusion list; (2) it was recommended in that batch; and (3) it was the best probeset for a marker interrogated by multiple probesets. We then generated a list of probesets per batch, created PLINK marker list binary files for each batch, and merged all batches together by using the PLINK merge command.

Reproducibility of Genotype Calling

To assess the consistency of genotype calls across time and vendors, we analyzed the discordance between 25,867 intentional duplicate samples that were sent blinded to the vendors. After confirming that these sample pairs were genetically identical through KING relatedness inference, we determined the number of minor-allele pairs (MAPs) for each marker. A MAP is any pair of genotypes for a marker where both genotypes are called in the sample pair and where the pair contains at least one minor allele. We then calculated the number of discordant genotyping pairs per MAP for each marker. Normalizing by the number of MAPs renders different minor-allele frequency (MAF) bins comparable in the discordance calculation. Otherwise, rare markers will always have extremely low discordance rates because most samples carry the homozygous major genotype.

Additionally, within the 485,856 samples genotyped in the MVP cohort, we included 2,064 positive control samples. We called the genotypes of the positive controls along with other MVP samples across 112 batches organized by genotyping scan date for 668,418 markers passing advanced marker quality control. These genotypes were compared to the consensus positive-control genotype.

To construct the consensus genotype sequence, we calculated the frequency of each marker across the panel of 2,064 positive-control samples. Markers with MAF of less than 1% were set to homozygous in the consensus sequence, and markers with a MAF of greater than 49% were set to heterozygous in the consensus sequence. For markers with a MAF greater than or equal to 1% and less than or equal to 49% (536, or 0.082% of markers) or that had no observed calls (18,158, or 2.76%), we set the consensus genotype to missing.

We calculated concordance across all common (MAF $\geq$ 5%) and low-frequency (MAF $<$ 5%) markers by assessing MAFs over the entire MVP sample. We then calculated concordance between the consensus sequence and each positive control. Concordance was defined as the number of matching called genotypes over the total number of called genotypes. Uncalled markers in either the positive control or the consensus sequence were not included in either the numerator or the denominator of the concordance calculation. We then plotted the concordance distribution for each batch's positive controls across time.

Comparing MVP Allele Frequencies to Those from gnomAD and the UK Biobank

Genome Aggregation Database (gnomAD) version 2.1 data were downloaded online (see Web Resources). Markers in both gnomAD and MVP were matched on chromosome, start position, end position, reference allele, and alternative allele. For any mismatch, we checked strands and indel notations. Reference and alternative alleles were corrected, and their frequencies were recomputed when strands were flipped. Indels had their genomic coordinates and alleles recoded and harmonized.

UK Biobank summary data were downloaded online (see Web Resources). Markers shared between the UK Biobank and MVP were matched through the use of SNP rsIDs. Because information on marker chromosome, genomic positions, reference alleles, and alternate alleles were not provided in the summary statistics, we were unable to explicitly check for strand flips. However, as we expected, variant annotation in MVP and the UK Biobank tended to be well harmonized because both were genotyped on Axiom arrays and followed the same standard Axiom marker QC workflow; thus, we compared allele frequencies as is without excluding any variants.

For this analysis, European Americans (EAs) were defined as samples with a GBR (British in England and Scotland) proportion greater than 0.9 on the basis of ADMIXTURE results (described below), resulting in a sample size of 311,365. We used PLINK to compute allele frequencies by genetic ancestry subgroup via the "–freq" command with default filters and quality control parameters.

Genetic Relatedness

We performed additional preprocessing of the MVP dataset before analyzing genetic relatedness. We applied standard PLINK 1.9 filters for genotype missingness ($>$ 5% removed), MAF ($<$ 1% removed), and sample missingness ($>$ 5% removed).[11] We then conducted pairwise relatedness inference by using KING 2.1 to identify related pairs.[10] KING explicitly accounts for population structure and is therefore an appropriate algorithm for our sample, which contains diverse genetic ancestry. However, KING is also known to overestimate relatedness in the presence of recent admixture. Therefore, we selected SNPs with a low load in principal components (PCs) 1–3 for a second round of KING, as was done in the UK Biobank.[12]

We ran the first round of KING with the command "–related–degree 3" to identify all potential pairs of individuals with closer than third-degree relatedness. From this result, we excluded all individuals with more than 200 third-degree relatives and also families with more than 100 members because we suspected they were artifacts of sample-processing errors such as low-level sample contamination. Then, a set of unrelated individuals was defined via the largest_independent_vertext_sets() function in the Python version of the igraph tool. Principal-component analysis (PCA) was then conducted with the unrelated samples. Only SNPs with a MAF greater than 0.01 and missingness less than 0.015 were considered for this PCA. 23 regions defined as having high linkage disequilibrium (LD) in the UK Biobank[13] were also excluded, and then SNPs were pruned according to an r^2 threshold of 0.1, a window of 1000 markers, and a step size of 80. In the end, 90,288 SNPs were selected for PCA, which was conducted with PLINK v2.00a2LM and the command "–pca var-wts approx" so that variant weights and fast PCA approximation could be obtained. We selected low-weight SNPs in PC1, PC2, and PC3 by adjusting the absolute weight threshold to keep at least two-thirds of the input SNPs, which led to 60,118 SNPs' being put forward for the next round of KING.

The second round of KING was again conducted with the command "–related–degree 3." The effect of using SNPs with low weights in PCs 1–3 on the distribution of the number of relatives per individual is shown in Figures S10A and S10B. We flagged 35 individuals with more than 200 third-degree relatives (UK Biobank reported nine individuals with more than 200 third-degree relatives), as well as all members of two clusters that were tightly interconnected with each other (Supplemental Materials and Methods and Figures S10C, S10D, and S11).

We defined genetically identical pairs as those having a kinship coefficient of 0.45 or greater (the maximum kinship coefficient output by KING is 0.5). However, given the large number of intentional duplicates samples in our dataset, we only considered genetically identical pairs as monozygotic twin pairs after cross-referencing EHR data as above. Parent-child pairs were defined as those having a kinship coefficient of greater than or equal to 0.19 and less than 0.45 and having less than 0.0025 percent of the genome held with zero alleles identical-by-state (IBS0). Sample pairs with a kinship coefficient greater than or equal to 0.19 and less than 0.45 and IBS0 greater than or equal to 0.0025 were designated as full siblings. Any pairs of participants with a kinship coefficient between 0.0884 and 0.19 were inferred to be second-degree or third-degree relatives. To identify potential trios in our sample, we extracted parent-child pairs in which a sample appeared twice. We then assessed the kinship coefficient between the other two participants. If the other two participants were not a related pair and consisted of one male and one female, we identified these three samples as a trio.

Genetic Ancestry

For genetic-ancestry analysis, we used the same set of markers used for relatedness analysis and applied LD pruning with PLINK (–indep-pairwise 1000 50 0.05), which left us with 50,000 markers.

Principal-Component Analysis

For the 1000 Genomes Project projection PCA, we merged the MVP dataset with the 1000 Genomes Project Phase 3 reference panel.[14] We first filtered the 1000 Genomes Project dataset to ensure scalable merging with the MVP dataset. Markers with MAF less than 1% and any samples constituting related pairs were removed prior to LD pruning via PLINK according to the same parameters as above. We then calculated PCs by using the 1000 Genomes Project dataset and projected the MVP samples onto them with EIGENSOFT, version 6.0.1.[15]

We also calculated the PCs on the filtered MVP dataset alone by using the FastPCA method from the EIGENSOFT package for within-cohort PCA. For this PCA, we excluded all related individuals, whereas we kept all related individuals in the 1000 Genomes project PCA.

ADMIXTURE Analysis

In order to quantify ancestry proportions in MVP, we ran the program ADMIXTURE, version 1.3, on the MVP samples in supervised mode with five reference populations from the 1000 Genomes Project dataset as training data.[16] We chose the five reference populations on the basis of their global geographic location to ensure global representativeness. The Yoruba in Ibadan, Nigeria (YRI) samples serve as a proxy for West African ancestry, the Luhya in Webuye, Kenya (LWK) for East African ancestry, the British in England and Scotland (GBR) for European ancestry, the Han Chinese in Beijing, China (CHB) for East Asian ancestry, and the Peruvians from Lima, Peru (PEL) for Native American ancestry (Figure S8C). Participants with more than 80% of their genetic ancestry attributed to one reference population were assigned to that reference. Remaining participants who had greater than 90% of their genetic ancestry derived from two reference populations were assigned to that pair of populations. Any participants not meeting the above two criteria were assigned to a separate subgroup (MVP_OTHER) and were assumed to contain admixture from three or more reference populations.

UMAP Analysis

We used Uniform Manifold Approximation Projection (UMAP), a dimensionality-reduction method that is useful for visualizing both global and local structure in data, to further visualize the genetic ancestry of the MVP cohort. A UMAP embedding was calculated on the basis of the first 10 principal components of unrelated samples with hyperparameters n_neighbors of 15 and min_distance of 0.1, which were suggested by a previous study on UK Biobank data.[17] We then visualized the population structure by projecting subpopulations identified by our ADMIXTURE analysis onto the UMAP embedding.

GWAS of Height

Height measurements, dates of measurement, and dates of birth for each participant were extracted from the VA healthcare system's EHR. Any height measurement outside the range of 48 to 84 inches was excluded,[18] and inches were converted to meters. Age at measurement was calculated by subtracting the date of birth from the date of height measurement. Individuals younger than 18 or older than 120 years old were excluded. Sex was genetically determined by PLINK.

Markers whose genotype missingness was greater than 1%, as well as non-autosomal markers, were removed. Samples whose missingness was over 5% were also excluded. By using the results of the relatedness analysis described below, we also removed all closely related pairs.

After marker and sample filtering, we ran association tests by using BOLT-LMM[13] with sex, age, age-squared, and the first 10 PCs as

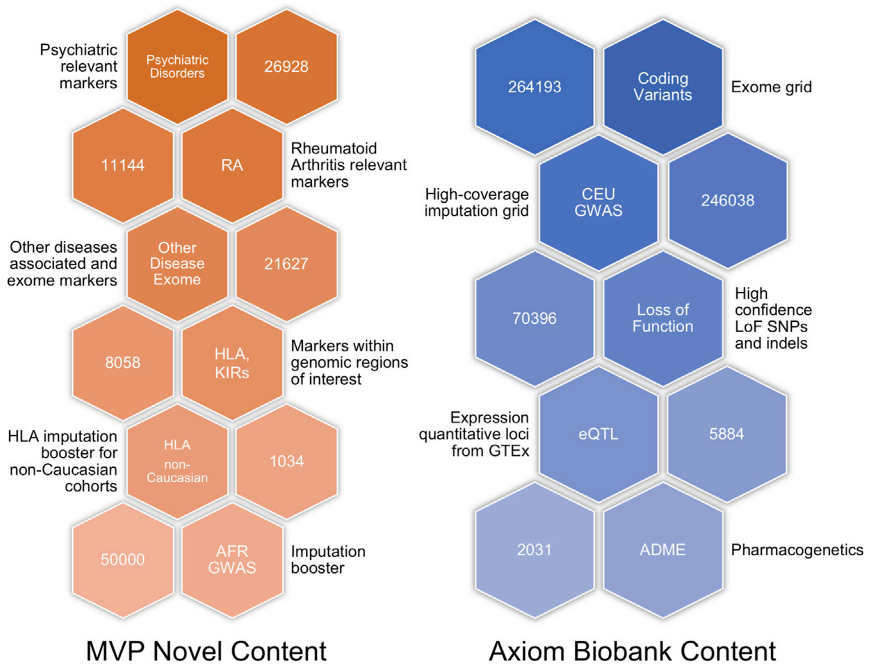

Psychiatric relevant markers
Psychiatric Disorders
26928
11144
RA
Rheumatoid Arthritis relevant markers
Other diseases associated and exome markers
Other Disease Exome
21627
8058
HLA, KIRs
Markers within genomic regions of interest
HLA imputation booster for non-Caucasian cohorts
HLA non-Caucasian
1034
50000
AFR GWAS
Imputation booster
MVP Novel Content
264193
Coding Variants
Exome grid
High-coverage imputation grid
CEU GWAS
246038
70396
Loss of Function
High confidence LoF SNPs and indels
Expression quantitative loci from GTEx
eQTL
5884
2031
ADME
Pharmacogenetics
Axiom Biobank Content

Figure 1. Key MVP 1.0 Genotyping Array Modules
The modules are divided into those shared with the Axiom Biobank Genotyping Array and those unique to the MVP 1.0 array, along with descriptions and counts of unique markers in each module. Counts represent the number of markers in the module, and markers can be in more than one module.

covariates. LD scores were calculated from the 1000 Genomes Project population subsets with ldsc 1.0.[19] We generated model SNPs with PLINK 2.0 by pruning unrelated samples with an R-squared threshold of 0.2 (–pairwise-indep 1000 50 0.2). We also generated PCs by using PLINK 2.0 (–pca approx) on the cohorts that had model SNPs extracted.

We extracted the effect size, direction of effect, and allele for each previously associated marker from the GWAS catalog on March 21, 2019 and then extracted the effects for the markers present in the MVP association analysis. We then scaled the effect values within each study to between 0 and 1 to account for different height units and plotted the previously derived effects against those inferred in MVP.

Results

The MVP 1.0 Array

Array Design and Content

The MVP 1.0 array was based on the Applied Biosystems Axiom Biobank Genotyping Array with additional custom content developed for MVP (Figure 1). The Axiom Biobank Genotyping Array incorporates multiple content categories that are important for translational medicine research and discovery; such categories include modules for genome-wide coverage of common European variants, rare coding SNPs and indels, pharmacogenomics markers, expression quantitative trait loci (eQTLs), and loss-of-function markers (further described in Supplemental Materials and Methods). The MVP-1.0-specific modules were mainly SNPs and indels known to be associated with diseases and traits of interest to MVP (especially psychiatric disorders and rheumatoid arthritis), as well as a set of SNPs selected to improve African American imputation performance (Supplemental Materials). In total, 723,305 probesets interrogating 686,682 unique bi-allelic markers (SNPs and indels) based on the GRCh37 genome build were tiled onto the MVP 1.0 array. Among these, 270 are mitochondrial markers, 142 are in the non-pseudoautosomal regions of the Y chromosome, 1,139 are in the pseudoautosomal regions (PAR1 and PAR2) of the X and Y chromosomes, 18,026 are in the non-pseudoautosomal regions of the X chromosome, and the remaining 667,105 markers are autosomal markers (Table S1).

MVP 1.0 Genotyping Quality Control and Assessment

Assessment of Overall Genotyping Performance

Figure S3 is an overview of the steps taken to ensure high-quality genotype data for the MVP cohort. Advanced genotype and sample QC were conducted in addition to the standard Affymetrix good-practice guidelines and are described in the Materials and Methods and Supplemental Materials and Methods. In addition, we further devised a batch variation correction step to apply to markers that showed significant allele frequency differences between releases (Supplemental Methods and Figures S4 and S6A).

We investigated multiple quality-control metrics for across and within the two assay vendors. Median Axiom DQC values for all genotyping batches were greater than 95 for both vendors (Figure S5A). Median QC call rate was also high, exceeding 99% for each genotyping batch (Figures S5B and S5C). Overall, sample call rates and other genotype quality-control metrics demonstrated high-quality genotype calls for MVP regardless of genotyping vendor (more detail is available in the Supplemental Materials and Methods).

Marker and Sample QC and Selection

The MVP 1.0 array contains a large amount of novel, custom marker content that has not been validated on other arrays. These markers were assayed with more than one probeset, so determining which probesets for a given marker performed best across all genotyped batches and removing systematically poor-quality probesets required advanced marker QC. Ultimately, we retained 668,418 markers representing 97.34% of the original markers and included 459,777 samples from a total of 485,856 unique

Table 1. Quarantine and Exclusion Criteria for MVP Samples and Sample Count per Category

Category	Number of Samples	Percentage of Samples
Starting MVP sample set for analysis	514,383	–
Intentionally duplicated samples	25,291	–
Uniquely genotyped individuals	485,856	100.00%
Samples with call rates below 98.5%	15,436	3.18%
Positive-control samples	3,236	0.66%
Samples with sex misclassification	1,450	0.29%
Samples on plates containing 4 or more sex misclassifications	2,619	0.53%
Unintentionally duplicated samples	1,149	0.23%
Samples on plates containing an intentional duplicate with high discordance	9,975	2.05%
Samples with high heterozygosity	248	0.05%
Samples with no or multiple unique participant identifiers	71	0.01%
Intentionally duplicated samples with high discordance	413	0.08%
Samples with 7 or more "relatives"	466	0.09%
Samples excluded from the dataset	28,527	5.87%
Samples quarantined from the dataset	31,836	6.55%
Sample set in current data release	459,777	–

Percentages are calculated from the total number of uniquely genotyped individuals (485,856). Categories are not mutually exclusive (i.e., a sample can be removed as a result of more than one category and is counted in each applicable category in the table).

genotyped samples in this data release. As expected, almost 98% of the markers that were previously tested on the Axiom biobank array were associated with a probeset that passed quality control, whereas 77% of the markers in the MVP 1.0 custom modules were associated with a probeset that remained after quality control. Additionally, although sample missingness (the fraction of missing genotype calls per individual; see Supplemental Materials and Methods) was slightly higher for vendor 1 than for vendor 2, almost all genotyped samples from both vendors exhibit missingness of less than 5% (Figure S6A).

We also either excluded or quarantined samples that did not meet sample QC criteria. Excluded samples include those expected to be removed by design or for known logistical or data errors. These samples include positive controls, samples with no or multiple unique participant identifiers, and samples in intentional duplicate pairs with the lower call rate. Quarantined samples are those that are temporarily removed from the dataset as a result of quality concerns. For instance, we investigated 1,149 pairs of samples with high relatedness to discriminate between misidentified intentional duplicates, technical duplicates (controls repeatedly genotyped by vendors), and monozygotic twins. Although we confirmed 49 monozygotic twins by cross-referencing with EHR data, the remaining 1,100 unintentional duplicate pairs could not be verified through independent means and were quarantined from data release as potentially mislabeled and will be re-genotyped. We also cross-checked genetically determined sample sex with EHR-reported gender information. Among the 485,856 unique genotyped samples, 2,000 (0.41%) did not have any reported gender information from either the EHR or self-reporting, and 2,073 (0.43%) of the remaining samples had a genetic sex that was opposite of the reported gender. We quarantined these samples for further analysis and potential re-genotyping (Table S2). The total number of samples that were excluded or quarantined from the current release of MVP genotype data and the reasons for exclusion are summarized in Table 1. All quarantined samples removed from the current data release will undergo further quality-control validation, be sent back to the vendors for re-genotyping, or will be otherwise verified before being included in subsequent data releases.

Marker Missingness and Discordance by MAF

We assessed marker missingness in correlation with MAF. Overall, the MAF distribution of MVP 1.0 is highly skewed toward rare variants; 42.89% of markers have a MAF below 1%, and 33.89% have a MAF below 0.1% (Figure 2A). This result is by design: the content of the MVP array focuses on markers associated with potential disease phenotypes. We find that MAF is correlated with marker missingness, as shown in Figures 2C and S6B, and lower frequency variants are missing in a larger fraction of samples. Despite this trend, genotyping call rate among low-frequency markers is still relatively high. For example, 87.29% of rare markers

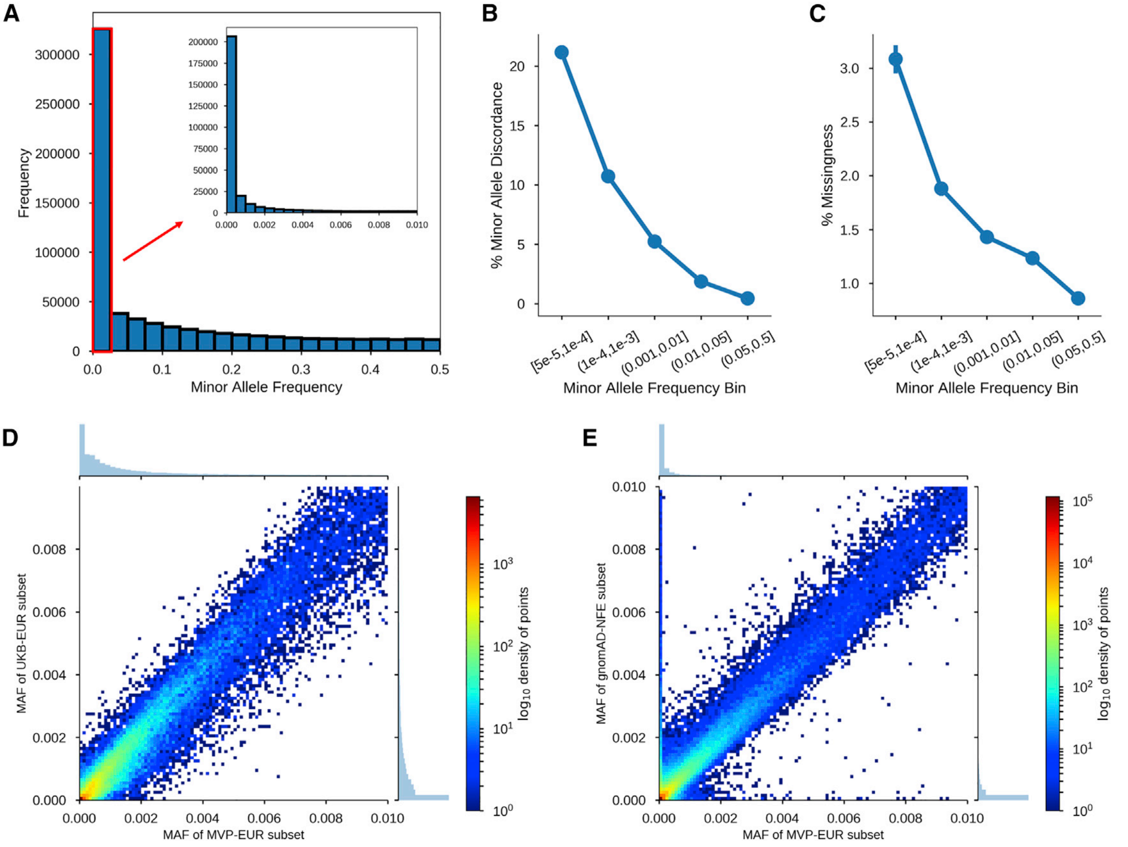


Figure 2. Quality-Control Assessments on the MVP Dataset after Performance of the Advanced Marker Quality Control Procedures
(A) MAF distribution after sample QC filtering. The inset diagram shows the distribution for markers with a MAF below 1%.
(B) Minor-allele discordance rates per MAF bin, based on intentionally duplicated samples.
(C) Marker missingness rates per MAF bin, after sample QC filtering.
(D) Comparison of MAFs between the EA subset of MVP (MVP-EUR) and the UK Biobank European subset (UKB-EUR).
(E) Comparison of MAFs between MVP-EUR and the non-Finnish European subset of gnomAD (gnomAD-NFE).

(MAF $<$ 0.1%) have a genotyping call rate of greater than 95%.

Additionally, we examined marker genotype discordance rates across intentional duplicate sample pairs with respect to MAF. Discordance is calculated per MAP for each marker, and markers are binned by MAF. We found a correlation between MAF and discordance rate, such that lower-frequency variants had a higher rate of minor-allele discordance (Figures 2B and S6C).

Duplicate and Positive Control Samples for Continuous Quality Assessment

Importantly, because we employed two separate vendors for genotyping, we intentionally included 25,291 duplicate samples that were blinded to the vendors for independent assessment of genotype quality. This amounts to a target of 5% of all genotyped samples and is an effort to accurately assess genotyping quality on a continuous basis. Sample concordance among intentional duplicates or positive controls was very high; the median concordance rate was greater than 99.8% across all comparisons (Figure S7A).

Assessing concordance in positive-control samples also provides valuable information about the consistency and reproducibility of the MVP 1.0 array's genotypes over time. Along with the MVP samples, 2,064 positive-control samples were genotyped on the MVP 1.0 array. As discussed in the Materials and Methods section, we constructed a consensus genotype sequence across 657,459 markers by using this panel of positive controls. For markers in the consensus sequence, 543,691 (82.70%) were homozygous, 95,079 (14.46%) were heterozygous, and 18,689 (2.84%) were uncalled. Concordance for each of the 2,064 positive-control samples is defined as the number of markers that agree with the consensus sequence divided by the number of called markers in the consensus sequence.

Overall positive-control concordance is shown in Figure S7A, and the distributions by batch of concordance values across all positive controls are shown in Figures S7B–S7D. The median concordance rate between each positive-control sample and the consensus sequence was 99.93% for all markers, 99.89% for common (MAF $\geq$ 5%) markers, and 100.00% for low-frequency (MAF $<$ 5%) markers. The minimum observed concordance rate between a positive control and the consensus occurs during analysis of common markers, but this concordance rate is still high at 99.05%.

Table 2. Concordance Rates across 96 HapMap Samples Genotyped on the MVP 1.0 Array

–	–	Metrics over Recommended[a] Markers		Metrics over All Markers	
Population	**Number of Samples**	**Average Sample Concordance (%)**	**Average Sample Call Rate (%)**	**Average Sample Concordance (%)**	**Average Sample Call Rate (%)**
ALL	96	99.70	99.85	99.35	99.49
CEU	28	99.70	99.85	99.34	99.47
CHB	20	99.70	99.86	99.37	99.51
JPT	20	99.68	99.84	99.35	99.51
YRI	28	99.71	99.86	99.34	99.49

[a] Recommended markers are those that were classified into one of the recommended SNP classes after execution of the Axiom Best Practices Genotyping workflow for the 96 co-clustered samples.

Concordance with HapMap Samples

To further test concordance and genotyping quality, we genotyped 96 HapMap samples (from Coriell cell lines) on the MVP 1.0 array. 210,630 markers are present in both the MVP 1.0 array and HapMap release 27, and among these markers, 205,647 (97.20%) are classified as recommended (see Supplemental Materials and Methods, Standard Marker Quality Control). When these 205,647 markers were analyzed over the 96 HapMap samples, and when HapMap and Axiom uncalled genotypes were removed from the numerator and denominator, the sample concordance across all population groups is 99.70% (Table 2). The Axiom sample call rate for recommended markers is 99.85%.

Assessing Rare-Allele Genotyping Quality

Given the importance of rare markers in clinically related studies, we evaluated the analytical validity of MVP 1.0 rare markers by observing the concordance of MAFs for rare markers with overlap between MVP 1.0 and either the gnomAD or the UK Biobank (Figures 2D and 2E). These databases are large enough for detection of very low MAFs, and agreement of MVP 1.0 marker MAFs with MAFs from these databases provides evidence for the accuracy of MVP 1.0 calls. MAFs were considered to agree when the lower bound of the regression slope's 95% confidence interval was ≥ 0.9. This value leaves some margin of error for expected differences between the databases in population structure (non-Finnish Europeans versus European Americans [EA]), technology (genotype arrays versus exome sequencing), technical processes (batch, user, etc.), and sample size. We used the MVP EA subgroup to benchmark performance because it has a larger sample size, which provides better confidence in assessing frequency of rare markers, and it has large complementary subgroups in gnomAD and the UK Biobank. We classified markers into three subgroups by MAF: rare variants ($< 1\%$), low-frequency variants (1%–5%), and common variants ($> 5\%$). The EA subgroup yielded 321,290 (48.1%) rare markers, 46,626 (6.97%) low-frequency markers, and 300,375 (44.9%) common markers.

From the gnomAD, we compared the allele frequencies derived from the non-Finnish European subgroup ($n = 55,860$) of the exome call set. This subgroup provided the largest cohort that was comparable in population structure. 69% (221,374 of 321,290 markers) of the rare variants in MVP were also found in gnomAD. Additionally, both MVP and gnomAD showed similar MAFs for these concordant rare variants (slope 0.9290, 95% CI: 0.9002, 0.9578).

From the UK Biobank, we compared allele frequencies derived from the self-reported white British ancestry group ($N > 330,000$). We found MAF agreement, as supported by the strong coefficient of determination (R^2) of 0.9864 and a slope of 0.9536 (95% CI: 0.9841, 0.9887) between 46,872 overlapping markers.

Although comparison against both sources met the ≥ 0.9 agreement threshold, we observed a small set of about 6,000 extremely discrepant markers (defined as having MAF > 0.001 in one database but MAF < 0.001 in the other) between MVP and gnomAD. About 53% of these markers were also present in the UK Biobank. For these discrepant markers, MAFs in the UK Biobank were much closer to MVP MAFs than those in gnomAD, and only one-quarter of the overlapping UK Biobank markers retained the "extremely discrepant" label. This is expected and consistent with previous observations that MAFs of MVP and the UK Biobank are in close agreement. The extremely discrepant markers between MVP and gnomAD might be attributed to smaller sample size of the gnomAD-exome database in comparison to the UK Biobank. The lowest MAF limit for MVP's EA subgroup is 1.6×10^{-6} (1 of 622,730 total alleles), 8.9×10^{-6} (1 of 111,720) for gnomAD's non-Finnish subgroup, and 1.4×10^{-6} (1 of 674,398) for UK Biobank. At very low frequencies, the absolute difference between rare variants, but not necessarily the relative difference, will be small. A given marker with a MAF of 0.001 in MVP and 0.01 in gnomAD will have an absolute difference of 0.009, but a relative difference of 10-fold. This is a common situation in our pairwise marker comparisons because overlapping marker MAFs are heavily clustered near zero (Figures 2D and 2E). This could also explain the relatively higher variance observed in the lower extremes when MVP is compared to gnomAD versus the UK Biobank. Overall, our results nonetheless show that our rare variant calls are highly consistent and within a reasonable range of agreement with overlapping markers

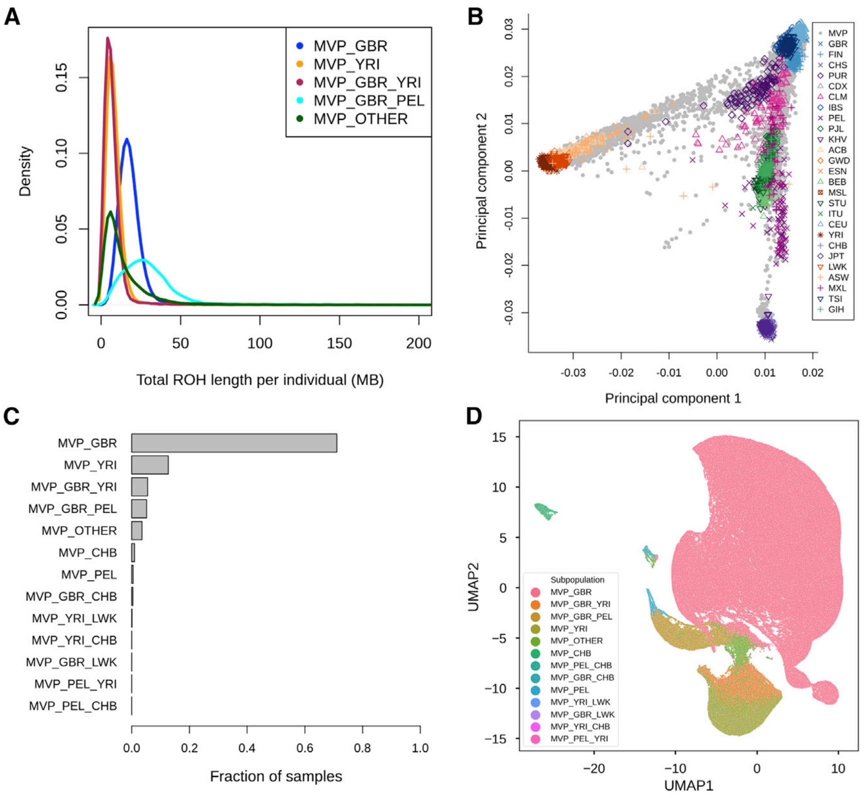


Figure 3. Analysis of Genetic Ancestry in the MVP Dataset

(A) Density plots of the total length of runs of homozygosity (ROH) per individual in each genetic-ancestry subgroup. Only the top five most common subgroups are shown.

(B) Principal-component analysis of the 1000 Genomes Project phase 3 dataset with MVP samples projected onto principal components 1 and 2.

(C) The number of MVP samples in each genetic-ancestry subgroup as inferred by ADMIXTURE percentages and our thresholds. For a single ancestry subgroup, such as MVP_GBR, the threshold is at least 80% inferred for that ancestry (e.g., MVP_GBR is GBR > 80%). For a pair of identified subgroups, the two ancestries must be at least 90% combined (e.g., MBP_GBR_YRI is GBR + YRI > 90%). MVP_OTHER includes all samples that had less than 80% ancestry aligned to any reference population and less than 90% combining any two populations.

(D) Visualization of ancestry subgroups via Uniform Manifold Approximation Projection (UMAP).

in gnomAD and the UK Biobank. However, it is important to note that the precision of calling very rare variants assayed with SNP chips has been reported to show variable quality.[20] Thus, visual inspection of calls underlying initial association results are always required.

Population Analysis of MVP Samples and a Test GWAS on Height

The MVP Cohort

In addition to quality assessment of MVP 1.0 genotyping results, we also performed exploratory analysis of the current population represented in the MVP samples. On the basis of data from the VistA EHR, the genotyped participants in the MVP cohort had a median age of 65 years at time of enrollment, and 8.33% are female. Although the percentage of female participants is low, reflecting the demographics of the Veteran population, this percentage corresponds to 46,924 female participants in the current release.

In light of the samples that have already been genotyped, the MVP cohort is relatively more diverse than other large biobanks on which data are available. For example, more than 94% of UK Biobank participants self-report as British, Irish, or "any other white background"[4,12], and 81% of individuals whose data are included in the Kaiser RPGEH biobank report as "white, non-Hispanic." On the other hand, 70.9% of MVP participants self-report as "white" and "non-Hispanic or Latino," in agreement with United States 2010 census information indicating that 63.7% of respondents self-report as "white alone" and "not Hispanic or Latino"[21].

Analysis of Relatedness

We examined the degree to which samples in the MVP population are related. Of the approximately 105.70 billion possible MVP sample pairings, 15,384 pairs appeared to be third-degree relatives or closer. The number of pairs for each type of relative pair, including trios, is shown in Table S8. Compared with the UK Biobank, this installment of MVP samples has a reduced fraction of related pairs.

Analysis of Genetic Ancestry

Assessing genetic ancestry for genotyped samples is an important tool for many applications, such as correcting for biases caused by population structure, constructing tests for natural selection, and determining disease risk by genetic ancestry, among other tasks.[22] To assess genetic ancestry in our sample, we visualized and then quantitatively assessed the genetic ancestry of MVP samples relative to external reference populations.

Runs of homozygosity (ROH) were measured via PLINK with a minimum ROH length of 1,000 Kb. The median total length of ROH is approximately 15.65 Mb, and the median number of blocks per sample is 10. In Figure 3A, we plotted the total length of ROH per individual by genetic ancestry subgroup for the five most common subgroups as defined in the Materials and Methods. MVP_GBR_PEL samples have a wide distribution of total ROH length but also some of the longest total lengths of all samples.

Samples that had African ancestry or that were admixed between three or more reference populations (MVP_OTHER) have the shortest total length of ROH per sample. Samples of mainly European ancestry have intermediate total ROH length. The total length of ROH per sample varies depending on the genetic-ancestry subgroup.

We also compared MVP samples to those in the 1000 Genomes Project. We first ran a PCA on the 1000 Genomes Project phase 3 samples and then projected the MVP samples onto these PCs. We found that most MVP samples lie close to reference populations of European origin. In addition, when we performed PCA on MVP samples alone, we found that genetic ancestry subgroups contain more complex intercontinental population structure, and a sizeable fraction of MVP samples exhibit admixture with respect to African and Asian references samples (Figures 3B and S9).

To assess ancestry proportion for each sample in MVP, we ran the program ADMIXTURE in supervised mode by using five 1000 Genomes Project phase 3 reference populations: Han Chinese in Beijing, China (CHB); British in England and Scotland (GBR); Luhya in Webuye, Kenya (LWK); Peruvians from Lima, Peru (PEL); and Yoruba in Ibadan, Nigeria (YRI).[16] For most participants, the largest percentage of their genome aligns with the GBR population (Figure S8C). However, a substantial fraction of samples contains a moderate amount of genetic ancestry similar to the YRI reference population. Examples were also found of participants who have almost 100% of their genetic ancestry aligning to each of the five reference populations except for LWK. By using ADMIXTURE analysis results, we grouped the MVP samples into 16 subgroups and determined the proportion of MVP samples belonging to each (Figure 3C). For example, 326,777 samples have over 80% of their genome aligning with the GBR reference population (MVP_GBR), whereas 58,267 samples have 80% or more of their genome aligning with YRI (MVP_YRI). Excluding samples with more than 80% of their genome aligning to one reference population, 25,295 of the samples have 90% or more of their genome aligning with a combination of GBR and YRI reference populations (MVP_GBR_YRI). Approximately 16,351 samples (MVP_OTHER) have neither 80% of their genome aligning with one reference population nor 90% aligning with a combined pair, indicating substantial admixture between three or more reference populations.

Finally, we visualized the diverse ancestry composition of MVP by using a non-parametric dimensionality reduction method called uniform manifold approximation projection (UMAP) (Figure 3D). As shown through PCA and ADMIXTURE, the largest cluster corresponds to samples with largely European ancestry. In this visualization, the distance between samples and clusters is not to be directly interpreted as genetic distance. Although there are distinct clusters (such as the tight cluster of individuals with Asian ancestry on the top left corner, and another small cluster of probable Polynesians in the middle of the plot), most MVP samples of different ancestries form a large single cluster rather than individual ancestry clusters with distinct breaks. This large cluster shows a continuum of ancestry proportion that transitions from GBR on the top right to different levels of admixture with YRI and PEL proportions. This is in line with a previous report based on 32,000 US individuals in the National Geographic Genographic Project cohort.[23]

GWAS of Height

To further validate the quality of our genotype data and the utility of MVP 1.0 array, we conducted a GWAS of height in both the EA and African American (AA) MVP subpopulations. EAs were defined as individuals with a GBR proportion greater than 90%, and AAs were defined as individuals with a YRI proportion greater than 60% and less than 40% GBR on the basis of ADMIXTURE results (Figures S8A and S8B). Our GWAS of height within EA and AA cohorts showed moderate inflation of $\lambda_{GC} = 1.12$ and $\lambda_{GC} = 1.13$ and pseudo-heritability of 0.396 and 0.378, respectively,[19,24,25] levels comparable to those found in previous height association studies without genotype imputation.[26]

Of the 822 reported height associations listed in the GWAS catalog,[27] 230 were present in the MVP EA GWAS, and 209 were present in the MVP AA GWAS. We assessed whether we could replicate effect sizes and direction of effects for markers present in MVP EA and AA GWASs by plotting these against the GWAS-catalog effect sizes and direction of effects (Figure 4). For the two subpopulations, the MVP associations perfectly replicated the directions of effect in most markers (two SNPs had an effect size near zero in EA). However, because most associations in the GWAS catalog are derived from Europeans, the overall correlation across all markers was lower for the AA cohort ($r = 0.69$) than for the EA cohort ($r = 0.85$).

Overall, we show that the performance of MVP 1.0 and the quality of its genotyping across 459,777 individuals of diverse ethnic backgrounds is very consistent and accurate by a variety of metrics.

Discussion

In this report, we provide an overview of the design of the MVP 1.0 genotyping array, the development of accompanying quality-control analyses, and our initial data exploration of an interim MVP genotyping dataset that consists of nearly 460,000 veterans. Our results demonstrate that the MVP 1.0 chip and the subsequent QC procedures have addressed notable challenges characteristic of large projects with individuals of diverse genetic backgrounds and that the resulting genotype calls are of high quality akin to that of other projects similar in scope. By using a single chip and unified quality control across the diverse cohort, we aimed to minimize batch effects between different ancestries and provide an initial genome-wide scan before whole-genome-sequenced samples become available.

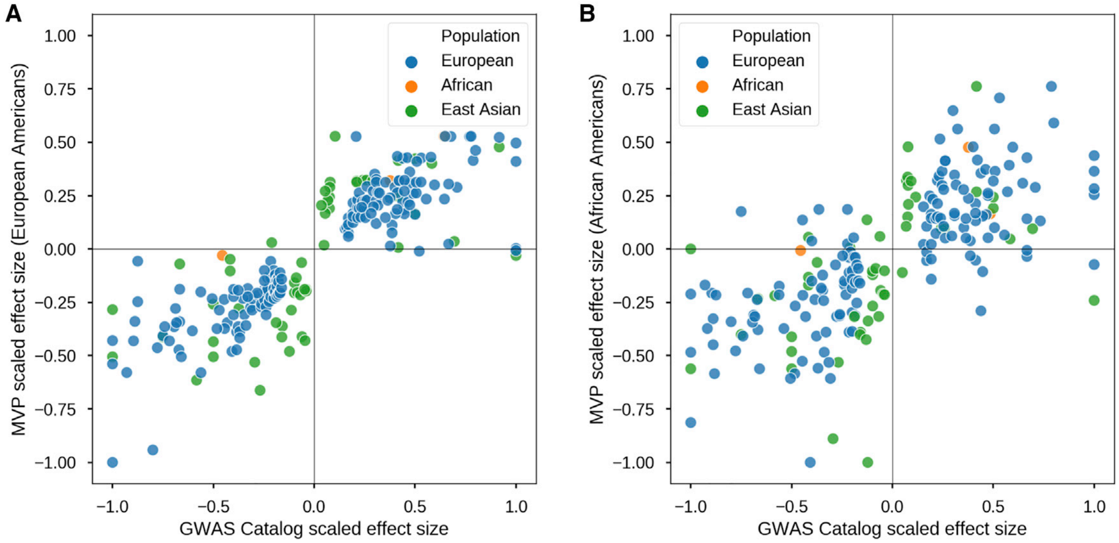


Figure 4. GWAS of Height with MVP Cohort
(A) Replication of the direction of effect for markers previously associated with height as annotated in the NHGRI-EBI GWAS Catalog in the MVP cohort of non-related European Americans (n = 291,609). Color coding denotes the genetic ancestry of the original cohort in which the markers were associated with height.
(B) Same as (A) except with the MVP cohort of non-related African Americans (N = 73,190).

Addressing the Challenges of MVP

MVP's large, diverse, and still-growing cohort poses numerous challenges for designing genotyping procedures and their subsequent quality-assessment and quality-control protocols. Genotyping large and ethnically diverse cohorts along with clinically relevant markers is even more challenging because of the finite number of probesets that can fit on a single array. However, using different arrays for different ethnic groups can also exacerbate the differences between these groups and lead to batch effects.

To address the limitations of array-based genotyping in diverse cohorts, we carefully selected array content to maximize clinical utility while at the same time ensuring both broad coverage of variants and robust imputation capabilities across different ethnic groups. We also developed comprehensive quality controls for markers and samples both before and after genotyping. These controls included intentional duplication of ~5% randomly selected samples over time, blinded-to-assay technicians so that batch variation could be detected and mitigated; assessment of genotyping concordance with positive control samples and HapMap samples (Figure S7A, Table 2); comparison of MVP 1.0 MAFs to those in gnomAD and the UK Biobank (Figure 2); and a height GWAS intended to replicate previously reported results (Figure 4). Overall, we retained and released 459,777 samples and 668,418 markers after QC for the initial release of data. Although QC metrics vary slightly over time and genotyping vendors, the final genotyped sample set shows consistently high call rates (98.5%) and genotype concordance over intentional duplicates (99.8%) both within and between vendors and over time. Furthermore, marker concordance is also high even for rare markers. Additionally, genotype concordance, MAF, and GWAS association results are generally in strong agreement with external or previously reported results. These results indicate that the design of the MVP 1.0 array and the associated quality-control and assessment procedures provide a robust, reliable method for genotyping common, low-frequency, and rare variants in a large, ethnically diverse cohorts.

Challenges remain however, and the MVP 1.0 array has several limitations. Notably, although concordance rates were high, our results demonstrate that low-frequency and rare variants are still more difficult to genotype accurately with the MVP 1.0 array than are common variants. Additionally, while although added markers to MVP 1.0 to increase coverage for AAs, we lack boosters for other ethnic groups, such as Asian and Native American populations, which currently comprise smaller but growing proportions of the MVP population. In addition, although a standard imputation to the 1000 Genomes reference has been completed, we have yet to quantify the effect of imputation across different ancestries within MVP to devise an optimized imputation strategy. The strategy implemented by the UK Biobank was to use the Haplotype Reference Consortium (HRC) as the main imputation reference panel and to supplement with variants from 1000 Genomes if those variants were missing in HRC. However, this is not viable for MVP, which has a large proportion of non-European individuals. Further understanding and analyses of imputation strategies in a multi-ethnic and admixed cohort will be required if we are to obtain an optimized strategy for MVP.

The MVP Dataset Is Ethnically and Genetically Diverse

Our exploratory analysis indicates that the MVP dataset and samples offer unique value for disease research. One particularly valuable aspect of the MVP dataset is the ethnic diversity it encompasses. Genetic ancestry analysis suggests that the MVP dataset contains sub-populations with both homogeneous and admixed genetic ancestry from multiple global populations. The largest sub-population corresponds to 71% of samples of mostly European descent, and the remaining samples show substantial African, East Asian, and Native American ancestry.

Because MVP recruits participants from United States veterans who receive care at VA hospitals, the demographics of the MVP dataset diverge from those of the United States population. Approximately 8.5% of MVP samples are female, which is similar to the fraction of women in the Veteran population.[28] With a median age of 68 years as opposed to 37.9 years, MVP participants are also substantially older than the United States population.[29] However, the demographics of MVP might change with increasing use of the VA by more recent veterans. The proportion of female veterans is projected to continuously grow and nearly double, to 16.5%, by 2043.[28] Meanwhile, the proportion of veterans from minority populations is expected to increase by approximately 50% over the same time period.[28] Thus, the VA and MVP is in a unique position for further inclusion of participants from diverse backgrounds.

The MVP Dataset Is an Invaluable Disease Research Resource

MVP has several unique features that make it an invaluable resource for researching human disease. As evidence of the general utility of this resource, initial reports using an earlier tranche of ~300,000 genotyped participants have reported substantial new findings regarding the genetics of blood lipids, a major cardiovascular risk factor.[30] Not only is MVP ideal for studying the burden of chronic disease, which increases with age, many of the clinical records in its EHR span several decades, allowing for robust longitudinal analysis. This is possible because patients using the VA health services do not lose coverage even after changing employers or residence. Additionally, MVP provides an opportunity to study diseases, such as PTSD and[31] alcohol- and substance-abuse disorders,[32] that disproportionately affect US veterans, as well as to study other deployment-related conditions and their impact on human health. The MVP phenotypes are collected in the VA EHR as part of the routine clinical care at VA national hospitals and clinics across the country, and clinical data spanning the past three decades are available to be integrated with genomic data on demand. A list of prevalent disease phenotypes surveyed from raw International Classification of Diseases (ICD) diagnostic codes in the VA EHR are provided in Table S9. Although ethnically diverse, the MVP cohort is not a true representation of the US general population, given the underrepresentation of women and requirements for military service. For instance, early-onset disease phenotypes (e.g., Mendelian diseases with large effect sizes) that might limit military service are underrepresented.

In conclusion, the high-quality genotype data generated with the MVP 1.0 array provides a valuable resource for researchers investigating the effect of both rare and common genetic variants within MVP. These quality-controlled genotype data as well as the results from genetic ancestry and relatedness analyses are made available to all approved researchers. MVP intends to make coded data available in a secure data and computing environment after periodic requests for proposals, where the data use intent and provenance will be clearly verified in accordance with the participant consent and MVP policies. The genotype data can be linked to participants' full EHR, often covering decades of care provided by the VA. MVP is a continuously expanding research cohort made available by participants with diverse backgrounds and altruistic intentions to support research that will benefit their fellow veterans and others.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.03.004.

Consortia

Million Veteran Program (MVP) Consortium

Michael Gaziano, MD, MPH, Rachel Ramoni, DMD, ScD, Jim Breeling, MD, Kyong-Mi Chang, MD, Grant Huang, PhD, Sumitra Muralidhar, PhD, Christopher J. O'Donnell, MD, MPH, Philip S. Tsao, PhD, Sumitra Muralidhar, PhD, Jennifer Moser, PhD, Stacey B. Whitbourne, PhD, John Concato, MD, MPH, Stuart Warren, JD, Pharm D, Dean P. Argyres, MS, Brady Stephens, MS, Mary T. Brophy MD, MPH, Donald E. Humphries, PhD, Xuan-Mai T. Nguyen, PhD, Saiju Pyarajan, PhD, Kelly Cho, MPH, PhD, Peter Wilson, MD, Rachel McArdle, PhD, Louis Dellitalia, MD, John Harley, MD, Jeffrey Whittle, MD, Jean Beckham, PhD, John Wells, PhD, Salvador Gutierrez, MD, Gretchen Gibson, DDS, Laurence Kaminsky, PhD, Gerardo Villareal, MD, Scott Kinlay, PhD, Junzhe Xu, MD, Mark Hamner, MD, Kathlyn Sue Haddock, PhD, Sujata Bhushan, MD, Pran Iruvanti, PhD, Michael Godschalk, MD, Zuhair Ballas, MD, Malcolm Buford, MD, Stephen Mastorides, MD, Jon Klein, MD, Nora Ratcliffe, MD, Hermes Florez, MD, Alan Swann, MD, Maureen Murdoch, MD, Peruvemba Sriram, MD, Shing Shing Yeh, MD, Ronald Washburn, MD, Darshana Jhala, MD, Samuel Aguayo, MD, David Cohen, MD, Satish Sharma, MD, John Callaghan, MD, Kris Ann Oursler, MD, Mary Whooley, MD, Sunil Ahuja, MD, Amparo Gutierrez, MD, Ronald Schifman, MD, Jennifer Greco, MD, Michael Rauchman, MD, Richard Servatius, PhD, Mary Oehlert, PhD, Agnes Wallbom, MD, Ronald Fernando, MD, Timothy Morgan, MD, Todd Stapley, DO, Scott Sherman, MD, Gwenevere Anderson, RN, Philip Tsao, PhD, Elif Sonel, MD, Edward Boyko, MD, Laurence Meyer, MD, Samir Gupta, MD, Joseph Fayad, MD, Adriana Hung, MD, Jack Lichy, MD, PhD, Robin Hurley, MD, Brooks Robey, MD, Robert Striker, MD.

Acknowledgments

This work was funded by the United States Veterans Administration (VA) Office of Research and Development (ORD), Million Veteran Program Grant number MVP000 to S.P. and the VA Special

Fellowship in Medical Informatics to H.H.-Z. We would like to thank all Veteran participants in the MVP for donating their samples, information, and time to this project. We additionally thank Anoop Sendamarai and Frederick Dong for critical discussions and providing editorial assistance and review. The content of this manuscript reflects the views of the authors and does not represent the views of the Department of Veterans Affairs, FDA, or the United States Government.

Declaration of Interests

The authors declare no competing interests.

Received: December 22, 2019
Accepted: March 6, 2020
Published: April 2, 2020

Web Resources

gnomAD, https://gnomad.broadinstitute.org/
UK Biobank, https://gbe.stanford.edu
UK Biobank, https://github.com/rivas-lab/public-resources/blob/master/uk_biobank/variant_filter_table.tsv

References

1. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223.
2. Banda, Y., Kvale, M.N., Hoffmann, T.J., Hesselson, S.E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L.A., Dispensa, B.P., Henderson, M., et al. (2015). Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics *200*, 1285–1295.
3. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A., et al. (2015). Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics *200*, 1051–1060.
4. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.
5. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. *375*, 655–665.
6. Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. *17*, 157.
7. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
8. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
9. Affimetrix. (2016). Axiom genotyping solution data analysis guide.
10. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
11. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
12. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
13. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
14. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
15. Galinsky, K.J., Loh, P.R., Mallick, S., Patterson, N.J., and Price, A.L. (2016). Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. Am. J. Hum. Genet. *99*, 1130–1139.
16. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
17. Diaz-Papkovich, A., Anderson-Trocme, L., and Gravel, S. (2019). Revealing multi-scale population structure in large cohorts. bioRxiv. https://doi.org/10.1101/423632.
18. Noël, P.H., Copeland, L.A., Pugh, M.J., Kahwati, L., Tsevat, J., Nelson, K., Wang, C.P., Bollinger, M.J., and Hazuda, H.P. (2010). Obesity diagnosis and care practices in the Veterans Health Administration. J. Gen. Intern. Med. *25*, 510–516.
19. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., Corvin, A., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
20. Weedon, M.N., Jackson, L., Harrison, J.W., Ruth, K.S., Tyrrell, J., Hattersley, A.T., and Wright, C.F. (2019). Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing. bioRxiv. https://doi.org/10.1101/696799.
21. United States Census Bureau (2011). The White Population: 2010–c2010. In 2010 Census Briefs, pp. 1–20.
22. Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. Front. Genet. *5*, 204.
23. Dai, C.L., Vazifeh, M.M., Yeang, C.-H., Tachet, R., Wells, R.S., Vilar, M.G., Daly, M.J., Ratti, C., and Martin, A.R. (2019). Population histories of the United States revealed through fine-scale migration and haplotype analysis. bioRxiv. https://doi.org/10.1101/57741.
24. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

25. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

26. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.

27. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

28. National Center for Veterans Analysis and Statistics (2014). Table 3L: Living veterans by race/ethnicity, gender, pp. 2013–2043.

29. World Factbook, C.I.A. (2017). Central Intelligence Agency (The World Factbook).

30. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.

31. Gelernter, J., Sun, N., Polimanti, R., Pietrzak, R., Levey, D.F., Bryois, J., Lu, Q., Hu, Y., Li, B., Radhakrishnan, K., et al. (2019). Genome-wide association study of post-traumatic stress disorder reexperiencing symptoms in >165,000 US veterans. Nat. Neurosci. *22*, 1394–1401.

32. Kranzler, H.R., Zhou, H., Kember, R.L., Vickers Smith, R., Justice, A.C., Damrauer, S., Tsao, P.S., Klarin, D., Baras, A., Reid, J., et al. (2019). Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. Nat. Commun. *10*, 1499.

The American Journal of Human Genetics, Volume *106*

Supplemental Data

Genotyping Array Design and Data Quality Control in the Million Veteran Program

Haley Hunter-Zinck, Yunling Shi, Man Li, Bryan R. Gorman, Sun-Gou Ji, Ning Sun, Teresa Webster, Andrew Liem, Paul Hsieh, Poornima Devineni, Purushotham Karnam, Xin Gong, Lakshmi Radhakrishnan, Jeanette Schmidt, Themistocles L. Assimes, Jie Huang, Cuiping Pan, Donald Humphries, Mary Brophy, Jennifer Moser, Sumitra Muralidhar, Grant D. Huang, Ronald Przygodzki, John Concato, John M. Gaziano, Joel Gelernter, Christopher J. O'Donnell, Elizabeth R. Hauser, Hongyu Zhao, Timothy J. O'Leary, VA Million Veteran Program, Philip S. Tsao, and Saiju Pyarajan

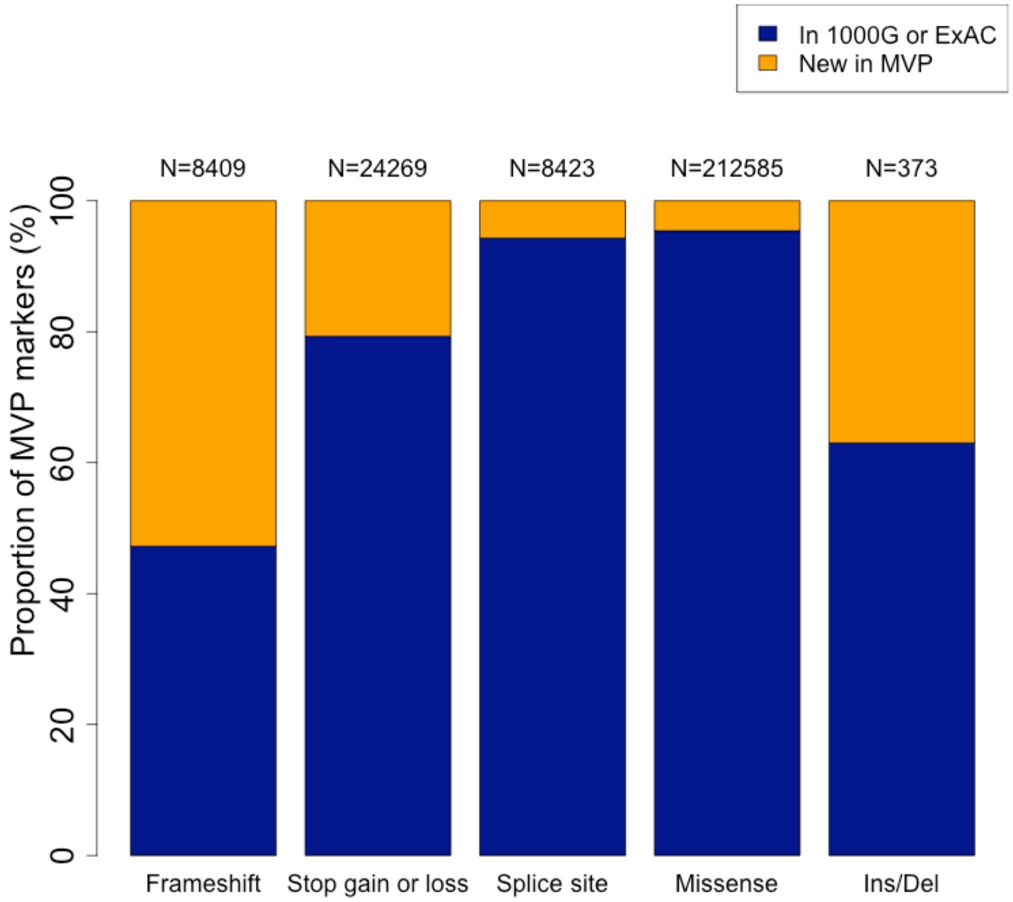


Figure S1. Comparison of MVP 1.0 markers to those in other databases. The distribution of MVP 1.0 Array markers by their functional annotation and whether they overlap with existing genetic variation databases, the 1000 Genomes Project and the ExAC database.

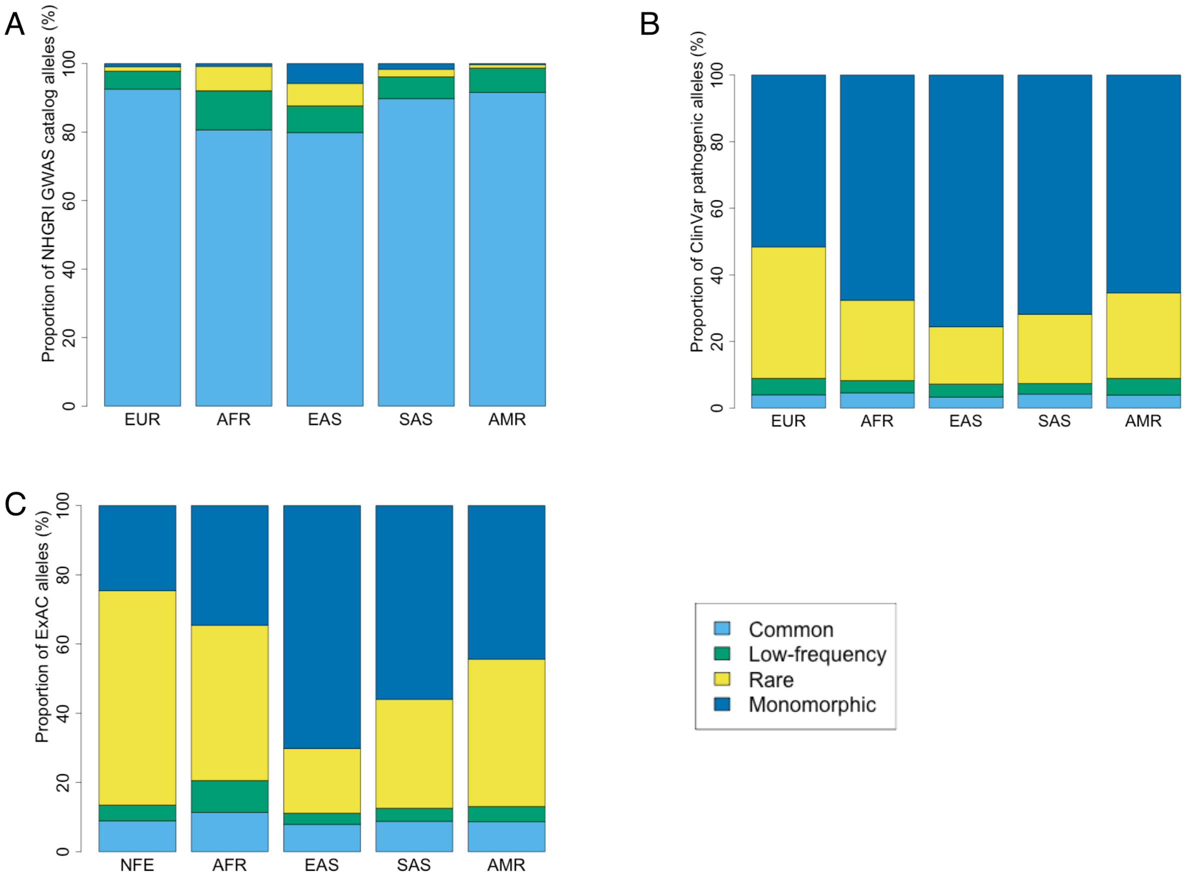


Figure S2. MVP 1.0 Array marker overlap with databases of functional and disease-related variants. Distribution of markers overlapping with (A) the NHGRI-EBI GWAS catalog; (B) the ClinVar database (2017 December release); and (C) the ExAC database by MAF bins and ancestral population groups. MAFs and ancestral population definitions were extracted for (A) and (B) from the 1000 Genomes Project and for (C) from the ExAC database. Common variants are markers with MAF ≥ 5%; low-frequency variants are markers with 1% ≤ MAF < 5%; rare variants are markers with MAF < 1%; and monomorphic variants are markers with MAF = 0%. (EUR: European, NFE: Non-Finnish European, AFR: African, EAS: East Asian, SAS: South Asian, AMR: Admixed American)

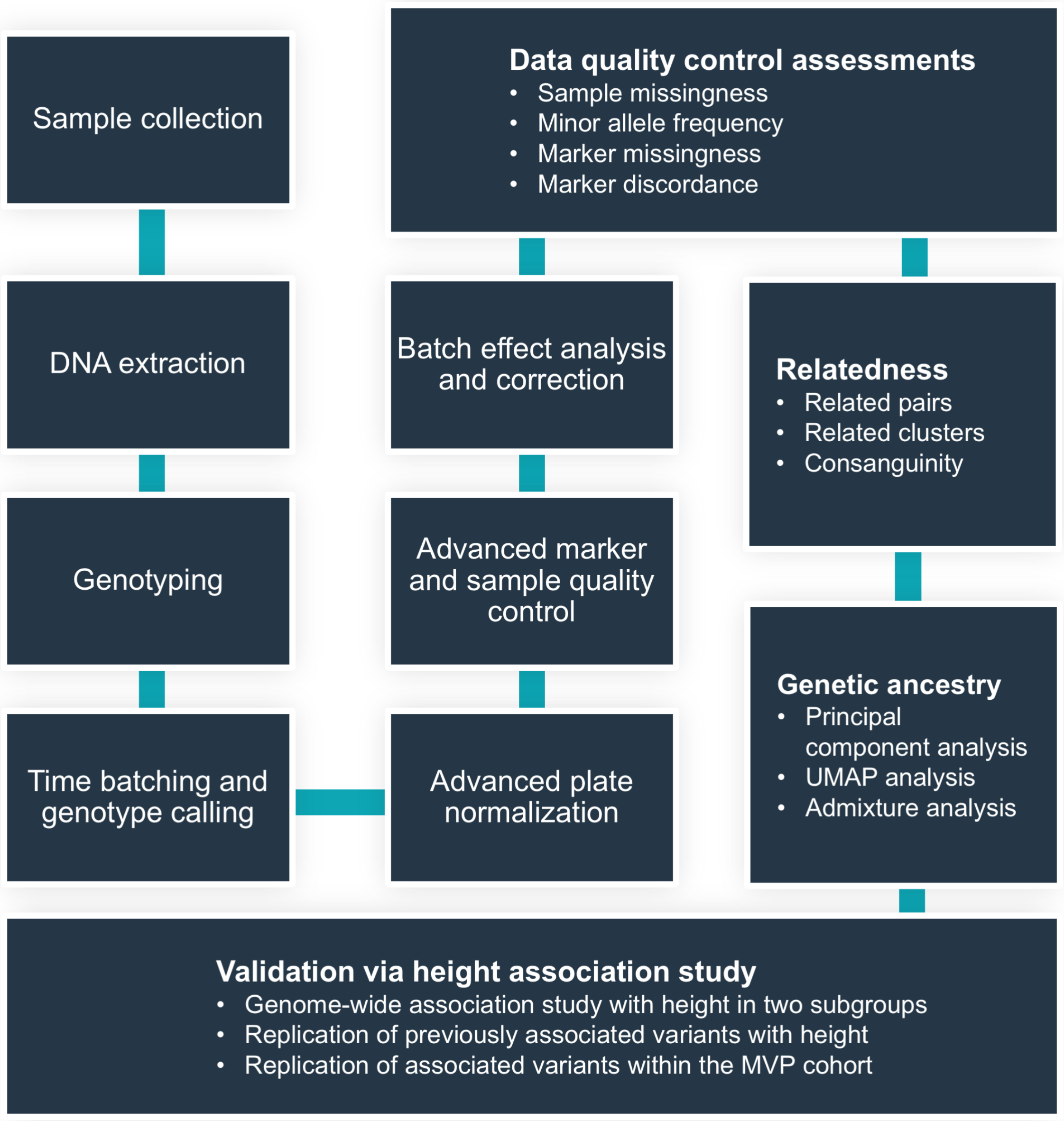
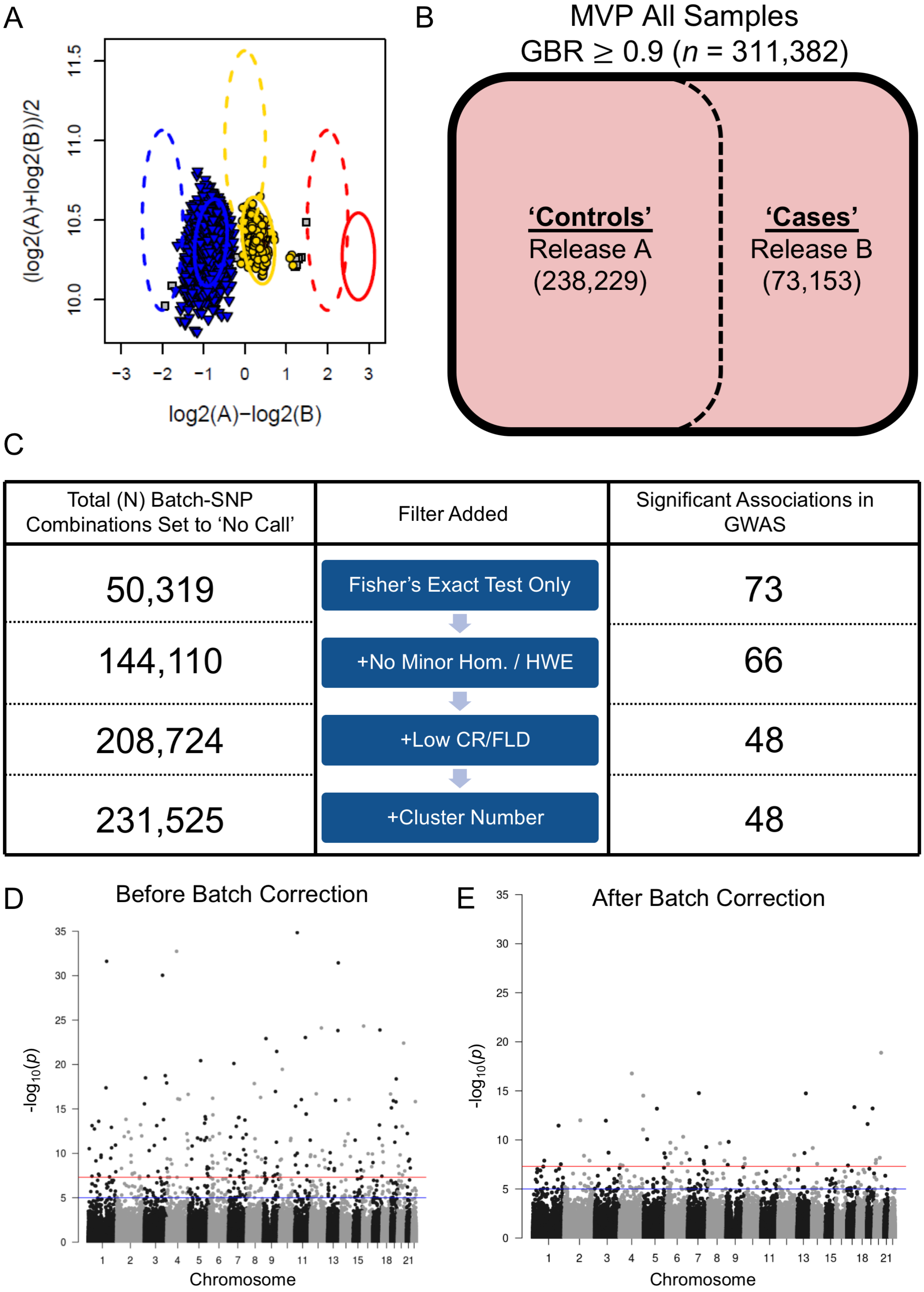

Sample collection
DNA extraction
Genotyping
Time batching and genotype calling
Data quality control assessments
• Sample missingness
• Minor allele frequency
• Marker missingness
• Marker discordance
Batch effect analysis and correction
Advanced marker and sample quality control
Advanced plate normalization
Relatedness
• Related pairs
• Related clusters
• Consanguinity
Genetic ancestry
• Principal component analysis
• UMAP analysis
• Admixture analysis
Validation via height association study
• Genome-wide association study with height in two subgroups
• Replication of previously associated variants with height
• Replication of associated variants within the MVP cohort

Figure S3. The overall genotyping quality control and MVP data analysis procedure. Overview of the steps of the genotype calling, quality control, data exploration, and final validation processes for the MVP dataset.



Total (N) Batch-SNP Combinations Set to 'No Call'	Filter Added	Significant Associations in GWAS
50,319	Fisher's Exact Test Only	73
144,110	+No Minor Hom. / HWE	66
208,724	+Low CR/FLD	48
231,525	+Cluster Number	48

Figure S4. **Batch variation correction.** (A) Example of a mis-clustered probeset in a batch detected using Fisher's Exact Test. (B) The GWAS experimental design used to detect batch effects across releases. Release A samples were designated "Controls," and Release B samples were designated "Cases." (C) Filters introduced in addition to Fisher's Exact Test, and the number of batch-SNP combinations removed and significant GWAS associations remaining after each step. (D) Manhattan plot showing 204 significant associations before batch correction. (E) Manhattan plot showing 48 remaining significant associations after application of batch correction procedure.

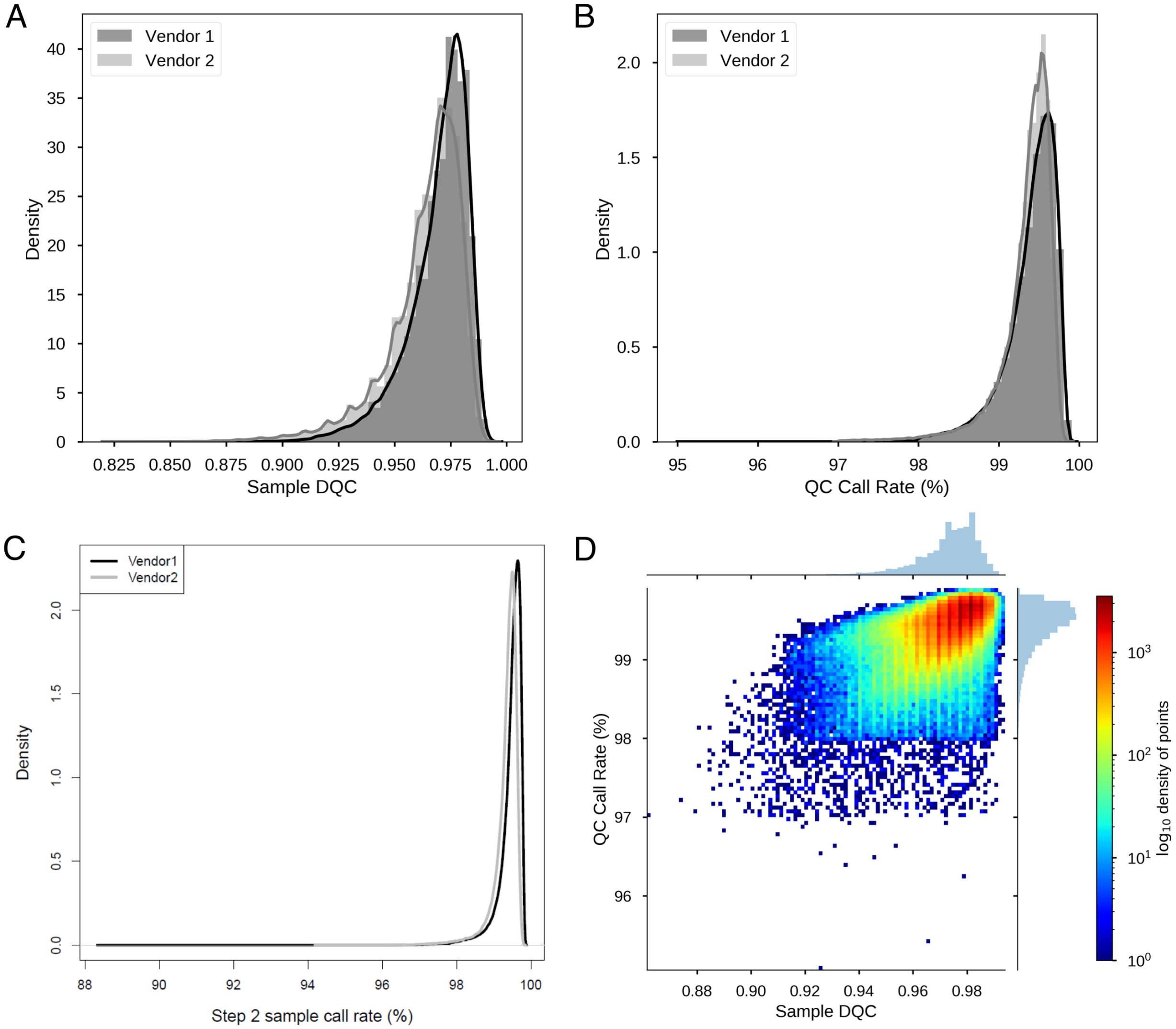


Figure S5. Axiom® genotyping DQC, sample QC call rate and sample call rate metrics. Metrics were reported directly from vendors, and distributions for each metric are for all samples genotyped before sample QC filtering. (A) Density curves of the distributions of the dish QC (DQC) metric by genotyping vendor. (B) Density curves of the distribution of sample QC call rate by genotyping vendor. Sample QC call rate is referred to as Step 1 call rate in the Axiom® Genotyping Solution Data Analysis Guide. (C) Density curves of the distribution of the sample call rate over all MVP 1.0 array probesets by genotyping vendor after re-batching and recalling genotypes at the VA. Sample call rate is referred to as Step 2 call rate in the Axiom® Genotyping Solution Data Analysis Guide. (D) Density plot of sample QC call rate against the DQC metric before sample quality control filtering. Color density is scaled by the log of the number of samples, with red indicating higher density and blue indicating lower density.

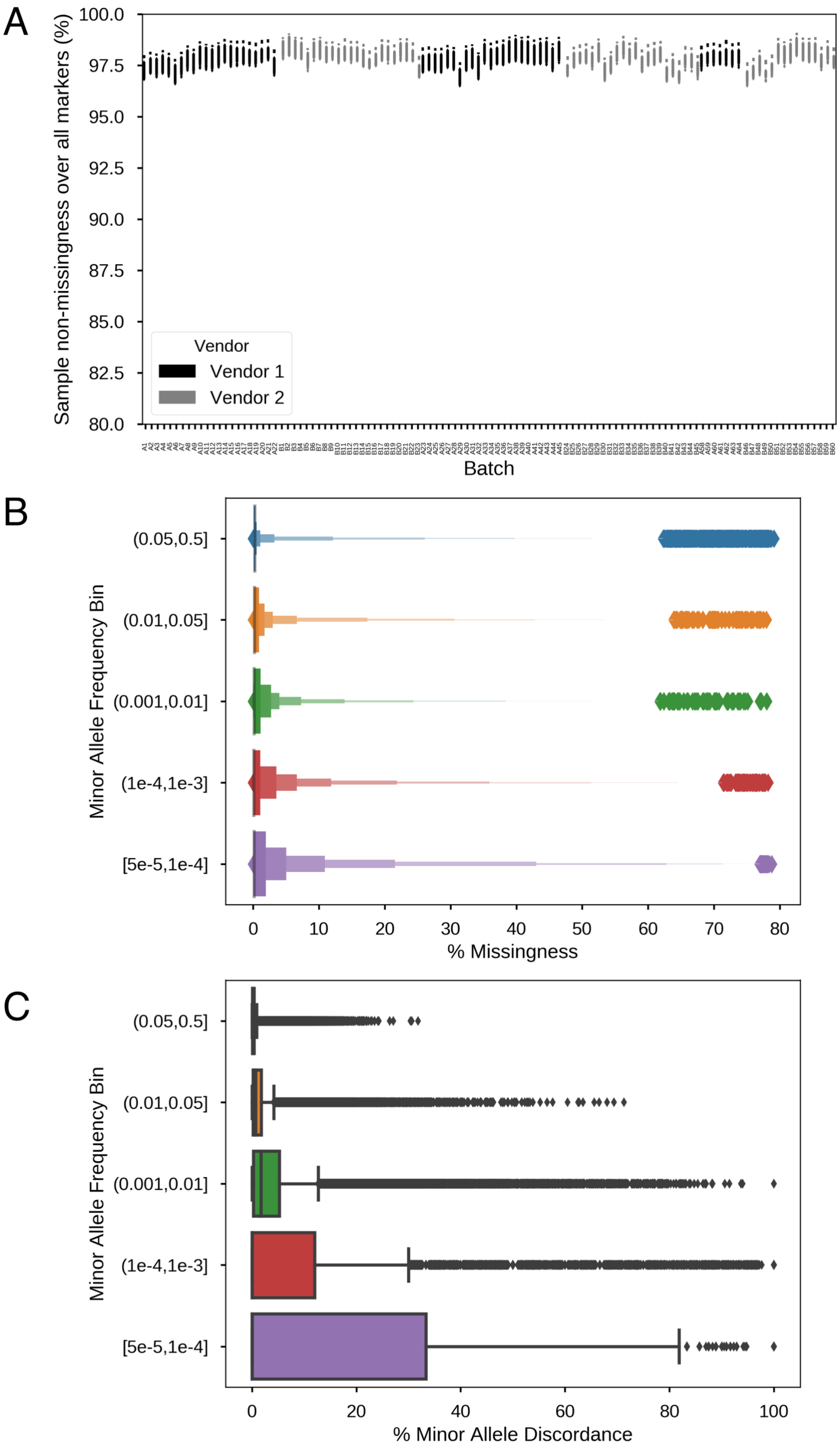

A
Sample non-missingness over all markers (%)
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
Vendor
Vendor 1
Vendor 2
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21 A22 B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20 B21 B22 B23 A23 A24 A25 A26 A27 A28 A29 A30 A31 A32 A33 A34 A35 A36 A37 A38 A39 A40 A41 A42 A43 A44 A45 B24 B25 B26 B27 B28 B29 B30 B31 B32 B33 B34 B35 B36 B37 B38 B39 B40 B41 B42 B43 B44 B45 A58 A59 A60 A61 A62 A63 A64 B46 B47 B48 B49 B50 B51 B52 B53 B54 B55 B56 B57 B58 B59 B60
Batch
B
Minor Allele Frequency Bin
(0.05,0.5]
(0.01,0.05]
(0.001,0.01]
(1e-4,1e-3]
[5e-5,1e-4]
0
10
20
30
40
50
60
70
80
% Missingness
C
Minor Allele Frequency Bin
(0.05,0.5]
(0.01,0.05]
(0.001,0.01]
(1e-4,1e-3]
[5e-5,1e-4]
0
20
40
60
80
100
% Minor Allele Discordance

Figure S6. Assessments of sample and marker missingness. (A) Sample non-missingness (i.e. 1 – missingness) for each batch. The X-axis conains identifiers for genotyping batches in sequential time order. (B) Percent marker missingness startified by MAF bin. (C) Minor allele discordance by MAF bin.

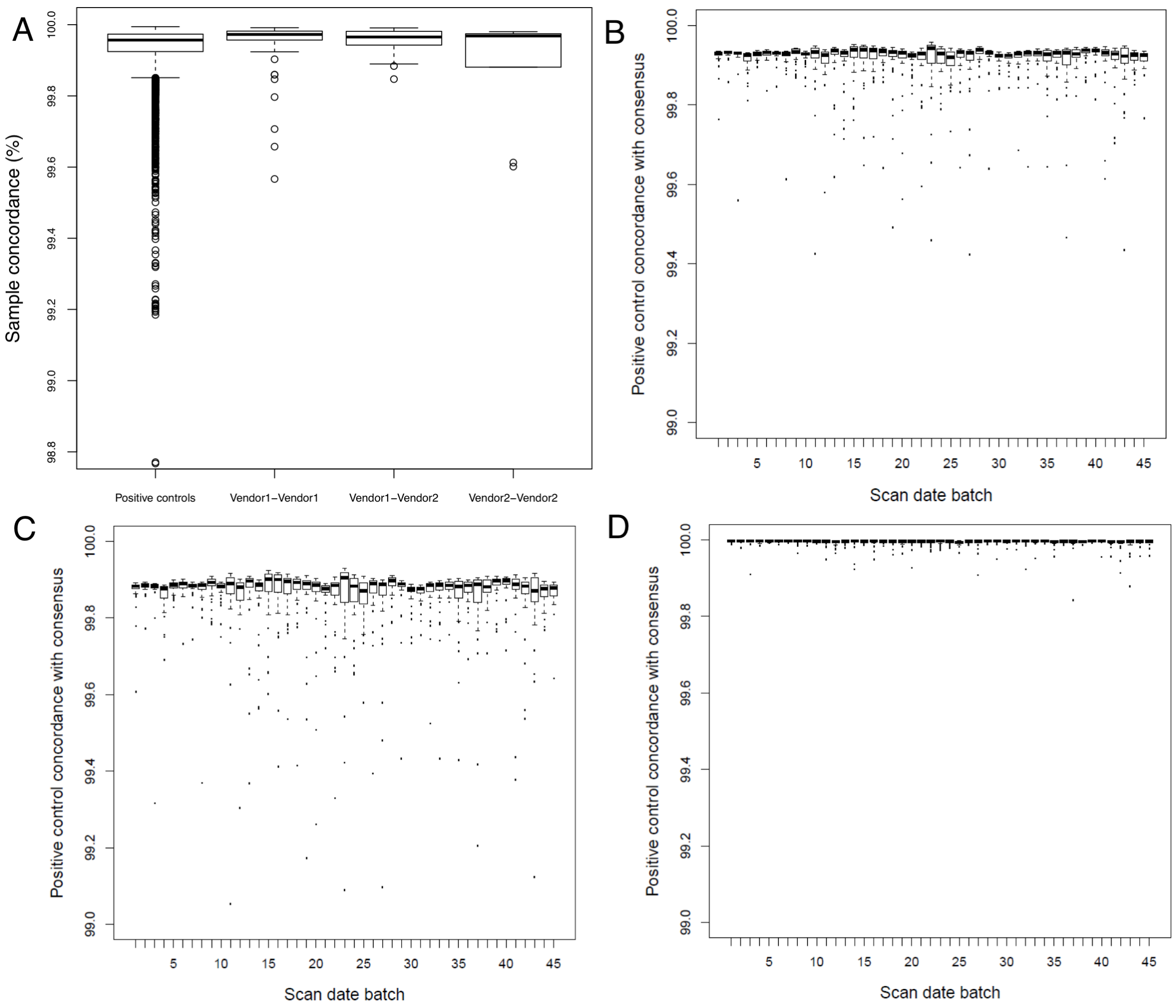


Figure S7. Sample and marker concordance. (A) Sample concordance over the marker set after the advanced marker QC procedure for pairs of positive controls and pairs of intentional duplicates genotyped both within and between vendors. Positive controls are only from Vendor 2, since Vendor 1 did not include positive controls on genotyped plates. Distributions are shown for a randomly selected 5,000 pairs for each category. (B-D) Marker concordance for (B) all markers, (C) common (MAF ≥ 5%) markers, and D) low-frequency (MAF < 5%) markers between positive control samples and the consensus positive control genotyping sequence, grouped by batch. Batches are ordered by time.

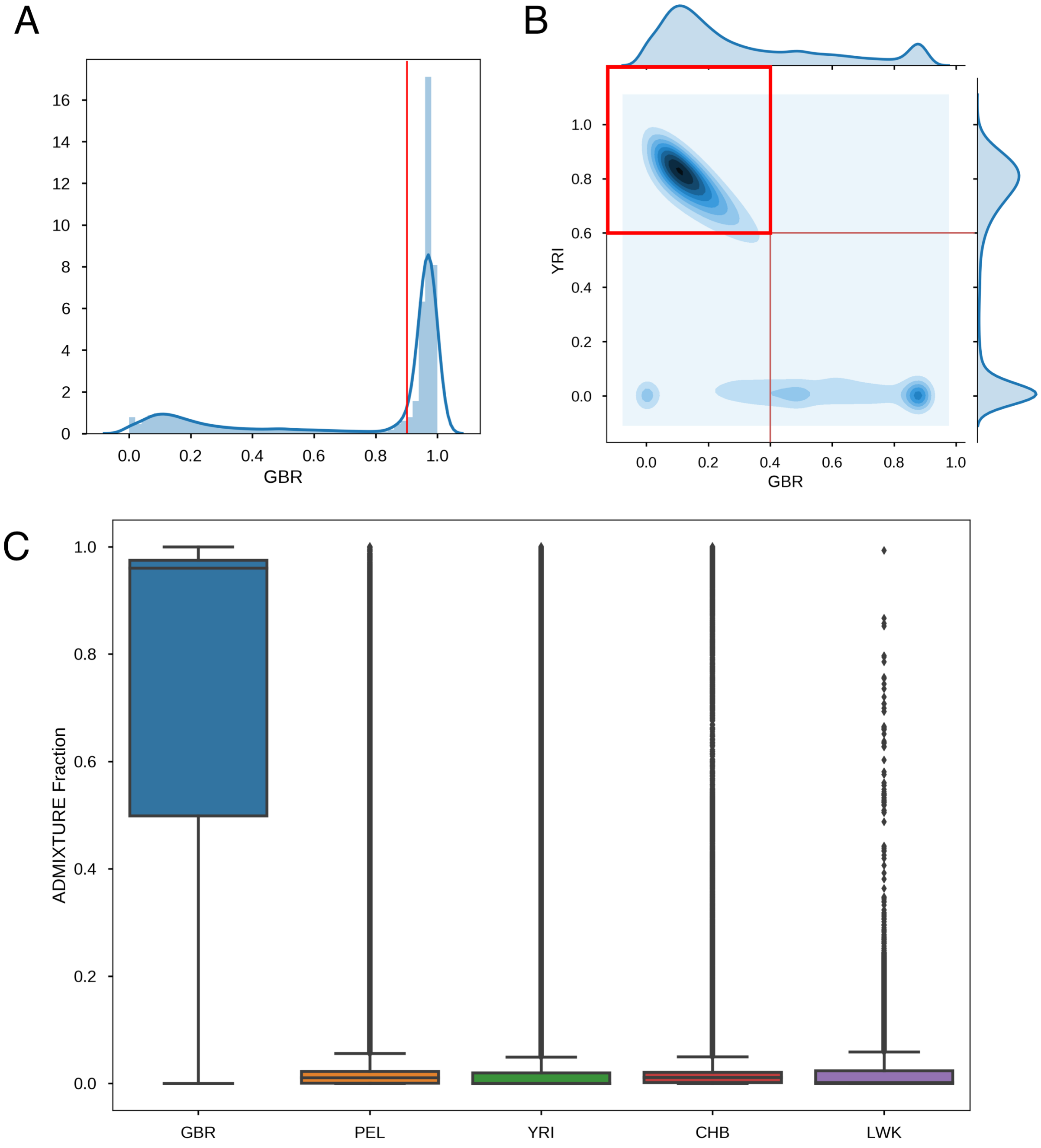


Figure S8. Genetic ancestry analysis of the MVP cohort using ADMIXTURE. (A) Distribution of British in England and Scotland (GBR) proportion in all MVP individuals. Individuals with more than 0.9 GBR proportion were defined as European American. (B) Kernel density estimates of Yoruba in Ibadan, Nigeria (YRI) and GBR proportion within individuals with less than 0.9 GBR proportion (i.e. non-European individuals). Individuals with more than 0.6 YRI and less than 0.4 GBR proportion were defined as African American. (C) Distribution of ADMIXTURE proportions for each of the five 1000 Genomes Project reference populations, sorted by average fraction, across all MVP samples. Reference population abbreviations: British in England and Scotland

(GBR); Yoruba in Ibadan, Nigeria (YRI); Peruvians from Lima, Peru (PEL); Han Chinese in Beijing, China (CHB); and Luhya in Webuye, Kenya (LWK).

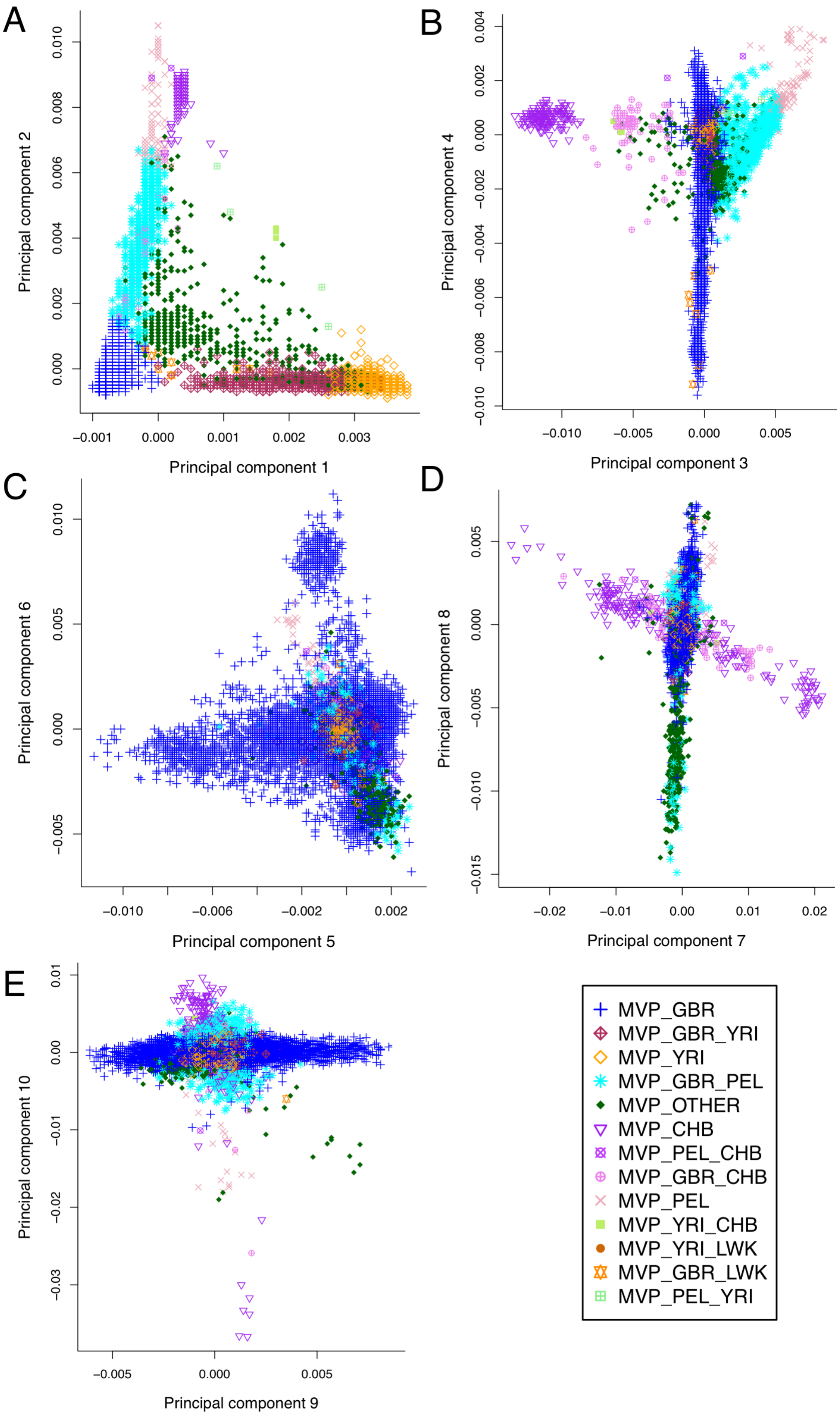

A
Principal component 2
0.010
0.008
0.006
0.004
0.002
0.000
−0.001
0.000
0.001
0.002
0.003
Principal component 1
B
Principal component 4
0.004
0.002
0.000
−0.002
−0.004
−0.006
−0.008
−0.010
−0.010
−0.005
0.000
0.005
Principal component 3
C
Principal component 6
0.010
0.005
0.000
−0.005
−0.010
−0.006
−0.002
0.002
Principal component 5
D
Principal component 8
0.005
0.000
−0.005
−0.010
−0.015
−0.02
−0.01
0.00
0.01
0.02
Principal component 7
E
Principal component 10
0.01
0.00
−0.01
−0.02
−0.03
−0.005
0.000
0.005
Principal component 9
MVP_GBR
MVP_GBR_YRI
MVP_YRI
MVP_GBR_PEL
MVP_OTHER
MVP_CHB
MVP_PEL_CHB
MVP_GBR_CHB
MVP_PEL
MVP_YRI_CHB
MVP_YRI_LWK
MVP_GBR_LWK
MVP_PEL_YRI

Figure S9. Ancestry Principal Component Analysis (PCA) of the MVP Cohort. Colors represent genetic ancestry subgroups as defined by ADMIXTURE analysis. (A) Principal components (PCs) 1 and 2 of the MVP cohort mimic patterns seen in PCA of the 1000 Genomes Project population. (B) PC3 splits samples aligning with the East Asian reference population (MVP_CHB) from the Native American reference population (MVP_PEL). (C) PC 6 reveals a separate cluster of samples with European Ancestry (MVP_GBR) that splits out distinctively from the main cluster of samples in this subgroup. (D) PC 7 reveals substantial diversity in the MVP_CHB subgroup, whereas PC 8 separates MVP_GBR participants from those with admixture from three or more global reference populations (MVP_OTHER). PC 8 also reveals substantial diversity within participants admixed between European and American ancestries (MVP_GBR_PEL). (E) PC 10 reveals a gradient between MVP_GBR and MVP_PEL samples, with MVP_GBR_PEL samples lying between the two clusters.

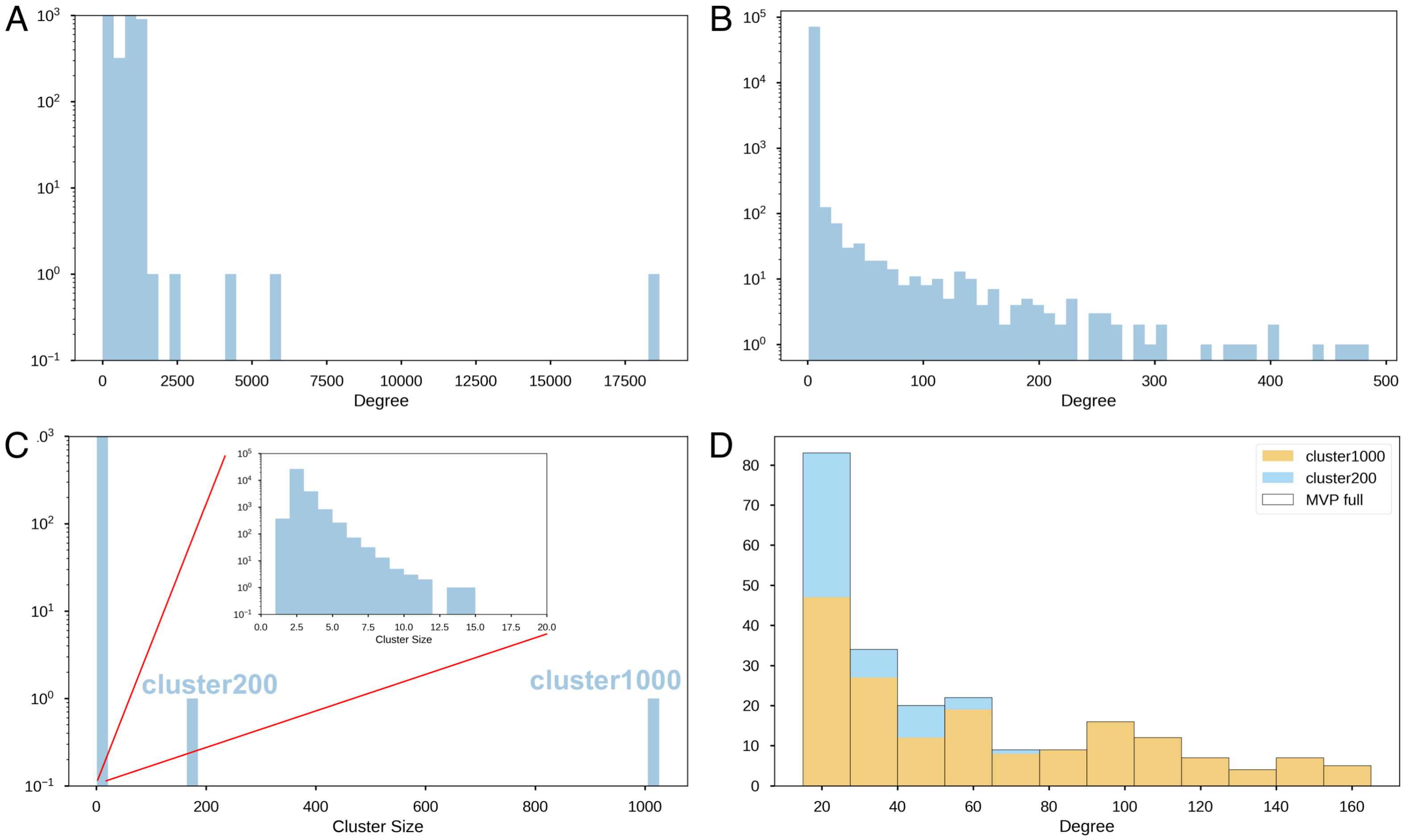


Figure S10. Analysis of relatedness in the MVP cohort using KING. (A-B) Number of related individuals per person in (A) round 1 of KING and (B) round 2. (C) Distribution of cluster size of 3rd degree-related individuals after round 2 of KING analysis. The two large clusters of related individuals are denoted as cluster200 and cluster1000. (D) Cluster membership of all individuals with more than 15 3rd degree relatives.

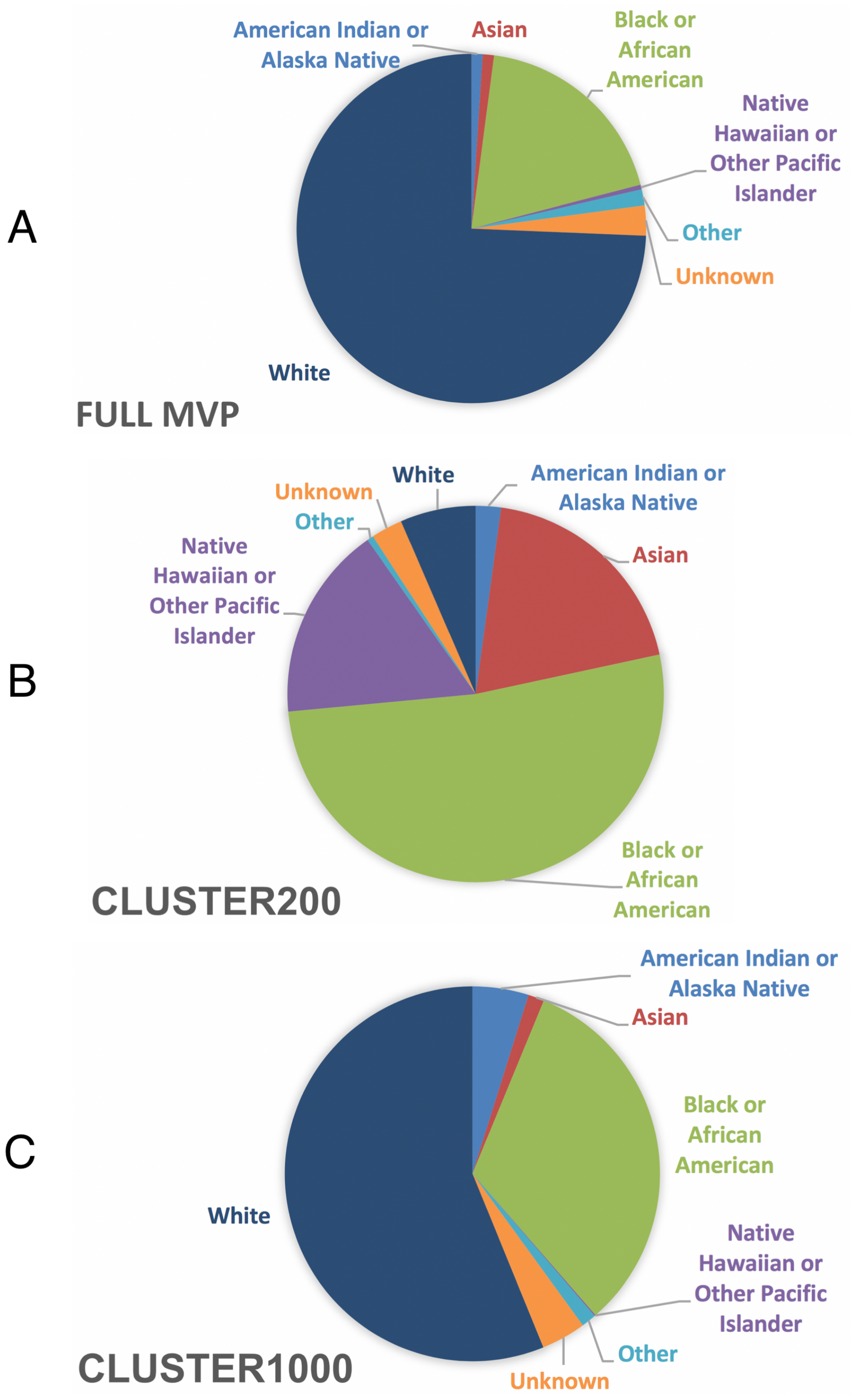


Figure S11. Race breakdown in the full MVP cohort (A) compared to breakdown in Cluster200 (B) and Cluster1000 (C). Pie charts show proportion of self-identified race in each group.

Table S1. Breakdown of MVP 1.0 Array markers by chromosome region.

Region	Number of markers
Mitochondrial chromosome	270
Non-pseudoautosomal Y chromosome	142
Pseudoautosomal X and Y chromosome	1,139
Non-pseudoautosomal X chromosome	18,026
Autosomal	667,105

Table S2. Comparison of PLINK-predicted sample gender to EHR-recorded sample gender.

Recorded gender / Result	Unknown	Male	Female	OK	Mismatch
Unknown	10	1,646	344	0	2,000
Male	292	440,653	753	440,653	1,045
Female	608	420	41,130	41,130	1,028
Total	910	442,719	42,227	481,783	4,073

Table S3. The top ten ACMG recommended genes for reporting incidental findings, and the number of overlapping MVP 1.0 Array markers.

Disease	Gene	Number of markers on the MVP 1.0 array[a]
Breast-ovarian cancer: familial 2 (MIM 612555)	*BRCA2*	662
Breast-ovarian cancer: familial 1 (MIM 604370)	*BRCA1*	541
Marfan's syndrome (MIM 154700)	*FBN1*	176
Dilated cardiomyopathy 1A (MIM 115200)	*MYBPC3*	120
Familial hypertrophic cardiomyopathy 4 (MIM 115197)	*MYBPC3*	120
Familial hypertrophic cardiomyopathy 1 (MIM 192600)	*MYH7*	101
Brugada syndrome 1 (MIM 601144)	*SCN5A*	77
Long QT syndrome 3 (MIM 603830)	*SCN5A*	77
Juvenile polyposis syndrome (MIM 174900)	*SMAD4*	66
Dilated cardiomyopathy 1A (MIM 115200)	*LMNA*	60

[a] We only included markers located in coding or regulatory regions having ClinVar annotation as either pathogenic or likely pathogenic.

Table S4. The top 10 disease categories by number of markers shared between the MVP 1.0 Array and the gene-disease associations reported in DisGeNET.

Disease	Number of markers	Example diseases
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	6,048	Charcot-Marie-Tooth Disease, Hemophilia A, Hereditary Breast and Ovarian Cancer Syndrome
Nutritional and Metabolic Diseases	4,223	Diabetes Mellitus, Hyperlipidemia, Obesity
Nervous System Diseases	3,994	Alzheimer's Disease, Common Migraine, Multiple Sclerosis
Cardiovascular Diseases	3,211	Abdominal Aortic Aneurysm, Coronary Heart Disease, Hypertensive Disease
Neoplasms	3,087	Adenocarcinoma, Leukemia, Osteosarcoma
Skin and Connective Tissue Diseases	2,566	Psoriasis, Rheumatoid Arthritis, Lupus Erythematosus
Immune System Diseases	2,260	Asthma, Hay Fever, Rheumatoid Arthritis
Pathological Conditions, Signs and Symptoms	2,251	Chronic pain, motion sickness, seizures
Digestive System Diseases	1,999	Crohn’s Disease, Inflammatory Bowel Disease, Esophagitis
Endocrine System Diseases	1,985	Hypothyroidism, Adrenal Cortical Adenoma, Dwarfism

Table S5. The Distribution of MVP 1.0 Markers by Minor Allele Frequency, Functional Annotation, and Ancestral Group as Extracted from the ExAC Database.

ExAC MAF		Frameshift	Stop gain/loss	Splice site	Missense	Ins/Del
NFE	MAF ≥ 5%	28	128	683	15,018	26
	5% > MAF ≥ 1%	24	110	246	9,288	8
	1% > MAF > 0%	1,533	10,805	4,260	138,093	94
	MAF = 0%	2,344	8,314	2,554	39,190	77
AFR	MAF ≥ 5%	522	2,670	1,296	33,791	40
	5% > MAF ≥ 1%	37	230	522	19,664	21
	1% > MAF > 0%	718	3,442	2,750	87,686	64
	MAF = 0%	2,619	13,014	3,175	60,445	80
EAS	MAF ≥ 5%	27	118	592	14,058	22
	5% > MAF ≥ 1%	24	79	174	6,453	5
	1% > MAF > 0%	915	4,304	1,662	35,304	47
	MAF = 0%	2,929	14,855	5,315	145,766	131
SAS	MAF ≥ 5%	32	127	650	14,934	23
	5% > MAF ≥ 1%	13	92	232	7,279	9
	1% > MAF > 0%	605	4,394	2,272	69,268	38
	MAF = 0%	3,243	14,744	4,589	110,103	135
AMR	MAF ≥ 5%	29	131	633	14,932	23
	5% > MAF ≥ 1%	22	74	263	8,090	11
	1% > MAF > 0%	401	2,907	2,490	101,308	84
	MAF = 0%	3,443	16,244	4,357	77,243	87
Not in ExAC		4,513	5,091	680	10,985	168

Table S6. MVP 1.0 Array markers categorized by PolyPhen and SIFT score categories based on predicted protein sequence changes.

—		SIFT[a]			
		deleterious	**deleterious Low confidence**	**tolerated**	**tolerated low confidence**
PolyPhen[a]	**probably_damaging**	47,954	3,005	12,447	716
	possibly_damaging	14,107	1,778	14,576	1,160
	benign	13,355	2,043	74,197	7,950
	unknown	752	612	1221	798

[a] PolyPhen and SIFT scores and categories were obtained from the Ensemble Variant Effect Predictor.

Table S7. MVP 1.0 array imputation accuracy for the five 1000 Genomes Project Phase 3 continental populations.

—	1% <MAF <5%		MAF ≥ 5%	
Population	**Number of imputed markers**	**Accuracy (mean.r^2)**	**Number of imputed markers**	**Accuracy (mean.r^2)**
EUR	2,902,508	0.777	7,244,340	0.931
AFR	7,339,754	0.767	10,097,459	0.874
EAS	2,297,176	0.656	6,693,655	0.890
SAS	3,190,047	0.710	7,437,310	0.904
AMR	4,140,903	0.824	7,464,928	0.925

Table S8. Breakdown and number of closely related pairs in the MVP Cohort and UK Biobank.

—	MVP	UK Biobank
Total sample size	459,777	488,410
Total pairs (billions)	105.70	119.27
Trios	28	1,066
Monozygotic twins	49	179
Parent-child	2,886	6,276
Full sibling	5,330	22,666
2nd or 3rd degree	9,055	78,041

Table S9. Thirty of the top ICD diagnostic code names, by prevalence in MVP, available in the VA EHR database. Code names are listed in alphabetical order.

Top ICD diagnostic code names
Acute kidney failure and chronic kidney disease (CKD)
Anemias
Benign neoplasms
Dermatophytosis
Diabetes mellitus
Diseases of esophagus, stomach and duodenum
Disorders of lipoid metabolism
Diseases of male genital organs
Diseases of oral cavity and salivary glands
Diseases of the ear and mastoid process
Diseases of the musculoskeletal system and connective tissue
Diseases of the nervous system
Diseases of the respiratory system
Diseases of the skin and subcutaneous tissue
Disorders of fluid electrolyte and acid-base balance
Disorders of the eye and adnexa
Disorders of thyroid gland
Diverticula of intestine
Hemorrhoids
Hypertension and heart disease
Hypotension
Malignant neoplasm of prostate
Mental, behavioral, and neurodevelopmental disorders
Metabolic disorders
Organic sleep disorders
Other dermatoses
Other diseases of intestines and peritoneum
Other disorders of urethra and urinary tract
Overweight, obesity and other hyperalimentation
Vitamin D deficiency

Supplemental Materials and Methods

Axiom Array Terminology

Here, we first define terms used in standard marker quality control. A marker refers to a genetic variation at a specific genomic location in the DNA of a sample. Both single nucleotide polymorphisms (SNPs) and insertion/deletions (indels) are markers and are genotyped. A set of one or more probe sequences from which array intensities are combined to interrogate a marker is referred to as a probeset. Most Axiom® markers are interrogated with one or two probesets derived from the forward strand sequence, the reverse strand sequence, or both.

Thermo Fisher Scientific Axiom® Genotyping Platform

The MVP 1.0 array is based on the Applied Biosystems Axiom® Biobank core variants. The Axiom® Biobank array modules (**A-F1**) include:

(1) Genome-wide coverage for common European variants *(N= 246,038; Module "High coverage imputation grid")*: Markers were selected using Affymetrix imputation aware marker choice algorithms to provide genome-wide coverage in European populations of common (MAF ≥ 5%) markers using the CEU panel from 1000 Genomes Project.

(2) Loss-of-function (LoF) SNPs and indels *(N= 70,396; Module "High confidence LoF SNPs and Indels")*: Markers with a high-confidence loss of function call, possible splice-site impact, or representing known, disease-associated mutations. Markers are extracted from an exome sequencing initiative of 26,000 people. This category of markers also includes known disease-causing mutations and potential splice variants from Human Gene Mutation Database (HGMD).

(3) Exome SNPs and indels *(N= 264,193; Module "Exome Grid")*: Non-synonymous coding markers in exonic regions of the genome. This content was derived from the Exome Chip Design Consortium, the NHLBI Exome Project, the Genetics of Type 2 Diabetes program (GoT2D), the 1000 Genomes Project, the Cancer Genome Atlas Project, the SardiNIA Medical Sequencing Study, the Autism Exome Sequencing Study, the UK10K project, and others.

(4) Expression quantitative trait locus *(N= 5,884; Module "Expression quantitative loci from GTEx")*: Markers with known associations to RNA expression traits taken from NCBI Genotype-Tissue Expression (GTEx) database.

(5) Pharmacogenomic/ADME *(N= 2,031; Module "Pharmacogenetics")*: Markers associated with phases of drug absorption, distribution, metabolism, and excretion (ADME). This content was derived from the PharmaADME and PharmGKB databases. The content is enriched for markers in Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines and those used in calling "star alleles" variants of genes that affect drug metabolism.

(6) *ApoE (N=2, not shown)* The two SNPs, rs429358 and rs7412, which define the ApoE isoforms known to be associated with risk of Alzheimer's disease and lipoprotein-associated conditions.

MVP-Specific Modules

(1) SNPs and indels based on known associations with diseases or traits of interest (N=21,627; module "*Other diseases associated and exome markers*"). Over 800 potential clinical conditions and 3,400 gene regions are covered by this module with

an emphasis on conditions of high prevalence in US Veterans such as substance abuse[1], amyotrophic lateral sclerosis (ALS [MIM: 105400]) [2], chronic obstructive pulmonary disease (COPD [MIM: 606963])[3], and diabetes[4]. Additionally, we included the 59 genes currently recommended for reporting of incidental findings by the American College of Medical Genetics and Genomics (ACMG)[5,6] in anticipation of future clinical applications.

(2) Human leukocyte antigen (HLA) and killer-cell immunoglobulin-like receptor (KIR) region markers for fine mapping and imputation (N=8,058; module *"Markers within genomic regions of interest")*. Genes in the *HLA* region are critical players in immune function, and some have known roles in histocompatibility[7,8]. We also included additional *KIR* markers that have an important role in infectious disease, autoimmune disorders, cancer, and reproduction by virtue of the region's interaction with the *HLA* region. Furthermore, to optimize coverage and improve imputation quality in non-European ancestry populations, we also included additional *HLA* and *KIR* markers in multiple ethnicities, including African ancestry (N= 1,043; Module "HLA imputation booster for non-Caucasian cohorts").

(3) Psychiatric condition-associated markers (N=26,928, module *"Psychiatric relevant markers"*). 11-20% of US military personnel who served in Operations Iraqi Freedom and Enduring Freedom are estimated to have diagnosed or undiagnosed post-traumatic stress disorder (PTSD)[9]. The lifetime prevalence of PTSD among Vietnam Veterans, who form the largest subset of the MVP cohort, is 15-20%[10]. To study genetic factors that affect the development of mental illness and the functional disability and cognitive impairment that accompany these conditions, we included additional markers extracted from the Psychiatric Genomics Consortium[11] (personal communications with PGC group leaders) that have been linked with common psychiatric disorders such as schizophrenia (MIM: 181500), bipolar disorder (MIM: 125480), autism spectrum disorders (MIM: 209850), attention deficit hyperactivity disorder (MIM: 143465), major depressive disorder (MIM: 608516), obsessive compulsive disorder (MIM: 164230), anorexia (MIM: 606788), and Tourette syndrome (MIM: 137580). Many of these variants are rare for some ethnicities and have minor allele frequency (MAF) < 1%.

(4) Candidate genes associated with rheumatoid arthriti*s* (MIM: 180300) (N=11,144; module *"Rheumatoid Arthritis relevant markers"*). Markers include a) rheumatoid arthritis-associated SNPs outside of the MHC region[12,13,14] (personal communications); b) SNPs with minor allele count (MAC) > 2 found in genes in rheumatoid arthritis risk loci in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) data (European ancestry cohort); and c) SNPs in the 1000 Genomes Project European population that overlap with the rheumatoid arthritis risk loci having MAF > 1% after linkage disequilibrium pruning to remove SNPs with r^2 >0.8.

(5) Genome-wide imputation booster for African ancestry population (N=50,000; module *"Imputation booster"*). The MVP population contains multi-ethnic groups, and African Americans comprise the second largest self-reported ethnic group in the MVP. To optimize the coverage of African genetic ancestry, an additional set of markers were tiled on the MVP array to boost corresponding imputation coverage. Markers were selected using Axiom® imputation aware marker choice algorithms[15] and the YRI panel from the 1000 Genomes Project to provide genome-wide coverage in African populations of common (MAF ≥ 5%) markers.

Overlap with databases of functional and disease-related variants

The MVP 1.0 array was designed to increase coverage of known functional and disease- or trait-associated variants across the genomes of diverse populations. To evaluate this coverage, we aimed to check the coverage of MVP 1.0 markers with markers from various data sources cataloging potential functional variants in multiple populations that were not considered during the design of the chip: the NHGRI-EBI GWAS catalog[16,17], ClinVar [18]; the Exome Aggregation Consortium (ExAC)[19,20]; and DisGeNET, which also contains the genes in the ACMG guideline recommendations for reporting incidental findings and gene-disease associations[5,6]. Markers in the MVP 1.0 chip overlapped with 7,199 variants in the GWAS catalog v1.0[16], 31,067 in ClinVar[18], and 294,092 in ExAC r.0.3.1[20]. As expected by the number of GWAS hits in the GWAS catalog, the top five diseases associated with variants shared between MVP 1.0 and the GWAS catalog are schizophrenia, rheumatoid arthritis, Crohn disease, type 2 diabetes, and systemic lupus erythematosus, and the top five biological traits are height, BMI, HDL, cholesterol, LDL cholesterol, and triglycerides. Furthermore, around half of the variants shared between MVP 1.0 and ClinVar (17,124 of 31,067) are classified by ClinVar as pathogenic or likely pathogenic and have low allele frequency in all ancestral groups.

Additionally, MVP 1.0 contains markers overlapping with pathogenic variants residing in all 59 ACMG genes[5,6]. The top ten ACMG genes and the corresponding number of MVP 1.0 markers are listed in Table S3. DisGeNET also contains one of the largest publicly available collections of genes and variants associated with human diseases[21], and the MVP 1.0 array contains 6,048 DisGeNET markers associated with congenital, hereditary, and neonatal diseases and abnormalities, which represents the largest overlapping disease category. The top ten DisGeNET categories and the corresponding number of MVP 1.0 markers are listed in Table S4.

Variants in the ExAC database predicted to have severe consequences (e.g. missense, frameshift, stop-gain, start-loss variants) were highly enriched in the rare and low frequency MAF ranges (Table S5). Furthermore, in addition to containing markers for variants observed in the 1000 Genomes Project and ExAC database, the Axiom Biobank Genotyping Array backbone has many variants not found in these databases. For example, more than half of the frame shift variants and one-third of the insertion or deletion variants were absent from these databases (Figure S1). We also examined MVP 1.0 nonsynonymous single nucleotide variants (nsSNVs) in detail by implementing functional annotation using the Ensembl Variant Effect Predictor (VEP), release 86[22]. 66,790 coding variants were predicted as deleterious by both PolyPhen[23] and Sift[24] (Table S6), and 16,598 non-coding variants were annotated as regulatory variants.

Finally, we examined the allele frequency distribution of variants on the MVP 1.0 chip to assess its coverage across ancestral population groups. Allele frequencies of variants in the GWAS catalog and ClinVar were assessed using 1000 Genomes, whereas ExAC variant frequencies were used directly. Four MAF categories were defined as: common variants (MAF ≥ 5%), low-frequency variants (MAF 1-5%), rare variants (MAF < 1%), and monomorphic variants (MAF = 0%). In general, markers in the MVP 1.0 chip showed similar distribution of MAFs across all ancestral groups in all subsets (Figure S2). As expected, most markers in the MVP 1.0 chip that overlaps NHGRI-EBI GWAS catalog variants were common (MAF ≥ 5%, Figure S2A**)**, whereas markers that overlap ClinVar variants were mostly uncommon (MAF < 5%, Figure S2B**)**. On the other hand, although the proportion of common variants in ExAC was fairly consistent across ancestral groups, the proportion of less common ones showed greater variability (Figure S2C).

Assessing imputation accuracy of the MVP v1 chip

Imputation was executed using IMPUTE29 and the 1000 Genomes Phase 3 reference panel10. To assess imputation accuracy, we performed ten-fold cross-validation, which corresponds to a sub-sample of the participants in each of the five populations. The 1000 Genomes Phase 3 genotypes of the held-out samples for the markers on the MVP 1.0 array were used to impute the genotypes of the markers in the modified reference panel. The accuracy of imputed variants was calculated as the squared Pearson correlation coefficient (r2) between imputed genotype dosages (0–2) and the genotypes from the 1000 Genomes Phase 3 release (0,1,2). The results were stratified into two non-overlapping minor allele frequency (MAF) bins MAF ≥ 5% and 1% < MAF < 5%, and accuracy was summarized as the mean r2 over the marker count in the MAF bin.

In addition to the aim of increasing coverage of functional variants, the MVP 1.0 array was designed to maximize imputation performance for GWAS studies. We analyzed imputation performance of the array on 1000 Genomes Project data by performing cross-validation with 10% of the samples at a time. We report the accuracy of imputed variants across two MAF bins in terms of the squared Pearson correlation coefficient (r2) between imputed genotype dosages (0-2) and the genotypes from the 1000 Genomes Project Phase III release (0, 1, 2) (discussed in Materials and Methods.) (Table S7). For example, the mean r2 for the EUR group is 0.931 in common variants and 0.777 in low frequency variants. Among the five ancestry groups, imputation of common variants was most accurate for the EUR population, while imputation of less common variants was most accurate for the AMR population. The high accuracy of imputation of less common variants in the AMR population reflects the admixture from European, African, and Native American ancestries, two of which were already included in the 1000 Genomes Project.

Vendor quality control (QC)

Two vendors, referred to as Vendor 1 and Vendor 2, performed initial genotyping of MVP samples for internal QC purposes. The vendors genotyped all MVP samples on the custom MVP 1.0 array. To batch samples in preparation for genotype calling, the vendors grouped approximately 5,000 samples into each batch and used either the Axiom® Analysis Suite (Windows graphical user interface) or Affymetrix® Power Tools (APT, command line software) to execute the *AxiomGT1* genotyping algorithm. Vendor 1 called genotypes with the Axiom® Analysis Suite versions 1.15, 1.16, 2.3, and 2.4. Vendor 2 called genotypes with the APT software and used versions 1.15.1 and 1.16.1. Both the Axiom® Analysis Suite and the APT software call genotypes in each batch independently of other batches.

The vendors followed the Axiom® Best Practices to filter the data for QC purposes[25] and so reprocessed samples with dish QC (DQC) values less than 0.82 or sample QC call rate less than 97%. Whereas Vendor 2 used a threshold of 97% to filter samples in all batches for re-genotyping, Vendor 1 began with a threshold of 97% but switched to a 98% threshold after genotyping several batches to increase genotyping quality. This is the reason why Vendor 1's QC call rate is consistently higher than that of Vendor 2 even after re-batching and re-genotyping on the full cohort (Figure S5 A-C). Only samples that pass internal vendor QC were sent to the VA.

The vendors also investigated plates with the plate QC metric (average QC call rate of passing samples over the plate) lower than 98.5% to identify any systematic row or column failure and removed these plates if found. Since both vendors re-processed failed samples immediately after performing the Best Practices analyses and did not always return the failed samples in the delivered data, we relied on the vendors' reports for the plate QC, sample DQC, and QC call rate values.

Genotype calling on the MVP cohort

All samples that passed vendor QC were then re-batched and recalled internally at the VA under a unified workflow. After comparing vendor and multiple additional batching strategies, we separated MVP samples by vendor and then batched samples in groups of 4,000 to 5,000 samples by each sample's plate scan date. We incorporated samples into each batch at the unit of a plate so that all samples on one plate would be within the same batch. Since experimental metadata captured in *CEL* files (Axiom® files that report the array intensity data) is considered the most reliable, we extracted the plate scan date from the CEL file header by using an APT program called calvin-extract. After batching samples, we had 112 batches with a median of 4,500 samples. Then, we called genotypes in each batch with the APT software, version 1.18, following the Axiom® Best Practices Genotyping workflow[25].

Initial QC

To provide an initial assessment of MVP sample genotyping quality, we analyzed several QC metrics in detail:

1) The DQC metric is an approximate measure of contrast between genotyping channels at non-polymorphic genomic locations[25]. Because the DQC metric is calculated only on non-polymorphic probesets, this metric assesses the signal to background noise ratio for each channel and should be close to 1 for high-quality samples.
2) We also analyzed QC call rates (call rates computed over a set of 20K QC probesets, also referred to as step 1 call rate) and sample call rates over all probesets (also referred to as step 2 call rate) by batch for all samples. The sample call rate is similar except it is calculated on genotype calls over all probesets in a second round of calling after samples that failed the QC call rate threshold were removed.
3) The final metric, sample missingness, is the fraction of genotype calls which are missing per individual after the advanced marker QC procedure over the number of markers that are called in at least one individual in the sample. The sample missingness is an indicator of sample genotype call completeness after the conclusion of all QC procedures.

Standard marker quality control

Standard marker quality control is executed by the Axiom® Best Practices Genotyping workflow. This workflow executes rules to select a *best* probeset for markers that are interrogated with more than one probeset, which is followed by the sorting of probesets into four *recommended* and three *not-recommended* SNP QC classes. Three recommended classes require probesets to produce well-resolved genotype clusters for probesets with three (the *PolyHighResolution* class), two (the *No Minor Homozygous* class), or one (the *Mono High Resolution* class) genotype clusters in the given batch (see the Axiom® Genotyping Solution Data Analysis Guide). The fourth recommended class is "Hemizygous" for Y chromosome and mitochondrial markers. The best practices guideline is to only use genotype calls from a given batch for marker probesets that are both *best* and *recommended*.

Advanced Marker Quality Control Procedure for the MVP 1.0 array

Approach 1: Exclude probesets that fail Advanced QC tests

The goal of the Advanced QC tests is to identify probesets that systematically underperform in most batches or that are interrogating multi-allelic markers. Such probesets were identified in five ways: (1) probesets for multi-allelic markers; (2) probesets with higher than expected variance in allele frequency between batches; (3) probesets that failed to cluster into recommended classes in a large fraction of batches; (4) probesets with large differences in allele frequency when compared to reference datasets; and (5) mitochondrial or Y chromosome probesets that fail visual inspection of their cluster plots.

Approach 1.1: Probesets for multi-allelic markers

Because the current *AxiomGT1* genotyping algorithm only calls the three genotypes for bi-allelic markers, probesets for multi-allelic markers, which have a minimum of six genotypes, are excluded. Only 0.22% of probesets interrogate multi-allelic markers.

Approach 1.2: Probesets with inconsistent allele frequencies across batches

Probesets exhibiting larger than expected variance in allele frequency across genotyping batches could indicate genotype calling issues. We used a statistic called the ratio A allele frequency (RatioAAF) to measure the stability of a probeset's allele frequency across batches[26]. The RatioAAF is defined as the expected allele frequency variance over the observed allele frequency variance:

$$RatioAAF = \frac{p'(1-p')}{\mathrm{var}(P)}$$

where P is a vector consisting of the marker allele frequencies across all recommended batches and p' is the average allele frequency across all recommended batches. A larger RatioAAF indicates a more stable allele frequency. Overall, 3.41% probesets have RatioAAF less than 1,000 when computed over recommended batches. These probesets are often recommended in individual batches but produce inconsistent allele frequencies when measured over multiple batches. Visual inspection of a sampling of cluster plots of low RatioAAF probesets confirms that such probesets produce poor genotype clusters.

Approach 1.3: Probesets that are recommended in few batches

Probesets that are not often classified into one of the four recommended categories (discussed above) over most batches are typically genotyping poorly. Visual inspection of cluster plots for probesets recommended in less than 20 batches out of 112 confirms that such probesets produce poor genotype clusters and suggests that probesets recommended in only a few batches are fundamentally not working. 2.68% of probesets were eliminated by this filter.

Approach 1.4: Probesets with allele frequencies diverging from those in reference datasets

A small fraction of probesets demonstrate large discrepancies in reference allele frequency between the MVP dataset and the 1000 Genomes Project, phase 1 CEU reference population. Such probesets were identified with the following thresholds on the reference allele frequencies: (MVP < 0.02 AND CEU ≥ 0.5) OR (MVP > 0.98 AND CEU ≤ 0.5). Allele frequency thresholds were set more stringently for the MVP samples because of the larger sample size. Only 0.82% of probesets were excluded by these criteria.

Approach 1.5: Probesets on the mitochondrial or Y chromosome that fail visual inspection

We conducted a manual review of cluster plots for probesets on the mitochondrial and Y chromosomes as recommended by the Best Practices for these hemizygous markers. Only a small fraction of probesets (0.01%) were excluded by this filter.

Approach 2: Exclude calls from a mis-clustering event by batch

After removing probesets that were fundamentally not working over all genotyping batches, we then turned to a batch level metric of performance. If a probeset was generally working but not one of the four recommended classes in a particular batch, the calls for that probeset in that batch were set to missing. In this way, we retained working probesets while removing calls in sporadic problematic batches.

Approach 3: Select one probeset per marker over all batches

There are 36,601 markers that are interrogated by more than one probeset on the MVP array. For markers that are interrogated by more than one probeset, the Axiom® Best Practices Workflow designates only one probeset as the best probeset per batch based on twelve performance metrics calculated as a part of the genotype calling procedure. However, the probeset that is designated as the best probeset can be different in different batches. We used five metrics to determine the best probeset per marker over all batches. These five metrics included the number of batches in which the probeset succeeded, the number of batches in which the probeset was marked best, the number of batches in which the probeset was classified as Poly High Resolution, the average probeset call rate over all batches, and the average Fisher Linear Discriminate (FLD) over all batches. We compared each probeset's value for each of these metrics sequentially and selected the first probeset that had a higher value for one of the metrics. If all probesets for a marker were equivalent for all five metrics, we selected the first probeset in alphanumeric order.

Correction of batch variation

Individuals in subsequent tranches (Release A and Release B) were randomly sampled from the same population and similarly processed. Therefore, we hypothesized no common genetic variants were significantly associated with assignment to release. We tested this hypothesis by conducting a preliminary GWAS. We restricted the GWAS to samples with European ancestry (GBR >= 0.9 in ADMIXTURE). We used genotypes called from both tranches combined and assigned samples to pseudo-control and case groups according to their Release tranches. Specifically, we assigned Release A samples to the control group (n=238,229) and Release B samples to the case group (n=73,153) as shown in Figure S4B. We ran the association using only common markers (MAF > .05) and used the first 10 principal components (projected to 1000 Genomes Project principal components) as covariates and a linear-mixed model association in BOLT-LMM, which adjusts for cryptic relatedness.

We found 204 significant ($p \leq 5 \times 10^{-8}$) genome-wide independent signals (Figure S4D). The signals were single variant associations without local LD structure based on manual inspection of locuszoom plots[27], suggesting batch effects. We subsequently ran an association adjusting for batch membership as a categorical variable (n = 112) and found only one genome-wide significant hit (in the HLA region). We also ran an association randomly assigning each individual to case or control and observed no significant associations.

For each variant that was genotyped, we then used Fisher's Exact Test to identify batches in which allele distributions significantly differed from that of all other batches. This tests the null hypothesis that a given batch has the same genotype frequencies as all other batches combined.

We compared distribution of allele frequencies across batch-SNP combinations (~60 million) with significant deviance defined as $p < 6.7 \times 10^{-10}$ (i.e., 0.05/batch-SNP combinations). 173 of the 204 genome-wide significant SNPs from the Bolt-LMM GWAS failed the Fisher's Exact Test. An example of a SNP-batch combination that was detected by the Fisher's Exact Test is shown in Figure S4A. We then set all failing SNP-batch combinations to 'no-call' (approximately ~50,000 of ~60,000,000 batch-SNP combinations) and re-ran the Bolt-LMM GWAS as above, this time finding 73 significant independent signals.

We then investigated the remaining 73 significant hits by examination of genotype cluster plots. We identified certain patterns for which Fisher's Test would fail and came up with a series of additional batch correction rules. For probes that fail Fisher's Test in at least one batch:

1) If the RatioAAF is less than 1,000, then exclude all batches for the probeset.
2) If the batches with passing probesets all have fewer clusters than the batches with failing probesets, then switch the pass/fail assignment.
3) If the batch has a call rate less than 98% and FLD less than 5, then fail the batch.
4) If a batch is classified as No Minor Homozygous and has an HWE p-value < $1x10^{-4}$, then fail it.

An overview of these steps and number of SNPs identified is summarized in Figure S4C**.**

After applying these additional batch correction rules, we re-ran the Bolt-LMM GWAS and found an overall reduction in the number of significant associations from 204 to 47 (Figure S4 D-E). These remaining associations were unlikely to be meaningful signals based on manual visual inspection of locuszoom plots and were set to no-call.

Relatedness analysis

Even after excluding 35 individuals with more than 200 relatives, we observed two clusters (denoted as cluster200 and cluster1000) with more than 100 family members in our KING analysis (Figure S10 C-D). Interestingly, all individuals with greater than 15 3rd degree relatives are in one of the two clusters. Members of each cluster are tightly interconnected with each other rather than there being a few individuals promiscuously interacting with other members of the cluster. It is unlikely that these two clusters are real families, and further investigations are warranted. We also observe an underrepresentation of individuals self-identifying as "white" in these two clusters (Figure S11), suggesting the low weight SNP strategy may not have fully removed the overestimation of relatedness in admixed individuals. We flagged all individuals in cluster200 and cluster1000 and redefined a new set of unrelated individuals using the largest_independent_vertext_sets() function to identify the largest set of unrelated individuals to use in our analyses.

Web Resources

Axiom Genotyping Solution Data Analysis Guide, https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf

Locuszoom, http://locuszoom.org/

References

1. Teeters, J.B., Lancaster, C.L., Brown, D.G., and Back, S.E. (2017). Substance use disorders in military veterans: prevalence and treatment challenges.

2. Tai, H., Cui, L., Shen, D., Li, D., Cui, B., and Fang, J. (2017). Military service and the risk of amyotrophic lateral sclerosis: A meta-analysis. Journal of Clinical Neuroscience *45*, 337–342.

3. Pugh, M.J., Jaramillo, C.A., Leung, K., Faverio, P., Fleming, N., Mortensen, E., Amuan, M.E., Wang, C.-P., Eapen, B., Restrepo, M., et al. (2016). Increasing Prevalence of Chronic Lung Disease in Veterans of the Wars in Iraq and Afghanistan. Mil Med *181*, 476–481.

4. Liu, Y. (2017). Prevalence of and Trends in Diabetes Among Veterans, United States, 2005–2014. Prev. Chronic Dis. *14*,.

5. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O’Daniel, J.M., Ormond, K.E., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med *15*, 565–574.

6. Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genetics in Medicine *19*, 249–255.

7. Parham, P., and Ohta, T. (1996). Population Biology of Antigen Presentation by MHC Class I Molecules. Science *272*, 67–74.

8. Manser, A.R., Weinhold, S., and Uhrberg, M. (2015). Human KIR repertoires: shaped by genetic diversity and evolution. Immunological Reviews *267*, 178–196.

9. Affairs UDoV (2016). How Common is PTSD?

10. Magruder, K., Serpi, T., Kimerling, R., Kilbourne, A.M., Collins, J.F., Cypel, Y., Frayne, S.M., Furey, J., Huang, G.D., Gleason, T., et al. (2015). Prevalence of Posttraumatic Stress Disorder in Vietnam-Era Women Veterans: The Health of Vietnam-Era Women’s Study (HealthVIEWS). JAMA Psychiatry *72*, 1127–1134.

11. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet *43*, 977–983.

12. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.

13. Roberson, E.D.O., and Bowcock, A.M. (2010). Psoriasis genetics: breaking the barrier. Trends in Genetics *26*, 415–423.

14. Sode, J., Vogel, U., Bank, S., Andersen, P.S., Hetland, M.L., Locht, H., Heegaard, N.H.H., and Andersen, V. (2015). Genetic Variations in Pattern Recognition Receptor Loci Are Associated with Anti-TNF Response in Patients with Rheumatoid Arthritis. PLOS ONE *10*, e0139781.

15. Hoffmann, T.J., Zhan, Y., Kvale, M.N., Hesselson, S.E., Gollub, J., Iribarren, C., Lu, Y., Mei, G., Purdy, M.M., Quesenberry, C., et al. (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics *98*, 422–430.

16. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res *45*, D896–D901.

17. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research *42*, 1001–1006.

18. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res *44*, D862–D868.

19. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

20. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res *45*, D840–D845.

21. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res *45*, D833–D839.

22. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology *17*, 122.

23. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res *30*, 3894–3900.

24. Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812–3814.

25. Affimetrix (2016). Axiom genotyping solution data analysis guide.

26. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A., et al. (2015). Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics *200*, 1051–1060.

27. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., Willer, C.J., and Frishman, D. (2011). LocusZoom: Regional visualization of genome-wide association scan results. In Bioinformatics, p.

Million Veteran Program (MVP) Consortium members and their Affiliations

MVP Executive Committee: Michael Gaziano, MD, MPH[1], Rachel Ramoni, DMD, ScD[2], Jim Breeling, MD[2], Kyong-Mi Chang, MD[3], Grant Huang, PhD[2], Sumitra Muralidhar, PhD[2], Christopher J. O'Donnell, MD[1], MPH, Philip S. Tsao, PhD[4]; **MVP Program Office**: Sumitra Muralidhar, PhD[2], Jennifer Moser, PhD[2]; **MVP Recruitment/Enrollment**: Stacey B. Whitbourne, PhD[1]; **MVP Coordinating Centers:** Clinical Epidemiology Research Center (CERC) - John Concato, MD, MPH[5]; Cooperative Studies Program Clinical Research Pharmacy Coordinating Center - Stuart Warren, JD, Pharm D[6], Dean P. Argyres, MS[6]; MVP Information Center - Brady Stephens, MS[7]; **Core Biorepository** - Mary T. Brophy MD, MPH[1], Donald E. Humphries, PhD[1]; **Data Operations/Analytics** - Xuan-Mai T. Nguyen, PhD[1]; **MVP Science** Christopher J. O'Donnell, MD, MPH[1]; **Genomics** - Saiju Pyarajan PhD[1], Philip S. Tsao, PhD[4]; **Phenomics** - Kelly Cho, MPH, PhD[1]; **Data and Computational Sciences** - Saiju Pyarajan, PhD[1]; **MVP Local Site Investigators**: Peter Wilson, MD[8], Rachel McArdle, PhD[9], Louis Dellitalia, MD[10], John Harley, MD[11], Jeffrey Whittle, MD[12], Jean Beckham, PhD[13], John Wells, PhD[14], Salvador Gutierrez, MD[15], Gretchen Gibson, DDS[16], Laurence Kaminsky, PhD[17], Gerardo Villareal, MD[18], Scott Kinlay, PhD[1], Junzhe Xu, MD[19], Mark Hamner, MD[20],

Kathlyn Sue Haddock, PhD[21], Sujata Bhushan, MD[22], Pran Iruvanti, PhD[23], Michael Godschalk, MD[24], Zuhair Ballas, MD[25], Malcolm Buford, MD[26], Stephen Mastorides, MD[27], Jon Klein, MD[28], Nora Ratcliffe, MD[29], Hermes Florez, MD[30], Alan Swann, MD[31], Maureen Murdoch, MD[32], Peruvemba Sriram, MD[33], Shing Shing Yeh, MD[34], Ronald Washburn, MD[35], Darshana Jhala, MD[36], Samuel Aguayo, MD[37], David Cohen, MD[38], Satish Sharma, MD[39], John Callaghan, MD[40], Kris Ann Oursler, MD[41], Mary Whooley, MD[42], Sunil Ahuja, MD[43], Amparo Gutierrez, MD[44], Ronald Schifman, MD[45], Jennifer Greco, MD[46], Michael Rauchman, MD[47], Richard Servatius, PhD[48], Mary Oehlert, PhD[49], Agnes Wallbom, MD[50], Ronald Fernando, MD[51], Timothy Morgan, MD[52], Todd Stapley, DO[53], Scott Sherman, MD[54], Gwenevere Anderson, RN[55], Philip Tsao, PhD[4], Elif Sonel, MD[56], Edward Boyko, MD[57], Laurence Meyer, MD[58], Samir Gupta, MD[59], Joseph Fayad, MD[60], Adriana Hung, MD[61], Jack Lichy, MD, PhD[62], Robin Hurley, MD[63], Brooks Robey, MD[64], Robert Striker, MD[65].

1. VA Boston Healthcare System, Boston MA
2. Office of Research and Development, Veterans Affairs Central Office, Washington, DC
3. Philadelphia Veterans Affairs Medical Center, Philadelphia, PA
4. VA Palo Alto Health Care System, Palo Alto, CA
5. VA Connecticut HealthCare System, West Haven, CT
6. Albuquerque VA Medical Center, Albuquerque, NM
7. Canandaigua VA Medical Center, Canandaigua, NY
8. Atlanta VA Medical Center, Atlanta, GA
9. Bay Pines VA Healthcare System, Bay Pines, FL
10. Birmingham VA Medical Center, Birmingham, AL
11. Cincinnati VA Medical Center, Cincinnati, OH
12. Clement J. Zablocki VA Medical Center, Milwaukee, WI
13. Durham VA Medical Center, Durham, NC
14. Edith Nourse Rogers Memorial Veterans Hospital, Bedford, MA
15. Edward Hines, Jr. VA Medical Center, Hines, IL
16. Fayetteville VA Medical Center, Fayetteville, AR
17. VA Health Care Upstate New York, Albany, NY
18. New Mexico VA Health Care System, Las Vegas, NM
19. VA Western New York Healthcare System, Buffalo, NY
20. Ralph H. Johnson VA Medical Center, Charleston, SC
21. Wm. Jennings Bryan Dorn VA Medical Center, Columbia, SC
22. VA North Texas Health Care System, Dallas, TX
23. Hampton VA Medical Center, Hampton, VA

24. Hunter Holmes McGuire VA Medical Center, Richmond, VA
25. Iowa City VA Health Care System, Iowa City, IA
26. Jack C. Montgomery VA Medical Center, Muskogee, OK
27. James A. Haley Veterans' Hospital, Tampa, FL
28. Louisville VA Medical Center, Louisville, KY
29. Manchester VA Medical Center, Manchester, NH
30. Miami VA Health Care System, Miami, FL
31. Michael E. DeBakey VA Medical Center, Houston, TX
32. Minneapolis VA Health Care System, Minneapolis, MN
33. N. FL/S. GA Veterans Health System, Gainesville, FL
34. Northport VA Medical Center, Northport, NY
35. Overton Brooks VA Medical Center, Shreveport, LA
36. Philadelphia VA Medical Center, Philadelphia, PA
37. Phoenix VA Health Care System, Phoenix, AZ
38. Portland VA Medical Center, Portland, OR
39. Providence VA Medical Center, Providence, RI
40. Richard Roudebush VA Medical Center, Indianapolis, IN
41. Salem VA Medical Center, Salem, VA
42. San Francisco VA Health Care System, San Francisco, CA
43. South Texas Veterans Health Care System, San Antonio, TX
44. Southeast Louisiana Veterans Health Care System, New Orleans, LA
45. Southern Arizona VA Health Care System, Tucson, AZ
46. Sioux Falls VA Health Care System, Sioux Falls, SD
47. St. Louis VA Health Care System, St. Louis, MO
48. Syracuse VA Medical Center, Syracuse, NY
49. VA Eastern Kansas Health Care System, Topeka, KS
50. VA Greater Los Angeles Health Care System, Los Angeles, CA
51. VA Loma Linda Healthcare System, Loma Linda, CA
52. VA Long Beach Healthcare System, Long Beach, CA
53. VA Maine Healthcare System, Augusta, ME
54. VA New York Harbor Healthcare System, New York, NY
55. VA Pacific Islands Health Care System, Honolulu, HI
56. VA Pittsburgh Health Care System, Pittsburgh, PA
57. VA Puget Sound Health Care System, Seattle, WA
58. VA Salt Lake City Health Care System, Salt Lake City, UT
59. VA San Diego Healthcare System, San Diego, CA
60. VA Southern Nevada Healthcare System, Las Vegas, NV
61. VA Tennessee Valley Healthcare System, Murfreesboro, TN
62. Washington, DC VA Medical Center, Washington, DC
63. W.G. (Bill) Hefner VA Medical Center, Salisbury, NC
64. White River Junction VA Medical Center, Hartford, Vermont
65. William S. Middleton Memorial Veterans Hospital, Shorewood Hills, WI