All of Us Research Program

Genomic Research Data Quality Report

All of Us Curated Data Repository (CDR) release C2022Q4R9

Overview

This document details the *All of Us* Genome Centers (GC) and Data and Research Center (DRC) quality control (QC) steps for genomic data in the research pipeline. This pipeline removes or flags samples and variants in the genomic data that fail quality thresholds. We apply these QC steps in the research pipeline before we release the genomic data for research use. We, the *All Of Us* DRC, only describe QC processes that are performed analytically (i.e., after the sample has been genotyped and sequenced). All descriptions and results are limited to the v7 data release made available in the Researcher Workbench April 20, 2023, which contains 312,945 genotyping array ("array") samples, 245,394 short read whole genome sequencing (srWGS) samples with single nucleotide polymorphism, insertion, and deletion variant calls (SNPs and Indels), 11,390 srWGS samples with structural variant (SV) calls, and 1,027 long read whole genome sequencing (lrWGS) samples with SNP, Indel, and SV calls. The srWGS SV samples and lrWGS samples are a subset of the srWGS SNP and Indel samples, which in turn are a subset of the array data. The samples in the genomic data correspond to the *All of Us* Curated Data Repository (CDR) release C2022Q4R9 ("v7"), though please see Known Issue #1, as 20 array samples (less than 0.01%) and six srWGS samples (less than 0.01%) are missing their corresponding CDR data. These pipelines are automated unless otherwise noted. This document covers all genomic data types made available to researchers at this time including small variants (SNPs and Indels), structural variants, raw data, and auxiliary data. Small variants are available for array samples, srWGS samples, and lrWGS samples. Structural variants are available for srWGS samples and lrWGS samples.

Audience: This document is intended for researchers using, or considering the use of, the genomic data in the Researcher Workbench (RW). This document assumes knowledge of sequencing, genotype arrays, common genomic data QC approaches, and the variant file formats released in *All of Us*. We recommend that at a minimum researchers read the **Known Issues** and the **FAQ** section below, even if they are not as concerned with the QC process.

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 document', 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. To register for access, please go to https://www.researchallofus.org/register/
- A small number of array and srWGS SNP & Indel samples are missing their corresponding CDR and Cohort Builder data and thus the sample counts are not the same. Please see Known Issue #1 for more details.

Executive Summary

On April 20, 2023, the *All of Us* Research Program released the genomic data of 312,945 array samples, 245,394 srWGS samples, and 1,027 lrWGS samples in the Researcher Workbench (RW) for use by researchers registered for Controlled Tier access. There are over 1.8M array SNP and Indel sites, over 1.03B srWGS SNP and Indel sites, over 64 million long read SNP and Indel sites on grch38_noalt, and over 73 million long read SNP and Indel sites on T2T-CHM13v2.0. There are over 515,427 SVs called on the 11,390 srWGS sample cohort and SVs called for each lrWGS sample. In addition to variant calls, raw data (IDAT files for array data, CRAM files for srWGS data, BAM files for lrWGS data) and auxiliary files (predicted ancestry, relatedness/kinship scores, functional annotation, and flagged samples) are available in the RW through Controlled Tier access. Quality control processes, performed both independently and across samples, indicate that these data are ready for general analysis. We suggest researchers, at a minimum, read the Known Issues and FAQ sections below before using the data.

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. As the program continues, the DRC will periodically release genomic data - in sync with planned CDR release timelines. This document describes the third release of genomic data to *All of Us* researchers (v7) made available in the RW on April 20, 2023. The genomic data contains 312,945 array samples, 245,394 srWGS samples, 11,390 srWGS samples with SV calls, and 1,027 lrWGS samples from a diverse set of participants (see Appendix A and Appendix B). Genomic data can be joined with other data types (e.g. survey data) for analysis (Appendix C), though please see Known Issue #1. In this document, we describe the QC processes applied to the array, srWGS, and lrWGS data. We describe which processes were performed at the GCs and which were performed at the DRC (see Appendix D), but for most researchers this demarcation has no practical significance.

This document is organized by data type and describes the QC processes performed. For each data type, we will outline the consistency, single sample QC, and joint callset QC.

1. Consistency is the uniformity of protocols at each GC that reduce the probability of batch effects and normalize the data across GCs. Descriptions in this document, for both QC and sample processing, apply to all GCs unless otherwise noted.
2. 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. We also use these tests to confirm internal consistency between the GCs and the DRC. These tests detect sample swaps, cross-individual contamination, and sample preparation errors.
3. Joint callset QC are the processes executed on the joint callset, which use information across samples to flag samples and variants. The QC steps are performed after single sample QC, during creation of the joint callset. The flagged samples and variants are not removed from the callset unless otherwise specified.

We have also performed data validation experiments, such as replicating GWAS results, but the results are shown in other, upcoming documentation (see the User Support Hub [1] or Tutorial Workspaces in the RW, the RW requires authorization to access).

Arrays

There are 312,945 array samples in the v7 release. The SNP and Indel variants from array samples are available in VCF, Hail, and PLINK formats. In addition, raw Array data is available in IDAT format. The data is described in the 'How the *All of Us* Genomic data are organized' article on the User Support Hub [1]. The QC process for array data includes consistency and single sample QC steps. Array data is not joint-called so no joint callset QC was performed.

Consistency across Genome Centers

The genome centers (GCs) established a consistent sample and data processing protocol for array data generation to attenuate the likelihood of batch effects across GCs.

The GCs generate variant calls (VCFs) that are submitted to the DRC. The GCs use the same lab protocols, scanners, software, and input files:

- GCs generate raw intensity data (.idat) using the same hardware (iSCAN scanners from Illumina). These files will still contain biases across GCs.
- GCs normalize the raw intensity data onto the same scale. This process yields a normalization transform for probe intensities, which are one of the inputs for variant calls. The array cluster definition file (.egt) was updated between this release and the prior release. This update was done to reduce variation across GCs. Each GC used the newly defined clusters to generate variant calls as well as reprocessing array samples from the prior release.
- GCs use identical pipelines to generate VCFs, including identical pipeline versions and input parameters, where applicable. As a result, the VCFs contain the same information, regardless of GC, including metadata about inputs.

Please see Appendix E for details.

Single Sample QC

For array samples, we perform sex concordance, call rate tests, and test cross-individual contamination. These tests are designed to detect sample swaps and sample preparation errors and are performed at the GCs. The list of specific QC processes and an overview of the results can be found in Table 1. Some srWGS QC processes, such as Fingerprint Concordance, use array data.

For more details about the array single sample QC process, including preparation, see Appendix E.

Table 1 -- Array Single Sample QC processes

QC process	Passing criteria	Error modes addressed	v7 release results

Sex concordance	Sex call is concordant with self-reported sex at birth. OR Self-reported sex at birth reported as “Other” or was not reported	-Sample swaps	All array samples are concordant. *Other refers to a participant self-reporting “Intersex”, “I prefer not to answer”, or “none of these fully describe me”
Call rate	> 0.98 (> 98%)	-Sample contamination -Sample preparation error	All array samples meet the threshold.
Cross-individual contamination rate	No passing criteria	-Sample contamination from another individual	For arrays, we only report the contamination rate, but do not filter array samples, since the call rate is a proxy for high levels of contamination.

Sex Concordance

We checked the computed sex against the self-reported sex assigned at birth for concordance. We used gencall to determine the computed sex and CDR data for the self-reported sex assigned at birth (Appendix F). If the two sources were not concordant, we assumed a potential sample swap, removed the sample, and investigated the source of the swap.

Method

We call the gencall tool [2] v3.0.0 to make a call on the sex of the sample from the array data. We use the Picard 2.26.0 tool, CollectArraysVariantCallingMetrics [3], to perform the actual concordance check against the self-reported sex assigned at birth. If we do not have a “male” or “female” for the sex assigned at birth, because the participant reported it as “Intersex”, “I prefer not to answer”, “none of these fully describe me”, or skipped the question, we passed the sex concordance check for that sample, regardless of the information from gencall. The sex assigned at birth data from the CDR is described in Appendix F.

To generate sex calls from the array, we call gencall from the Illumina Array Analysis Platform Genotyping Command Line Interface (iaap-cli):

Parameter	Value	Notes
Tool name	“gencall”	
Manifest file	Bead pool manifest (BPM)	Illumina-supplied file that contains metadata (alleles, mapping information, source, etc.) for all of the probes on the genotyping array.
Cluster file	Cluster file (EGT)	Used for normalization of intensities across GCs
-f	Location of the IDAT (.idat) files	
-i	“1”	Algorithm version

--gender-estimate-call-rate-threshold	-0.1	This effectively disables the sex estimation.

To ensure concordance with the self-reported sex assigned at birth, we call CollectArraysVariantCallingMetrics with the following parameters from the Picard toolkit:

Parameter	Value
Tool name	“CollectArraysVariantCallingMetrics”
INPUT	Array single sample VCF
DBSNP	"gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.dbsnp138.vcf"

Results

Since we catch sex concordance failures before including a sample in the release, all array samples in the v7 release passed a sex concordance check. Note that 2.09% of array samples passed the sex concordance check solely because they did not answer “male” or “female” on the self-reported sex assigned at birth question. Appendix F has more details on this CDR question and responses.

Call Rate

Method

The call rate is the number of successful variant calls divided by the number of probes. We invoke the gencall tool [2] v3.0.0, as described above in the Sex Concordance QC process. The gencall tool generates both sex calls and the call rate. We also invoke CollectArraysVariantCallingMetrics with the same parameters as the above section to extract the call rate metric from the VCF header.

We applied a threshold of 0.98 to the call rate for inclusion in the v7 release.

Results

As seen in Figure 1, we did not include any samples that were below the call rate threshold of 0.98. See Figure 2 for cross-GC call rate frequencies. Please note that differences in call rates between males and females will cause a double peak in call rate frequencies, since sites on chrY will have a lower call rate for females.

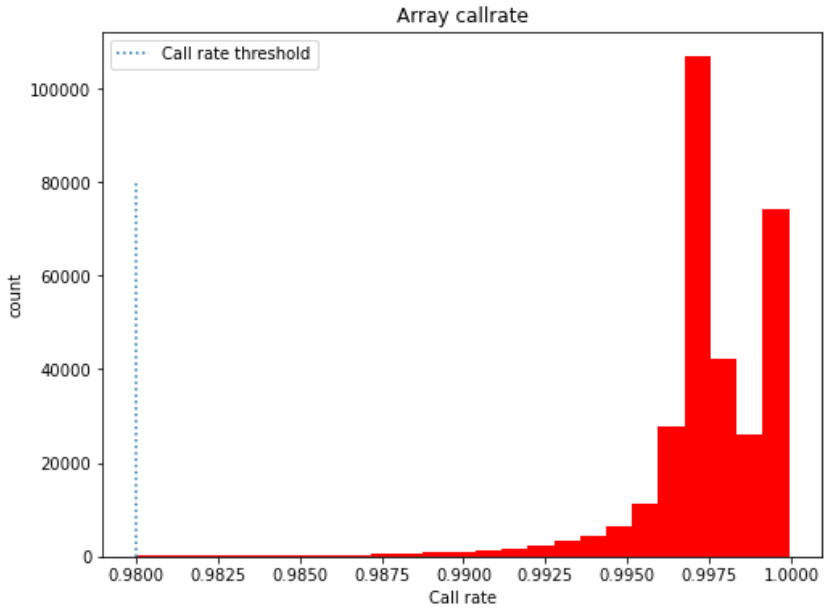


Figure 1 -- Histogram of the array call rate for the v7 release.

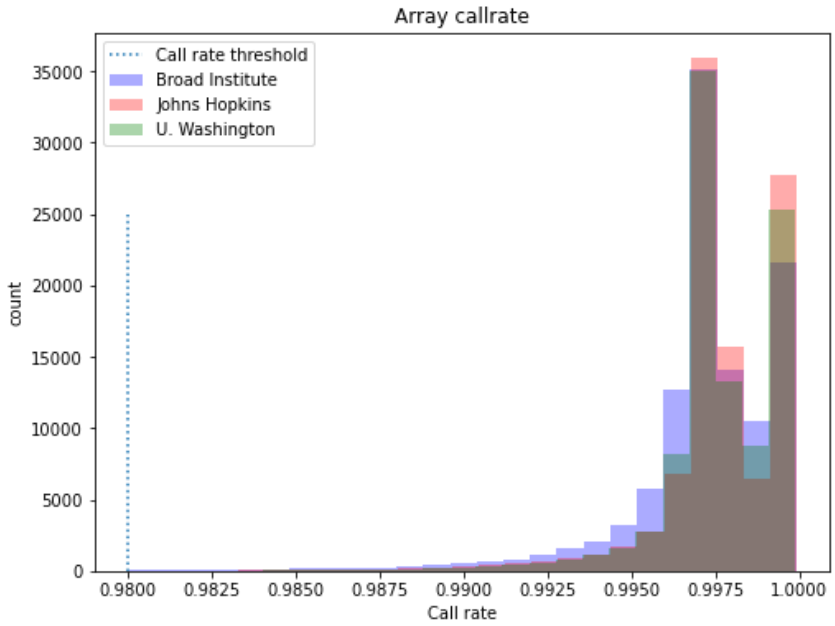


Figure 2 -- Call rate across each GC.

Cross-Individual Contamination Rate

For all samples, we estimate the proportion of data coming from an individual other than the one being processed, referred to as the contamination rate. For array samples, as the contamination rate increases, we expect a lower call rate. We fail array samples for a call rate that does not meet the threshold.

Method

We use BAFRegress [4] to estimate the contamination rate in our array data. We do not use the cross-individual contamination rate to filter array samples, and we do not process the corresponding WGS aliquots for any array sample with a contamination greater than 10%. We filter samples based on the call rate, which is a proxy for contamination and other errors, such as sample preparation errors. Note that most samples with a contamination rate greater than 10% will also not meet the call rate threshold.

We extract allele frequency information from the array VCF and convert it into the file format expected by BAFRegress. We then invoke BAFRegress with the following parameters:

Parameter	Value
task	“estimate”
freqfile	Allele frequency information for all sites, which was extracted from the single sample array VCF.

Results

We estimated the contamination rate below 0.11 for all array samples. As the contamination rate increased, we did see a small decrease in the call rate (see Figure 3). Of the 312,945 array samples, 309,972 (99.0%) had an estimated contamination rate below 3.0% and 303,044 (96.8%) had a contamination rate less than 1%.

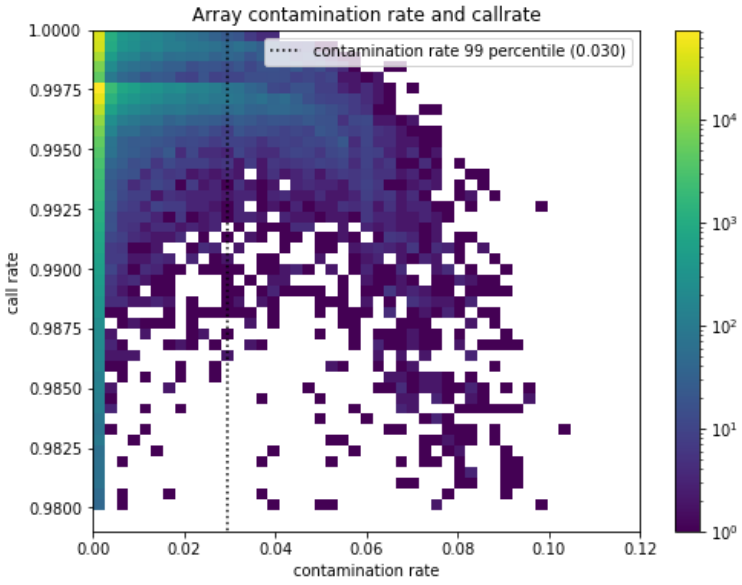


Figure 3 -- Histogram of the array contamination rate estimates vs call rate. As the contamination rate increases, the call rate decreases.

Short Read Whole Genome Sequencing (srWGS)

Consistency across Genome Centers

The GCs use the same protocol for library construction (PCR Free Kapa HyperPrep), sequencer (NovaSeq 6000), software (DRAGEN v3.4.12), and software configuration. The software produces the metrics that are consumed by the sample QC processes. For more information about the sequencing processes used by the GCs, see previous work [5] and the NIH *All of Us* Research Program's Return of Genetic Results FDA IDE (G200165).

Single Sample QC

The list of specific QC processes for srWGS samples and an overview of the results can be found in Table 2. Our WGS single sample QC uses the same sequencing process described previously [5] and in the NIH *All of Us* Research Program's Return of Genetic Results FDA IDE (G200165). Most thresholds in our single sample QC process are identical to the clinical pipeline described previously [5], except for a higher contamination rate.

In some cases, we perform these tests twice for two reasons: 1) to confirm internal consistency between the GCs and the DRC and 2) to mark samples as passing (or failing) QC based on the research pipeline criteria. In this document, we are focused on downstream analytical QC processes after a sample has been sequenced, thus, there are some upstream processes not described here. The list of specific QC processes and an overview of the results can be found in Table 2.

Table 2 -- srWGS Single Sample QC processes

QC process	Calculated at the DRC or GCs?	Passing criteria	Error modes addressed	v7 release results
Fingerprint concordance	Both	log-likelihood ratio > -3	-Sample swaps -Large amount of sample contamination	All srWGS samples are concordant with array samples.
Sex concordance	Both	Sex call is concordant with self-reported sex at birth. OR Self-reported sex at birth reported as "Other" or was not reported	-Sample swaps	All srWGS samples are concordant. *Other refers to a participant self-reporting "Intersex", "I prefer not to answer", or "none of these fully describe me"
Cross-individual contamination rate	Both	< 0.03 (< 3%)	Sample contamination from another individual	All srWGS samples meet the threshold. srWGS samples with corresponding arrays that have a contamination rate above

				10% were not released.
Coverage	GCs only	≥ 30x mean coverage ≥ 90% of bases at 20x coverage ≥8e10 aligned Q30 Bases ≥ 95% at 20x in regions of the 59 AoU Hereditary Disease Risk genes (AoUHDR) See Appendix G for more information	-Sample preparation error -Poor sensitivity and precision of variant calling	All srWGS samples meet the thresholds.

Fingerprint Concordance

Method

We filter variant calls to 113 sites ("fingerprint") for both the array and srWGS SNP & Indel variants. We measure the concordance between the array and WGS data, using a log-likelihood ratio (fingerprint LOD) based on reads. We chose the threshold value, -3.0, to split a bimodal distribution (not shown). If the calls are not concordant (i.e., the fingerprint LOD does not meet the threshold), then there has likely been a sample processing error. A detailed description of fingerprint concordance is described in the Genome Analysis Toolkit documentation [6].

Note: *One GC (Broad Institute) performed an internal check against a different fingerprint (Fluidigm SNP genotyping (SNPtype chemistry) using the 96.96 Dynamic Array), which did not use the same fingerprint sites as the array. The DRC treated these samples the same as from the other GCs and ran the array concordance as described in the main text of this document.

We call the fingerprint concordance tool "CheckFingerprint" using Picard (version 2.23.9) with the following parameters:

Parameter	Value
program name	"CheckFingerprint"
INPUT	The WGS cram to check concordance
REFERENCE_SEQUENCE	"gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta"
GENOTYPES	VCF from corresponding array file
HAPLOTYPE_MAP	"gs://gcp-public-data--broad-references/hg38/v0/aou/fp/aou.fp.haplotype_database.txt"
IGNORE_READ_GROUPS	"true"

SAMPLE_ALIAS	Chipwell barcode from the header of the array file (array file passed in the GENOTYPES parameter)

Note: Quoted parameters are exact values, but quotes were not included in the actual call to the tool.

Results

All samples in the v7 release passed the fingerprint concordance check based on arrays. As seen in Figure 4, the passing samples exceeded the threshold. 1490 samples had a fingerprint LOD [6] less than 45 and the minimum fingerprint LOD was 13.

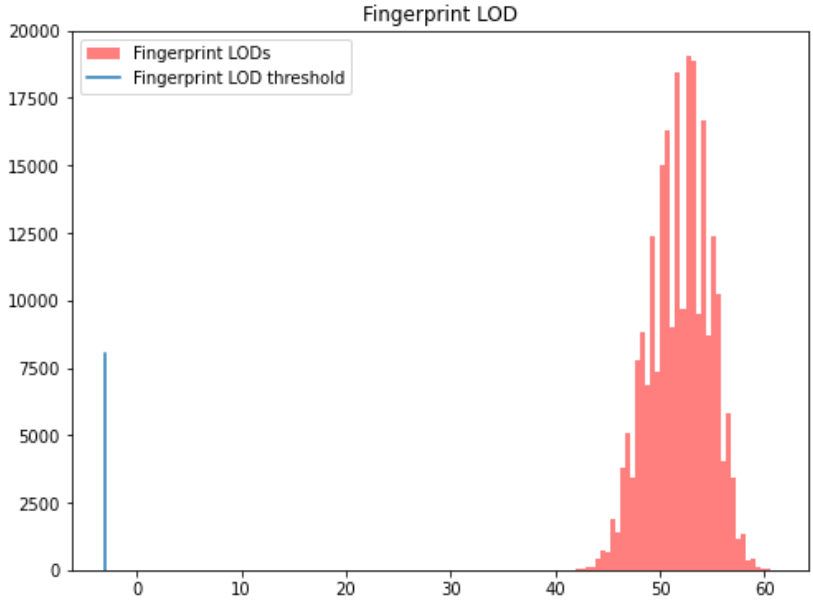


Figure 4 -- Distribution of the Fingerprint LODs for srWGS v7 samples

Sex Concordance

For srWGS data, we compared the computed sex from DRAGEN (Appendix H) and peddy [7] against the self-reported sex assigned at birth (Appendix F). If the two sources were not concordant, we assumed a potential sample swap, removed the sample, and investigated the source of the swap.

Method

We compared variant and ploidy calls for chromosome X and Y against the self-reported sex assigned at birth for the sample. We check the sex ploidy call (e.g., XY or XX) from the DRAGEN pipeline (v 3.4.12) and use heterozygous chrX variant calls from peddy [7]. If the concordance test fails against either of these calls, the sample fails QC and is not included in the release. If we do not have a “male” or “female” for the sex assigned at birth, because the participant reported it as “Intersex”, “I prefer not to answer”, “none of these fully describe me”, or skipped the question, we passed the sex concordance check for that sample, regardless of the

information from peddy and DRAGEN. The sex assigned at birth data from the CDR is described in Appendix F.

DRAGEN invocations include a wide breadth of functionality, including ploidy calls (see Appendix H for the parameters).

The DRAGEN pipeline outputs a single sample VCF, which is primarily used in the clinical pipeline (for individual samples)[5], but we use it for our call to peddy. We call peddy with the following parameters:

Parameter	Value
vcf	Single sample VCF from DRAGEN (hard-filtered)
Pedigree file	We create this file dynamically based on the single sample and its sex call. Please note: This implies that we do not use pedigree information in our peddy call.

Results

We do not include any srWGS samples that fail the sex concordance check in the released samples. It is important to note that some samples automatically passed this check solely because they did not answer “male” or “female” on the self-reported sex assigned at birth question (2.07% of srWGS samples). Appendix F has more details on this CDR question and the possible responses.

Cross-Individual Contamination Rate

For all srWGS samples, we estimate the proportion of data coming from an individual other than the one being processed, referred to as the contamination rate.

Method

We estimate the percent contamination from another individual by counting the number of reads at common homozygous alternate SNP sites. If there is a small amount of cross-individual contamination, we expect to see small numbers of reads supporting SNPs at these sites. We determine the percentage of the sample that may have come from a different individual using VerifyBamID2 [8], and the DRAGEN 3.4.12 pipeline. Contamination rate is a float value from 0.0 to 1.0, which represents 0 to 100%.

We use the following parameters for VerifyBamID2:

Parameter	Value
NumPC	“4”

BamFile	WGS cram file
Reference	“gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta”
UDPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.UD”
BedPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.bed”
MeanPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.mu”
Verbose	specified

Please see Appendix H for the DRAGEN command line parameters, as the command line contains multiple functions, including calculating contamination.

Results

The hard threshold for contamination was 0.03 for the research pipeline, higher than 0.01 for the clinical pipeline [5].

We did not include any samples with a contamination larger than 0.018 and only three samples greater than 0.015. Figure 5 demonstrates the frequency of the contamination estimates for samples in the v7 release.

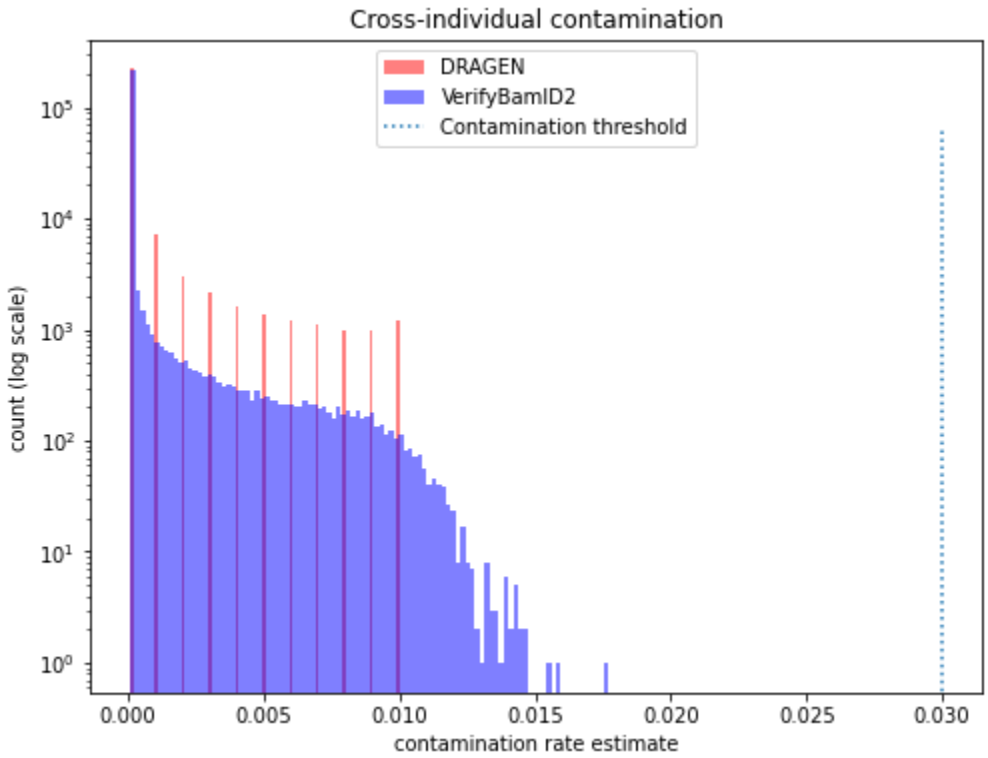


Figure 5 -- srWGS contamination estimates from both sources (DRAGEN and VerifyBamID2). DRAGEN rounds the contamination estimate to three decimal places. Note the log scale of the counts (y-axis). Over 89.0% and 91.4% of srWGS samples had contamination estimates lower than 1e-4 by VerifyBamID2 and DRAGEN, respectively.

Coverage

Method

Coverage is defined as the number of reads covering the bases of the genome. Maintaining coverage is important for consistent statistical power and accurate variant calling. We apply several thresholds (summarized from the FDA IDE (G200165)):

- Mean coverage (threshold ≥30x) - This is the mean number of overlapping reads at every targeted base of the genome. Accuracy steadily decreases as mean coverage decreases, with a rapid decrease below 20x coverage, supporting a stringent threshold selection of a minimum of 30x.
- Genome coverage (threshold ≥90% at 20x) - Accuracy steadily decreases as the percent of bases with at least 20x coverage drops. Drop-off of performance is initially gradual, supporting a threshold of 90%.
- All of Us Hereditary Disease Risk gene (AoUHDR) coverage (threshold ≥95% at 20x) - For clinically relevant areas of the genome, we insist on higher mean coverage to ensure a higher calling accuracy. As we reduce the coverage in the AoUHDR region, the reduction in performance is slow initially but increases rapidly below 40%, showing that the threshold of 95% is conservative.
- Aligned Q30 bases (threshold ≥8e10) - All bases in the sequencing reads get a quality assignment, which is phred scaled (Q30 → probability of error is 0.001) [9]. As lower base quality counts increase, we see a reduction in accuracy with an inflection point starting around 6e10.

Results

As seen in Figure 6, all srWGS samples exceed the thresholds that we set as part of the research pipeline. We had 281 (0.1%) samples with mean coverage greater than 70x.

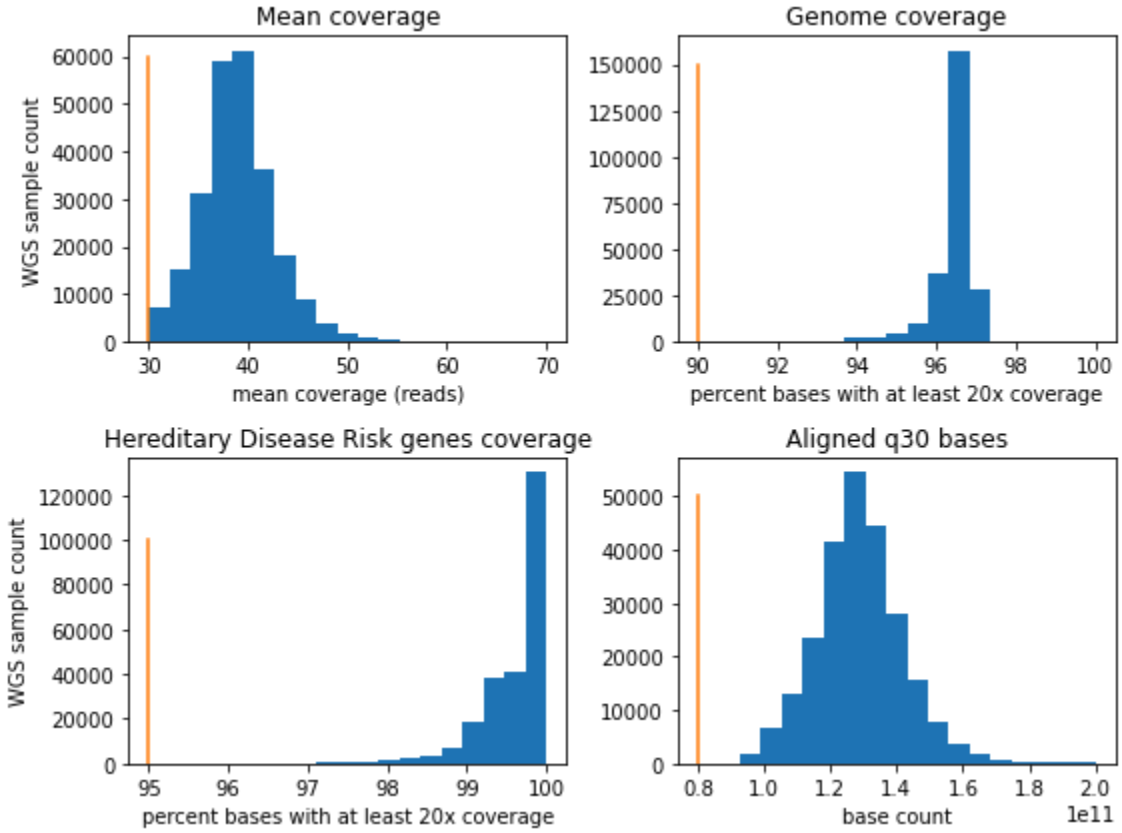


Figure 6 -- Coverage metrics for the v7 release srWGS samples. The orange line is the threshold for each metric. There are 281 samples (0.1%), with mean coverage greater than 70x, that are not included in the mean coverage (upper left) nor aligned q30 bases (lower right) plots. As expected, these samples were outliers in the number of aligned q30 bases (i.e., higher base count than samples with lower mean coverage).

Short-read WGS SNP & Indel Joint Callset QC

The srWGS small variants are delivered as a joint callset and the QC steps in this section are performed on the joint callset, not individual samples [10]. Please note that the QC steps described here apply during creation of the srWGS joint callset, after single sample QC. The joint callset is available in the RW and other auxiliary information (including lists of flagged samples) is available through the User Support Hub [1]. The joint callset QC process is similar to that of gnomAD 3.1 [11], though not exactly the same. See a summary of the joint callset QC steps in Table 3. Sample QC is performed before Variant QC.

We flag samples or variants as failing QC, rather than removing them from the callset, since we could not validate whether samples (especially population outliers) were problematic or were

just a part of a poorly-sampled ancestry. Flagged variants can also be a result of poorly-sampled ancestry.

Table 3 -- srWGS SNP & Indel joint callset QC summary

QC process	Sample or variant QC	Error modes addressed	v7 release results
Sample Hard Threshold Flag	sample	Extremely noisy samples	No samples flagged.
Sample Population Outlier Flag	sample	Noisy samples	551 samples flagged (0.2%). Based on regressing out the PCAs from callset metrics, such as snp_count.
Variant Hard Threshold Filters	variant	Artifacts that cannot be detected in a single sample	This has a simple implementation with high precision, which saves compute for downstream variant filtering. 59,496,403 were filtered 972,115,272 were not filtered
Allele-Specific VariantQualityScore Recalibration (AS-VQSR)	variant	Artifacts that cannot be detected in a single sample	See [12].
Sensitivity and Precision Evaluation	both	Poor variant detection	See Appendix I for a list of samples.
Auxiliary processes			
Ancestry	sample	Flagging sample outliers and allows calculation of population level metrics, such as allele frequency (AF).	Error rate from holdout set (incl. Other): 0.046 Error rate from holdout set (not incl. Other): 0.002 Concordance vs self-reported: 0.915 See Appendix A. Number of independent, bi-allelic sites ("high-quality sites") used: 151159 See Appendix J.
Relatedness and maximal independent set of samples	sample	Related samples, which confound analyses	19374 related pairs and 15376 samples in the maximal independent set. See Appendix K. This process produces a list of the sample pairs with kinship score, calculated by Hail [13]. No samples are removed from the callset, but this allows researchers to easily remove a minimal set of samples to eliminate related samples in the callset.

Sample Hard Threshold Flag

We flag srWGS individual samples based on these sample-level QC metrics. The flagged samples can be found in a list on the User Support Hub [1].

Method

We initially flagged any samples with strong erroneous signals. We calculated all metrics using autosomal territory only. The criteria for being eliminated as “obviously erroneous”:

- number of SNPs: < 2.4M and > 5.0M
- number of variants not present in gnomAD 3.1: > 100k
- heterozygous to homozygous ratio (SNPs and Indel separately): > 3.3

Results

We did not flag any samples for failing hard thresholds.

Sample Population Outlier Flag

We flag srWGS individual samples based on the population outlier data. The flagged samples can be found in a list on the User Support Hub [1].

Method

As part of ancestry prediction (see Appendix A), we regressed out sixteen principal component features computed and used the residuals to determine the outliers. We define outlier samples as being eight median absolute deviations (MADs) away from the median residual in any of the following metrics:

i. number of deletions
ii. number of insertions
iii. number of SNPs
iv. number of variants not present in gnomAD 3.1
v. insertion : deletion ratio
vi. transition : transversion (TiTv) ratio
vii. heterozygous to homozygous ratio (SNPs and Indel separately)

Results

We flagged 551 (0.2%) samples as outliers based on at least one of the above criteria (See Table 4). Plots of the first principal components against these eight metrics can be found in Appendix L.

Table 4 -- srWGS SNP & Indel population outlier sample counts

Metric(s) considered	Flagged sample count
Indel heterozygous to homozygous ratio	307
Deletion count + Indel heterozygous to homozygous ratio + Insertion count + SNP count	74
Indel heterozygous to homozygous ratio + SNP	48

heterozygous to homozygous ratio	
Indel heterozygous to homozygous ratio + SNP count	41
Variants not present in gnomAD 3.1 count	28
Deletion count + Indel heterozygous to homozygous ratio + SNP count	26
SNP heterozygous to homozygous ratio	17
Ti/Tv ratio + Variants not present in gnomAD 3.1 count	5
Indel heterozygous to homozygous ratio + Variants not present in gnomAD 3.1 count	3
Ins/del ratio	1
Indel heterozygous to homozygous ratio + SNP count + SNP heterozygous to homozygous ratio	1

Total	551

Variant Hard Threshold Filters

These site-level QC metrics for the srWGS SNP & Indel callset will flag variants, appearing as filtered in the site level filters of the VDS and VCF (`filters` in the VDS, `FILTER` in the VCF). These variants will still be included in cohorts, including in the Cohort builder.

Method

If a variant does not meet the following criteria, it will be filtered:

- No high-quality genotype (`GQ≥20`, `DP≥10`, and `AB≥0.2` for heterozygotes) called for the variant.
    - Allele Balance (AB) is calculated for each heterozygous variant as the number of bases supporting the least-represented allele over the total number of base observations. In other words, `min(AD)/DP` for diploid GTs.
    - Filter field value: `NO_HQ_GENOTYPES`
- `ExcessHet < 54.69`
    - `ExcessHet` is a phred-scaled p-value. We cutoff of anything more extreme than a z-score of -4.5 (p-value of 3.4e-06), which phred-scaled is 54.69
    - Filter field value: `ExcessHet`
- `QUAL` score is too low (lower than 60 for SNPs; lower than 69 for Indels)
    - QUAL tells you how confident we are that there is some kind of variation at a given site. The variation may be present in one or more samples.

- Filter field value: `LowQual`

Results

Unfiltered variants will have "." or `PASS` in the site level filters fields in the srWGS joint callset SNP & Indel VCFs, VDS, and Hail MTs. Filtered variants will have the filter name in the site level filters of the VCF, VDS, or Hail MT (`FILTER` or `filters`). We recommend that researchers do not include variant sites that were filtered in their analyses. The variant counts can be found in Table 5.

Table 5 -- srWGS SNP & Indel variant hard threshold filter counts

Filters	Numbers
None	972115272
'NO_HQ_GENOTYPES'	33526160
'NO_HQ_GENOTYPES', 'LowQual'	22245268
'LowQual'	3064830
'ExcessHet'	659051
'NO_HQ_GENOTYPES', 'ExcessHet'	1094

Allele-Specific Variant Quality Score Recalibration (AS-VQSR)

These genotype-level QC metrics for the srWGS SNP & Indel callset will flag variants. The AS-VQSR tool scores genotypes, at some sites only some genotypes are filtered whereas at other sites all genotypes are filtered. We do not report the AS-VQSR scores in the srWGS SNP & Indel callset, we only report whether or not a genotype or variant is filtered.

A filtered genotype will appear as filtered in the genotype level filter (`FT`) in the VCF, VDS, and Hail MT. In the VDS, `FT` will contain `True` for PASS and `False` for FAIL. In the VCF or Hail MT, `FT` will contain `PASS` or `FAIL`. If all genotypes fail the AS-VQSR filtering at a variant site, the site will be filtered in the VDS filter field (`filters`) or the VCF/Hail MT filter field (`FILTER`). All variants will still be included in cohorts, including in the Cohort builder.

Method

As part of the joint calling, we will filter variants with Allele-Specific Variant Quality Score Recalibration (AS-VQSR or VQSR) [12]. This filtering technique uses machine learning to identify variants across samples that are likely artifacts. We used the following annotations as features for training:

- Variant Confidence/Quality by Depth (`AS_QD`)
- Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities (`AS_MQRankSum`)

- Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias (AS_ReadPosRankSum)
- Phred-scaled p-value using Fisher's exact test to detect strand bias (AS_FS)he
- RMS Mapping Quality of reference vs alt reads (AS_MQ) [SNPs only]
- Symmetric Odds Ratio of 2x2 contingency table to detect strand bias (AS_SOR)

We used the default training sets as described in the GATK documentation [14] and Table 6. Training sets are flagged as true or training sites and assigned an initial prior likelihood score. Details of these parameters can be found in the GATK documentation [12,14], and the sites can be found as public resource downloads for the GATK [15].

Table 6 – srWGS SNP and Indel AS-VQSR training and truth datasets

Training Set Name	SNP or Indel	Truth	Training	Prior Likelihood	Description
Omni [16]	SNP	True	True	Q12 (93.69%)	This resource is a set of polymorphic SNP sites produced by the Omni genotyping array.
HapMap [17]	SNP	True	True	Q15 (96.84%)	This resource is a SNP callset that has been validated to a very high degree of confidence.
1000 Genomes [18]	SNP	False	True	Q10 (90%)	This resource is a set of high-confidence SNP sites produced by the 1000 Genomes Project.
Mills [19]	Indel	True	True	Q12 (93.69%)	This resource is an Indel callset that has been validated to a high degree of confidence.
Axiom [18]	Indel	False	True	Q10 (90%)	This resource is an Indel callset based on the Affymetrix Axiom array on 1000 Genomes Project samples.

Sensitivity and Precision Evaluation

Method

In the callset, we included four well-characterized control samples (four Genomes-in-a-Bottle samples (GiaB) [20] from HapMap [17] and Personal Genome Project; see Appendix I), which we can use to determine sensitivity and precision. The samples were sequenced with the same protocol as the *All of Us* samples. These samples do not appear in any user data (e.g., cohorts built using the RW).

We use the high confidence calling region, defined by GiaB v4.2.1, as the source of ground truth. In order to be called a true positive, a variant must match the chromosome, position,

reference allele, and alternate allele. In cases of sites with multiple alternate alleles, each alternate allele is considered separately.

Results

Sensitivity and precision results can be seen in Table 7.

Table 7 -- Sensitivity and precision measurements for control samples using the *All of Us* sequencing protocol

Variant type	Sample	Sensitivity	Precision
SNV	HG-001	0.995	>0.999
	HG-003	0.988	>0.999
	HG-004	0.988	>0.999
	HG-005	0.989	>0.999
Indel	HG-001	0.987	0.996
	HG-003	0.985	0.997
	HG-004	0.986	0.998
	HG-005	0.994	0.999

Short Read Structural Variants (srWGS SVs)

Short read structural variant calling was performed on 11,390 srWGS samples. The samples are a subset of the v7 srWGS samples with SNP and Indel variant calls. All srWGS samples followed the Consistency across Genome Centers and Single Sample QC processes. See those sections for an overview of the CRAM-level QC processes before structural variant calling and analysis. We used GATK-SV to call structural variants, which has been previously described [21]. Further technical information can be found in Appendix M.

GATK-SV discovers structural variants (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 [22] for more information on SV types and their evidence signatures.

Sample Selection for srWGS SVs

We initially selected 12,000 samples for SV calling from the v6 srWGS release and v7 lrWGS samples. We generated the participant list for srWGS SV calling before the final v7 srWGS SNP and Indel list was created. We did not expect to release all 12,000 selected samples due to stricter QC criteria for srWGS SV calling and dropping samples that were in the srWGS v6 sample set but not in v7 (e.g., participant withdrew) (Table 8). A total of 570 v7 srWGS SNP and Indel samples failed the QC criteria of the srWGS SV callset.

Among the 12,000 selected samples, 1,010 samples were selected because they were in the v7 lrWGS cohort. Among these lrWGS samples, 989 passed all SV QC and made it into the v7 srWGS SV callset. An additional 1,253 samples were included because they were selected for future long read sequencing. The remaining 9,737 samples of the 12,000 were randomly selected from the v6 srWGS data (C2022Q2R2).

Table 8 -- Selected samples that were excluded from v7 srWGS SV calling

srWGS SV sample exclusion steps	Number of samples filtered from initial count (N=12000)	Notes
Single sample QC	561	See Table 9 and Table 10
Joint SV callset refinement and QC	9	Only the Outlier Removal step filters samples.
Other	40	These are v6 srWGS samples that were not included in v7 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 9 and Table 10, on all 12,000 selected samples. We removed a total of 561 samples during srWGS SV single sample QC, which left 11,439 samples remaining in the callset for downstream processing.

Basic Filters

Method

As seen in Table 9:

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 0.5%.
2. We checked the mean insert size of each srWGS sample using the Picard tool CollectInsertSizeMetrics and removed five samples that were outside of the range 370-700.
3. We checked the whole genome dosage (WGD) [21] 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 [23], Wham [24], and MELT [25] 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, and samples that were outliers in 30 or more categories were removed from the callset.
6. We dropped any samples that could not successfully complete our workflows.

We removed all samples that failed any of these filters, counted in Table 9. Note that some samples failed multiple filters.

Results

The results for all six basic single-sample filtering steps are summarized in Table 9. Three samples had an issue with the DRAGEN 3.4.12 pipeline, which the GCs used to generate all v7 srWGS data. This rare issue caused incorrect formatting in gVCF fields that are required by the srWGS SV evidence collection steps. As a result, our pipelines reported an error for these samples. We did not believe that the effort to recover the samples was commensurate with the gain in sample count (0.03%) for this release. This issue has been fixed in subsequent versions of the DRAGEN pipeline (e.g., 3.7.8).

Table 9 -- srWGS SV single sample QC: Basic filters

QC process	Passing criteria	Error modes addressed	Number of samples removed	Notes
Cross-individual contamination	≤ 0.005 (≤ 0.5%)	Sample contamination from another individual	316	Same method as srWGS SNP and Indel QC, see Cross-individual contamination rate
Mean insert size	Mean insert size in range [320, 700]	Insert size outliers, which could skew distributions of discordant pairs	5	Picard's CollectInsertSizeMetrics within GATK's CollectMultipleMetrics
Whole genome dosage (WGD)	WGD within 6*MAD of the median, approx. [-0.159, 0.131]	Samples with high variability in coverage across the genome, which could lead to unreliable CNV calling from depth evidence	175	Method can be found in [21]
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	205	
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	11	
Processing failure	Sample must complete processing	Upstream issues from srWGS SNV calling that affect the srWGS SV deliverables	3	All of these failures were due to a (rare) issue in the DRAGEN 3.4.12 pipeline where records in a gVCF can be improperly formatted. This issue has been addressed in subsequent versions of the DRAGEN pipeline.

Ploidy estimation

Method

We estimated ploidy per chromosome across all 12,000 samples by binning read counts in 1Mb intervals and normalizing by half the genome-wide median. We only performed filtering based on ploidy (Table 10) on samples that passed the basic filters (Table 9).

Results

We filtered less than 20 samples because they had an estimated copy number greater than 2.3 on one autosomal chromosome. Plots of binned coverage across these chromosomes demonstrated that these samples appeared to have mosaic autosomal aneuploidies. We do not provide the exact count of samples in this document to comply with *All of Us* policy [26]. If you need the exact count and/or the specific samples, we provide the list of samples with mosaic aneuploidy on autosomes in the Controlled Tier; for more details see the 'Controlled CDR directory document' on the User Support Hub [1].

Sex Concordance

Method

Using the ploidy estimation process detailed above, we estimated the ploidy for allosomes to infer sex for each of the 12,000 samples. 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 with males or females as described below. 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 for more details.

Results

All samples passed this check, indicating no sample swaps based on the computed sex.

We observed likely mosaic loss of chrX and chrY in 100 samples; these samples had an estimated copy number of 0.1-0.8 on chrY and 1.5-1.8 on chrX. These samples 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. These samples were retained in the callset and grouped with males (if chrX rounded ploidy = 1 and chrY ploidy > 0.1) and females (if chrX rounded ploidy = 2) for the sex-specific steps of the GATK-SV pipeline. A list with the 100 samples is available; for more details see the 'Controlled CDR directory document' on the User Support Hub [1].

We found less than 20 samples with predicted germline sex chromosome aneuploidies (i.e. computed sex ploidy other than XX, XY, or mosaic) and a list of these samples is available; for more details see the 'Controlled CDR directory document' on the User Support Hub [1]. 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.

Table 10 -- 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)	≤ 2.3	Samples with mosaic autosomal aneuploidies, which could skew distributions of SV evidence classes	≤ 20	Calculated after applying all above filters. Method can be found in [21]
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 samples into 24 batches with an average of 477 samples in each batch for the batched analysis steps of the GATK-SV pipeline, depicted in Figure 8. Batching controls for technical variability between samples and parallelizes computation. The batching procedure was as follows:

1. Split by chrX ploidy (<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 whole genome dosage (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 [21], 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 (Figure 7). The multimodal distribution of mean insert size across samples was investigated and found to stem from differences between Chemagen and Autogen extraction protocols. Differences in insert size could impact the distribution of discordant pair counts across samples during genotyping, so samples with similar mean insert sizes were batched together to improve genotyping accuracy.

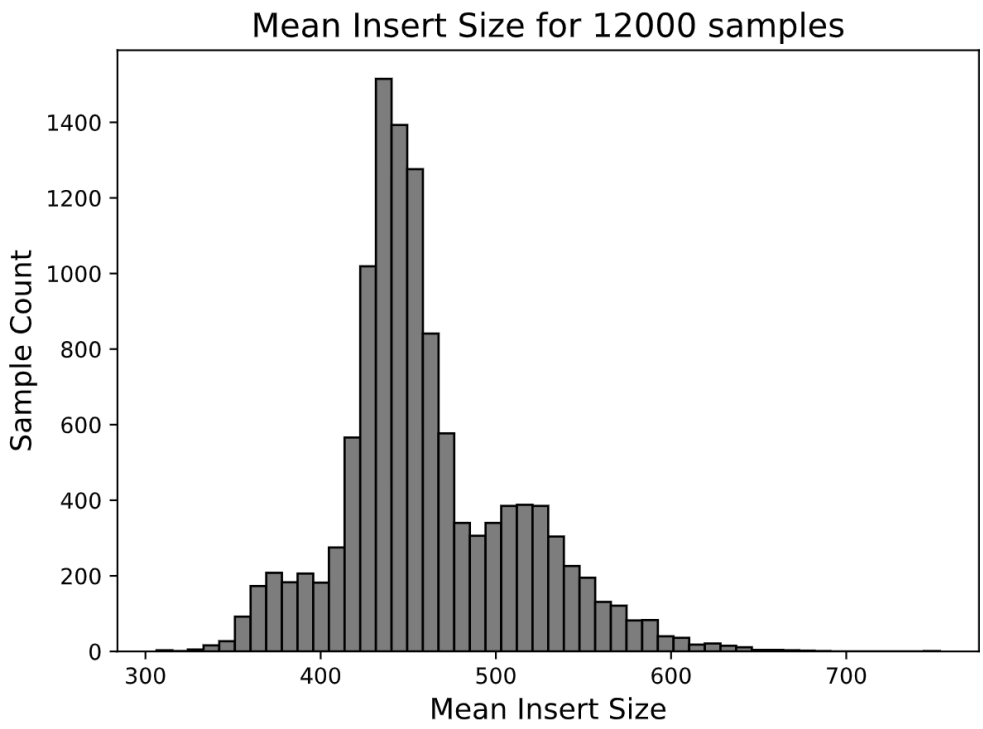


Figure 7 -- Mean insert size across the 12,000 srWGS samples initially selected for SV calling.

Joint Callset Refinement and QC for srWGS SVs

The steps to generate the GATK-SV joint callset, including refinement and filtering, are described in Figure 8 and Appendix M. Below, we describe refinement and filtering steps introduced in the v7 srWGS release (blue steps in Figure 8) that were not published previously or are modifications to canonical GATK-SV pipelines. These steps include both hard and soft filters at the sample, site, and genotype level (Table 11).

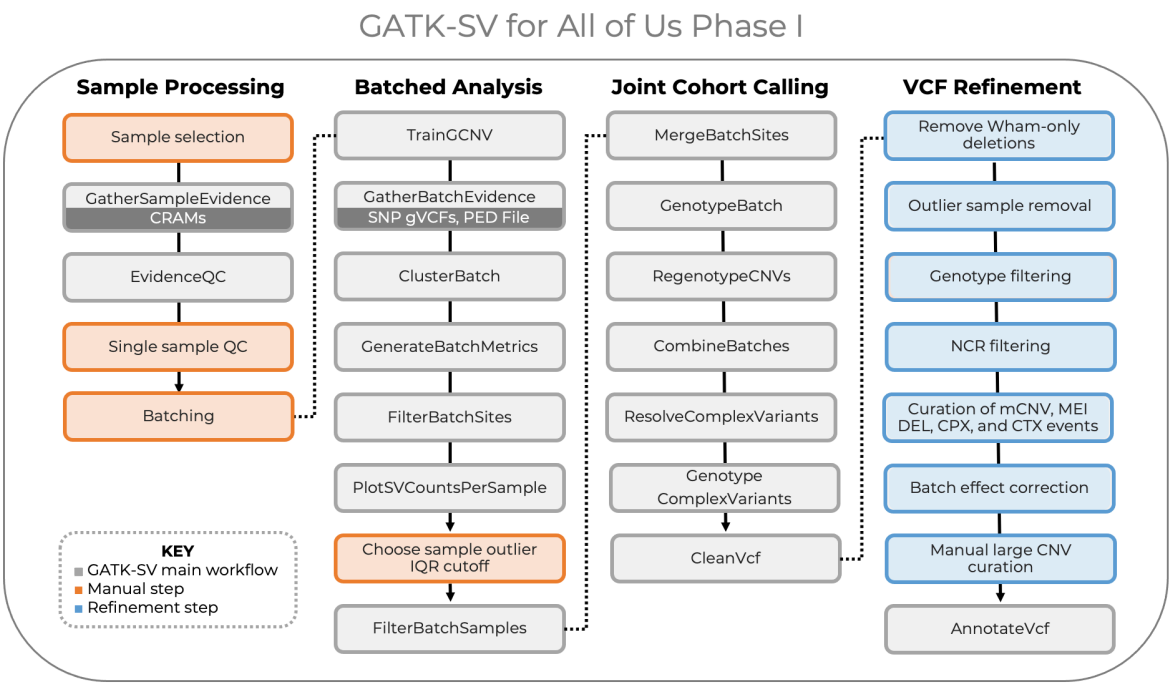


Figure 8 -- GATK-SV Pipeline Schematic. GATK-SV automated workflows are shown in gray and the names correspond to the name of the Workflow Definition Language (WDL) file. Manual steps performed in notebooks or requiring user input are shown in orange. Steps in blue are custom VCF refinement and QC steps for All of Us.

Table 11 -- 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 unique Wham deletions	Variant		False positive deletions	Variants removed from callset
Outlier removal	Sample		Noisy samples	Samples removed from callset
Genotype filter	Genotype		False positive genotypes for INS, INV, DEL, and DUP	Filtered genotypes are set to no-call (`./.`)
No-call rate (NCR)	Variant	HIGH_NCR	INS, INV, DEL, or DUP sites that have many false positives and are likely to be technical artifacts or difficult to genotype	

Batch Effect Correction	Variant	VARIABLE_ACROSS_BATCHES	Technical artifacts from batch effects	
Multiallelic CNVs	Variant		False positive MCNVs	Multiallelic CNVs <5 kilobases (kb) in length were removed from the callset.
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
CPX and CTX precision improvement	Genotype		False positive CTX and CPX	Filtered genotypes are set to no call (`./.`)
Manual curation of large CNVs	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

Remove unique Wham deletions

We used the lrWGS validation tool VaPoR [27] to assess the evidence for srWGS SV calls in matched lrWGS data. During prior analyses such as in the 1000 Genomes Project [37], we discovered very high false-positive rates for deletions that were uniquely contributed by the Wham algorithm [24], one of the SV calling algorithms used by GATK-SV. As expected, we again observed here that 97% of non-reference genotypes in deletion sites uniquely discovered by Wham (13,557,361 out of 13,915,316) were not supported by VaPoR genotypes from lrWGS and 95% of unique deletions in this category (214,675 out of 225,547) had no non-reference genotypes supported by VaPoR. We applied a hard filter for these variants (i.e. these variants will not appear in the v7 srWGS SV callset).

Outlier Removal

We calculated the distribution of SV counts across all samples stratified by SV type and observed a subset of samples (n = 9) that carried significantly more SVs in a single SV class than the rest of the cohort. These samples were likely to have a higher false positive rate and were removed from the v7 srWGS SV callset. We defined outlier samples as the subset that have SV counts greater than $Q_3 + 1.5 * IQR$ (where Q_3 is the third quartile and IQR is the interquartile range) for any SV type. Among the 9 outlier samples identified, 6 were outliers for duplication counts and 3 were outliers for deletion counts.

Genotype Filter (SL Filter)

We filtered genotypes of bi-allelic SVs using a machine learning model, described below, to reduce the number of false positive INS, INV, DEL, and DUP while minimizing loss of sensitivity.

Method

Training data

We selected true positive and false positive training sites for the machine learning model based on comparisons against long read data or genotyping arrays, depending on length and type of the SV. 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. Genotyping arrays provide a good source of orthogonal data to confirm large CNVs (>10kb) but are too sparse to be sensitive to smaller CNVs and cannot be used to detect other SV types. Therefore, we trained INS, INV, DEL <5kb, and DUP <5kb on lrWGS data and we trained DEL >10kb and DUP >10kb on SNP arrays. There was a gap in the training data for CNVs between 5kb and 10kb in size; these CNVs were grouped with large CNVs >10kb for filtering because both size ranges rely on read depth evidence, so their genotyping error modes are expected to be similar.

lrWGS training data

A subset of 606 samples with matched lrWGS data were selected for model training, and an additional 67 were held out as a test set to validate the model (the remaining 316 samples with matched long read data had not completed long read sequencing and QC at the time). 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 using VaPoR and overlap with SV calls from lrWGS data from the tools PAV [28], PBSV [29], and sniffles [30]. The GATK tool SVCluster was used to compute overlap between SV calls from srWGS and lrWGS [10].

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 (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 9A).

Genotyping array training data

Using array data, we evaluated deletions and duplications of at least 10 kb on the autosomes with the Genome STRiP IntensityRankSumAnnotator (IRS) [31,32]. The IRS tool compares the array probe intensity values between samples predicted to carry the CNV and those predicted to be non-carriers (according to genotypes in the SV VCF), using all probes that are within the CNV interval. Using a non-parametric test, the IRS tool assigns a p-value to each CNV which indicates if the CNV genotypes are supported by the intensity data.

We ran the IRS test on batches of 500 samples. Prior to running the IRS test for each batch, we dropped any probes for which samples in the batch had missing data at that probe. We examined the IRS test results for all CNVs of at least 10 kb for which there were at least 5 overlapping probes tested in the batch. Sites with an IRS p-value (IRS_PVALUE) less than 1e-6 were chosen as positive training sites for the carrier samples (as determined by the genotypes in the SV VCF). Sites with an IRS p-value greater than 0.2 were chosen as negative training sites for the carrier samples.

Filtering model

We employed a method for re-calculating SV genotype quality to reduce false positive variants. This filter tool (XGBoostMinGqVariantFilter) is implemented in GATK [33]. The filter applies a decision tree from the XGBoost library for gradient boosted machine learning to predict the quality of a given genotype [34].

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 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 (GQ) 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.

The minimum SL scores required for each genotype to pass the model were selected with the goal to maximize F1 scores. Failed genotypes were revised to no-call (./.). Homozygous reference genotypes were also filtered by inverting the SL score and applying the same cutoffs.

Results

Analysis of the training samples from lrWGS and genotyping arrays yielded a total of 664,997 trainable genotypes, while labels for 1,006,729 genotypes (60% of the total) could not be determined (Figure 9 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 9 B). The model also predicted that the majority of unlabeled variants were of poor quality, likely reflecting the low signal-to-noise ratio of many SVs called from srWGS.

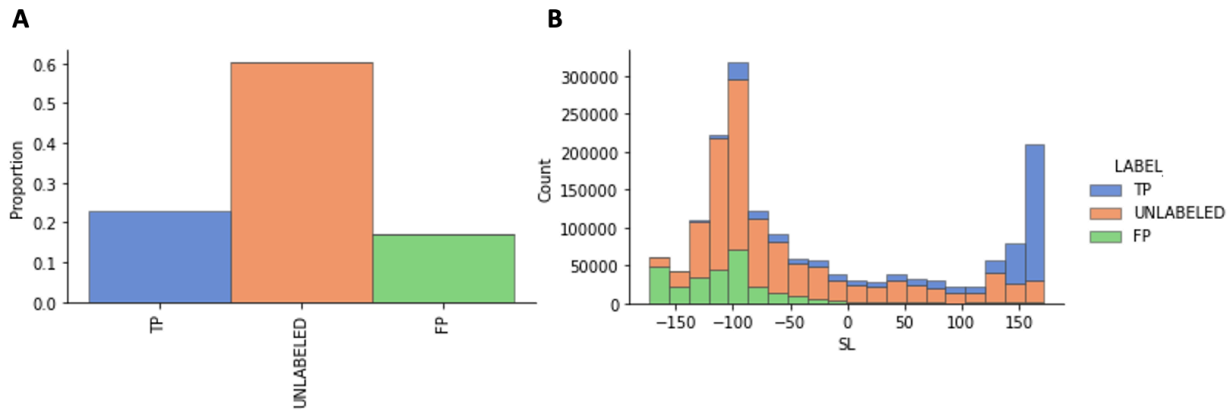


Figure 9 -- Training data for genotype filtering. (A) The proportion of each training label out of all SV genotypes in the training data, and (B) the SL score distribution produced by the trained model.

The genotype filtering performance was evaluated in the test set of 67 held-out samples with matched lrWGS data. Figure 10 shows that precision decreases consistently as a function of recall when thresholding on SL. This demonstrates that the method is effective for tuning callset accuracy. These results also indicate comparable performance across the spectrum of SV classes, with the exception of lower performance for medium (0.5-10 kb) duplications. Optimal

cutoffs for SL filtering were determined using the training set as described above and are shown in Appendix Table N.1.

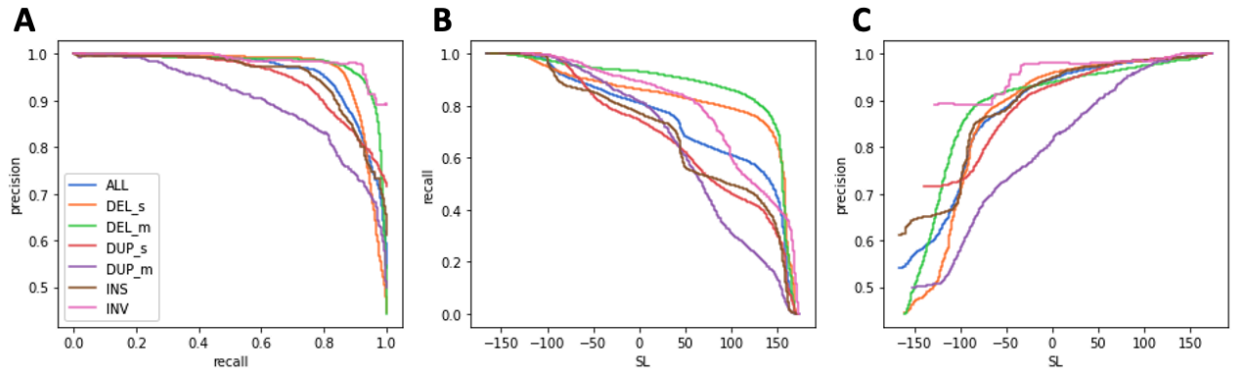


Figure 10 -- SL genotype filtering performance assessed against 67 lrWGS labeled test samples. (A) Precision-recall curves for all filtering classes, (B) recall as a function of the SL cutoff value, and (C) precision as a function of the SL cutoff value.

We report the performance of the SL genotype filter combined with the No-call rate filtering (see below) in Appendix N.

No-call rate (NCR) Filtering

To further refine the SV sites, we also filtered on the no-call rate (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 sites with SV types CPX, CTX, and MCNV, which were not included in this filtering process. A filter status of "HIGH_NCR" was applied to every variant exceeding an NCR cutoff of 7%. This cutoff of 7% was chosen so as to remove noisy sites while preserving as many non-reference genotypes as possible across samples (Figure 11). To calculate the inflection point for proportion of non-reference genotypes retained vs. NCR cutoff value, we minimized the following quantity:

$$\beta^2 r^2 + (1 - v)^2,$$

where r is the NCR, v is the proportion of retained variants, and β=1.5 is a weighting parameter. With the chosen cutoff of r=0.07, v=85% of non-reference genotypes across were retained.

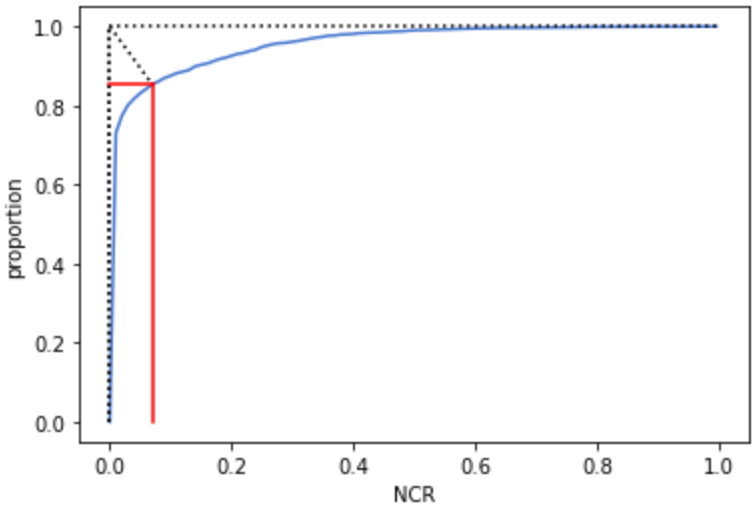


Figure 11 -- Proportion of retained variants against the NCR filter cutoff in the training set. Red lines indicate the chosen NCR cutoff point, where 85% of non-reference genotypes are retained.

We report the performance of the NCR filtering combined with the SL genotype filter (see above) in Appendix N.

Batch Effect Correction

To correct for batch effects among the 24 batches used for the batched steps of the GATK-SV pipeline, each variant was evaluated for batch effects. 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 [21]. Please note that PCR-amplified samples are not part of the AoU cohort, so we applied only the pairwise and one-vs-all comparisons described in Collins et al.

Multiallelic CNVs

Read depth signal is less reliable in events smaller than 5 kb [35]. We removed 1,051 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".

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 that overlap annotated mobile elements by greater than 50% and have SVLEN > 5kb, STRANDS = +-, and PE evidence were changed

back to SV type DEL. 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. This method added back 883 LINE1 deletions and 74 HERVK deletions in the srWGS SV callset.

Complex SVs and complex inter-chromosomal translocations

Specific alignment patterns and discordant paired end reads (PE) are expected for complex (CPX) and translocation (CTX) SVs [21]. 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 and GATK-gCNV).

Over half of the non-reference CPX genotypes (53.1%) had the expected PE evidence across all breakpoints, 24.8% had PE support only for some breakpoints, and 22.1% lacked PE support for all breakpoints. 50.7% of samples with non-reference genotypes for CPX events that involve large CNVs had overlapping CNVs in the raw depth callers. When requiring both PE evidence for all breakpoints and RD evidence when applicable, 48.2% of CPX genotypes failed this assessment and the genotypes were revised to no-call (./.).

Out of 26 CTX events, 8 failed our filters: 3 were carried at an allele frequency >1% and 5 lacked sufficient PE evidence for all non-reference genotypes. Failed genotypes were revised to no-call (./.) and failed sites were flagged with the filter status "UNRESOLVED".

Manual Curation of Large CNVs

We performed a visual inspection of read depth across all 225 CNVs (deletions and duplications) larger than 1 Mb observed in our final VCF using a visualization tool found in GATK-SV [36]. After inspection, we confirmed the presence of 219 CNVs (97.3%). We determined that the six CNVs (2.7%) that were not confirmed were all caused by a bug in the GATK-SV genotyping module, which has now been corrected and will not affect any version of GATK-SV from v0.27-beta. We also found that 12 (5.3%) of the CNVs larger than 1MB appeared to have multiple copy states, so we applied the multiallelic filter tag (MULTIALLELIC). Finally, for 35 CNVs (15.6%) 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.

Structural Variant QC Results

Below we detail several metrics of interest for this SV callset. 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. Figure 12 shows the SV counts, stratified by SV type, within the callset. Figure 13 shows the distribution of SV counts per genome, stratified by SV type, in the full cohort and for different predicted ancestries. Figure 14 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 elements. Figure 15 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 v7 short read structural variant calls,” available in the User Support Hub [1].

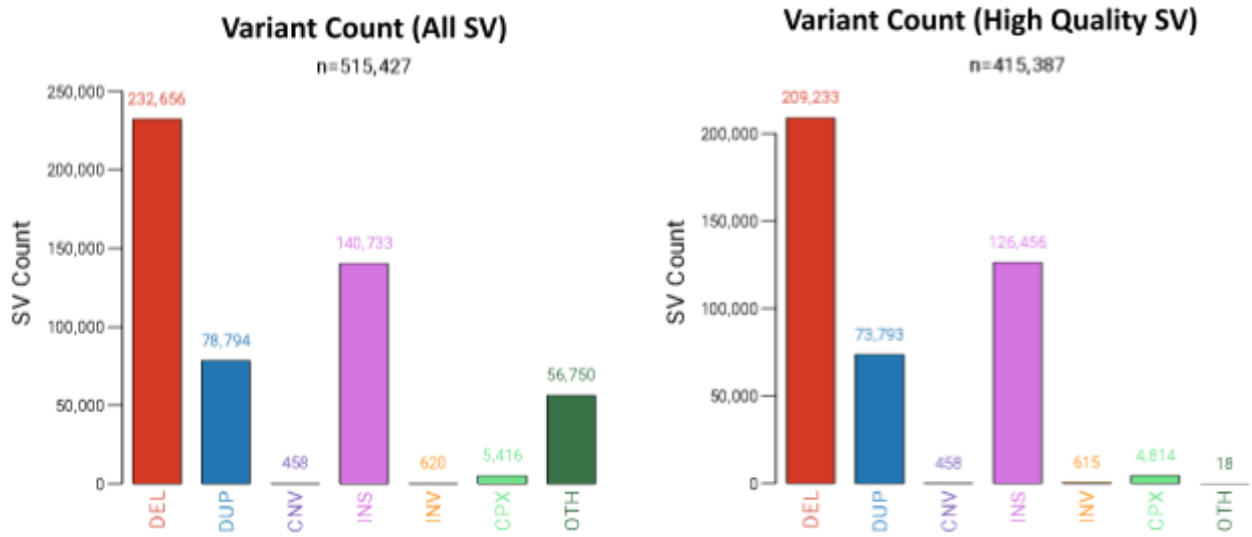


Figure 12 – We observed 515,427 total SVs of which we determined 415,387 to be of high quality. These counts are consistent with previous studies of similar sample size including gnomAD V2 [21].

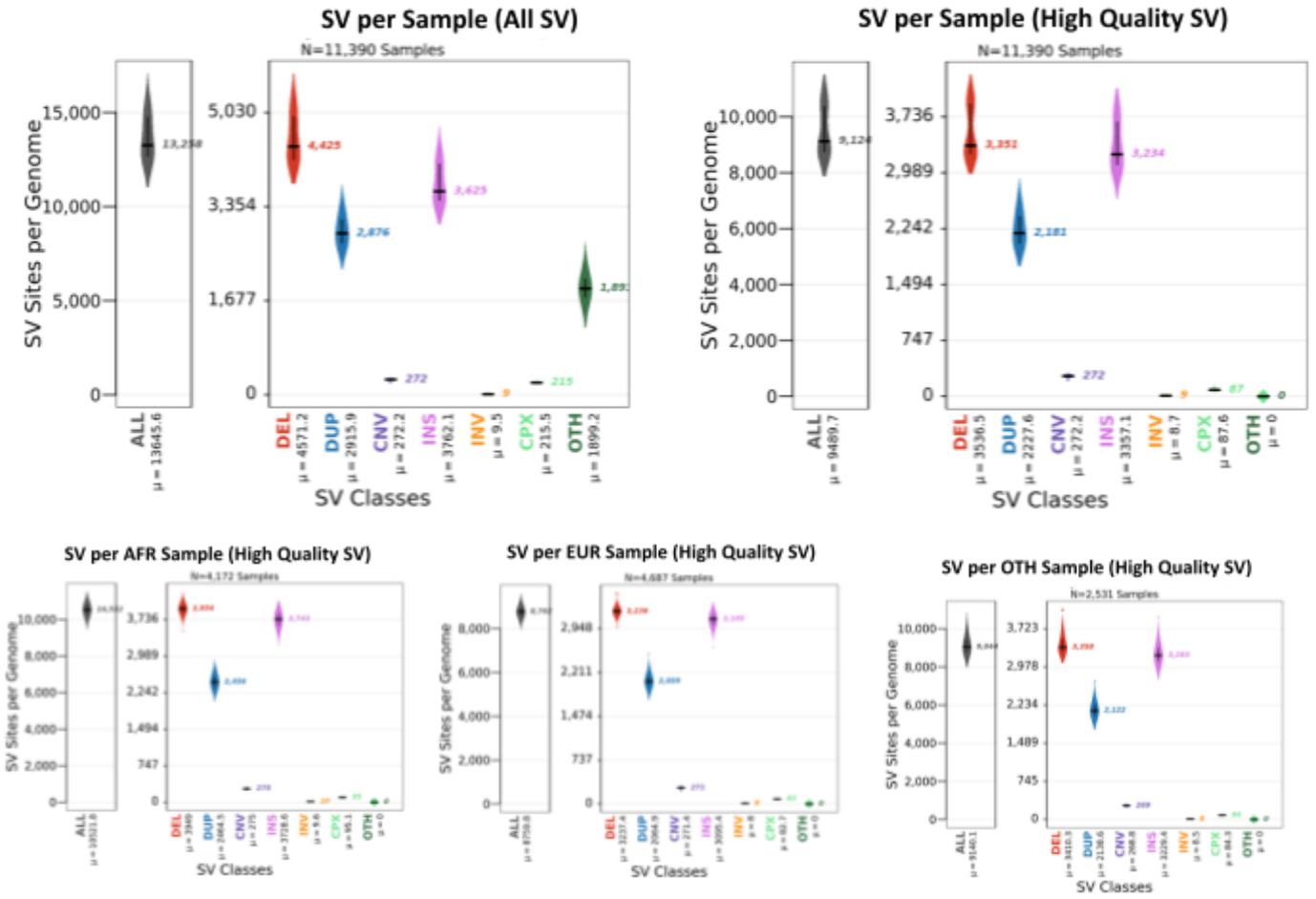


Figure 13 – We observed approximately 9k high quality SVs per person, which is consistent with SVs recently generated on the 1000 Genomes Project samples [37]. As expected, samples with predicted African ancestry had the highest SV counts while those with European ancestry had the lowest. Non-African and non-European samples were grouped together for this figure due to lower sample counts. (AFR=predicted African ancestry, EUR=predicted European ancestry, OTH=predicted Non-African and non-European ancestry)

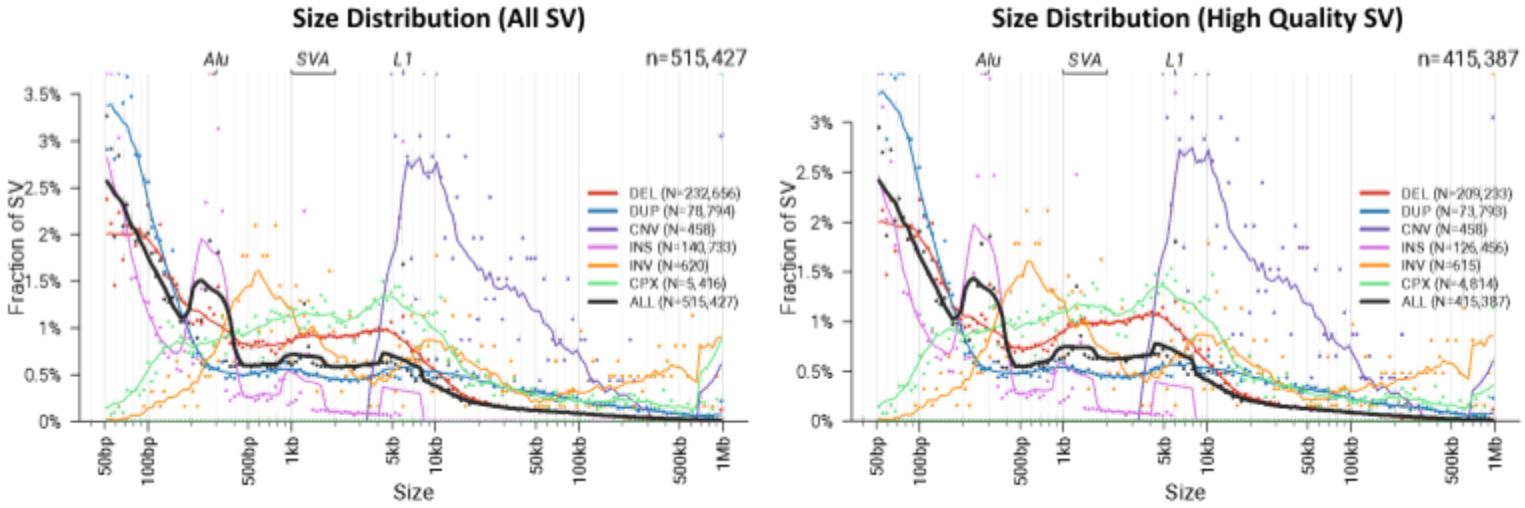


Figure 14 – SV size distribution matches previous expectations with notable insertion peaks corresponding to ALU, SVA, and LINE-1 insertions.

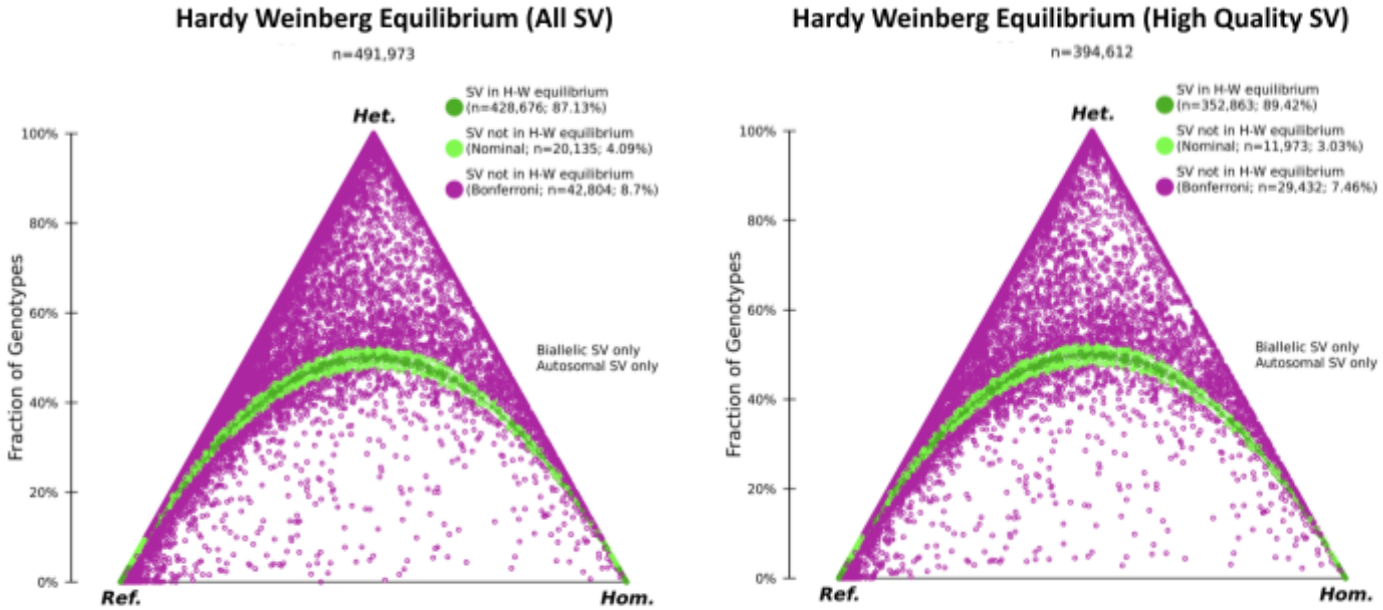


Figure 15 – Among high quality variants, only 7.5% fail Hardy Weinberg Equilibrium (HWE). Most of these failures appear to be driven by a bias towards genotyping variants as heterozygous. For the high quality calculation we included only the 11,306 unrelated samples.

Long Reads

There are 1,027 long read sequencing (lrWGS) samples in the v7 genomic data release. We applied a consistent sequencing protocol and performed QC checks for all samples. We also did quality checks on the joint SNP and Indel callset using information across samples. SNP and Indel variants and SVs are available for long read sequences in VCF format. In addition, long read sequences aligned to the grch38_noalt and T2Tv2.0 references are available in BAM format [38]. De novo assemblies are available in both FASTA format and graphical fragment assemblies (GFA) format. The data is described in the 'How the *All of Us* Genomic data are organized' article on the User Support Hub [1].

Data generation for lrWGS

Only one sequencing center generated the long read whole genome sequencing (lrWGS) data with a single technology, so we did not perform any cross-sequencing center consistency checks.

Samples were selected for lrWGS sequencing if they had srWGS and Array data in this v7 release, are self-reported African American participants, and are unrelated. Please see Appendix B for the self-reported ancestry report for participants with long reads data.

HudsonAlpha was the single sequencing facility ("sequencing center") commissioned for performing long reads sequencing for this cohort. The sequencing center used the Single Molecule Real Time (SMRT) sequencing technology [39] from Pacific Biosciences (PacBio). A SMRT cell is a chip conceptually similar to a flowcell in short read sequencing. Each SMRT cell is made of millions of zero-mode waveguides, where the modules are trapped and sequenced. Each molecule is scanned multiple times. The scanned data goes through consensus-processing (Circular Consensus Sequencing or CCS), which combines multiple lower-accuracy readings of a single molecule to produce a high accuracy consensus sequence. The sequencing instrument outputs the CCS-processed reads for each SMRT cell as an unaligned BAM (uBAM). Please refer to the PacBio glossary [40] for a more thorough definition of terms.

The uBAM files containing data for one SMRT cell are then delivered to the DRC along with metadata (e.g. which samples are sequenced with this SMRT cell, barcodes used if applicable, etc). The DRC extracts the CCS reads with a quality score greater than Q20 (HiFi reads) from the uBAM.

Single Sample QC for lrWGS

Samples can be sequenced across multiple SMRT cells, which are then demultiplexed and aggregated into BAMfiles for each sample. For each sample, we run QC on each individual demultiplexed SMRT cell and the aggregated sample bam file (Table 12). Three of the QC processes are performed on both the demultiplexed SMRT cell data and the aggregated sample data. As expected, no aggregated samples failed these three QC checks.

The HiFi reads of the demultiplexed SMRT cells are aligned to two references: “grch38_noalt” and “T2Tv2.0”. grch38_noalt corresponds to the GRCh38 reference with no alternate sequences [41,42]. T2Tv2.0 corresponds to the T2T-CHM13v2.0 reference, with the EBV contig added from the grch28_noalt reference [43]. We check the grch38_noalt BAM files before downstream processing, described in Table 12. In total, the DRC performed QC on 2,597 demultiplexed SMRT cell samples for a total of 1,027 aggregated samples.

Running three QC steps at both the demultiplexed SMRT cell level and aggregated sample level allows us to identify errors early in the process and to find quality issues at both levels. Please note that we do not use cross-sample information in the Single Sample QC process. Appendix O provides an overview of the processing steps for the lrWGS data, including processes after QC, such as variant calling and assemblies.

Table 12 -- QC processes performed on single sample data in each SMRT cell and after aggregation

QC Process	Aggregated sample or demultiplexed SMRT cell?	Passing criteria	Error modes addressed	V7 release results
Fingerprint concordance	Both	Log-likelihood ratio > 6	- Sample swaps - Large amounts of cross-individual contamination	All lrWGS samples are concordant with array samples.
Sex concordance	Both	Sex call is concordant with self-reported sex at birth. OR Self-reported sex at birth reported as “Other” or was not reported	- Sample swaps	All lrWGS samples are concordant. *Other refers to a participant self-reporting “Intersex”, “I prefer not to answer”, or “none of these fully describe me”
Cross-individual contamination	Both	< 0.03 (<3%)	- Sample contamination from another individual	All lrWGS samples meet the threshold (Aggregated samples plotted in Figure 17).
Coverage	Aggregated sample	≥ 5x mean coverage	- Sample preparation errors - Poor sensitivity and precision of variant calling	All lrWGS samples passed this check.
Read length median	Aggregated sample	≥ 10,000 bp	- Shorter fragments significantly impacting variant calling performance	All lrWGS samples passed this check.
Outlier sample	Aggregated sample	Manual threshold	-Poor variant calling	Five total samples

filtering			performance	were outliers due to variant counts and are not included in the v7 lrWGS data release (See Figure 20).

Fingerprint Concordance

Method

Each grch38_noalt BAM is checked against a fingerprint VCF to verify their marked identity from the sequencing metadata. This is applied to both the SMRT cell demultiplexed reads for the sample and the aggregated sample reads. We use the same fingerprint VCFs that are used by the srWGS fingerprint verification pipeline. Please refer to the Fingerprint Concordance method of the srWGS SNP & Indel QC process, as we follow the same method for lrWGS data, with some engineering adaptations that do not change the algorithm. The HAPLOTYPE_MAP file for lrWGS fingerprint concordance can be found at "gs://gcp-public-data--broad-references/hg38_noalt/v0/aou/fp/lr.aou.fp.haplotype_database.no_alt.txt", which differs from the srWGS HAPLOTYPE_MAP file only in the header section.

If a BAM failed fingerprint concordance, we launched a pipeline to find its identity by exhaustively testing to find a match against the other lrWGS fingerprint VCFs, whether planned or already sequenced. The true identity is the identity of the fingerprint VCF that returns the highest LOD that is above 6.0. We did not encounter any cases where a BAM file matched more than one other sample. If no matching VCF was found, then the lrWGS data was not released. The DRC also returns findings about swaps and their computationally-resolved identities back to the sequencing center for cross verification.

Results

All lrWGS samples (including all corresponding demultiplexed SMRT cells) in the v7 release passed the fingerprint concordance check based on the corresponding array VCF (Figure 16).

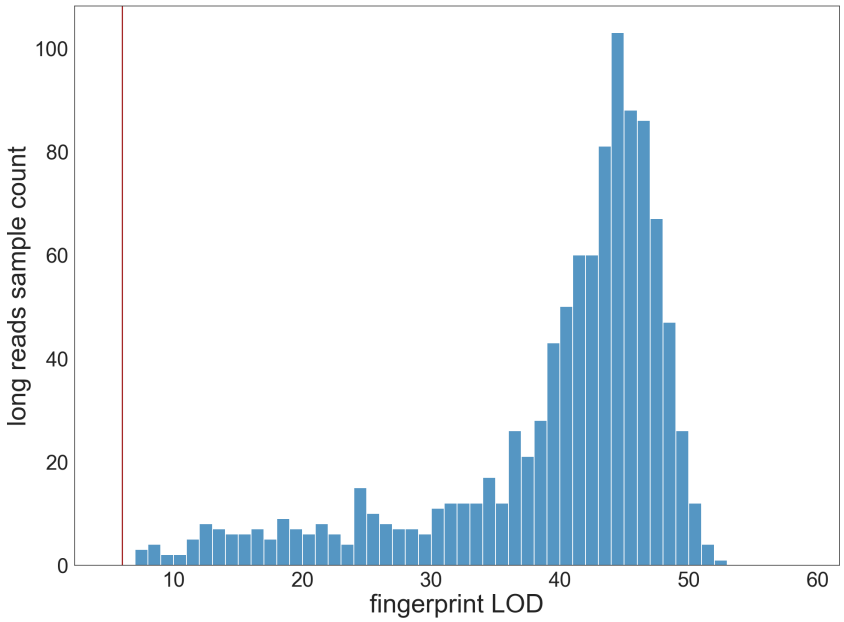


Figure 16 -- Fingerprint LOD histogram of the v7 lrWGS samples. All lrWGS v7 samples exceeded the fingerprint LOD threshold of 6.0.

Sex Concordance

Method

A simple sex concordance check is also performed on the grch38_noalt version of each BAM. The tool `mosdepth` (v0.3.3) [44] is run to calculate coverage across the whole genome and over each chromosome. A formula is then used for inferring the sex ploidy of the sample

```
Ploidy_x = round( 2 * cov(chrX) / cov(chr1) )
Ploidy_y = round( 2 * cov(chrY) / cov(chr1) )
```

The self-reported sex assigned at birth from the CDR data for each sample is then checked for contradictions with the inferred sex chromosome ploidies. If we do not have a “male” or “female” for the sex assigned at birth for the sample from the CDR data, because the participant reported it as “Intersex”, “I prefer not to answer”, “none of these fully describe me”, or skipped the question, we passed the sex concordance check, regardless of the information from the inferred sex ploidy. The sex assigned at birth data from the CDR is described in Appendix F.

Results

We do not include any lrWGS samples that fail the sex concordance check in the v7 release samples, but one sample was included after a manual inspection. This sample, whose self-reported sex assigned at birth was male, received normalized X and Y coverage of 0.97 and 0.22. This happened in two separate SMRT cells for the sample. As a comparison, the mean normalized chrY coverage of female samples who passed the lrWGS sex concordance

check (samples whose self-reported sex at birth is female and the lrWGS calculated sex is female) is 0.05 with a 0.01 standard deviation. With rounding, the sample's chromosome X and Y ploidy was reported as 1 and 0, respectively; but because of the normalized 0.22 coverage on the Y chromosome, we manually marked the sample as passing and included it in the release.

Cross-individual Contamination Rate

Method

VerifyBamID2 (version 2.0.1) was adapted into a pipeline for estimating cross-individual contamination for lrWGS data. VerifyBamID2 was originally designed for short read sequencing. To make the process scalable, we converted the grch38_noalt BAM to a pileup format at selected sites, where VerifyBamID2 genotypes the input, in parallel per-chromosome.

We also evaluated the accuracy of VerifyBamID2 around the 3% contamination cutoff for lrWGS data to determine if the tool would erroneously pass or fail samples. We did this through an *in silico* mixture of samples, simulating different contamination scenarios and at different levels:

- Cross contamination from a sample from a different population and of opposite sex.
- Cross contamination from a sample from a different population and of the same sex.
- Cross contamination from a sample from the same population of different sex.
- Cross contamination from within a family, i.e. parent-child contamination.

We did not have publicly accessible long reads data for assessing the case where the contaminant is from the same population and the same sex. Given that the sites used by VerifyBamID2 for estimation are all autosomal sites, we don't believe this case will have any effect. All *in silico* mixed BAMs have coverage ~8X to emulate the production coverage.

We tested six levels of contamination (3%, 9%, 17%, 33%, and 50%). At 3%, 9%, and 17%, the error between VerifyBamID2 and our *in silico* mixture was never over 10% of the testing contamination level (eg, error was < 0.3% when testing an *in silico* mixture of 3%). At higher tested contamination levels (33% and 50%), the error stayed within 20%. Note that if contamination were to be this high, fingerprint verification would have failed the sample.

We observed from this experiment that for unrelated samples, VerifyBamID2's estimations are in line with the expected contamination level. For related samples, VerifyBamID2 tends to significantly underestimate the contamination level. This does not affect the v7 lrWGS release because all samples are unrelated.

Results

We did not include any lrWGS samples with a cross-individual contamination rate higher than 2.5% (Figure 17). This is below the error (2.7%) we observed in the *in silico* testing of VarifyBamID2 at 3% contamination, which indicates that all released samples have a cross-contamination rate below 3%.

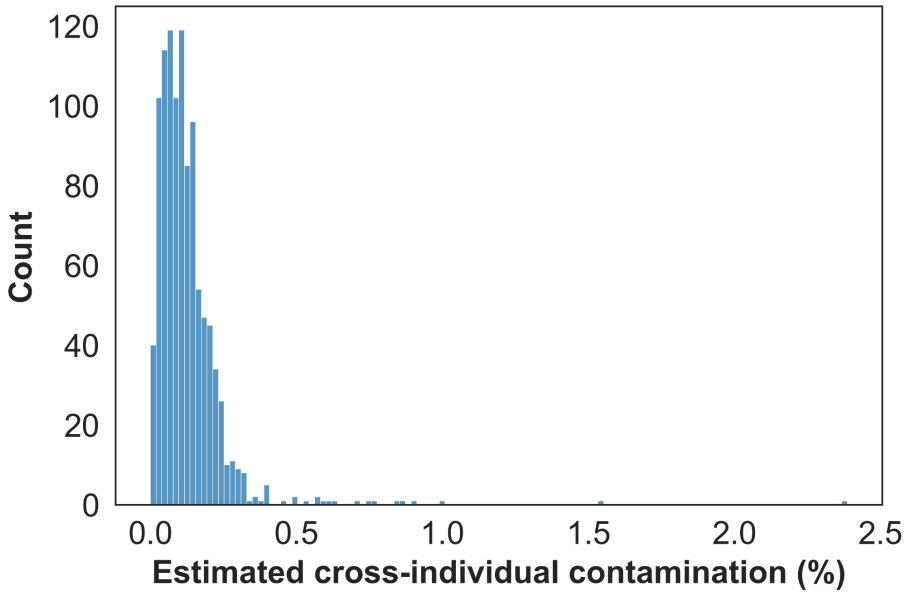


Figure 17 -- Estimated cross-individual contamination for the lrWGS samples, calculated by VerifyBamID2. All samples met the 3% threshold, even when accounting for possible errors in VerifyBamID2.

Coverage

Method

Coverage is defined as the number of reads covering the bases of the genome. Maintaining coverage is important for consistent statistical power and accurate variant calling. We applied a coverage threshold of ≥ 5x mean coverage for the lrWGS samples in order to remove any samples that did not have enough coverage for statistical power and accurate variant calling.

Results

We did not release any samples that did not meet the mean coverage threshold. The frequency of mean coverage can be seen in Figure 18.

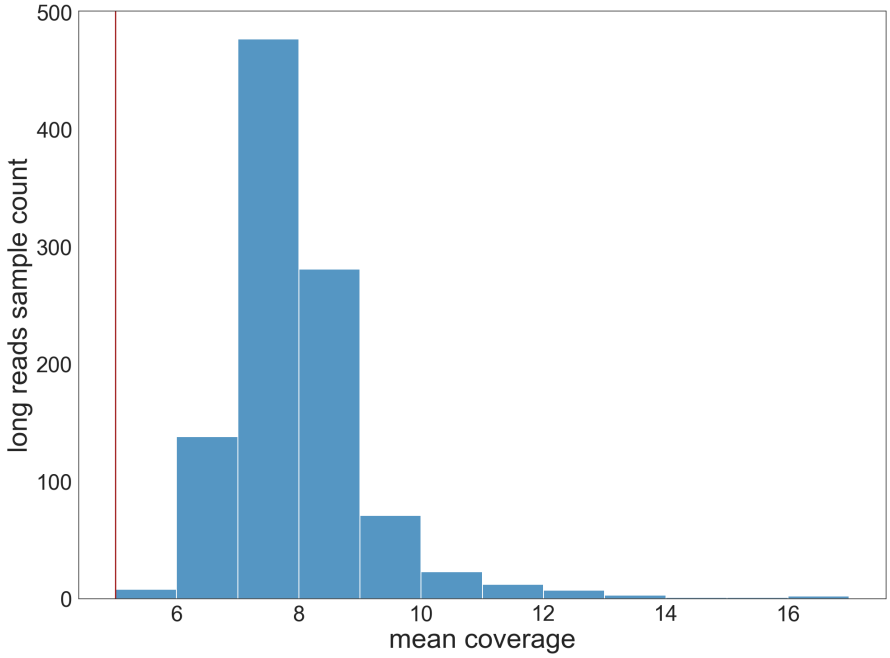


Figure 18 -- Coverage histogram of the lrWGS samples in the v7 release. Note that a mean coverage of 5x was the minimum to be included in the release.

Read Length Median

Method

We calculated the read length median to determine if any samples had shorter fragments that would significantly impact the variant calling performance. The threshold read length median was ≥ 10,000 base pairs and all lrWGS samples passed this check.

Results

We did not release any samples that did not meet the read length median threshold. A distribution of the read length median can be seen in Figure 19.

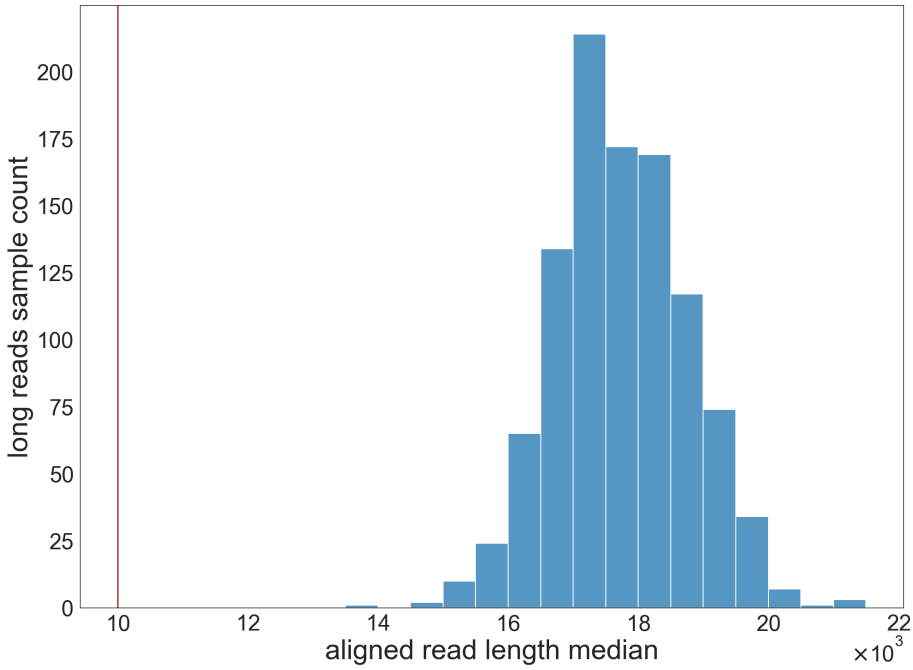


Figure 19 -- Read length median histogram of the lrWGS samples in the v7 release. Note that a read length median of 10,000 base pairs was the minimum to be included in the release.

Outlier Sample Filtering

Method

Variant calling was performed on single sample BAM files for long reads data. SNP and Indel single sample callsets were generated using the PEPPER-Margin-DeepVariant pipeline [45] (version 0.4.1 for PEPPER and version 1.3.0 for DeepVariant). We filtered out events with a QUAL score less than 40. We called SVs with Sniffles2 [46] and PBSV [29].

We used two independent criteria when detecting outlier samples using their single sample VCFs: abnormal SV counts and SNP and Indel counts, given their coverage. We found outliers by plotting the variant counts versus the coverage and manually evaluating the distribution across the entire cohort, seen in Figure 20.

Results

We removed four samples from the v7 release due to their low SV counts and we removed one sample from the release due to the combined SV and small variant count being an outlier. We do not include these samples in the v7 lrWGS release.

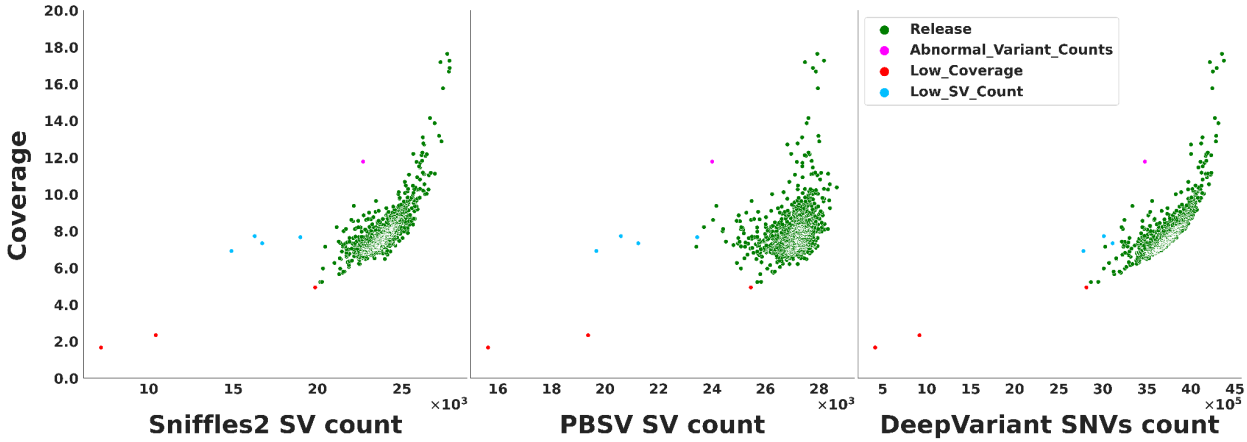


Figure 20 -- We compared the variant count from long read samples to the sample coverage and manually identified outliers. The Sniffles2 and PBSV plots are SV counts and the DeepVariant plot represents SNP and Indel counts. Each dot represents one sample. Green dots (1027) are included in the v7 Release. Cyan dots (4) are excluded because they appear to be outliers when judged by their SV counts. Magenta dots (1) are excluded because they appear to be outliers when judged by their SV and SNP and Indel counts. Red dots (3) are excluded due to low coverage.

Joint Callset QC for the lrWGS SNP/Indel callsets

Joint callset QC is performed on the joint SNP and Indel callset from the long reads data (Table 13). The single sample VCFs were joined using GLNexus (version 1.4.1) [47], parallelized per chromosome. Final joint-called SNP and Indel callsets were converted to Hail MatrixTable (Hail MT) format for analysis. For the Hail MT joint callset QC relatedness and sample population outlier analyses, we used high quality sites that can be called accurately (Appendix J). If a sample was an outlier, we did not include it in the v7 release, as opposed to the srWGS joint callset QC, which flagged but did not remove outlier samples.

Table 13 -- lrWGS joint callset QC steps

QC process	Error modes addressed	V7 release results
Relatedness	- Sample swaps - Related samples, which confound analyses	No sample pairs had kinship scores > 0.1, indicating that no samples were related.
Sample Population Outlier	- Noisy samples	All outlier samples were removed from the v7 release and are not included in the lrWGS joint callset

Relatedness

Method

We used the Hail pc_relate function to determine the kinship score of pairs, as we did with the short reads data in Appendix K. We ran `pc_relate` on the Hail MT based on the lrWGS SNP/Indel joint callset.

Results

We found that no lrWGS sample pairs had kinship scores > 0.1. This indicated that no lrWGS samples were related to a 2nd degree or closer, as we expected.

Sample Population Outlier

Method

We used the high quality sites from the lrWGS SNP/Indel joint callset (Appendix J) to do a population analysis and find population outliers. Following a similar method to the srWGS SNP & Indel sample population outlier flag analysis, we regressed out sixteen principal component features for the joint lrWGS genotype matrix in Hail.

Results

We manually inspected samples in PC1 and PC2 and discovered two outlier samples in the callset, which were excluded from the release (Figure 21).

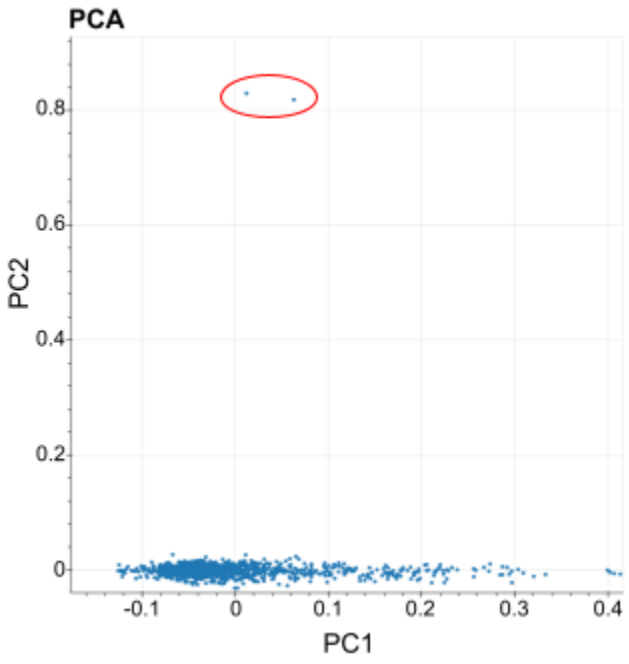


Figure 21 -- Principal component analysis of the population structure of the lrWGS joint callset in order to determine outlier samples. The two outliers were identified manually and removed from the v7 lrWGS release.

Known Issues

The issues below apply to the v7 release genomic data (arrays, srWGS, srWGS SVs, long reads, and auxiliary data). 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

Update as of April 18, 2024*. The 20 array samples have been removed from the array Hail MT and PLINK files, released as version 7.1 on April 18, 2024.*

Six srWGS samples and 20 Array samples in this release are missing their corresponding CDR data. The affected participants are consented to appear in the genomic data.

Affects:

- srWGS SNP & Indel samples: VDS, VCF, PLINK, and Hail MT formats
- Array samples: VCFs, PLINK, and Hail MT

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 and array samples without corresponding data in the CDR.

Remediation:

- We will provide two lists (srWGS and array) of the affected samples through the CDR.

Known Issue #2: 11 samples were affected by a sample swap incident

Update as of April 18, 2024*. The 11 array samples have been removed from the array Hail MT and PLINK files, released as version 7.1 on April 18, 2024.*

We identified a sample swap incident from the GCs which affects 11 samples with genomic data, one srWGS sample and 11 array samples in the v7 release. This was due to an internal issue that we have now remediated.

Affects:

- srWGS SNP & Indel samples: VDS, VCF, PLINK, and Hail MT formats
- Array samples: VCFs, PLINK, and Hail MT

Suggested action:

- Remove samples that are affected by the sample swap from your analysis. We provide two list files containing the research IDs of the affected samples, one file for srWGS and one file for array.

Remediation:

- We have fixed these sample swaps for future releases.

Known issue #3: Array samples (N=416) from previous release are missing in this release

A total of 416 array samples from the previous release are not included in the newest v7 release. This happened for multiple reasons:

1. For this release (v7 C2022Q4R9), we have reprocessed the arrays (details of the new approach are found in Appendix E). As a result, some array samples passed QC in the previous release, but did not pass QC checks in this release.
2. The participant withdrew consent between releases.
3. The sample was not reprocessed in time for this release. These samples will appear in a future release.

At this time, we cannot provide a breakdown of the counts for each of the above reasons.

Affects:

- Array samples: VCFs, PLINK, and Hail MT.

Suggested action:

- If you are not using any array data from the previous release (C2022Q2R2), then no action is necessary.
- If your cohort includes one of the samples that is missing you should upgrade your cohort to include only samples in the v7 set.

Remediation:

- We have released a list of missing array samples on the User Support Hub [1].
- We will continue to re-process the array samples (ETA 2023). All new array samples will use the new process as described in Appendix E.

Known Issue #4: Single array sample missing from Array Hail MT and PLINK files

Update as of April 18, 2024*. The sample has been included in the array Hail MT and PLINK files, released as version 7.1 on April 18, 2024.*

For sample 3518297, we only have a corresponding array in VCF. This sample was not included in the array Hail MT and PLINK files. This was due to an internal issue in synchronization between our srWGS and array sample lists.

Affects:

- Array Hail MT and PLINK files

Suggested action:

- No action. For array analyses, we recommend that you proceed with this one sample missing. We believe that the effort to add one sample (0.0003%) to the MT or PLINK files is not worth the expense.

Remediation:

- This will be fixed in the next release.

Known Issue #5: Larger than expected changes in ancestry predictions from previous release

We have found that the predicted ancestry for 7744 participants changed from the previous release (7.9 % of the v6 C2022Q2R2 data release). Most of the changes were to and from the "Other" classification (7519 participants, 7.6% of the v6 release). This was due to a change in the SNP set used as the features in the PCA for prediction. We used chr20 and chr21 in v6 and the autosomal exome in the current v7 release (see Appendix A). The ancestry changes may affect your analysis if you migrate from the previous release to this release. In future releases, we do not expect large ancestry changes (see Remediation below). The VAT uses these computed ancestries to generate *All of Us* population (gvs_*_*) annotations. The genomic data in the public Data Browser are also dependent on the ancestry predictions for populating population information about variants.

Affects:

- Ancestry predictions, if you are migrating from v6 ancestry to v7.
- Variant Annotation Table (VAT)
- Public Data Browser

Suggested action:

- If you are not using ancestry predictions from the previous v6 release, then no action
- If you are migrating your analysis from the previous v6 release to the v7 release and you used ancestry predictions, we recommend that you rerun downstream analyses that are affected by computed ancestry.

Remediation:

- We are planning to continue to use the current (v7) SNP set as features for ancestry prediction and thus, we do not expect large ancestry changes in future releases. In an internal test, we used the new feature set with the previously released v6 data and compared the ancestry predictions to the current v7 data with the new feature set. We found that only 1890 samples (1.9%) changed ancestry assignment. Note that of these samples, all samples switched to/from the "Other" classification.

Known issue #6: Ancestry prediction has higher error rates for Middle Eastern ancestry

A paucity of labeled Middle Eastern samples reduced the performance of the random forest classifier. This caused the confidence to dip when predicting ancestry for Middle Eastern samples, which caused a larger proportion of these samples relative to other computed ancestries, to be classified as Other ("oth"). The VAT uses these computed ancestries to generate *All of Us* population (gvs_mid_* and gvs_oth_*) annotations. The genomic data in the public Data Browser are also dependent on the ancestry predictions for populating population information about variants.

See Table A.2 for details of the ancestry prediction performance

Affects:

- Ancestry predictions
- Variant Annotation Table (VAT)
- Public Data Browser

Suggested Action:

- When limiting cohorts to samples with computed ancestry of Middle Eastern (“mid”), use the ancestry predictions that do not include “other”. In other words, use the “ancestry_pred” column, instead of “ancestry_pred_other”.

Remediation:

- We have completed investigating other approaches and have minimized this error in the data. However, we do not have enough samples of middle eastern ancestry in our training data to fully remediate this issue. We will be adding this information in the FAQ section for all future releases.

Known Issue #7: VDS issue with GT

In the srWGS SNP & Indel VDS, we have precalculated the genotypes (GT) from the local genotype field (LGT). If you filter any participants from the VDS, you will need to recalculate the GT field. Examples of how to do this are found in example notebooks in the Researcher Workbench.

Affects:

- srWGS SNP & Indel VDS

Suggested Action:

- If you are not using the VDS, no action.
- If you filter participants from the VDS, make sure to recalculate the GT fields.

Remediation:

- We have provided documentation in Researcher Workbench showing how to recalculate GT fields from the VDS.
- In future releases, we will also provide a FAQ question and eliminate this Known Issue.

Known Issue #8: AS_VQSLOD is incorrect in the VDS and dropped in callset data with all participants

There is inconsistency in the AS_VQSLOD annotation in the srWGS SNP & Indel VDS. This annotation indicates information regarding the allele-specific filtering model and we generally do not recommend that researchers use it for their analysis. Instead, we recommend using the filters annotations. In the VDS, AS_VQSLOD is included but incorrect. In the reduced SNP & Indel callset VCF, Hail MT, and PLINK files, AS_VQSLOD is dropped. AS_VQSLOD is correct in the cohort builder.

Affects:

- srWGS SNP & Indel samples: VDS

Suggested action:

- Ignore (or drop) AS_VQSLOD from any analyses involving the VDS.

- Update your analyses to not use this annotation

Remediation:

- We will provide the correct AS_VQSLOD in future releases.

Known Issue #9: Smaller callset ChrM VCFs are empty

We provide the srWGS SNP and Indel callset in VCF format over limited genomic regions, sharded by chromosome. The chrM VCFs are empty and do not contain any data. We cross-checked the complete callset VDS and verified that there are no calls on chrM.

Affects:

- srWGS SNP & Indel samples: VCF format

Suggested action:

- No action.
- If you are using the srWGS SNP & Indel smaller callsets in VCF format, do not include the chrM.vcf.bz file in your analysis.

Remediation:

- We will not release these empty files in the next release.

Known issue #10: srWGS SNP & Indel VDS and VCFs from the Cohort Browser will have extraneous INFO field (AS_YNG)

The srWGS joint callset includes AS_YNG (an INFO field) in the VDS, which should be ignored by researchers. We have removed AS_YNG from all other genomic srWGS SNP & Indel VCFs and Hail MTs, but it will appear when creating a VCF from the Cohort Browser

Affects:

- srWGS SNP & Indel VCF files created from the Cohort Browser
- srWGS SNP & Indel variants: VDS

Suggested Action:

- Do not include the AS_YNG field in any analyses when creating a VCF from the Cohort Browser.
- Ignore (or drop) AS_YNG from any analyses involving the VDS.

Remediation:

- We will update the VDS to remove the AS_YNG field (ETA 2024)
- We will update the Cohort Browser data to remove the AS_YNG field (ETA 2023)

Known issue #11: srWGS SNP & Indel variant calls on chromosome Y need additional filtering

We see variants with heterozygous calls in chromosome Y, which cannot be correct germline calls. After manual review, we believe that regions of chromosome Y are prone to misalignment artifacts (low mappability). This will cause heterozygous calls in chrY that are likely artifacts. We have not investigated whether these are somatic mutations.

Affects:

- srWGS SNP & Indel variants: VDS, VCF, Hail MT formats

Suggested Action:

- If you do not use variant calls on chrY, then no action.
- Otherwise, we recommend that you use AD, GQ, and GT to filter variants on chromosome Y.

Remediation (ETA 2023):

- We will provide a set of regions (via a BED file) that researchers can use to mask regions of the genome with poor calling accuracy for chromosome Y. It is not currently available with the v7 release.

Known Issue #12: QUAL information has been removed for srWGS SNP & Indel variants

The srWGS SNP & Indel variants are now released in VDS format (see the VDS article on the User Support Hub [1]). In the VDS format, the actual QUALApprox annotation is not included. This will affect other srWGS short variant files, such as the smaller callsets (eg, exome). The filters used for srWGS SNP & Indel variants are the same and correct, but the annotation is not included.

Affects:

- srWGS SNP & Indel variants: VDS, VCF, Hail MT, and PLINK formats

Suggested action:

- Use the filter field to determine the quality of variants
- If this annotation is important to your (current or future) analyses, please contact the User Support Team [1]. We do not plan to remediate this unless we hear from researchers.

Remediation:

- This will not be remediated, this known issue will move to a FAQ for all future releases.

Known Issue #13: srWGS callset using new convention for genotype filtering flag

The srWGS SNP & Indel variants are now released in VDS format (see the VDS article on the User Support Hub [1]). In order to reduce the genotype metadata stored and reduce the size of the srWGS SNP & Indel variant dataset, the genotype filtering reason is no longer available. Genotype filtering, which is in the VCF and the VDS as the `FT` annotation, is reported as PASS, FAIL, or ".". Treat "." as PASS. In previous AoU releases, we reported more filtering information, including annotations about VQSR filtering.

Affects:

- srWGS SNP & Indel variants: VDS, VCF, PLINK, and Hail MT formats

Suggested action:

- Update your analyses to check the genotype filter (FT) for "FAIL" instead of "low_VQSLOD_INDEL" or "low_VQSLOD_SNP" to determine whether or not to use a srWGS SNP & Indel variant in your analysis.

Remediation:

- This will not be remediated, this known issue will move to a FAQ for all future releases.

Known Issue #14: Data processing issue affecting array data

Update as of April 18, 2024. We identified a sample swap incident that affects 63 samples with array data in the v7 release.

Affects:

- Array samples: IDATs, VCFs, PLINK bed, and Hail MT

Suggested action:

- Update to the new 7.1 merged array Hail MT or PLINK bed files.
- Remove affected single samples IDAT or VCFs from analysis. We provided a list file containing the research IDs of the affected samples.

Remediation:

- We generated a new array merged Hail MT and merged PLINK bed dataset, released April 18, 2024.

Known Issue #15: Array and srWGS data with bone marrow transplant history

Update as of April 18, 2024. The DRC has identified 1430 participants with a history of an allogeneic bone marrow transplant or a bone marrow transplant of an unknown type. This affects 1430 array samples and 12 srWGS samples.

Affects:

- Array samples: IDATs, VCFs, PLINK bed, and Hail MT formats
- srWGS samples: CRAMs, VDS, VCFs, PLINK bed, and Hail MT formats

Suggested action:

- Update to the new 7.1 merged array Hail MT or PLINK bed files.
- Remove affected samples from your analysis. We provide two list files containing the research IDs of the affected samples, one file for srWGS and one file for array.

Remediation:

- We generated a new array merged Hail MT and merged PLINK bed dataset, released April 18, 2024.

FAQ

1. Why do you fail samples based on contamination rate for srWGS, but not for array samples?

   srWGS analyses (e.g., mosaicism) rely on other signals, such as read counts, which are affected by contamination. Low rates of contamination do not affect array calls and problematic levels of contamination will be reflected in the array call rate.

2. Did you remove samples from participants with bone marrow transplants?

   Yes, we removed both array and srWGS samples associated with participants that have received bone marrow transplants from allogeneic transplantation (transplantation from another person), according to the corresponding electronic health record (EHR) and survey responses provided by participants (Overall Health). We did not remove samples who received bone marrow transplants from autologous transplantation (transplantation from themselves).

3. Are all samples in the srWGS joint callset sourced from blood?

   Yes. Although the program does have saliva srWGS samples, we did not include these samples in the v7 release. Once we identify any batch effects between saliva and blood samples (ETA 2023), we will reassess the inclusion of saliva samples in the joint srWGS callset. If we decide that the batch effects will have minimal impact, we will include saliva samples in the srWGS joint callsets in 2024.

   Update as of April 18, 2024. The DRC has identified 1430 participants in the v7 data release with a history of allogeneic transplantation (transplantation from another person) or a bone marrow transplant of an unknown type. Please see Known Issue #15 for more information.

References

[1] ***All Of Us User Support Hub*** https://aousupporthelp.zendesk.com/hc/en-us
[2] Illumina GenCall Data Analysis Software. (n.d.). Retrieved October 21, 2021, from https://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf
[3] CollectArraysVariantCallingMetrics (Picard), Retrieved October 21, 2021 , from https://gatk.broadinstitute.org/hc/en-us/articles/360037593871-CollectArraysVariantCallingMetrics-Picard-
[4] G. Jun et al., ***Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data***, American journal of human genetics doi:10.1016/j.ajhg.2012.09.004 (volume 91 issue 5 pp.839 - 848)
[5] E Venner, D Muzny, et al., ***Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us Research Program***, *Genome Medicine* (2022). https://doi.org/10.1186/s13073-022-01031-z
[6] ***Detecting sample swaps with Picard tools – GATK.*** (n.d.). Retrieved October 21, 2021, from https://gatk.broadinstitute.org/hc/en-us/articles/360041696232-Detecting-sample-swaps-with-Picard-tools
[7] Pedersen and Quinlan, **Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy** The American Journal of Human Genetics (2017) http://dx.doi.org/10.1016/j.ajhg.2017.01.017
[8] Zhang F, et al. **Ancestry-agnostic estimation of DNA sample contamination from sequence reads**. *Genome Research* (2020). https://doi.org/10.1101/gr.246934.118
[9] ***Phred-scaled quality scores – GATK.*** (n.d.). Retrieved January 31, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores.
[10] 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
[11] ***gnomAD v3.1 New Content, Methods, Annotations, and Data ....*** (n.d.). Retrieved February 1, 2022, from https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability.
[12] 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.166
[13] ***Relatedness - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/relatedness.html#hail.methods.pc_relate.
[14] ***Which training sets arguments should I use for running VQSR ....*** (n.d.). Retrieved February 1, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/4402736812443-Which-training-sets-arguments-should-I-use-for-running-VQSR-.
[15] ***Resource bundle – GATK.*** (n.d.). Retrieved February 1, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle.
[16] ***The Omni Family of Microarrays.*** (n.d.). Retrieved February 16, 2022, from https://www.illumina.com/Documents/products/datasheets/datasheet_gwas_roadmap.pdf.

[17] International HapMap Consortium. **The International HapMap Project**. Nature. 2003 Dec 18;426(6968):789-96. doi: 10.1038/nature02168. PMID: 14685227.
[18] The 1000 Genomes Project Consortium, ***A global reference for human genetic variation***, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393
[19] Mills R.E. et al. ***An initial map of insertion and deletion (INDEL) variation in the human genome***. Genome Res. 2006;16:1182–1190. doi:10.1101/gr.4565806.
[20] Krusche, P., Trigg, L., Boutros, P.C. *et al.* ***Best practices for benchmarking germline small-variant calls in human genomes***. *Nat Biotechnol* **37,** 555–560 (2019). https://doi.org/10.1038/s41587-019-0054-x
[21] 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
[22] ***Structural Variants*** (n.d.). Retrieved March 3, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022476791323-Structural-Variants
[23] 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
[24] 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
[25] 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.
[26] ***All of Us Research Program Data and Statistics Dissemination Policy*** (May 2020) Retrieved March 5, 2023 from https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf
[27] 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.
[28] P. Ebert, P. A. Audano, Q. Zhu et al., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
[29] **PacBio structural variant calling and analysis tools (PBSV)**, Retrieved March 3, 2023, from https://github.com/PacificBiosciences/pbsv.
[30] Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018 Jun;15(6):461-468. doi: 10.1038/s41592-018-0001-7. Epub 2018 Apr 30. PMID: 29713083; PMCID: PMC5990442.
[31] Mills, Ryan E et al. Mapping copy number variation by population-scale genome sequencing. Nature vol. 470,7332 (2011): 59-65. doi:10.1038/nature09708
[32] Handsaker, R., Van Doren, V., Berman, J. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296-303 (2015). https://doi.org/10.1038/ng.3200
[33] **XGBoostMinGqVariantFilter** (n.d.) Retrieved March 4, 2023, from unreleased GATK branch https://github.com/broadinstitute/gatk/tree/tb_recalibrate_gq

[34] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016
[35] 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.
[36] **VisualizeCnvs.wdl** (n.d.) Retrieved March 4, 2023, from https://github.com/broadinstitute/gatk-sv/blob/v0.26.5-beta/wdl/VisualizeCnvs.wdl
[37] 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
[38] Li, H. and Handsaker, B. et al. "The Sequence Alignment/Map format and SAMtools." Bioinformatics, 25 (2009): 2078–2079, https://doi.org/10.1093/bioinformatics/btp352
[39] Wenger, A.M., Peluso, P., Rowell, W.J. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37, 1155–1162 (2019). https://doi.org/10.1038/s41587-019-0217-9
[40] **Pacbio glossary**, Retrieved March 8, 2023, from https://www.pacb.com/wp-content/uploads/2015/09/Pacific-Biosciences-Glossary-of-Terms.pdf
[41] Li, Heng. "Which Human Reference Genome to Use?" *Heng Li's Blog*, 13 Nov. 2017, https://lh3.github.io/. Accessed 2 Mar. 2023.
[42] Schneider, Valerie A et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly." *Genome research* vol. 27,5 (2017): 849-864. doi:10.1101/gr.213611.116
[43] Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, et al. The complete sequence of a human Y chromosome. bioRxiv, 2022.
[44] Pedersen, B.S. and Quinlan, A.R. "Mosdepth: quick coverage calculation for genomes and exomes", Bioinformatics, 34(2018):867–868 https://doi.org/10.1093/bioinformatics/btx699
[45] Shafin, K., Pesout, T., Chang, PC. et al. "Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads." *Nat Methods* 18, 1322–1332 (2021). https://doi.org/10.1038/s41592-021-01299-w
[46] **Sniffles2 (PBSV)**, Retrieved March 3, 2023, from https://github.com/fritzsedlazeck/Sniffles
[47] Yun, T., et al. "Accurate, scalable cohort variant calls using DeepVariant and GLnexus" Bioinformatics, 36 (2020): 5582–5589, https://doi.org/10.1093/bioinformatics/btaa1081
[48] 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
[49] 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
[50] 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.
[51] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *Journal of Machine Learning Research* 12, pp. 2825-2830, (2011).
[52] ***Downloads | gnomAD.*** (n.d.). Retrieved February 1, 2021, from https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg.

[53] 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
[54] ***Genetics - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/genetics.html#hail.methods.hwe_normalized_pca.
[55] Laurie CC, Doheny KF, et al. ***Quality control and quality assurance in genotypic data for genome-wide association studies***. Genet Epidemiol. 2010 Sep;34(6):591-602. doi: 10.1002/gepi.20516. PMID: 20718045; PMCID: PMC3061487.
[56] Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ***ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.*** Genet Med. 15:565–574. (2013)
[57] Erdős, P. **On cliques in graphs**, Israel Journal of Mathematics, 4 (4): 233–234, (1966), doi:10.1007/BF02771637, MR 0205874, S2CID 121993028
[58] 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.
[59] 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.
[60] ***Structural variant (SV) discovery*** (n.d.). Retrieved March 15, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022487952155-Structural-variant-SV-discovery
[61] **WDL Specification**, from https://github.com/openwdl/wdl
[62] **Long Reads Pipeline**, Retrieved March 7, 2023 from https://github.com/broadinstitute/long-read-pipelines

Appendix A: Ancestry

We computed categorical ancestry for all of the srWGS SNP & Indel samples in *All of Us* and made these available to researchers. These predictions are also the basis for population allele frequency calculations in the Variant Annotation Table (e.g. gvs_afr_ac) and data in the Genomic Variants section of the public Data Browser. We used the high-quality set of sites (HQ sites), described in Appendix J, to determine an ancestry label for each sample. The ancestry categories are based on the same labels used in gnomAD [48], Human Genome Diversity Project [49], and 1000 Genomes [18]:

- African (afr)
- Latino/Native American/Ad Mixed American (amr)
- East Asian (eas)
- Middle Eastern (mid)
- European (eur) -- Composed of Finnish (FIN) and Non-Finnish European (NFE)
- Other (oth) -- not belonging to one of the other ancestries or is an admixture.
- South Asian (sas)

We trained a random forest classifier [50,51] on a training set of the HGDP and 1kg samples variants on the autosomal exome, obtained from gnomAD [52]. This exome was derived from the exon regions of all autosomal, basic, protein-coding transcripts in GENCODE v42 [53]. We generated the first 16 principal components (PCs) of the training sample genotypes (using the hwe_normalized_pca in Hail [54]) at the high-quality variant sites (see Appendix J) 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 determine the ancestry of *All of Us* samples, we project the *All of Us* samples into the PCA space of the training data and apply the classifier (see Figure A.1). Since we do not have truth labels, we can not determine the accuracy of our *All of Us* predictions. As a proxy for the accuracy of our *All of Us* predictions, we look at the concordance between the survey results and the predicted ancestry. The ancestry predictions can be found in Table A.1.

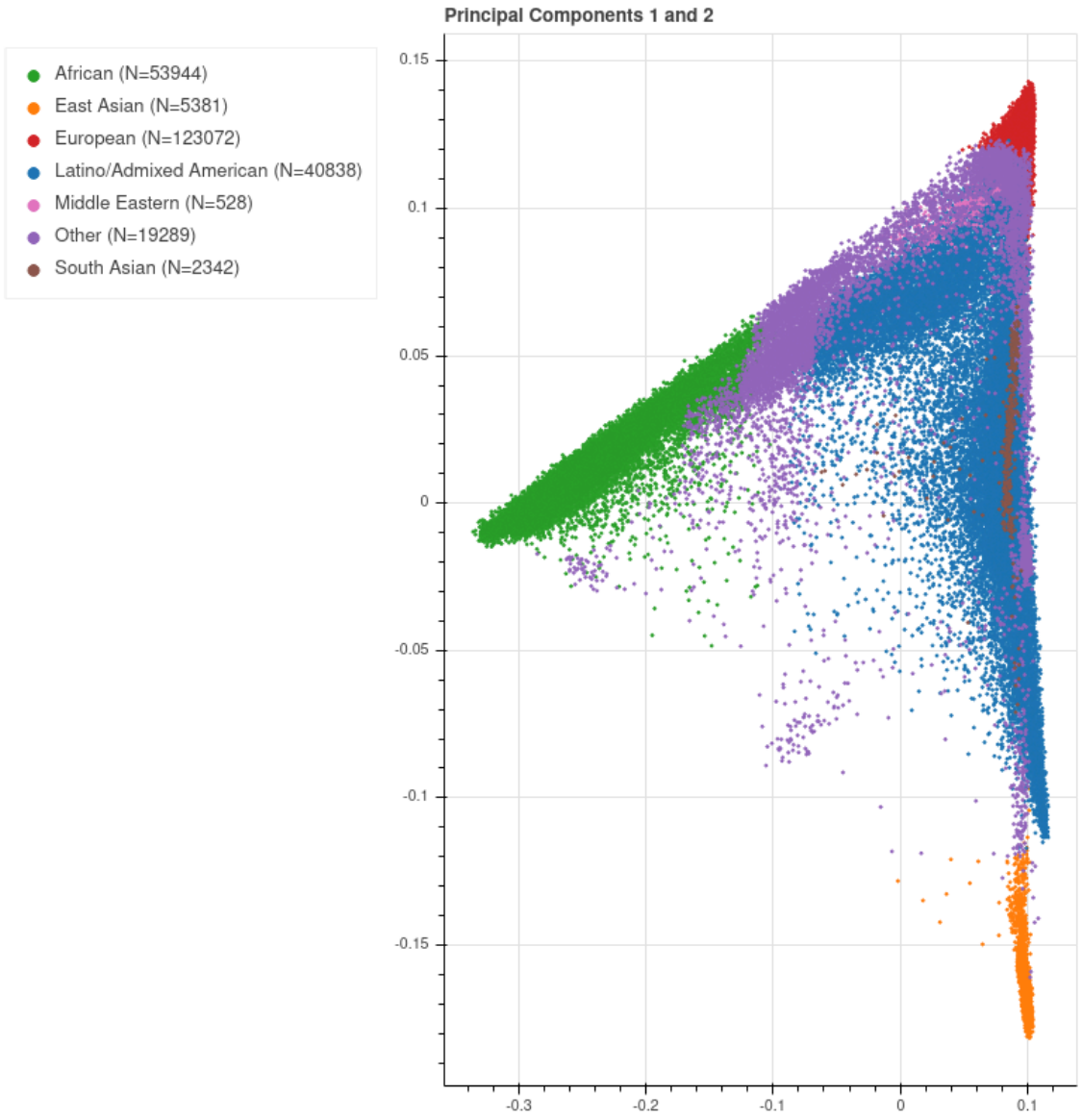


Figure A.1 -- Ancestry predictions for the *All of Us* srWGS samples plotted on the first two principal components (PC1 on x-axis and PC2 on the y-axis) of the genotype calls.

Table A.1 -- Breakdown of the computed ancestries in *All Of Us* srWGS data

Computed Ancestry (sorted by percentage)	Count (percentage)
European	123072 (50.2%)
African	53944 (22.0%)
Latino/Admixed American	40838 (16.6%)
Other	19289 (7.9%)

East Asian	5381 (2.2%)
South Asian	2342 (1.0%)
Middle Eastern	528 (0.2%)

Total:	245394 (100.0%)

We evaluated the performance of the ancestry predictions using two different test datasets:

1. A holdout set of training samples. We tested performance with and without the "Other" ancestry
    a. Error rate (incl Other): 0.050
        i. See Table A.2
        ii. Please see Known Issue #6, since the error rate is higher for Middle Eastern (mid) ancestry. Our classifier conflates Middle Eastern and Other.
    b. Error rate (not incl Other): 0.002
        i. See Table A.3

Table A.2 -- Error rate (incl. Other) on labeled training data using holdout set

	Predicted						
Actual	AFR	AMR	EAS	EUR	MID	OTH	SAS
AFR	198	0	0	0	0	2	0
AMR	0	50	0	0	0	0	0
EAS	0	0	199	0	0	1	0
EUR	0	0	0	198	0	2	0
MID	0	0	0	0	49	1	0
OTH	0	2	2	3	25	7	8
SAS	0	0	0	0	0	2	198

Table A.3 -- Error rate (not incl. Other) on labeled training data using holdout set

	Predicted					
Actual	AFR	AMR	EAS	EUR	MID	SAS
AFR	199	1	0	0	0	0
AMR	0	50	0	0	0	0
EAS	0	0	199	0	0	1

EUR	0	0	0	200	0	0
MID	0	0	0	0	50	0
SAS	0	0	0	0	0	200

2. We measured the concordance of the ancestry predictions against the self-reported race and ethnicity of the *All of Us* samples. The concordance should be worse than the holdout HGDP samples, but this is expected. Self-reported race and ethnicity does not correspond to the populations listed above (race and ethnicities are social constructs and the ancestry predictions are computed from the genotypes). We expect these to be correlated, but not corresponding. This concordance evaluation is meant to discover large errors in our ancestry predictions.

    "Concordant" labeling between HGDP/1kg populations and *All of Us* race/ethnicities:

    1. African (AFR) → Black
    2. Latino/Ad Mixed American (AMR) → Hispanic
    3. East Asian (EAS) → Asian
    4. Finnish (FIN) → White
    5. Middle Eastern (MID) → MENA
    6. Non-Finnish European (NFE) → White
    7. Other (OTH) → Other (do not include skipped)
    8. South Asian (SAS) → Asian

We do not include any samples where the self-reported race/ethnicity is "Skip", "Prefer not to answer", or was not filled in. If a participant selected that their race/ethnicity was not a possible selection ("NoneOfThese"), we counted them as "Other".

Based on the procedure above, the concordance between self-reported race/ethnicity and the ancestry predictions: 0.898.

Appendix B: Self-reported race/ethnicity

As seen in Table B.1, the race/ethnicity breakdown of the genomic data is similar to all participants *All of Us* CDR release C2022Q4R9. Samples with "Skip" responses include participants that answered "prefer not to answer", entered blank text, or did not respond to the survey question.

Please see Known Issue #1, as six srWGS samples and 10 array samples are missing CDR data. All other array, srWGS, and lrWGS samples have corresponding survey data (Appendix C).

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

Self-reported Race/Ethnicity	Survey response counts (%)	Array counts (%)	srWGS counts (%)	srWGS SV counts (%)	lrWGS counts (%)
Asian	13838 (3.3%)	9605 (3.1%)	7422 (3%)	260 (2.3%)	–
Asian, White	1894 (0.5%)	1284 (0.4%)	992 (0.4%)	32 (0.23%)	–
Black	77069 (18.6%)	62514 (20.0%)	50064 (20.4%)	3938 (34.6%)	911 (88.7%)
Black, White	2390 (0.6%)	1743 (0.6%)	1351 (0.6%)	158 (0.14%)	54 (4.8%)
Hispanic	64672 (15.6%)	52904 (16.9%)	41938 (17.1%)	1586 (13.9%)	–
Hispanic, White	6512 (1.6%)	4718 (1.5%)	3682 (1.5%)	131 (0.12%)	–
Other	15294 (3.7%)	11264 (3.6%)	8709 (3.5%)	457 (4%)	49 (4.8%)
Skip	91422 (2.2%)	6872 (2.2%)	5387 (2.2%)	237 (2.1%)	13 (1.3%)
White	222646 (53.8%)	162021 (51.8%)	125843 (51.3%)	4591 (40.3%)	–
Total	**413457 (100.0%)**	**312925 (100.0%)**	**245388 (100.0%)**	**11390 (100.0%)**	**1027 (100.0%)**

Appendix C: Data type availability with genomic data

Please see Known Issue #1, as six srWGS samples and 10 array samples are missing CDR data. All other array, srWGS, and lrWGS samples have corresponding survey data. The srWGS samples are a subset of the array data. The srWGS SV samples are a subset of the srWGS SNP & Indel data (all srWGS samples with SNP and Indel data have SV data). The long reads samples are a subset of the srWGS SNP & Indel and array data.

Additionally, array (Table C.1), srWGS SNP & Indel (Table C.2), srWGS SV (Table C.3), and lrWGS (Table C.4) data have other corresponding non-genomic data. This can be one or more of the following:

- Electronic Health Records (EHR)
- Physical Measurements (PM)
- Participant Provided Information (PPI/surveys)
- Fitbit (FB)

Descriptions of the non-genomic data can be found on the *All of Us* Data Sources page.

Table C.1 -- Array overlap with non-genomic data types

Data Combination	Description	Participant Count
Array	any Array data	312925
Array and PPI	any Array AND any PPI	312925
Array and PPI and PM	any Array AND any PPI AND any PM	303064
Array and EHR	any Array AND any EHR	255052
Array and PPI and EHR	any Array AND any PPI AND any EHR	255052
Array and PPI and EHR and PM	any Array AND any EHR AND any PM AND any PPI	254203
Array and Fitbit	any Array AND any Fitbit	11763
Array and PPI and Fitbit	any Array AND any PPI AND Fitbit	11763
Array and PPI and PM and Fitbit	any Array AND any PPI AND any PM AND any Fitbit	10442
Array and Fitbit and PPI and EHR	any Array AND any Fitbit AND and PPI AND any EHR	8867
Array and PPI and EHR and PM and Fitbit	any Array AND any EHR AND and PM AND any PPI AND any Fitbit	8778

Table C.2 -- srWGS SNP & Indel overlap with non-genomic data types

Data Combination	Description	Participant Count
srWGS	any srWGS data	245388

srWGS and PPI	any srWGS AND any PPI	245388
srWGS and PPI and PM	any srWGS AND any PPI AND any PM	245149
srWGS and EHR	any srWGS AND any EHR	206173
srWGS and PPI and EHR	any srWGS AND any PPI AND any EHR	206173
srWGS and PPI and EHR and PM	any srWGS AND any EHR AND any PM AND any PPI	206109
srWGS and Fitbit	any srWGS AND any Fitbit	8812
srWGS and PPI and Fitbit	any srWGS AND any PPI AND Fitbit	8812
srWGS and PPI and PM and Fitbit	any srWGS AND any PPI AND any PM AND any Fitbit	8798
srWGS and Fitbit and PPI and EHR	any srWGS AND any Fitbit AND and PPI AND any EHR	7445
srWGS and PPI and EHR and PM and Fitbit	any srWGS AND any EHR AND and PM AND any PPI AND any Fitbit	7444

Table C.3 -- srWGS SV overlap with non-genomic data types

Data Combination	Description	Participant Count
srWGS SV	any srWGS SV data	11390
srWGS SV and PPI	any srWGS SV AND any PPI	11390
srWGS SV and PPI and PM	any srWGS SV AND any PPI AND any PM	11385
srWGS SV and EHR	any srWGS SV AND any EHR	9744
srWGS SV and PPI and EHR	any srWGS SV AND any PPI AND any EHR	9744
srWGS SV and PPI and EHR and PM	any srWGS SV AND any EHR AND any PM AND any PPI	9743
srWGS SV and Fitbit	any srWGS SV AND any Fitbit	425
srWGS SV and PPI and Fitbit	any srWGS SV AND any PPI AND Fitbit	425
srWGS SV and PPI and PM and Fitbit	any srWGS SV AND any PPI AND any PM AND any Fitbit	425
srWGS SV and Fitbit and PPI and EHR	any srWGS SV AND any Fitbit AND and PPI AND any EHR	361
srWGS SV and PPI and EHR and PM and Fitbit	any srWGS SV AND any EHR AND and PM AND any PPI AND any Fitbit	361

Table C.4 -- lrWGS overlap with non-genomic data types

Data Combination	Description	Participant Count
lrWGS	any lrWGS data	1027

lrWGS and PPI	any lrWGS AND any PPI	1027
lrWGS and PPI and PM	any lrWGS AND any PPI AND any PM	1027
lrWGS and EHR	any lrWGS AND any EHR	985
lrWGS and PPI and EHR	any lrWGS AND any PPI AND any EHR	985
lrWGS and PPI and EHR and PM	any lrWGS AND any EHR AND any PM AND any PPI	985
lrWGS and Fitbit	any lrWGS AND any Fitbit	38
lrWGS and PPI and Fitbit	any lrWGS AND any PPI AND Fitbit	38
lrWGS and PPI and PM and Fitbit	any lrWGS AND any PPI AND any PM AND any Fitbit	38
lrWGS and Fitbit and PPI and EHR	any lrWGS AND any Fitbit AND and PPI AND any EHR	37
lrWGS and PPI and EHR and PM and Fitbit	any lrWGS AND any EHR AND and PM AND any PPI AND any Fitbit	37

Appendix D: Genome Centers and Data and Research Center

Below is the listing of the three Genome Centers (GCs), the Data and Research Center (DRC), and the Biobank.

Role	Principal Investigator(s)
Genome Center	Richard Gibbs - Baylor College of Medicine, Johns Hopkins University Eric A. Boerwinkle - Baylor College of Medicine, Johns Hopkins University Kimberly F. Doheny - Baylor College of Medicine, Johns Hopkins University Stacey Gabriel - Broad Institute Gail Jarvik - Northwest Genomics Center at the University of Washington Evan Eichler - Northwest Genomics Center at the University of Washington
Data and Research Center	Paul Harris - Vanderbilt University Medical Center Dan M. Roden - Vanderbilt University Medical Center Melissa Basford - Vanderbilt University Medical Center Anthony Philippakis - Broad Institute David Glazer - Verily Life Sciences
Biobank	Stephen Norman Thibodeau - Mayo Clinic

Appendix E: Array processing overview

See Figure E.2 for an overview of the array genotyping process for the *All of Us* Research Program. The three GCs used identical array products, scanners, resource files, and genotype calling software. The GCs used the Illumina Global Diversity Array (GDA) (https://www.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html).

For the v7 data release (C2022Q4R9), new cluster definition files (.egt) was created at Johns Hopkins using raw data from 12,983 samples from all 3 genotyping centers (3,782-Broad, 4,342-Johns Hopkins, 4,859-UW) in order to reduce batch effects. Manual review and editing of cluster boundaries was performed for 67,812 assays including all X, MT and Y SNPs, rare variant calls with “new hets” detected by z-call (new hets > 2, total hets >=4, and MAF <=0.0025) GEM trait SNPs, fingerprint sites for array concordance to WGS datasets and all assays within the bed file regions for health-related return of results. 11,916 assays were dropped based on manual review and 75,237 assays were dropped based on call rate <99% and/or cluster separation <0.4. 681 trios were examined for mendelian segregation errors, 15 SNPs were dropped due to >1 mendelian error. A homogeneous subset of 7,511 samples was defined using PCA and MCD (minimum covariance determinant method). Using this homogeneous sample subset, HWE and sex differences in allele frequency were evaluated. 4,005 SNPs were dropped due to Hardy Weinberg equilibrium p-value less than 10^{-4} and 258 SNPs were dropped due to a sex difference in allele frequency of >0.2. \Batch effects were evaluated by comparing allele frequencies between genotyping centers within the homogenous sample subset. Chi-square statistics were Broad 0.73, Johns Hopkins 0.74, UW 0.74.

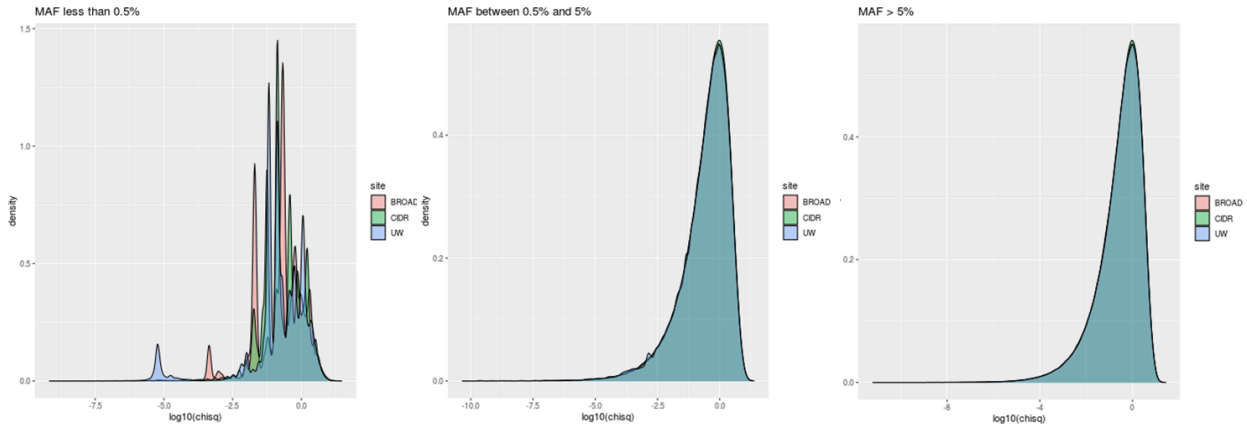


Figure E.1 Comparisons shown in Figure E.1 broken out into MAF bins.

	Different Genotypes		Missing_1		Missing_2		Same	
	original	reclustered	original	reclustered	original	reclustered	original	reclustered
CIDR_Broad	0.0045%	0.0003%	0.4283%	0.0337%	0.0533%	0.0135%	99.51%	99.95%
CIDR_UW	0.0053%	0.0004%	0.3671%	0.0337%	0.1767%	0.0484%	99.45%	99.92%
Broad_UW	0.0032%	0.0001%	0.0498%	0.0135%	0.2346%	0.0484%	99.71%	99.94%

Figure E.2 Data for the program control sample HG001 was compared to evaluate the performance of the new cluster file. When comparing data between the 3 genotyping centers, missing data rates were decreased and concordance rates were increased.

Array product details:

- Bead pool file: `GDA-8v1-0_D1.bpm`
- EGT cluster file: `GDA-8v1-1_A1_AoUupdated.08.17.21_ClusterFile.egt`
- gentrain v.3
- reference hg19 (Note: We liftover to hg38 before publishing array data in the RW. The IDAT files are raw files and thus have no reference.)
- gencall cut-off 0.15
- 1,814,226 assays
    - 1,767,452 SNVs
    - 36,839 indels
    - 9,934 IntensityOnly (probes intended only for Copy Number Variant (CNV) calling)

Chemistry: Illumina Infinium LCG using automated protocol
Liquid handling robotics: Various platforms across the genome centers
Scanners: Illumina iSCANs with Automated Array Loader
Software:

- Illumina IAAP Version: iaap-cli-linux-x64-1.1.0-sha.80d7e5b3d9c1fdfc2e99b472a90652fd3848bbc7.tar.gz
    - IAAP converts raw data (.idat files – 2 per sample) into a single `.gtc` file per sample using the `.bpm` file (defines strand, probes sequences, and illumicode address) and the `.egt` file (defines the relationship between intensities and genotype calls)
- Picard-2.26.4
    - Picard tool, GTCtoVCF, converts the `.gtc` file into a vcf file.
- BAFRegress version 0.9.3 [4]
    - BAFRegress measures the within species DNA sample contamination using B allele frequency data from Illumina genotyping arrays using a regression model

Quality Control:

Each genome center ran the GDA array under Clinical Laboratory Improvement Amendments (CLIA) compliant protocols. We generated `.gtc` files and uploaded metrics to in-house Laboratory Information Management Systems (LIMS) systems for quality control review. At batch level (each set of 96 well plates run together in the laboratory at one time), each GC included positive control samples, which were required to have > 98% call rate and >99% concordance to existing data, in order to approve release of the batch of data. At the sample level, the call rate and sex are the key quality control determinants [55]. Contamination is also measured using BAFRegress [4] and reported out as metadata. Any sample with a call rate below 98% is repeated one time in the laboratory. Genotyped sex is determined by plotting normalized X versus normalized Y intensity values for a batch of samples [55]. Any sample

discordant with 'sex assigned at birth' reported by an *All of Us* participant is flagged for further detailed review. If multiple sex discordant samples are clustered on an array or on a 96 well plate, the entire array or plate will have data production repeated. Samples identified with sex chromosome aneuploidies are also reported back as metadata (XXX, XXY, XYY, etc). A final processing status of "PASS," "FAIL" or "ABANDON" is determined before release of data to the DRC. An array sample will PASS if the call rate is > 98% and the genotyped sex and sex assigned at birth are concordant. If we do not have a "male" or "female" for the sex assigned at birth, because the participant reported it as "Intersex", "I prefer not to answer", "none of these fully describe me", or skipped the question, the array sample is marked as PASS. The sex assigned at birth data from the CDR is described in Appendix F. An array sample will FAIL if the genotyped sex and the sex assigned at birth are discordant or if the call rate is less than 98% on the first run of the sample. An array sample will have the status ABANDON if the call rate is less than 98% after at least 2 attempts at the GC.

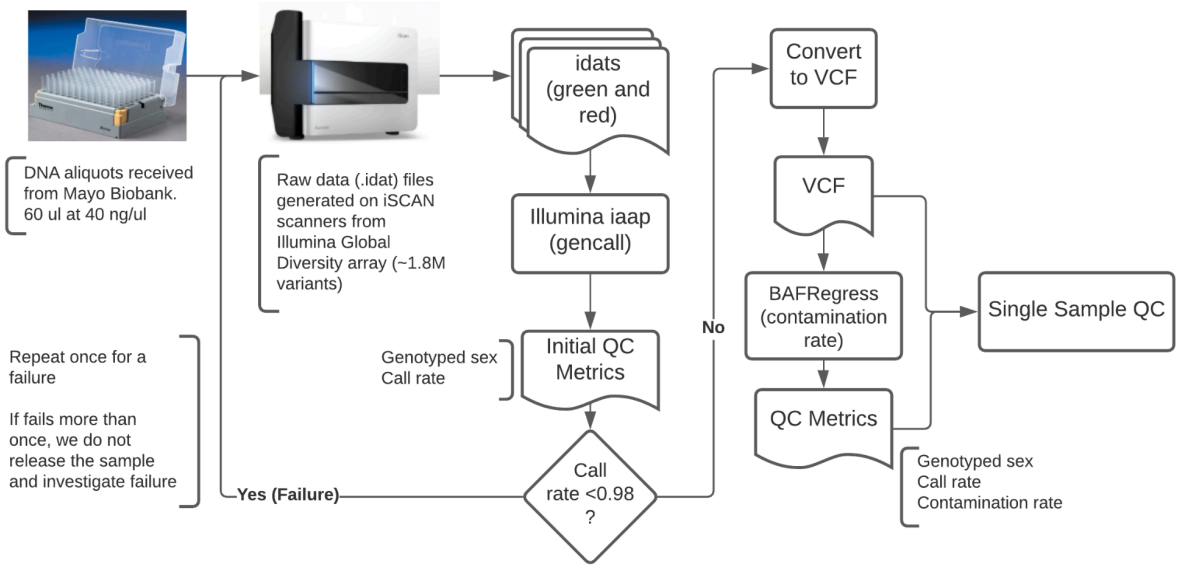


Figure E.3 -- Overview of the array processing pipeline.

Appendix F: Self-reported sex assigned at birth

See Table F.1 for the counts and percentages of participant responses to “What was your biological sex assigned at birth?” in the Basics survey (based on *All of Us* CDR release C2022Q4R9). The CDR code for this question is sex_at_birth. These participant responses are used for the participant self-reported sex at birth information used in sex concordance checks.

Please see Known Issue #1, as six srWGS samples and 10 array samples are missing CDR data. All other array, srWGS, and lrWGS samples have corresponding survey data (Appendix C).

Table F.1 -- v7 release participants response breakdown to sex assigned at birth question

V7 release	Array		srWGS SNP & Indel		srWGS SV		lrWGS	
Sex assigned at birth responses	counts	percent	counts	percent	counts	percent	counts	percent
Female	185210	59.19	145563	59.32	6928	60.83	712	69.33
Male	121173	38.72	94756	38.61	4234	37.17	294	28.63
Other responses*	6542	2.09	5069	2.07	228	2.00	21	2.04
Total	312925		245388		11390		1027	

Percentages may not add to 100 due to rounding.

*The *Other responses count includes any or no response for sex_at_birth. The available options in the CDR are “I prefer not to answer”, “None of these fully describe me”, “Intersex”, “No matching concept”, and “PMI: Skip”. “No matching concept” and “PMI: Skip” are separate counts both referring to no response for sex_at_birth. These are separate because participants in “No matching concept” did select a gender option for this survey question. The terms used here are the Concept Names as they appear in the CDR.

Appendix G: *All of Us* Hereditary Disease Risk genes

The following gene symbols are in the *All of Us* Hereditary Disease Risk (AoUHDR) genes. We have additional srWGS QC criteria in the regions covered by these genes, described in Table 2 of the main text. In the v7 callset, the AoUHDR genes are the same as the American College of Medical Genetics and Genomics' list of 59 genes where incidental findings should be reported (ACMG59) [56]. The AoUHDR gene list may change in future releases.

ACTA2, ACTC1, APC, APOB, ATP7B, BMPR1A, BRCA1, BRCA2, CACNA1S, COL3A1, DSC2, DSG2, DSP, FBN1, GLA, KCNH2, KCNQ1, LDLR, LMNA, MEN1, MLH1, MSH2, MSH6, MUTYH, MYBPC3, MYH11, MYH7, MYL2, MYL3, NF2, OTC, PCSK9, PKP2, PMS2, PRKAG2, PTEN, RB1, RET, RYR1, RYR2, SCN5A, SDHAF2, SDHB, SDHC, SDHD, SMAD3, SMAD4, STK11, TGFBR1, TGFBR2, TMEM43, TNNI3, TNNT2, TP53, TPM1, TSC1, TSC2, VHL, and WT1

Appendix H: DRAGEN invocation parameters

Table H.1 summarizes the parameters used by the GCs to generate GVCFs, contamination estimates, and sex ploidy calls from the DRAGEN for srWGS data.

Table H.1 DRAGEN 3.4.12 parameters run at all GCs

Parameter	Parameter Value	Description
-f	n/a	Overwrite if output exists
-r	<hg38-ref-dir>	The reference to use
--fastq-list	<path-to>/fastq_list.csv	A list of fastq files to use as input for this sample
--fastq-list-sample-id	<sampleID>	The sample ID to use for naming this sample
--output-directory	<output-dir>	The location of the final output files
--intermediate-results-dir	<int-results-dir>	The location to write intermediate outputs
--output-file-prefix	[CenterID]_[Biobankid_Sampleid]_[LocalID:optional]_[Rev#]	Standardized naming prefix for each output file
--enable-variant-caller	TRUE	Turn on variant call outputs
--enable-duplicate-marking	TRUE	Mark duplicate reads during alignment
--enable-map-align	TRUE	Produce an alignment from unaligned read input
--enable-map-align-output	TRUE	Store the output of the alignment
--output-format	CRAM	Store the alignment as a CRAM file
--vc-hard-filter	DRAGENHardQUAL:all:QUAL<5.0;LowDepth:all:DP<=1'	This parameter setting changes the threshold on the quality to 5.
--vc-frd-max-effective-depth	40	Setting this parameter puts a limit on the penalty value that is applied for variant calls that deviate from the expected 50% allele fraction for heterozygous variants.
--qc-cross-cont-vcf	<path-to/SNP_NCBI_GRCh38.vcf>	Marker sites to use for contamination estimation
--qc-coverage-region-1	<path-to/wgs_coverage_regions.bed>	Regions to use for coverage analysis (whole genome)
--qc-coverage-reports-1	cov_report	The type of reports requested for qc-coverage-region-1
--qc-coverage-region-2	<path-to/HDRR_regions.bed>	Regions to use for coverage analysis (HDR reportable regions)
--qc-coverage-reports-2	cov_report	The type of reports requested for qc-coverage-region-2
--qc-coverage-region-3	<path-to/PGx_regions.bed>	Regions to use for coverage analysis

		(PGx reportable regions)
--qc-coverage-reports-3	cov_report	The type of reports requested for qc-coverage-region-3

Appendix I: Samples used in the Sensitivity and Precision Evaluation

In order to calculate the sensitivity and precision of the srWGS SNP and Indel joint callset, we included four well-characterized samples in the v7 callset (Table I.1). We sequenced the NIST reference materials (DNA samples) from Genome in a Bottle (GiaB) and performed variant calling as described in the main text. We used the corresponding published set of variant calls for each sample as the ground truth in our sensitivity and precision calculations [20].

Please note that the control samples do not appear in the data released to researchers.

Table I.1 -- Samples used in sensitivity and precision evaluation

Control Sample	Ground Truth	Genome Center	GVCF origin	Notes
HG-001	GiaB	BI	DRAGEN 3.4.12	NA12878
HG-003	GiaB	UW	DRAGEN 3.4.12	Ashkenazi Trio NA24149 - Father
HG-004	GiaB	BI	DRAGEN 3.4.12	Ashkenazi Trio NA24143 - Mother
HG-005	GiaB	BI	DRAGEN 3.4.12	Han ancestry NA24631- Son

Genome Center:
BI -- Broad Institute
UW -- University of Washington

Appendix J: High quality site determination (srWGS)

In order to do relatedness and ancestry checks, we identified a corpus of sites that can be called accurately in both our ancestry training set (HGDP+1KG) and our target data (*All of Us* srWGS callset). We used a similar methodology that gnomAD used to determine high-quality sites [11]:

1. Autosomal, bi-allelic single nucleotide variants (SNVs) only
2. Allele frequency > 0.1%
3. Call rate > 99%
4. LD-pruned with a cutoff of r2 = 0.1

Our aim was to assemble a set of independent sites where we can be confident of the accuracy.

We identified 150229 high-quality (HQ) sites in the v7 callset. These were HQ sites in both the HGDP+1kg training VCF and the *All of Us* v7 callset. A sites-only VCF of the HQ sites is available in the RW (access required).

Appendix K: Relatedness (srWGS)

We used the Hail pc_relate function to determine the kinship score to report any pairs with a kinship score over 0.1. This analysis was done with the srWGS SNP and Indel data and the lrWGS SNP and Indel data. The kinship score is half of the fraction of the genetic material shared (ranges from 0.0 - 0.5).

- Parent-child or siblings: 0.25
- Identical twins: 0.5

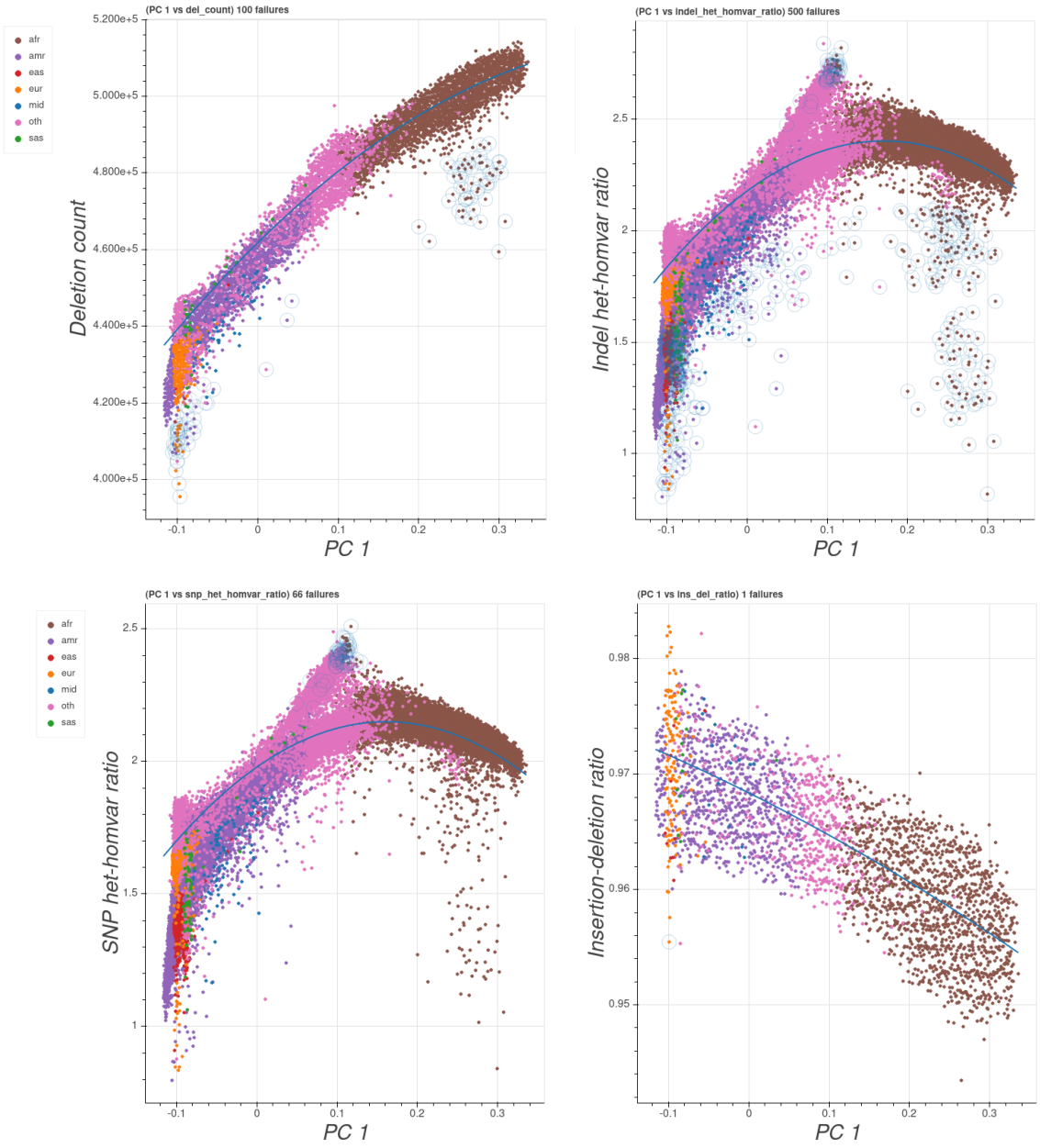
Please see the Hail pc_relate function [13] documentation for more information, including interpretation.

We will determine the maximal independent set [57] for related samples to minimize the number of samples that would need pruning. Using the HQ sites identified in Appendix J, researchers can remove first and second degree relatives.

We estimated 19,374 related pairs and 15,376 samples in the maximal independent set for kinship scores above 0.1. The sample pairs, with kinship score, and the set are available in the RW (access required).

Appendix L: Plots of the first principal component against population outlier QC metrics

Figure L.1 contains the plots of the first principal component against metrics used for determining sample population outliers in srWGS sample QC. Note that we use sixteen principal components for determining which samples should be flagged for being outliers in a metric. The blue line shows the linear regression fit in the first dimension (residuals are calculated as the distance from this hyperplane). The failure count over these plots will sum higher than the 551 flagged samples, since samples can get flagged for multiple criteria. Please see the next page for Figure L.1.


(PC 1 vs del_count) 100 failures
(PC 1 vs indel_het_homvar_ratio) 500 failures
(PC 1 vs snp_het_homvar_ratio) 66 failures
(PC 1 vs ins_del_ratio) 1 failures
afr
amr
eas
eur
mid
oth
sas
Deletion count
Indel het-homvar ratio
SNP het-homvar ratio
Insertion-deletion ratio
PC 1

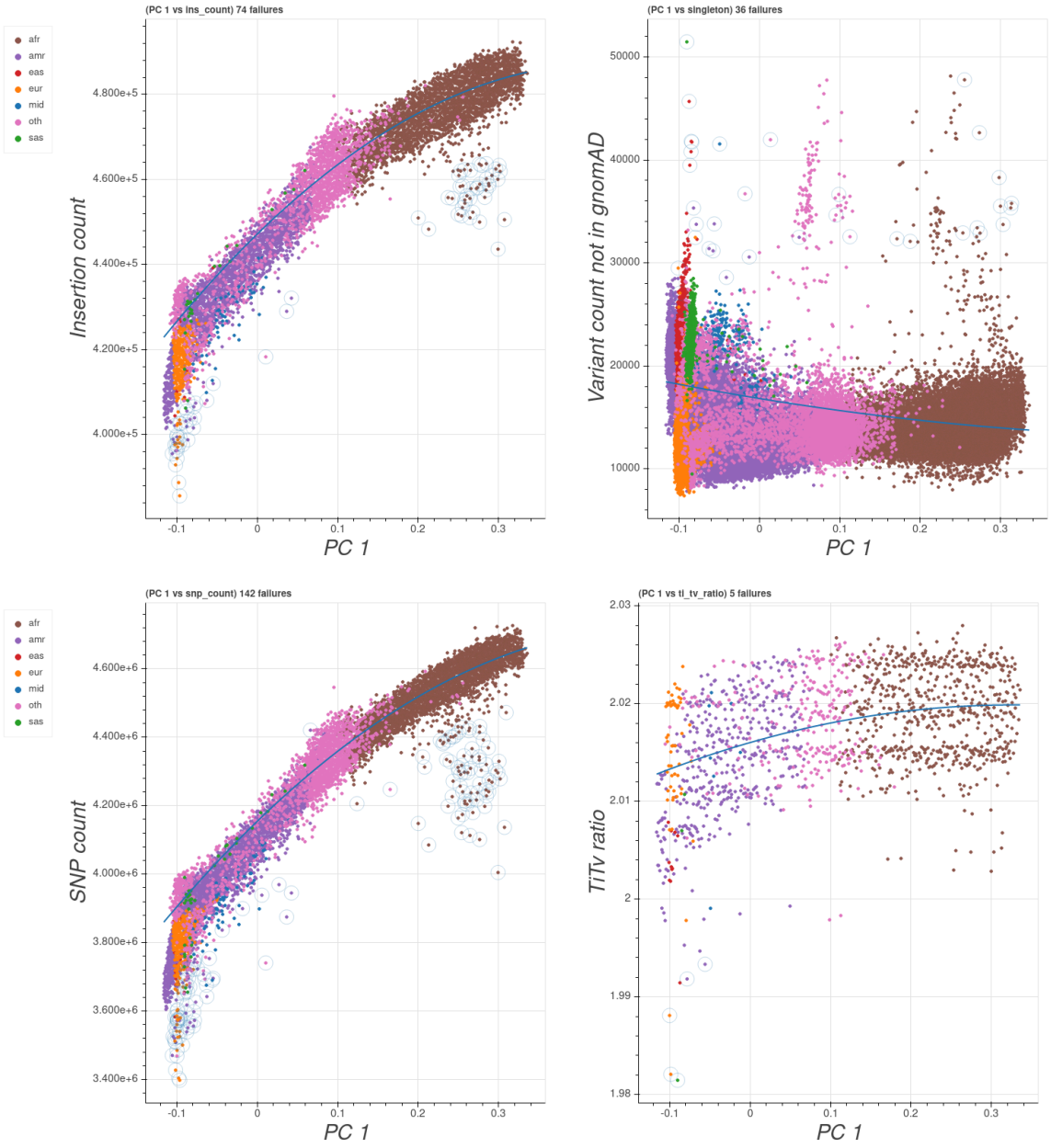


Figure L.1 -- Sample population outlier plots for eight metrics (see Population Outlier Flagging). Each metric (y-axis) is plotted against the first (of sixteen) principal components (x-axis). Outliers are identified by regressing out the principal components and determining if the residual is over 8 MADs from the sample population.

Appendix M: srWGS Structural Variant Pipeline

The GATK-SV pipeline was applied to detect SVs from srWGS data [21]. 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 [23] and Wham [24] to leverage paired-end (PE) and split-read (SR) evidence, MELT [25] to specifically target mobile elements, and GATK-gCNV [58] and cn.MOPS [59] 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 [60], and details of the method can be found in Collins et al 2020 [21]. 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) [61].

Figure 8 depicts the steps of the pipeline as it was run in AoU. Table M.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 M.1-- GATK-SV Pipeline Versions and Notes

Workflow/Step Name	Version Used	Entity	Notes
Sample selection	Notebook		See Sample Selection
GatherSampleEvidence	v0.21-beta	Sample	SV callers used: Manta, Wham, and MELT. All 12,000 samples completed this step, with a 0.74% initial failure rate.
EvidenceQC	v0.21-beta	Arbitrary partition of samples	Run on arbitrary partitions of samples.
Single sample QC	Notebook		See Single Sample QC
Batching	Notebook		See Batching
TrainGCNV	v0.23-beta	Batch	Batches of samples were created according to the scheme described in the main text under Batching

GatherBatchEvidence	v0.23-beta	Batch	Depth-based CNV callers used: GATK g-CNV and cn.MOPS. gVCFs were reblocked to resolve a minor formatting issue prior to this step.
ClusterBatch	v0.21-beta	Batch	Following this step, SV counts per sample were visualized with PlotSVCountsPerSample (v0.21-beta) as a QC checkpoint. No strong outliers were observed, so no samples were removed.
GenerateBatchMetrics	v0.21-beta	Batch	
FilterBatchSites	v0.21-beta	Batch	
PlotSVCountsPerSample	v0.21-beta	Batch	No SV count outliers observed.
FilterBatchSamples	v0.21-beta	Batch	No outlier samples were removed at this stage (nIQR cutoff = 10000).
MergeBatchSites	v0.21-beta	Cohort	For cohort-level steps, data from all samples across all batches was merged.
GenotypeBatch	v0.24-beta	Batch	
RegenotypeCNVs	v0.24-beta	Cohort	
CombineBatches	v0.24-beta	Cohort	
ResolveComplexVariants	v0.24-beta	Cohort	
GenotypeComplexVariants	v0.24.1-beta	Cohort	
CleanVcf	v0.24-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 5265fec)	Cohort	A developmental version of AnnotateVcf was used for improved scaling
MainVcfQc	v0.26.9-beta	Cohort	

Appendix N: Overall precision and recall after SL and NCR filtering

Table N.1 summarizes performance after SL and NCR filtering across SV classes. Overall recall/precision were 0.820/0.940 in the training set and 0.817/0.920 in the test set with similar performance observed across the spectrum of SV classes. These results indicate that the model will generalize accurately to unseen data.

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

Filtering class	Min size (bp)	Max size (bp)	SL cutoff	Train		Test	
				Recall	Precision	Recall	Precision
Small DEL	50	500	-37	0.809	0.973	0.805	0.963
Medium DEL	500	10,000	-8	0.856	0.979	0.821	0.967
Large DEL	10,000	inf	-60	0.997	0.975	NA*	NA*
Small DUP	50	500	-74	0.840	0.870	0.864	0.813
Medium DUP	500	10,000	-47	0.598	0.815	0.712	0.770
Large DUP	10,000	inf	-99	1.00	0.999	NA*	NA*
INS	50	inf	-45	0.813	0.936	0.813	0.907
INV	50	inf	-36	0.949	0.989	0.763	0.961
BND**	NA	NA	-48	NA	NA	NA	NA

*Large DEL and DUP variants were tested in a separate analysis. The results will be reported in the supplementary SV QC doc, Benchmarking and quality analyses on the All of Us v7 short read structural variant calls, which can be found on the User Support Hub [1].
**BNDs lacked training data, so the SL cutoff for BNDs was set as the optimal value for all training variants across SV types.

Appendix O: Long Read Workflow Diagrams

The following figures summarize the workflows utilized to process the AoU Long Read samples. A standard sample will be processed in the following sequence of workflows:

1) Preprocessing CCSed SMRT cells (HiFi reads extraction, demultiplex if applicable) [Figure N.1]
2) Post-processing SMRT cells (alignment, QC controls at atomic unit level) [Figure N.2]
3) Whole Genome Variant Calling (data aggregation, de novo assembly, and variant calling) [Figures N.3 and N.4]

The workflows are available as Workflow Definition Language (WDL) files in a public github repository [62].

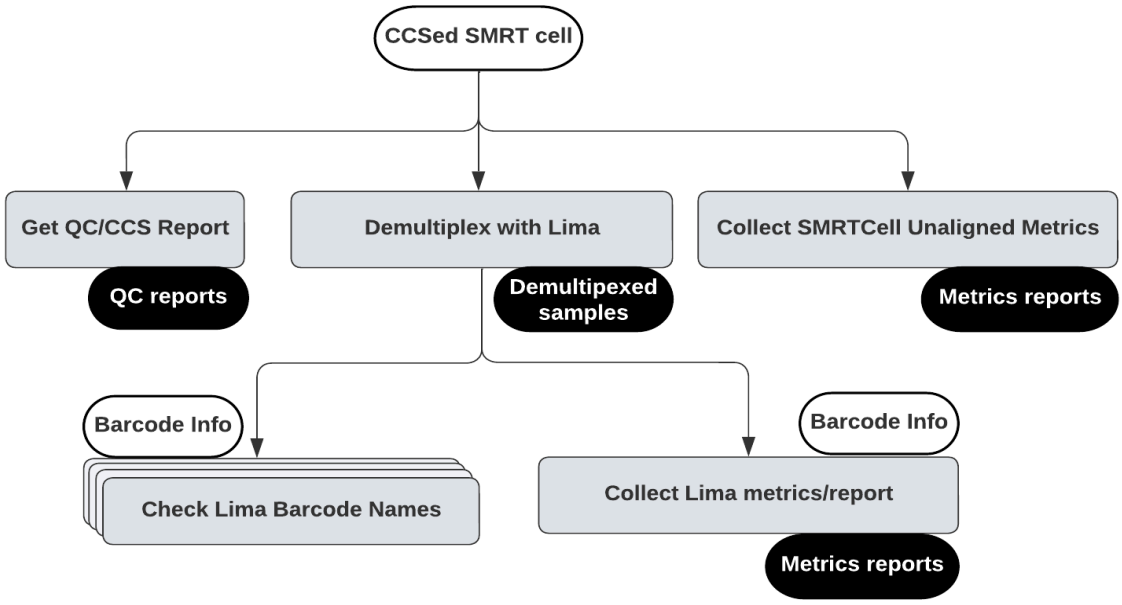


Figure N.1 Preprocessing CCSed SMRT cell. A workflow for preprocessing barcoded (potentially multiplexed) SMRTCell. The cell is assumed to be CCSed on-instrument, and the whole data folder is mirrored onto a cloud bucket.

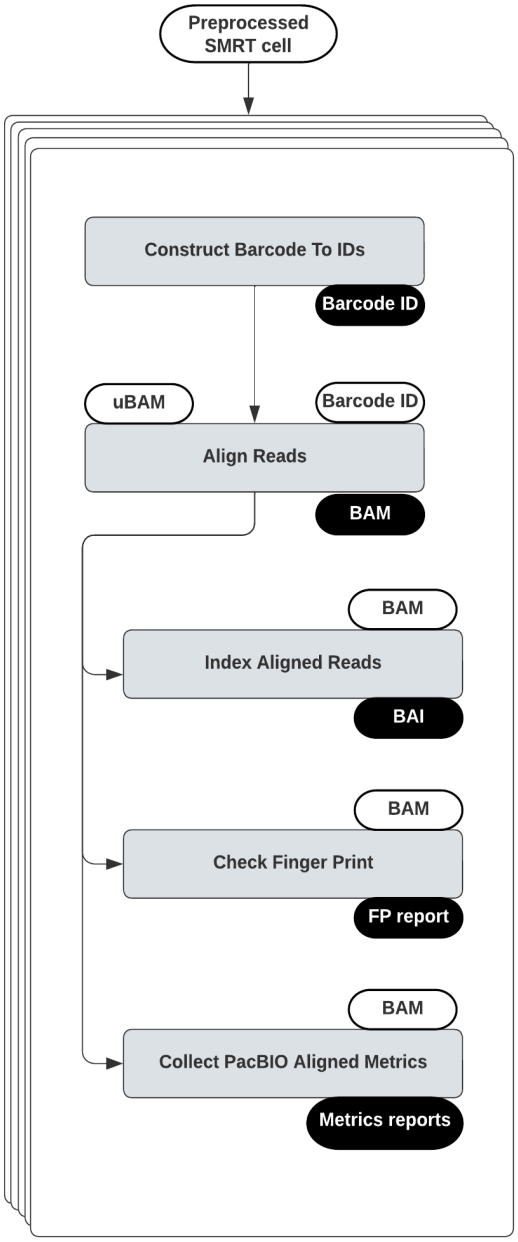


Figure N.2 Post-processing demultiplexed SMRTCell (CCS already performed). The various IDs are assumed to be in-phase with the barcode names. Alignment to the two references (grch38_noalt and T2Tv2.0) are done independently and only the grch38_noalt version goes through the three sample SMRT cell (pre-sample-aggregation) QC processes.

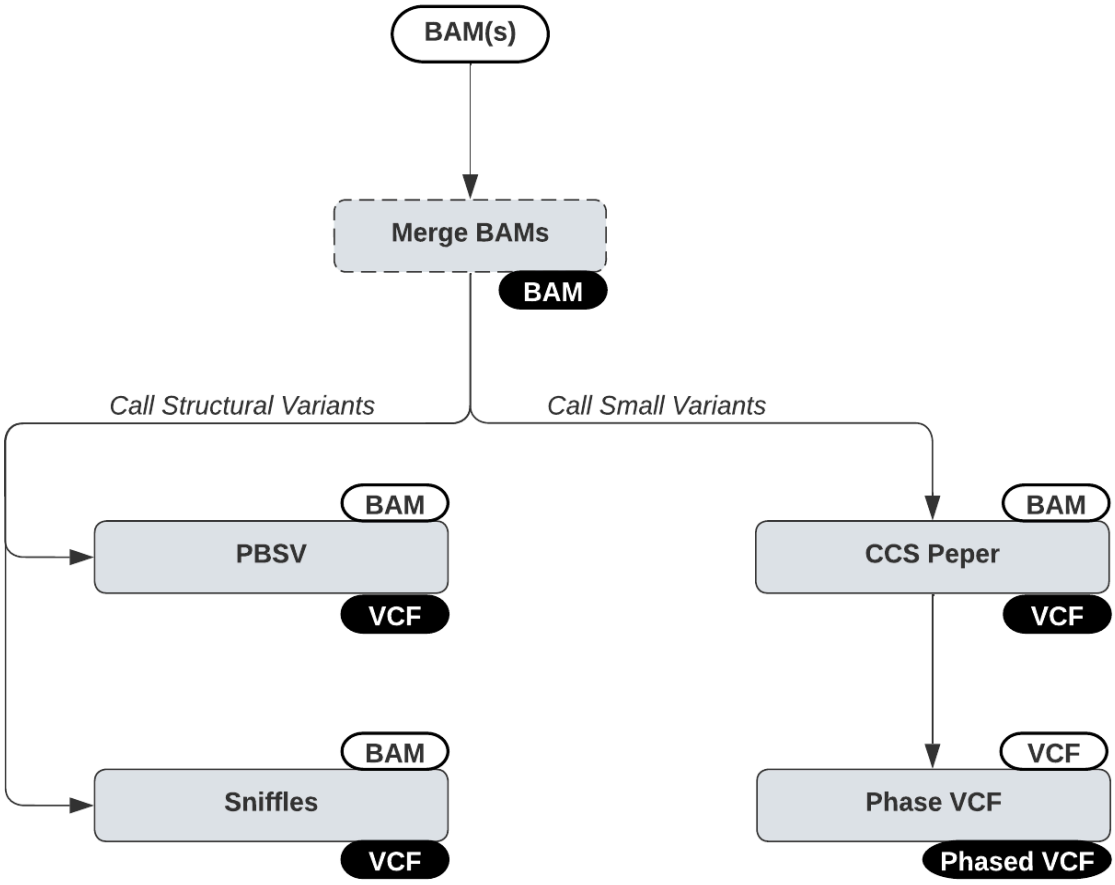


Figure N.3 Single sample whole genome variant calling. The workflow merges information from each sample's SMRT cell into a single BAM prior to variant calling. We run the entire workflow separately for each reference.

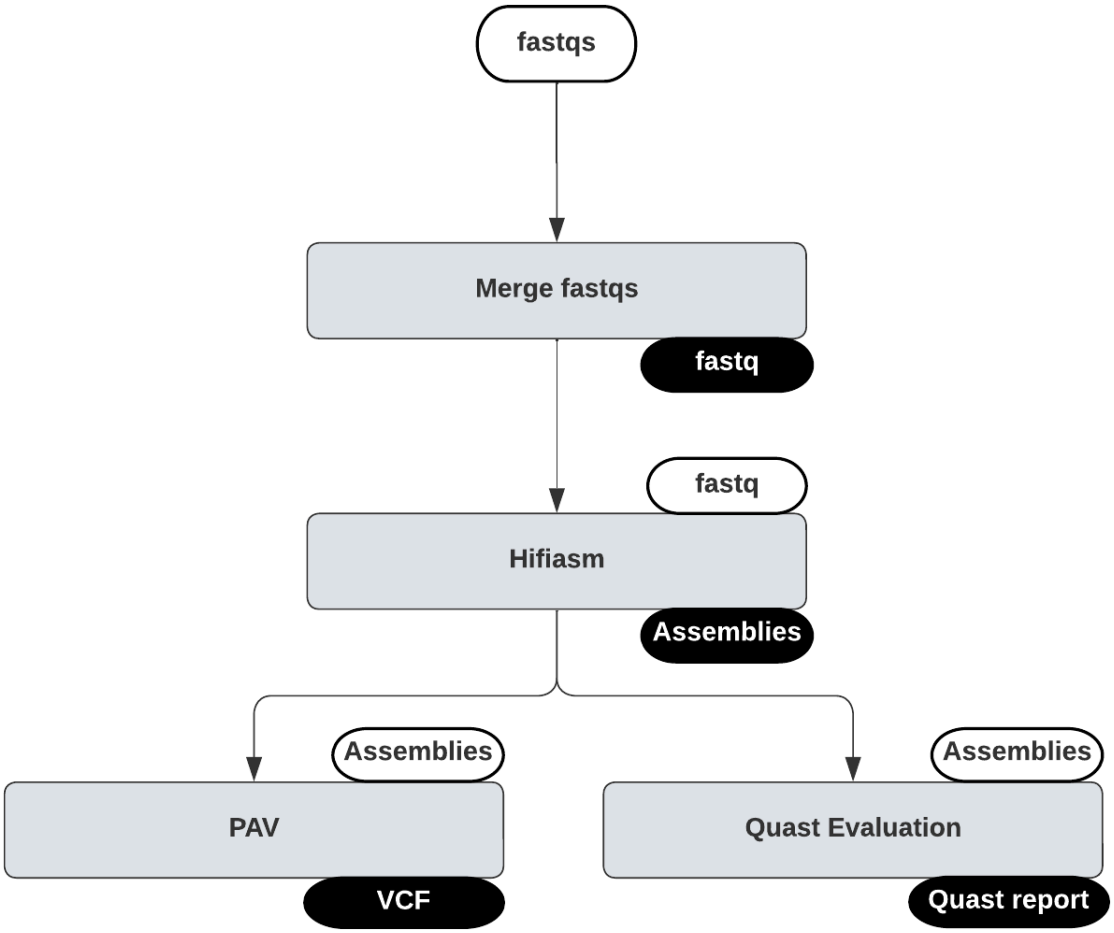


Figure N.4 HiFi FASTQ files for a sample are merged before sent for de novo assembly with Hifiasm. The haplotype-resolved assembly is then evaluated with QUAST. This assembly will then be used by PAV for calling variants.

Appendix P: lrWGS analysis versions and parameters

Table P.1 – lrWGS pipeline software versions and parameters

Software	Version used	Functionality	Invocation parameters
`extracthifi`	1.0.0	Extracting HiFi reads from CCS bam.	`extracthifi <unaligned.ccs.bam> <unaligned.hifi.bam>`
`lima`	2.6.0	SMRT cell demultiplexing.	`lima --dump-removed --split-bam-named --hifi-preset SYMMETRIC <unaligned.hifi.bam> Sequel_16_barcodes_v3.fasta <demux.bam>`
`pbmm2`	1.4.0	HiFi reads alignment.	`pbmm2 align <unaligned.hifi.bam><reference.fasta> --preset CCS --sample <sample_name> --strip --sort --unmapped`

`samtools`	1.13	BAM aggregation and conversion to FASTQ.	Aggregation
			`samtools merge -p -c --no-PG -@ 2 --write-index -o <agg.bam> <input.bam>[,<input.bam>,...]`
			
			Conversion
			`samtools fastq <input.bam> <output.fastq>`

`Hifiasm`	0.16.1	*de novo* assembly.	Primary and alt assembly
			`hifiasm -o <output_prefix> -t <cpu_cores_to_use> –primary <input.fastq>[,<input.fastq>,...]`
			
			Haplotype resolved assembly
			`hifiasm -o <output_prefix> -t <cpu_cores_to_use> <input.fastq>[,<input.fastq>,...]`

`pbsv`	2.6.0=h9ee0642_0	Single sample SV calling per chromosome. After this step, chromosomes are merged.	`pbsv discover --tandem-repeats <trf.bed> <aligned.bam> <output.svsig.gz>`
			
			`pbsv call –ccs <reference.fasta> <input.svsig.gz> <output.vcf>`

`sniffles2`	2.0.6	Single sample SV calling	`sniffles -i <input.bam> --minsvlen 50 --sample-id <sample_id> --vcf <output.vcf> --snf <output.snf>`

`pav` (the tool)	Branch aou with hash `fa43453` in repo https://github.com/EichlerLab/pav	The specific `pav` docker that we ran	
`pav` (WDL pipeline)		Single sample SV and small variant calling from assembly	`pav` pipeline at https://github.com/broadinstitute/pav-wdl/tree/sh_more_resources_pete It is currently in development. We ran the pipeline in the state that is documented in the git commit hash `5558ebdbd0be3f2eb722b10774a1e305a20fa569`

`Pepper`	Docker `kishwars/pepper_deepvariant:r0.4.1`	Prepping BAM for small variant calling	`run_pepper_margin_deepvariant \`
			`  call_variant \`
			`  -b <input.bam> \`
			`  -f <reference_fasta> \`
			`  -s <sample_name> \`
			`  -o <output_dir> \`
			`  -p <output_prefix> \`
			`  --phased_output \`
			`  --ccs`

`DeepVariant`	1.3.0	Single sample SNP and Indel variant calling	`/opt/deepvariant/bin/run_deepvariant \`
			`  --model_type=PACBIO \`
			`  --ref=<reference_fasta> \`
			
			`--reads=<pepper.prepared.input.bam> \`
			`  --output_vcf=<output.vcf.gz> \`
			
			`--output_gvcf=<output.g.vcf.gz> \`
			`  --use_hp_information`

`Margin`	Docker `kishwars/pepper_deepvariant:r0.4.1`	Single sample SNP and Indel phasing	`margin phase \`
			`  <input.haplotagged.bam> \`
			`  <reference_fasta> \`
			`  <unphased_vcf> \`
			
			`/opt/margin_dir/params/misc/allParams.phase_vcf.json \`
			`  -M \`
			`  -o <output_dir>/<prefix>`