All of Us Research Program

Genomic Research Data Quality Report

All of Us Curated Data Repository (CDR) Structural Variant Supplemental Dataset release C2022Q4R9_offcycle

Overview

This document details the *All of Us* Data and Research Center (DRC) quality control (QC) steps for the generation and release of a structural variant (SV) callset that includes **1,506,805 high-quality SVs across 97,940 participants** with short-read whole genome sequencing (srWGS). We have applied these QC steps in the research pipeline before release of the genomic data for research use. We only describe QC processes that are performed analytically herein (i.e., after the sample has been processed, genotyped, and sequenced).

The descriptions and results in this document are limited to the Curated Data Repository (CDR) version 7 (v7) SV supplemental dataset made available in the *All of Us* Researcher Workbench on June 17, 2024. The CDRv7 SV supplemental dataset contains srWGS SV calls from 97,940 participants, all of whom have srWGS single nucleotide polymorphism and small insertion and deletion variant calls (SNPs and Indels) included in the CDRv7 dataset.

Previously in the CDRv7 C2022Q4R9 data release in 2023, we released srWGS SVs for 11,390 samples. The QC descriptions and results for the complete CDRv7 dataset including all other genomic data types are available on the User Support Hub [1].

Audience: This document is intended for researchers using, or considering the use of, the genomic data in the Researcher Workbench. This document assumes knowledge of sequencing, genotype arrays, common genomic data QC approaches, and the variant file formats released in *All of Us*.

Notes:

- Details of the processing (e.g., algorithms) are out of scope for this document.
- The locations of raw data are in the Controlled CDR Directory, published on the User Support Hub [1]. Auxiliary data sample lists are also published on the User Support Hub.
- The genomic data mentioned in this document requires Controlled Tier access to view. Register for access.

Executive Summary

On June 17, 2024, the *All of Us* Research Program released the structural variant (SV) genomic data representing 97,940 participants in the Researcher Workbench for use by registered researchers with Controlled Tier access. There are over 1.5 million SV sites in the dataset. With this release of SV data, we increased the percent of srWGS samples that have SV data to 40%. The data complements the existing genomic data available on the Workbench, where researchers can analyze 312,945 array samples, 245,394 srWGS SNP and Indel samples, and 1,027 long-read whole genome sequencing (lrWGS) samples. The genomic data is paired with other health and survey data available on the Workbench. Quality control processes, performed both independently and across samples, indicate that these data are ready for general analysis.

Introduction

All of Us is collecting biospecimens and generating genomic data for all participants who have consented among its target of 1,000,000 participants. This document describes the off-cycle data release of 97,940 samples with srWGS SVs made available in the Workbench on June 17, 2024. Genomic data can be joined with other data types for analysis on the Workbench. In this document, we describe the QC processes applied to the SV data.

The srWGS SV calling was performed on 97,940 srWGS samples, which are a subset of the 245,394 CDRv7 srWGS samples with SNP and Indel variant calls. Prior to SV calling, all samples followed the Consistency across Genome Centers and Single Sample QC processes in the srWGS QC pipeline. These steps are documented in the CDRv7 *All of Us* Genomic Research Data Quality Report available on the User Support Hub [1].

We used GATK-SV to call SVs, which has been previously described [2]. Further technical information can be found in Appendix A. GATK-SV discovers SVs of the following types: deletion (DEL) and duplication (DUP), which can together be described as copy number variants (CNV); insertion (INS); inversion (INV); translocation (CTX); complex event (CPX); unresolved breakend (BND); and multiallelic CNV (we refer to them as MCNV in this document but their SV type in the VCF is CNV). See [3] for additional information on SV types and their evidence signatures.

We outline the sample selection process, the single sample QC, and the joint callset QC. Single sample QC are the QC processes for each sample independently to catch major errors. If a sample fails these tests, it is excluded from the release and not reported in this document. Joint callset QC are the processes executed on the joint callset, which use information across samples to flag samples and variants.

We have also performed data validation experiments and benchmarking and the results are shown in other, upcoming documentation (see the User Support Hub [1]).

Sample Selection for srWGS SVs

We initially selected 100,321 samples for SV calling. The samples were selected from participants who had srWGS data in the Controlled Tier CDRv6 (C2022Q2R2) dataset or participants who have been selected for previous or future long-read sequencing. Of these initially selected samples, we excluded 2,381 (2.37%) from the final callset (Table 1). Of these 2,381, some were removed between the CDRv6 and CDRv7 (e.g., participant withdrew) (Table 1). Additionally, we use stricter QC criteria for srWGS SV calling than for srWGS SNP and Indel calling and as a result, some samples were dropped during the QC steps. The final CDRv7 off-cycle srWGS SV callset contains 97,940 samples.

The 100,321 selected samples contain 11,439 samples selected for the CDRv7 srWGS SV callset that passed single-sample SV QC. For a full description of the sample selection criteria, see the CDRv7 QC report [1]. The remaining 88,882 samples that were not in the CDRv7 srWGS SV callset are the samples from the CDRv6 srWGS release that were not previously selected for SV calling.

Table 1 -- Number of samples that were excluded from SV calling

srWGS SV sample exclusion steps	Number of samples filtered from initial count (N=100,321)	Notes
Single sample QC	2066	See Table 2 and Table 3. 2,005 samples were removed by basic filters and 61 were removed during ploidy estimation.
Joint SV callset refinement and QC	11	Outlier samples were removed following ClusterBatch (see Appendix A).
Other	304	These are CDRv6 srWGS samples that were not included in CDRv7 for reasons unrelated to SV calling (e.g., participant withdrew between releases)

Single Sample QC for srWGS SVs

We performed single sample QC, as described in Table 2 and Table 3, on all 88,882 newly selected samples for the CDRv7 off-cycle srWGS SV callset. We removed a total of 2,066 samples during srWGS SV single sample QC, which left 86,816 new samples and 98,255 total samples remaining in the callset for downstream processing.

Basic filters

Method

As seen in Table 2:

1. We performed a cross-individual contamination check following the same protocol that we used for the srWGS SNP and Indel analysis but with a more stringent passing criteria of 1%. Previously in the CDRv7 srWGS SV release, this filter was 0.5%. We increased this filter to avoid removing too many samples.

2. We checked the mean insert size of each srWGS sample using the Picard tool CollectInsertSizeMetrics within GATK's CollectMultipleMetrics and removed samples that were outside of the range 320-700.

3. We checked the whole genome dosage (WGD) [2] to identify samples that were outliers for dosage bias, i.e. whose coverage across the genome was highly variable. Non-uniformity of coverage negatively impacts copy number variant (CNV) calling. Samples with a WGD score more than six times the median absolute deviation (MAD) outside the median were removed, where $MAD = median(|WGD_i - median(WGD)|)$.

4. We counted the number of non-diploid 1 megabase (Mb) bins in each sample. If the number of bins exceeded our threshold (500), we believed that the coverage would be too variable for accurate CNV calling,

5. We filtered samples with outlier SV counts from the SV calling tools Manta [4], Wham [5], and MELT [6] relative to the other samples in the cohort. Higher than typical SV counts may signify technical artifacts. SV counts were stratified by SV caller, chromosome, and SV type. Samples that were outliers in 30 or more categories were removed from the callset.

We removed all samples that failed any of these filters, in total 2,005 (Table 2). Note that some samples failed multiple filters.

Results

The results for all six basic single-sample filtering steps are summarized in Table 2.

Table 2 -- srWGS SV single sample QC: Basic filters

QC process	Passing criteria	Error modes addressed	Number of samples removed
Cross-individual contamination	≤ 0.01 (≤ 1%)	Sample contamination from another individual	296
Mean insert	Mean insert size in range	Insert size outliers, which could skew	30

size	[320, 700]	distributions of discordant pairs	
WGD	WGD within 6*MAD of the median, approx. [-0.162, 0.136]	Samples with high variability in coverage across the genome, which could lead to unreliable CNV calling from depth evidence	1,337
Number of non-diploid 1Mb bins	≤ 500	Samples with high variability in coverage across the genome, which could lead to unreliable CNV calling from depth evidence	1,508
SV count outliers	Sample is an outlier < 30 times across bins of SV caller, SV type, and chromosome	Samples with unusually high raw SV counts after initial SV discovery, which could introduce large numbers of false positive calls to the callset	89

Ploidy estimation

Method

We estimated ploidy per chromosome across all 88,882 new samples by binning read counts in 1Mb intervals and normalizing by half the genome-wide median. We only performed filtering based on ploidy on the 86,877 samples that passed the basic filters (Table 2).

We observed likely mosaic loss of chrX and chrY in some samples, as described in previous studies [7] [8]. These samples had an estimated copy ratio of 0.1-0.8 on chrY and 1.2-1.8 on chrX and are likely to have mosaic loss of chrX or chrY, but the low copy number could also be due to large deletions on these chromosomes. For the sex-specific steps of the GATK-SV pipeline, these samples were classified as follows:

- Grouped with males if chrX rounded ploidy = 1 and chrY ploidy > 0.1
- Grouped with females if chrX rounded ploidy = 2
- Classified as “other” and no calls made on allosomes if chrX rounded ploidy = 1 and chrY ploidy = 0.

For each sample, the computed sex was compared to the self-reported sex at birth to evaluate concordance as a check for potential sample swaps. Samples with mosaic loss of chrX or chrY were grouped as described above.

Samples passed this check if the computed sex matched the self-reported sex assigned at birth, if there was a predicted germline aneuploidy of an allosome, or if the participant did not respond or selected an answer other than “male” or “female” for the sex assigned at birth question in the Basics survey. Because we were looking for sample swaps, we chose these cutoffs in order to prevent unnecessarily removing samples. Participants can report “Male”, “Female”, “Intersex”, “I prefer not to answer”, “none of these fully describe me”, or skip the sex_at_birth question. Please refer to Appendix F in the CDRv7 QC report for additional details [1].

Results

We filtered 61 samples because they had an estimated copy ratio greater than 2.3 or less than 1.8 on at least one autosomal chromosome (Table 3). Plots of binned coverage across these chromosomes confirmed that these samples may represent mosaic autosomal aneuploidies. In addition, we discovered 849 samples with a likely mosaic loss of chrX or chrY among the 86,877 new samples that passed basic filters, though in-depth analyses and validation of somatic and mosaic variation was outside of the scope of activities for this callset. All samples passed the comparison check between computed sex and self-reported sex at birth, indicating no sample swaps based on the computed sex.

Among the 86,877 new samples that passed basic filters and the samples previously examined during CDRv7 srWGS SV processing, we identified 106 samples with predicted germline sex chromosome aneuploidies (i.e. computed sex ploidy other than XX, XY, or mosaic). These samples were classified as “other” for the sex-specific steps of the GATK-SV pipeline and SV calls were not made on chrX or chrY for these samples.

Lists of the samples identified to have likely mosaic autosomal aneuploidies, likely mosaic loss of chrX or chrY, and germline sex chromosome aneuploidies are available; for additional details, read the Controlled CDR Directory on the User Support Hub [1]. These lists include samples identified from both the 86,877 new samples that passed basic filters and the samples previously examined during CDRv7 srWGS SV processing.

Table 3 -- srWGS SV single sample QC: Ploidy estimation filters

QC process	Passing criteria	Error modes addressed	Number of samples removed	Notes
Estimated copy number per autosome (Ploidy estimation)	1.8 ≤ copy ratio ≤ 2.3	Samples with mosaic autosomal aneuploidies, which could skew distributions of SV evidence classes	61	Calculated after applying all above filters. Method can be found in [2]
Sex concordance	Computed sex is concordant with self-reported sex at birth. OR Computed sex is neither male nor female. OR Self-reported sex at birth reported as “Other”* or was not reported	Sample swaps	0	All samples passed this check *Other refers to a participant self-reporting “Intersex”, “I prefer not to answer”, or “none of these fully describe me”

Batching

We divided the 88,882 new samples into 168 batches with an average of 517 samples in each batch for the batched analysis steps of the GATK-SV pipeline, depicted in Figure 1. Batching controls for technical variability between samples and parallelizes computation. The batching procedure was as follows:

1. Split by chrX copy ratio (<1.5 and ≥1.5)
2. Split each partition of samples from the previous step four ways by mean insert size
3. Split each partition three ways by WGD score
4. Split each partition two ways by median coverage
5. Merge corresponding partitions by chrX ploidy to balance chrX ploidy within batches

The batching scheme was based on previously described methods [2], except for the addition of the mean insert size as a batching parameter. We added this to address an observed multimodal distribution of mean insert size, described previously in the CDRv7 QC report [1].

Joint Callset Refinement and QC for srWGS SVs

The steps to generate the GATK-SV joint callset are described in Figure 1 and Appendix A. Appendix A also includes a summary of GATK-SV pipeline improvements that have been implemented since the CDRv7 srWGS SV release. Below, we describe refinement and filtering steps introduced in the *All of Us* srWGS SV dataset that were not published previously or are modifications to canonical GATK-SV pipelines (blue steps in Figure 1). These steps include both hard and soft filters at the sample, site, and genotype level (Table 4).

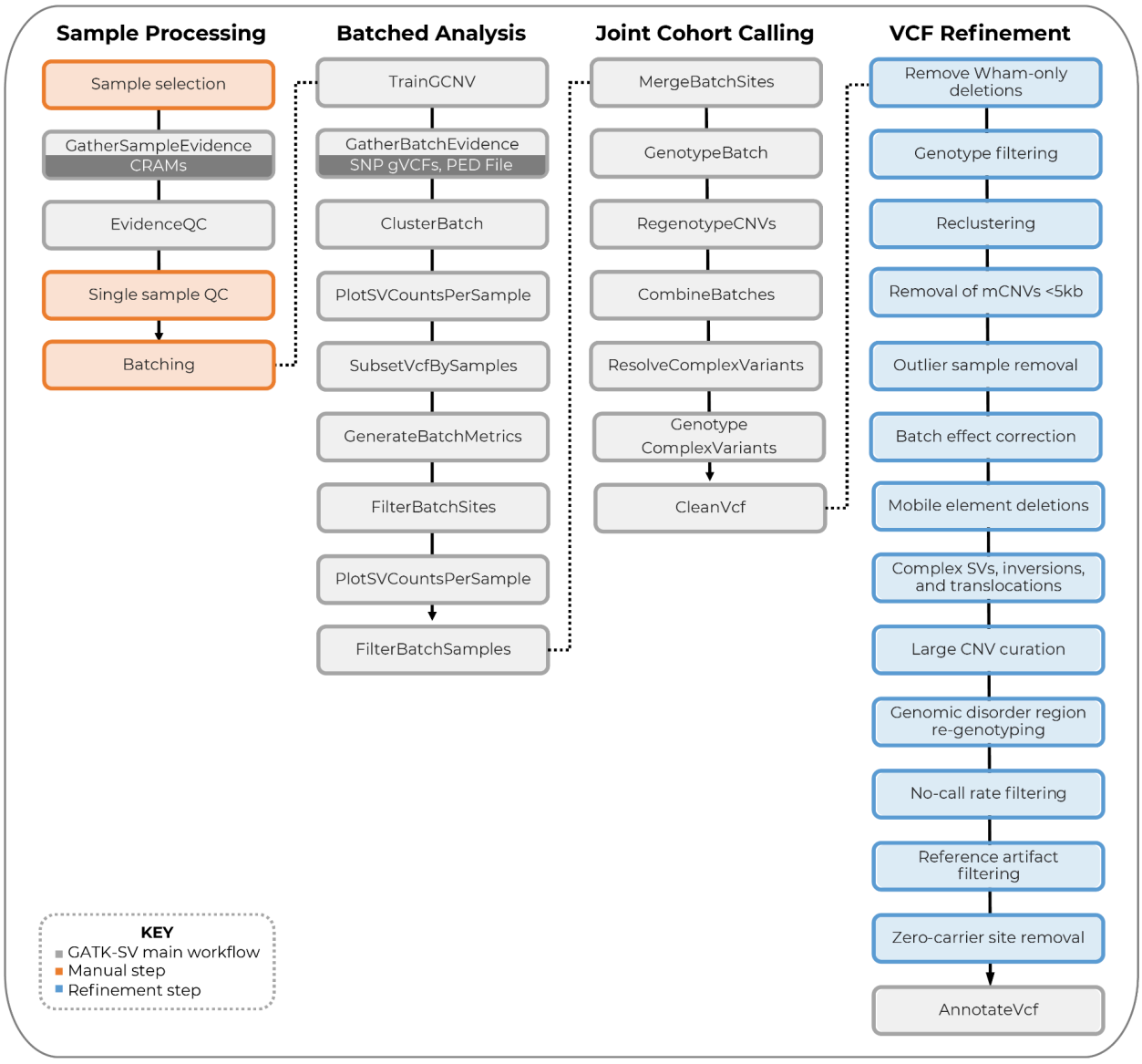


Figure 1 -- GATK-SV Pipeline Schematic. GATK-SV automated workflows are shown in gray and the names correspond to the name of the Workflow Definition Language (WDL) file. Manual steps performed in notebooks are shown in orange. Steps in blue are custom VCF refinement and QC steps for the *All of Us* SV callset.

Table 4 -- GATK-SV VCF refinement and filtering steps unique to *All of Us*

QC process	Sample, variant, or genotype QC	Filter tag	Error modes addressed	Notes
Remove	Variant		False positive	Unique Wham deletions were removed

Wham-only deletions			deletions	from the callset.
Genotype filtering	Genotype		False positive genotypes for INS, INV, DEL, and DUP	We used a machine learning model to filter bi-allelic genotypes with a scaled logit (SL) score. Filtered genotypes are set to no-call (`./.`)
Reclustering			Redundant sites in repetitive regions	No filtering at this step
Removal of mCNVs <5kb	Variant		False positive MCNVs	Multiallelic CNVs less than 5 kilobases (kb) in length were removed from the callset.
Outlier sample removal	Sample		Noisy samples	No samples were removed from the callset at this stage.
Batch effect correction	Variant	VARIABLE_ACROSS_BATCHES	Technical artifacts from batch effects	
Mobile element deletions	Variant		Rescue mobile element deletions previously marked UNRESOLVED	Mobile element deletions detected in this step were revised to PASS, the SVTYPE field was set to DEL, and the ALT field was set to describe the type of mobile element deletion
Complex SVs, inversions, and translocations curation	Variant and genotype		False positive CTX, INV, and CPX	Filtered genotypes are set to no call (`./.`). Revisions are found in the INFO field MANUAL_REVIEW_TYPE
Large CNV curation	Variant and genotype		Large CNVs that are false positives, have inaccurate breakpoints, or are multiallelic	Revisions are found in the INFO field MANUAL_REVIEW_TYPE
Genomic disorder region re-genotyping	Variant and genotype		False positive and false negative calls overlapping genomic disorder regions	Genomic disorder regions were re-genotyped to improve sensitivity and specificity. Manual revisions are found in the INFO field MANUAL_REVIEW_TYPE
No-call rate (NCR) filtering	Variant	HIGH_NCR	False positives, technical artifacts, sites that are difficult to genotype	
Reference artifact filtering	Variant	LIKELY_REFERENCE_ARTIFACT	Sites that are homozygous in >99% of samples, indicating a likely reference artifact	

Zero-carrier site removal	Variant		Sites are removed if no carriers remain after filtering	Variant sites are removed if no carriers remain after filtering.

Remove Wham-only deletions

As described in the CDRv7 QC report, we observed very high false-positive rates for deletions that were uniquely called by the Wham algorithm [5], one of the SV calling algorithms used by GATK-SV. These variants were removed from the callset.

Genotype filtering (SL filter)

We filtered genotypes of bi-allelic SVs using a machine learning model trained on lrWGS data. This model recomputes genotype qualities (GQs), enabling us to reduce false positive INS, INV, DEL, and DUP variant calls while minimizing loss of sensitivity.

Method

lrWGS training data

We selected true positive and false positive training sites for the machine learning model based on comparisons against long read data. Long read SV calls are ideal for confirming SV events with accurate breakpoint resolution but are not sensitive to large CNVs (>5kb) that must be detected by read depth signatures. Therefore, the training labels based on lrWGS were applied only to DEL and DUP variants less than 5kb in length, as well as INS and INV variants.

A subset of 893 samples with matched lrWGS data were selected for model training, and an additional 97 were held out as a test set to validate the model. For each sample, non-reference genotypes for eligible variants (SV type DEL, DUP, INS, or INV, restricting to below 5 kb in length for CNVs) were assessed against lrWGS. Calls were first evaluated using the lrWGS validation tool VaPoR [9]. In addition, the lrWGS variant calling was performed using the tools PAV [10], PBSV [11], and sniffles2 [12]. The GATK tool SVConcordance was then used to compute overlap between SV calls from srWGS and lrWGS [13].

Variants were labeled as positive training examples if:

- The variant had at least two reads supporting the alternate allele according to VaPoR. We counted a read as supporting the alternate allele if the VaPoR_Rec score (a confidence score for each long read; positive values indicate support for the alternate structure described by the SV call) was greater than zero AND
- The variant had at least one long read SV call with at least 10% reciprocal overlap (ratio of total overlap to the size of the larger call) and 50% size similarity (ratio of the smaller to larger call size).

Variants were labeled as negative training examples if:

- The variant had at least 5 reads that VaPoR was able to evaluate in the sample and no reads had a positive VaPoR_Rec score AND
- The variant was not within 5 kb of a breakpoint of a lrWGS SV call with a matching SV type.

Variants that did not meet either the positive or negative criteria were dropped from the training set (Figure 2A).

Filtering model

We trained a model to re-calculate SV genotype qualities based on the training data. This produced more accurate quality scores to use for filtering low-quality genotypes. We used XGBoostMinGqVariantFilter, a GATK tool [14], to perform the quality score recalibration. This tool applies a decision tree from the XGBoost library for gradient boosted machine learning to predict the quality of a given genotype [15].

The model was trained to assess the probability that a genotype is true given a set of features that include:

- SV class
- SV size
- allele frequency
- existing genotype quality scores
- read evidence support
- source callers
- concordance with raw calls
- overlap with segmental duplication, simple repeat, mappability, and RepeatMasker track intervals

The filtering model was trained on labeled non-reference genotypes described in the lrWGS training data section. The filtering tool annotates each genotype with a scaled logit (SL) score, for which lower (more negative) scores reflect a low probability of being non-reference, higher scores (more positive) a higher probability, and a score of 0 being equally likely. Genotype quality scores were also updated according to SL using the formula:

$$GQ = -10log_{10}\left[\frac{1}{(0.52/0.48)^{SL}+1}\right].$$

Precision and recall were then calculated across a range of SL cutoffs using the following equations:

$$precision = \frac{n_{TRUE}^{PASS}}{n_{TRUE}^{PASS}+n_{FALSE}^{PASS}},$$

$$recall = \frac{n_{TRUE}^{PASS}}{n_{TRUE}^{PASS}+n_{TRUE}^{FAIL}},$$

Where n_X^Y is the number of non-reference srWGS genotypes with truth label *X* and filter status *Y*. Note that a recall of 1 corresponds to retaining all srWGS SV calls with lrWGS support and therefore does not account for false negatives in the initial srWGS SV callset.

Genotype filtering was applied to the same variant types that were used for training (DEL, DUP, INS, and INV). See lrWGS training data for additional details. However, the size restriction on DEL and DUP variants was increased from 5 to 10 kb for filtering, as the variants in this range are expected to have error modes similar to those used for training (under 5 kb). Filtering was not applied to CNVs that were either multi-allelic or over 10 kb in size because those categories lacked training labels.

We filtered each genotype based on a minimum SL cutoff for its SV type and size category. We selected the SL cutoffs to balance gains in precision with losses in recall. For each SV type and size category, we calculated the F score, which is a measure of model performance based on both the precision and recall:

$$F = \left(1 + \beta^2\right)\frac{precision \cdot recall}{\beta^2 precision + recall}$$

where β is an adjustable parameter. We chose cutoffs to maximize the F scores and attain a minimum precision of 90% within each SV type and size category. Failed genotypes were revised to no-call (./.).

We believe that the precision and recall of the filtered callset is high enough for most applications. Researchers who require a higher-precision callset may apply more stringent GQ cutoffs, but should be aware that GQ was calculated under a different model than the SNP and Indel callsets, so typical filtering cutoffs may not produce the desired results.

Results

Analysis of the training samples from lrWGS and genotyping arrays yielded a total of 27,437,577 trainable genotypes, while labels for 15,611,637 genotypes (36% of the total) could not be determined (Figure 2 A). SL scores from the trained model largely recapitulated truth labels, with false positives (FP) and true positives (TP) generally having lower and higher scores, respectively (Figure 2 B).

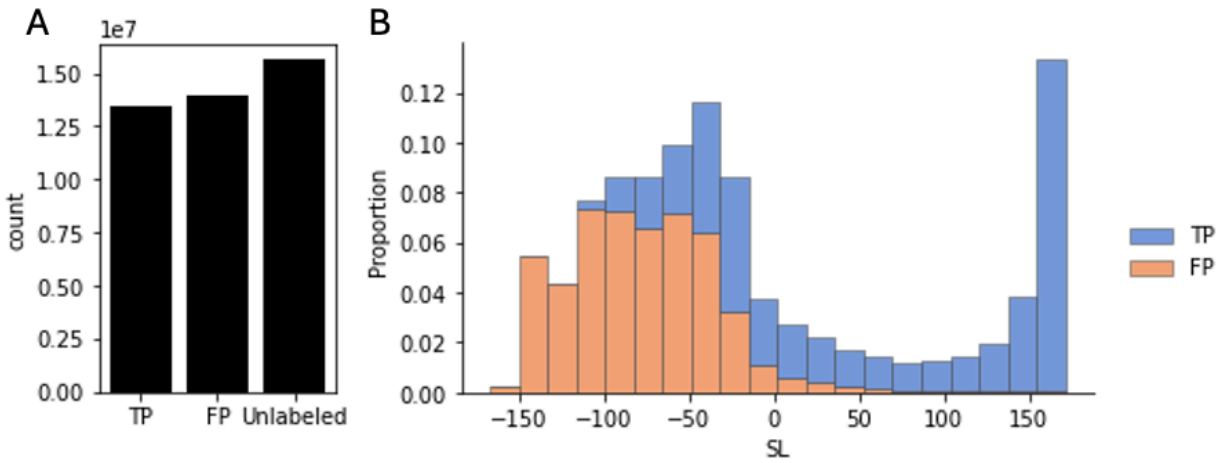


Figure 2 -- Training data for genotype filtering. (A) The proportion of each training label out of all SV genotypes in the training data, and (B) the SL score distribution produced by the trained model.

The genotype filtering performance was evaluated in the test set of 97 held-out samples with matched lrWGS data. We observed that precision decreases consistently as a function of recall when thresholding on SL (Figure 3). This demonstrates that the method is effective for tuning callset accuracy. These results also indicate comparable performance across the spectrum of SV classes. Optimal cutoffs for SL filtering were determined using the training set as described above and are shown in Appendix Table B.1.

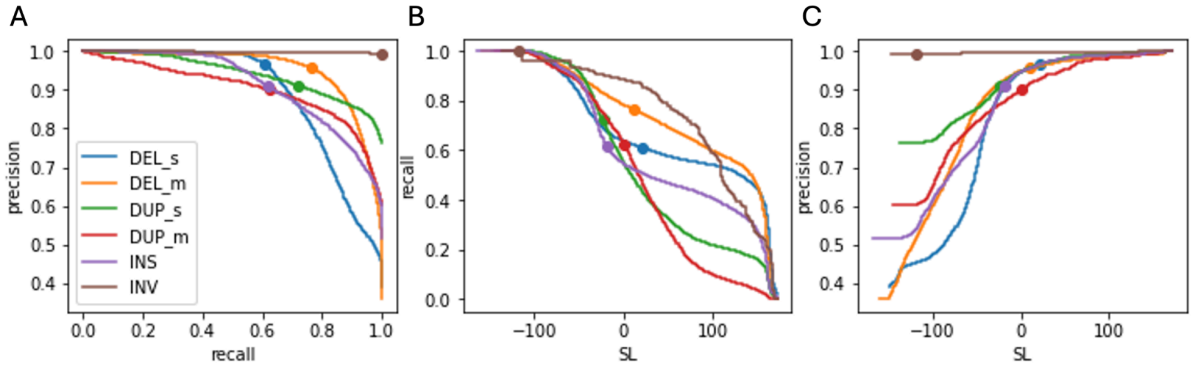


Figure 3 -- SL genotype filtering performance assessed against 97 lrWGS labeled test samples. (A) Precision-recall curves for all filtering classes, (B) recall as a function of the SL cutoff value, and (C) precision as a function of the SL cutoff value. Markers depict cutoffs used for genotype filtering.

We report the performance of the SL genotype filter in Appendix B.

Reclustering in repetitive regions

We applied additional clustering to SVs in repetitive genomic contexts in order to reduce the number of redundant calls. For insertions in simple repeat regions and deletions and duplications under 5 kb in length in simple repeat regions or repeat-masked sequences, we clustered SVs that had 50% reciprocal overlap, had breakpoints within 100 base pairs (bp), and shared 10% of their carrier samples. We further reclustered the subset of deletions 1-5 kb in length in simple repeat regions and repeat-masked sequences that had 70% reciprocal overlap, had breakpoints within 1 kb, and shared 10% of their carrier samples. For deletions and duplications over 5 kb in length in segmental duplications, we clustered SVs that had 30% reciprocal overlap and shared 10% of their carrier samples.

Removal of mCNVs <5kb

Read depth signal is less reliable in events smaller than 5 kb [16]. We removed all MCNVs under 5 kb in length from the callset, so they will not appear in the VCF file. We report MCNVs of greater than 5 kb with the “MULTIALLELIC” filter tag. Therefore, all MCNVs in the final callset will have a length greater than 5 kb and be tagged as “MULTIALLELIC”.

Outlier sample removal

We calculated the distribution of SV counts across all samples stratified by SV type and did not observe any outlier samples, so no samples were removed due to unusually high or low SV counts at this stage.

Batch effect correction

We evaluated each variant for batch effects among the 192 batches used for the batched steps of the GATK-SV pipeline. The filter “VARIABLE_ACROSS_BATCHES” was applied to variants with statistically significant batch effects.

Details of the statistical methods for batch effect correction can be found in the “Assessment of batch effects” paragraph in the supplementary methods of Collins et al 2020 [2]. Please note that PCR-amplified samples are not part of the AoU cohort, and 36,672 pairwise comparisons were not feasible, so we applied only the one-vs-all comparisons described in Collins et al.

Mobile element deletions

GATK-SV requires read depth support for biallelic CNVs greater than 5 kb in size; candidate large CNVs that lack read depth support are retained in the callset but the SV type is revised to breakend (BND) and the filter “UNRESOLVED” is applied. However, deletions of large mobile elements, such as LINE1 and HERVK, are not expected to show significant decreases in sequencing depth due to the presence of reads from other mobile elements across the genome. To rescue these deletions, records of SV type BND were revised to SV type DEL if they met the following criteria: overlap annotated mobile elements by greater than 50%, are less than or equal to 10 kb in size, match the breakpoint orientation indicating a deletion (STRANDS=+-), and are supported by PE evidence. In addition to being annotated as DEL in the SVTYPE field in INFO, the mobile element class was annotated in the ALT field, i.e. DEL:ME:LINE1.

Complex SVs, large inversions, and inter-chromosomal translocations curation

Translocation sensitivity

To improve the sensitivity for inter-chromosomal translocations (CTX) in this callset, we re-evaluated the raw translocation calls from Manta [4]. We clustered the translocation variants across batches of around 500 samples and we retained only the rare variants (<1% allele frequency). We next removed redundant translocations that were within 100 bp of a translocation site already called by GATK-SV within the batch. We manually reviewed the discordant paired end read (PE) evidence for each non-reference genotype as described below. Translocations with sufficient PE evidence were added to the GATK-SV callset.

Filtering complex SVs and translocations

Specific alignment patterns and discordant paired end reads are expected for complex (CPX) and translocation SVs [2]. For example, CPX events involving inversions are expected to have clusters of +/+ and -/- stranded alignments, while those that involve duplications are expected to have -/+ stranded clusters. In addition, read depth (RD) changes are expected if large copy number variants (>5kb) are involved. For CTX, discordant read pairs that link the involved chromosomes are expected.

To improve the precision of the CPX and CTX calls from GATK-SV, the PE and RD evidence was assessed and compared against these expectations. For each CPX and CTX non-reference genotype, the PE evidence within a window of 100-1000 bp around the breakpoints was extracted and compared to the expectation for each sample genotyped as non-reference. We validated the CPX events involving large CNVs for each sample by comparing the non-reference genotypes with the CNV calls generated by raw depth algorithms (i.e. cnMOPS [17] and GATK-gCNV [18]).

For each CPX and CTX genotype, we required PE evidence for all breakpoints and RD evidence when applicable. Genotypes that did not meet these criteria were revised to no-call (./.). Sites with at least 50% of samples lacking depth support with PE evidence at some but not all breakpoints were flagged with the filter status “UNRESOLVED”.

Manual curation of translocations, large inversions, and large complex SVs

To further verify the accuracy of the inter-chromosomal translocations and large inversions and large complex SVs greater than 1 Mb in size, we manually reviewed the PE evidence for these SVs. We evaluated the PE evidence for each carrier sample within a window of 100-1000 bp around the breakpoints according to the following criteria:

1. Each breakpoint should have at least 4 supporting discordant pairs
2. All breakpoints in an event should have a sum of at least 10 supporting discordant pairs
3. The supporting discordant pairs should follow certain patterns:
    a. For deletions, the forward-facing (+) reads should be upstream of the reverse-facing (-) reads, and vice versa for duplications
    b. For translocations with both breakpoints on the same side of the centromere (both on p arms or both on q arms), we expect +- pairs followed by -+ pairs
    c. For translocations with breakpoints on different sides of the centromere (one on a p arm and one on a q arm), we expect ++ pairs followed by -- pairs
4. The supporting reads across each breakpoint should span a minimum of 50 bases
5. Translocation sites should not have a high background level of discordant pairs (greater than or equal to 4 discordant pairs in at least 10 non-carrier samples). This filter was applied because translocation events are expected to be rare, and to remove sites with potential mapping artifacts

Failed genotypes were revised to no-call (./.) and all revisions resulting from manual review are described in the INFO field MANUAL_REVIEW_TYPE.

Large CNV curation

We performed a visual inspection of read depth across all 1,322 CNVs (deletions and duplications) larger than 1 Mb observed in our final VCF using a visualization tool found in GATK-SV [19]. After inspection, we confirmed the presence of 1,310 CNVs (99.1%). We observed that 4 of the CNVs larger than 1Mb appeared to have multiple copy states, so we applied the multiallelic filter tag (MULTIALLELIC). Finally, for 415 CNVs (31.4%) that had at least one sample with inaccurate breakpoints, we manually reassigned breakpoints using the more precise sample level depth calls derived from preceding modules in the pipeline. All revisions resulting from manual review are described in the INFO field MANUAL_REVIEW_TYPE.

Genomic disorder region re-genotyping

Genomic disorders are human diseases largely arising from recurrent CNVs mediated by segmental duplications containing homologous sequences [20]. To improve variant discovery and genotyping accuracy in known genomic disorder (GD) regions [21], we applied local depth-based re-genotyping to large CNVs. The purpose of this step is to ensure that these complex and repeat-mediated events are accurately profiled and not fragmented into smaller events during variant clustering and defragmentation. Briefly, depth evidence of all bi-allelic DEL and DUP sites overlapping at least 40% of a GD region were reassessed to refine breakpoints, remove false positives, and recover false negatives.

Each GD region was padded by 100% of its total length on either side and divided into up to 30 equally-sized bins, which were then genotyped in all samples using the same depth-based methods as the GATK-SV genotyping module. Existing calls were then evaluated across the genotyped bins and either removed or revised depending on the extent of depth support. In addition, samples exhibiting strong depth-based CNV support across at least 50% of a GD region but without a corresponding CNV call triggered creation of rescued variants across the supported intervals. However, variant rescue was not performed if the entirety of the GD region and its flanking regions were fully supported, as these are evidence of a spanning event that would not correspond to the given GD.

This process was implemented as a fully automated workflow, and a subset of the data was reviewed manually for quality control. Revisions resulting from manual review are described in the INFO field MANUAL_REVIEW_TYPE. All DEL and DUP variants with at least 50% reciprocal overlap of a GD region were manually reviewed and annotated with the GD region name in the “GD” field if determined to sufficiently match known GD breakpoints.

No-call rate filtering

To further refine the SV sites, we also filtered on the NCR, which is defined as the proportion of no-call genotypes (./.) among all genotypes. The NCR for each site is annotated in the INFO field, with the exception of MCNVs, which do not use the genotype field. A filter status of “HIGH_NCR” was applied to every variant exceeding an NCR cutoff of 5%.

Reference artifact filtering

We applied the REFERENCE_ARTIFACT filter status to sites at which 99% of samples have homozygous alternate genotypes.

Zero-carrier site removal

We removed sites from the callset if no carriers remained after filtering.

Structural Variant QC Results

Below we detail several metrics of interest for this SV callset. Figure 4 shows the SV counts, stratified by SV type, within the callset. In this figure, we include measures from both the total callset (all variants in the callset, regardless of filter tag) as well as a high-quality callset composed of only variants with a filter tag of PASS or MULTIALLELIC. The remaining figures focus on the high-quality callset. Figure 5 shows the distribution of SV counts per genome, stratified by SV type, in the full cohort and grouped by *All of Us* genetic ancestry groups (see Appendix C). Figure 6 shows the distribution of SV lengths for each SV type; the fraction of SVs decreases with increasing SV size, except for MCNVs, which are always over 5 kb, and INS, which have peaks representing Alu, SVA, and LINE-1 mobile genetic elements [22]. Figure 7 shows the ratios of homozygous reference, heterozygous, and homozygous alternate genotypes at each SV site and the fraction of SV sites that are in Hardy-Weinberg equilibrium.

Additional QC analyses are described in a supplementary document, "Benchmarking and quality analyses on the *All of Us* CDRv7 short read structural variant calls," available in the User Support Hub [1].

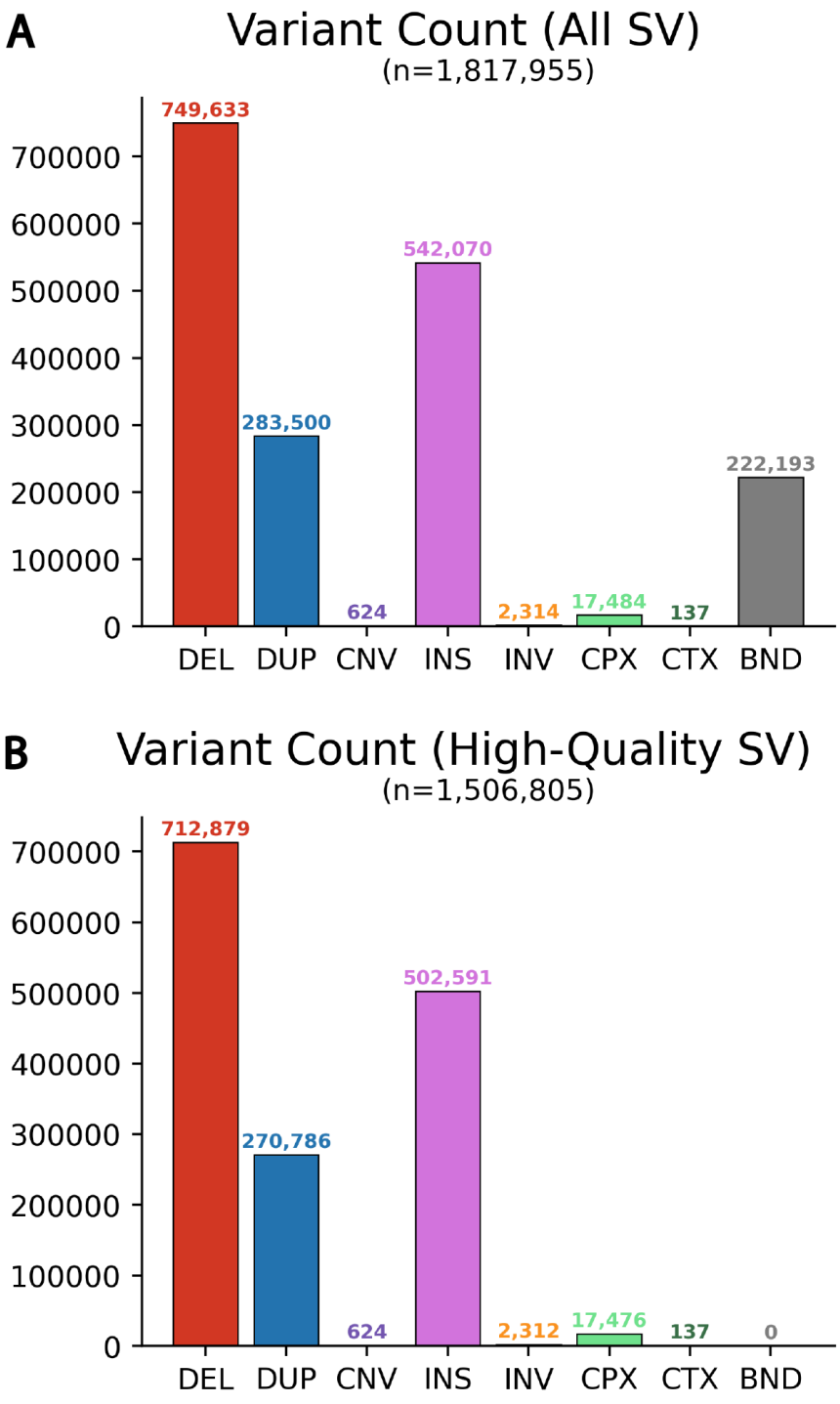
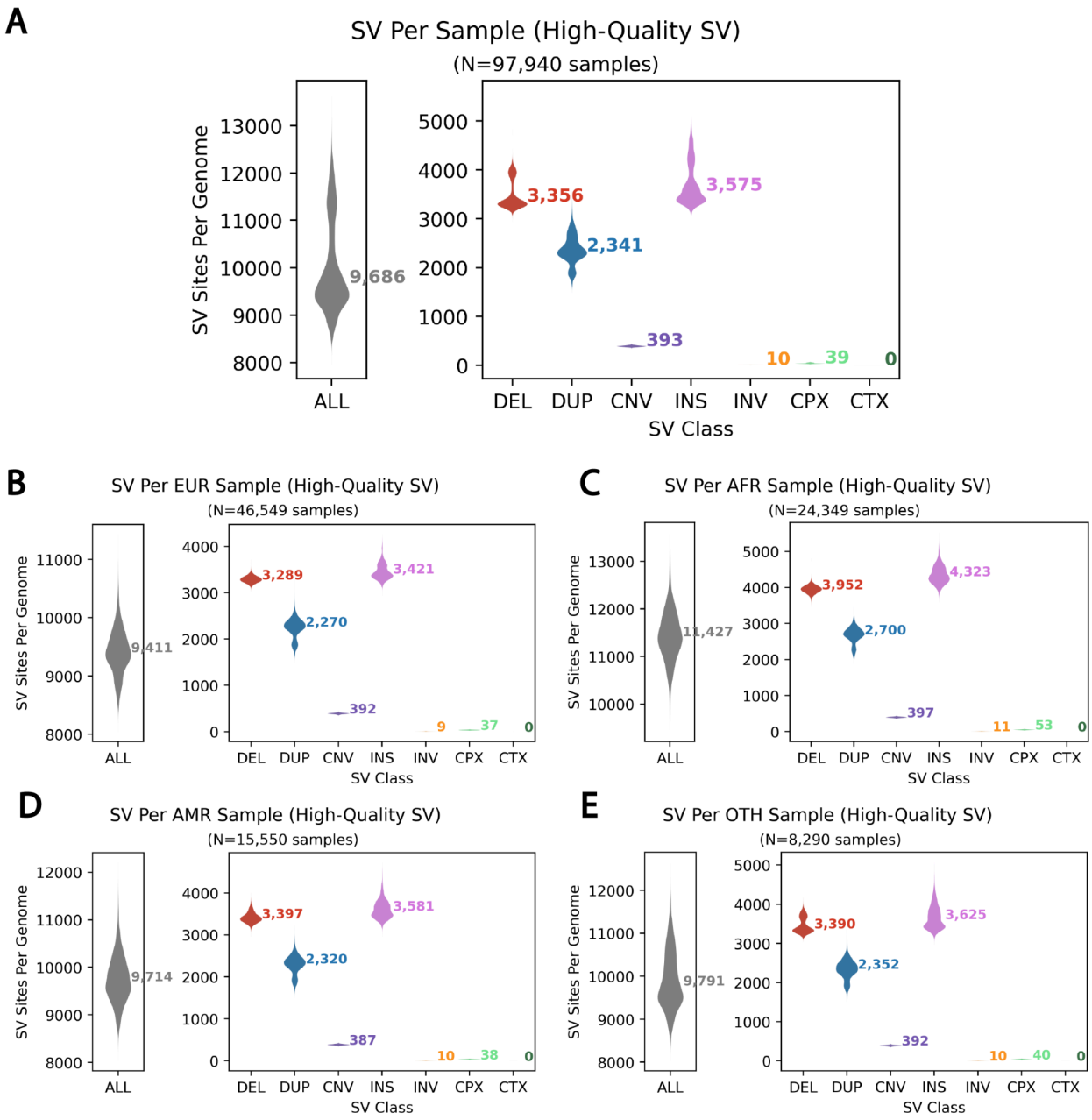


Figure 4 – SV counts in the complete callset and the high-quality SV callset. We observed 1,817,955 total SVs of which we determined 1,506,805 (82.9%) to be of high quality. (A) The total callset includes all variants in the callset regardless of the filter status. (B) The high-quality SV callset only contains variants with the PASS or MULTIALLELIC filter status. Note that all BND sites have the filter UNRESOLVED, so they are not included in the high-quality callset.


A
SV Per Sample (High-Quality SV)
(N=97,940 samples)
SV Sites Per Genome
13000
12000
11000
10000
9000
8000
9,686
ALL
5000
4000
3000
2000
1000
0
3,356
2,341
393
3,575
10
39
0
DEL DUP CNV INS INV CPX CTX
SV Class
B
SV Per EUR Sample (High-Quality SV)
(N=46,549 samples)
9,411
3,289
2,270
392
3,421
9
37
0
C
SV Per AFR Sample (High-Quality SV)
(N=24,349 samples)
11,427
3,952
2,700
397
4,323
11
53
0
D
SV Per AMR Sample (High-Quality SV)
(N=15,550 samples)
9,714
3,397
2,320
387
3,581
10
38
0
E
SV Per OTH Sample (High-Quality SV)
(N=8,290 samples)
9,791
3,390
2,352
392
3,625
10
40
0

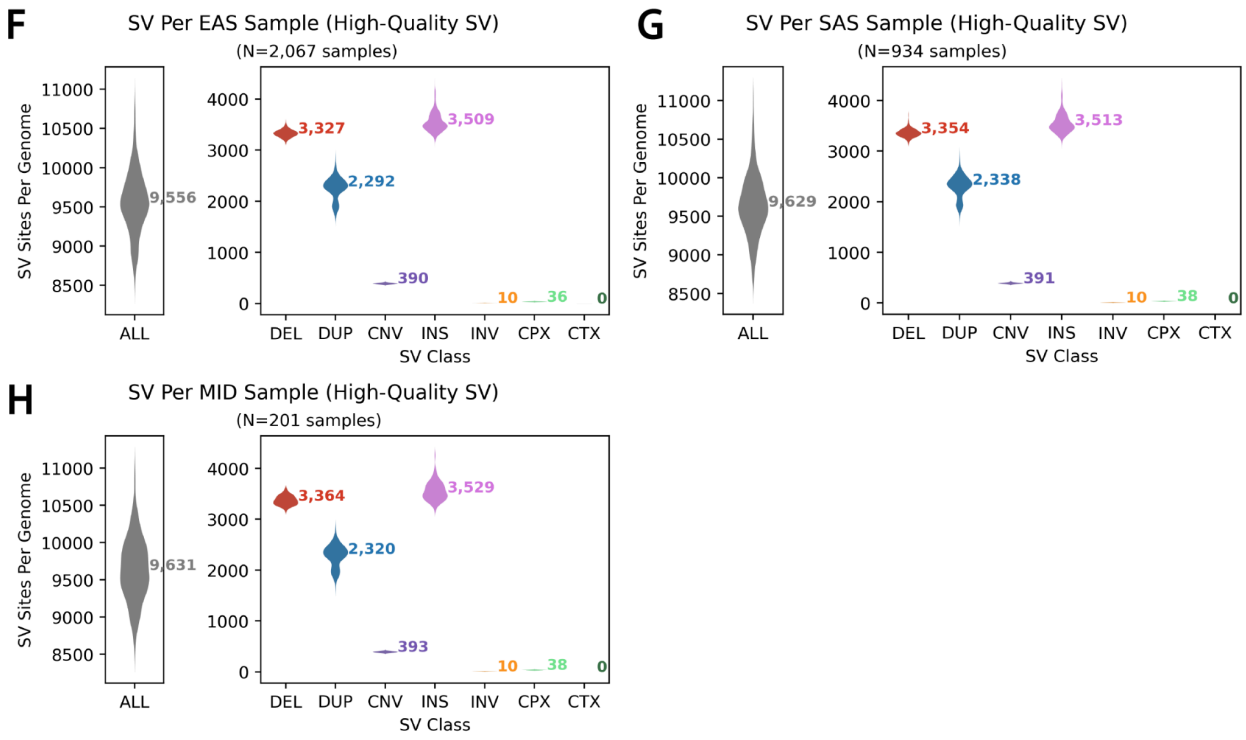


Figure 5 – We observed a median of 9,686 high-quality SVs per person, which is consistent with SVs recently generated on the 1000 Genomes Project samples [23]. We display here the overall SVs per genome and per SV type per genome in the high-quality callset (A) as well as stratifying by the *All of Us* predicted genetic ancestry group in order of prevalence in the callset (B-H). See Appendix C for the *All of Us* genetic ancestry groupings. The median of each distribution is labeled on the plot. As expected, samples in the *All of Us* African/African American genetic ancestry group (AFR) had the highest SV counts while those in the *All of Us* European genetic ancestry group (EUR) had the lowest SV counts.

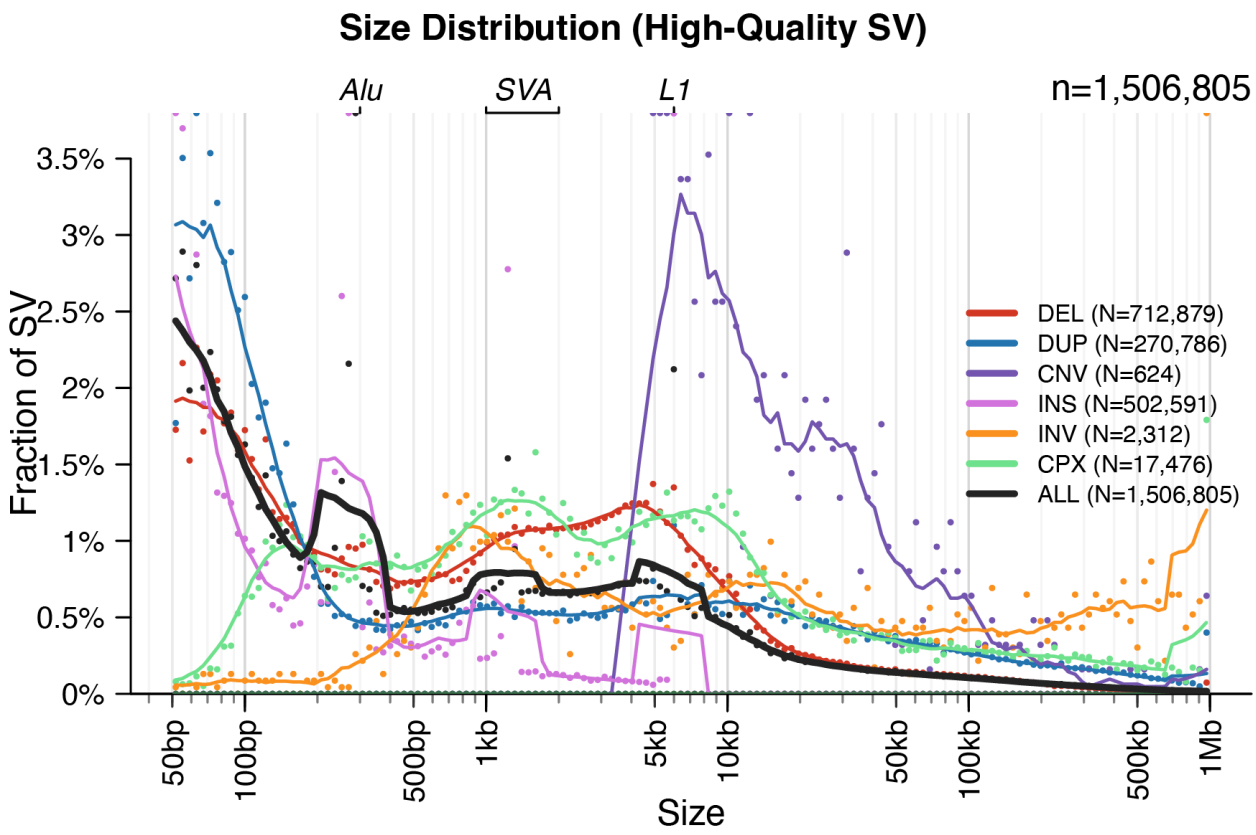


Figure 6 – SV size distribution matches previous expectations with notable insertion peaks corresponding to Alu, SVA, and LINE-1 insertions. Points represent the fraction of each SV type occupied by a given size range. Lines represent the rolling 10-bin average (the size ranges are divided into 150 bins).

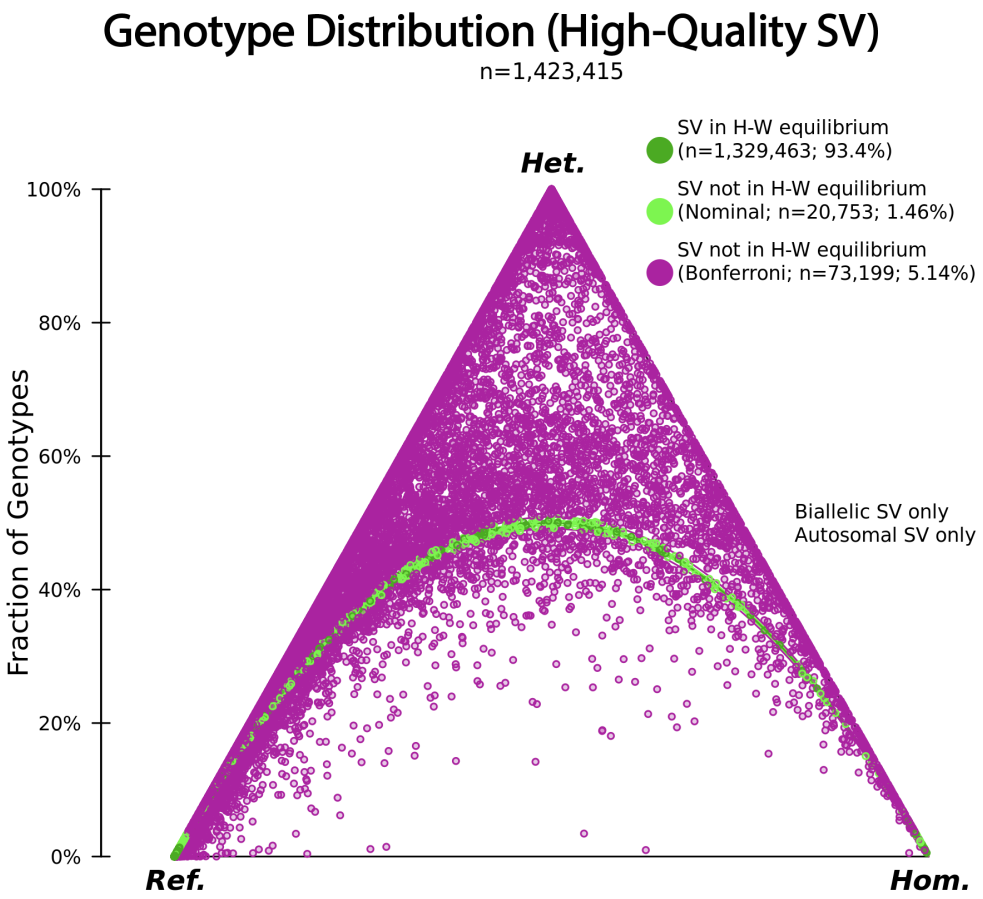


Figure 7 – Among high quality variants, 93.4% are in Hardy Weinberg Equilibrium (HWE). Of the 5.14% that fail, most of these failures appear to be driven by a bias towards genotyping variants as heterozygous. For this calculation, we included only the 94,181 unrelated samples and only biallelic SV sites.

Known Issues

The issue below applies to the CDRv7 SV supplemental dataset. We have provided suggested actions for researchers to workaround the issues and provided remediation plans when necessary. Sample lists relevant to these issues can be found in the User Support Hub [1].

Known Issue #1: Small subset of samples missing corresponding CDR data

Three srWGS SV off-cycle samples are missing their corresponding CDR data. The affected participants are consented to appear in the genomic data.

Affects:

- SV variant files: joint-called VCF

Suggested action:

- If you are not using CDR data (e.g., surveys, EHR), then no action.
- Otherwise, remove samples without corresponding CDR data. We will provide the lists of srWGS off-cycle SV samples without corresponding data in the CDR.

Remediation:

- We will provide lists of srWGS off-cycle SV affected samples through the CDR.

References

[1] ***All Of Us User Support Hub*** https://aousupporthelp.zendesk.com/hc/en-us
[2] Collins, R.L., Brand, H., Karczewski, K.J. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020). https://doi.org/10.1038/s41586-020-2287-8
[3] ***Structural Variants*** (n.d.). Retrieved March 3, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022476791323-Structural-Variants
[4] Chen, X. *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32, 1220-1222. doi:10.1093/bioinformatics/btv710
[5] Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, *et al.* (2015) Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol 11(12): e1004572. https://doi.org/10.1371/journal.pcbi.1004572
[6] Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Mills, R. E., Pittard, W. S., 1000 Genomes Project Consortium & Devine, S. E. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*, 2017. **27**(11): p. 1916-1929.
[7] Jakubek YA, Zhou Y, Stilp A, *et al.* Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. Nat Genet. 2023 Nov;55(11):1912-1919. doi: 10.1038/s41588-023-01553-1. Epub 2023 Oct 30. PMID: 37904051; PMCID: PMC10632132.
[8] Forsberg LA, Rasi C, Malmqvist N, *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet. 2014 Jun;46(6):624-8. doi: 10.1038/ng.2966. Epub 2014 Apr 28. PMID: 24777449; PMCID: PMC5536222.
[9] Zhao X, Weber AM, Mills RE. A recurrence-based approach for validating structural variation using long-read sequencing technology. Gigascience. 2017 Aug 1;6(8):1-9. doi: 10.1093/gigascience/gix061. PMID: 28873962; PMCID: PMC5737365.
[10] P. Ebert, P. A. Audano, Q. Zhu et al., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
[11] **PacBio structural variant calling and analysis tools (PBSV)**, Retrieved March 3, 2023, from https://github.com/PacificBiosciences/pbsv.
[12] **Sniffles2 (PBSV)**, Retrieved March 3, 2023, from https://github.com/fritzsedlazeck/Sniffles
[13] Van der Auwera GA & O'Connor BD. (2020). ***Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)***. O'Reilly Media. P.400
[14] **XGBoostMinGqVariantFilter** (n.d.) Retrieved March 4, 2023, from unreleased GATK branch https://github.com/broadinstitute/gatk/tree/tb_recalibrate_gq
[15] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016
[16] Werling DM, Brand H, An JY et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018 Apr

26;50(5):727-736. doi: 10.1038/s41588-018-0107-y. PMID: 29700473; PMCID: PMC5961723.
[17] Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012 May;40(9):e69. doi: 10.1093/nar/gks003. Epub 2012 Feb 1. PMID: 22302147; PMCID: PMC3351174.
[18] Babadi M, Fu JM, Lee SK, Gauthier LD, Walker M, Benjamin DI, Karczewski KJ, Wong I, Collins RL, Sanchis-Juan A, Brand H, Banks E, Talkowski ME. GATK-gCNV: A Rare Copy Number Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank. bioRxiv, 2022.
[19] **VisualizeCnvs.wdl** (n.d.) Retrieved March 4, 2023, from https://github.com/broadinstitute/gatk-sv/blob/v0.26.5-beta/wdl/VisualizeCnvs.wdl
[20] Carvalho, C., Lupski, J. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**, 224–238 (2016). https://doi.org/10.1038/nrg.2015.25
[21] Collins, R. L., Glessner, J. T., Porcu, et al. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell, 185*(16), 3041–3055.e25. https://doi.org/10.1016/j.cell.2022.06.036
[22] Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics, 12*, 187–215. https://doi.org/10.1146/annurev-genom-082509-141802
[23] Byrska-Bishop, Marta et al. "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios." *Cell* vol. 185,18 (2022): 3426-3440.e19. doi:10.1016/j.cell.2022.08.004
[24] ***Structural variant (SV) discovery*** (n.d.). Retrieved March 15, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022487952155-Structural-variant-SV-discovery\
[25] **WDL Specification**, from https://github.com/openwdl/wdl
[26] Karczewski, K.J., Francioli, L.C., Tiao, G. *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7
[27] M'Charek, A. **The Human Genome Diversity Project: An Ethnography of Scientific Practice** (Cambridge Studies in Society and the Life Sciences). Cambridge: Cambridge University Press. (2005) doi:10.1017/CBO9780511489167
[28] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393
[29] Ho, TK. ***Random Decision Forests.*** Proceedings of the 3rd International Conference on
Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
[30] Scikit-learn: Machine Learning in Python, Pedregosa *et al., Journal of Machine Learning Research* 12, pp. 2825-2830, (2011).
[31] Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 581**,** 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

[32] Frankish A, Diekhans M, Jungreis I, et al. **GENCODE 2021**, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D916–D923, https://doi.org/10.1093/nar/gkaa1087
[33] ***Genetics - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/genetics.html#hail.methods.hwe_normalized_pca.

Appendix A: srWGS Structural Variant Pipeline

The GATK-SV pipeline was applied to detect SVs from srWGS data [2]. GATK-SV is an ensemble method which applies multiple SV callers to increase sensitivity and leverages different types of evidence to refine SV calls and remove false positives. The SV callers used for this callset were Manta [4] and Wham [5] to leverage PE and split-read (SR) evidence, MELT [6] to specifically target mobile elements, and GATK-gCNV [18] and cn.MOPS [17] to detect large copy-number variants (CNVs) from read depth (RD) evidence. Following candidate SV discovery by these algorithms, GATK-SV re-evaluates the PE, SR, RD, and B-Allele Frequency (BAF) evidence for each variant from the raw reads to improve precision. Each candidate SV is jointly genotyped in every sample in the cohort, and then SV signatures are integrated to resolve complex variants involving more than one SV type. An overview of the GATK-SV algorithms and evidence types can be found at [24], and details of the method can be found in Collins et al 2020 [2]. Code and technical documentation can be found on GitHub (https://github.com/broadinstitute/gatk-sv). This includes automated workflows written in Workflow Definition Language (WDL) [25].

Notable improvements to the GATK-SV pipeline since the CDRv7 srWGS SV release include:

- More precise SR-based genotyping and breakpoint determination for INS variants
- Refined functional consequence annotations for CPX variants
- Added annotations of allele frequency from gnomAD-v4.1 SVs for variants present in both callsets [26]
- Improved the depth-based genotyping method for very large CNVs to address an issue observed and manually fixed in the v7 srWGS SV callset
- Performance and scaling enhancements

The full release history for GATK-SV can be found at https://github.com/broadinstitute/gatk-sv/releases.

Figure 1 depicts the steps of the pipeline as it was run in AoU. Table A.1 provides further details on the software versions and how the steps were run. The software versions vary from step to step because the latest version of each workflow available at the time was used in order to incorporate the latest improvements. The main pipeline modules were run as Terra workflows, in which case the GitHub release version and entity to which the workflow was applied (sample, arbitrary partition of samples, batch, cohort) is noted. Steps for which there was not an established workflow, such as QC and batching, were performed in Jupyter notebooks in Terra in Python.

Table A.1-- GATK-SV Pipeline Versions and Notes

Workflow/Step Name	Version Used	Entity	Notes
Sample selection	Notebook		See Sample Selection
GatherSampleEvidence	v0.24-beta	Sample	SV callers used: Manta, Wham, and

			MELT. All 88,882 samples completed this step, with a 0.00% initial failure rate.
EvidenceQC	v0.26.6-beta	Arbitrary partition of samples	Run on arbitrary partitions of samples.
Single sample QC	Notebook		See Single Sample QC
Batching	Notebook		See Batching
TrainGCNV	v0.24-beta	Batch	Batches of samples were created according to the scheme described in the main text under Batching
GatherBatchEvidence	v0.26.7-beta	Batch	Depth-based CNV callers used: GATK g-CNV and cn.MOPS.
ClusterBatch	v0.25.1-beta	Batch	
PlotSVCountsPerSample	v0.27.1-beta	Batch	
SubsetVcfBySamples	v0.27.1-beta	Batch	We removed the 11 significant outliers identified for duplication and deletion counts (nIQR cutoff = 10).
GenerateBatchMetrics	In development (git commit 769811f2)	Batch	This version has since been merged and released as v0.28-beta
FilterBatchSites	v0.24.3-beta	Batch	
PlotSVCountsPerSample	v0.27.1-beta	Batch	No SV count outliers observed.
FilterBatchSamples	v0.26.10-beta	Batch	No outlier samples were removed at this stage (nIQR cutoff = 10000).
MergeBatchSites	v0.24-beta	Cohort	For cohort-level steps, data from all samples across all batches was merged.
GenotypeBatch	v0.28.1-beta	Batch	
RegenotypeCNVs	v0.28.1-beta	Cohort	
CombineBatches	v0.24-beta	Cohort	
ResolveComplexVariants	v0.28.2-beta	Cohort	
GenotypeComplexVariants	In development (git commit 424ca4f)	Cohort	A developmental version of GenotypeComplexVariants was used for improved scaling

CleanVcf	v0.28.3-beta	Cohort	
Filtering and refinement	Multiple steps	Cohort	See Joint Callset Refinement & QC. Filtering and refinement was performed in a series of workflows and notebooks.
AnnotateVcf	In development (git commit 71e73c6)	Cohort	A developmental version of AnnotateVcf was used for improved scaling

Appendix B: Overall precision and recall after SL filtering

Table B.1 summarizes performance after SL filtering across SV classes. Overall recall/precision were 0.646/0.926 in the training set and 0.648/0.927 in the test set with similar performance observed across the spectrum of SV classes. These results indicate that the model generalizes accurately to unseen data.

Table B.1 -- Genotype filtering performance after applying SL and NCR cutoffs

Filtering class	Min size (bp)	Max size (bp)	SL cutoff	Corresp-onding GQ	Train		Test	
					Recall	Precision	Recall	Precision
Small DEL	50	500	21	42	0.604	0.964	0.610	0.965
Medium DEL	500	5,000	11	38	0.759	0.955	0.765	0.955
Large DEL*	5,000	inf	NA	NA	NA	NA	NA	NA
Small DUP	50	500	-23	26	0.719	0.910	0.722	0.910
Medium DUP	500	5,000	1	35	0.621	0.901	0.625	0.900
Large DUP*	5,000	inf	NA	NA	NA	NA	NA	NA
INS	50	inf	-19	28	0.619	0.907	0.619	0.908
INV	50	inf	-118	0	0.999	0.994	0.999	0.994

*Large DEL and DUP variants were tested in a separate analysis. The results will be reported in the supplementary SV QC document, Benchmarking and quality analyses on the *All of Us* CDRv7 short read structural variant calls, which can be found on the User Support Hub [1].

Appendix C: *All of Us* genetic ancestry groups

We assigned genetic ancestry labels to all participants with CDRv7 srWGS SNP and indel data, as described in Appendix A in the CDRv7 QC report [1]. The labeling is based on gnomAD [26], the Human Genome Diversity Project (HGDP) [27], and the 1000 Genomes (1KG) [28] genetic ancestry labels (Table C.1).

We used the high-quality set of sites (described in Appendix J in the CDRv7 QC report [1]) and trained a random forest classifier [29,30] on a training set of the HGDP and 1KG sample variants on the autosomal exome, obtained from gnomAD [31]. This exome was derived from the exon regions of all autosomal, basic, protein-coding transcripts in GENCODE v42 [32].

We generated the first 16 principal components (PCs) of the training sample genotypes (using the hwe_normalized_pca in Hail [33]) at the high-quality variant sites for use as the feature vector for each training sample. We used the truth labels from the sample metadata, which can be found alongside the VCFs. Note that we do not train the classifier on the samples labeled as "Other." We use the label probabilities ("confidence") of the classifier on the other ancestries to determine ancestry of "Other".

To assign genetic ancestry groups for each participant, we project the genotypes at the high-quality set of variant sites of the *All of Us* samples into the PCA space of the training data. We then apply the classifier (see Figure A.1 in the CDRv7 QC report [1]). Since we do not have truth labels, we can not determine the accuracy of our *All of Us* predictions.

Table C.1 The *All of Us* genetic ancestry groups with descriptions

***All of Us* genetic ancestry group**	***All of Us* v7 genetic ancestry group label**	**Notes**
African/African American	AFR	
Admixed American	AMR	
East Asian	EAS	
Middle Eastern	MID	
European	EUR	
South Asian	SAS	
Remaining (Other)	OTH	Not belonging to one of the other genetic ancestries or is a balanced admixture

Appendix D: Self-reported race/ethnicity

As seen in Table D.1, the race/ethnicity breakdown of the structural variant genomic data is similar to all participants *All of Us* CDR release C2022Q4R9_offcycle. Samples with “Skip” responses include participants that answered “prefer not to answer”, entered blank text, or did not respond to the survey question.

*Corresponding survey data are missing for three participants. Please see Known Issue #1 for more information.

Table D.1 -- Self-reported Race/Ethnicity breakdown of the genomic data

Self-reported Race/Ethnicity	srWGS SV counts (%)
Asian	2,888 (2.90%)
Asian, White	384 (0.40%)
Black	22,446 (22.90%)
Black, White	627 (0.60%)
Hispanic	16,778 (17.10%)
Hispanic, White	1,324 (1.40%)
MENA	499 (0.50%)
Other	2,968 (3.00%)
Skip	2,150 (2.20%)
White	47,873 (48.90%)
Total	**97,937 (99.99%)***