All of Us Research Program

Genomic Research Data Quality Report

All of Us Curated Data Repository (CDR) release C2024Q3R3

Overview

This document details the *All of Us* Genome Centers (GC) and Data and Research Center (DRC) quality control (QC) steps for the genomic data made available in the Researcher Workbench February 3, 2025 in the CDRv8 data release. 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 researchers. We, the *All Of Us* DRC, only describe QC processes that are performed analytically (i.e., after the sample has been sequenced).

The samples in the genomic data correspond to the *All of Us* Curated Data Repository (CDR) release C2024Q3R3 ("CDRv8"). All descriptions and results are limited to the CDRv8 data, which contains 447,278 genotyping array ("array") samples, 414,830 short-read whole genome sequencing (srWGS) samples with single nucleotide polymorphism, insertion, and deletion variant calls (SNPs and Indels), 97,061 srWGS samples with structural variant (SV) calls, and 2,800 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 (25 (<0.01%) exceptions exist, see Known Issue #2).

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:

- We have received an exception to the Data and Statistics Dissemination Policy from the *All of Us* Resource Access Board for the contents of this report.
- 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' and descriptions of the file formats for the genomic data are available in the 'How the All of Us Genomic data are organized', both published on the User Support Hub [1].
- 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/
- 22 lrWGS samples are missing their corresponding phenotypic and Cohort Builder data and thus the sample counts are not the same. Please see Known Issue #2 for more details.

Executive Summary

On February 3, 2025, the *All of Us* Research Program released the genomic data of 447,278 array samples, 414,830 srWGS samples with SNP & Indels, 97,061 srWGS samples with SV calls, and 2,800 lrWGS samples in the Researcher Workbench (RW) for use by researchers registered for Controlled Tier access. As described previously [2], this high-quality genetic data along with comprehensive health data will enable health research and catalog the genetic variation that leads to human health and disease. For a snapshot of the data, see Table 1.

Table 1 -- Snapshot of *All of Us* CDRv8 genomic dataset

Dataset	Number of participants	Number of variants	Highlights
Array	447,278	More than 1.8 million	- We now have Array data from participants that self-identify as American Indian or Alaska Native
Short-read WGS SNP and Indel	414,830	More than 1.2 billion	- We added ~150,000 new participants in CDRv8 which resulted in over 200 million new variants - We now have srWGS data from participants that self-identify as American Indian or Alaska Native
Short-read WGS structural variants (SVs)	97,061	Nearly 1.5 million	
Long-read WGS	2,800	11 cohorts with SNPs, Indels, and structural variants	- Variants are called to both the grch38_noalt and T2Tv2.0 references

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 (variant annotations, pharmacogenomics, genetic ancestry categories, genetic ancestry admixture estimates, and relatedness/kinship scores) 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 [2]. As the program continues, the DRC will periodically release genomic data - in sync with planned CDR release timelines. This document describes the CDRv8 release of genomic data to *All of Us* researchers made available in the RW on February 3, 2025. The genomic data contains 447,278 array samples, 414,830 srWGS samples, 97,061 srWGS samples with SV calls, and 2,800 lrWGS samples which can be joined with other data types (e.g. survey data) for analysis, though please see Known Issue #2. In this document, we describe the QC processes applied to the array, srWGS, and lrWGS data.

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 (See Appendix A for the GCs and DRC locations).
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 that are outliers or do not meet thresholds. 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.

Arrays

There are 447,278 array samples in the v8 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. Please see Appendix B for details.

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 prior to the CDRv7 release 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.

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 2. 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 B.

Table 2 -- Array Single Sample QC processes

QC process	Passing criteria	Error modes addressed	v8 release results
Sex concordance	Sex call is concordant with self-reported sex at birth.	-Sample swaps	All array samples are concordant.

	OR Self-reported sex at birth reported as “Other” or was not reported		*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 C). 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 [3] 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 [4], 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 C.

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-	-0.1	This effectively disables the sex

threshold		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 v8 release passed a sex concordance check. Note that 1.06% 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 C 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 [3] 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 v8 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.

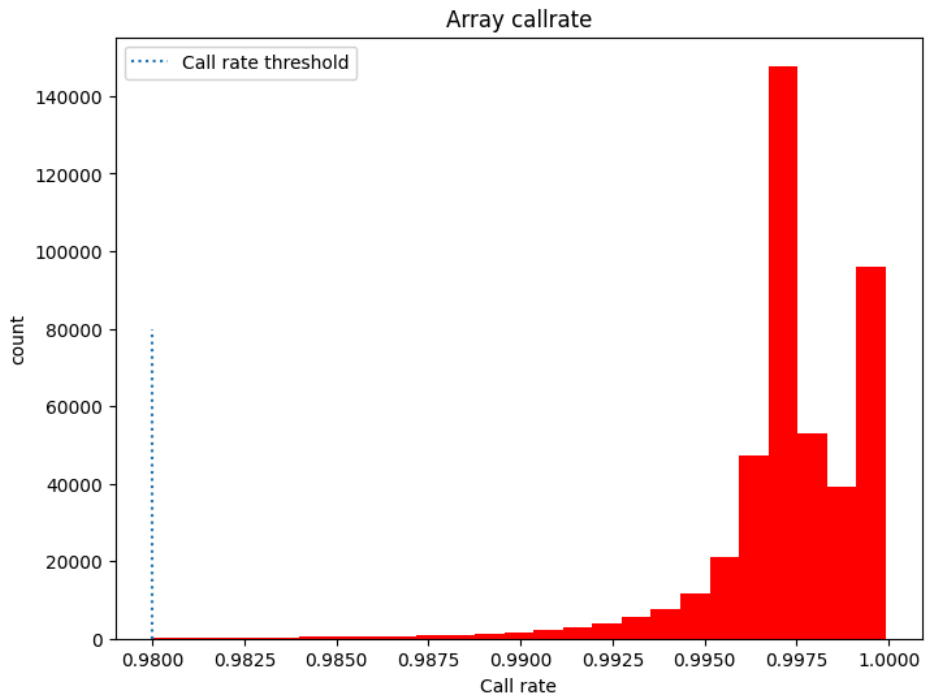


Figure 1 -- Histogram of the array call rate for the v8 release.

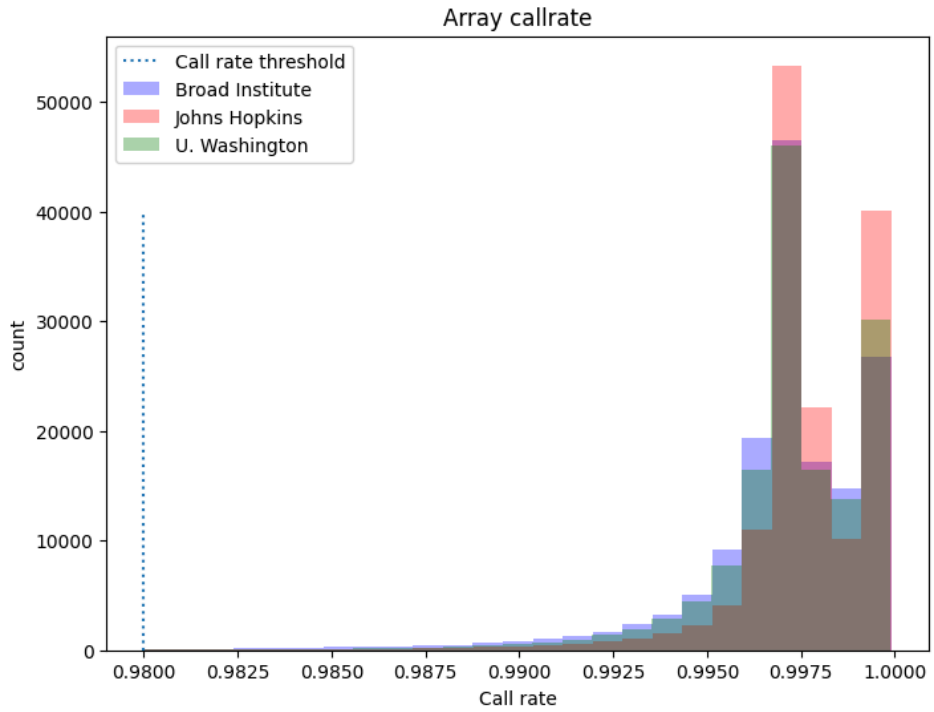


Figure 2 -- Call rate across each GC.

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 [5] to estimate the contamination rate in our array data. We do not use the cross-individual contamination rate to filter array samples, but we do not process the corresponding srWGS 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 0.12 for all array samples. As the contamination rate increased, we did see a small decrease in the call rate (see Figure 3). Of the 447,278 array samples, 443,937 (99.3%) had an estimated contamination rate below 3% and 436,332 (97.7%) had a contamination rate less than 1%.

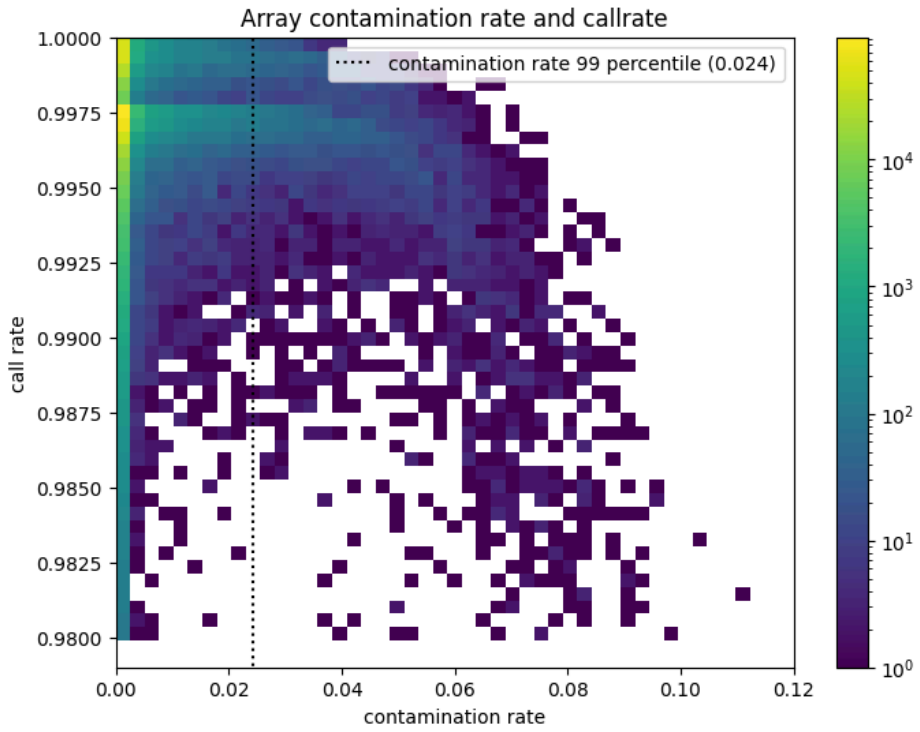


Figure 3 -- Histogram of the array contamination rate estimates vs call rate. As the contamination rate increases, the call rate decreases.

Short-Read Whole Genome Sequencing (srWGS)

The *All of Us* srWGS dataset is a high-quality comprehensive dataset of 414,830 participants [2], available as raw reads, variant data, and annotated variants. Please read the article 'How the *All of Us* Genomic Data are Organized' for more information about the srWGS data available.

Consistency across Genome Centers

The GCs use the same protocol for library construction (PCR Free Kapa HyperPrep), sequencer (NovaSeq 6000), software (DRAGEN v3.7.8), and software configuration.

The srWGS CDRv8 samples were processed on DRAGEN 3.7.8. For samples that were originally processed for previous releases on DRAGEN 3.3.12, they were reprocessed from DRAGEN 3.4.12 to 3.7.8. Some reprocessed samples had new coverage metrics, which caused them to fall below the coverage threshold and were dropped (See Known Issue #2). 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 [6] 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 3. Our srWGS single sample QC uses the same sequencing process described previously [2] [6] 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 [6], except for a higher threshold for contamination.

In some cases, we perform these tests at both the DRC and the GCs 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. There are some upstream processes not described here because in this document, we are focused on downstream analytical QC processes after a sample has been sequenced. The list of specific QC processes and an overview of the results can be found in Table 3.

Table 3 -- srWGS Single Sample QC processes

QC process	Calculated at the DRC or GCs?	Passing criteria	Error modes addressed	CDRv8 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 D for more information	-Sample preparation error -Poor sensitivity and precision of variant calling	All srWGS samples meet the thresholds. For the CDRv8 release, all samples were reprocessed from DRAGEN 3.4.12 to 3.7.8 (Appendix E). [Update as of June 2025] Please see Known Issue #10, as we have determined samples that failed coverage metrics were not removed.

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 [7].

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 CDRv8 release passed the fingerprint concordance check based on arrays. As seen in Figure 4, the passing samples exceeded the threshold. 8920 samples had a fingerprint LOD [7] less than 45 and the minimum fingerprint LOD was 38.

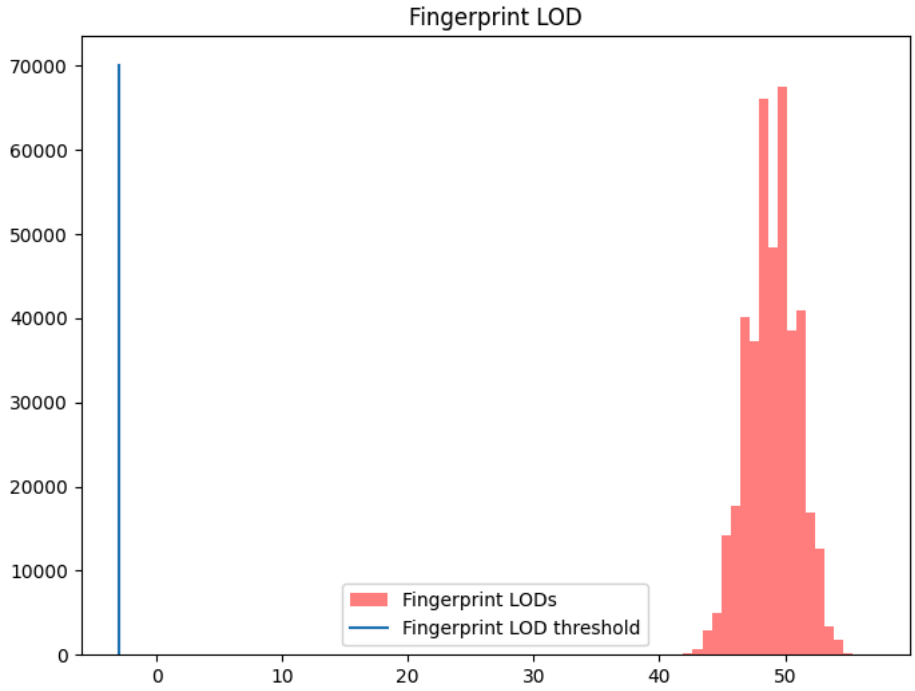


Figure 4 -- Distribution of the Fingerprint LODs for srWGS CDRv8 samples

Sex Concordance

For srWGS data, we compared the computed sex from DRAGEN (Appendix E) and peddy [8] against the self-reported sex assigned at birth (Appendix C). 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 3.7.8, Appendix E) and use heterozygous chrX variant calls from peddy [8]. If the concordance test fails against either of these calls, the sample fails QC and is not included in the release. If the DRAGEN ploidy is not XY or XX, we pass the sample. 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 C.

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

The DRAGEN pipeline outputs a single-sample VCF, which is primarily used in the clinical pipeline (for individual samples)[6], but we use it as input to the peddy tool, with the following parameters. We run peddy in single-sample mode so we do not use pedigree information with relatedness for multiple samples.

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.

Results

We do not include any srWGS samples that fail the sex concordance check in the released samples. Please 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 (1.06% of srWGS samples). Appendix C 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 [9], and the DRAGEN 3.7.8 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	srWGS 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 E 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 [6].

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 CDRv8 release.

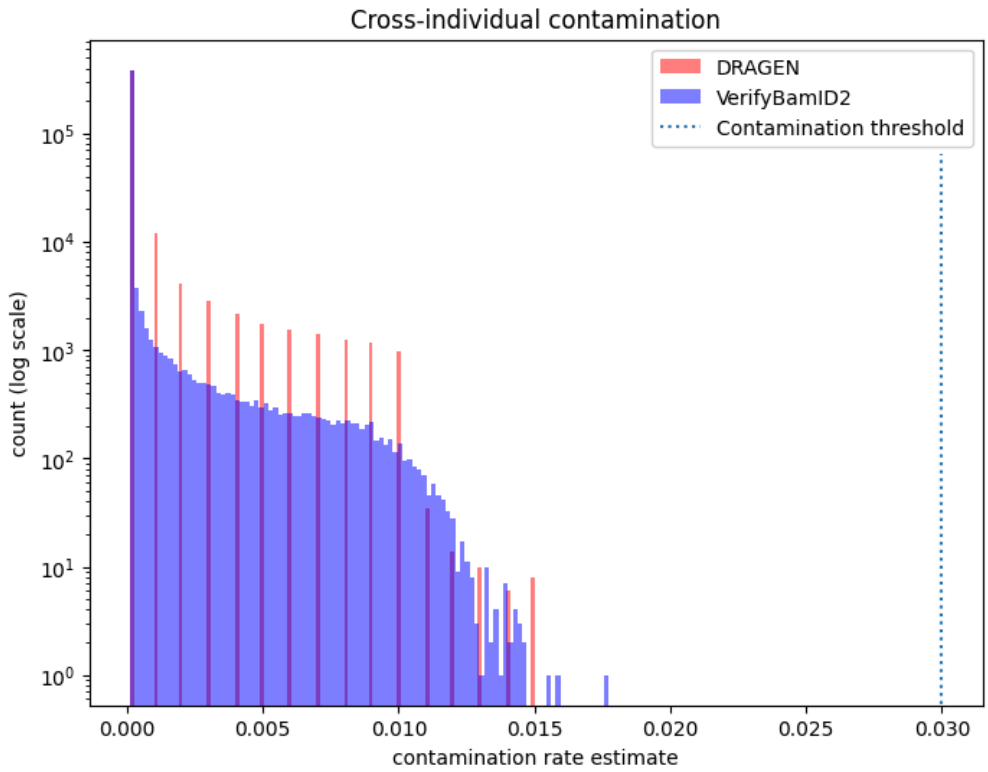


Figure 5 -- srWGS contamination estimates from both sources (DRAGEN and VerifyBamID2). DRAGEN rounds the contamination estimate to three decimal places. Note the log scale of the counts (y-axis). Over 90.3% and 92.0% of srWGS samples had contamination estimates lower than 1e-4 by VerifyBamID2 and DRAGEN, 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) [10]. As lower base quality counts increase, we see a reduction in accuracy with an inflection point starting around 6e10.

Results

All srWGS v8 samples were reprocessed from DRAGEN 3.4.12 to 3.7.8, which affected sample coverage metrics. Some samples had new coverage metrics. Samples that fell below the mean coverage threshold were excluded from the callset (see Known Issue #2). As of June 2025, we have identified a Known Issue affecting the sample coverage metrics (Known Issue #10). 4,044 samples that fell below the coverage thresholds were included in the CDRv8 release and should have been removed.

As seen in Figure 6, we had 395 (0.1%) samples with mean coverage greater than 70x.

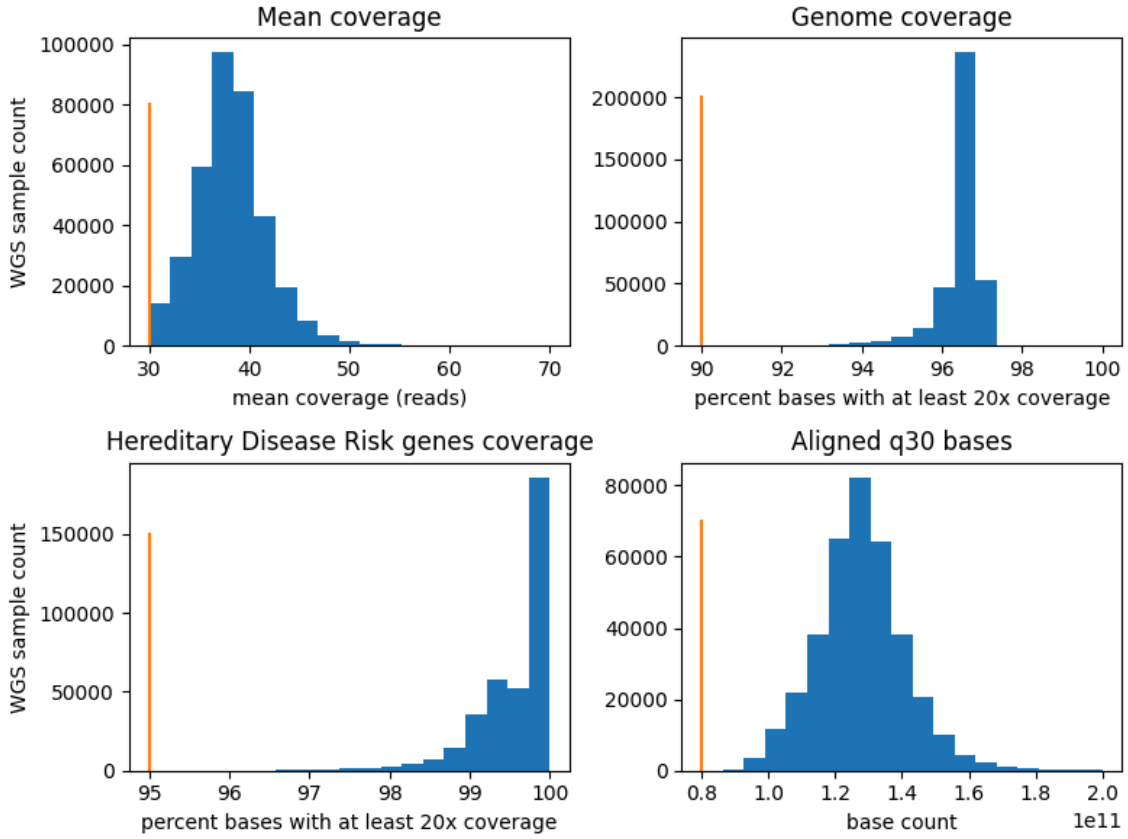


Figure 6 -- Coverage metrics for the CDRv8 release srWGS samples. The orange line is the threshold for each metric. There are 395 samples (0.1%), with mean coverage greater than 70x, that are not included in the mean coverage (upper left) nor aligned q30 bases (lower right) plots. As expected, these 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 [11]. Please note that the QC steps described here apply during creation of the srWGS joint callset, after single sample QC. Sample QC is performed before variant QC. The joint callset QC process is similar to that of gnomAD 3.1 [12], though not exactly the same. See a summary of the joint callset QC steps in Table 4.

We flag samples or variants as failing QC, rather than removing them from the callset, since we cannot validate whether samples (especially population outliers) are problematic or are just a part of a poorly-sampled ancestry. Flagged variants can also be a result of poorly-sampled ancestry.

Table 4 -- srWGS SNP & Indel joint callset QC summary

QC process	Sample or variant QC	Error modes addressed	CDRv8 release results
Sample Hard Threshold Flag	sample	Extremely noisy samples	No samples flagged.
Sample Population Outlier Flag	sample	Noisy samples	987 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. 67,694,029 were filtered 1,192,874,611 were not filtered
Variant Extract-Train-Score Filtering (VETS)	variant	Artifacts that cannot be detected in a single sample	See [13]
Sensitivity and Precision Evaluation	both	Poor variant detection	See Appendix F 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): 0.001 Concordance vs self-reported: 0.884 See Appendix G. Number of independent, bi-allelic sites (“high-quality sites”) used: 130660 See Appendix H.
Relatedness and maximal independent set of samples	sample	Related samples, which confound analyses	39,682 related pairs and 30,585 samples in the maximal independent set. See Appendix I. This process produces a list of the sample pairs with kinship score, calculated by Hail [14]. 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 the RW, listed in the Controlled CDR directory document.

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

All samples that failed these hard thresholds also failed the sample population outlier flag. See the section below for the results.

Sample Population Outlier Flag

We flag srWGS individual samples based on the population outlier data. The flagged samples can be found in the RW and Genomic QC metrics used in the joint-callset QC are available for all samples. Locations for where to find these files are in the Controlled CDR directory document.

Method

As part of ancestry prediction (see Appendix G), we regressed out sixteen principal component features computed and used the residuals to determine the outliers. We calculated sample features using the gnomAD QC methods compute_stratified_metrics_filter and compute_qc_metrics_residuals with version 0.5.0.

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
    - Del count
ii. number of insertions
    - Ins count
iii. number of SNPs
    - SNP count

iv. number of variants not present in gnomAD 3.1
    - Not in gnomAD
v. insertion : deletion ratio
    - Ins/Del ratio
vi. transition : transversion ratio
    - Ti/Tv ratio
vii. SNP heterozygous to homozygous ratio
    - SNP Het/Hom
viii. Indel heterozygous to homozygous ratio
    - Indel Het/Hom

Results

We flagged 987 (0.2%) samples as outliers based on at least one of the above criteria (See Table 5). Plots of the first principal components against these eight metrics can be found in Appendix J.

Table 5 -- srWGS SNP & Indel population outlier sample counts

Metric(s) considered	Flagged sample count
Indel Het/Hom	477
Del count + Indel Het/Hom + Ins count + SNP count	123
Not in gnomAD	103
Indel Het/Hom + SNP Het/Hom	95
Indel Het/Hom + SNP count	68
Del count + Indel Het/Hom + SNP count	45
SNP Het/Hom	27
Ti/Tv ratio	19
Ti/Tv ratio + Not in gnomAD	8
Del count + Ins count + SNP count + Ti/Tv ratio + Not in gnomAD	3
Del count + Ins count + SNP count + SNP Het/Hom + Not in gnomAD	3
Del count + Ins count + SNP count + Not in gnomAD	3
Del count + Ins count + SNP count + SNP Het/Hom + Ti/Tv ratio + Not in gnomAD	2
Del + Indel Het/Hom + Ins/Del ratio + Ins count + SNP count + SNP Het/Hom + Ti/Tv ratio + Not in gnomAD	2
SNP count + Not in gnomAD	2
SNP count + Ti/Tv ratio + Not in gnomAD	1
Del count + SNP count + Not in gnomAD	1
Del count	1

Del count + Indel Het/Hom ratio + Ins count + SNP count + SNP Het/Hom + Ti/Tv ratio + Not in gnomAD	1
Indel Het/Hom + Not in gnomAD	1
Indel Het/Hom + SNP count + Not in gnomAD	1
Indel Het/Hom + SNP count + SNP Het/Hom	1

Total	987

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, and `filters` in the Hail MT). 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`
- If a site has more than 100 alternate alleles
    - We count the alternate alleles at each site and filter out sites with more than 100 alternate alleles
    - Filter field value: `EXCESS_ALLELES`

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 6.

Table 6 -- srWGS SNP & Indel variant hard threshold filter counts

Filters	Variant Count
'EXCESS_ALLELES'	89,853 (<0.01%)
'EXCESS_ALLELES', 'ExcessHet'	7,981 (<0.01%)
'EXCESS_ALLELES', 'ExcessHet', 'NO_HQ_GENOTYPES'	1 (<0.01%)
'EXCESS_ALLELES', 'NO_HQ_GENOTYPES'	1,004 (<0.01%)
'ExcessHet'	608,593 (0.05%)
'NO_HQ_GENOTYPES', 'ExcessHet'	429 (<0.01%)
'LowQual'	3,553,875 (0.28%)
'NO_HQ_GENOTYPES', 'LowQual'	24,983,788 (1.98%)
'NO_HQ_GENOTYPES'	38,448,505 (3.05%)
Total variants	1,260,586,640
Total variants filtered	67,694,029 (5.37%)
Total not filtered	**1,192,874,611**

Variant Extract-Train-Score Filtering (VETS)

We flag variants using the Variant Extract-Train-Score (VETS) method, which is a genotype-level filtering algorithm. At some sites only some genotypes are filtered whereas at other sites all genotypes are filtered. We do not report the variant score directly, only if the variant is filtered.

VETS implements the same basic algorithm as VQSR, which was used in previous *All of Us* srWGS SNP and Indel datasets, and so we do not expect researchers to see a change in the dataset.

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 VETS 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. Though please see Known Issue #3 about a small number of variants missing from the Cohort Builder Variant Search.

In all datasets, we report the filtering status in the genotype filters (FT) field. In the VDS, `FT` will contain `True` for PASS and `False` for FAIL. In the Hail MT, FT will contain PASS or FAIL. In the VCF, a filtered genotype will be annotated with high_CALIBRATION_SENSITIVITY_SNP or high_CALIBRATION_SENSITIVITY_INDEL. All variants will still be included in cohorts, including in the Cohort Builder. Though please see Known Issue #3 about a small number of variants missing from the Cohort Builder Variant Search.

Method

The VETS algorithm uses an isolation-forest outlier detection model 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`)
- 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 [15] and Table 7. 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 [15], and the sites can be found as public resource downloads for the GATK [16].

Table 7 – srWGS SNP and Indel VETS training and truth datasets

Training Set Name	SNP or Indel	Truth	Training	Prior Likelihood	Description
Omni [17]	SNP	True	True	Q12 (93.69%)	This resource is a set of polymorphic SNP sites produced by the Omni genotyping array.
HapMap [18]	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 [19]	SNP	False	True	Q10 (90%)	This resource is a set of high-confidence SNP sites produced by the 1000 Genomes Project.
Mills [20]	Indel	True	True	Q12 (93.69%)	This resource is an Indel callset that has been validated to a high degree of confidence.
Axiom [19]	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 eight well-characterized Genomes-in-a-Bottle (GiaB) control samples from HapMap [18] and Personal Genome Project; (see Appendix F), which we can use to determine sensitivity and precision [21] . The samples were sequenced with the same protocol as the *All of Us* samples. These control samples are available to researchers in GVCF format on 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 8.

Table 8 -- Sensitivity and precision measurements for control samples using the *All of Us* sequencing protocol

Variant type	Sample	Sensitivity	Precision
SNV	HG-001_A	0.989	>0.999
	HG-001_B	0.989	>0.999
	HG-002_A	0.986	>0.999
	HG-002_B	0.986	>0.999
	HG-003_A	0.985	>0.999
	HG-003_B	0.986	>0.999
	HG-004	0.986	>0.999
	HG-005	0.985	>0.999
Indel	HG-001_A	0.983	0.998
	HG-001_B	0.983	0.998
	HG-002_A	0.987	0.999
	HG-002_B	0.987	0.999
	HG-003_A	0.988	0.998
	HG-003_B	0.989	0.999

	HG-004	0.989	0.999
	HG-005	0.990	0.999

srWGS Structural Variant (SV) Callset

The srWGS SV callset represents 97,061 participants with SVs called from srWGS data. All participants with srWGS SV calls are within the srWGS SNP and Indel dataset. Prior to SV calling, all samples followed the Consistency across Genome Centers and Single Sample QC processes in the srWGS QC pipeline.

We used GATK-SV to call SVs, which has been previously described [22]. Further technical information can be found in Appendix L. GATK-SV discovers SVs of the following types: deletion (DEL) and duplication (DUP), which can together be described as copy number variants (CNV); insertion (INS); inversion (INV); translocation (CTX); complex event (CPX); unresolved breakend (BND); and multiallelic CNV (we refer to them as MCNV in this document but their SV type in the VCF is CNV). See [23] for additional information on SV types and their evidence signatures.

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

We have also performed data validation experiments and benchmarking and the results are shown in other, upcoming documentation (see the Benchmarking and quality analyses on the All of Us short read structural variant calls).

The dataset is a refresh of the CDRv7 off-cycle srWGS SV dataset, where we released srWGS SV data for 97,940 samples. To prepare the dataset for the CDRv8 release, we did not redo variant calling. We removed any samples that were dropped between releases and performed extra steps to refine the callset, described at the end of this report in CDRv8 updates. Importantly, this means that the CDRv8 srWGS SVs were called from CRAMs aligned with DRAGEN version 3.4.12, which is different from the other data derived from srWGS in CDRv8, which used DRAGEN version 3.7.8 for alignment.

The documentation of SV calling methods and QC processes used to generate the CDRv7 off-cycle srWGS SV dataset is included in this document for convenience. However, these processes were not performed again for the CDRv8 release. Refer to the CDRv8 Updates section for details of the changes since the CDRv7 off-cycle release.

Sample Selection for srWGS SVs

We initially selected 100,321 samples from participants who had srWGS data in the Controlled Tier CDRv6 (C2022Q2R2) dataset or participants who have been selected for previous or future long-read sequencing. Of these initially selected samples, we excluded 3,260 (3.25%) from the final callset (Table 9). Of these 3,260, some were removed between the CDRv6 and CDRv7, and some were removed between CDRv7 and CDRv8 (e.g., participant withdrew) (Table 9).

Additionally, we use stricter QC criteria for srWGS SV calling than for srWGS SNP and Indel calling and as a result, some samples were dropped during the QC steps. The final CDRv8 srWGS SV callset contains 97,061 samples.

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

Table 9 -- Number of samples that were excluded from SV calling

srWGS SV sample exclusion steps	Number of samples filtered from initial count (N=100,321)	Notes
Single sample QC	2066	See Table 10 and Table 11. 2,005 samples were removed by basic filters and 61 were removed during ploidy estimation.
Joint SV callset refinement and QC	11	Outlier samples were removed following ClusterBatch (see Appendix L).
Removed between CDRv6 and CDRv7	304	These are CDRv6 srWGS samples that were not included in CDRv7 for reasons unrelated to SV calling (e.g., participant withdrew between releases)
Removed between CDRv7 and CDRv8	879	These are CDRv7 srWGS samples that were not included in CDRv8 for reasons unrelated to SV calling (e.g., participant withdrew between releases or sample was missing mainline CDR data due to a known issue, CDRv7 off-cycle Known Issue #1, CDRv8 Known Issue #2)

Single Sample QC for srWGS SVs

We performed single sample QC, as described in Table 10 and Table 11, on all 88,882 newly selected samples for the CDRv7 off-cycle srWGS SV callset. We removed a total of 2,066 samples during srWGS SV single sample QC, which left 86,816 new samples and 98,255 total samples remaining in the callset for downstream processing.

Basic filters

Method

As seen in Table 10:

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

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

3. We checked the whole genome dosage (WGD) [22] 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 [24], Wham [25], and MELT [26] relative to the other samples in the cohort. Higher than typical SV counts may signify technical artifacts. SV counts were stratified by SV caller, chromosome, and SV type. Samples that were outliers in 30 or more categories were removed from the callset.

We removed all samples that failed any of these filters, in total 2,005 (Table 10). Note that some samples failed multiple filters.

Results

The results for all six basic single-sample filtering steps are summarized in Table 10.

Table 10 -- srWGS SV single sample QC: Basic filters

QC process	Passing criteria	Error modes addressed	Number of samples removed
Cross-individual contamination	≤ 0.01 (≤ 1%)	Sample contamination from another individual	296
Mean insert size	Mean insert size in range [320, 700]	Insert size outliers, which could skew distributions of discordant pairs	30

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

Ploidy estimation

Method

We estimated ploidy per chromosome across all 88,882 new samples by binning read counts in 1Mb intervals and normalizing by half the genome-wide median. We only performed filtering based on ploidy on the 86,877 samples that passed the basic filters (Table 10).

We observed likely mosaic loss of chrX and chrY in some samples, as described in previous studies [27] [28]. These samples had an estimated copy ratio of 0.1-0.8 on chrY and 1.2-1.8 on chrX and are likely to have mosaic loss of chrX or chrY, but the low copy number could also be due to large deletions on these chromosomes. For the sex-specific steps of the GATK-SV pipeline, these samples were classified as follows:

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

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

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

Results

We filtered 61 samples because they had an estimated copy ratio greater than 2.3 or less than 1.8 on at least one autosomal chromosome (Table 11). Plots of binned coverage across these

chromosomes confirmed that these samples may represent mosaic autosomal aneuploidies. In addition, we discovered 849 samples with a likely mosaic loss of chrX or chrY among the 86,877 new samples that passed basic filters, though in-depth analyses and validation of somatic and mosaic variation was outside of the scope of activities for this callset. All samples passed the comparison check between computed sex and self-reported sex at birth, indicating no sample swaps based on the computed sex.

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

Lists of the samples identified to have likely mosaic autosomal aneuploidies, likely mosaic loss of chrX or chrY, and germline sex chromosome aneuploidies are available; for additional details, read the Controlled CDR Directory on the User Support Hub [1]. The analysis was performed on the 86,877 new samples that passed basic filters and joined with the results from the samples previously examined during CDRv7 srWGS SV processing. Samples that were removed from the CDRv8 callset were removed from the lists of samples with probable aneuploidies, so the sample counts may differ from those represented here.

Table 11 -- srWGS SV single sample QC: Ploidy estimation filters

QC process	Passing criteria	Error modes addressed	Number of samples removed	Notes
Estimated copy number per autosome (Ploidy estimation)	1.8 ≤ copy ratio ≤ 2.3	Samples with mosaic autosomal aneuploidies, which could skew distributions of SV evidence classes	61	Calculated after applying all above filters. Method can be found in [22]
Sex concordance	Computed sex is concordant with self-reported sex at birth. OR Computed sex is neither male nor female. OR Self-reported sex at birth reported as “Other”* or was not reported	Sample swaps	0	All samples passed this check *Other refers to a participant self-reporting “Intersex”, “I prefer not to answer”, or “none of these fully describe me”

Batching

We divided the 88,882 new samples into 168 batches with an average of 517 samples in each batch for the batched analysis steps of the GATK-SV pipeline, depicted in Figure 7. Batching controls for technical variability between samples and parallelizes computation. The batching procedure was as follows:

1. Split by chrX copy ratio (<1.5 and ≥1.5)

2. Split each partition of samples from the previous step four ways by mean insert size
3. Split each partition three ways by WGD score
4. Split each partition two ways by median coverage
5. Merge corresponding partitions by chrX ploidy to balance chrX ploidy within batches

The batching scheme was based on previously described methods [22], except for the addition of the mean insert size as a batching parameter. We added this to address an observed multimodal distribution of mean insert size, described previously in the CDRv7 QC report [1].

Joint Callset Refinement and QC for srWGS SVs

The steps to generate the GATK-SV joint callset are described in Figure 7 and Appendix L. Appendix L also includes a summary of GATK-SV pipeline improvements that have been implemented since the CDRv7 srWGS SV release. Below, we describe refinement and filtering steps introduced in the *All of Us* srWGS SV dataset that were not published previously or are modifications to canonical GATK-SV pipelines (blue steps in Figure 7). These steps include both hard and soft filters at the sample, site, and genotype level (Table 12).

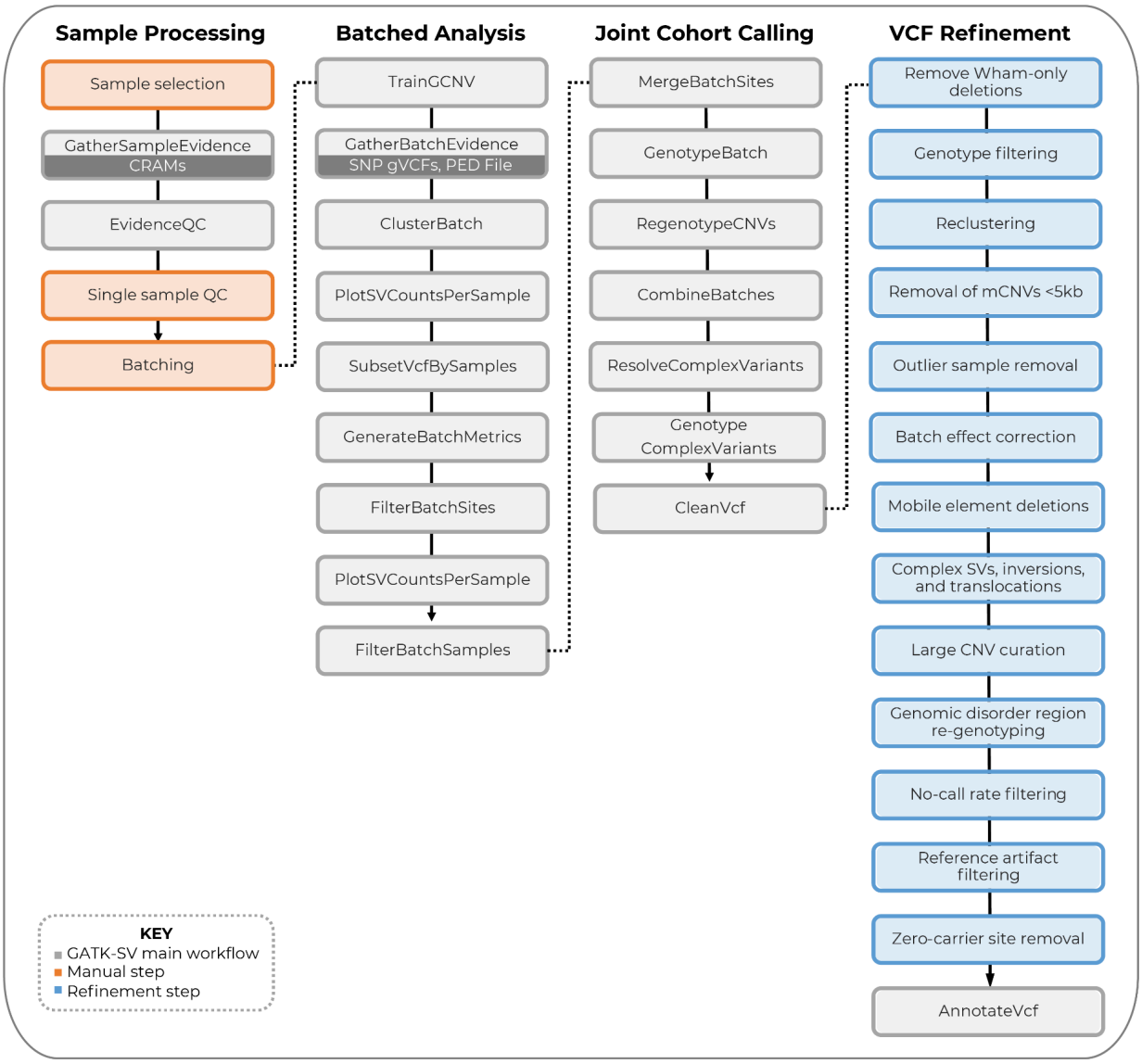


Figure 7 -- GATK-SV Pipeline Schematic. GATK-SV automated workflows are shown in gray and the names correspond to the name of the Workflow Definition Language (WDL) file. Manual steps performed in notebooks are shown in orange. Steps in blue are custom VCF refinement and QC steps for the *All of Us* SV callset.

Table 12 -- 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 Wham-only	Variant		False positive deletions	Unique Wham deletions were removed from the callset.

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

removal			removed if no carriers remain after filtering	remain after filtering.

Remove Wham-only deletions

As described in the CDRv7 QC report, we observed very high false-positive rates for deletions that were uniquely called by the Wham algorithm [25], one of the SV calling algorithms used by GATK-SV. These variants were removed from the callset.

Genotype filtering (SL filter)

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

Method

lrWGS training data

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

A subset of 893 samples with matched lrWGS data were selected for model training, and an additional 97 were held out as a test set to validate the model. For each sample, non-reference genotypes for eligible variants (SV type DEL, DUP, INS, or INV, restricting to below 5 kb in length for CNVs) were assessed against lrWGS. Calls were first evaluated using the lrWGS validation tool VaPoR [29]. In addition, the lrWGS variant calling was performed using the tools PAV [30], PBSV [31], and sniffles2 [32]. The GATK tool SVConcordance in GATK version 4.6.0.0 was then used to compute overlap between SV calls from srWGS and lrWGS [33].

Variants were labeled as positive training examples if:

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

Variants were labeled as negative training examples if:

- The variant had at least 5 reads that VaPoR was able to evaluate in the sample and no reads had a positive VaPoR_Rec score AND

- The variant was not within 5 kb of a breakpoint of a lrWGS SV call with a matching SV type.

Variants that did not meet either the positive or negative criteria were dropped from the training set (Figure 8A).

Filtering model

We trained a model to re-calculate SV genotype qualities based on the training data. This produced more accurate quality scores to use for filtering low-quality genotypes. We used XGBoostMinGqVariantFilter, a GATK tool [34], to perform the quality score recalibration. This tool applies a decision tree from the XGBoost library for gradient boosted machine learning to predict the quality of a given genotype [35].

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

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

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

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

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

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

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

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

Genotype filtering was applied to the same variant types that were used for training (DEL, DUP, INS, and INV). See lrWGS training data for additional details. However, the size restriction on

DEL and DUP variants was increased from 5 to 10 kb for filtering, as the variants in this range are expected to have error modes similar to those used for training (under 5 kb). Filtering was not applied to CNVs that were either multi-allelic or over 10 kb in size because those categories lacked training labels.

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

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

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

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

Results

Analysis of the training samples from lrWGS and genotyping arrays yielded a total of 27,437,577 trainable genotypes, while labels for 15,611,637 genotypes (36% of the total) could not be determined (Figure 8A). 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 8B).

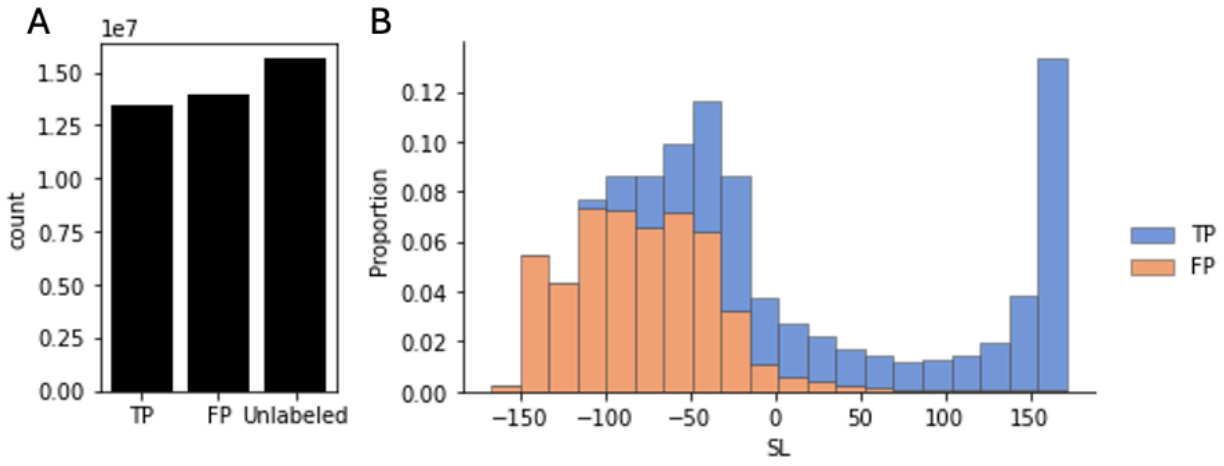


Figure 8 -- Training data for genotype filtering. (A) The proportion of each training label out of all SV genotypes in the training data, and (B) the SL score distribution produced by the trained model.

The genotype filtering performance was evaluated in the test set of 97 held-out samples with matched lrWGS data. We observed that precision decreases consistently as a function of recall when thresholding on SL (Figure 9). This demonstrates that the method is effective for tuning callset accuracy. These results also indicate comparable performance across the spectrum of SV classes. Optimal cutoffs for SL filtering were determined using the training set as described above and are shown in Appendix Table M.1.

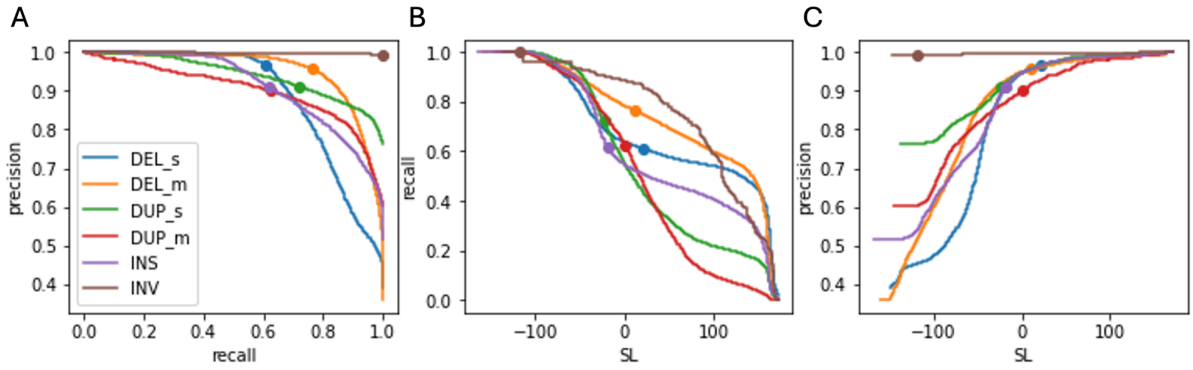


Figure 9 -- SL genotype filtering performance assessed against 97 lrWGS labeled test samples. (A) Precision-recall curves for all filtering classes, (B) recall as a function of the SL cutoff value, and (C) precision as a function of the SL cutoff value. Markers depict cutoffs used for genotype filtering.

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

Reclustering in repetitive regions

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

Removal of mCNVs <5kb

Read depth signal is less reliable in events smaller than 5 kb [36]. We removed all MCNVs under 5 kb in length from the callset, so they will not appear in the VCF file. We report MCNVs of greater than 5 kb with the “MULTIALLELIC” filter tag. Therefore, all MCNVs in the final callset will have a length greater than 5 kb and be tagged as “MULTIALLELIC”.

Outlier sample removal

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

Batch effect correction

We evaluated each variant for batch effects among the 192 batches used for the batched steps of the GATK-SV pipeline (See Appendix L). 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 [22]. Please note that PCR-amplified samples are not part of the AoU cohort, and 36,672 pairwise comparisons were not feasible, so we applied only the one-vs-all comparisons described in Collins et al.

Mobile element deletions

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

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

Translocation sensitivity

To improve the sensitivity for inter-chromosomal translocations (CTX) in this callset, we re-evaluated the raw translocation calls from Manta [24]. We clustered the translocation variants across batches of around 500 samples and we retained only the rare variants (<1% allele frequency). We next removed redundant translocations that were within 100 bp of a translocation site already called by GATK-SV within the batch. We manually reviewed the discordant paired end read (PE) evidence for each non-reference genotype as described below. Translocations with sufficient PE evidence were added to the GATK-SV callset.

Filtering complex SVs and translocations

Specific alignment patterns and discordant paired end reads are expected for complex (CPX) and translocation SVs [22] . 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 [37] and GATK-gCNV [38]).

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

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

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

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

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

Large CNV curation

We performed a visual inspection of read depth across all 1,322 CNVs (deletions and duplications) larger than 1 Mb observed in our final VCF using a visualization tool found in GATK-SV [39]. After inspection, we confirmed the presence of 1,310 CNVs (99.1%). We observed that 4 of the CNVs larger than 1Mb appeared to have multiple copy states, so we applied the multiallelic filter tag (MULTIALLELIC). Finally, for 415 CNVs (31.4%) that had at least one sample with inaccurate breakpoints, we manually reassigned breakpoints using the more precise sample level depth calls derived from preceding modules in the pipeline. All revisions resulting from manual review are described in the INFO field MANUAL_REVIEW_TYPE.

Genomic disorder region re-genotyping

Genomic disorders are human diseases largely arising from recurrent CNVs mediated by segmental duplications containing homologous sequences [40]. To improve variant discovery and genotyping accuracy in known genomic disorder (GD) regions [41], we applied local depth-based re-genotyping to large CNVs. The purpose of this step is to ensure that these complex and repeat-mediated events are accurately profiled and not fragmented into smaller events during variant clustering and defragmentation. Briefly, depth evidence of all bi-allelic DEL and DUP sites overlapping at least 40% of a GD region were reassessed to refine breakpoints, remove false positives, and recover false negatives.

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

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

No-call rate filtering

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

Reference artifact filtering

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

Zero-carrier site removal

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

CDRv8 Updates

This section describes the changes that were applied to the CDRv7 off-cycle srWGS SV callset to produce the CDRv8 callset.

Sample removal

We removed the 879 samples that were removed between CDRv7 and CDRv8 that were in the CDRv7 off-cycle SV callset. We also removed all variant sites for which only the dropped samples were carriers.

Insertion reclustering

A high degree of redundancy was observed in the insertion sites in the CDRv7 off-cycle srWGS SV callset, particularly in and around simple repeat regions. To reduce this redundancy, we applied additional clustering to insertions. For all insertions, we clustered sites that had 50% reciprocal overlap and had breakpoints within 10 base pairs (bp), regardless of the fraction of carrier samples shared. We further reclustered the subset of insertions in simple repeat regions and within 100 bp of simple repeat regions that had 50% reciprocal overlap and had breakpoints within 100 bp, regardless of the fraction of carrier samples shared.

Complex SV filtering

We identified an issue that resulted in the PE and depth evidence assessments and genotype filters described in Filtering complex SVs and translocations not being applied to a subset of complex SVs smaller than 1 Mb in size. We applied those filters to the remaining complex SVs that were not previously assessed.

Merging redundant CNVs in genomic disorder regions

Redundant CNV records overlapping genomic disorder regions were observed. Four pairs of CNV records were merged to address this redundancy.

Final updates

To account for the changes we applied, we redid No-call rate filtering, Reference artifact filtering, Zero-carrier site removal, allele frequency annotation, and QC and benchmarking.

Structural Variant QC Results

Below we detail several metrics of interest for this SV callset. Figure 10 shows the SV counts, stratified by SV type, within the callset. In this figure, we include measures from both the total callset (all variants in the callset, regardless of filter tag) as well as a high-quality callset composed of only variants with a filter tag of PASS or MULTIALLELIC. The remaining figures focus on the high-quality callset. Figure 11 shows the distribution of SV counts per genome, stratified by SV type, in the full cohort and grouped by *All of Us* genetic ancestry groups (see Appendix G). Figure 12 shows the distribution of SV lengths for each SV type; the fraction of SVs decreases with increasing SV size, except for MCNVs, which are always over 5 kb, and INS, which have peaks representing Alu, SVA, and LINE-1 mobile genetic elements [42]. Figure 13 shows the ratios of homozygous reference, heterozygous, and homozygous alternate genotypes at each SV site and the fraction of SV sites that are in Hardy-Weinberg equilibrium.

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

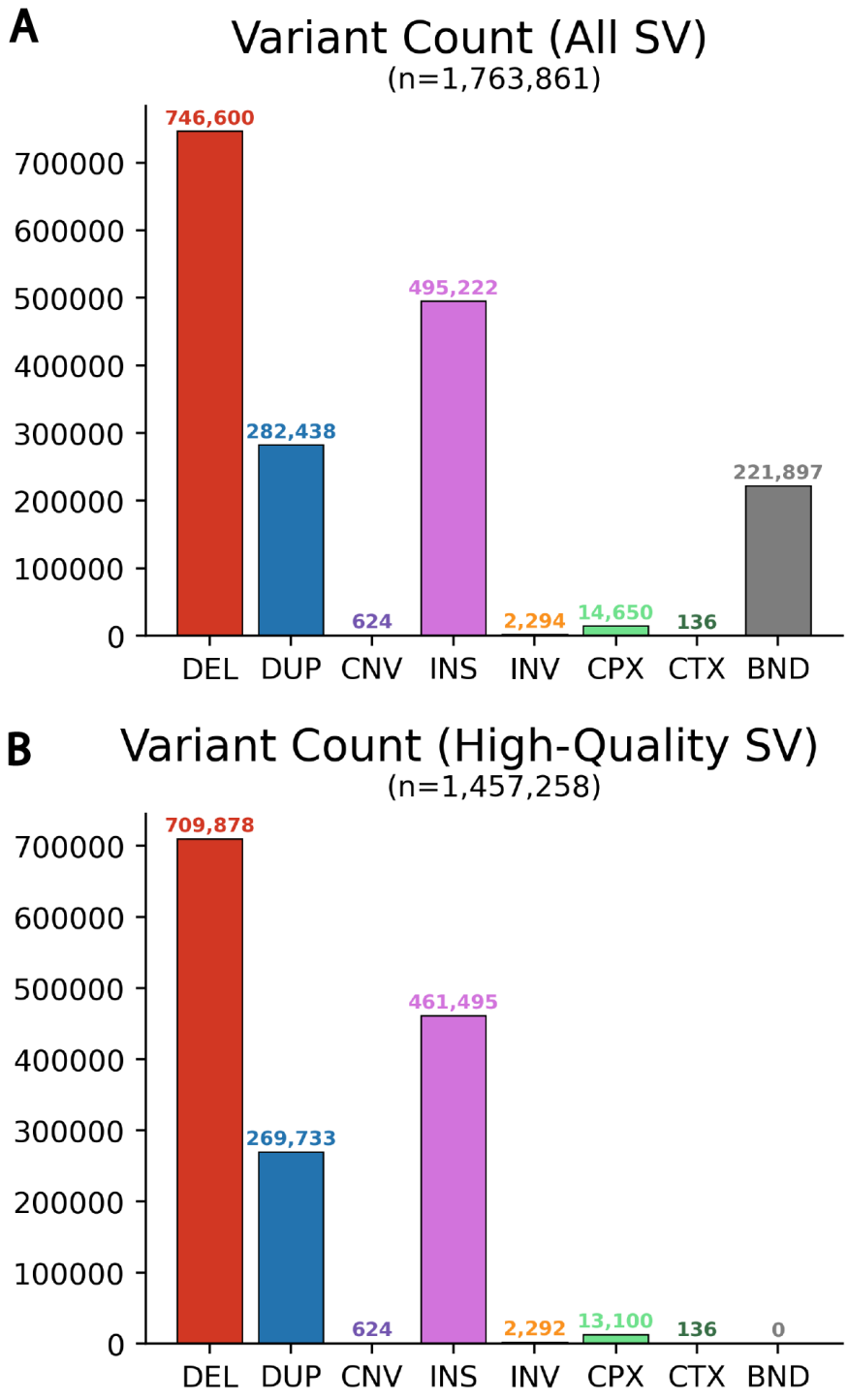
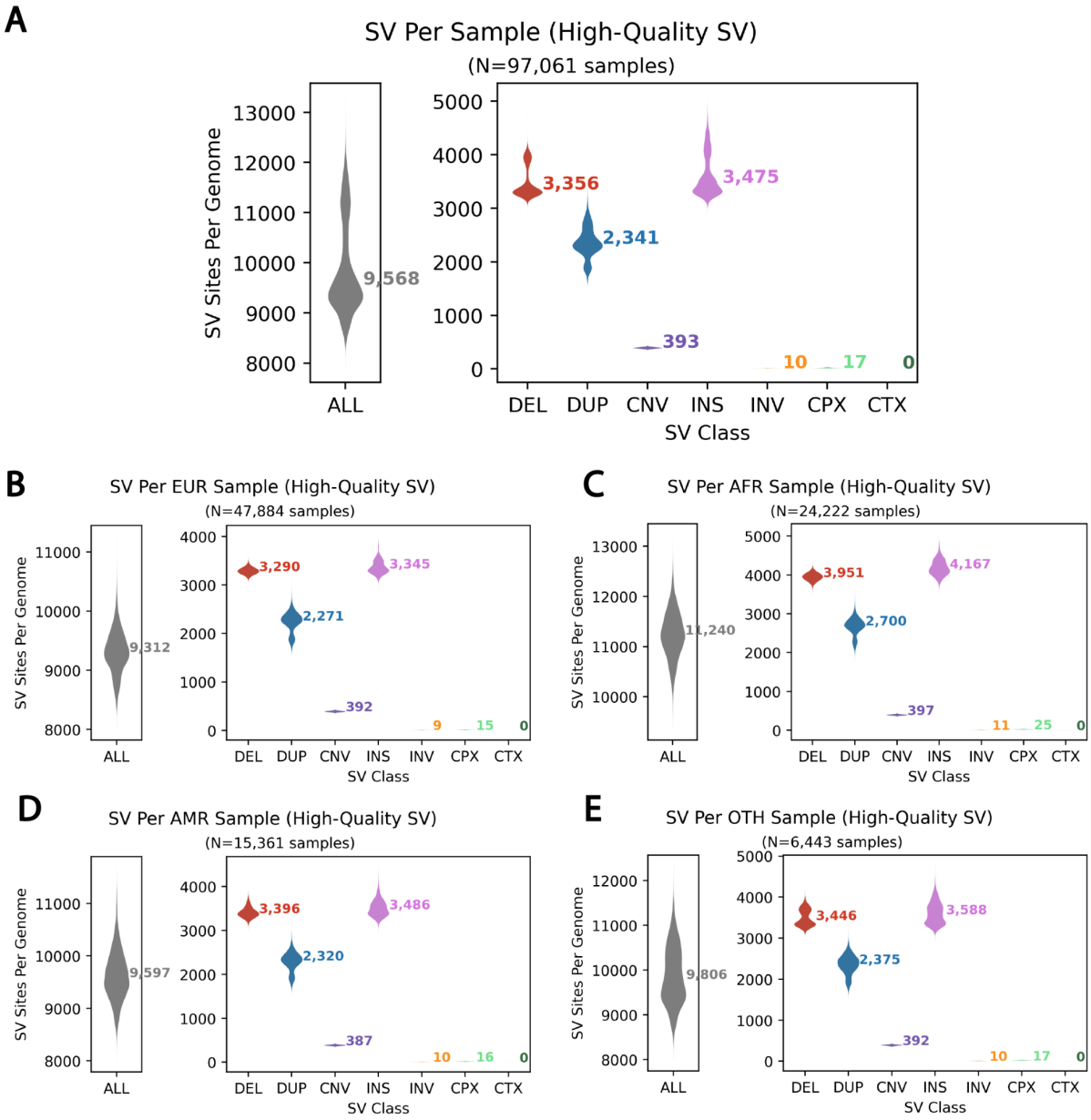


Figure 10 – SV counts in the complete callset and the high-quality SV callset. We observed 1,763,861 total SVs of which we determined 1,457,258 (82.6%) to be of high quality. (A) The total callset includes all variants in the callset regardless of the filter status. (B) The high-quality SV callset only contains variants with the PASS or MULTIALLELIC filter status. Note that all BND sites have the filter UNRESOLVED, so they are not included in the high-quality callset.


A
SV Per Sample (High-Quality SV)
(N=97,061 samples)
SV Sites Per Genome
9,568
3,356
2,341
393
3,475
10
17
0
ALL
DEL DUP CNV INS INV CPX CTX
SV Class
B
SV Per EUR Sample (High-Quality SV)
(N=47,884 samples)
9,312
3,290
2,271
392
3,345
9
15
0
C
SV Per AFR Sample (High-Quality SV)
(N=24,222 samples)
11,240
3,951
2,700
397
4,167
11
25
0
D
SV Per AMR Sample (High-Quality SV)
(N=15,361 samples)
9,597
3,396
2,320
387
3,486
10
16
0
E
SV Per OTH Sample (High-Quality SV)
(N=6,443 samples)
9,806
3,446
2,375
392
3,588
10
17
0

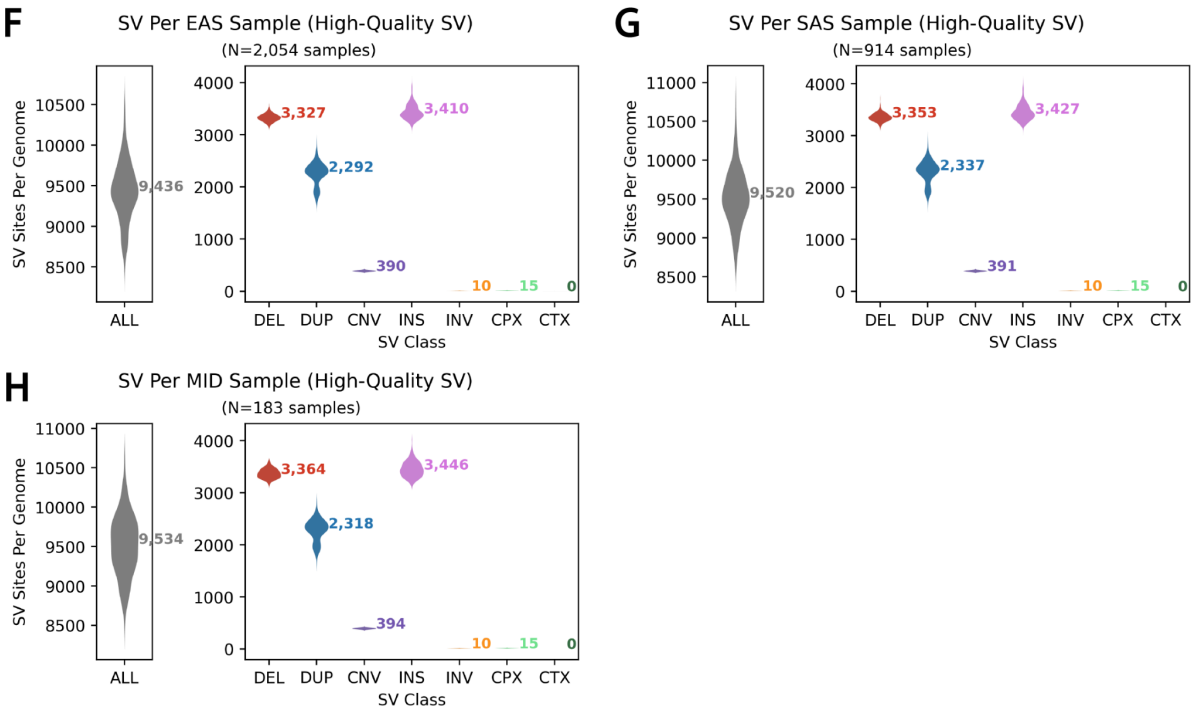


Figure 11 – We observed a median of 9,568 high-quality SVs per person, which is consistent with SVs recently generated on the 1000 Genomes Project samples [43]. We display here the overall SVs per genome and per SV type per genome in the high-quality callset (A) as well as stratifying by the *All of Us* predicted genetic ancestry group in order of prevalence in the callset (B-H). See Appendix G for the *All of Us* genetic ancestry groupings. The median of each distribution is labeled on the plot. As expected, samples in the *All of Us* African/African American genetic ancestry group (AFR) had the highest SV counts while those in the *All of Us* European genetic ancestry group (EUR) had the lowest SV counts.

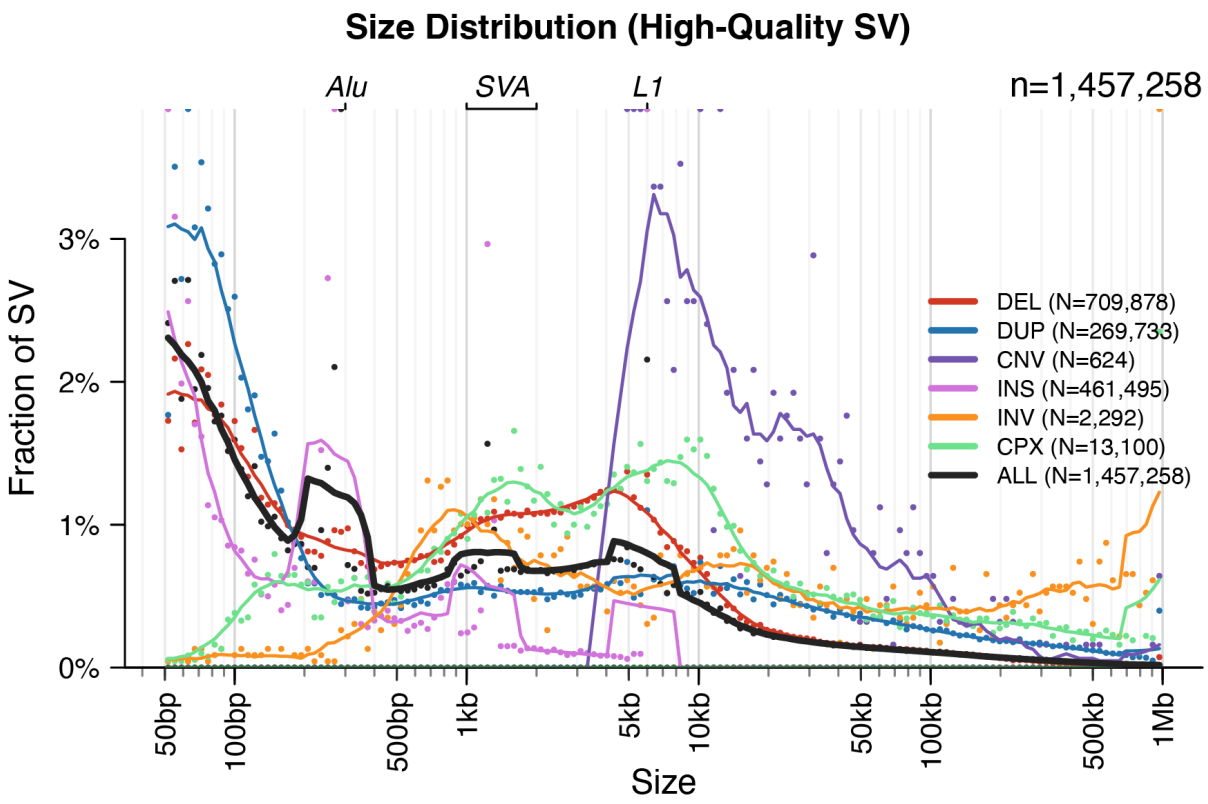


Figure 12 – SV size distribution matches previous expectations with notable insertion peaks corresponding to Alu, SVA, and LINE-1 insertions. Points represent the fraction of each SV type occupied by a given size range. Lines represent the rolling 10-bin average (the size ranges are divided into 150 bins).

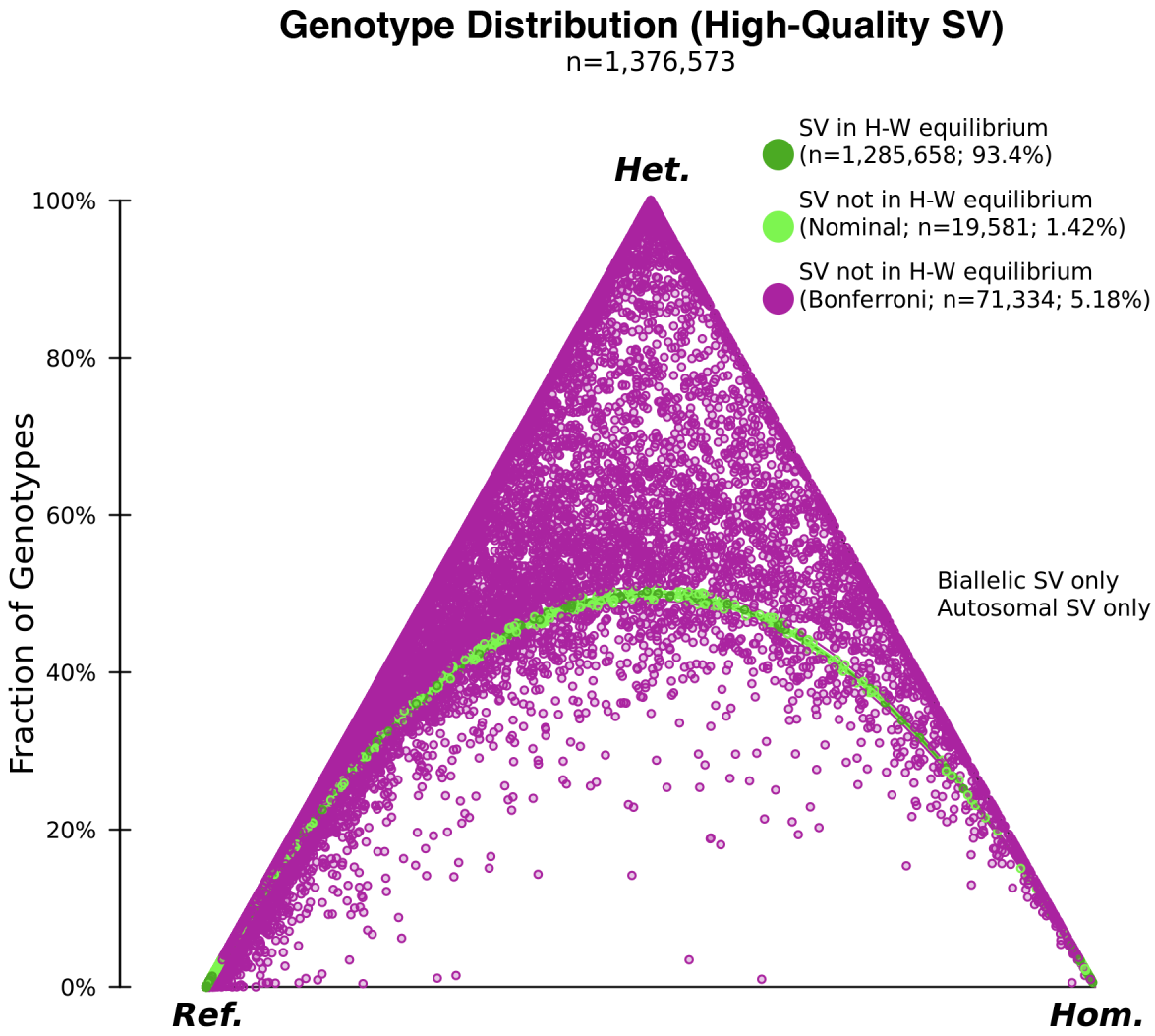


Figure 13 – Among high quality variants, 93.4% are in Hardy Weinberg Equilibrium (HWE). Of the 5.18% that fail, most of these failures appear to be driven by a bias towards genotyping variants as heterozygous. For this calculation, we included only the 93,360 unrelated samples and only biallelic SV sites on autosomes.

Long-Read Whole Genome Sequencing (lrWGS)

We have data representing 2,800 participants in the long-read genomic dataset. These data are particularly useful for resolving complex genomic regions, structural variants, and phasing of alleles, to provide a more comprehensive view of the genome. We have added lrWGS data from 1,773 participants in the CDRv8 release to accompany the lrWGS data from 1,027 participants in CDRv7. To maximize the diversity of the lrWGS dataset, non-european participants are over-represented. The self-reported race and ethnicity data for the participants with lrWGS data can be found in Appendix H.

This report covers the QC steps for the new lrWGS samples representing 1,773 participants. For the QC results for the CDRv7 1,027 samples, please see the CDRv7 QC report. While we generally follow the same QC steps, because the data types are different, some of our QC processes are different.

Please see the overview of our lrWGS pipeline in Appendix N for how we perform QC, generate SNP and Indel variants, call SVs, and perform *de novo* assembly. Our sequencing data is from two different sequencing technologies, Pacific Biosciences (PacBio) High-Fidelity (HiFi) and Oxford Nanopore Technologies (ONT).

The lrWGS data are aligned to the grch38_noalt and T2Tv2.0 references. The QC steps are performed on the read data, then at the single sample level, and then for each data type, including *de novo* assembly, SNP and Indel variants, and structural variants. The data is described in more detail in the How the *All of Us* Genomic data are organized article on the User Support Hub [1].

The following are the general QC steps we performed:

1. Data generation: PacBio Hifi and ONT sequencing
2. Single sample QC: At the read group and single sample level
3. *De novo* assembly: generated for all PacBio HiFi data
4. SNP and Indel joint callset QC
5. Structural variant individual sample QC

During the QC process, we flagged some samples that displayed abnormal behaviors. The sample IDs are available in RW as a 3-column CSV, where the 3 columns are: sample ID, sequencing facility, and reasons for flagging (there could be multiple reasons for a sample).

Data generation

The lrWGS data were generated at five sequencing facilities, including Baylor College of Medicine (BCM), Broad Institute (BI), Johns Hopkins University (JHU), and University of Washington (UW), and HudsonAlpha Institute (HA).

The lrWGS data are aligned to the grch38_noalt and T2Tv2.0 references [44]. grch38_noalt corresponds to the GRCh38 reference with no alternate sequences [45,46]. T2Tv2.0

corresponds to the T2T-CHM13v2.0 reference with a few modifications [47]. The EBV contig is added from the grch38_noalt reference, Chromosome Y is hardmasked with N bases in the Human Pseudoautosomal Region (PAR) region, and the mitochondrial genome is updated to the revised Cambridge Reference Sequence (rCRS). We updated the T2Tv2.0 reference for this CDRv8 release and so it is different from the previous CDRv7 T2Tv2.0 version. For more information see Known Issue #7.

Participants were selected for long-read sequencing by each sequencing facility. The criterion for a sample being selected was that it had matching srWGS data. The sequencing facilities used both PacBio HiFi sequencing and ONT sequencing (Table 13), which both generate single molecule sequences that are typically longer than 10kbp. Their base qualities and other systematic artifacts, however, can differ.

PacBio HiFi sequencing uses DNA molecules circularized with bell adapters that are repeatedly sequenced to generate consensus sequences [48]. Some HiFi sequences were generated on Sequel IIe, also called Sequel, whereas some were generated on a Revio. The Revio is a newer HiFi machine than the Sequel IIe. No sample is sequenced on both Sequel IIe and Revio. The ONT sequencing platform measures changes in ionic current as nucleic acids pass through a nanopore [49]. The particular ONT platform used at the sequencing facilities was the ONT R10.4 on PromethION.

The CDRv8 dataset represents 1,773 participants, though we have a total of 1,815 samples, since 41 participants are sequenced on both PacBio and ONT. In addition, one participant was sequenced at both BI and UW, though to different coverage.

Sample cohorts

We separated the lrWGS samples into cohorts based on their sequencing facility, the sequencing technology (HiFi vs ONT), the minimum coverage, and for HiFi, the generation of the machine (e.g. Revio vs Sequel IIe) (Table 13).

Samples were sequenced with different minimum coverages, which is the minimum coverage for each sample in each cohort. A cohort is either high-pass, with a minimum coverage of 25x or mid-pass, with a minimum coverage of 12x. This is an increase of the minimum coverage from CDRv7, where the coverage cutoff was 8x. The minimum coverages were chosen to balance the number of samples and the depth of each sample to achieve high power enabling downstream analyses.

The sample cohorts were used for QC steps and SNP and Indel joint-calling. SVs were called for single samples. Because of batch effect concerns, we analyzed these effects, reported in Appendix R. Because we joint-called only within batches, we also did not joint-call the CDRv7 data with the CDRv8 data.

Note that due to the fast-evolving nature of long-read sequencing technologies, small variations exist even within a particular generation of a technology. For example, for the PacBio machines, the instrument control software (ICS) being updated could cause variations in the data. The variations within a generation of technology likely have only minor effects when compared to the

factors listed above. To keep larger cohorts to boost the power of joint-calling, we did not further divide the cohorts.

Table 13 -- Sample cohorts for all 2,800 participants with lrWGS data

Cohort name	Sequencing facility	Sequencing platform	Number of samples	Minimum coverage	Notes
HA_Rev_mid	HA	PacBio Revio	65	Mid-pass (12x)	
HA_Seq_CDRv7	HA	PacBio Sequel IIe and Sequel II	1027	Mid-pass (8x)	The CDRv7 data
BI_Seq_high	BI	PacBio Sequel IIe	84	High-pass (25x)	
BI_Seq_mid	BI	PacBio Sequel IIe	198	Mid-pass (12x)	
BI_Rev_mid	BI	PacBio Revio	803	Mid-pass (12x)	
BCM_Seq_high	BCM	PacBio Sequel IIe	77	High-pass (25x)	
BCM_Rev_high	BCM	PacBio Revio	111	High-pass (25x)	
BCM_ONT_high	BCM	ONT R10.4 on PromethION	196	High-pass (25x)	
JHU_ONT_high	JHU	ONT R10.4 on PromethION	128	High-pass (25x)	
UW_Seq_high	UW	PacBio Sequel IIe	100	High-pass (25x)	
UW_Rev_high	UW	PacBio Revio	53	High-pass (25x)	
Total samples			**2842**		42 CDRv8 participants were sequenced in two different samples

Single Sample QC

We perform the following QC methods at both the read group level and the single sample level, summarized in Table 14. These QC steps are performed with the grch38_noalt aligned data. These QC methods are consistent across all the sequencing protocols and cohorts, though we adjust parameters when applicable, as noted. LrWGS samples are typically sequenced across multiple read groups. A read group is a set of sequencing reads that have the same technical properties and conditions, like being run on the same sequencing machine and prepared in the same way.

We first perform the QC steps at the read group level to check for any read groups that show signs of quality issues, before aggregating the read groups by sample and running the same QC

steps (Table 14). Read groups and samples that fail any QC checks are dropped and not further analyzed or released.

Table 14 -- QC processes performed for read groups and single samples

QC Process	Read groups or samples?	Passing criteria	Error modes addressed	CDRv8 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.
Coverage	Sample	- Samples were evaluated to see if they met their intended coverage (Table 13)	- Sample preparation errors - Poor sensitivity and precision of variant calling	All except a small amount of HA samples and one BI sample meet the intended coverage threshold. These samples are included as they are not far below the minimum coverage for their corresponding cohorts.
Read length median	Sample	≥ 10,000 bp	- Shorter fragments significantly impacting variant calling and assembly performance	All lrWGS samples passed this check.

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 individual read groups and the aggregated

sample reads. We use the same fingerprint VCFs that are used by the srWGS fingerprint verification pipeline and the same method, described above. The HAPLOTYPE_MAP parameter is the only parameter that differs, with only a difference in the header section.

Parameter	Value
HAPLOTYPE_MAP	"gs://gcp-public-data--broad-references/hg38_noalt/v0/aou/fp/lr.aou.fp.haplotype_database.no_alt.txt"

Results

All lrWGS samples in the CDRv8 release passed the fingerprint concordance check. Fingerprint LOD results are displayed with coverage in Figure 16.

Sex Concordance

Method

We performed a sex concordance check on the grch38_noalt version of each BAM, using mosdepth [50] to calculate coverage across the whole genome and over each chromosome. Tool parameters are listed in Appendix O. We used the following formula to infer the sex ploidies for each read group and sample.

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

We compared the inferred sex chromosome ploidies to each participant's self-reported sex assigned at birth (Appendix C). If the two sources were not concordant, we assumed a potential sample swap, removed the sample, and investigated the source of the swap. 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, regardless of the information from the inferred sex ploidy. The sex assigned at birth data is described in Appendix C.

Results

We do not include any lrWGS samples that fail the sex concordance check in this release.

We performed a manual review for several samples that initially failed the sex concordance check because of the ploidy formula. These samples had low Y-chromosome coverage and due to rounding, were marked as having no Y-chromosome coverage. We manually checked in IGV for reads that mapped to the Y chromosome and marked the samples as passing if they had Y chromosome coverage higher than that expected if the participant was female. There is a possibility that these samples have mosaic loss of the Y chromosome. These samples have been flagged in the lrWGS flagged samples list, in the RW.

One sample marked as female had relatively low chrX coverage, though using the fingerprint LOD score of > 25, we found it unlikely that there was a sample swap. This sample is also flagged and available in the flagged sample list in the RW.

Cross-Individual Contamination Rate

Method

We performed a Cross-Individual Contamination check to remove any samples that had a high level of contamination from another individual. The complete method, described in the CDRv7 Genomic Data QC Report, converts the VerifyBamID2 tool to work with long-read data by using a pileup format of the grch38_alt alignment at selected sites [9].

We have identified that this method underestimates cross-individual contamination when the contaminant is from a related sample, as described in Appendix P. This could impact the Broad high-coverage samples.

Results

We did not include any lrWGS samples with a cross-individual contamination rate higher than 3% (Figure 14).

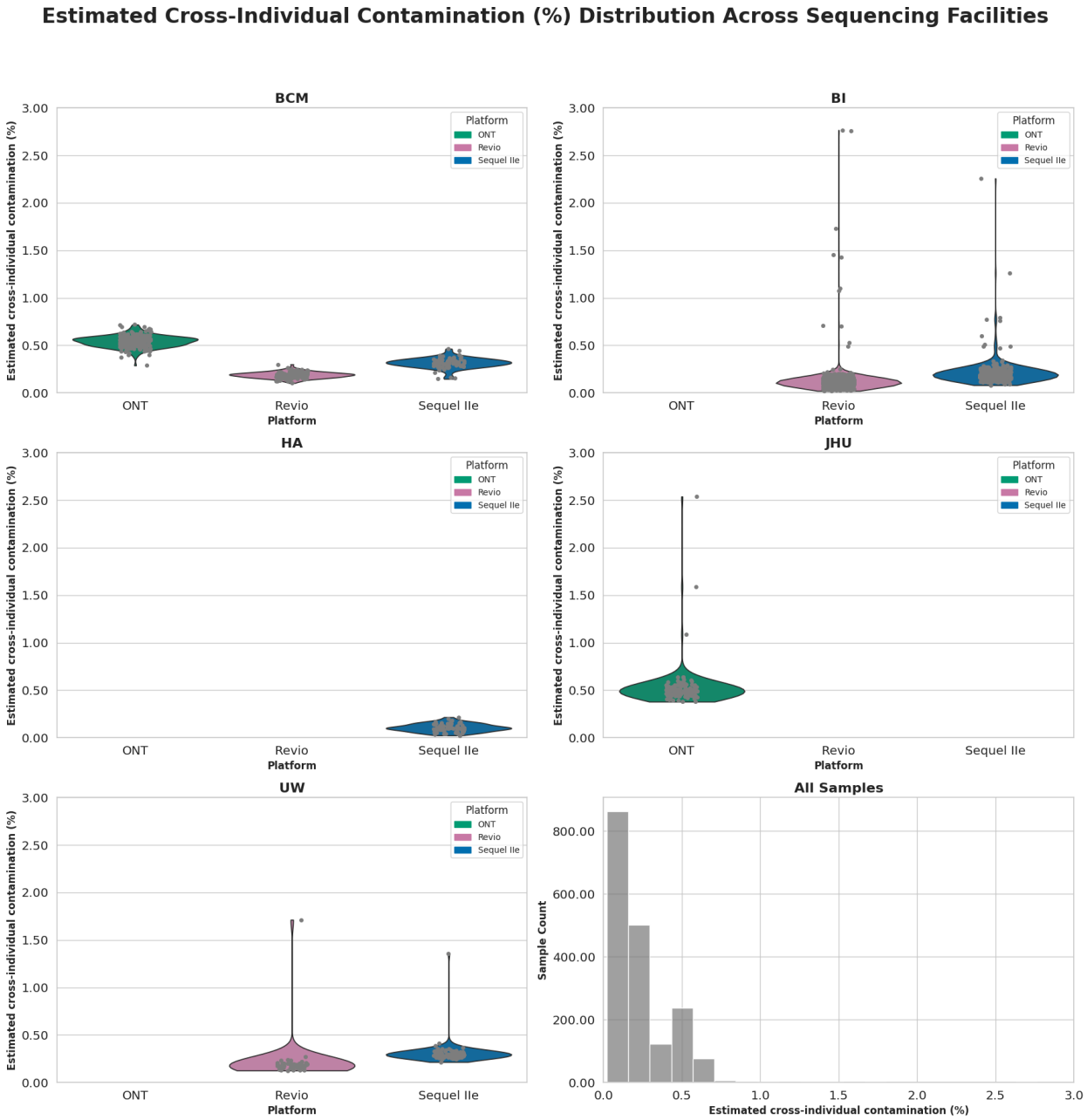


Figure 14 -- The distribution of the cross-individual contamination rates across sequencing facilities and platforms (Sequel IIe, Revio, ONT). Each subplot represents a different sequencing facility, with the x-axis showing the sequencing platforms and the y-axis indicating the contamination percentages. The violin plots illustrate the distribution of the contamination estimates, while the overlaid dots represent individual data points.

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. Since

samples were selected by and sequenced at different facilities, no universal coverage threshold applies to all samples in this CDRv8 cohort. We did not filter by coverage but used the metric as an indicator along with the other QC methods performed.

The mean coverage of each sample is collected with the tool mosdepth [50]. Tool parameters are listed in Appendix O.

Results

Most samples meet their minimum coverage, except a few samples in the cohorts BI_Seq_high and HA_Rev_mid (Table 13). We decided to include them in the release because they are close to the minimum coverage. A detailed breakdown of coverage by sub-cohorts is available in Figure 15.

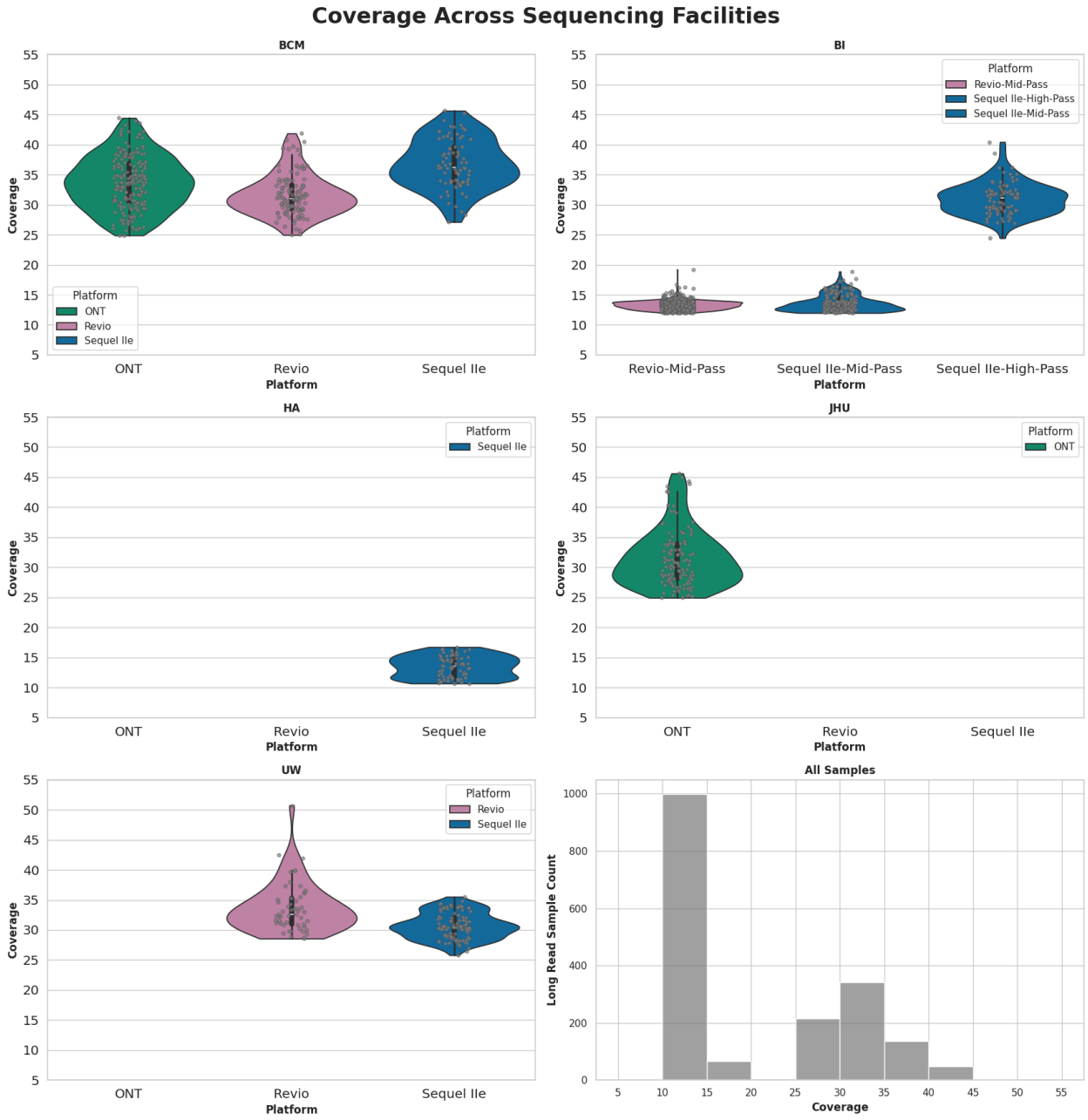


Figure 15 -- Coverage for each new lrWGS sample in the CDRv8 release. Each subplot represents the coverage distribution from one sequencing facility. Coverage for all of the new lrWGS samples in the CDRv8 release is displayed in the last subplot.

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 17.

We also compared the read length median to the coverage at each sequencing facility and with every sequencing technology, described in Appendix Q. The analysis demonstrated that there was no clear correlation between the read length median and the coverage.

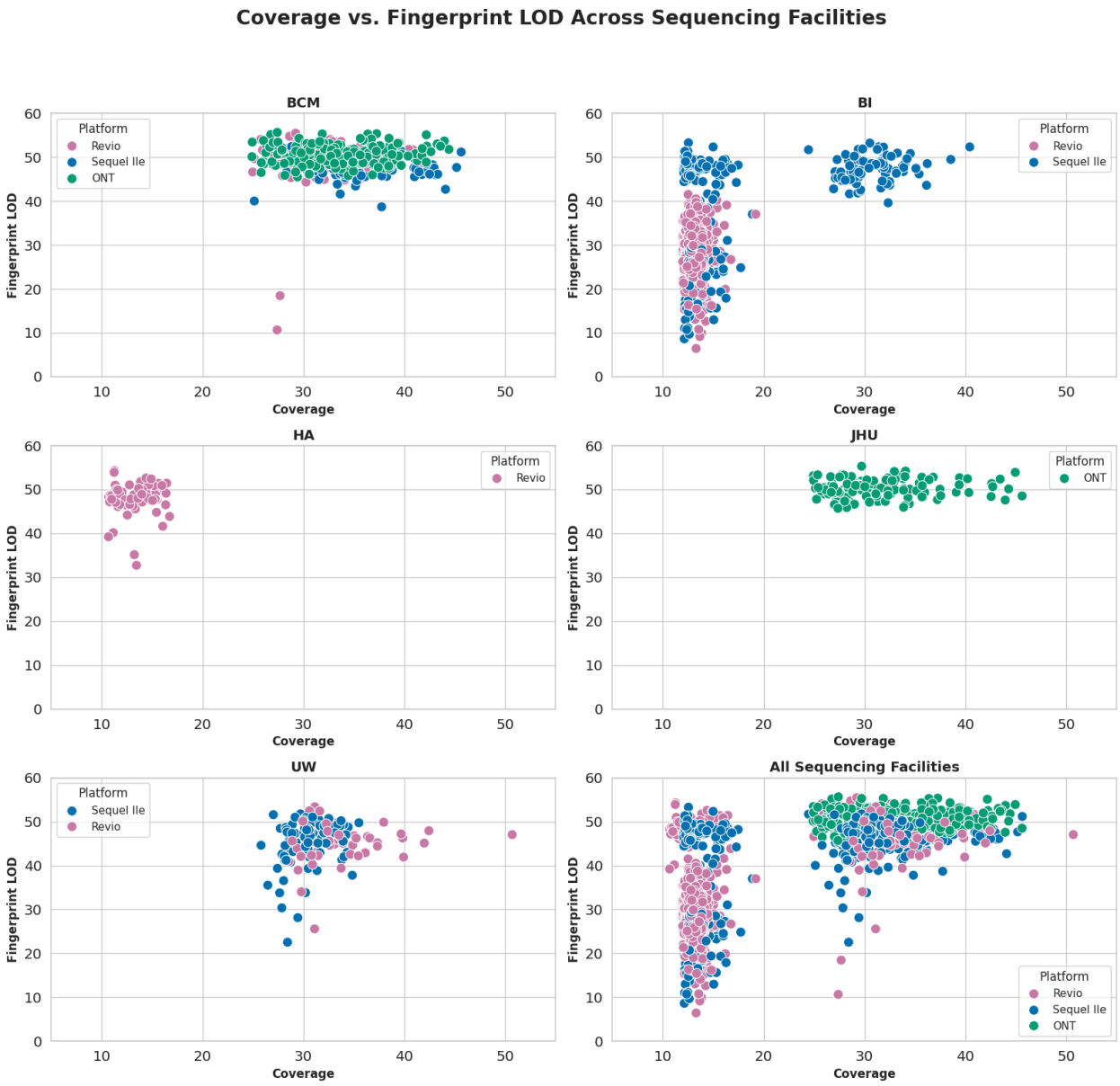


Figure 16 -- Coverage vs. fingerprint LOD for each sequencing facility. The x-axis displays the coverage values and the y-axis displays the fingerprint LOD score. Data points are color-coded based on the sequencing platform used, including Sequel IIe, Revio, and ONT, as indicated by the legend in the upper right. Though LOD scores vary with coverage, all samples pass this QC check.

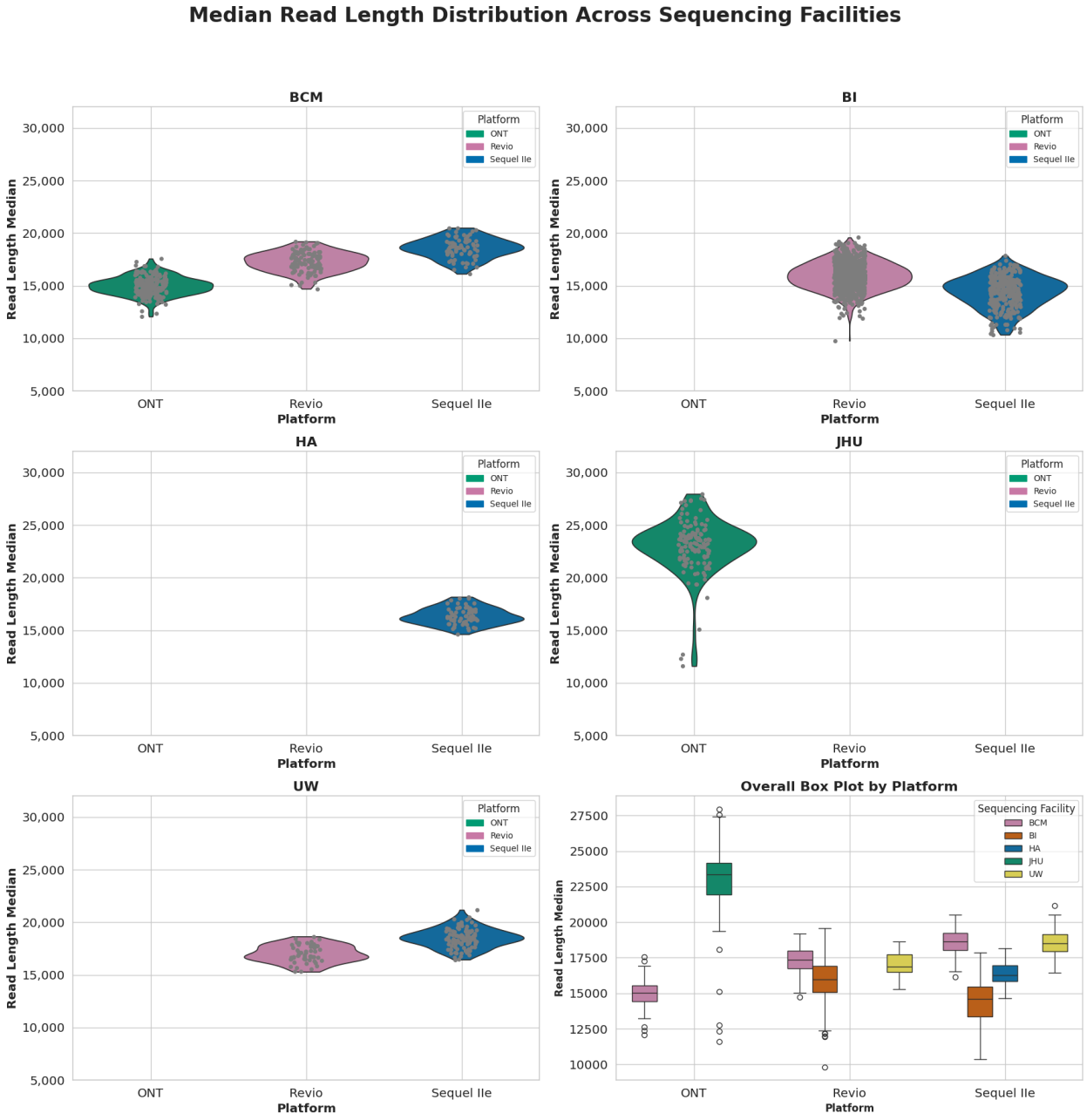


Figure 17 -- Read length median across sequencing facilities and platforms. Each subplot demonstrates the distribution of read length medians for one sequencing facility. Read length medians are displayed for all sequencing facilities in the last subplot.

De Novo Assembly

Method

We performed haplotype-resolved *de novo* assembly for all PacBio HiFi samples (Table 13), using the tool hifiasm [51]. We did not generate *de novo* assemblies for ONT samples (Appendix N). To evaluate the quality of the *de novo* assemblies, we used the tool QUAST [52]. Each *de novo* assembly has two haplotypes, which represent the genome that is inherited from each parent. Two metrics are calculated for each haplotype, auN and assembled genome length, which will help diagnose major *de novo* assembly issues. Assembled genome length is a proxy measure of the completeness of the assembly through the length of assembled sequences. We looked into using BUSCO [53] for the completeness evaluation but did not use it in production based on scalability concerns observed during testing. auN is a measure of contiguity of the assembly contigs that is less sensitive to large jumps in contig length [54].

After generating the *de novo* assembly, variants were called on the *de novo* assembly using PAV [34] for each sample on T2Tv2.0 and grch38_noalt (Appendix N, Figure N.3) to generate phased SNPs, Indels, and structural variants.

Results

When looking at *de novo* assembly genome lengths, we flagged 12 total samples. Flagged samples are circled in red in Figure 18. We flagged one sample in the UW plot that has its haplotype 2 assembled genome length shorter than its peers in the same cohort. In the BI cohorts, 10 samples are flagged because their assembled genome lengths are distant from the expected 3.0 Gbp. One sample in the HA_Rev_mid cohort is flagged for the same reason.

We observed that the samples that had high coverage, i.e. the high-pass samples, have *de novo* assembly lengths closer to the expected value (~3Gbp), whereas the samples that have mid coverage, i.e. the mid-pass samples, have shorter than expected assembled genome length. This is most likely due to coverage requirements from the hifiasm tool.

For continuity measures, we only saw one outlier sample, in the UW cohort, with lower continuity (Figure 19). We flagged but did not remove this sample.

The flagged samples are available in a list on the RW.

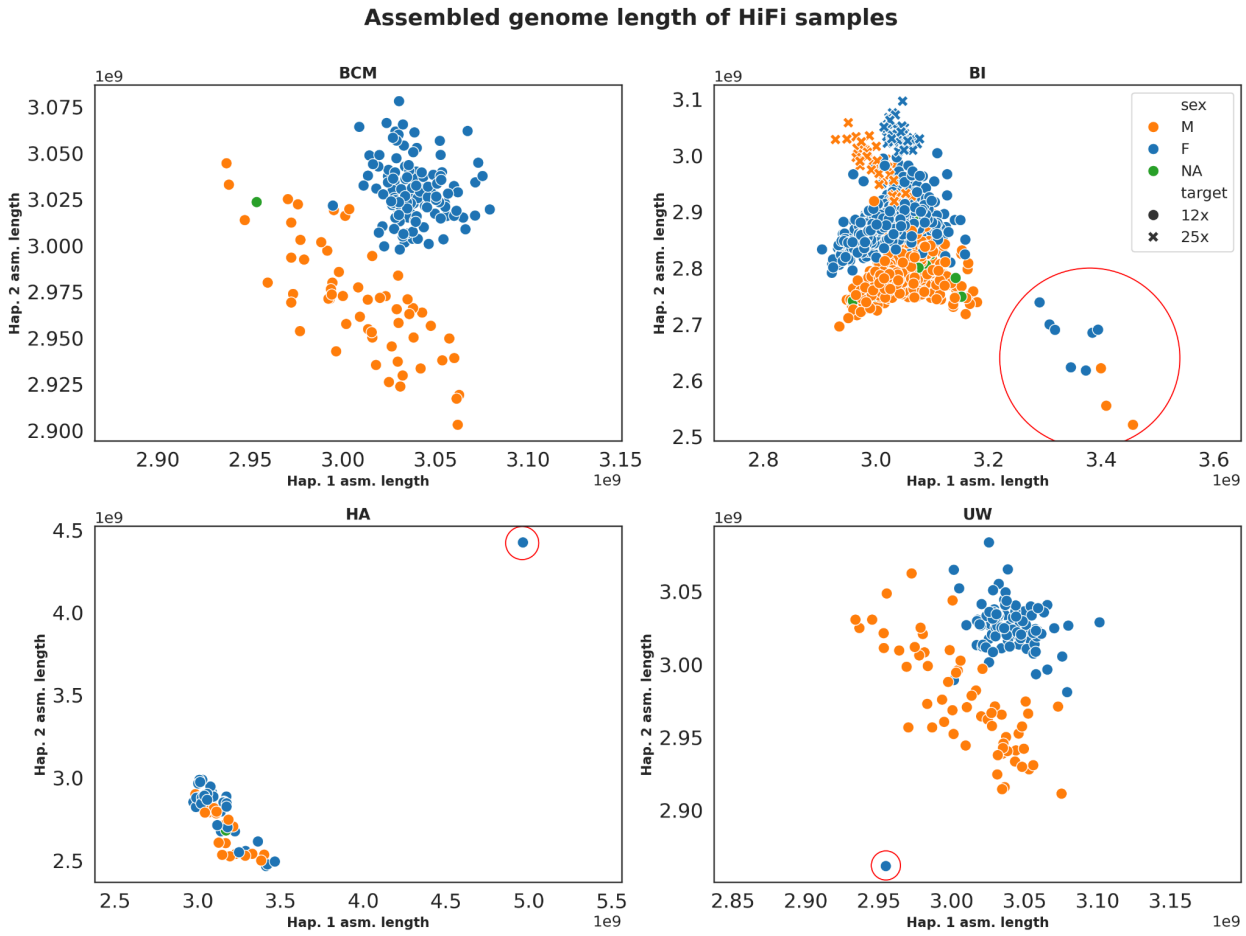


Figure 18 -- *De novo* assembled genome lengths of HiFi samples. Outlier samples are circled in red. Note that all HA samples are mid-pass (12x) and all BCM and UW samples are high-pass (25x).

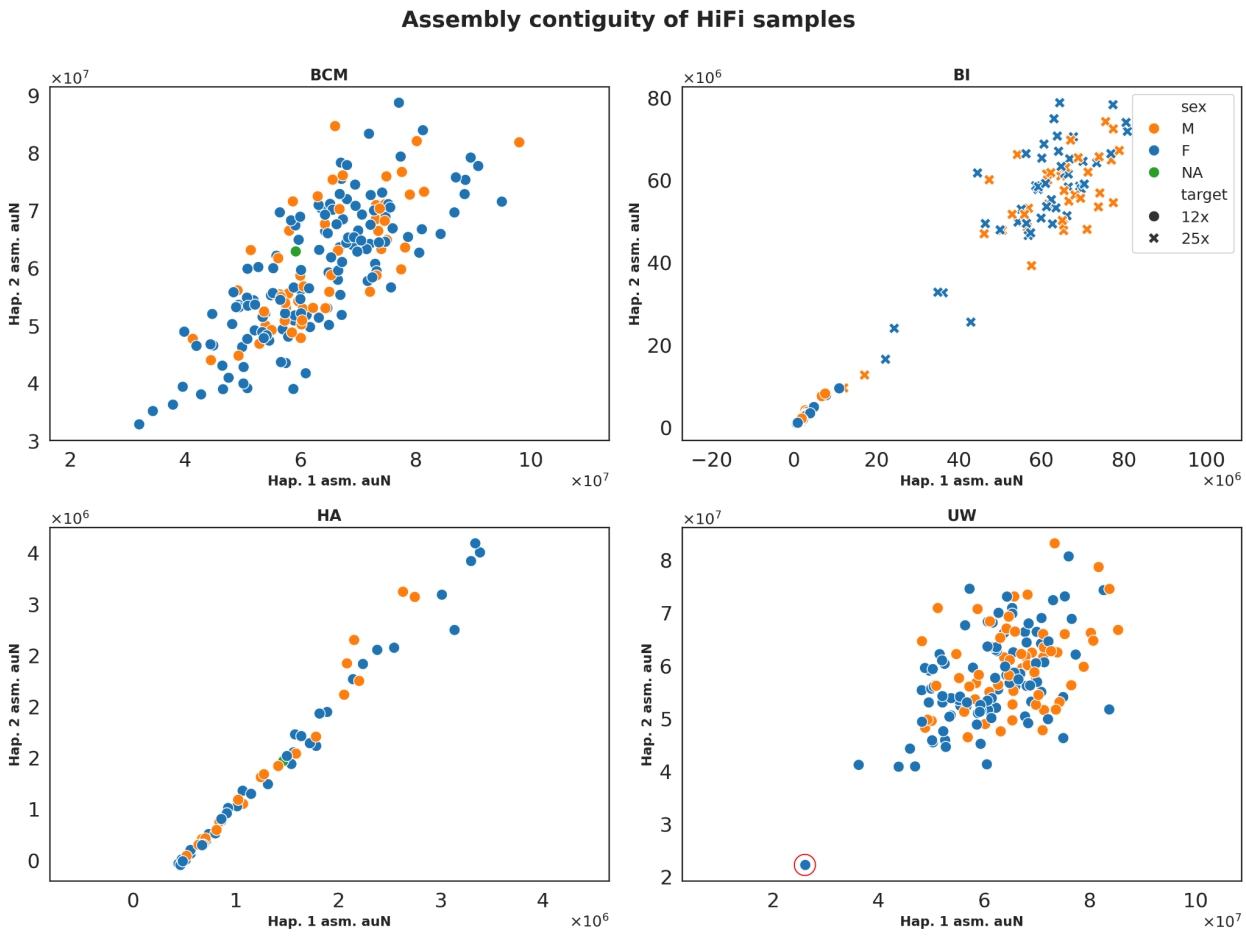


Figure 19 -- *De novo* assembly contiguity of HiFi samples. Outlier samples are circled in red. Note that all HA samples are mid-pass (12x), and all BCM and UW samples are high-pass (25x).

SNP and Indel QC

We performed SNP and Indel calling with DeepVariant [55] for each reference version for each sample individually [56]. Each GVCF per sample was then used to create joint callsets for grch38_noalt and T2Tv2.0, using GLNexus [57]. Final joint-called SNP and Indel callsets were converted to Hail MatrixTable (Hail MT) format for analysis. Joint callset QC is performed on the joint SNP and Indel callset from the long-read data on the grch38_noalt callset (Table 15).

Table 15 -- lrWGS joint callset QC steps

QC process	Error modes addressed	CDRv8 release results
Variant Hard Filter	- Low quality variants	

Variant Hard Filter

Method

We used the QUAL annotation at variant sites to filter out low quality variants. We evaluated four metrics for each cohort to determine the QUAL threshold that we would use: SNP Het/Hom ratio, Ti/Tv ratio, Ins/Del ratio, and variant count on the autosomes. These four metrics are demonstrated for each cohort in Appendix S.

Results

These metrics under different QUAL filter thresholds are displayed in Figures S.1-S.10 (Appendix S), for the various cohorts.

After the evaluation, we decided to use a QUAL score cutoff of any variant sites under 40 for PacBio HiFi samples and 34 for ONT samples.

Structural Variant QC

Method

We called SVs with Sniffles2 [32], PBSV, [31] and PAV [30]. For ONT samples, we did not use the PAV variant caller since it depends on the availability of haplotype-resolved assemblies.

Outliers are flagged by plotting the variant counts versus the coverage and manually evaluating the distribution for each cohort. The results are demonstrated in Appendix T and outlined in Table 16.

Results

See Table 15 for the total number of samples flagged as a result of the Structural Variant QC. We flagged four samples from the CDRv8 release due to their low SV counts (Table 16). The referenced figures are in Appendix T.

Table 16 -- SV results for each cohort

Cohort name	SV callers	Number of outliers	Figure	Notes
HA_Rev_mid	Sniffles, PBSV, PAV	3	Figure T.1, Figure T.2	
HA_Seq_CDRv7	Sniffles, PBSV, PAV	N/a	N/a	See CDRv7 QC report
BI_Seq_high	Sniffles, PBSV, PAV	0	Figure T.3, Figure T.4	
BI_Seq_mid				
BI_Rev_mid				

BCM_Seq_high	Sniffles, PBSV, PAV	0	Figure T.5, Figure T.6	
BCM_Rev_high				
BCM_ONT_high	Sniffles, PBSV	0	Figure T.7, Figure T.8	
JHU_ONT_high				
UW_Seq_high	Sniffles, PBSV, PAV	1	Figure T.9, Figure T.10	
UW_Rev_high				

Known Issues

The issues below apply to the CDRv8 release genomic data (arrays, srWGS, srWGS SVs, lrWGS, 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 on the Researcher Workbench, locations are in the Controlled CDR directory document.

Known Issue #1: Three samples were affected by a data quality issue

Three samples in the srWGS CDRv8 data release were affected by a data quality issue. We have provided a list file of the research IDs of affected samples that can be accessed via the RW.

Affects:

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

Suggested action:

- Remove affected samples from analysis. We provide a list file of research IDs of affected samples in the CDR, see the CDR Directory Document.

Remediation:

- The samples will be removed in the CDRv9 release.

Known Issue #2: Samples from previous release are missing in this release (N=2,684)

A total of 2,684 srWGS samples from the CDRv7 previous release were not included in the CDRv8 release. This happened for multiple reasons:

1. We reprocessed srWGS samples in the CDRv8 data release from DRAGEN 3.4.12 to 3.7.8, which resulted in some samples with new coverage metrics. As a result, some srWGS samples passed QC in the previous release, but did not pass QC in this release. This was expected behavior, but is being called out here for completeness.
2. The participants withdrew consent between releases.
3. Samples were dropped from the CDRv8 release due to an internal data harmonization issue (See CDRv8 Controlled Tier Release Notes). These participants will be added back in CDRv9.

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

Due to these issues, the lrWGS and srWGS SV callsets were affected.

There are 22 participants with lrWGS data that do not have matching CDR data in CDRv8. While there are 2800 participants with lrWGS data, when using the cohort builder or when using phenotypic data for the CDRv7 lrWGS data, you may see 2778. If you access the CDRv7

lrWGS callset within a CDRv8 workspace, you will not be able to access the matching phenotypic data for those 22 participants.

Additionally, the srWGS SV callset was affected by this issue. There were no added participants between CDRv7 off-cycle and CDRv8 and samples were removed from the callset due to the above reasons. The total number of participants for CDRv7 off-cycle was 97,937 and the total number of participants for CDRv8 is 97,061.

Affects:
- srWGS data, srWGS SV data, lrWGS data

Suggested action:
- No action necessary

Remediation:
- Where possible, we will add these samples back in CDRv9

Known Issue #3: Variants missing from variant search

A total of 3595 expected variants are missing from the SNP/Indel variant search function in the Cohort Builder due to a data comparison issue. None of the variants overlap ClinVar dataset or have pathogenic significance, and only 7% are within the exome.

The variant search uses data from the Variant Annotation Table (VAT) and the raw variant storage and there are some cases where the variants do not properly overlap causing them to not appear in the variant search. The variants affected by this issue are still available in the VAT and within the other genomic data files.

Affects:
- Cohort & Dataset Builder Variant Search

Suggested action:
- Use Cohort & Dataset Builder Variant Search as normal, and if an expected variant is missing, search the Variant Annotation Table (VAT)

Remediation:
- We will monitor this issue to determine if more variants become affected

Known Issue #4: ClinVar annotation missing for some variants in the VAT

We have identified a small number of variants that are missing their ClinVar annotation in the Variant Annotation Table (VAT). These variants are in the ClinVar database as the reverse complement of what we would expect. As a result, the variant annotator does not correctly provide their annotation. We have found a solution to this issue and will provide a fix in the next data release. Variants that are incorrectly missing their ClinVar classification will also be missing from the ClinVar smaller callsets.

Affects:
- Variant Annotation Table (VAT)
- srWGS ClinVar smaller callsets: Hail MT, VCF, PLINK

- Variant Search in the Cohort Builder

Suggested action:

- You can check if a ClinVar annotation of interest is in the reverse complement by if the gene appears on ClinVar as stop_codon<-start_codon

Remediation:

- We will provide a fix for this issue in the next data release.

Known issue #5: 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:

- 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 CDRv8 release.

Known issue #6: One site missing for all srWGS samples

Due to a variant calling issue in one sample, a single variant site is missing for all samples. The site is chromosome 4, position 190181387. The rsid of the sample with the variant calling issue is available in a list file in the Researcher Workbench. The variant information is still available for the rest of the samples, but requires extra steps on the part of researchers.

Affects:

- srWGS SNP & Indel callsets: VDS, VCF, Hail MT, and PLINK formats
- Variant Annotation Table (VAT)
- Variant Search in the Cohort and Dataset Builder
- Public Data Browser

Suggested Action:

- If you are interested in this variant site, access the callset in VDS format and filter out the individual sample with the Hail method `filter_samples`.
    - The rsid of the sample is in a list file, which can be found in the CDR Directory Document.
- Otherwise, no action.

Remediation:

- We will convert this genotype into a no call in the next release to remove the issue.

Known Issue #7: lrWGS CDRv8 T2Tv2.0 reference is different than the lrWGS CDRv7 T2Tv2.0 reference

The reference used for the lrWGS T2Tv2.0 reference has changed in the CDRv8 data release. The changes are that Chromosome Y is hardmasked with N bases in the PAR region and the mitochondrial genome in the reference is changed to the revised Cambridge Reference Sequence (rCRS). All other aspects of the CDRv8 T2Tv2.0 reference version stay the same as the reference used in CDRv7.

Affects:
- lrWGS variant data and alignment data on T2Tv2.0 reference

Suggested action:
- When comparing variants between releases, do not compare chrM, chrX, and chrY calls or use extra care

Remediation:
- The callsets are already separate and reflect this change

[SOLVED] Known Issue #8: BGEN ‘rsid’ is empty

In the original CDRv8 smaller callsets BGEN files, the ‘rsid’ field was empty, causing issues for some software, including PLINK and Regenie. We have re-created these BGEN files with the correct ‘rsid’ field. Researchers can access the corrected files in the existing BGEN file paths after 02/11/2025.

Affects:
- srWGS SNP & Indel smaller callsets: Binary GEN (BGEN) format

Suggested action:
- Utilize the remediated dataset for your research by accessing the files after 02/11/2025

Remediation:
- This issue has already been remediated. In future releases, we will perform automated testing to make sure that this issue does not occur again.

[SOLVED] Known Issue #9: Genomic extraction chromosome 5 files empty

In the original CDRv8 genomic extraction results, the files for chromosome 5 were empty. We have provided a fix for this issue and researchers can access the corrected files by running the genomic extraction after 03/10/2025.

Affects:
- srWGS data from the genomic extraction tool

Suggested action:
- Re-run your genomic extraction after 03/10/2025

Remediation:

- This issue has already been remediated. In future releases, we will perform automated testing to make sure that this issue does not occur again.

Known Issue #10: srWGS samples were affected by a data quality issue (N=4,044)

We have identified a data quality issue affecting 4,044 CDRv8 samples. These samples did not pass the srWGS coverage metrics mean coverage threshold of ≥30x and were erroneously included in the callset when they should have been removed. These samples will be removed from the next release.

Affects:

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

Suggested action:

- Remove affected coverage samples from analysis. We provide a list file of research IDs of affected samples in the CDR, see the CDR Directory Document.

Remediation:

- The samples will be removed in the next release.

Known Issue #11: PLINK BED and BGEN issues on the X and Y chromosomes

There is a known issue affecting variants on the X and Y chromosomes in the smaller callset PLINK BED and BGEN formats. Due to a bug in the data creation pipeline, haploid genotype (GT) calls at affected sites are stored incorrectly, resulting in inaccurate GT values. GTs for chrX and chrY variants may correctly appear as either haploid or diploid—even for chrY and chrX in males. This issue only affects haploid calls. Researchers working with X and Y chromosome variants, including non-pseudoautosomal region (nPAR) variants, are advised to use the PLINK 2 binary genotype table (PGEN), Hail MatrixTable (MT), or VCF formats instead.

Affects:

- srWGS SNP & Indel smaller callsets: PLINK BED and Binary GEN (BGEN) format

Suggested action:

- If you are working with variants on the X and Y chromosomes, especially in the non-pseudoautosomal (nPAR) regions, use other data formats such as PGEN or Hail MT.

Remediation:

- We will solve this bug in the next dataset release.

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. Do you have blood or saliva srWGS samples?

   We include both saliva and blood srWGS samples. We have performed an investigation into batch effects, which will be documented in a future report, posted on the User Support Hub [1]. Previously in CDRv7 and earlier, we only included blood samples. You can find the source of the sample in the genomic metrics auxiliary file.

3. 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).

4. Are there any genomic duplicates in the dataset?

   There are a small number of samples that we have identified as genomic duplicates, which were identified by a kinship score >0.40, see Appendix I. These samples may represent true individuals or may represent the same person submitting data multiple times.

   In the CDRv8 release, 43 samples are within clusters of three or more, and based on the rarity that all identical siblings are represented in *All of Us,* we believe that some of these clusters represent one individual who submitted data multiple times.

   To remove related samples from your analysis, you can use the maximal set of unrelated samples (see Appendix J). We have example code for this step in our featured workspace on working with genomic data, within the Hail GWAS notebook.

5. Do you call the challenging medically relevant autosomal genes (CMRG)?

   As identified in a previous report in Nature Biotechnology [58], challenging medically relevant autosomal genes (CMRG) are missing from many callsets due to limitations of current methods. We currently see reduced sensitivity in the srWGS dataset for these genes.

   As described previously in this report, we used the hg38 reference and GATK for variant calling. We have addressed 271 total genes by creating a separate callset with these genes using GATK and the masked hg38 reference.

6. Who does the genetic ancestry group ‘Americas’(1KGP-HGDP-AMR-like) include?

   This genetic ancestry group includes people who may be able to trace at least some of their distant ancestors back to North, Central, or South America. However, many of these people may also have some ancestors who came from other places, like Europe and Africa. People with combinations of Indigenous American genetic ancestry with European and/or African genetic ancestry are included in this category. It is important to acknowledge that these combinations are common in large part because of the shameful history of colonization and slavery in the Americas.

   It’s also important to recognize that having American genetic ancestry does not necessarily mean someone is a citizen of a Tribal Nation or a member of a Tribal community. Only Tribes and Tribal communities decide how to define their membership.

7. Why does the 1KGP-HGDP-MID-like genetic ancestry group have higher error rates?

   The 1KGP-HGDP-MID-like (MID or Middle Eastern) genetic ancestry group has higher error rates because of limitations of existing truth data in the 1KGP-HGDP-MID genetic ancestry group. We use existing truth datasets to train the random forest classifier (see Appendix G) and with a small 1KGP-HGDP-MID dataset, the confidence of the classifier dips within that group. The result is that a larger proportion of individuals in the 1KGP-HGDP-MID-like genetic ancestry group are classified as Remaining (OTH) when compared to other genetic ancestry groups.

   The VAT uses these genetic ancestry groups to generate the *All of Us* population annotations (gvs_mid_* and gvs_oth_*). When limiting cohorts to samples with 1KGP-HGDP-MID-like genetic ancestry (“mid”), use the ancestry predictions that do not include “Remaining Individuals”. In other words, use the “ancestry_pred” column, instead of “ancestry_pred_other”.

   This affects the variant annotations in the public Data Browser and the Variant Annotation Table (VAT).

   We have completed investigating other approaches and have minimized this error in the data. However, we do not have enough individuals in the 1KGP-HGDP-MID training dataset to fully remediate this issue.

8. Why do the genetic ancestry groups change between releases?

   In CDRv8, we have 6,192 CDRv8 participants (2.52% of the total CDRv7 srWGS samples) whose genetic ancestry classification changed to or from the “Remaining” category (see Appendix G). We believe that this is due to repeating the process of defining HQ sites for the training model (Appendix I) for each CDR data release.

   These ancestry changes may affect your analysis if you migrate from the previous release to a new release. We recommend that you re-run your downstream analyses

that are affected by the genetic ancestry categories. These include the VAT population level annotations (gvs_*_*) and genomic data in the public Data Browser.

9. How do you find a failed genotype for srWGS data?

   The srWGS SNP & Indel variants are released in VDS format (see the Variant Dataset (VDS) article). 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 (CDRv6 and earlier), we reported more filtering information.

10. Where is the QUAL field for the srWGS SNP & Indel variants?

    In the VDS format, the actual QUALApprox annotation is not included, which affects the VDS and also the smaller callsets (e.g., exome). Instead of using QUAL to filter variants, we recommend using the filter field to determine the quality of variants. Please see the Variant Dataset (VDS) article for more information.

References

[1] ***All Of Us User Support Hub*** support.researchallofus.org/
[2] The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024). https://doi.org/10.1038/s41586-023-06957-x
[3] 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
[4] CollectArraysVariantCallingMetrics (Picard), Retrieved October 21, 2021 , from https://gatk.broadinstitute.org/hc/en-us/articles/360037593871-CollectArraysVariantCallingMetrics-Picard-
[5] 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)
[6] 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
[7] ***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
[8] 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
[9] 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
[10] ***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.
[11] 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
[12] ***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.
[13] S. K. Lee, K. Degatano, G. Grant, G. Brandt, **New Variant Filtration Algorithm for the Genomic Variant Store** Published August 11, 2023. Retrieved August 21, 2024 from https://github.com/broadinstitute/gatk/blob/ah_var_store/scripts/variantstore/docs/release_notes/VETS_Release.pdf
[14] ***Relatedness - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/relatedness.html#hail.methods.pc_relate.
[15] ***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-.

[16] ***Resource bundle – GATK.*** (n.d.). Retrieved February 1, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle.
[17] ***The Omni Family of Microarrays.*** (n.d.). Retrieved February 16, 2022, from https://www.illumina.com/Documents/products/datasheets/datasheet_gwas_roadmap.pdf.
[18] International HapMap Consortium. **The International HapMap Project**. Nature. 2003 Dec 18;426(6968):789-96. doi: 10.1038/nature02168. PMID: 14685227.
[19] The 1000 Genomes Project Consortium, ***A global reference for human genetic variation***, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393
[20] 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.
[21] 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
[22] 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
[23] ***Structural Variants*** (n.d.). Retrieved March 3, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022476791323-Structural-Variants
[24] 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
[25] 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
[26] 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.
[27] Jakubek YA, Zhou Y, Stilp A, *et al.* Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. Nat Genet. 2023 Nov;55(11):1912-1919. doi: 10.1038/s41588-023-01553-1. Epub 2023 Oct 30. PMID: 37904051; PMCID: PMC10632132.
[28] Forsberg LA, Rasi C, Malmqvist N, *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet. 2014 Jun;46(6):624-8. doi: 10.1038/ng.2966. Epub 2014 Apr 28. PMID: 24777449; PMCID: PMC5536222.
[29] 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.
[30] P. Ebert, P. A. Audano, Q. Zhu et al., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
[31] **PacBio structural variant calling and analysis tools (PBSV)**, Retrieved March 3, 2023, from https://github.com/PacificBiosciences/pbsv.
[32] **Sniffles2 (PBSV)**, Retrieved March 3, 2023, from https://github.com/fritzsedlazeck/Sniffles
[33] ***SVConcordance - GATK*** (June 29, 2024). Retrieved September 13, 2024, from https://gatk.broadinstitute.org/hc/en-us/articles/27007917991707-SVConcordance-BETA.

[34] **XGBoostMinGqVariantFilter** (n.d.) Retrieved March 4, 2023, from unreleased GATK branch https://github.com/broadinstitute/gatk/tree/tb_recalibrate_gq
[35] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016
[36] 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.
[37] 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.
[38] Babadi, M., Fu, J. M., Lee, S. K., Smirnov, A. N., Gauthier, L. D., Walker, M., Benjamin, D. I., Zhao, X., Karczewski, K. J., Wong, I., Collins, R. L., Sanchis-Juan, A., Brand, H., Banks, E., & Talkowski, M. E. (2023). GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature genetics, 55(9),* 1589–1597. https://doi.org/10.1038/s41588-023-01449-0
[39] **VisualizeCnvs.wdl** (n.d.) Retrieved March 4, 2023, from https://github.com/broadinstitute/gatk-sv/blob/v0.26.5-beta/wdl/VisualizeCnvs.wdl
[40] Carvalho, C., Lupski, J. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**, 224–238 (2016). https://doi.org/10.1038/nrg.2015.25
[41] Collins, R. L., Glessner, J. T., Porcu, et al. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell, 185*(16), 3041–3055.e25. https://doi.org/10.1016/j.cell.2022.06.036
[42] Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics, 12*, 187–215. https://doi.org/10.1146/annurev-genom-082509-141802
[43] 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
[44] 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
[45] Li, Heng. “Which Human Reference Genome to Use?” *Heng Li's Blog*, 13 Nov. 2017, https://lh3.github.io/. Accessed 2 Mar. 2023.
[46] 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
[47] Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, et al. The complete sequence of a human Y chromosome. bioRxiv, 2022.
[48] 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
[49] Wang, Y., Zhao, Y., Bollas, A. *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 39, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

[50] 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
[51] Cheng, Haoyu, et al. "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm." *Nature methods* 18.2 (2021): 170-175.
[52] Gurevich, Alexey, et al. "QUAST: quality assessment tool for genome assemblies." **Bioinformatics** 2013 Apr 15;29(8):1072-5.
[53] Simão Felipe, et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." **Bioinformatics** 2015 Oct 1;31(19):3210-2.
[54] Heng Li, **auN: a new metric to measure assembly contiguity** Published April 08, 2020. Retrieved August 21, 2024 from https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity
[55] Poplin, Ryan, et al. "A universal SNP and small-indel variant caller using deep neural networks." **Nat Biotechnol.** 2018 Nov;36(10):983-987.
[56] 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
[57] 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
[58] Wagner, J., et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol 40, 672–680 (2022).
[59] 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.
[60] 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)
[61] National Academies of Sciences, Engineering, and Medicine. 2023. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field.* Washington, DC: The National Academies Press. https://doi.org/10.17226/26902.
[62] 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
[63] Katherine Chao and gnomAD Production Team, **Genetic Ancestry**. Published November 01, 2023. Retrieved August 21, 2024 from https://gnomad.broadinstitute.org/news/2023-11-genetic-ancestry/
[64] ***Downloads | gnomAD.*** (n.d.). Retrieved February 1, 2021, from https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg.
[65] 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.
[66] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *Journal of Machine Learning Research* 12, pp. 2825-2830, (2011).
[67] 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

[68] Erdős, P. **On cliques in graphs**, Israel Journal of Mathematics, 4 (4): 233–234, (1966), doi:10.1007/BF02771637, MR 0205874, S2CID 121993028
[69] ***Structural variant (SV) discovery*** (n.d.). Retrieved March 15, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022487952155-Structural-variant-SV-discovery\
[70] **WDL Specification**, from https://github.com/openwdl/wdl
[71] 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

Appendix A: 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 Eric Banks, Lee Lichtenstein - Broad Institute David Glazer - Verily Life Sciences
Biobank	Mine Cicek - Mayo Clinic

Appendix B: Array processing overview

See Figure B.3 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 CDRv7 data release (C2022Q4R9), cluster definition files (.egt) were 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.

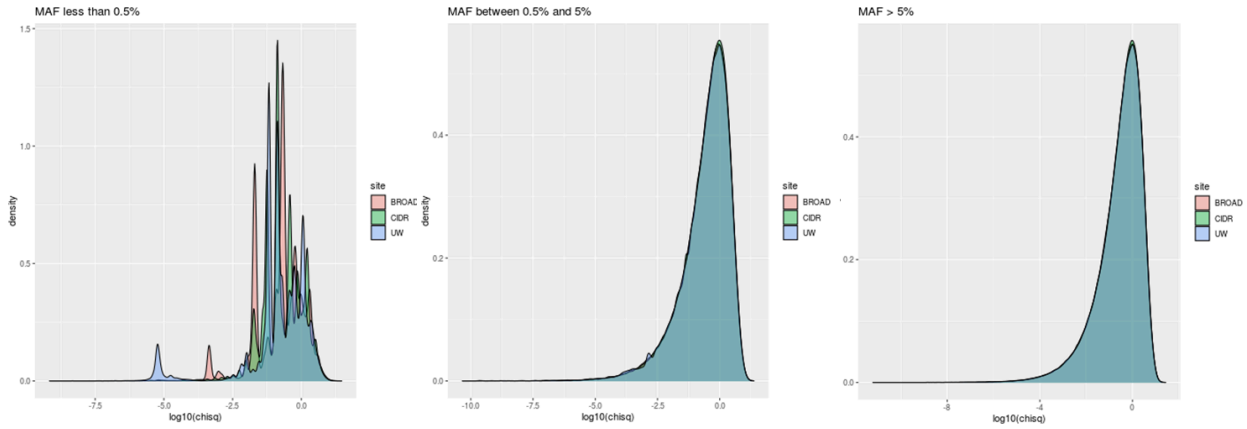


Figure B.1 Comparisons shown in Figure B.1 broken out into MAF bins.

	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 B.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 [5]
    - 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 [59]. Contamination is also measured using BAFRegress [5] 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 chrX versus normalized chrY intensity values for a batch of samples [59]. Any

sample discordant with 'sex assigned at birth' reported by an *All of Us* participant (see Appendix C) 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 C. 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.

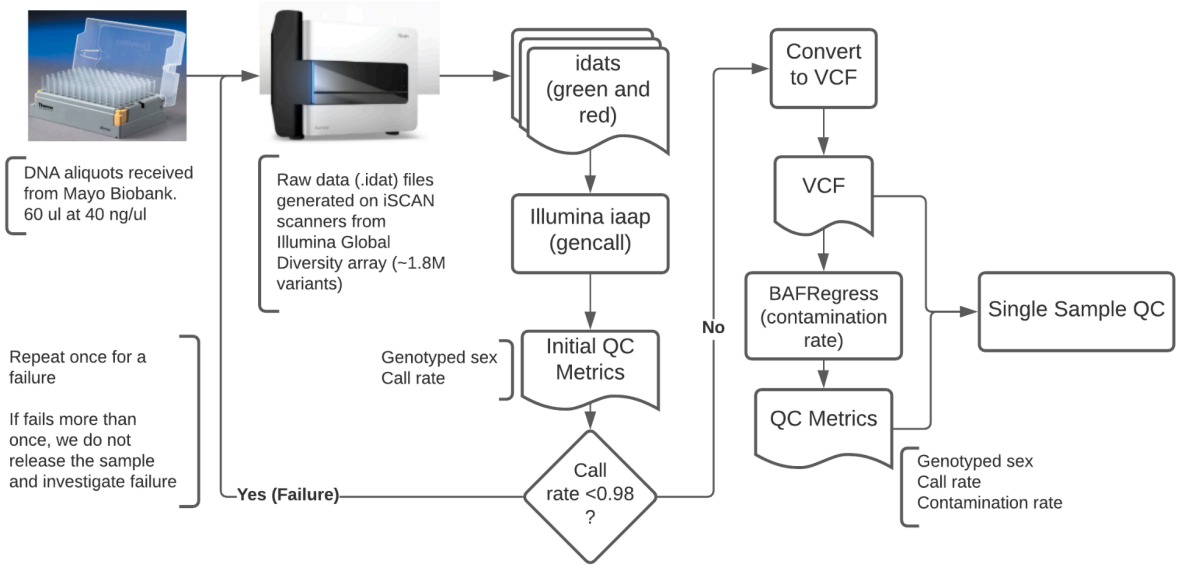


Figure B.3 -- Overview of the array processing pipeline.

Appendix C: Self-reported sex assigned at birth

See Table C.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.

Table C.1 -- CDRv8 release participants response breakdown to sex assigned at birth question

Sex assigned at birth responses	Array counts (%)	srWGS counts (%)	srWGS SV counts (%)	lrWGS counts (%)
Female	269848 (60.33%)	250071 (60.28%)	58514 (60.29%)	1759 (63.32%)
Male	172710 (38.61%)	160374 (38.66%)	37441 (38.57%)	985 (35.46%)
Other responses*	4720 (1.06%)	4385 (1.06%)	1106 (1.14%)	34 (1.22%)
Total	447278 (100.00%)	414830	97061	2778(100.00%)**

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.

**Please see Known Issue #2, as some lrWGS samples are missing CDR data and so are not reflected in this table.

Appendix D: *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 3 of the main text. In the CDRv8 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) [60]. 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 E: DRAGEN invocation parameters

Table E.1 summarizes the parameters used by the GCs to generate GVCFs, contamination estimates, and sex ploidy calls from the DRAGEN for srWGS data. All srWGS CDRv8 samples were reprocessed from DRAGEN 3.4.12 to 3.7.8.

Table E.1 DRAGEN 3.7.8 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 F: 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 eight well-characterized samples in the CDRv8 callset (Table F.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 [21].

The control samples are available for researchers on the Researcher Workbench. Please see the 'Controlled CDR directory document' for the locations.

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

Control Sample	Ground Truth	Genome Center	GVCF origin	Notes
HG-001_A	GiaB	BCM	DRAGEN 3.7.8	NA12878
HG-001_B	GiaB	UW	DRAGEN 3.7.8	NA12878
HG-002_A	GiaB	UW	DRAGEN 3.7.8	Ashkenzi Trio NA24385 - Son
HG-002_B	GiaB	BI	DRAGEN 3.7.8	Ashkenzi Trio NA24385 - Son
HG-003_A	GiaB	UW	DRAGEN 3.7.8	Ashkenazi Trio NA24149 - Father
HG-003_B	GiaB	BI	DRAGEN 3.7.8	Ashkenazi Trio NA24149 - Father
HG-004	GiaB	UW	DRAGEN 3.7.8	Ashkenazi Trio NA24143 - Mother
HG-005	GiaB	UW	DRAGEN 3.7.8	Han ancestry NA24631- Son

Genome Center:
BCM – Baylor College of Medicine
BI -- Broad Institute
UW -- University of Washington

Appendix G: Genetic Ancestry

Background

Genetic ancestry, as defined by the National Academies of Sciences, Engineering, and Medicine (NASEM), is “the paths through an individual’s family tree by which they have inherited DNA from specific ancestors” [61]. Each individual in the *All of Us* cohort necessarily has their own unique genetic relationship both to other members of the cohort and to previously sampled individuals from across the globe, determined by the familial relationships driven by chance encounters and forced or voluntary migration of ancestors across the history of the Americas.

Genetic ancestry is inferred by measuring the relative genetic similarity of each participant to global reference populations. As described by Katherine Chao and the gnomAD Production Team, “Genetic ancestry is a continuous measure, so any methods of creating discrete groups of individuals will inherently be inadequate.” [63] Although these groups have limitations, we believe that there are benefits to the broader scientific community to be able to study variants within populations [63]. In *All of Us,* we use genetic ancestry predictions in the population allele frequency calculations for annotated variants, which indicate how often a variant occurs in different populations. These calculations are available in the Variant Annotation Table (e.g. gvs_afr_ac) and data in the Genomic Variants section of the public Data Browser.

All of Us genetic ancestry methods

Genetic ancestry is inferred by measuring the genetic similarity of each participant to global reference populations. We compute these categorical groupings of genetic similarity to reference populations using harmonized continental metadata labels from the Human Genome Diversity Project (HGDP) [62] and 1000 Genomes Project training data [19] (N=3,942) for all srWGS samples in *All of Us*. We used the high-quality set of sites (or HQ sites, see Appendix I) to determine a genetic similarity label for each sample.

As genetic similarity is continuous, the groupings of the genetic similarity categories presented here are used to highlight genetic similarity between individuals to aid in variant classification and risk. The categories are based on the labels used in gnomAD [63,64], the HGDP and 1000 Genomes: We use the following acronyms or terms to describe genetic similarity to a reference population: 1KGP-HGDP-AFR-like (AFR or African); 1KGP-HGDP-AMR-like (AMR or Americas); 1KGP-HGDP-EAS-like (EAS or East Asian); 1KGP-HGDP-EUR-like (EUR or European); 1KGP-HGDP-MID-like (MID or Middle Eastern); 1KGP-HGDP-SAS-like (SAS or South Asian); and not belonging to one of the other ancestries or is an admixture (OTH or remaining individuals) (see Table G.1).

Table G.1 -- The *All of Us* genetic ancestry groups and the counts in each group

All of Us genetic ancestry group	Group acronym	*All of Us* Data Browser Genetic Ancestry Population name	CDR v8 Count (percentage)	Notes

1KGP-HGDP-AFR-like	AFR	African	79,826 (19.2%)	
1KGP-HGDP-AMR-like	AMR	Americas	71,854 (17.3%)	Who does the genetic ancestry group 'Americas'(1KGP-HGDP-AMR-like) include?
1KGP-HGDP-EAS-like	EAS	East Asian	9,440 (2.3%)	
1KGP-HGDP-EUR-like	EUR	European	223,350 (53.8%)	
1KGP-HGDP-MID-like	MID	Middle Eastern	810 (0.2%)	
1KGP-HGDP-SAS-like	SAS	South Asian	4,046 (1.0%)	
Remaining individuals	OTH	Remaining	25,504 (6.1%)	Not belonging to one of the other genetic ancestries or is an admixture
Total count			**414,830**	

We trained a random forest classifier [65,66] on a training set of the HGDP and 1000 Genomes samples variants (the HQ sites) on the autosomal exome, obtained from gnomAD (Table G.2). This exome was derived from the exon regions of all autosomal, basic, protein-coding transcripts in GENCODE v42 [67].

We generated the first 16 principal components (PCs) of the training sample genotypes (using the hwe_normalized_pca method in Hail) at the HQ sites for use as the feature vector for each training sample. We used the truth labels from the sample metadata, which can be found alongside the VCFs. Note that we do not train the classifier on the samples labeled as 'remaining individuals'. We use the label probabilities ('confidence') of the classifier to determine genetic similarity of these individuals. In cases where the confidence does not exceed 0.75 for any label, we apply the OTH/remaining individuals label.

Table G.2 -- The training set of HGDP and 1000 Genomes data

All of Us genetic ancestry group	Project	Count
1KGP-HGDP-AFR-like	1000 Genomes	841
1KGP-HGDP-AFR-like	HGDP	55
1KGP-HGDP-AMR-like	1000 Genomes	481
1KGP-HGDP-AMR-like	HGDP	60
1KGP-HGDP-EAS-like	1000 Genomes	581
1KGP-HGDP-EAS-like	HGDP	220

1KGP-HGDP-EUR-like	1000 Genomes	619
1KGP-HGDP-EUR-like	HGDP	148
1KGP-HGDP-MID-like	HGDP	126
1KGP-HGDP-SAS-like	1000 Genomes	596
1KGP-HGDP-SAS-like	HGDP	168
Remaining individuals (Note: we do not train on this category, we only test on this category)	1000 Genomes	5
Remaining individuals (Note: we do not train on this category, we only test on this category)	HGDP	42

As seen in Figure G.1, the projection shows a continuum of diversity in the *All of Us* cohort. Of individuals in the CDRv8 srWGS dataset, we estimate that 19.2% were similar to the 1KGP-HGDP-AFR individuals; 17.3% were similar to 1KGP-HGDP-AMR individuals; 2.3% were similar to 1KGP-HGDP-EAS individuals; 1.0% were similar to 1KGP-HGDP-SAS individuals; <1% were similar to 1KGP-HGDP-MID individuals; and 53.8% were similar to 1KGP-HGDP-EUR individuals (Table G.1).

We evaluated the performance of the ancestry predictions using a holdout set of the training samples. We tested performance with and without the Remaining individuals group.

1. Error rate (including Remaining individuals): 0.049 (See Table G.3)
    a. Please see the FAQ, Why does the 1KGP-HGDP-MID-like genetic ancestry group have higher error rates?, since the error rate is higher for 1KGP-HGDP-MID-like genetic ancestry. Our classifier conflates 1KGP-HGDP-MID-like and Remaining individuals.
2. Error rate (not including Remaining individuals): 0.001 (See Table G.4)

Please see the FAQ section for two FAQs involving genetic ancestry:

Who does the genetic ancestry group 'Americas'(1KGP-HGDP-AMR-like) include?
Why does the 1KGP-HGDP-MID-like genetic ancestry group have higher error rates?
Why do the genetic ancestry groups change between releases?

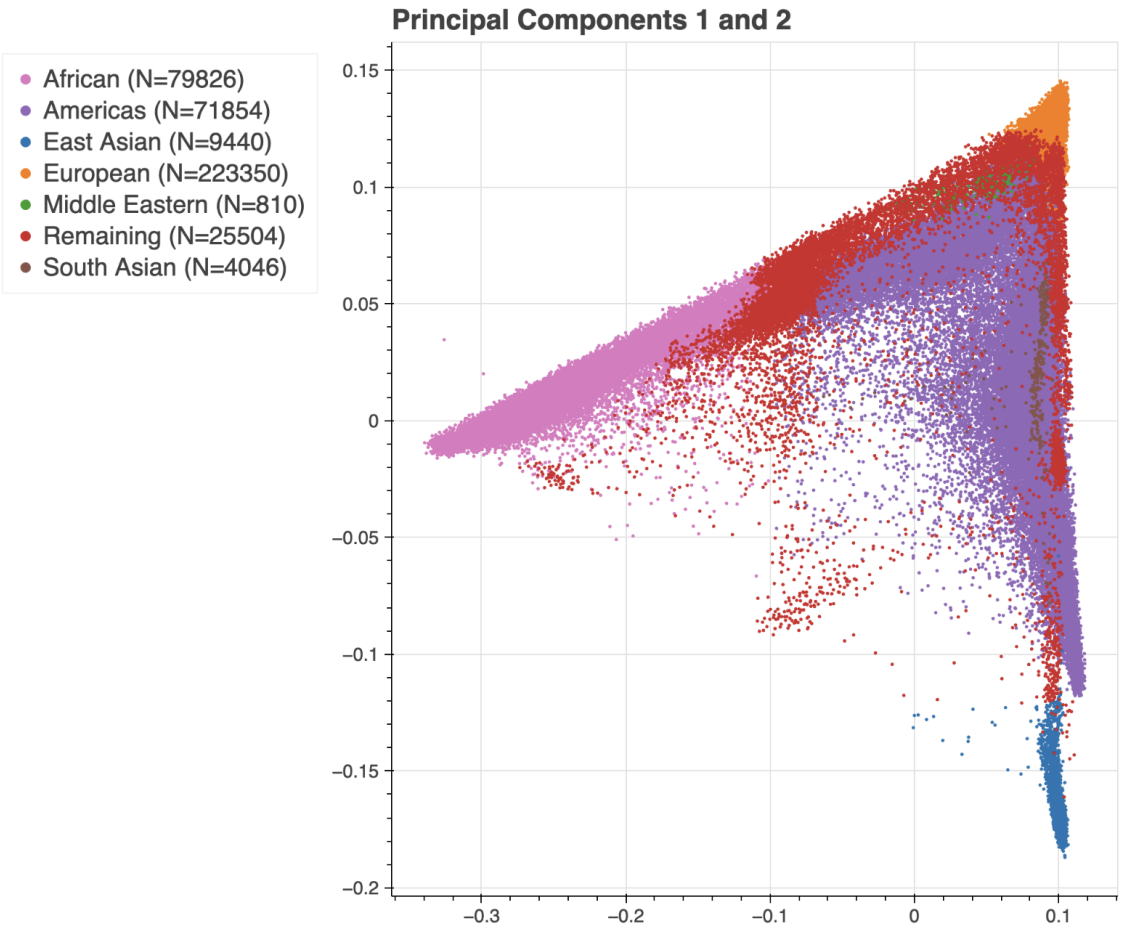


Figure G.1 -- Projection of the *All of Us* srWGS onto the PCA space of the 1000 Genomes and HGDP samples plotted on the first two principal components (PC1 on x-axis and PC2 on the y-axis) derived from genotype calls. Colored points represent six genetic ancestry groups.

Table G.3 -- Error rate (including Remaining individuals) on labeled training data using holdout set

	Predicted						
Actual	1KGP-HGDP -AFR-like	1KGP-HGDP -AMR-like	1KGP-HGDP -EAS-like	1KGP-HGDP -EUR-like	1KGP-HGDP -MID-like	Remaining individuals	1KGP-HGDP -SAS-like
1KGP-HGDP -AFR-like	198	0	0	0	0	2	0
1KGP-HGDP -AMR-like	0	50	0	0	0	0	0
1KGP-HGDP -EAS-like	0	0	199	0	0	1	0
1KGP-HGDP -EUR-like	0	0	0	197	0	3	0

1KGP-HGDP-MID-like	0	0	0	0	47	3	0
Remaining individuals	0	2	2	3	22	11	7
1KGP-HGDP-SAS-like	0	0	0	0	0	1	199

Table G.4 -- Error rate (not including Remaining individuals) on labeled training data using holdout set

	Predicted					
Actual	1KGP-HGDP-AFR-like	1KGP-HGDP-AMR-like	1KGP-HGDP-EAS-like	1KGP-HGDP-EUR-like	1KGP-HGDP-MID-like	1KGP-HGDP-SAS-like
1KGP-HGDP-AFR-like	200	0	0	0	0	0
1KGP-HGDP-AMR-like	0	50	0	0	0	0
1KGP-HGDP-EAS-like	0	0	199	0	0	1
1KGP-HGDP-EUR-like	0	0	0	200	0	0
1KGP-HGDP-MID-like	0	0	0	0	50	0
1KGP-HGDP-SAS-like	0	0	0	0	0	200

Appendix H: Self-reported race/ethnicity

As seen in Table H.1 and Table H.2, 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.

*The “Remaining” category refers to participants whose response did not match with the other categories shown in the table.

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

Self-reported Race/Ethnicity	Array counts (%)	srWGS counts (%)	srWGS SV counts (%)
AI/AN	4704 (1.05%)	4261 (1.03%)	0
AI/AN, White	5612 (1.25%)	5123 (1.23%)	0
Asian	14230 (3.18%)	13113 (3.16%)	2892 (2.98%)
Asian, White	2029 (0.45%)	1866 (0.45%)	384 (0.40%)
Black	76468 (17.10%)	71161 (17.15%)	22390 (23.07%)
Black, White	2520 (0.56%)	2358 (0.57%)	630 (0.65%)
Hispanic	72253 (16.15%)	66809 (16.11%)	16718 (17.22%)
Hispanic, White	7268 (1.62%)	6720 (1.62%)	1328 (1.37%)
MENA	2375 (0.53%)	2199 (0.53%)	495 (0.51%)
MENA, White	1451 (0.32%)	1333 (0.32%)	289 (0.30%)
White	233879 (52.29%)	217277 (52.38%)	48007 (49.46%)
Remaining*	19026 (4.25%)	17489 (4.22%)	2672 (2.75%)
Skip	5463 (1.22%)	5121 (1.23%)	1256 (1.29%)
Total	447278 (100.00%)	414830 (100%)	97061 (100%)

**Please see Known Issue #2, as 22 lrWGS samples are missing CDR data and so are not reflected in this table.

***The “Remaining” category in Table H.2 includes the categories “Asian, White”, and “MENA, White” in order to follow the *All of Us* Data and Data and Statistics Dissemination Policy

Table H.2 -- Self-reported Race/Ethnicity breakdown of the lrWGS samples

Self-reported Race/Ethnicity	lrWGS counts (%)
AI/AN	0
AI/AN, White	0
Asian	198 (7.13%)

Black	1209 (43.52%)
Black, White	73 (2.63%)
Hispanic	783 (28.19%)
Hispanic, White	43 (1.55%)
MENA	24 (0.86%)
White	261 (9.40%)
Remaining***	155 (5.58%)
Skip	33 (1.19%)
Total	2778 (100%)**

Appendix I: 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 [12]:

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 130660 high-quality (HQ) sites in the CDRv8 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 J: 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 approximately half of the fraction of the genetic material shared (ranges from 0.0 - 0.5, though the value can be higher than 0.5 for identical twins).

- Parent-child or siblings: 0.25
- Identical twins: greater than 0.45

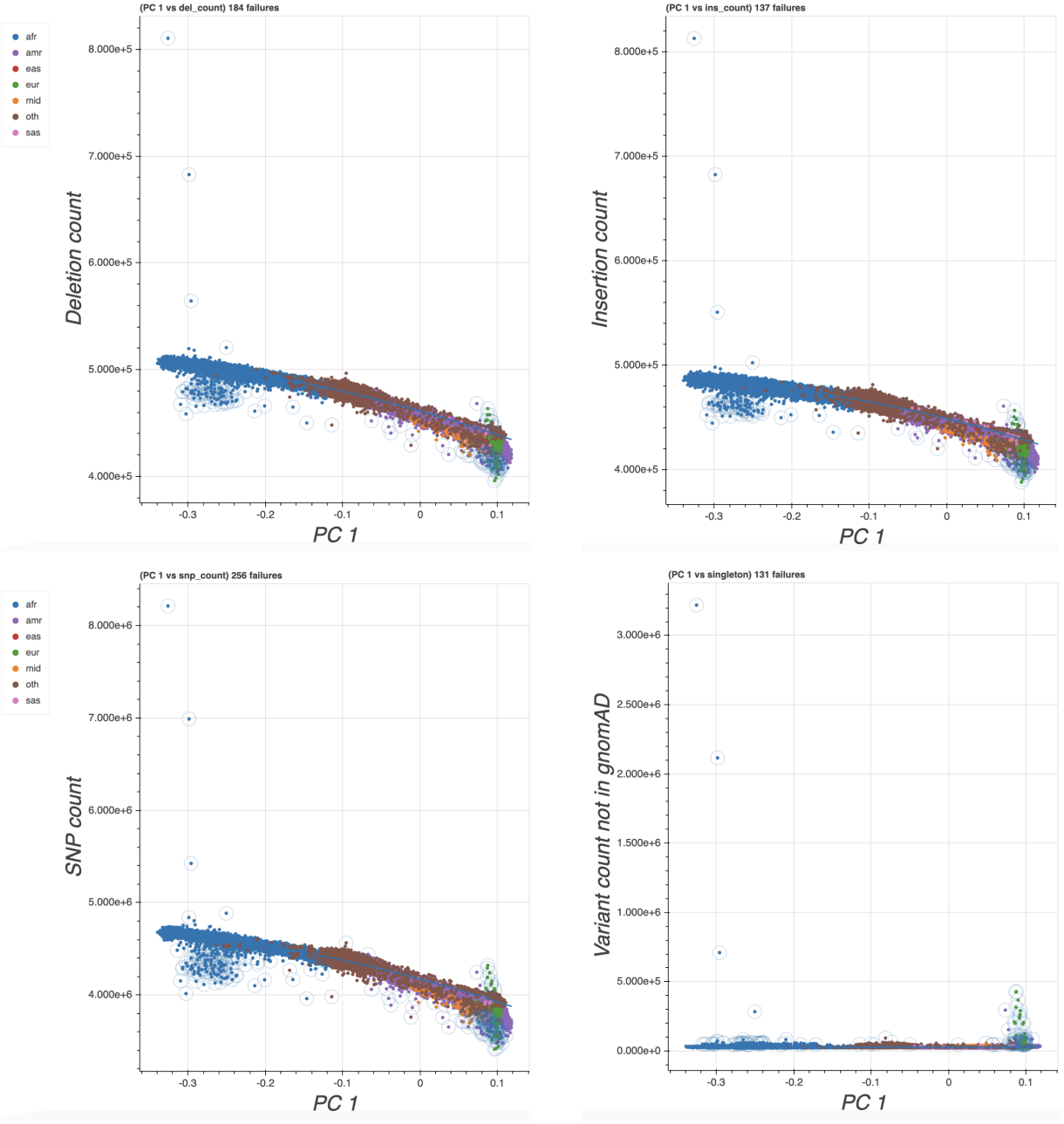
Please see the Hail pc_relate function [14] documentation for more information, including interpretation.

We will determine the maximal independent set [68] for related samples to minimize the number of samples that would need pruning. Using the HQ sites identified in Appendix I, researchers can remove first and second degree relatives.

We estimated 39,682 related pairs and 30,585 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 K: Plots of the first principal component against population outlier QC metrics

Figure K.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 K.1.



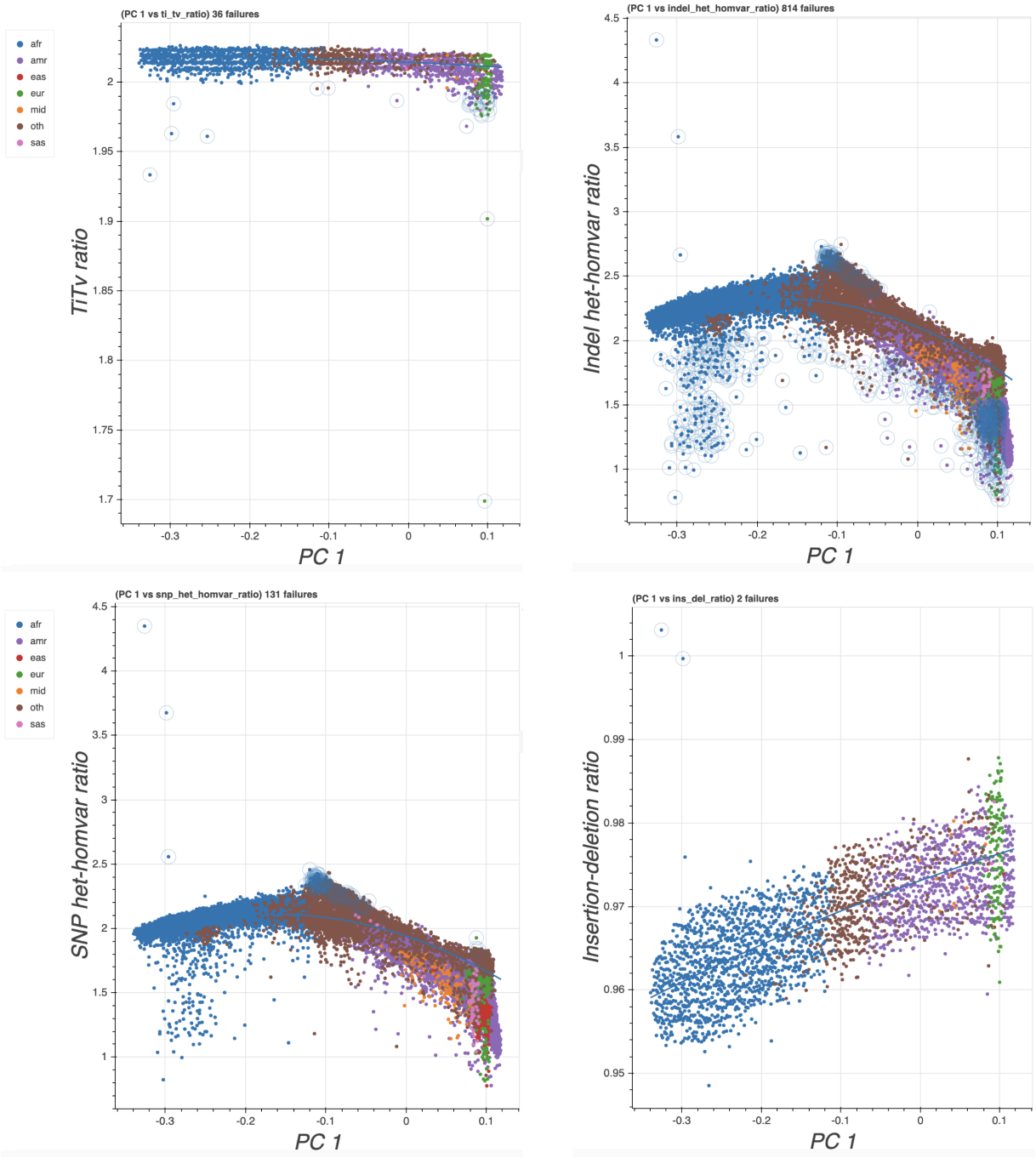


Figure K.1 -- Sample population outlier plots for eight metrics (see Population Outlier Flagging). Each metric (y-axis) is plotted against the first (of sixteen) principal components (x-axis). Outliers are identified by regressing out the principal components and determining if the residual is over 8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

Appendix L: srWGS Structural Variant Pipeline

The GATK-SV pipeline was applied to detect SVs from srWGS data [41]. 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 [24] and Wham [25] to leverage PE and split-read (SR) evidence, MELT [26] to specifically target mobile elements, and GATK-gCNV [38] and cn.MOPS [37] 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 [69], and details of the method can be found in Collins et al 2020 [41]. 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) [70].

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

- More precise SR-based genotyping and breakpoint determination for INS variants
- Refined functional consequence annotations for CPX variants
- Added annotations of allele frequency from gnomAD-v4.1 SVs for variants present in both callsets [71]
- Improved the depth-based genotyping method for very large CNVs to address an issue observed and manually fixed in the v7 srWGS SV callset
- Performance and scaling enhancements

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

Figure 7 depicts the steps of the pipeline as it was run for *All of Us*. Table L.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 L.1-- GATK-SV Pipeline Versions and Notes

Workflow/Step Name	Version Used	Entity	Notes
Sample selection	Notebook		See Sample Selection
GatherSampleEvidence	v0.24-beta	Sample	SV callers used: Manta, Wham, and MELT. All 88,882 samples completed this step, with a 0.00% initial failure rate.

EvidenceQC	v0.26.6-beta	Arbitrary partition of samples	Run on arbitrary partitions of samples.
Single sample QC	Notebook		See Single Sample QC
Batching	Notebook		See Batching
TrainGCNV	v0.24-beta	Batch	Batches of samples were created according to the scheme described in the main text under Batching
GatherBatchEvidence	v0.26.7-beta	Batch	Depth-based CNV callers used: GATK g-CNV and cn.MOPS.
ClusterBatch	v0.25.1-beta	Batch	
PlotSVCountsPerSample	v0.27.1-beta	Batch	
SubsetVcfBySamples	v0.27.1-beta	Batch	We removed the 11 significant outliers identified for duplication and deletion counts (nIQR cutoff = 10).
GenerateBatchMetrics	In development (git commit 769811f2)	Batch	This version has since been merged and released as v0.28-beta
FilterBatchSites	v0.24.3-beta	Batch	
PlotSVCountsPerSample	v0.27.1-beta	Batch	No SV count outliers observed.
FilterBatchSamples	v0.26.10-beta	Batch	No outlier samples were removed at this stage (nIQR cutoff = 10000).
MergeBatchSites	v0.24-beta	Cohort	For cohort-level steps, data from all samples across all batches was merged.
GenotypeBatch	v0.28.1-beta	Batch	
RegenotypeCNVs	v0.28.1-beta	Cohort	
CombineBatches	v0.24-beta	Cohort	
ResolveComplexVariants	v0.28.2-beta	Cohort	
GenotypeComplexVariants	In development (git commit 424ca4f)	Cohort	A developmental version of GenotypeComplexVariants was used for improved scaling
CleanVcf	v0.28.3-beta	Cohort	

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

Appendix M: srWGS SV overall precision and recall after SL filtering

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

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

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

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

Appendix N: Long-read workflow overview

The following figures summarize the workflows utilized to process the lrWGS data.

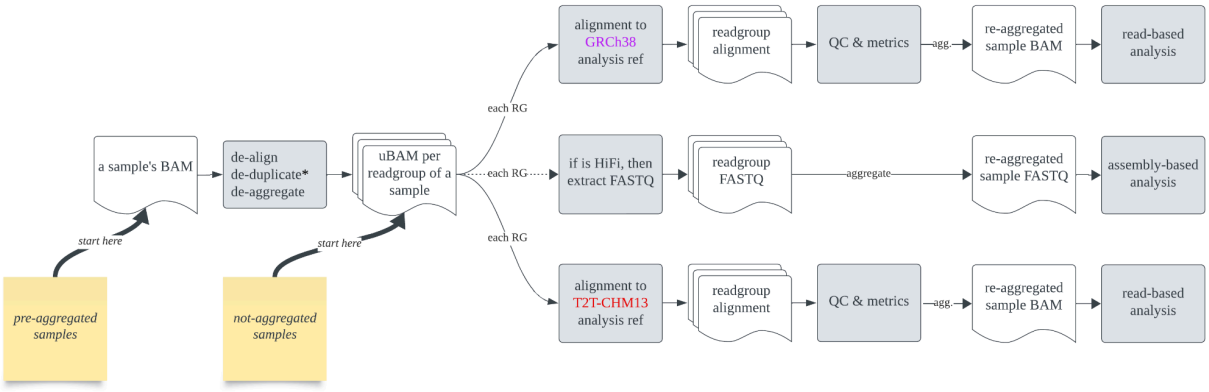


Figure N.1 -- The pre-processing and processing steps at the DRC for each lrWGS sample.

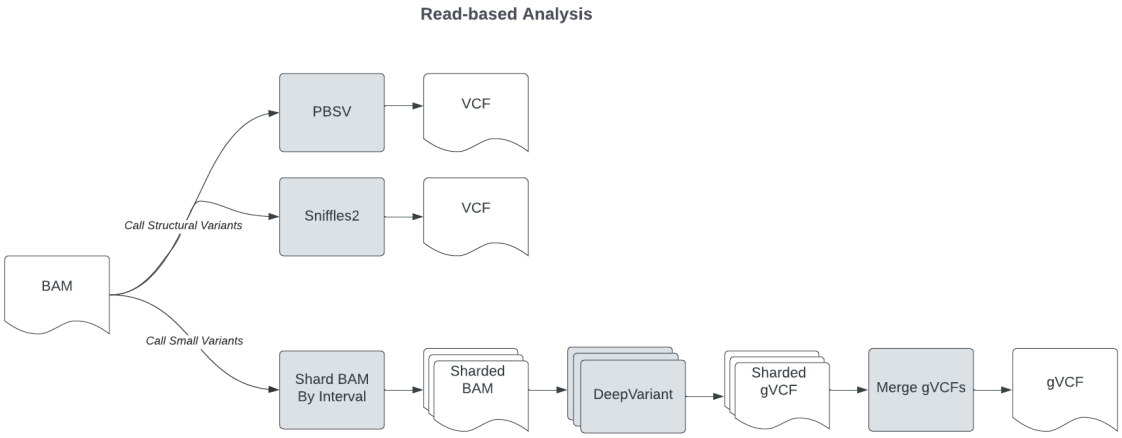


Figure N.2 -- The lrWGS variant calling steps, applied on both the grch38_noalt and the T2Tv2.0 references.

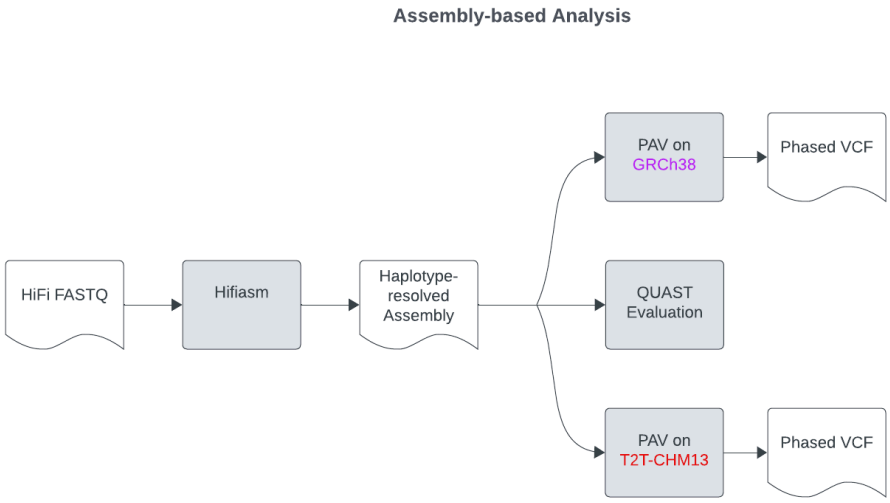


Figure N.3 -- lrWGS *de novo* assembly steps, for all cohorts with PacBio HiFi sequencing data.

Appendix O: Long-read pipeline tool versions and parameters

Table O.1 – lrWGS pipeline software versions and parameters

Software	Version used	Functionality	Invocation parameters

`minimap2`	2.26 (r1175)	ONT reads alignment.	`minimap2 \`
			`  -ayYL --MD --eqx --cs \`
			`  -x map-ont \`
			`  <reference.fasta> \`
			`  <unaligned.ont.fastq>`

`pbmm2`	packaged in smrtlink 12.0.0.176214	HiFi reads alignment.	`pbmm2 align \`
			`  <unaligned.hifi.bam> \`
			`  <reference.fasta> \`
			`  --preset CCS \`
			`  --sample <sample_name> \`
			`  --strip --sort --unmapped`

`gatk CheckFingerprint`	4.2.0.0	Check lrWGS BAM fingerprint	`gatk CheckFingerprint \`
			`  --INPUT <aligned_bam> \`
			`  --GENOTYPES <fingerprint_vcf> \`
			`  --EXPECTED_SAMPLE_ALIAS <vcf_sample_name> \`
			`  --HAPLOTYPE_MAP <haplotype_map> \`
			`  --OUTPUT <prefix>`

`VerifyBamID`	2.0.1	Estimate lrWGS BAM cross-individual contamination	`/VerifyBamID/bin/VerifyBamID \`
			`  --SVDPrefix /VerifyBamID/resource/1000g.phase3.10k.b38.vcf.gz.dat \`
			`  --Reference <reference.fasta> \`
			`  --PileupFile <pileup_converted_from_BAM>`

`mosdepth`	0.3.1=h4dc83fb_1	Coverage estimation	`mosdepth \`
			`  -x -n -Q1 \`
			`  <prefix> \`
			`  <bam_file>`

`samtools`	1.18	BAM aggregation, conversion to FASTQ, indexing of BAM, and BAM file MD tag calculation	Aggregation
			`samtools merge \`
			`  -p \`
			`  -c \`
			`  --no-PG -@ 2 --write-index \`
			`  -o <agg.bam> \`
			`  <input.bam>[,<input.bam>,...]`
			
			Conversion

			samtools fastq \
			  -0 <output.fastq> \
			  <input.bam>
			
			Indexing
			samtools index \
			  <bam>
			
			Aggregation
			samtools calmd \
			  -b <input.bam> \
			  <reference.fasta> \
			  > <agg.bam>

hifiasm	0.19,5	*de novo* assembly.
		
		Note that we generate BIN files first, then later when hifiasm resumes, it automatically detects these BIN files to resume assembly. This helps saving computational costs.
			Bin generation
			hifiasm \
			  --bin-only \
			  -o <output_prefix> \
			  -t <cpu_cores_to_use> \
			
			<input.fastq>[,<input.fastq>,...]
			
			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	packaged in smrtlink 12.0.0.176214	Single sample SV calling.
		
		For each sample, the svsig files are generated per chromosome, followed by VCF generation using all svsig files from all chromosomes.
			pbsv discover \
			  --tandem-repeats <trf.bed> \
			  <one_chromosome.bam> \
			  <output.svsig.gz>
			
			pbsv call \
			
			  –ccs \
			<reference.fasta> \
			
			  <chr.svsig.gz>, ..,
			<chr.svsig.gz> \
			
			  <output.vcf>

sniffles2	2.2	Single sample SV calling	sniffles \
			  -i <input.bam> \
			  --minsvlen 20 \
			  --tandem-repeats <trf.bed> \

			`--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 was used	
`pav` (WDL pipeline)		Single sample SV and small variant calling from hifiasm-generated 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`

`DeepVariant`	1.6.0	Single sample SNP and Indel variant calling.
		
		Model_type for ONT reads is “ONT_R104”, and for HiFi reads is “PACBIO”.
			For male samples:
			`/opt/deepvariant/bin/run_deepvariant \`
			`--model_type=~{model_type} \`
			`--ref=<reference.fasta> \`
			`--haploid_contigs “chrX,chrY” \`
			`--par_regions_bed <PAR.bed> \`
			`--reads=<BAM> \`
			`--output_vcf=<output_vcf.gz> \`
			`--output_gvcf=<output_gvcf.gz> \`
			`--num_shards=<num_core>`
			
			For female samples:
			`/opt/deepvariant/bin/run_deepvariant \`
			`--model_type=~{model_type} \`
			`--ref=<reference.fasta> \`
			`--reads=<BAM> \`
			`--output_vcf=<output_vcf.gz> \`
			`--output_gvcf=<output_gvcf.gz> \`
			`--num_shards=<num_core>`

`QUAST`	`5.2.0`	Assembly quality evaluation	`quast \`
			`--no-icarus \`
			`--large \`
			`<assembly.fa>, [<assembly.fa>, ...]`

`nanoplot`	Git hash e0028d85ec9e61f8c96b	Various alignment metrics collection	`NanoPlot \`
			`-c orangered \`
			`--N50 \`

	ea240ffca65b713e3385		`--tsv_stats \`
			`--no_supplementary \`
			`--verbose \`
			`--bam <BAM>`

`GLnexus`	1.4.3	Joint-calling SNPs and InDels from single sample gVCFs.	`glnexus_cli \`
			`--config 'DeepVariantWGS' \`
			`--bed <range.bed> \`
			`--list [gVCF, gVCF, ...] \`
			`> <output.bcf>`

`Hail`	0.2.130	Convert joint-called VCF to Hail MatrixTable.	`Hail python API is used to read the joint-called VCF into memory (via Hail function 'import_vcf') as a Hail MatrixTable, then it is written to disk (via Hail function 'write').`

Appendix P: Long-read contamination pipeline analysis

We evaluated VerifyBamID2 for its performance on lrWGS data, since it is a tool originally made for short read data. We tested the 3% contamination cutoff 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:

1. Cross contamination from a sample from a different population and of opposite sex.
2. Cross contamination from a sample from a different population and of the same sex.
3. Cross contamination from a sample from the same population of different sex.
4. 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 around 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 could impact the CDRv8 samples that are in the BI_Seq_25 cohort, as those samples are related. Following sequencing facility protocols, the relevant sequencing facility took efforts to randomize the order samples are prepared into libraries and carried onto sequencers, further minimizing the chance of related-sample contaminations.

Appendix Q: Long-read comparison of read length vs coverage

We investigated the relationship between read length and coverage across the sequencing facilities and sequencing technologies for the lrWGS CDRv8 newly released data. We found that there was no strong correlation between coverage and median read length across the sequencing facilities and platforms (Figure S.1).

For example, for BCM and BI, despite the variation in coverage, the median read length remains relatively stable. Similarly, in the HA and UW datasets, distinct clusters of samples exhibit varying levels of coverage, yet these changes do not correspond to significant shifts in median read length. The lack of a distinct pattern indicates that within this dataset, coverage and median read length are independent variables, with changes in one not directly influencing the other.

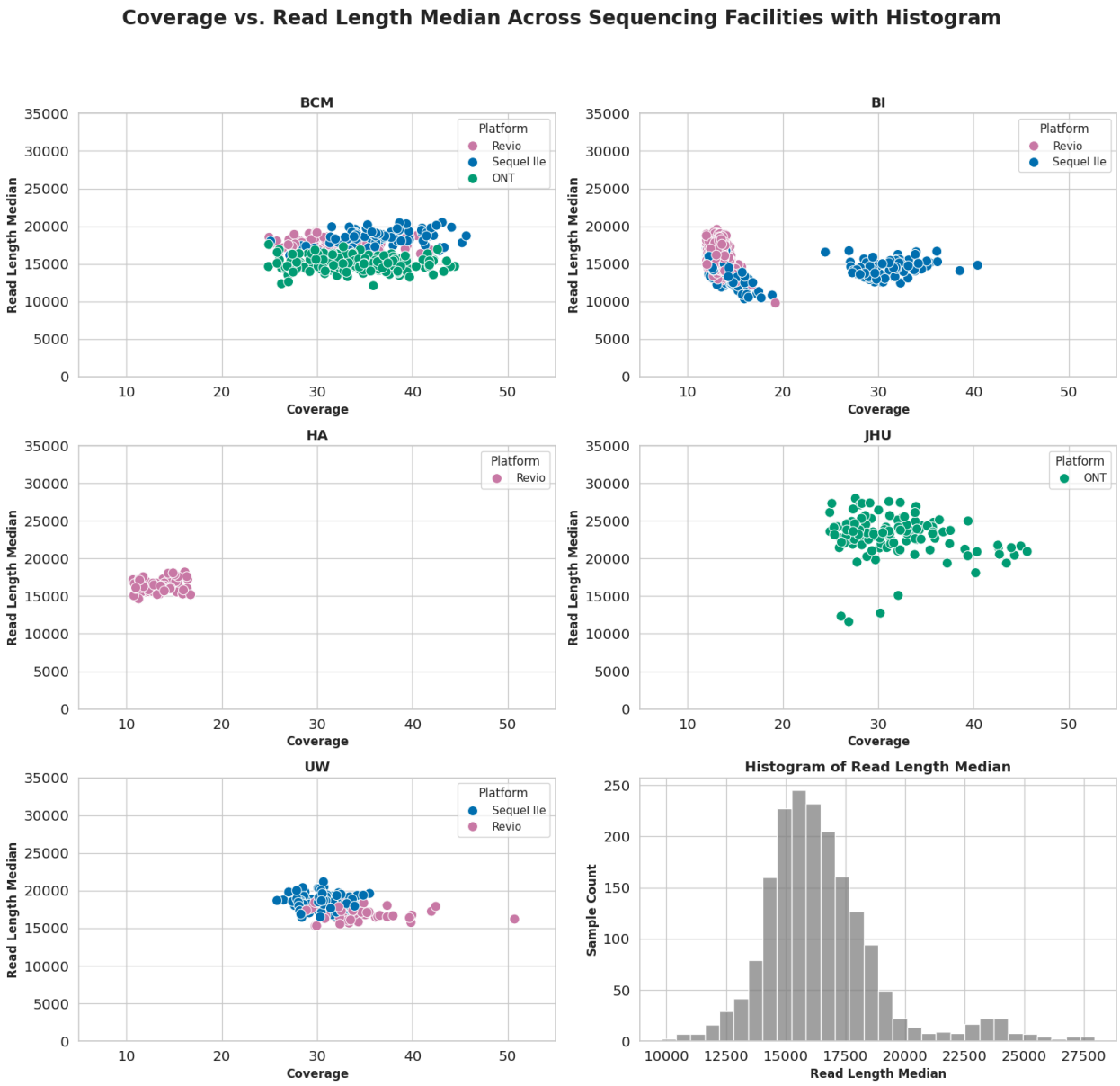


Figure Q.1 -- Coverage vs. read length median across sequencing facilities and platforms. Each subplot corresponds to a specific sequencing facility, with data points color-coded by sequencing platform (Sequel IIe, Revio, ONT). The last subplot provides an overview of all samples.

Appendix R: Long-read batch effect analysis

First, we checked several factors (mapping rate, chimera rate, error rate on the T2Tv2.0 reference, and the read length) in the CDRv8 lrWGS dataset that may impact variant calling (Figure R.1). As expected, we found that there were multiple factors in the lrWGS dataset that would lead to batch effects, which is why we separated the data into cohorts for joint-calling.

We observed that the ONT data had a lower mapping rate than the HiFi data, regardless of the reference used.

All three platforms show distinct patterns of “chimera” rate, where “chimera” is defined as those molecules that have non-contiguous mapping (or supplementary alignments, using the SAM specification terminology). A systematically different “chimera” rate will impact SV calling performance since supplementary alignments is a major source of signal used by most SV detection algorithms. Depending on how DeepVariant treats supplementary alignments exactly, it may also be impacted by these patterns.

Expectedly, the three platforms show different read error rate patterns, which will impact SNP and Indel qualities since base qualities are essential there.

Read-lengths aren’t obviously correlated with particular platforms, although JHU ONT samples tend to have longer reads overall. Read length impacts SV calling performance because longer read lengths tend to empower larger-sized SV detection.

Lastly, we observed that coverage impacts structural variant discovery, with higher coverage samples tending to have more SVs called with the read-based analysis (see Appendix T). It is tricky to evaluate the effect of coverage on SNPs and Indels discovery, since they are joint-called and impacted by cohort size as well as ancestry of the samples. That being said, it is reasonable to expect differences in genotyping qualities with different coverages. But because each lrWGS sample has corresponding srWGS data, this problem can be remedied, at least partially.

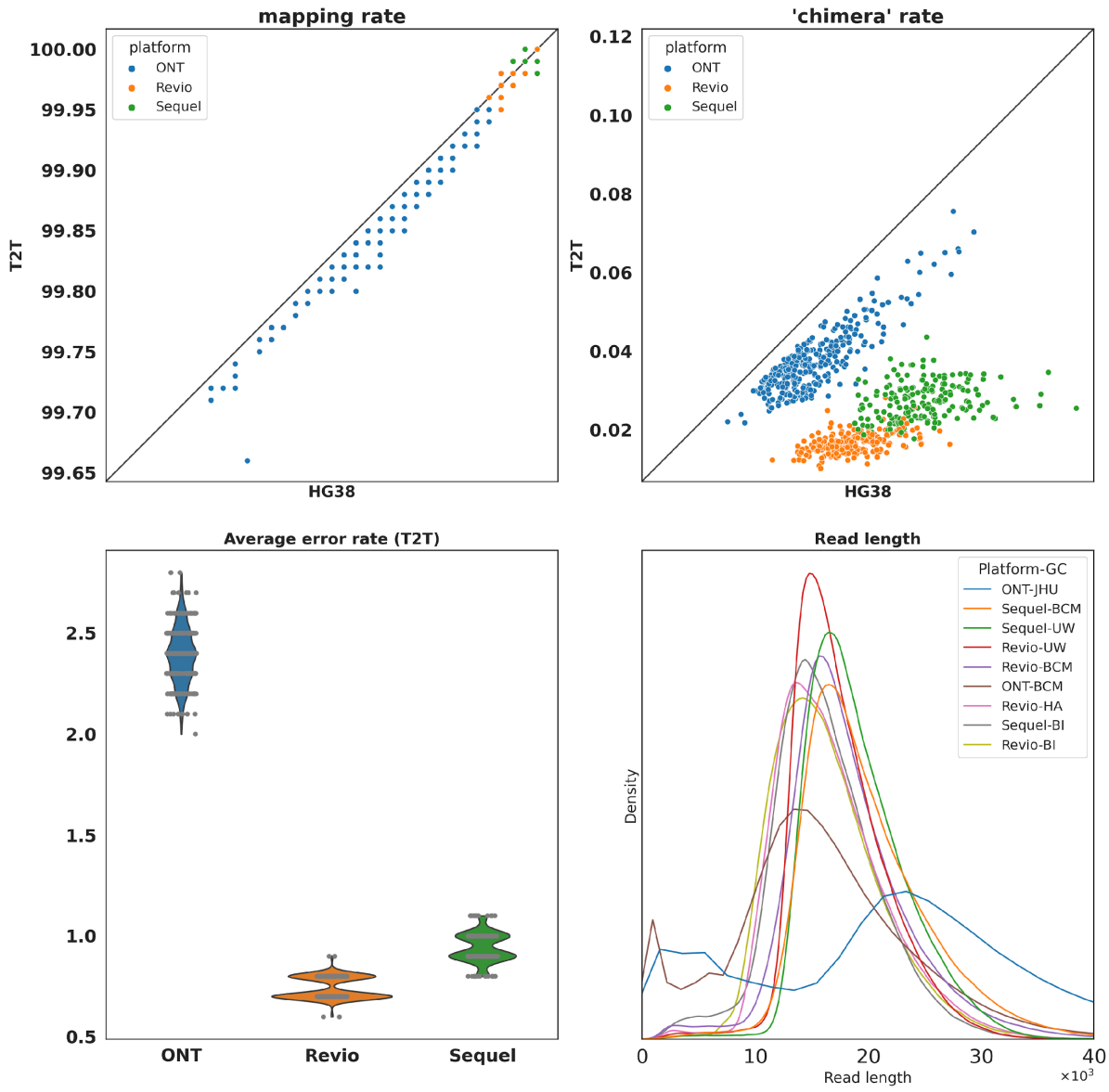


Figure R.1 – Mapping rate, 'chimera' rate, error rate, and read length values compared between sequencing facilities and sequencing platform. Read error rate is estimated from the long-read sequences aligned to the T2Tv2.0 reference.

Appendix S: Long-read QUAL score cutoff determination

We used the QUAL annotation of variants to filter out low quality variants during the Variant Hard Filter SNP and Indel QC. However, due to the various factors such as sequencing platform and coverage, it is not clear what threshold is appropriate and if different cohorts need different thresholds. Hence different thresholds are attempted, and the filtered callset is evaluated based on four metrics for each cohort: SNP heterozygous to homozygous variant ratio (Het/Hom ratio), transition to transversion ratio (Ti/Tv ratio), short insertion to deletion counts ratio (Ins/Del ratio), and variant count. These metrics under different QUAL filter thresholds are displayed in Figures S.1-S.10, for the various cohorts.

After analysis, we decided to use a QUAL score cutoff of 40 for PacBio HiFi data, and 34 for ONT data.

Note that the exact number of variants in each cohort is not directly comparable due to cohort size differences, coverage differences, and sequencing platforms used.

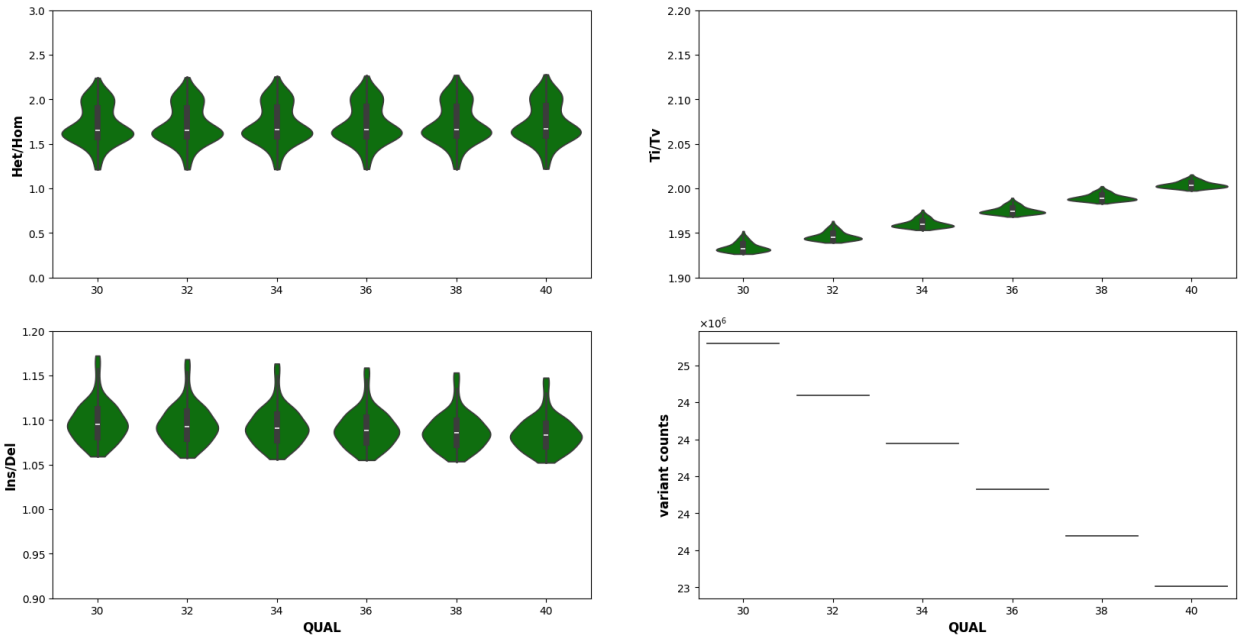


Figure S.1 -- BCM_Rev_mid cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds. The Het/Hom and Ins/Del ratios change only slightly when applying different QUAL thresholds. Ti/Tv ratio increases to the generally accepted range of 2.0 and 2.2 with 40 as the QUAL threshold. The total number of variants decreases, as expected, while the threshold increases, but not significantly (8%).

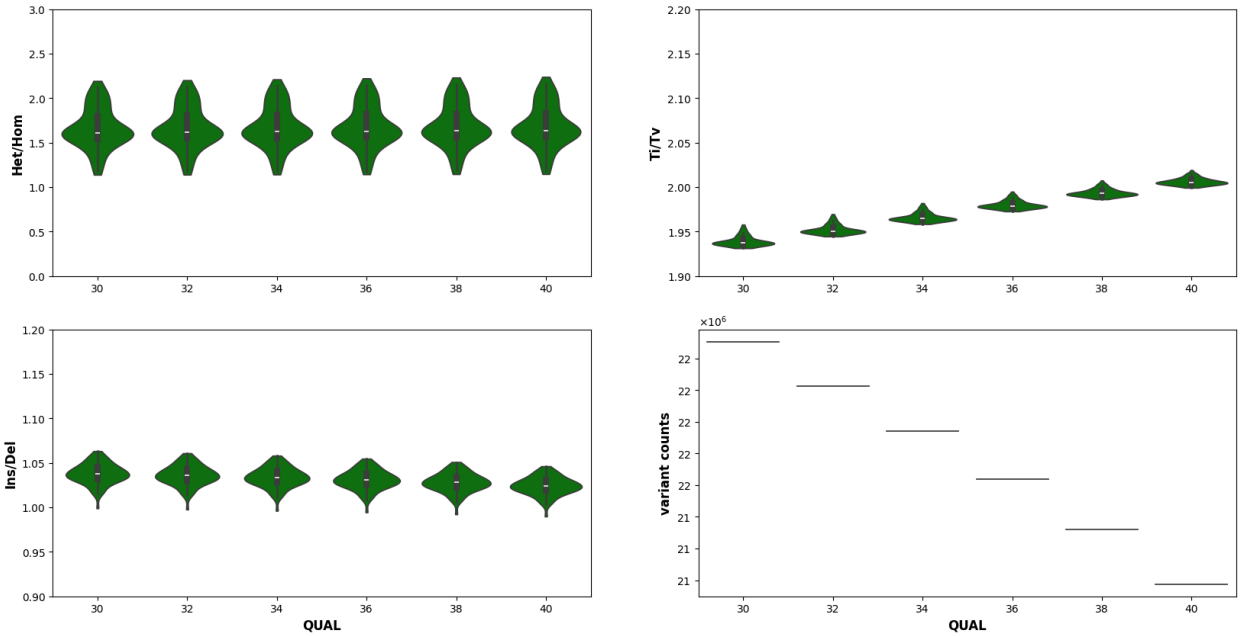


Figure S.2 -- BCM_Seq_mid cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds. We see similar patterns as Figure S.1.

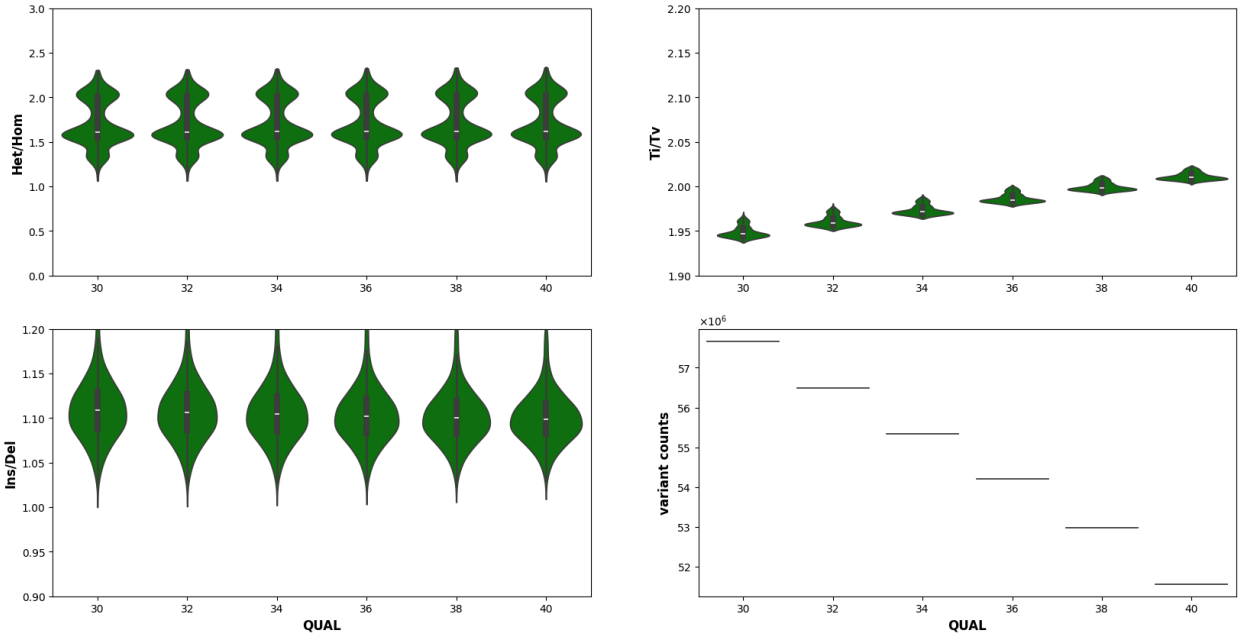


Figure S.3 -- BI_Rev_mid cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds. Note that the number of variants here are significantly higher than other cohorts due to its cohort size.

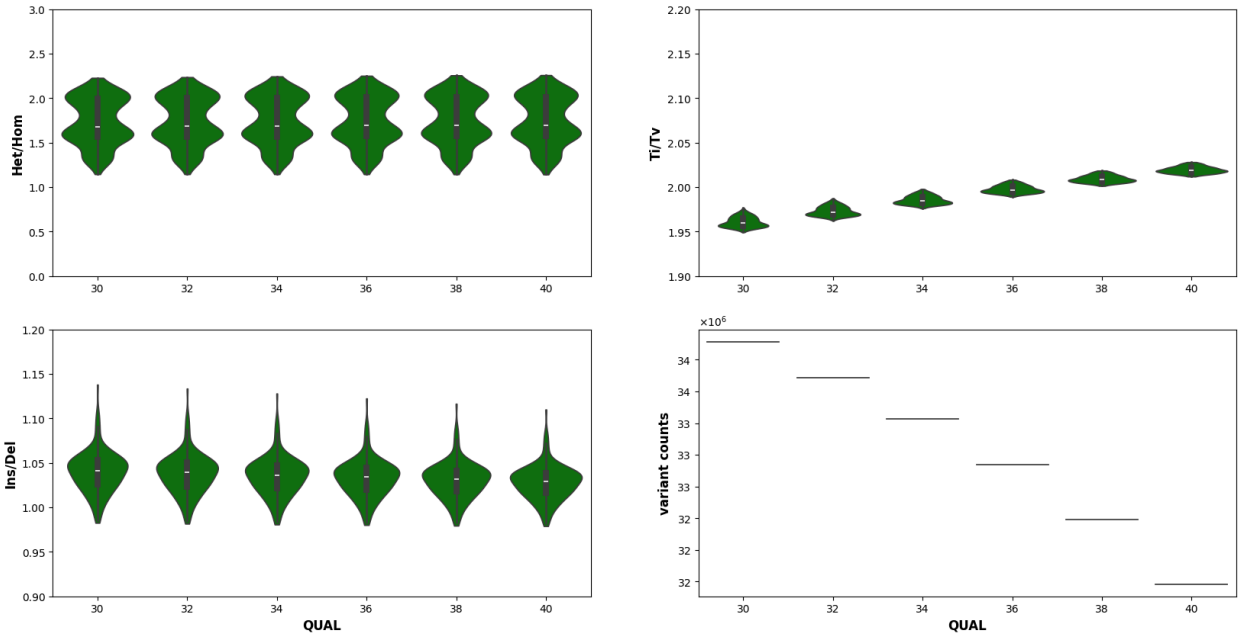


Figure S.4 -- BI_Seq_mid cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds.

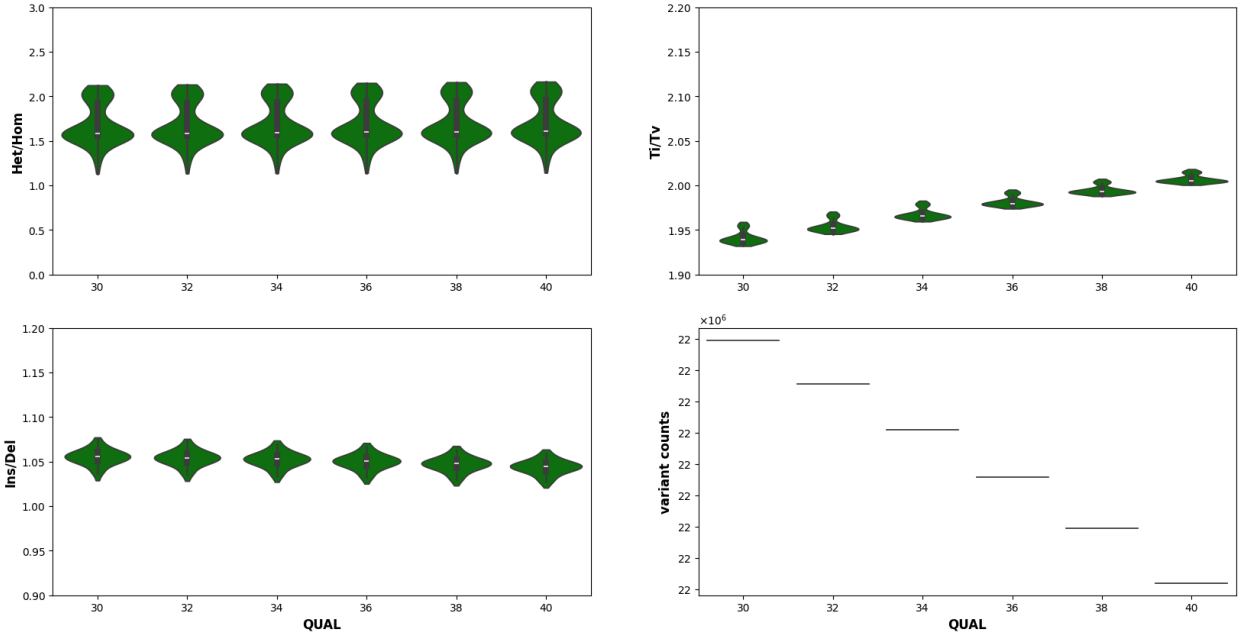


Figure S.5 -- BI_Seq_high cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds.

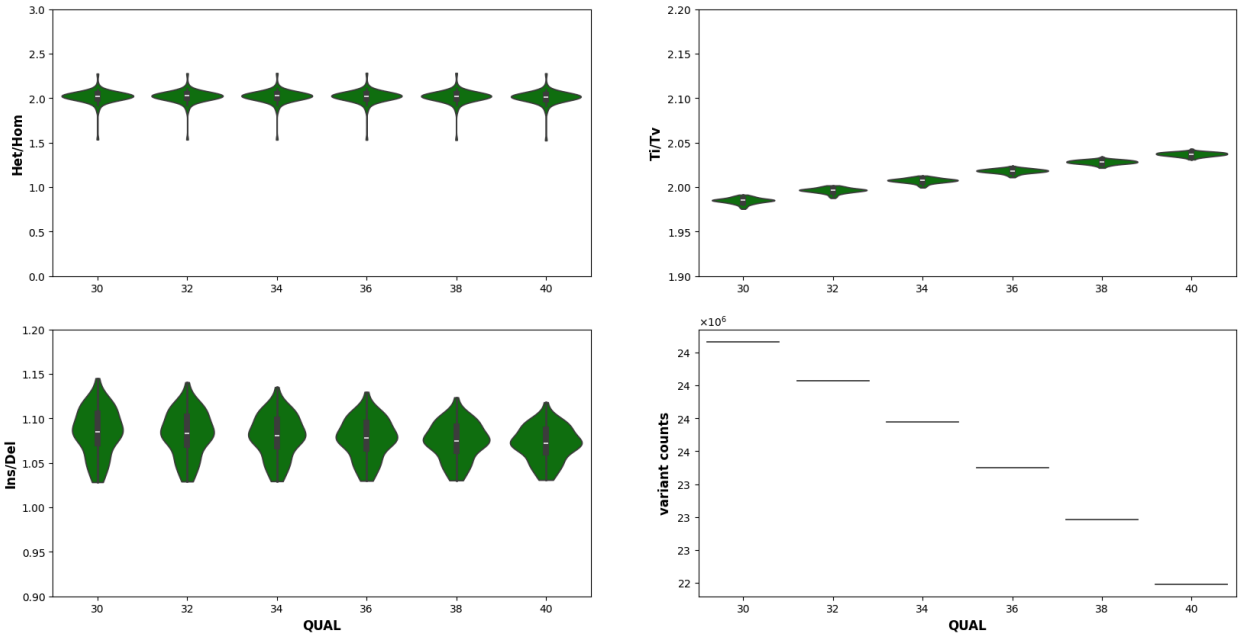


Figure S.6 -- HA_Rev_mid cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds.

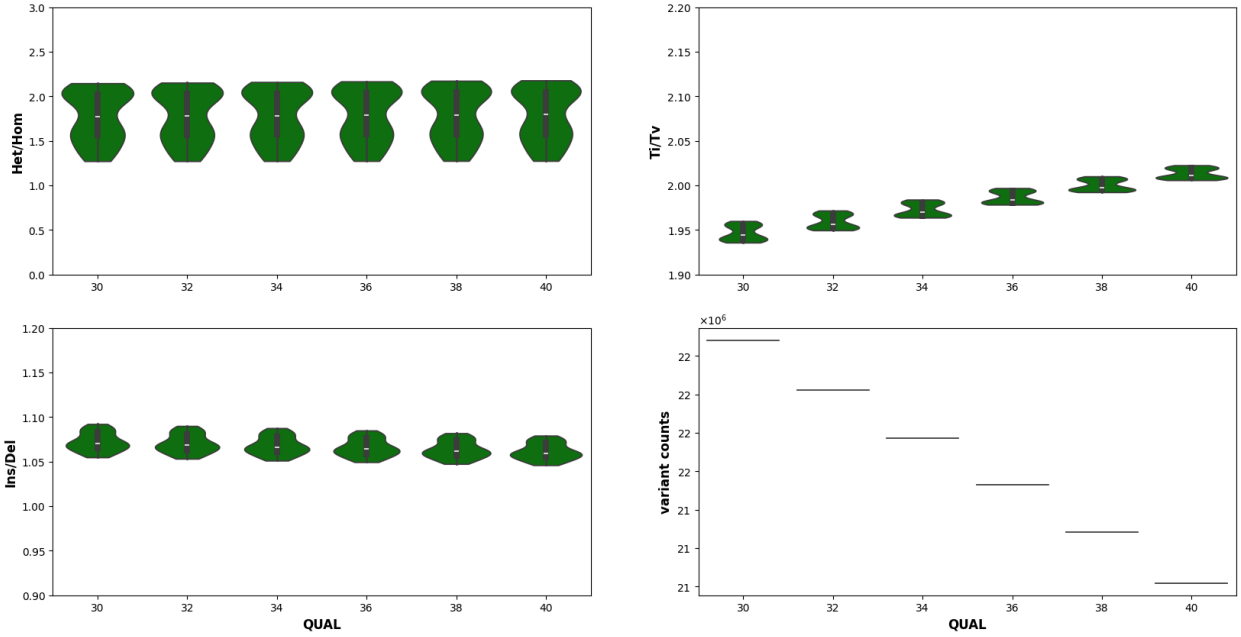


Figure S.7 -- UW_Rev_25 cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds.

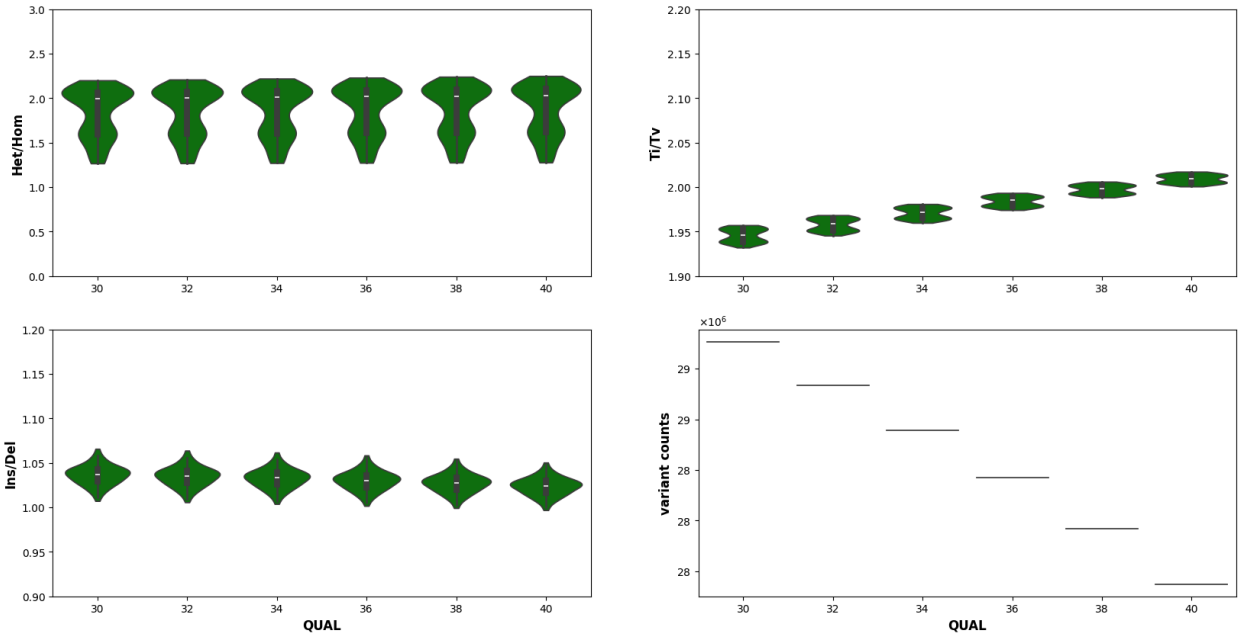


Figure S.8 -- UW_Seq_25 cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds.

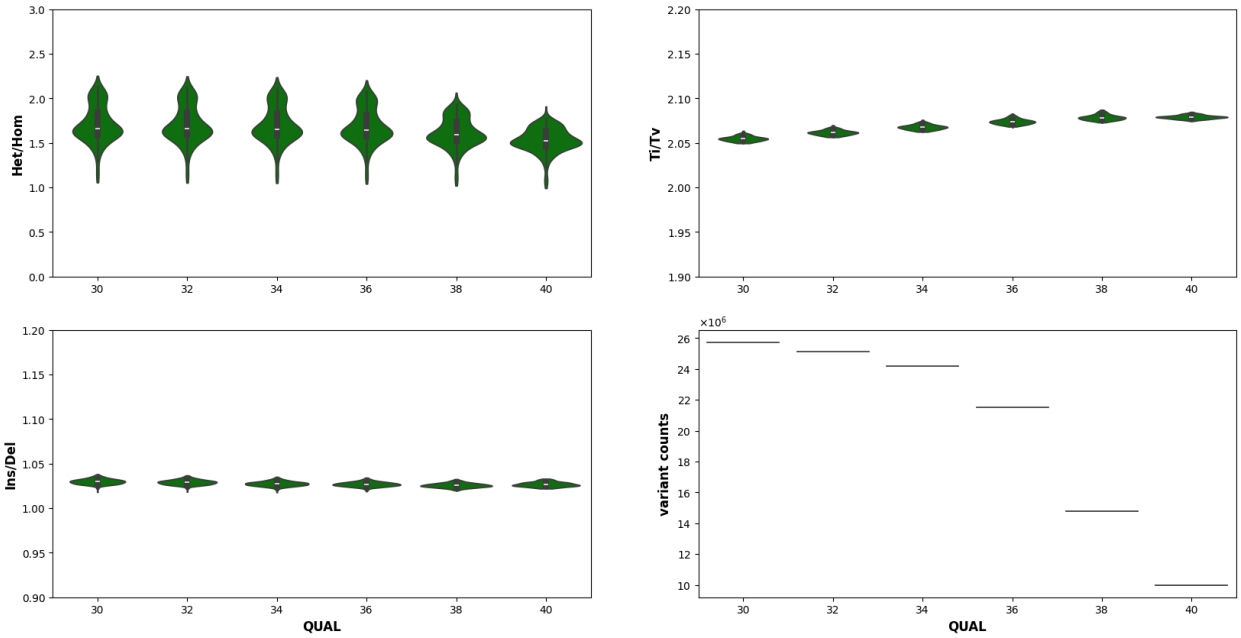


Figure S.9 -- BCM_ONT_high cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds. Note that the number of variants dropped significantly after the QUAL 34 threshold.

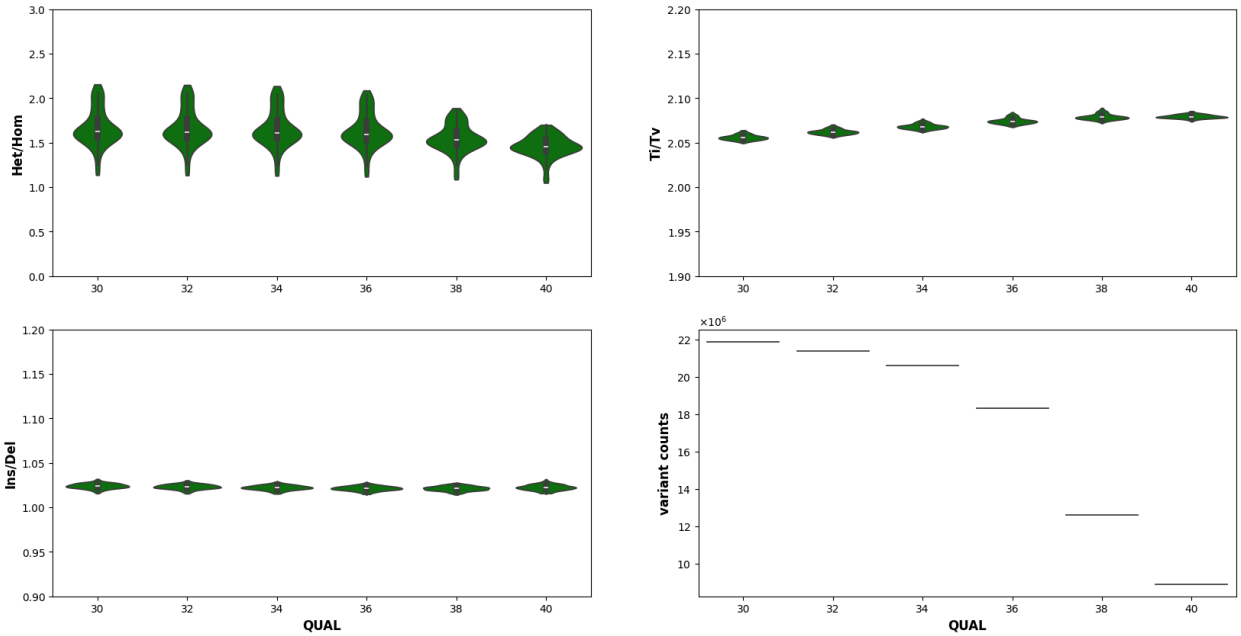


Figure S.10 -- JHU_ONT_high cohort Het/Hom, Ti/Tv, Ins/Del, and variant counts compared to different QUAL thresholds. Note that the number of variants dropped significantly after the QUAL 34 threshold.

Appendix T: Long-read SV results

As described in Structural Variant QC, we performed variant calling with Sniffles2, PBSV, and PAV, aligned to the grch38_noalt and T2Tv2.0 reference. The following plots (on both references) show the number of variants (at or above 50bp) called for each sample versus the sample coverage.

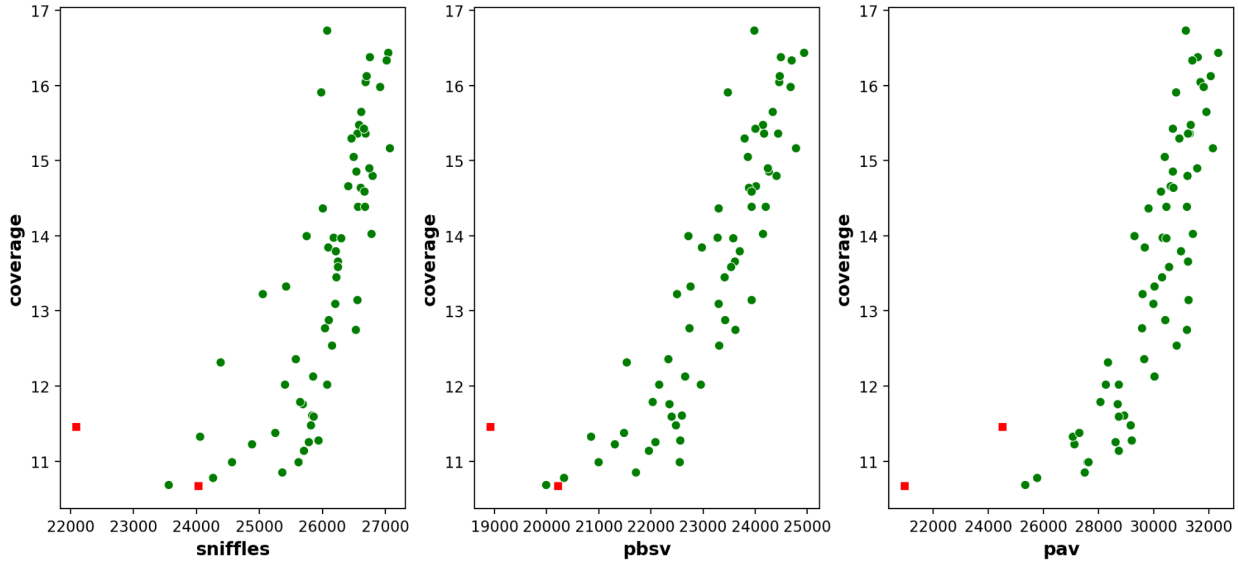


Figure T.1 -- SV counts for each sample in the HA_Rev_mid cohort (for the grch38_noalt reference), compared for each SV caller, Sniffles2, PBSV, and PAV. We manually identified three outliers, highlighted in red.

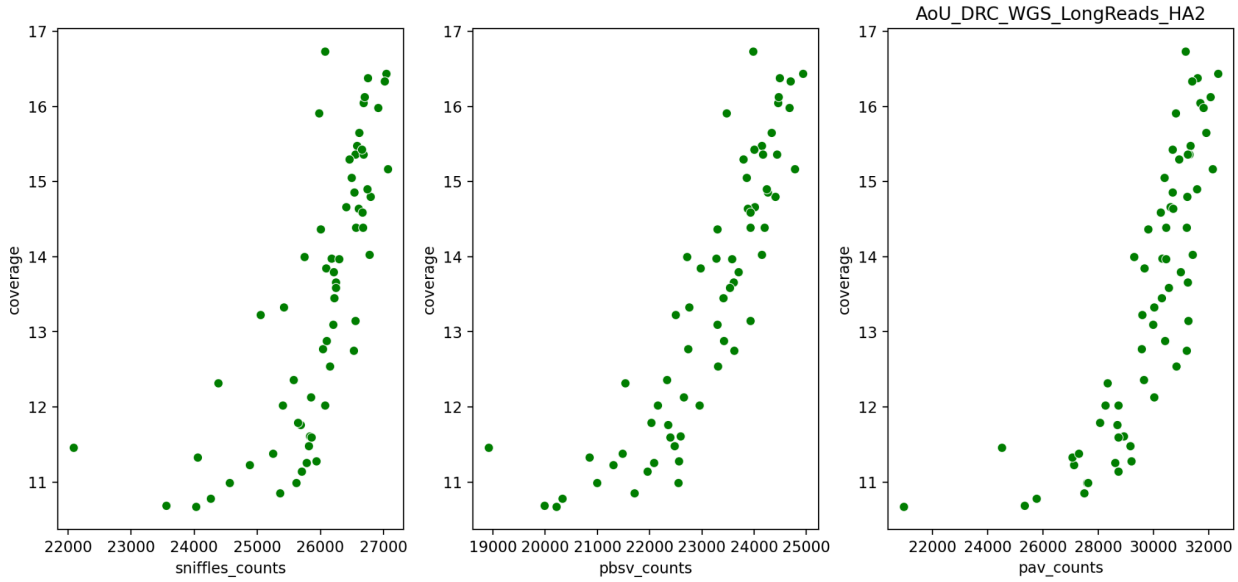


Figure T.2 -- SV counts for each sample in the HudsonAlpha HA_Rev_mid cohort (for the T2Tv2.0 reference), compared for each SV caller, Sniffles2, PBSV, and PAV. We manually identified three outliers, highlighted in red.

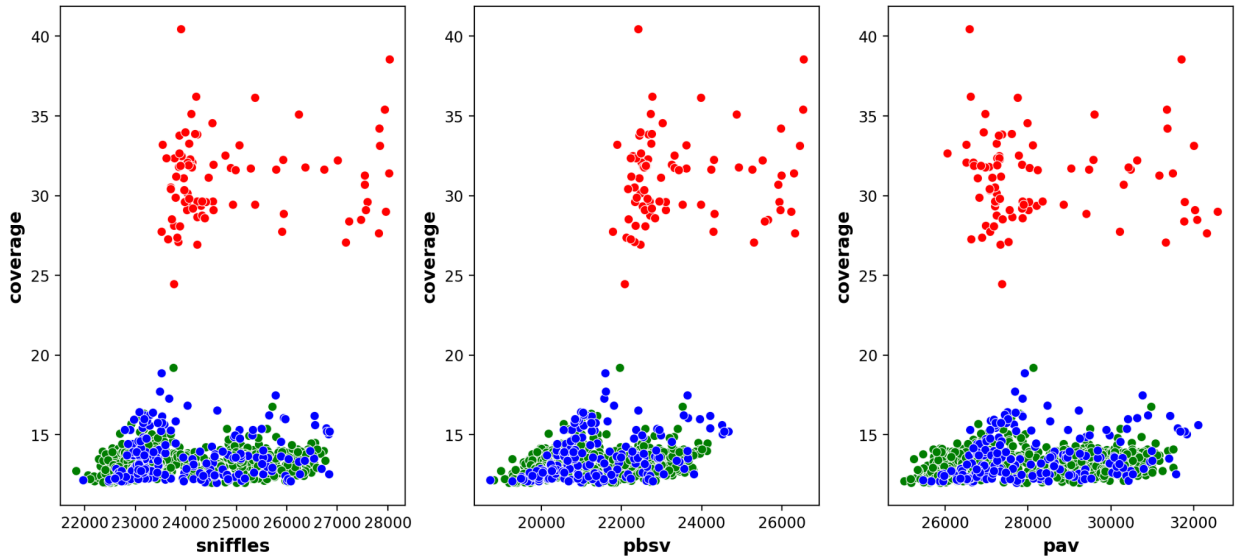


Figure T.3 -- SV counts (for the grch38_noalt reference) for each sample in all three BI cohorts: BI_Seq_high (red), BI_Seq_mid (blue), BI_Rev_mid (green), compared for each SV caller, Sniffles2, PBSV, and PAV. No outliers are detected.

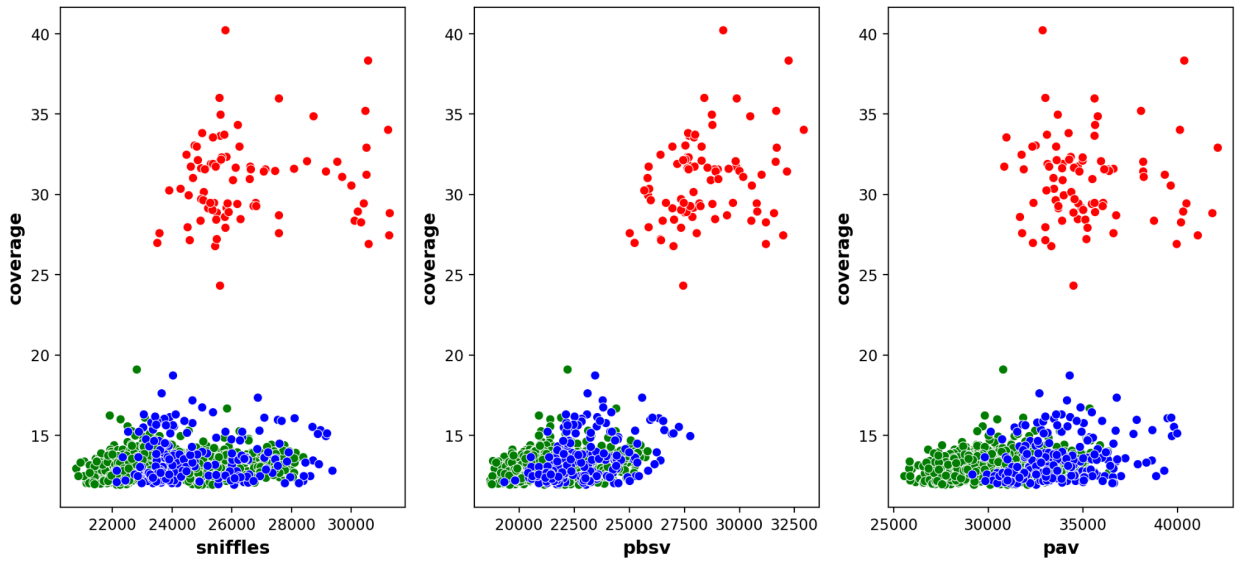


Figure T.4 -- SV counts (for the T2Tv2.0 reference) for each sample in all three BI cohorts: BI_Seq_high (red), BI_Seq_mid (blue), BI_Rev_mid (green), compared for each SV caller, Sniffles2, PBSV, and PAV. No outliers are detected.

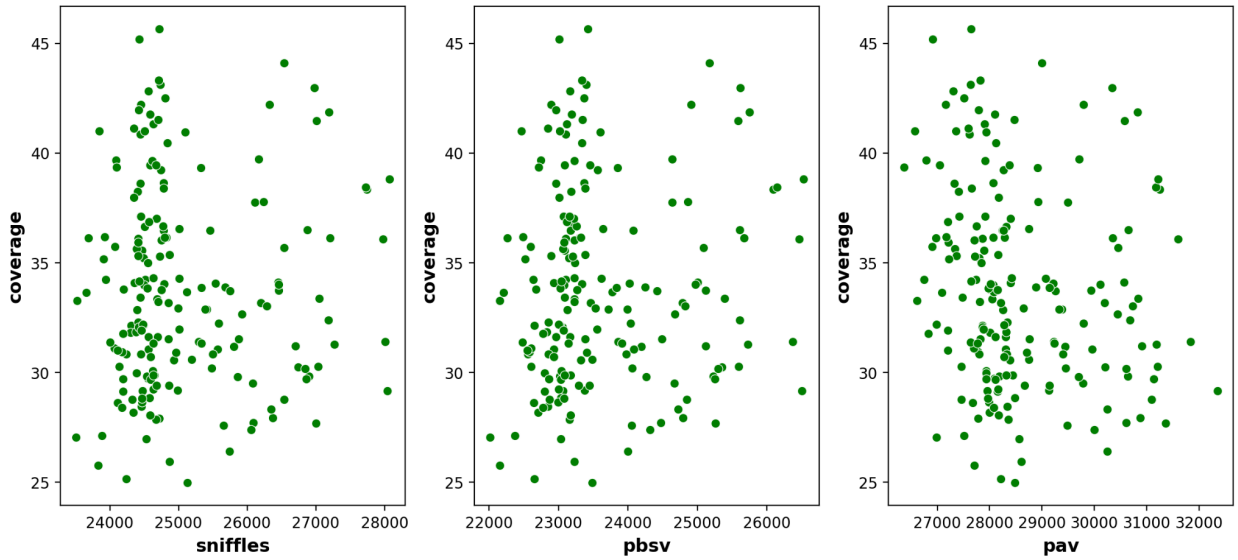


Figure T.5 -- SV counts for each sample (for the grch38_noalt reference) in both BCM PacBio cohorts: BCM_Seq_high and BCM_Rev_high, compared for each SV caller, Sniffles2, PBSV, and PAV. We found no outliers in these cohorts.

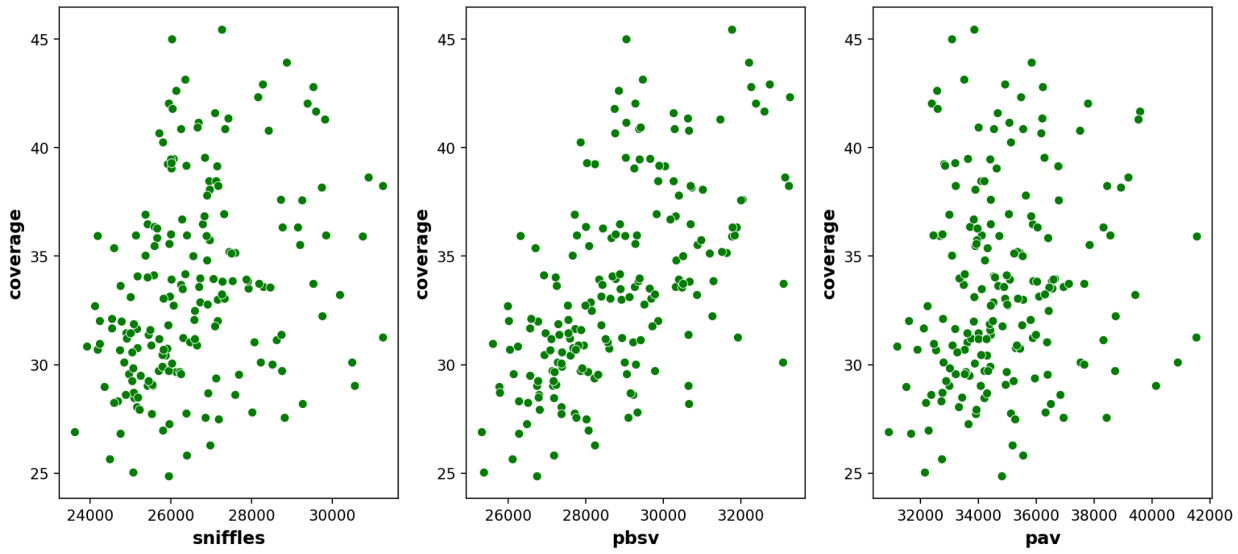


Figure T.6 -- SV counts (for the T2Tv2.0 reference) for each sample in both BCM PacBio cohorts: BCM_Seq_high and BCM_Rev_high, compared for each SV caller, Sniffles2, PBSV, and PAV. We found no outliers in these cohorts.

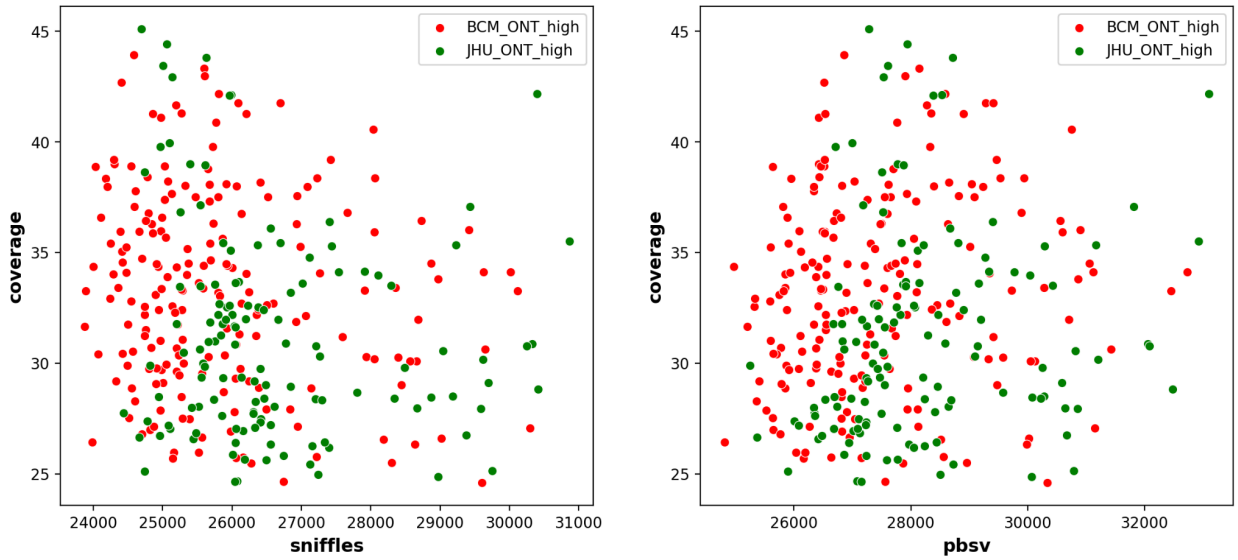


Figure T.7 -- SV counts (for the grch38_noalt reference) for each sample in the BCM_ONT_high and JHU_ONT_high cohorts, compared for each SV caller, Sniffles2 and PBSV. We found no outliers in these cohorts.

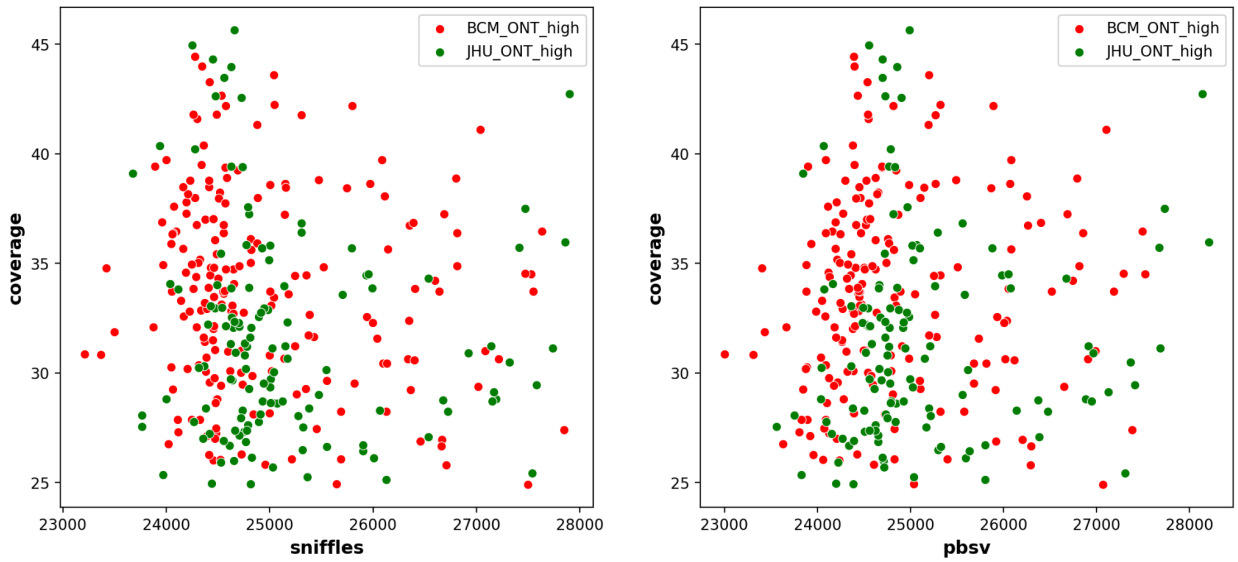


Figure T.8 -- SV counts (for the T2Tv2.0 reference) for each sample in the BCM_ONT_high and JHU-ONT_high cohorts, compared for each SV caller, Sniffles2 and PBSV. We found no outliers in these cohorts.

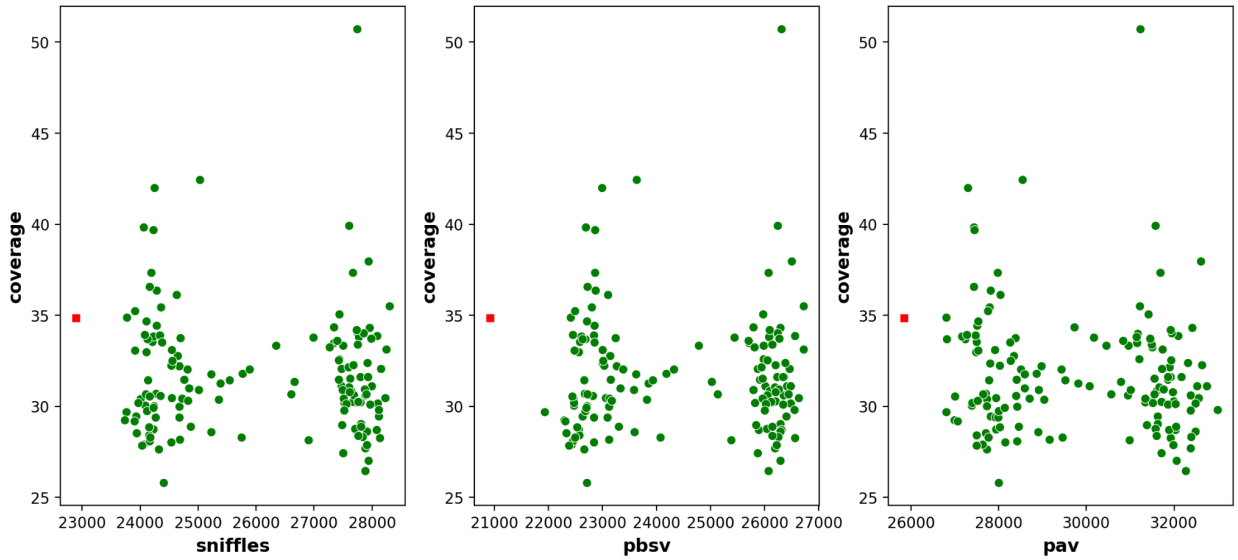


Figure T.9 -- SV counts (for the grch38_noalt reference) for each sample in the UW cohorts: UW_Seq_high and UW_Rev_high, compared for each SV caller, Sniffles2, PBSV, and PAV. One outlier sample was identified with low SV counts in all three callers (red rectangle). We also observed two clusters across some of the callers which could be attributed to differences in genetic ancestry between the samples.

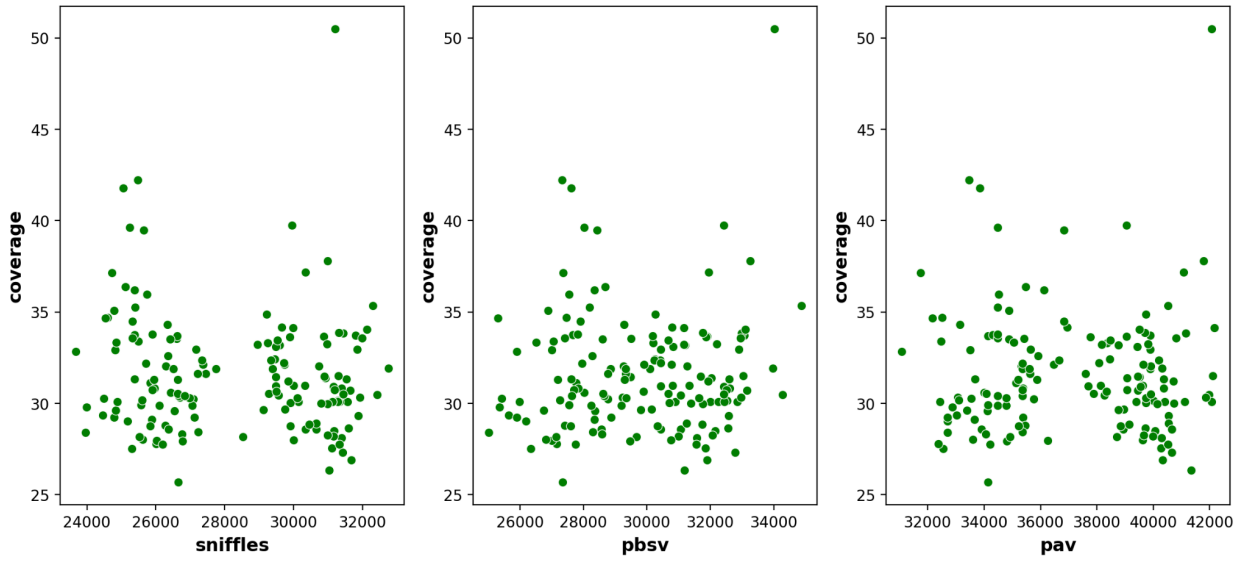


Figure T.10 -- SV counts (for the T2Tv2.0 reference) for each sample in the UW cohorts: UW_Seq_high and UW_Rev_high, compared for each SV caller, Sniffles2, PBSV, and PAV. We also observed two clusters across some of the callers which could be attributed to differences in genetic ancestry between the samples.