All of Us Research Program

Genomic & Multi-omic Research Data Quality Report

All of Us Curated Data Repository (CDR) release C2025Q4R6

Overview

This document details the *All of Us* Genome Centers (GC) and Data and Research Center (DRC) quality control (QC) steps for the genomic and multi-omic data made available in the new Researcher Workbench platform June 26, 2026 in the CDRv9 data release. This pipeline removes or flags samples and variants that fail quality thresholds. We apply these QC steps in the research pipeline before we release the genomic and multi-omic data for researchers. This document only describes QC processes that are performed analytically (i.e., after the sample has been sequenced).

The participants' samples in the genomic and multi-omic data correspond to the *All of Us* Curated Data Repository (CDR) release C2025Q4R6 ("CDRv9"). All descriptions and results are limited to the CDRv9 data, which contains 553,949 genotyping array ("array") data from 553,949 participants, short-read whole genome sequencing (srWGS) data from 535,662 samples with single nucleotide polymorphism, insertion, and deletion variant calls (SNPs and Indels), srWGS data from 96,405 participants' samples with structural variant (SV) calls, RNA Seq data from 8,980 participants' samples, 9,969 proteomics samples, and 14,521 long-read whole genome sequencing (lrWGS) samples with SNP, Indel, and SV calls. Many of these data types overlap with each other, allowing co-analysis (25 exceptions exist, see Known Issue #3).

Audience: This document is intended for researchers using, or considering the use of, the genomic and multi-omic data in the new Researcher Workbench (RW). This document assumes knowledge of sequencing, genotype arrays, common multi-omic 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 algorithms are out of scope for this document. A selection of processing pipelines used for data analysis are available in a public GitHub repository along with accompanying documentation.
- The locations of raw data are in the Data Dictionary and descriptions of the file formats for the multi-omic data are available in the 'How the All of Us Genomic data are organized', both published on the User Support Hub [1].
- The multi-omic data mentioned in this document requires Controlled Tier access to view. To register for access, please go to https://www.researchallofus.org/register/

Executive Summary

On June 26, 2026, the *All of Us* Research Program released the genomic and multi-omic data of 553,949 array samples, 535,662 srWGS samples with SNP & Indel calls, 96,405 srWGS samples with SV calls, 8,980 RNA Seq samples, 9,969 proteomics samples and 14,521 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* CDRv9 Data

Dataset	Number of participants	Highlights
Array	553,949	- More than 1.9 million variants - We added more than 100,000 new participants in CDRv9
Short-read WGS SNP and Indel	535,662	- More than 1.3 billion variants - We added more than 120,000 new participants with srWGS data in CDRv9 - There are more than 125 million new variants as compared to the previous *All of Us* dataset - The *All of Us* srWGS dataset is now one of the largest srWGS datasets
Short-read WGS structural variants (SVs)	96,405	- Nearly 1.5 million variants
Long-read WGS	14,521	- We added more than 11,000 new participants with long-read WGS data in CDRv9 (more than 2.5 million variants)
RNA Seq	8,980	- NEW data type in the CDRv9 release - Gene counts and per-ancestry quantitative trait loci (QTL) analyses
Proteomic samples	9,969	- NEW data type in the CDRv9 release - Expression counts for more than 5,000 proteins

In addition to variant calls, raw data (IDAT files for array data, CRAM files for srWGS data, BAM files for lrWGS data and RNA data), and auxiliary files (including variant annotations, pharmacogenomics, genetic ancestry categories, relatedness/kinship scores, HLA variant calls, and challenging medically relevant gene calls) 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 and multi-omic 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 and multi-omic data - in sync with planned CDR release timelines. This document describes the CDRv9 release of genomic and multi-omic data to *All of Us* researchers made available in the RW on June 26, 2026. The genomic and multi-omic data contains 553,949 array samples, 535,662 srWGS samples, 96,405 srWGS samples with SV calls, 14,521 lrWGS samples, 8,980 RNA sequencing, and 9,969 proteomics samples which can be joined with other data types (e.g. survey data) for analysis, though please see Known Issue #3. In this document, we describe the QC processes applied to the array, srWGS, lrWGS, RNA sequencing, and proteomics 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 553,949 array samples in the CDRv9 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	v9 release results
Sex concordance	Sex call is concordant with self-reported sex at birth. OR	-Sample swaps	All array samples are concordant or passed investigation (see FAQ #10 below).

	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 and investigated the source of the swap. If, after thorough investigation, we determined no evidence that a swap occurred, then we included the sample in the release. Please see FAQ #10 for more details: How did we investigate samples that failed the sex concordance check? Were any samples that failed the sex concordance check added back?

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-threshold	-0.1	This effectively disables the sex estimation.

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

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

Results

Since we catch sex concordance failures before including a sample in the release, all array samples in the CDRv9 release passed a sex concordance check or passed after investigation. We added back 933 array samples (0.17%) after initial failure of the sex concordance check and investigation (See FAQ #10).

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 CDRv9 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. You will see two separate peaks in the data because the Y chromosome is only present in XY individuals. Since XX individuals lack a Y chromosome, their call rates for those specific areas will be near zero.

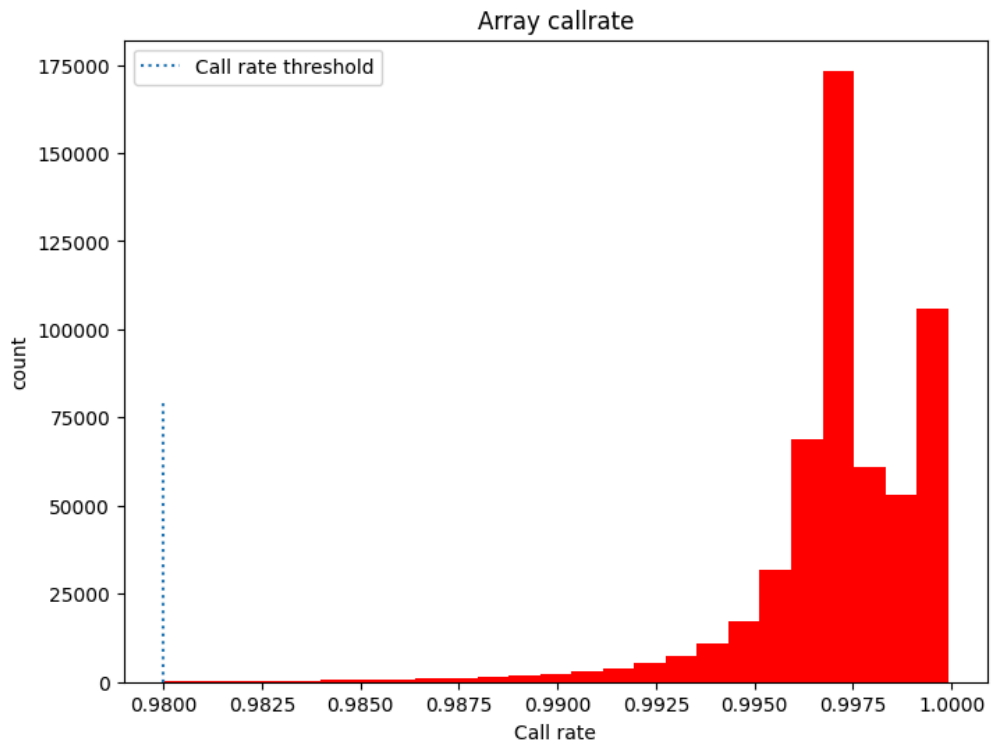


Figure 1 -- Histogram of the array call rate for the v9 release.

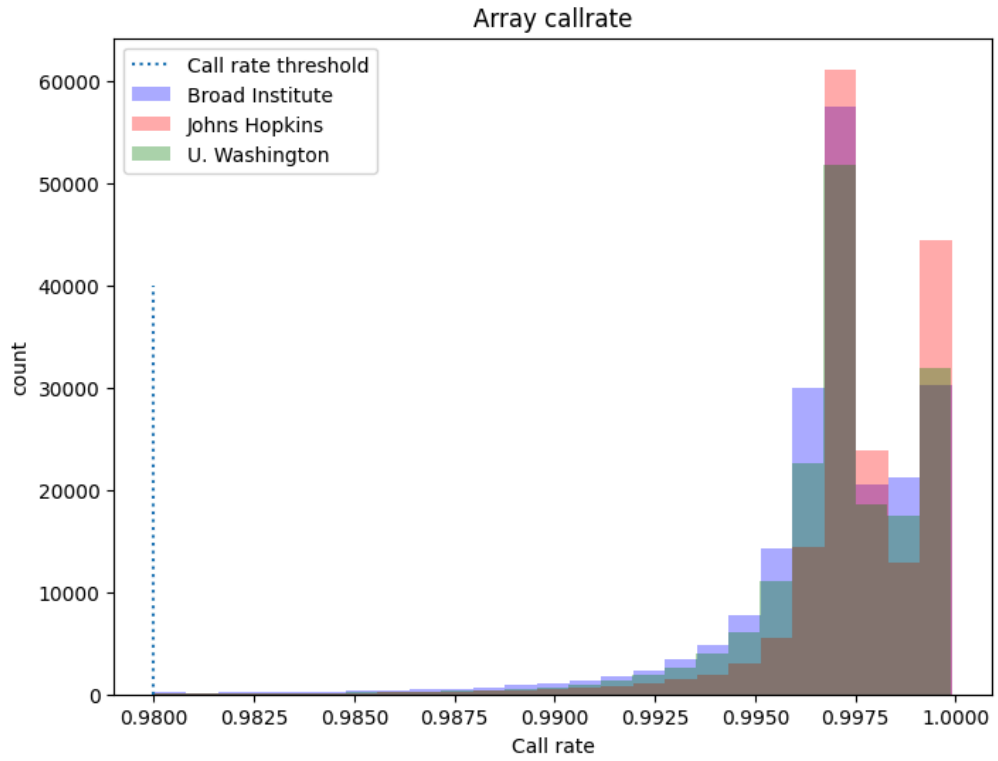


Figure 2 -- Call rate across each GC.

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 553,949 array samples, 550,293 (99.3%) had an estimated contamination rate below 3% and 541,997 (97.8%) had a contamination rate less than 1%.

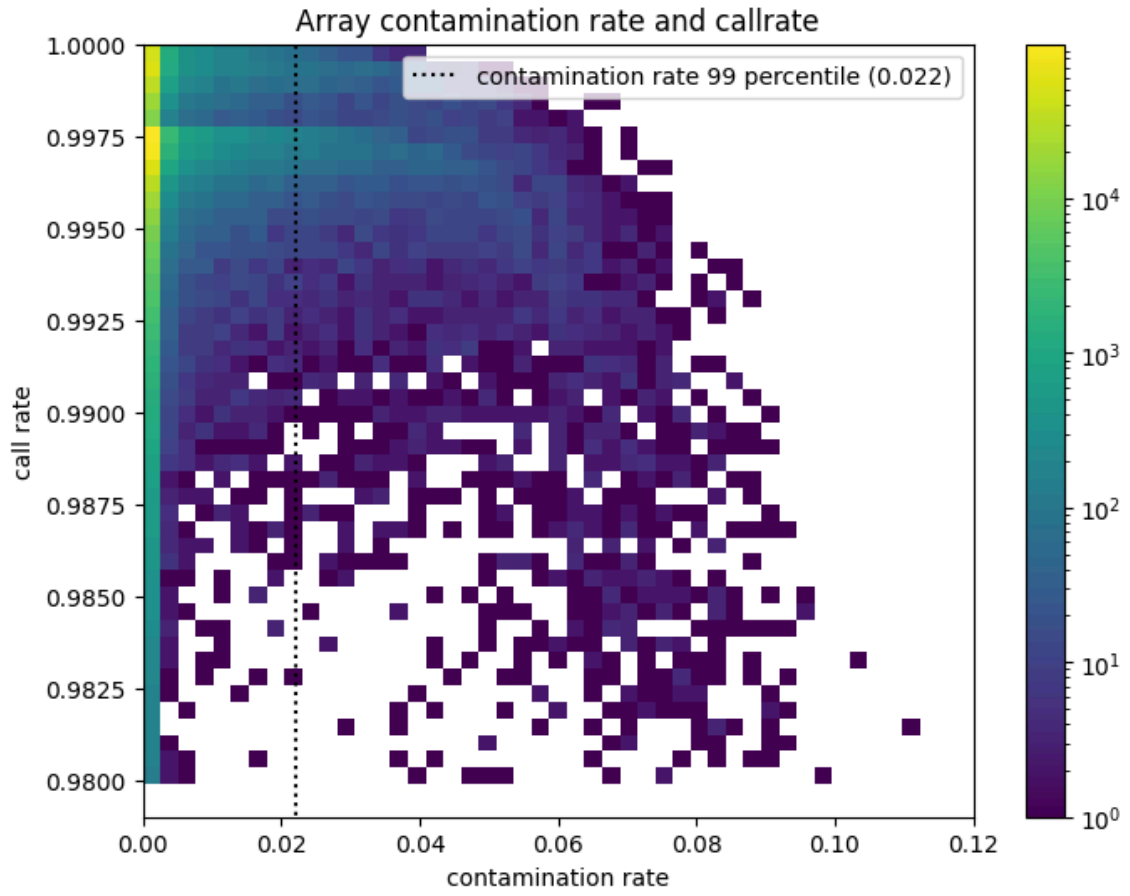


Figure 3 -- Histogram of the array contamination rate estimates vs call rate. As the contamination rate increases, the call rate decreases.

Short-Read Whole Genome Sequencing (srWGS)

The *All of Us* srWGS dataset is a high-quality comprehensive dataset of 535,662 participants [2], available as raw reads, variant data, and annotated variants. We also provide substantial auxiliary data to accompany the srWGS dataset. 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), software (DRAGEN v3.7.8), and software configuration. As of CDRv9, the GCs used multiple sequencers (NovaSeq 6000 N=507,913 (94.8%) and NovaSeqX N=27,749 (5.2%)). We have performed a batch effect analysis between sequencer versions. Please see Appendix U for the characterization of batch effects. For more information about the sequencing processes used by the GCs, on the NovaSeq6000, see previous work [2] [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. Most thresholds in our single sample QC process are identical to the clinical pipeline described previously [6], except for a higher threshold for contamination and resolution of sex discordance by investigation (see FAQ #10)

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 we are focused, in this document, 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	CDRv9 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 or passed investigation (see FAQ #10 below). . *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	152 CDRv9 samples did not meet the coverage threshold, due to reprocessing of samples, and were included in the callset in order to retain samples with matching multiomics data (RNA sequencing, proteomics, and/or lrWGS). All other samples that did not meet the coverage threshold were removed. (Please see Known Issue #1).

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), on NovaSeq 6000 samples only, 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 CDRv9 release passed the fingerprint concordance check based on arrays. As seen in Figure 4, the passing samples exceeded the threshold. 12,332 samples had a fingerprint LOD [7] less than 45 and the minimum fingerprint LOD was 36.

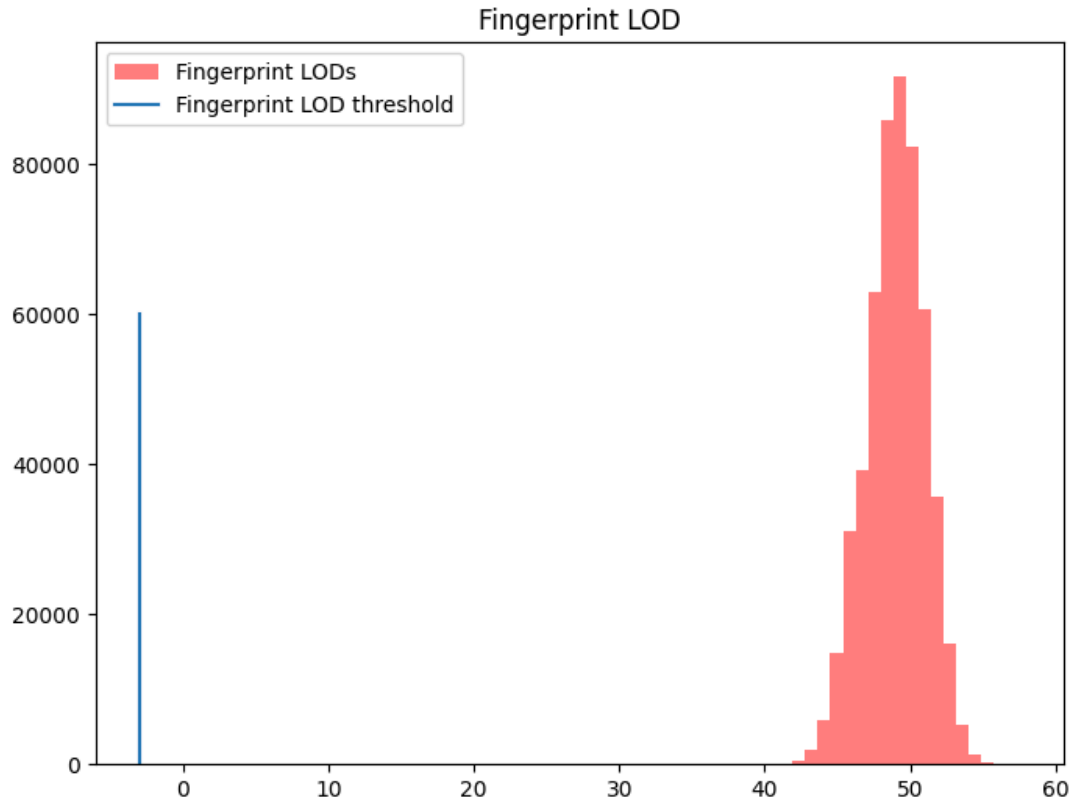


Figure 4 -- Distribution of the Fingerprint LODs for srWGS CDRv9 samples

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 and investigated the source of the swap. If we determined no swap, we included the sample in the release. Please see FAQ #10 for more details: How did we investigate samples that failed the sex concordance check? Were any samples that failed the sex concordance check added back?

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

Since we catch sex concordance failures before including a sample in the release, all srWGS samples in the CDRv9 release passed a sex concordance check or passed after investigation. We added back 522 srWGS samples (0.10%) after initial failure of the sex concordance check and investigation (See FAQ #10).

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 0.03 (according to DRAGEN) and 2,948 samples greater than 0.015. Figure 5 demonstrates the frequency of the contamination estimates for samples in the CDRv9 release.

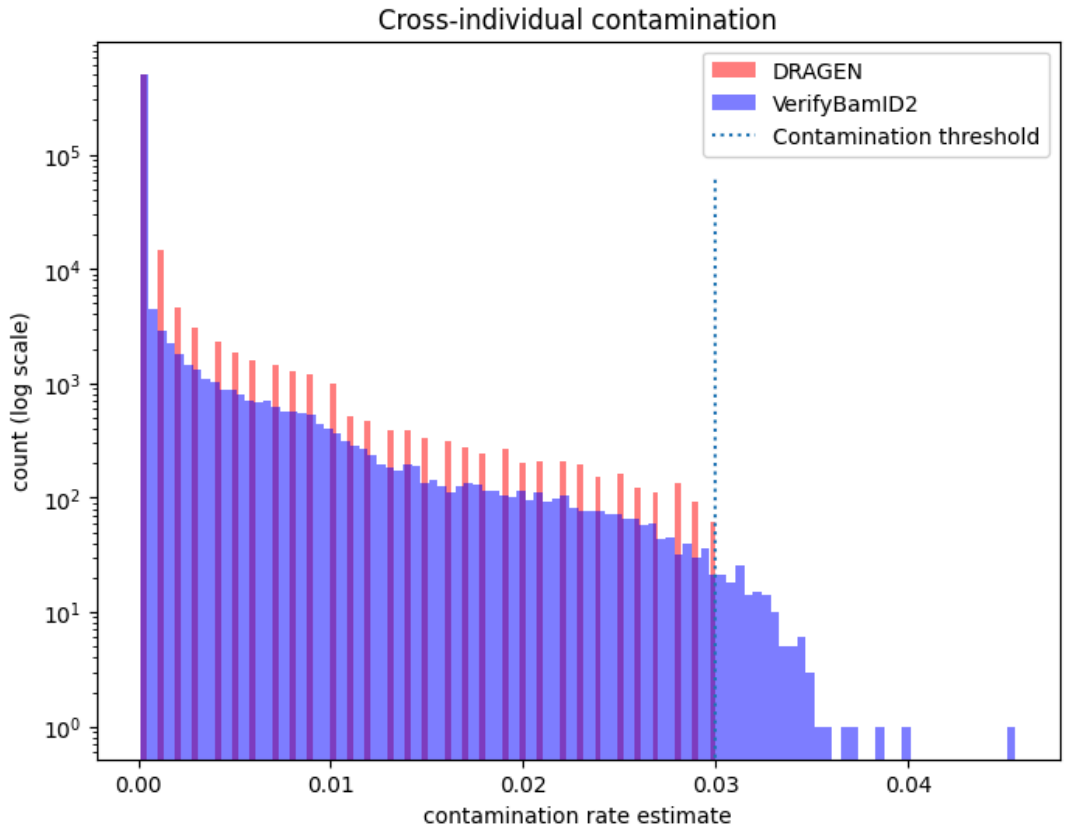


Figure 5 -- srWGS contamination estimates from both sources (DRAGEN and VerifyBamID2). DRAGEN rounds the contamination estimate to three decimal places. Note the log scale of the counts (y-axis). Over 89.7% and 93.0% of srWGS samples had contamination estimates lower than 1e-4 by VerifyBamID2 and DRAGEN, 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

Samples that fell below the mean coverage threshold were excluded from the callset. Please see Known Issue #1, as 147 CDRv9 samples did not meet the mean coverage threshold (min: 28.5x) and were included in the callset in order to include samples with matching multiomics data (RNA sequencing, proteomics, and/or lrWGS). We had 634 (0.1%) samples with mean coverage greater than 70x.

Plots of the srWGS metric histograms are seen in Figure 6.

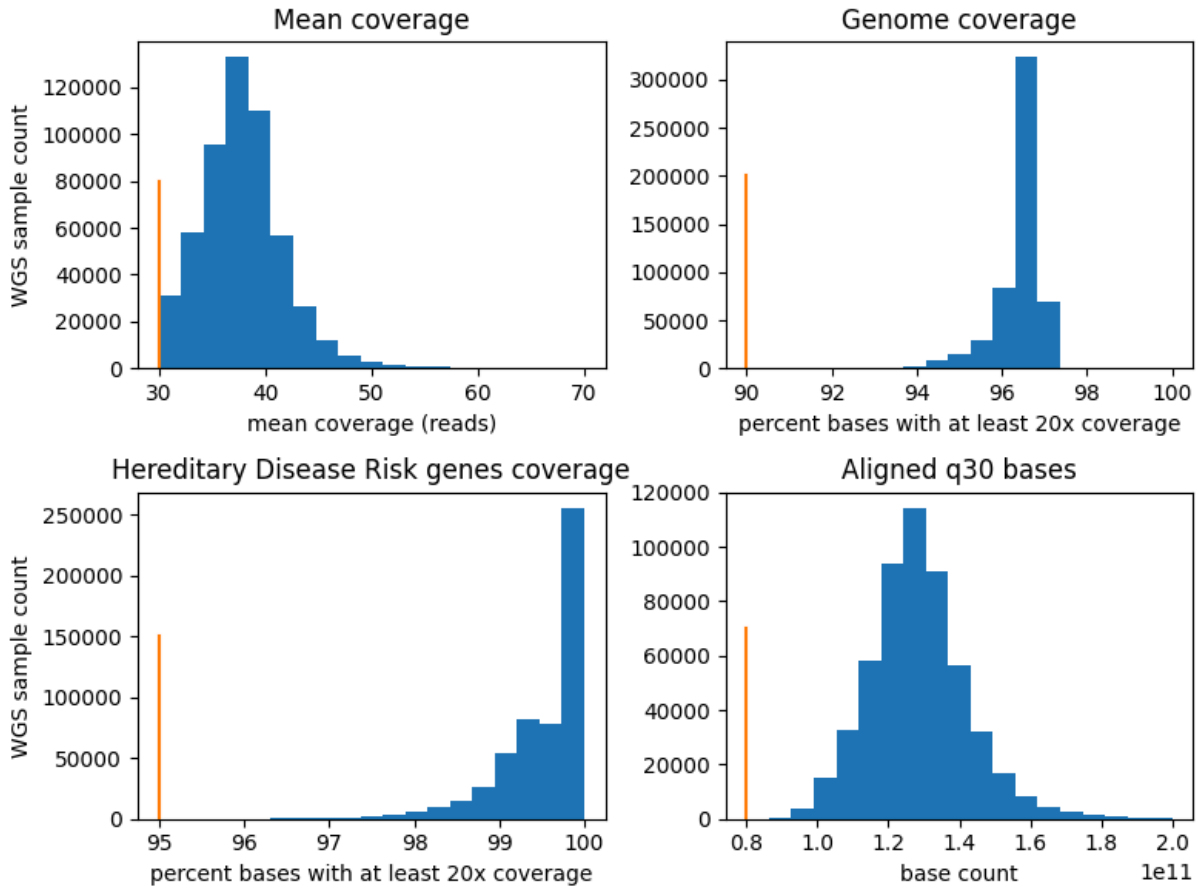


Figure 6 -- Coverage metrics for the CDRv9 release srWGS samples. The orange line is the threshold for each metric. There are 634 samples (0.1%), with mean coverage greater than 70x, that are not included in the mean coverage (upper left) nor aligned q30 bases (lower right) plots. As expected, these samples were outliers in the number of aligned q30 bases (i.e., higher base

count than samples with lower mean coverage). Please note that 152 samples did not meet minimum mean coverage and are not shown in this figure (see Known Issue #1)

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.

Please note one known issue affecting the srWGS SNP & Indel callset, Known Issue #2.

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

QC process	Sample or variant QC	Error modes addressed	CDRv9 release results
Sample Hard Threshold Flag	sample	Extremely noisy samples	All samples that were flagged were flagged as part of the Sample Population Outlier Flag (see below).
Sample Population Outlier Flag	sample	Noisy samples	1012 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. 71,024,812 were filtered 1,311,400,566 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): 0.02 Error rate from holdout set (not incl. Other): 0.006 See Appendix G. Number of independent, bi-allelic sites ("high-quality sites") used: 123,171 See Appendix I.
Relatedness and	sample	Related samples,	55,907 related pairs and 42,863 samples in the

maximal independent set of samples		which confound analyses	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 Data Dictionary.

Method

We initially flagged any samples with strong erroneous signals. We calculated all metrics using autosomal territory only. The criteria for being flagged 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 Data Dictionary.

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 3.1 count
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 1012 (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	466
Del count + Indel Het/Hom + Ins count + SNP count	153
Not in gnomAD 3.1 count	120
Indel Het/Hom + SNP count	87
Del count + Indel Het/Hom + SNP count	61
Indel Het/Hom + SNP Het/Hom	50
SNP Het/Hom	26
Ti/Tv ratio	14
Del count + Ins count + Not in gnomAD 3.1 count + SNP Het/Hom + SNP count	6
Del count + Ins count + Not in gnomAD 3.1 count + SNP count + Ti/Tv ratio	6
Not in gnomAD 3.1 count + Ti/Tv ratio	5
Not in gnomAD 3.1 count + SNP count	4
Del count + Ins count + Not in gnomAD 3.1 count + SNP count	4
Del count + Indel Het/Hom + Ins count + Ins/Del ratio + Not in gnomAD 3.1 count + SNP Het/Hom + SNP count + Ti/Tv ratio	3
Not in gnomAD 3.1 count + SNP count + Ti/Tv ratio	2
Del count + Ins count + Not in gnomAD 3.1 count + SNP Het/Hom + SNP count	2

+ Ti/Tv ratio	
Del count	1
Del count + Indel Het/Hom + Ins count + Not in gnomAD 3.1 count + SNP Het/Hom + SNP count + Ti/Tv ratio	1
Indel Het/Hom + SNP Het/Hom + SNP count	1

Total	1012

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'	108,905 (<0.01%)
'EXCESS_ALLELES', 'ExcessHet'	9,630 (<0.01%)
'EXCESS_ALLELES', 'ExcessHet', 'NO_HQ_GENOTYPES'	1 (<0.01%)
'EXCESS_ALLELES', 'NO_HQ_GENOTYPES'	1,270 (<0.01%)
'ExcessHet'	625,883 (0.05%)
'NO_HQ_GENOTYPES', 'ExcessHet'	397 (<0.01%)
'LowQual'	3,633,421 (0.26%)
'NO_HQ_GENOTYPES', 'LowQual'	26,042,540 (1.88%)
'NO_HQ_GENOTYPES'	40,602,765 (2.94%)
Total variants	1,382,425,378
Total variants filtered	71,024,812 (5.14%)
Total not filtered	**1,311,400,566**

Variant Extract-Train-Score Filtering (VETS)

We flag variants using the Variant Extract-Train-Score (VETS) method, which is an allele filtering algorithm. The VETS filters are implemented in the genotype filtering field. In other words, some sites may have only certain genotypes filtered, whereas other sites may have all genotypes filtered.

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 PASS or FAIL. In the VCF, a filtered genotype will be annotated with high_CALIBRATION_SENSITIVITY_SNP or high_CALIBRATION_SENSITIVITY_INDEL.

Variants that do not pass filtering do not appear in the Variant Annotation Table (VAT) or the Cohort Builder.

The variant score is available in the VDS, within the `as_vets` row field for each variant. The cutoff score is a global field in the VDS as `truth_sensitivity_snp_threshold` and `truth_sensistivity_indel_threshold`. Please see How the *All of Us* Genomic data are organized for more information.

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. In this analysis, we used two replicates for three of the samples. Each replicate was sequenced independently and at separate genome centers.

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	0.984	0.9993
	HG-001_B	0.9839	0.9992
	HG-002_A	0.9863	0.9996
	HG-002_B	0.9865	0.9997
	HG-003_A	0.9861	0.9991
	HG-003_B	0.9864	0.9994
	HG-004	0.9863	0.9995
	HG-005	0.986	0.9997
Indel	HG-001_A	0.9723	0.9969
	HG-001_B	0.9708	0.9961
	HG-002_A	0.9874	0.9987
	HG-002_B	0.9875	0.9988
	HG-003_A	0.9876	0.9977
	HG-003_B	0.9885	0.9983
	HG-004	0.9874	0.9981
	HG-005	0.9904	0.9987

srWGS Structural Variant (SV) Callset

The srWGS SV callset represents 96,405 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 documentation (see the Benchmarking and quality analyses on the All of Us short read structural variant calls).

The dataset is a refresh of the previous CDRv7 off-cycle release of srWGS SV data. For this current dataset release, the data was refreshed and we did not redo variant calling. We removed any samples that were dropped between releases and performed extra steps to refine the callset. Importantly, this means that the CDRv9 srWGS SVs were called from CRAMs aligned with DRAGEN version 3.4.12, which is different from the other data derived from srWGS in CDRv9, 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 CDRv9 release. For specific details on the changes since the CDRv7 off-cycle release, please consult the CDRv8 updates and CDRv9 updates sections.

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 3,916 (3.90%) from the final callset (Table 9). Of these 3,916, some were removed between CDRv6 and CDRv9 (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 CDRv9 srWGS SV callset contains 96,405 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)
Removed between CDRv8 and CDRv9	656	These are CDRv8 srWGS samples that were not included in CDRv9 for reasons unrelated to SV calling

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 read depth across the genome was highly variable. Non-uniformity of read depth 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 read depth 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 read depth 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 read depth 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.

We performed a sex concordance check as part of the CDRv7 off-cycle SV QC processes that differs from the process that was introduced in CDRv9, described in FAQ #10. This check did not remove any samples. For completeness, we retained the description of this check here: 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 read depth 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 Data Dictionary 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 CDRv9 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 14 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).

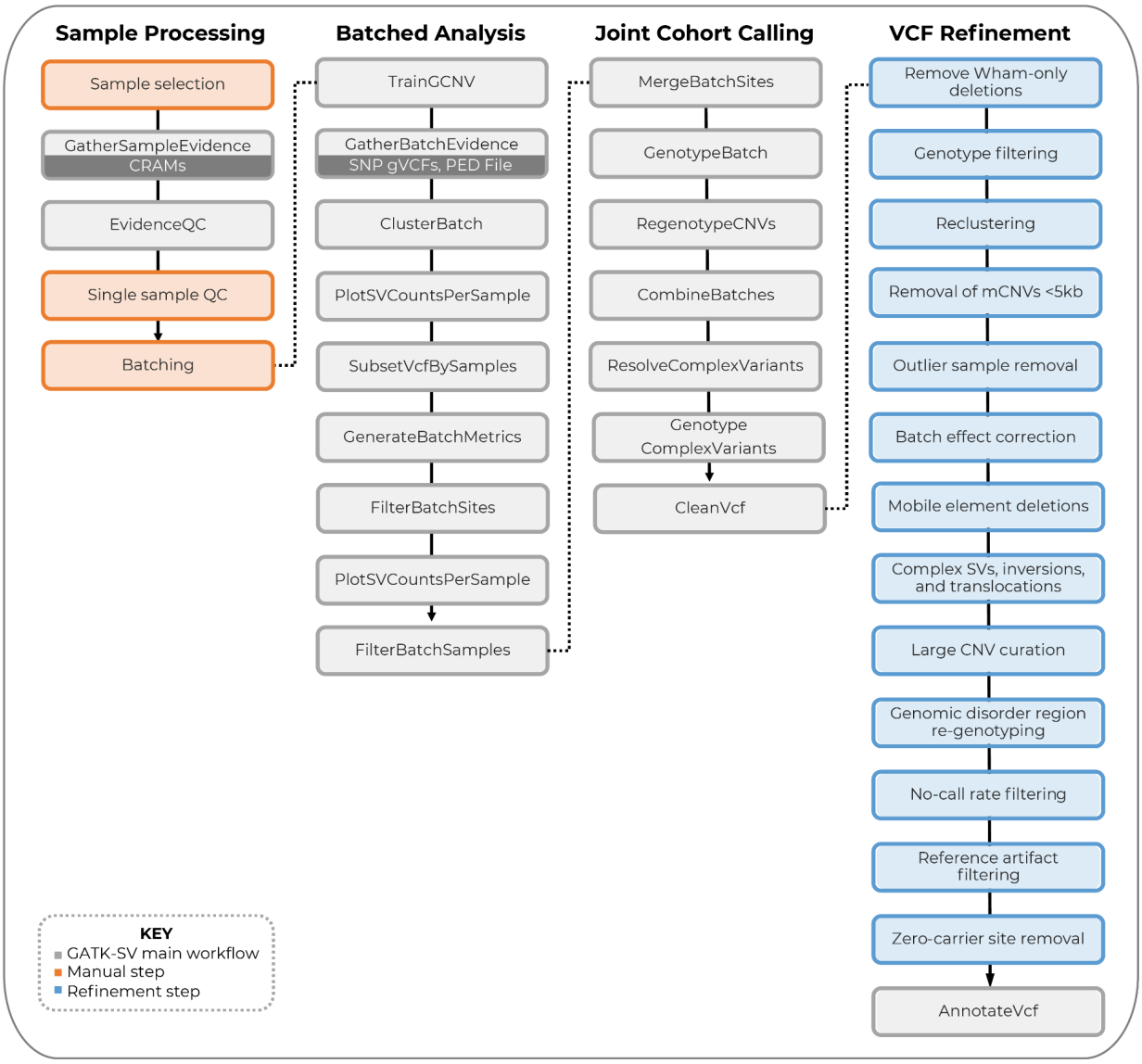


Figure 7 -- GATK-SV Pipeline Schematic. GATK-SV automated workflows are shown in gray and the names correspond to the name of the Workflow Definition Language (WDL) file. Manual steps performed in notebooks are shown in orange. Steps in blue are custom VCF refinement and QC steps for the *All of Us*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

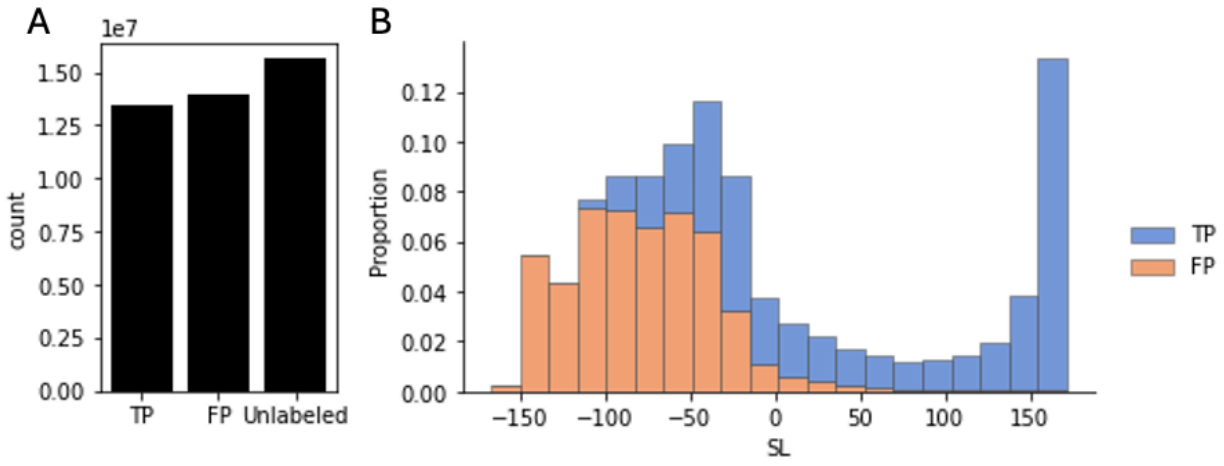


Figure 8 -- Training data for genotype filtering. (A) The proportion of each training label out of all SV genotypes in the training data, and (B) the SL score distribution produced by the trained model.

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.

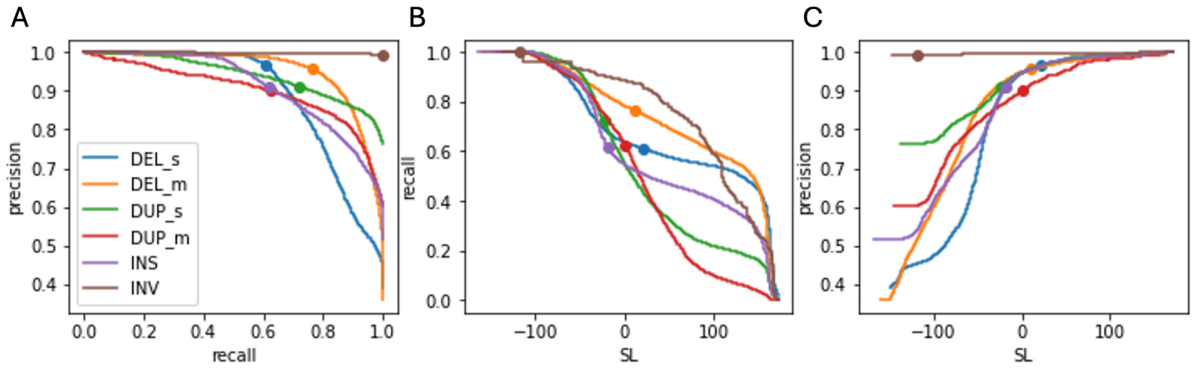


Figure 9 -- SL genotype filtering performance assessed against 97 lrWGS labeled test samples. (A) Precision-recall curves for all filtering classes, (B) recall as a function of the SL cutoff value, and (C) precision as a function of the SL cutoff value. Markers depict cutoffs used for genotype filtering.

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.

CDRv9 Updates

This section describes the changes that were applied to the CDRv8 callset to produce the CDRv9 callset.

Sample removal

We removed 656 samples from the CDRv8 dataset that were not included in the CDRv9 *All of Us* dataset, resulting in a final count for the CDRv9 srWGS SV dataset of 96,405 participants. We also removed all variant sites for which only the dropped samples were carriers.

Alpha-globin SV recovery

A known 5-kb deletion affecting *HBA1* and *HBA2* was genotyped but dropped from the callset due to the large number of overlapping CNVs. This deletion (AoU_srWGS_SV.v9.DEL_chr16_shard0_127) was recovered and added back to the callset as of CDRv9. Note that filtering and refinement was not applied to this variant.

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].

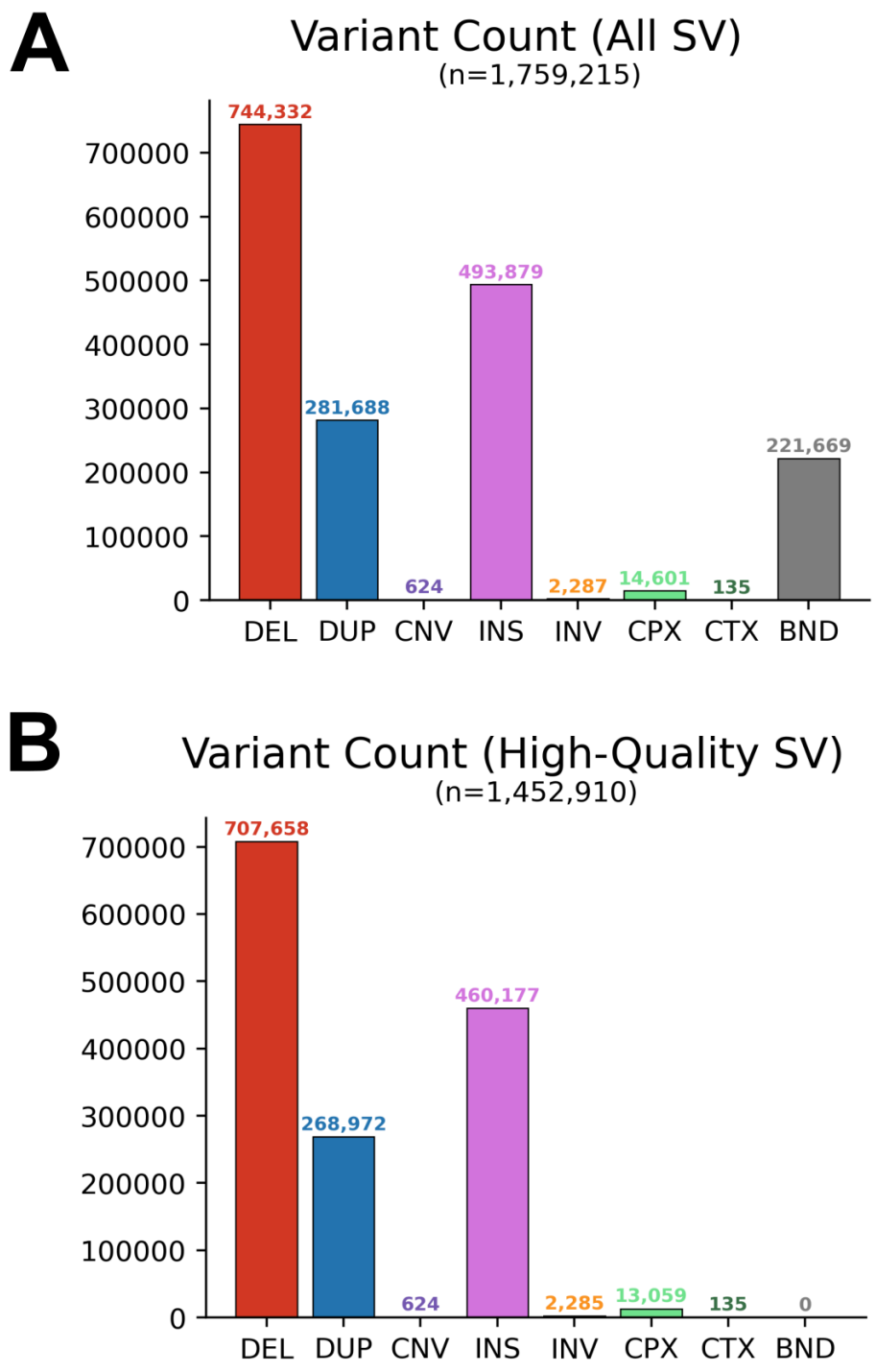


Figure 10 – SV counts in the complete callset and the high-quality SV callset. We observed 1,759,215 total SVs of which we determined 1,452,910 (82.6%) to be of high quality. (A) The total callset includes all variants in the callset regardless of the filter status. (B) The high-quality SV callset only 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.

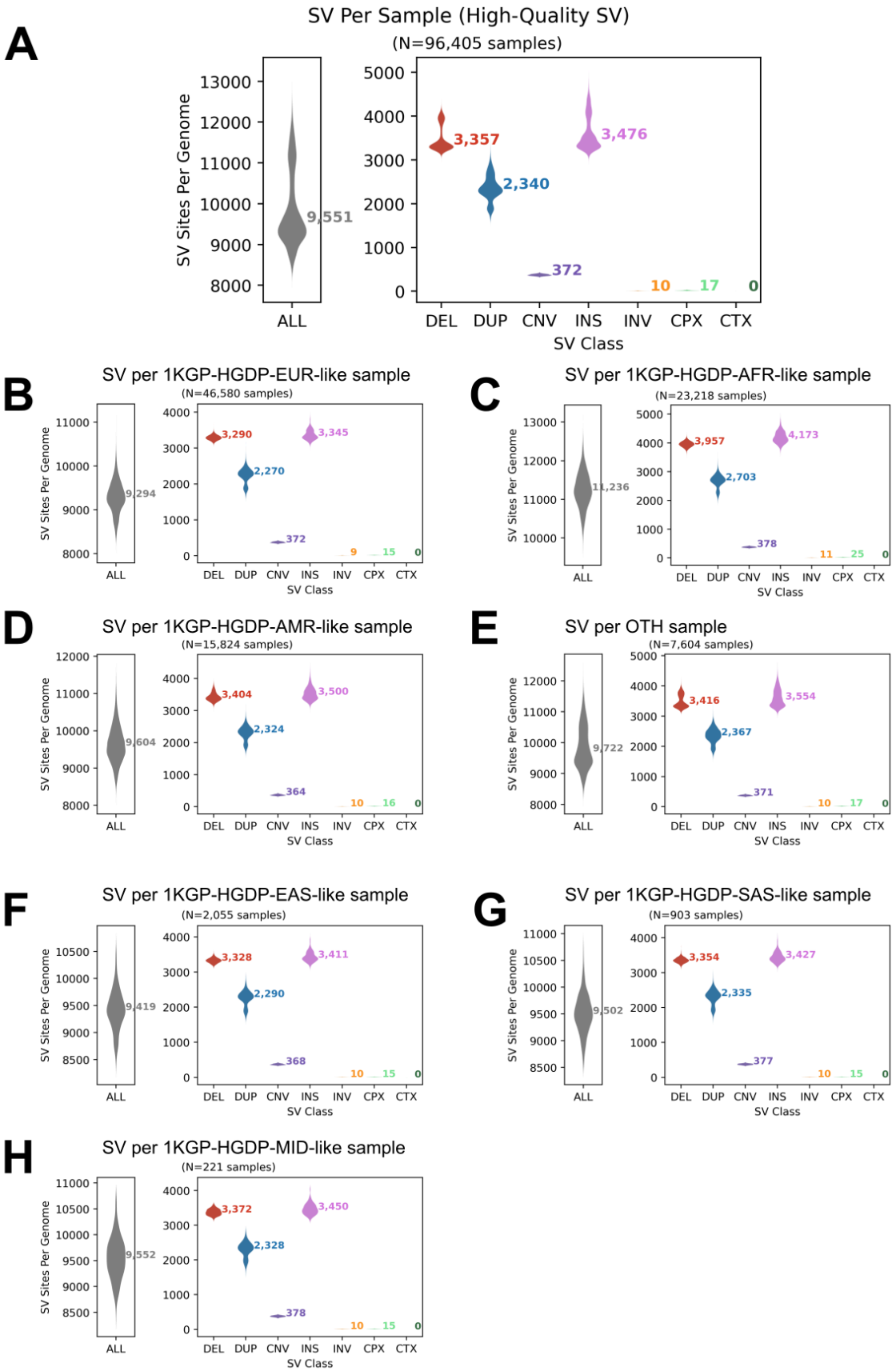

A
SV Per Sample (High-Quality SV)
(N=96,405 samples)
SV Sites Per Genome
SV Class
ALL
DEL DUP CNV INS INV CPX CTX
9,551
3,357
2,340
372
3,476
10
17
0
B
SV per 1KGP-HGDP-EUR-like sample
(N=46,580 samples)
9,294
3,290
2,270
372
3,345
9
15
0
C
SV per 1KGP-HGDP-AFR-like sample
(N=23,218 samples)
11,236
3,957
2,703
378
4,173
11
25
0
D
SV per 1KGP-HGDP-AMR-like sample
(N=15,824 samples)
9,604
3,404
2,324
364
3,500
10
16
0
E
SV per OTH sample
(N=7,604 samples)
9,722
3,416
2,367
371
3,554
10
17
0
F
SV per 1KGP-HGDP-EAS-like sample
(N=2,055 samples)
9,419
3,328
2,290
368
3,411
10
15
0
G
SV per 1KGP-HGDP-SAS-like sample
(N=903 samples)
9,502
3,354
2,335
377
3,427
10
15
0
H
SV per 1KGP-HGDP-MID-like sample
(N=221 samples)
9,552
3,372
2,328
378
3,450
10
15
0

Figure 11 – We observed a median of 9,551 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 1KGP-HGDP-AFR-like genetic ancestry group had the highest SV counts while those in the 1KGP-HGDP-EUR-like had the lowest SV counts.

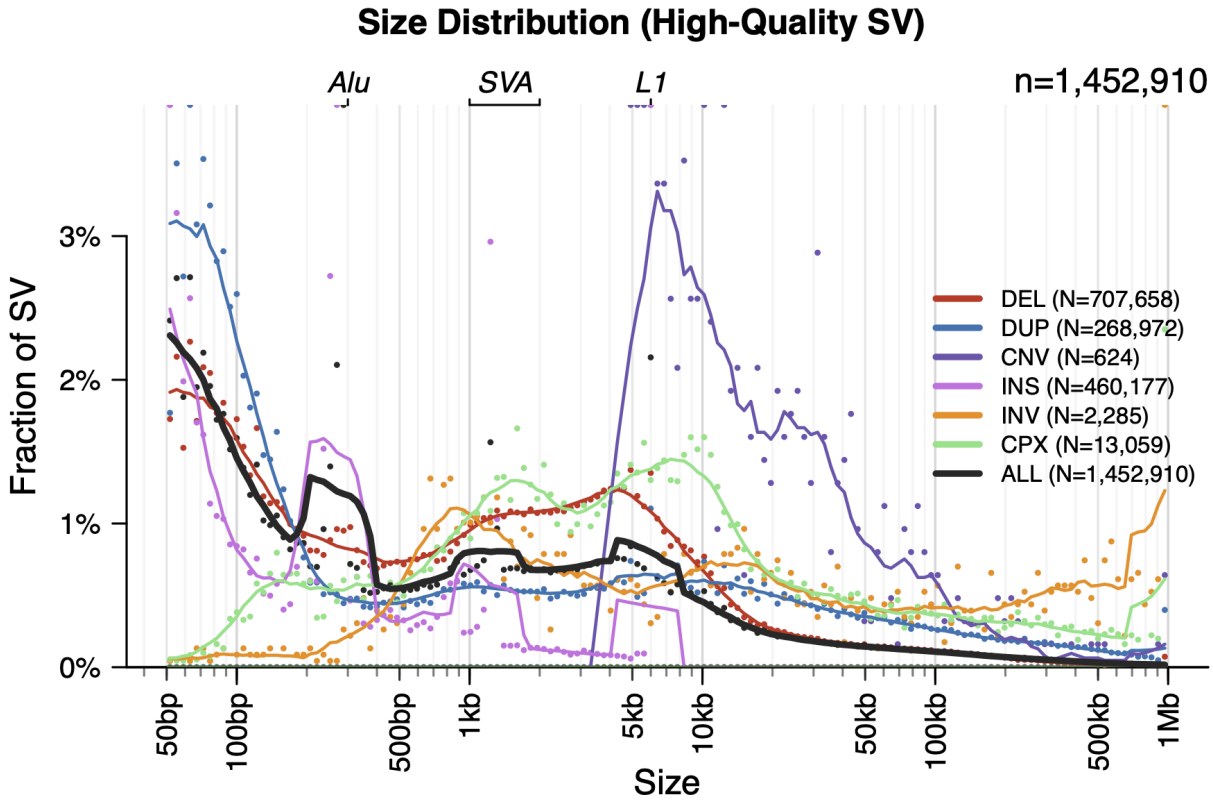


Figure 12 – SV size distribution matches previous expectations with notable insertion peaks corresponding to Alu, SVA, and LINE-1 insertions. Points represent the fraction of each SV type occupied by a given size range. Lines represent the rolling 10-bin average (the size ranges are divided into 150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

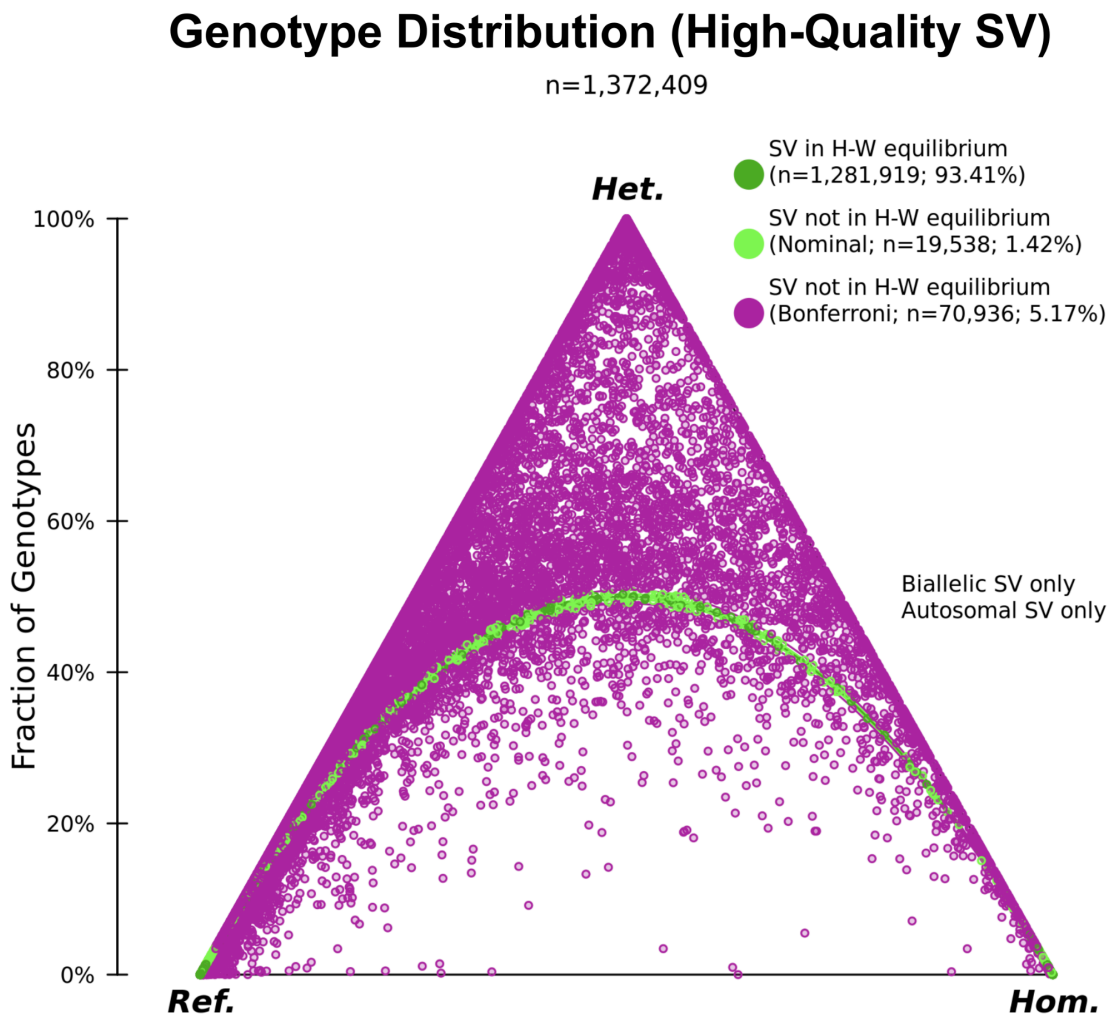


Figure 13 – Among high quality variants, 93.4% are in Hardy Weinberg Equilibrium (HWE). Of the 5.17% that fail, most of these failures appear to be driven by a bias towards genotyping variants as heterozygous. For this calculation, we included only the 92,740 unrelated samples and only biallelic SV sites on autosomes.

Long-Read Whole Genome Sequencing (lrWGS)

We have data representing 14,521 participants in the long-read genomic dataset. These data are particularly useful for resolving complex genomic regions, structural variants, and phasing of alleles, and 5mC methylation status to provide a more comprehensive view of the genome. The long-read genomic dataset includes three dataset releases, including the CDRv7 data, the CDRv8 data, and the current CDRv9 data. Some samples included in prior releases have been removed from CDRv9 due to participant withdrawals.

This report covers the QC steps for the new lrWGS samples representing 13,530 participants. For the previous QC results, please see the CDRv8 QC report and 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. We performed benchmarking for some parts of the analysis pipeline, which are described in the CDRv8 QC report. We link to this report when applicable.

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. The self-reported race and/or ethnicity data for the participants with lrWGS data can be found in Appendix H. Our sequencing data is from two different sequencing technologies, Pacific Biosciences (PacBio) High-Fidelity (HiFi) and Oxford Nanopore Technologies (ONT).

The lrWGS data are primarily aligned to the grch38_noalt reference, with select samples additionally aligned to the T2Tv2.0 reference. The QC steps are performed on the grch38_noalt aligned 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).

All lrWGS samples are aligned to the grch38_noalt reference and some lrWGS samples are aligned to the T2Tv2.0 reference [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).

Participants were selected for long-read sequencing by each genome center according to site-specific criteria, with shared requirements that all participants had existing short-read whole genome sequence (srWGS) data in CDRv7 and/or CDRv8 and sufficient high-molecular-weight (HMW) DNA available at the biobank. Electronic health record (EHR) availability was used as a prioritization factor but was not a strict inclusion criterion. Many participants also have matching RNA sequencing and proteomics data.

Site-specific selection criteria and sequencing strategies differed across centers. At HudsonAlpha, participants were selected from individuals who self-identified as African American or Black, and libraries were sequenced at mid-pass coverage (~12×). At Baylor College of Medicine (BCM) and Johns Hopkins University (JHU), participants were selected from self-identified hispanic samples primarily based on SNV and indels on chromosome 21 to maximize genetic diversity. This cohort was sequenced at high-pass coverage (~25×). At the University of Washington (UW), selection criteria were broader and included individuals with known disease status expected to benefit from long-read sequencing. All UW participant samples were sequenced on both PacBio HiFi and Oxford Nanopore Technology (ONT) platforms at high-pass coverage, and two multi-generational pedigrees were included to leverage phasing and variant resolution across platforms and family members. At the Broad Institute (BI), participants were selected based on computed genetic ancestry to maximize coverage of genetic ancestral diversity present in the short-read WGS data, with an emphasis on inclusion of admixed individuals, with similarity to multiple external continental references. Pedigrees (trios, quartets, and larger families) identified in the short-read data through kinship and identity-by-descent (IBD) analyses were prioritized and are typically represented at high-pass coverage. Library preparation and sequencing at BI used PacBio HiFi at mid-pass coverage, with additional high-pass sequencing for select pedigrees. To ensure inclusion of all available participants in the selection process, self-identified American Indian/Alaska Native (AI/AN) participants were included in all cohorts except HA after the program allowed inclusion of these participants following tribal consultation. Due to the timing of this process, the selection and sequencing of these participant samples by each site occurred as a later batch at each facility.

In this release, many of the mid-pass long-read sequencing participant samples have matching RNA sequencing and proteomics 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 utilizes DNA molecules circularized with SMRTbell adapters, which are repeatedly sequenced to generate highly accurate consensus reads [48]. Depending on the sample, HiFi sequencing was performed using either the Sequel IIe or the newer Revio platform. Oxford Nanopore Technologies (ONT) sequencing, which measures changes in ionic current as nucleic acids pass through a nanopore, was also used for a subset of samples [49]. The ONT workflows utilized the R10.4 PromethION and the R9 platforms. While some participants' samples were sequenced using only one of these approaches, some participants' samples were sequenced with both PacBio HiFi and ONT platforms.

Both long-read sequencing platforms are able to read methylation status by characterizing nucleotide kinetics data (PacBio) or ionic current data (ONT). These calls were not generated for the initial long-read samples in CDRv7, owing to the lack of vendor-supplied tools for processing the raw kinetics/current data. As vendor-supplied methylation callers are now available as part of the on-sequencer data processing, all long-read samples (both PacBio and ONT) subsequently generated and added to CDRv8 and CDRv9 all have 5mC methylation calls. These are provided within the aligned BAM files using standard tags MM (base modifications / methylation) and ML (base modification probabilities). Specifics on per-sample PacBio methylation calling and methylation-capable Nanopore basecalling are provided in the sample manifest.

Participant Sample cohorts

We separated the lrWGS samples into cohorts based on their sequencing facility and the sequencing technology (HiFi vs ONT) (Table 13). The CDRv9 dataset includes prior released data and represents 14,521 participants, though we have a total of 15,424 samples, since some participants are sequenced on multiple technologies or centers.

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. The cohort of mid-pass at 8x corresponds to the CDRv7 release. 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, where applicable. SVs were called for single samples. A batch effect analysis was conducted on factors influencing variant calling; these findings are detailed in the CDRv8 QC report, Appendix R. Joint-calling was performed strictly within designated technology and cohort groupings; consequently, select V8 samples were grouped and joint-called together with the V9 data.

Table 13 -- Sample cohorts for all 14,521 participants with lrWGS data

Cohort name	Sequencing facility	Sequencing platform	Number of samples	Minimum coverage
HA_PacBio	HA	PacBio Sequel IIe	991	Mid-pass (8x)
HA_PacBio	HA	PacBio Revio	1,210	Mid-pass (12x)

BI_PacBio	BI	PacBio Sequel IIe, PacBio Revio	9,934	Mid-pass (12x)
BCM_PacBio	BCM	PacBio Sequel IIe, PacBio Revio	748	High-pass (25x)
UW_PacBio	UW	PacBio Sequel IIe, PacBio Revio	397	High-pass (25x)
BCM_ONT	BCM	ONT R10.4 on PromethION	1,031	High-pass (25x)
JHU_ONT	JHU	ONT R10.4 on PromethION	717	High-pass (25x)
UW_ONT	UW	ONT R10.4 on PromethION	272	High-pass (25x)
UW_ONT_R9	UW	ONT R9.4 on PromethION	124	High-pass (25x)
Total number of samples			15,424	
Total number of participants			14,521	

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	CDRv9 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	We flagged samples if the sex chromosome coverage did not match expected coverage	- Sample swaps	We flagged 84 samples based on the ploidy formula
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_PacBio and BI_PacBio samples 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 CDRv9 release passed the fingerprint concordance check.

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). We flag samples if the sex chromosome coverage does not match the expected coverage.

Results

We performed a manual review for several samples that initially failed the sex concordance check because of the ploidy formula. We flagged samples that had low Y chromosome coverage or if they had non-canonical allosome coverage. We flagged 84 samples (82 participants) based on the metric and the flagged sample list is available 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 of the CDRv8 QC report.

Results

Most samples were under a cross-individual contamination rate threshold of 3%, other than one sample (Figure 14). We decided to include all samples in the release because they are close to the minimum threshold. The sample is available in the flagged samples list. The flagged samples are available in a list on the RW and the path can be found in the Data Dictionary.

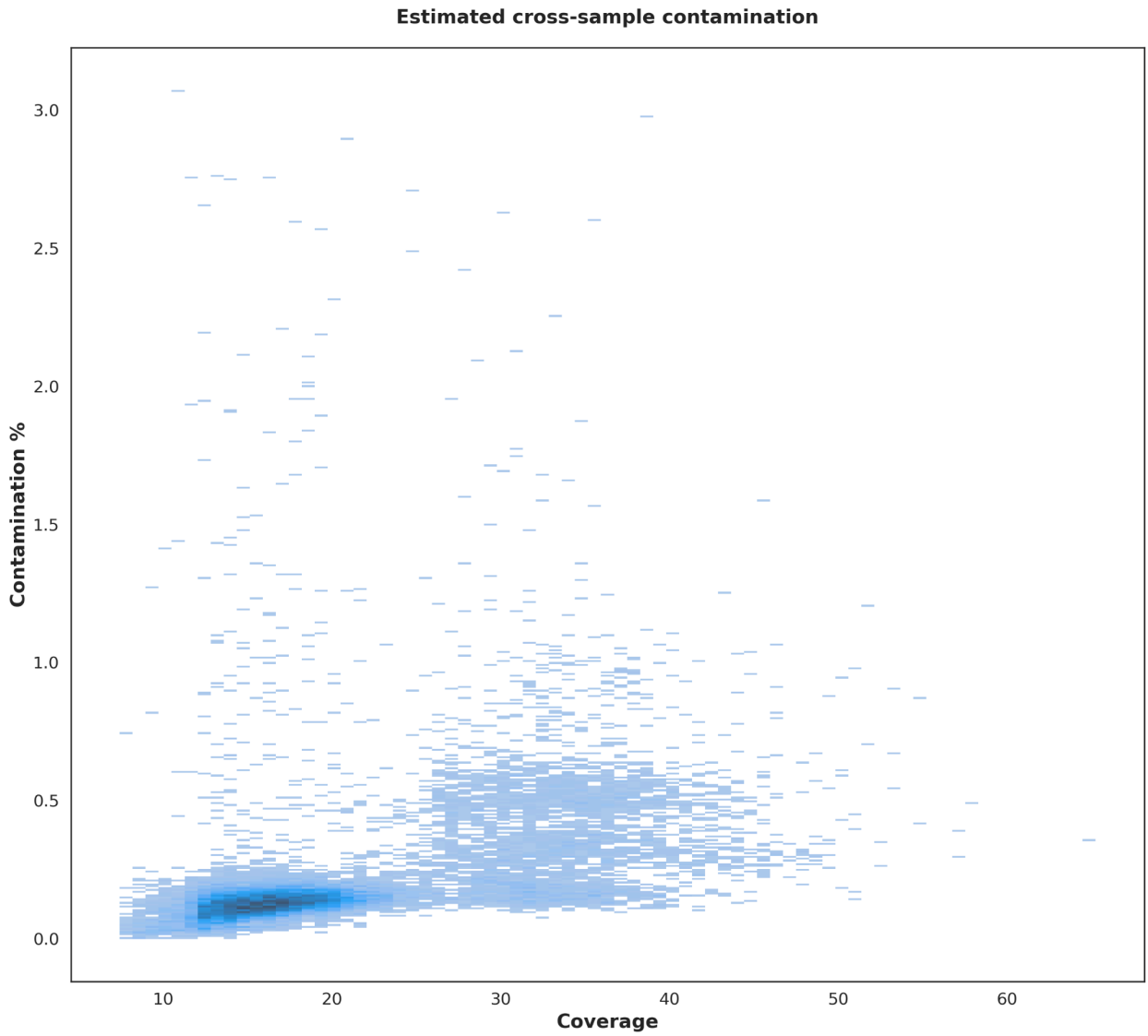


Figure 14 -- The cross-individual contamination for CDRv9 lrWGS samples.

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 CDRv9 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_PacBio and HA_PacBio (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.

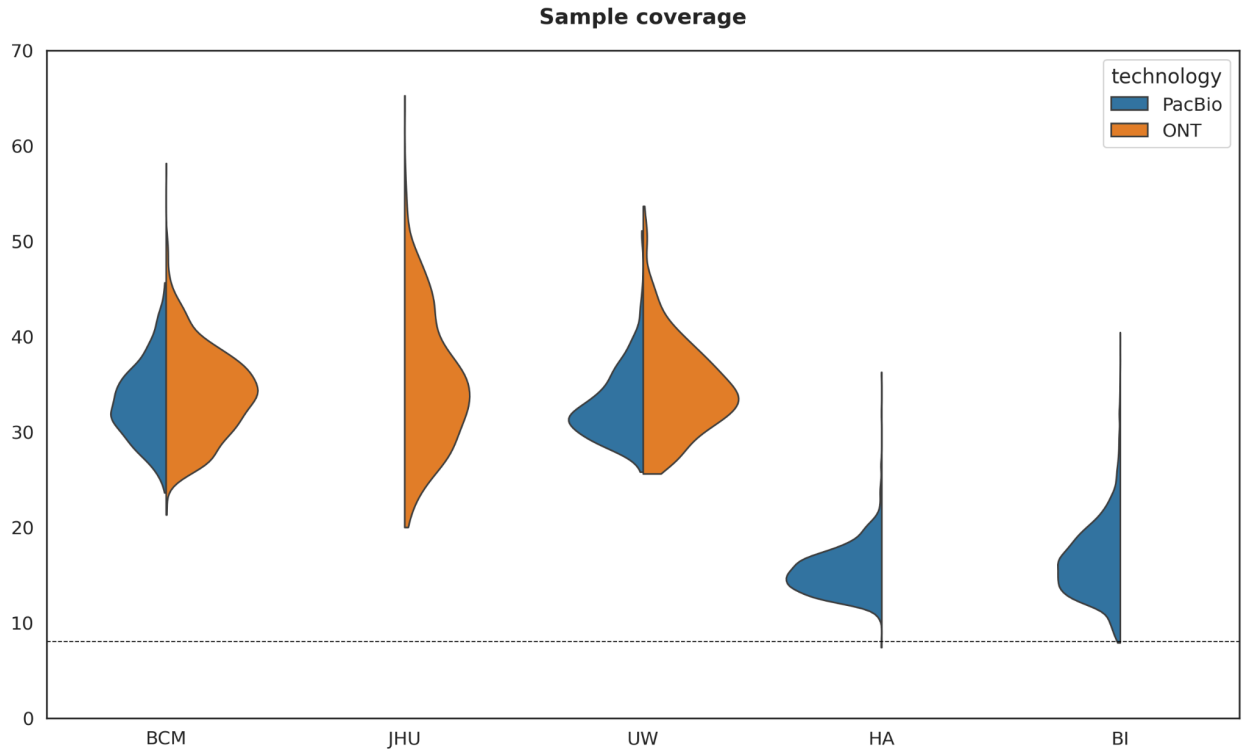


Figure 15 -- Coverage for each lrWGS sample in the CDRv9 release, showing the coverage distribution for each sequencing facility, for each sequencing technology.

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. Samples not meeting this requirement were flagged accordingly.

Results

A distribution of the read length median can be seen in Figure 16.

We also compared the read length median to the coverage at each sequencing facility and with every sequencing technology, described in Appendix Q of the CDRv8 QC report. The analysis demonstrated that there was no clear correlation between the read length median and the coverage.

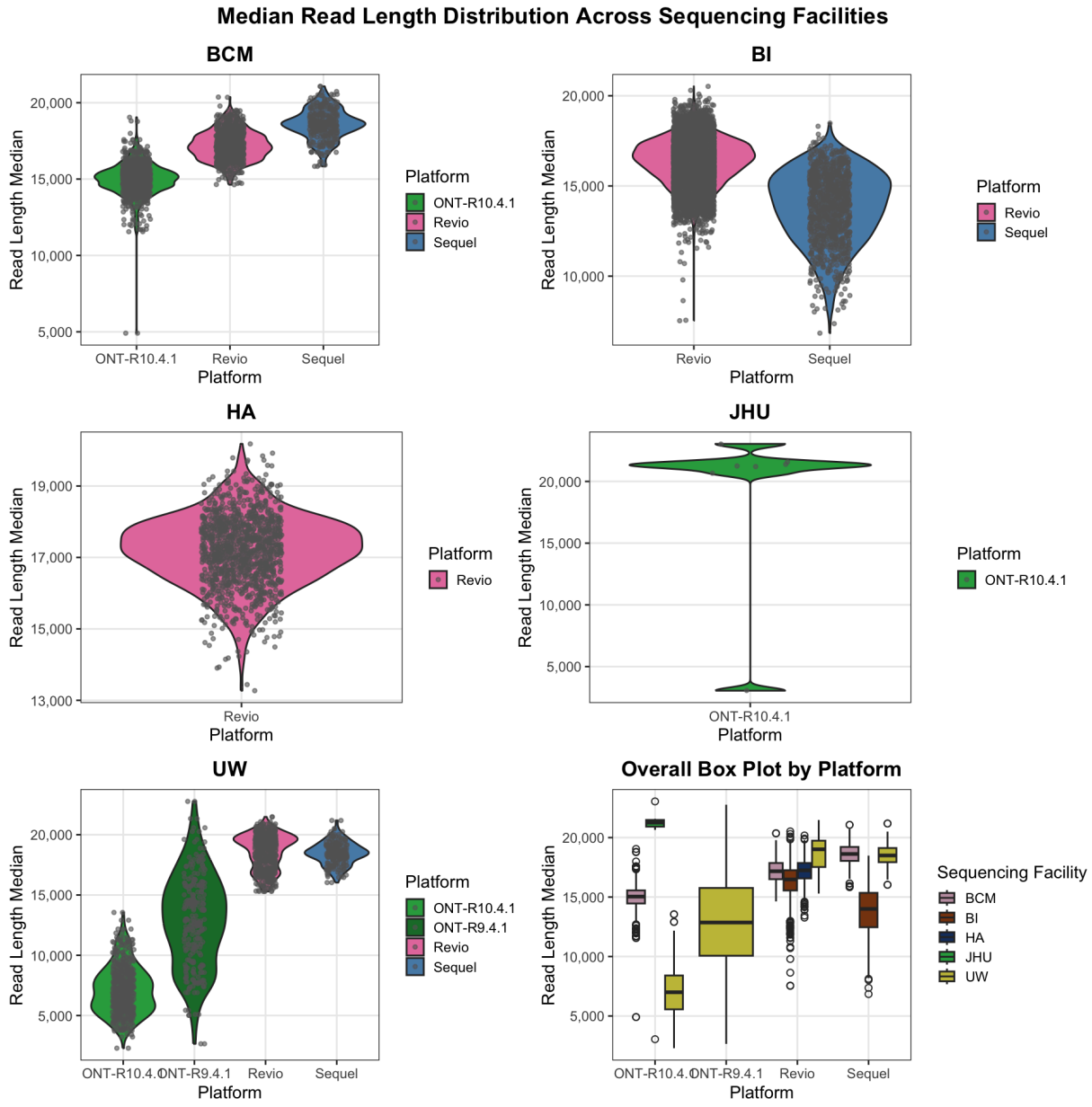


Figure 16 -- Read length median across sequencing facilities and platforms. Each subplot demonstrates the distribution of read length medians for one sequencing facility. Read length medians are displayed for all sequencing facilities in the last subplot.

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 grch38_noalt (Appendix N, Figure N.3) to generate phased SNPs, Indels, and structural variants. For some cohorts, variants were also called on T2Tv2.0. Please see the How the *All of Us* Genomic data are organized article on the User Support Hub [1] for an overview of the data available for each sample.

Results

When looking at *de novo* assembly genome lengths, we flagged 26 total samples. Flagged samples are circled in red in Figure 17. Flagged samples in the HA cohort exhibited unusually short haplotype 2 assemblies. Similarly, flagged samples in the BI cohort exhibited substantially longer haplotype assemblies than expected.

For some samples, we updated the assembler version (hifiasm version 0.19.5 to 0.20.0) because of a bug in the tool. The samples that used an updated assembler are available in the flagged samples list. The flagged samples are available in a list on the RW and the path can be found in the Data Dictionary. The assembler version does not change the downstream variant call format for analysis between versions

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 or longer than expected assembled genome length. This is most likely due to coverage requirements from the hifiasm tool, and misallocation by the assembler of contigs to one haplotype or the other.

The flagged samples are available in a list on the RW and the path can be found in the Data Dictionary.

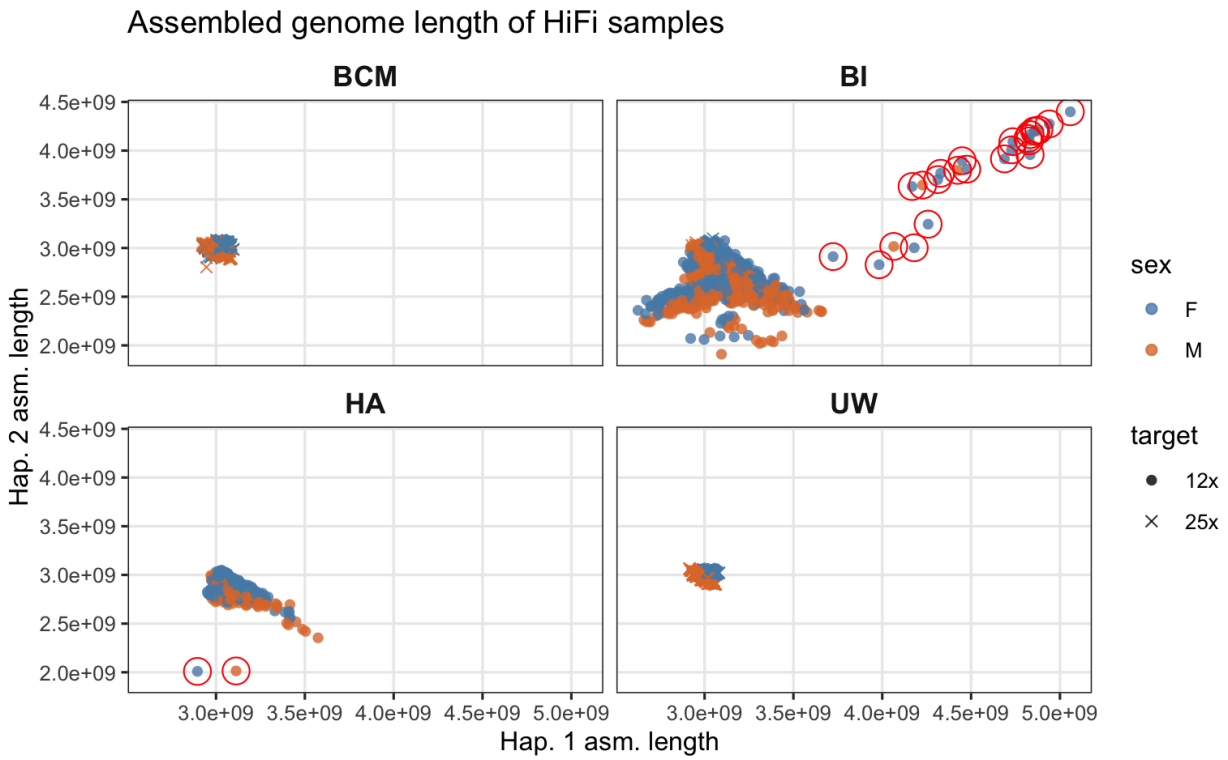


Figure 17 -- *De novo* assembled genome lengths of HiFi samples. Outlier samples are circled in red. Mid-pass (12x) and high-pass (25x) samples are denoted with circles and crosses, respectively.

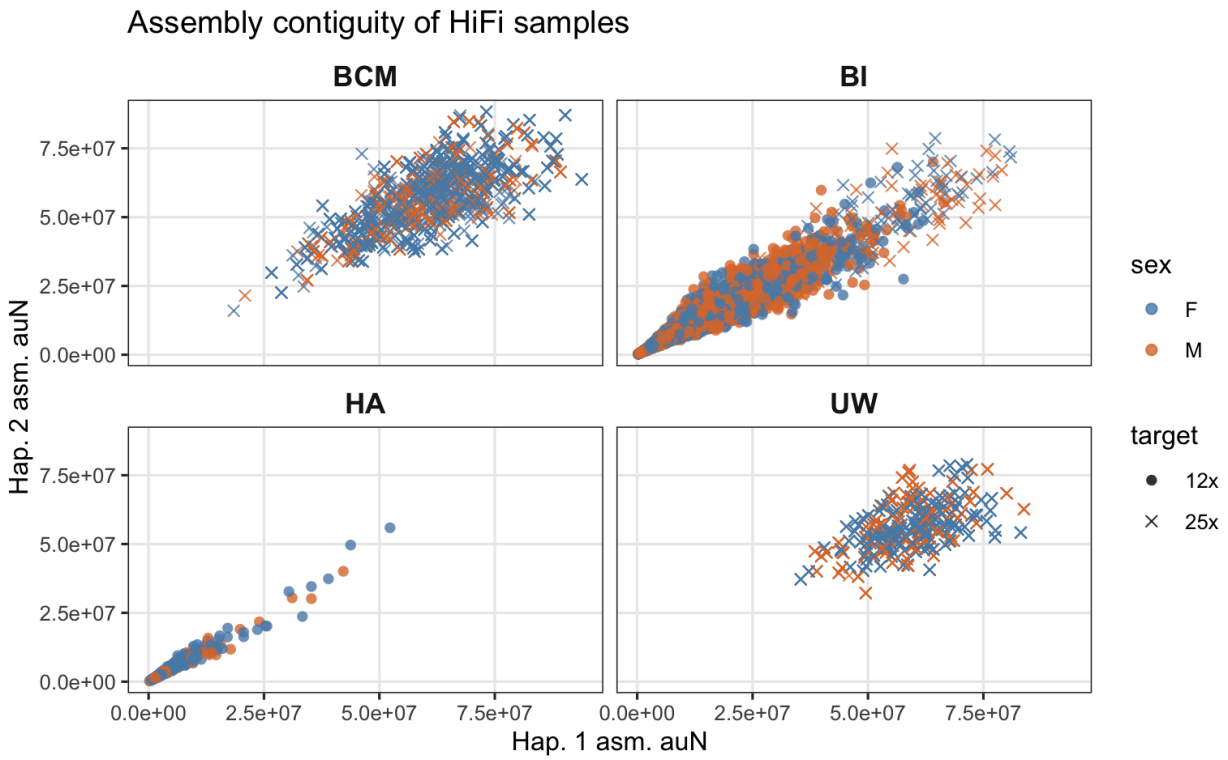


Figure 18 -- *De novo* assembly continuity of HiFi samples. Outlier samples are circled in red. Mid-pass (12x) and high-pass (25x) samples are denoted with circles and crosses, respectively.

SNP and Indel Calling

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 per cohort for grch38_noalt, using GLNexus [57]. Final joint-called SNP and Indel callsets were converted to Hail MatrixTable (Hail MT) format for analysis.

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.

We perform QC only on the grch38_noalt reference. Outliers are flagged by plotting variant counts versus coverage and manually evaluating the distribution for each cohort. The results are detailed in Appendix P. The focus for QC is mostly on insertions and deletions and distinguishing between ≥50bp calls and ≥20bp calls.

Results

Table 15 contains the total number of samples flagged as a result of the Structural Variant QC. The referenced figures are in Appendix P. The flagged samples are available in a list on the RW and the path can be found in the Data Dictionary.

In the BI_PacBio cohort we manually identified two outliers, one with high PAV insertion and deletion counts outside tandem repeats (Figure P.1) and one with low total PBSV and Sniffles insertions and deletions (Figure P.2). We see a consistency between the number of insertion and deletion counts for each genome region and variant size (Figure P.3). Evaluating the duplications and inversions, we did not identify any other outliers in the cohort (Figure P.4, Figure P.5, Figure P.6, Figure P.7).

We did not identify any outliers in the HA_PacBio, the BCM_PacBio, or the UW_PacBio cohorts.

The ONT UW cohorts have both R9 and R10 technology. The number of insertions and deletions are plotted for Sniffles (Figure P.8) and for PBSV (Figure P.9). We see 28 outlier samples in the PBSV insertions and deletions, which are all from the R9 technology (Table 15). We also see an outlier when plotting Sniffles inversions (Figure P.10).

We see the Sniffles variant counts versus coverage for Sniffles deletions duplications for the BCM_ONT and JHU_ONT cohorts (Figure P.11 and Figure P.12). No samples from the JHU_ONT cohorts are identified as an outlier. There are 50 samples identified as an outlier for BCM with Sniffles deletions (Figure P.11) and 45 samples identified as an outlier for BCM_ONT for Sniffles duplications (Figure P.12). The two rows for BCM_ONT have 44 samples in common (Table 15).

Table 15 -- Summary of SV QC for each cohort

Cohort name	Number of outliers	Reason	Figure
BI_PacBio	1	High PAV INS and DEL outside TRs.	Figure P.1
BI_PacBio	1	Low PBSV and Sniffles INS and DEL.	Figure P.2
UW_ONT_R9	28	PBSV outside tandem repeats: not on a linear (n_ins, n_del) locus.	Figure P.9
UW_ONT	1	High Sniffles INV.	Figure P.10
BCM_ONT	50	Low Sniffles INS and DEL ≥20bp.	Figure P.11
BCM_ONT	45	High Sniffles DUP.	Figure P.12

RNA Sequencing (RNA seq)

The RNA seq dataset consists of paired-end sequencing reads from 8,980 whole blood samples processed with Watchmaker RNA Library Prep kit with Polaris Depletion. Raw data is available in BAM file format containing STAR-aligned [58] reads to the hg38 reference that excludes ALT, HLA, and decoy and the GENCODE v48 GTF. The RNA sequencing metrics are provided from RNA-SeQC2 and RSEM results. eQTL and sQTL files are available for each sample. The BAM files provide a mapped record of gene activity, while the eQTL and sQTL files reveal how specific genetic differences actually influence that activity. These samples undergo a standardized alignment and quantification pipeline to ensure consistency across the release. The data is described in the 'How the *All of Us* Genomic data are organized' article on the User Support Hub [1].

Please see Appendix Q for an overview of the RNA processing pipeline following the QC steps.

Single sample QC

We performed an initial QC on the DRAGEN 4.2.4 aligned reads (Table 16). Samples were checked for RNA quality, composition, and mapping rate. Duplicate samples were identified and removed. The QC thresholds included a 5.5 RNA Quality Score (RQS), 80% alignment rate, 20% mRNA bases, and 8 million aligned read pairs.

Table 16 -- RNA seq single sample QC processes

QC process	Passing criteria	Error modes addressed	CDRv9 release results
RNA Quality	5.5 RNA Quality Score (RQS)	- Sample preparation error - Alignment issues	All RNA seq samples pass the cutoff
Alignment rate	> 80%	- Sample preparation error - Alignment issues - Contamination	All RNA seq samples pass the cutoff
mRNA bases	> 20%	- rRNA contamination	All RNA seq samples pass the cutoff
Aligned read pairs	> 8 million	-Sample preparation error - Poor mapping for eQTL and sQTL	All RNA seq samples pass the cutoff

Proteomics data

The proteomics dataset consists of Olink normalized protein expression data from 10,096 blood samples representing 9,969 participants. These 10,096 blood samples include 127 technical replicates that can be used for downstream normalizations. Blood samples were processed with the Olink HT kit. The Olink platform uses a high-precision technology called Proximity Extension Assay (PEA) to measure hundreds of proteins simultaneously from a single drop of blood or tissue. The proteomics participants largely overlap with participants represented in the RNA sequencing data and srWGS data. The data is described in the 'How the *All of Us* Genomic data are organized' article on the User Support Hub [1]. The Normalized Protein Expression (NPX) data is available for each participant. The data is also integrated with genomic information and provided in protein quantitative trait loci (pQTL) data.

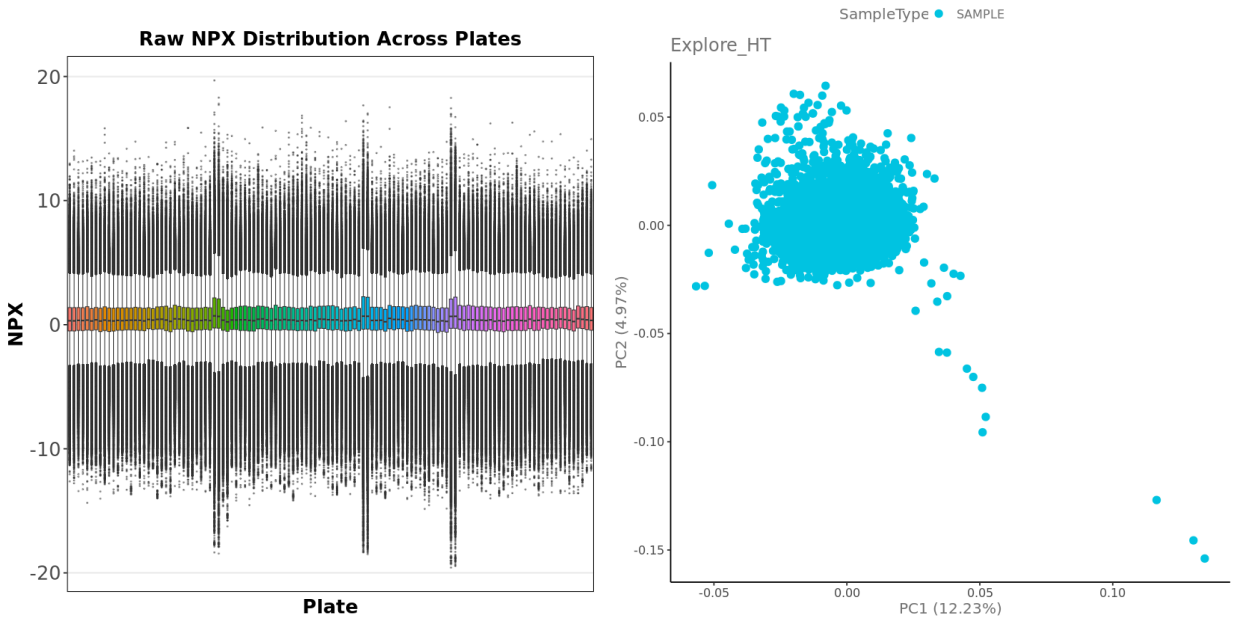
Consistency across batches

To evaluate the consistency across batches and identify the quality of the proteomics dataset, we analyzed the distribution of NPX values, performed a PCA analysis of the NPX scores, calculated coefficients (CVs) of variations for each protein, and evaluated the proportion of protein assays below the limit of detection (LOD). The proteomics sequencing pipeline is described in Appendix R.

Samples were organized into 61 Olink projects, which are samples that are run together on the sequencer. A project consists of two plates. The projects were processed across two primary batches, one from year 2024 and one from year 2025, as demarcated in the PlateIDprovided in the data and also available in the project-specific reports. Replicate samples were used to address potential batch effects. The project-specific reports are in PDF format available to researchers and provide software versions and sample counts for each batch.

To evaluate the presence of technical bias, we analyzed the distribution of the NPX values across all plates prior to further normalization as described in Appendix R (Figure 19, panel A). The NPX scores are demonstrated after plate-control normalization. Overall, the medians are consistent across plates, with some plates exhibiting higher technical variation, indicating overall that the batch effects are moderate.

We performed a PCA analysis on the proteomics NPX scores to project the data into a two-dimensional space and visualize the primary drivers of variance (Figure 19, panel B). We see no global batch-driven effects. A small subset of outlier samples is observed along PC1 and PC2, consistent with potential technical artifacts.



Fig**ure 19 --** Plate control Normalized NPX distribution and NPX PC analysis. (A) NPX distribution after plate-control normalization. (B) PCA analysis on the proteomics NPX scores.

To further investigate the quality, we calculated coefficients of variations (CVs) for each protein (Figure 20, Panel A). We identified a median CV of 0.222. We evaluated the proportion of protein assays below the limit of detection (LOD) (Figure 20, Panel B). Overall, we see high precision in the proteomics data.

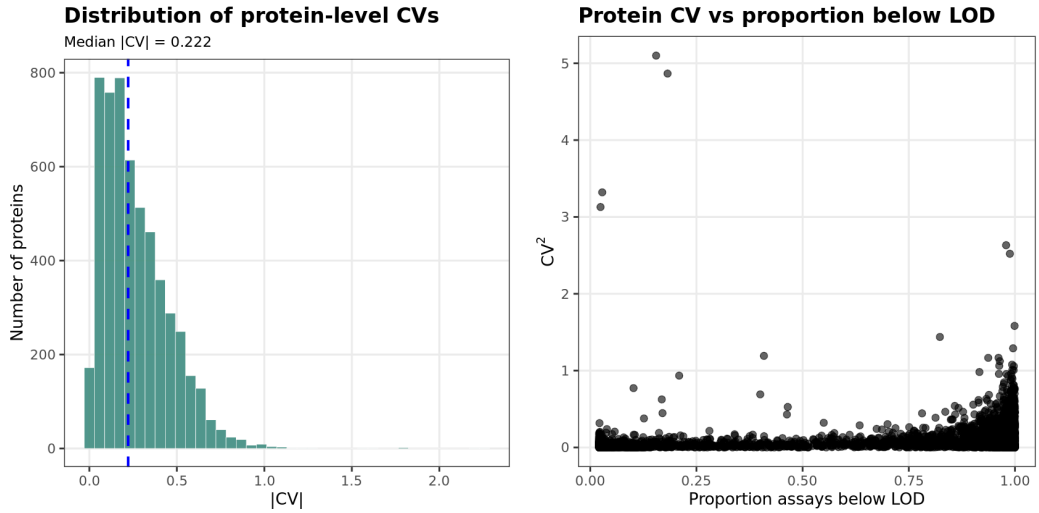


Figure 20 -- Distribution of protein-level CVs (A) and Protein CV versus proportion below limit of detection (LOD) (B).

Known Issues

The issues below apply to the CDRv9 release multi-omic data (arrays, srWGS, srWGS SVs, lrWGS, RNA sequencing, proteomics 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 Data Dictionary.

Known Issue #1: srWGS samples were affected by a data quality issue (N=152)

We have identified a data quality issue affecting 152 CDRv9 samples. These samples did not pass the srWGS coverage metrics mean coverage threshold of ≥30x and were included in the callset in order to retain samples with matching multiomics data (RNA sequencing, proteomics, and/or lrWGS). The minimum mean coverage retained was 28.5x. These samples will be retained in future releases. Note that per the selection criteria described for the Long Read samples, these samples had passed srWGS at the time of selection. Coverage fell below the threshold upon reprocessing of srWGS on DRAGEN version 3.7.8 where previously processed and released on DRAGEN version 3.4.12 in CDRv7

Affects:

- srWGS SNP & Indel samples

Suggested action:

- We do not believe that this will have an appreciable effect on most downstream analyses, since the mean coverage was close to the cutoff (min: 28.5x vs threshold of 30x). If you believe that your analysis is affected and you are not using matching multiomic data types, remove the samples from your analysis. We provide a list file of research IDs of affected samples in the CDR, see the CDR Directory Document.

Remediation:

- These samples will likely be retained in future releases and continued to be documented.

Known Issue #2: srWGS variant sites (0.015%) missing for 8000 samples in known genomic regions

We have identified two regions of missing variant sites (0.015% of the genome) missing on chromosomes 4 and 19 for 8000 participants (1.49% of participants with srWGS data). This was caused by a transient processing issue. The affected regions are chromosome 4: 56,580,000 - 57,040,000 and chromosome 19: 40,097,000 - 40,650,000. Each of the two regions has 4000 participants affected. For each participant, this is 0.015% of their genomic variants.

We will provide lists of research IDs for each chromosome (chromosome 4 and chromosome 19). Each list contains the participants who have the missing genomic variants in that region.

Additionally, we are providing two bed files for each missing region and a list of genes that intersect with these regions. See the CDR Directory Document for these lists.

Cohort extractions, VCFs, and PGENs are unaffected by this issue.

Affects:

- srWGS SNP & Indel VDS
- srWGS SNP & Indel VAT
- srWGS smaller callsets (ACAF, ClinVar, and exome), in the Hail Multi MT, Hail split MT, BGEN, and PLINK bed
    - This issue does not affect VCFs and PGENs
- Multiomics eQTL, sQTL, and pQTL

Suggested action:

- We recommend that if researchers are interested in these regions, they remove the affected participants from their analysis.

Remediation:

- We will remediate this issue in the next dataset release.

Known Issue #3: Small amount of multi-omic data missing matching genomic data (N=25)

We have identified research IDs in the srWGS SV, lrWGS, RNA-seq, and proteomics datasets that do not have matching array or WGS data. Please see a list of research IDs of the affected samples in the CDR.

Affects:

- srWGS SV data
    - 2 participants have no matching array data, 3 participants have no matching srWGS SNP/Indel data (<0.01% of srWGS SV data)
- lrWGS data
    - 22 participants have no matching array data, 23 participants have no matching srWGS SNP/Indel data (0.16% of lrWGS data)
- Multi-omic data
    - 1 participant with proteomic and RNA-seq data has no matching array or srWGS SNP/Indel data (0.01% of proteomics data)

Suggested action:

- If matched srWGS or array data is required in your multi-omic data analysis, remove the multi-omic samples from your analysis. We provide a list file of research IDs of affected samples in the CDR, see the CDR Directory Document.

Remediation:

- The cause is under investigation. It is possible the srWGS and/or arrays will be restored in future releases. If not, the multi-omic samples will likely be retained in future releases and continued to be documented

Known Issue #4: Small differences in genetic ancestry categories between multi-omic datasets

Due to a sample processing issue, some samples with multi-omic and/or srWGS SV data will have different genetic ancestry categorizations in these datasets compared to the srWGS SNV/indel data. All differences are to or from the “Remaining individuals” category and affects samples where the confidence was close to the 0.75 threshold (see Appendix G Methods)

For the multiomic data, this impacts 25 samples (0.28%) of RNA-seq and proteomics samples. The differences between srWGS genetic ancestry classifications and the genetic ancestry in pQTLs and eQTLs is expected. However, we believe the differences do not appreciably affect analyses.

For the srWGS SV data, the variant AF fields were calculated using these different genetic ancestry categorizations, which affected 523 (0.5%) srWGS SV samples. These samples were close to the threshold and we believe that these differences are jitter from a second training of the random forest classifier. We believe the differences do not appreciably affect analyses.

Affects:

- The per-ancestry protein quantitative trait loci (pQTL) data and expression quantitative trait loci (eQTL) data of the RNASeq and proteomics datasets.
- srWGS SV VCF fields that use genetic ancestry categories (eg, the population AF fields),

Suggested action:

- We do not believe that any action is necessary.

Remediation:

- The sample processing issue has been resolved and will be remediated for the next dataset 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 take no action.
- Filter the chrY variants using publicly available bed files, such as those published by Genomes-in-a-Bottle (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.6/GRCh38@all/XY/)

- Otherwise, we recommend that you use AD, GQ, and GT to filter variants on chromosome Y.

Remediation:

- We will address this in a future 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 is documented in Appendix S: Saliva and blood batch effect analysis. 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 >0.45, see Appendix I. These samples may represent true individuals or may represent the same person submitting data multiple times.

   In the CDRv9 release, we found 69 clusters of three or more genomic duplicate samples. Based on the rarity of identical triplets (and quadruplets, etc), we believe that each cluster likely represents 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 [59], 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 30 protein-coding genes, including 7 CMRG genes such as KCNE1, CBS, and MAP2K3 in a separate callset using GATK and the

masked hg38 reference, which is released as a CMRG callset in the auxiliary data. Please see the 'How the All of Us Genomic data are organized' article for more information about the CMRG callset.

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 do the genetic ancestry groups change between releases?

   In CDRv9, we have 12,234 participants (2.98% of the total CDRv8 srWGS samples) whose genetic ancestry classification changed to or from the "Remaining individuals" category from the previous release (see Table below and Appendix G). In this release, we updated the metadata used for categorial ancestry to gnomAD version 3.1.2. Additionally, with each release, we repeat the process of defining HQ sites for the training model (Appendix I).

   Both these factors can lead to changes in the ancestry classification for a sample.

   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.

Genetic ancestry group changes

Genetic ancestry group name	Percent change
1KGP-HGDP-AFR-like	3.22
1KGP-HGDP-AMR-like	1.79
1KGP-HGDP-EAS-like	0.04
1KGP-HGDP-EUR-like	2.32
1KGP-HGDP-MID-like	2.51

1KGP-HGDP-SAS-like	0.87
Remaining individuals	12.80

8. 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 VDS as the FT annotation, 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.

9. 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.

10. How did we investigate samples that failed the sex concordance check? Were any samples that failed the sex concordance check added back?

    A subset of array and srWGS samples were designated as sex concordance exceptions after failing the participant-reported sex at birth vs. genetically-inferred sex concordance check, while passing all genotyping and sequencing quality control metrics (887 array samples (0.16%) and 495 srWGS samples (0.09%)). The biobank conducted an independent validation for these samples and confirmed that there is no evidence of a sample swap. This external verification further supports the integrity of the specimens and the associated genomic data.

    Given the otherwise high-quality data and the absence of additional QC indicators consistent with sample mix-up, cross-contamination, or processing errors, these discordances are considered unlikely to reflect true sample identity issues. Instead, they may reflect data entry errors in the reported sex field at the time of sample collection or other differences.

    As a result, these samples have been retained in downstream analyses as sex concordance exceptions. Including these samples as noted exceptions preserves valuable data for the research community while maintaining confidence in sample integrity based on comprehensive QC evaluation and independent validation.

The exception samples will be provided as two separate lists (one for array and one for srWGS). The corresponding file paths for these lists are documented in the Data Dictionary.

11. What are the differences between genetic ancestry, genetic admixture, and race and/or ethnicity?

    Please see the following definitions from *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* from NASEM [60].
    - **Ethnicity:** a sociopolitically constructed system for classifying human beings according to claims of shared heritage often based on perceived cultural similarities (e.g., language, religion, beliefs); the system varies globally.
    - **Race:** a sociopolitically constructed system for classifying and ranking human beings according to subjective beliefs about shared ancestry based on perceived innate biological similarities; the system varies globally.
    - **Ancestry:** a person's origin or descent, lineage, "roots," or heritage, including kinship.
    - **Genetic ancestry:** the paths through an individual's family tree by which they have inherited DNA from specific ancestors. Genetic ancestry can be thought of in terms of lines extending upwards in a family tree from an individual through their genetic ancestors. Shared genetic ancestry arises from having genetic ancestors in common (that is, overlapping lines of ancestry). In practice, shared genetic ancestry is typically inferred by some measure(s) of genetic similarity.
    - **Admixture:** an individual is described as admixed when they have lines of ancestry that trace back to multiple distant geographic origins on a recent timescale: as an example, individuals of Central and South America whose ancestry 600 years ago traces to individuals mostly living in western Europe, west Africa and Central/South America (Winkler et al., 2010). A difficulty with the concept is the often-implicit timescale being considered. All humans are admixed, but not everyone is recently admixed: for some, ancestry lines will trace back to geographically distant ancestors within a few generations, whereas for others, the same process occurs on much longer timescales. A further challenge is the framing of admixture as the blending of "source populations," which may erroneously imply the existence of homogeneous populations in the past.

TABLE 5-1 Recommended Approaches for the Use of Population Descriptors by Genomics Study Type

This table should be read and interpreted in conjunction with the report text. Consult the decision tree in Appendix D for more information and Chapter 5 text for best practices for each study type. See also the terminology box preceding the table and descriptions of each study type in Chapter 1 section "Classification of Genomics Study Types." For any given study, the use of multiple descriptors may be preferable.

LEGEND

+ Preferred population descriptor(s)	− Should not be used
? In some cases; refer to Ch. 5 text and the decision tree in Appendix D	E Descriptors could be used if appropriate proxies for environmental, not genetic, effects

GENOMICS STUDY TYPE	Race	Ethnicity/ Indigeneity	Geography	Genetic Ancestry	Genetic Similarity	Notes
1: Gene Discovery - Mendelian Traits	−	?	?	?	+	Similarity suffices as a genetic measure; at fine-scale, other variables may be useful
2: Trait Prediction - Mendelian Traits	−	E	E	?	+	No population descriptors may be necessary for analysis
3: Gene Discovery - Complex Traits	−	E	E	?	+	Similarity suffices as a genetic measure
4: Trait Prediction - Complex Traits	−	E	E	?	+	Similarity suffices as a genetic measure
5: Cellular and Physiological Mechanisms	−	E	E	−	?	No population descriptors may be necessary for analysis
6: Health Disparities with Genomic Data	E	E	E	?	+	Not all health disparities studies rely on descent-associated population groupings, so none may be necessary for analysis
7: Human Evolutionary History	−	?	+	+	+	Reconstructing genetic ancestry may be of central interest

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] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.
[59] Wagner, J., et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol 40, 672–680 (2022).
[60] 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.
[61] 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.
[62] 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)
[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] 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
[65] ***Downloads | gnomAD.*** (n.d.). Retrieved February 1, 2021, from https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg.
[66] 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.

[67] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *Journal of Machine Learning Research* 12, pp. 2825-2830, (2011).
[68] 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
[69] Erdős, P. **On cliques in graphs**, Israel Journal of Mathematics, 4 (4): 233–234, (1966), doi:10.1007/BF02771637, MR 0205874, S2CID 121993028
[70] ***Structural variant (SV) discovery*** (n.d.). Retrieved March 15, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022487952155-Structural-variant-SV-discovery\
[71] **WDL Specification**, from https://github.com/openwdl/wdl
[72] 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
[73] Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics*. DOI: 10.1093/bioinformatics/bts606.
[74] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

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 Niall Lennon - Broad Institute Heidi Rehm - Massachusetts General Hospital 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 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).

We use cluster definition files (.egt) that were created for the CDRv7 data release (C2022Q4R9) 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.

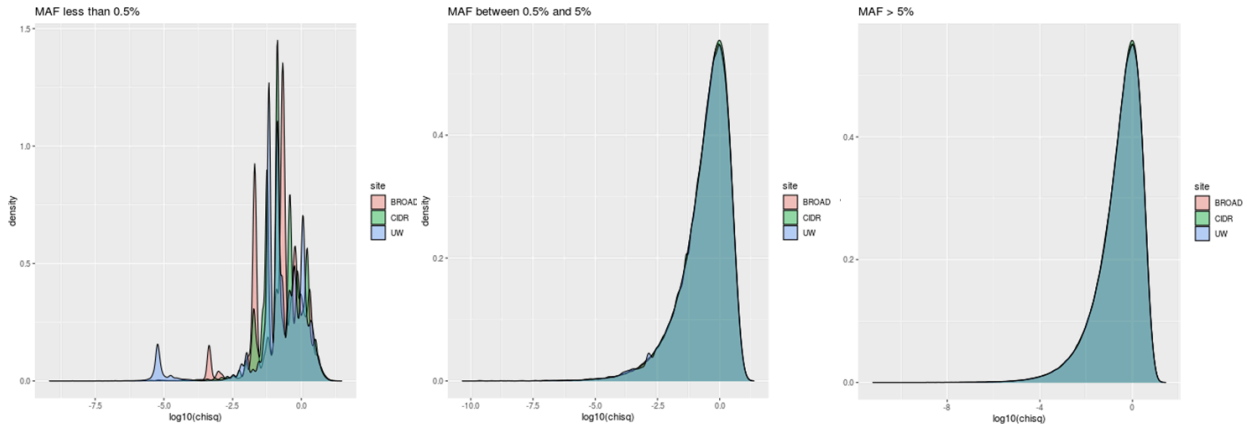


Figure B.1 Comparisons shown in Figure B.1 broken out into MAF bins.

	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 [61]. 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 [61]. 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 (though see FAQ # 10 for an exception). An array sample will have the status ABANDON if the call rate is less than 98% after at least 2 attempts at the GC.

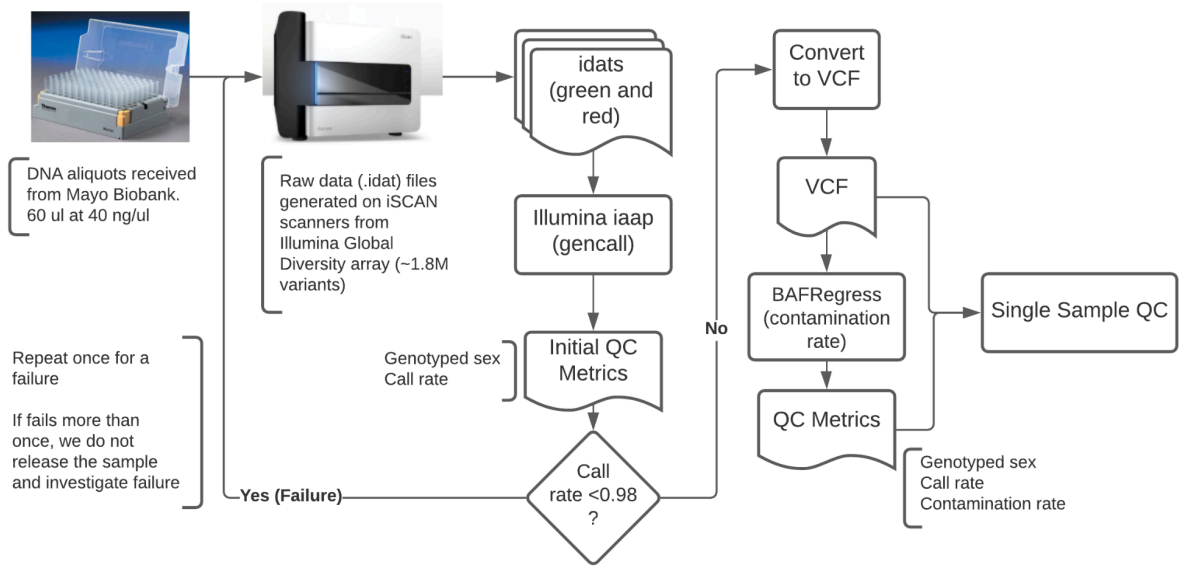


Figure B.3 -- Overview of the array processing pipeline.

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 C2025Q4R6). 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 -- Breakdown on sex assigned at birth for CDRv9 participants with multi-omics Data

Sex assigned at birth responses	Array counts (%)	srWGS counts (%)	srWGS SV counts (%)	lrWGS counts (%)	RNA Seq counts (%)	Proteomics counts (%)
Female	335,725 (60.61%)	324,751 (60.63%)	58,102 (60.27%)	8,101 (55.79%)	4,432 (49.35%)	4,924 (49.35%)
Male	212,765 (38.41%)	205,634 (38.39%)	37,183 (38.57%)	5,854 (40.31%)	4,060 (45.21%)	4,516 (45.21%)
Other responses*	5,459 (0.99%)	5,277 (0.99%)	1120 (1.16%)	566 (3.90%)	488 (5.43%)	529 (5.43%)
Total	553,949	535,662	96,405	14,521	8,980	9,969

Table Notes:

- Percentages may not add to 100 due to rounding.
- *The “Other responses” count includes any or no response for sex_at_birth. The available options in the CDR are “I prefer not to answer”, “None of these fully describe me”, “Intersex”, “No matching concept”, and “PMI: Skip”. “No matching concept” and “PMI: Skip” are separate counts both referring to no response for sex_at_birth. These are separate because participants in “No matching concept” did select a gender option for this survey question. The terms used here are the Concept Names as they appear in the CDR.

Appendix 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 CDRv9 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) [62]. 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. As of the CDRv8 release (C2024Q3R3), all samples are from DRAGEN 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 CDRv9 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 Data Dictionary 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” [60]. 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]. Additionally, please see our FAQ on population descriptors: What are the differences between genetic ancestry, genetic admixture, and race and/or ethnicity?

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) [64] 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 [63,65], 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-like or African); 1KGP-HGDP-AMR-like (AMR-like or Americas); 1KGP-HGDP-EAS-like (EAS-like or East Asian); 1KGP-HGDP-EUR-like (EUR-like or European); 1KGP-HGDP-MID-like (MID-like or Middle Eastern); 1KGP-HGDP-SAS-like (SAS-like or South Asian); and not belonging to one of the other ancestries or is an admixture (REM or remaining individuals) (see Table G.1). Because the HGDP and 1000 Genomes reference dataset has uneven representation across global populations, ancestry and admixture estimates may be less precise for individuals from underrepresented or poorly sampled groups, and should be interpreted with appropriate caution.

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 v9 Count (percentage)	Notes
1KGP-HGDP-AFR-like	AFR	African	94,465 (17.6%)	
1KGP-HGDP-AMR-like	AMR	Americas	97,261 (18.2%)	Who does the genetic ancestry group 'Americas'(1KGP-HGDP-AMR-like) include?
1KGP-HGDP-EAS-like	EAS	East Asian	12,980 (2.4%)	
1KGP-HGDP-EUR-like	EUR	European	284,004 (53.0%)	
1KGP-HGDP-MID-like	MID	Middle Eastern	1,292 (0.2%)	
1KGP-HGDP-SAS-like	SAS	South Asian	5,631 (1.1%)	
Remaining individuals	OTH	Remaining	40,029 (7.5%)	Not belonging to one of the other genetic ancestries or is an admixture
Total count			**535,662**	

We trained a random forest classifier [66,67] 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). The autosomal exome was derived from the exon regions of all autosomal, basic, protein-coding transcripts in GENCODE v42 [68].

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	542
1KGP-HGDP-AFR-like	HGDP	461
1KGP-HGDP-AMR-like	1000 Genomes	393

1KGP-HGDP-AMR-like	HGDP	159
1KGP-HGDP-EAS-like	1000 Genomes	378
1KGP-HGDP-EAS-like	HGDP	447
1KGP-HGDP-EUR-like	1000 Genomes	347
1KGP-HGDP-EUR-like	HGDP	441
1KGP-HGDP-MID-like	HGDP	126
1KGP-HGDP-SAS-like	1000 Genomes	498
1KGP-HGDP-SAS-like	HGDP	292
Remaining individuals (Note: we do not train on this category, we only test on this category)	1000 Genomes	0
Remaining individuals (Note: we do not train on this category, we only test on this category)	HGDP	30

As seen in Figure G.1, the projection shows a continuum of diversity in the *All of Us* cohort. Of individuals in the CDRv9 srWGS dataset, we estimate that 17.6% were similar to the 1KGP-HGDP-AFR individuals; 18.2% were similar to 1KGP-HGDP-AMR individuals; 2.4% were similar to 1KGP-HGDP-EAS individuals; 1.1% were similar to 1KGP-HGDP-SAS individuals; <1% were similar to 1KGP-HGDP-MID individuals; and 53.0% were similar to 1KGP-HGDP-EUR reference 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): 0.02 (See Table G.3)
2. Error rate (not including Remaining individuals): 0.006 (See Table G.4)

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

Who does the genetic ancestry group 'Americas'(1KGP-HGDP-AMR-like) include?
Why do the genetic ancestry groups change between releases?
What are the differences between genetic ancestry, genetic admixture, and race and/or ethnicity?

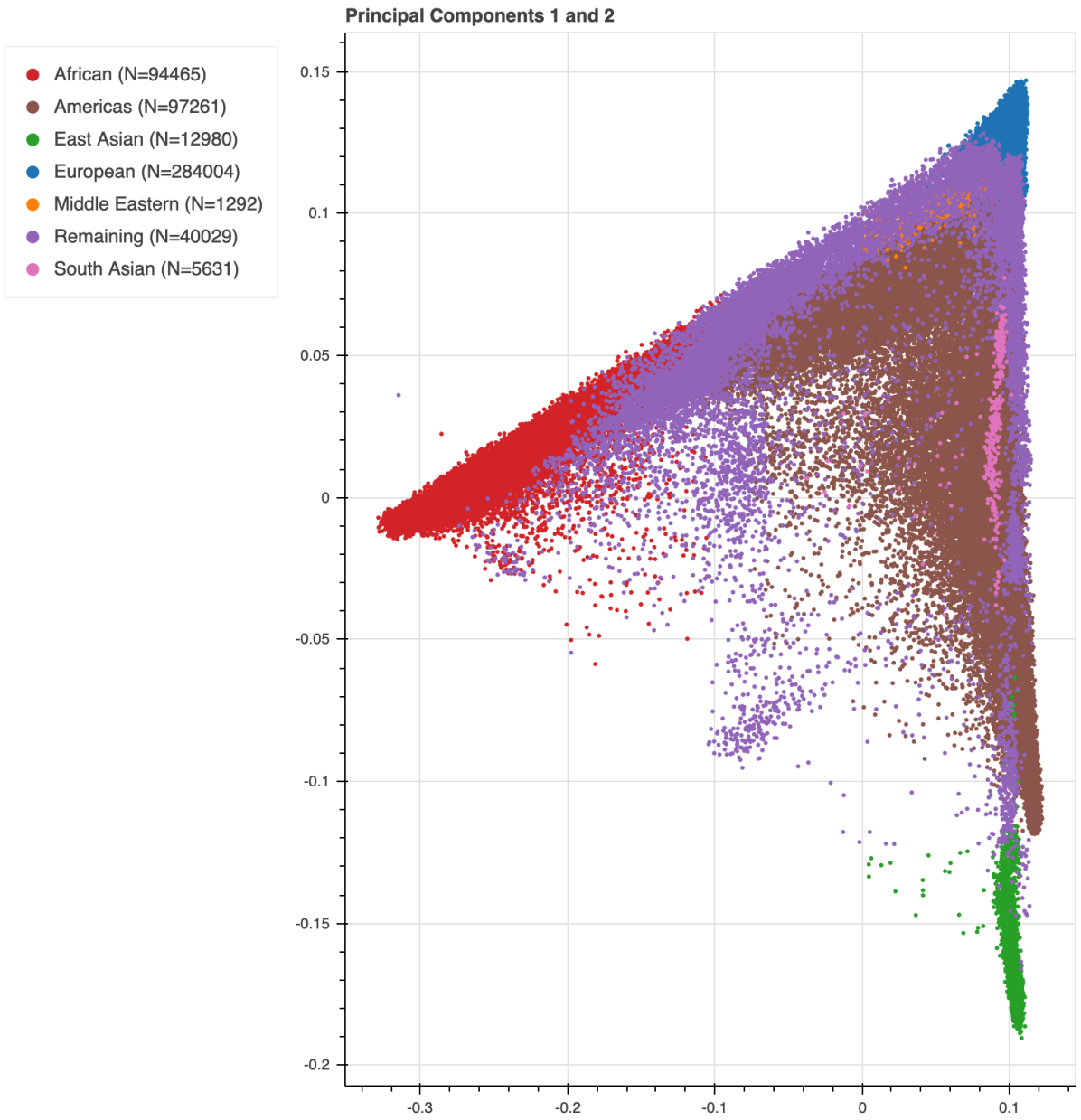


Figure G.1 -- Projection of the *All of Us* srWGS onto the PCA space of the 1000 Genomes and HGDP samples plotted on the first two principal components (PC1 on x-axis and PC2 on the y-axis) derived from genotype calls. Colored points represent proximity to one of six reference populations or “Remaining” depending on classifier confidence.

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	1	0	0	0	1	0
1KGP-HGDP-AMR-like	0	194	0	0	0	6	0
1KGP-HGDP-EAS-like	0	0	197	0	0	3	0
1KGP-HGDP-EUR-like	0	0	0	200	0	0	0
1KGP-HGDP-MID-like	1	0	0	0	49	0	0
Remaining individuals	0	0	0	0	0	26	4
1KGP-HGDP-SAS-like	0	0	0	0	0	8	192

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	199	1	0	0	0	0
1KGP-HGDP-AMR-like	2	198	0	0	0	0
1KGP-HGDP-EAS-like	0	0	200	0	0	0
1KGP-HGDP-EUR-like	0	0	0	200	0	0
1KGP-HGDP-MID-like	1	0	0	0	49	0
1KGP-HGDP-SAS-like	0	0	2	0	0	198

Appendix H: Self-reported race and/or ethnicity

As seen in Table H.1, the race and/or ethnicity breakdown of the genomic and multi-omic data is similar to all participants *All of Us* CDR release C2025Q4R6. Samples with "Skip" responses include participants that answered "prefer not to answer", entered blank text, or did not respond to the survey question.

Additionally, please see our FAQ on population descriptors: What are the differences between genetic ancestry, genetic admixture, and race and/or ethnicity?

Table H.1 -- Breakdown on self-reported race and/or ethnicity for CDRv9 participants with multi-omics data

Self-reported Race and/or Ethnicity	Array counts (%)	srWGS counts (%)	srWGS SV counts (%)	lrWGS counts (%)	RNA Seq counts (%)	Proteomics counts (%)
AI/AN	5,668 (1.02%)	5,422 (1.01%)	0**	246 (1.69%)	94 (1.05%)	124 (1.24%)
AI/AN, White	7,101 (1.28%)	6,849 (1.28%)	0**	243 (1.67%)	103 (1.15%)	134 (1.34%)
Asian	18,702 (3.38%)	18,087 (3.38%)	2,882 (2.99%)	2,388 (16.45%)	2,055 (22.88%)	2,319 (23.26%)
Asian, White	2,711 (0.49%)	2,637 (0.49%)	382 (0.40%)	153 (1.05%)	146 (1.63%)	161 (1.62%)
Black	90,011 (16.25%)	86,736 (16.19%)	22,232 (23.06%)	3,375 (23.24%)	1,119 (12.46%)	1,251 (12.55%)
Black, White	3,206 (0.58%)	3,103 (0.58%)	626 (0.65%)	250 (1.72%)	132 (1.47%)	138 (1.38%)
Hispanic	91,383 (16.50%)	88,473 (16.52%)	16,630 (17.25%)	3,384 (23.30%)	1,685 (18.76%)	1,846 (18.52%)
Hispanic, White	9,381 (1.69%)	9,112 (1.70%)	1,321 (1.37%)	303 (2.09%)	164 (1.83%)	176 (1.77%)
MENA	3,025 (0.55%)	2,903 (0.54%)	489 (0.51%)	485 (3.34%)	438 (4.88%)	476 (4.77%)
MENA, White	1,845 (0.33%)	1,792 (0.33%)	285 (0.30%)	74 (0.51%)	67 (0.75%)	70 (0.7%)
White	290,122 (52.37%)	280,790 (52.42%)	47,633 (49.41%)	2,272 (15.65%)	2,092 (23.30%)	2,278 (22.85%)
Remaining*	23,991 (4.33%)	23,147 (4.32%)	2,678 (2.78%)	1,078 (7.42%)	647 (7.2%)	728 (7.3%)
Skip	6,803 (1.23%)	6,611 (1.23%)	1,247 (1.29%)	270 (1.86%)	238 (2.65%)	268 (2.69%)
Total	553,949	535,662	96,405	14,521	8,980	9,969

Table Notes:

- Percentages may not add to 100 due to rounding.
- * The “Remaining” category refers to participants whose response did not match with the other categories shown in the table.
- ** The srWGS SV callset is based on a subset of participants from CDRv6, which did not include self-identified AI/AN participants.

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 123,171 high-quality (HQ) sites in the CDRv9 callset. These were HQ sites in both the HGDP+1kg training VCF and the *All of Us* CDRv9 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.

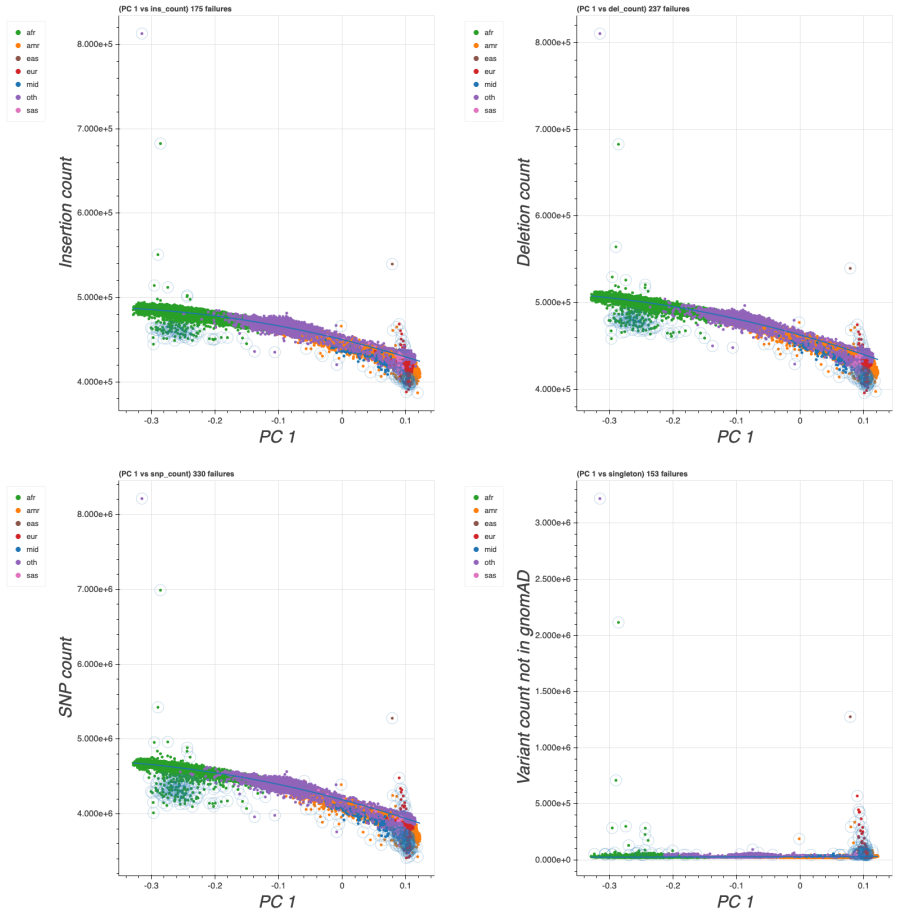
We will determine the maximal independent set [69] for related samples to minimize the number of samples that would need pruning. Using the HQ sites identified in Appendix I, researchers can remove first and second degree relatives.

We estimated 55,907 related pairs and 42,863 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).

Additionally, please see the FAQ "Are there any genomic duplicates in the dataset?"

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 1,010 flagged samples, since samples can get flagged for multiple criteria. Please see the next page for Figure K.1.



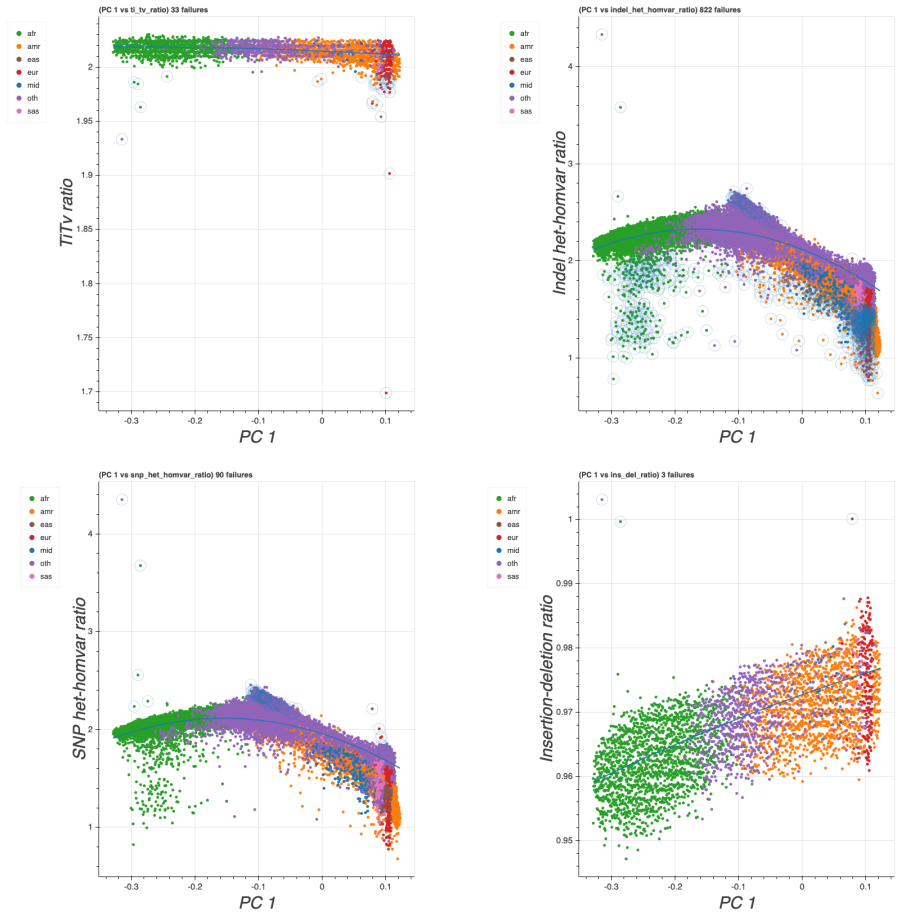


Figure K.1 -- Sample population outlier plots for eight metrics (see Population Outlier Flagging). Each metric (y-axis) is plotted against the first (of sixteen) principal components (x-axis). Outliers are identified by regressing out the principal components and determining if the residual is over 8 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 [70], 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) [71].

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 [72]
- Improved the depth-based genotyping method for very large CNVs to address an issue observed and manually fixed in the v7 srWGS SV callset so it is now fixed in all releases
- 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.

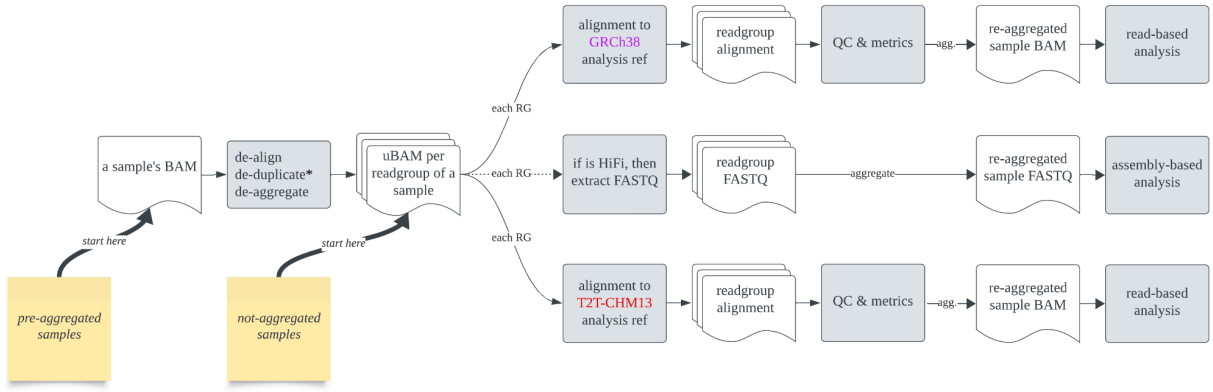


Figure N.1 -- The pre-processing and processing steps at the DRC for each lrWGS sample.

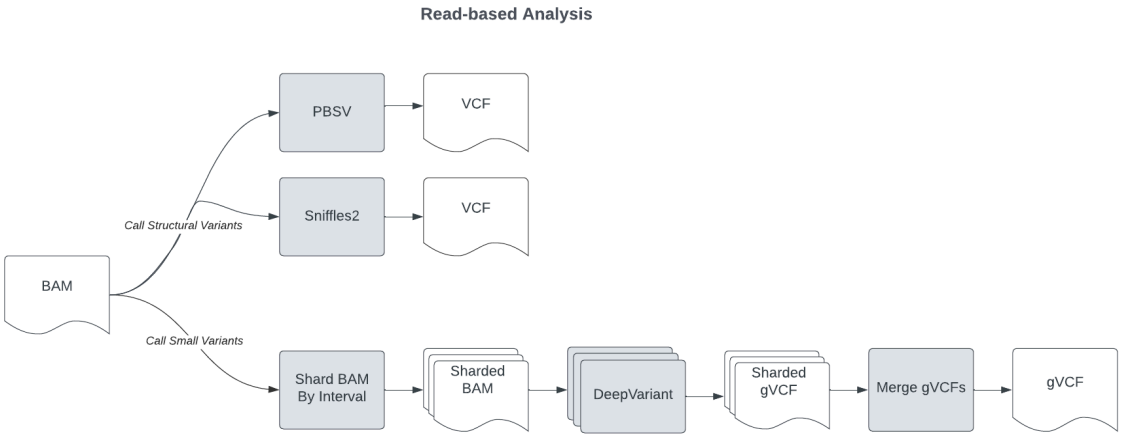


Figure N.2 -- The lrWGS variant calling steps, applied on both the grch38_noalt and the T2Tv2.0 references.

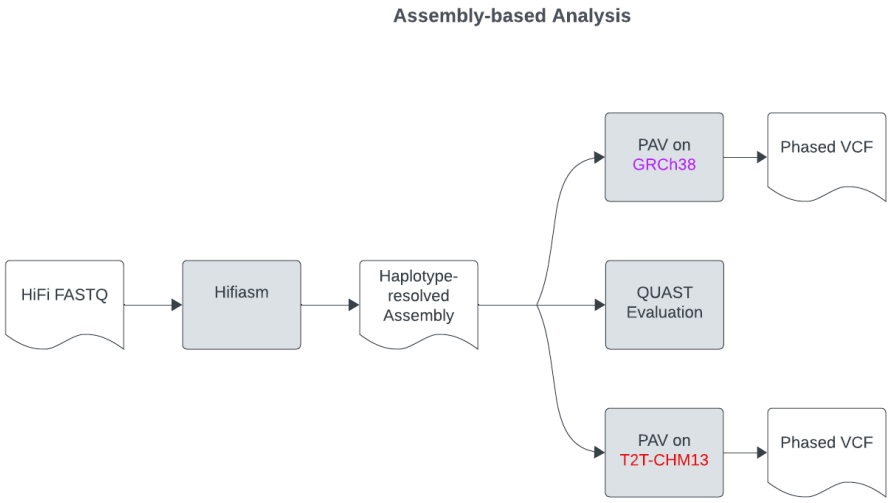


Figure N.3 -- lrWGS *de novo* assembly steps, for all cohorts with PacBio HiFi sequencing data.

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 and 0.20.0*	*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.
		
		* Note that some samples used an updated version of hifiasm. The samples with the updated version are available in the flagged samples list (See *de novo* assembly)
			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”.
			Based on participant self-reported sex assigned at birth data, described in Appendix C. If the answer was not male, the pipeline defaulted to the female sample command.
			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	quast \

		evaluation	`--no-icarus \`
			`--large \`
			`<assembly.fa>, [<assembly.fa>, ...]`

`nanoplot`	Git hash e0028d85ec9e61f8c96bea240ffca65b713e3385	Various alignment metrics collection	`NanoPlot \`
			`-c orangered \`
			`--N50 \`
			`--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 SV results

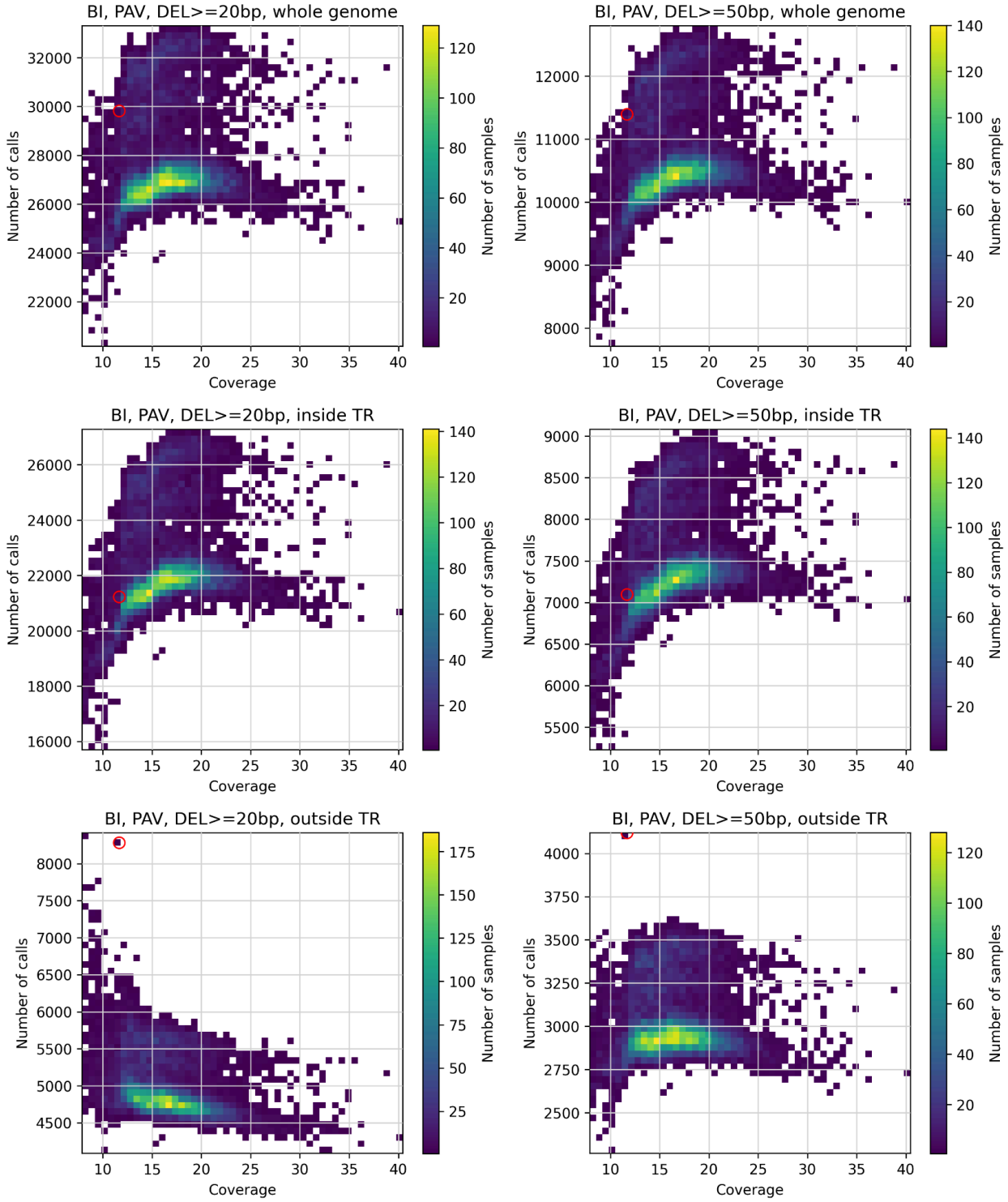


Figure P.1 -- BI_PacBio cohort PAV deletions. Outliers are circled in red and insertion plots show similar trends.

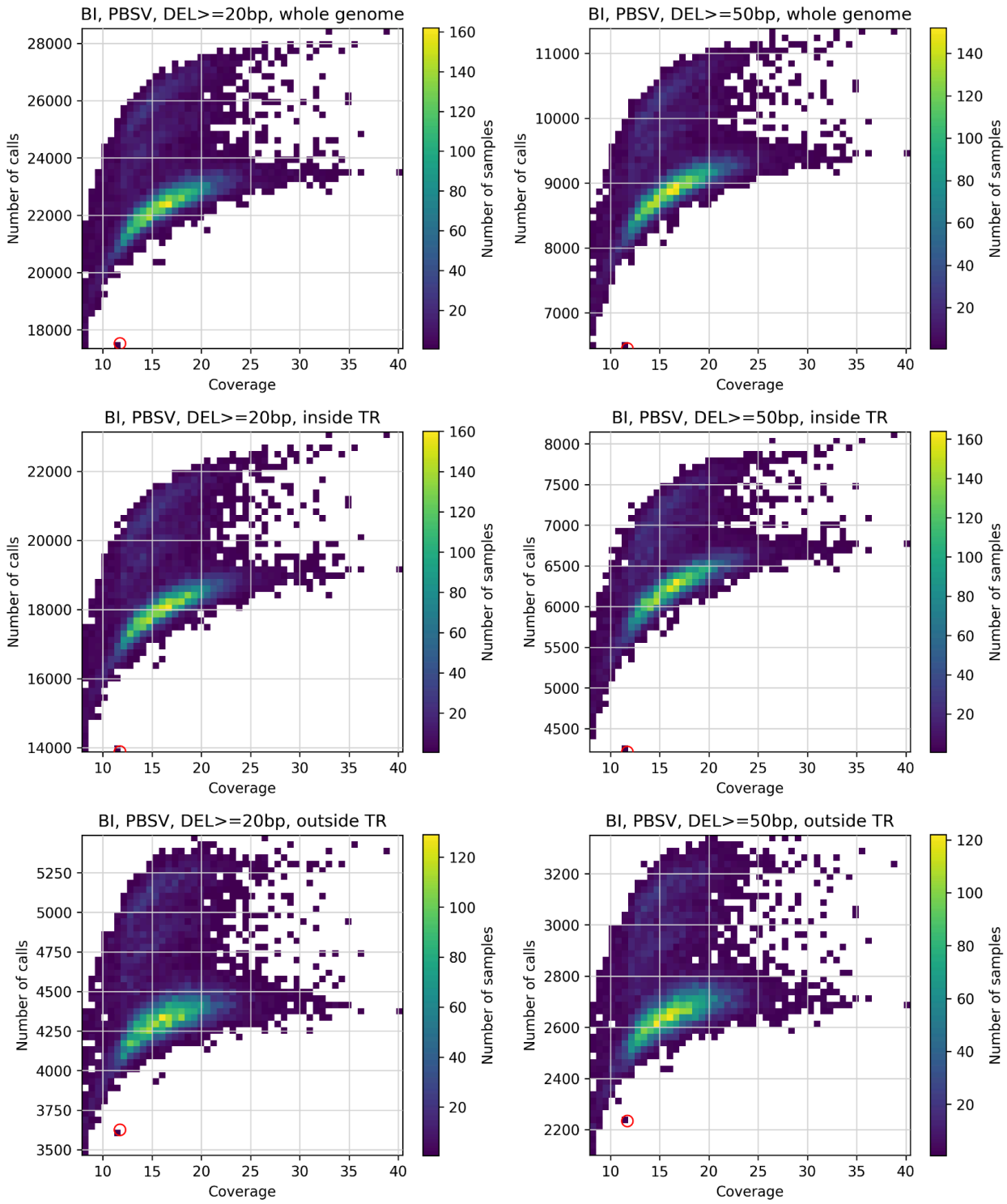


Figure P.2 -- BI_PacBio cohort, PBSV deletions. The outlier is circled in red.

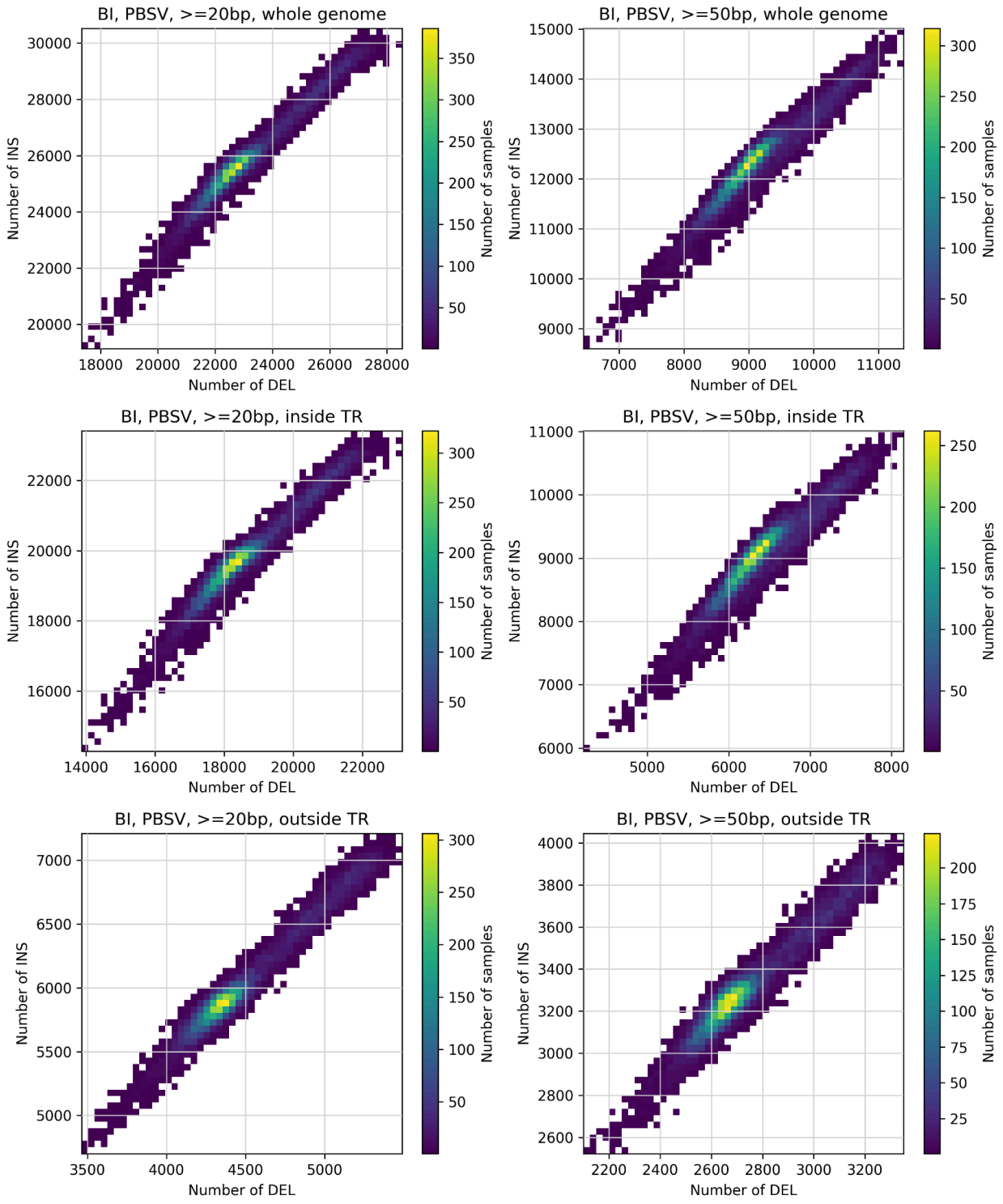


Figure P.3 -- BI_PacBio cohort, PBSV deletions and insertions.

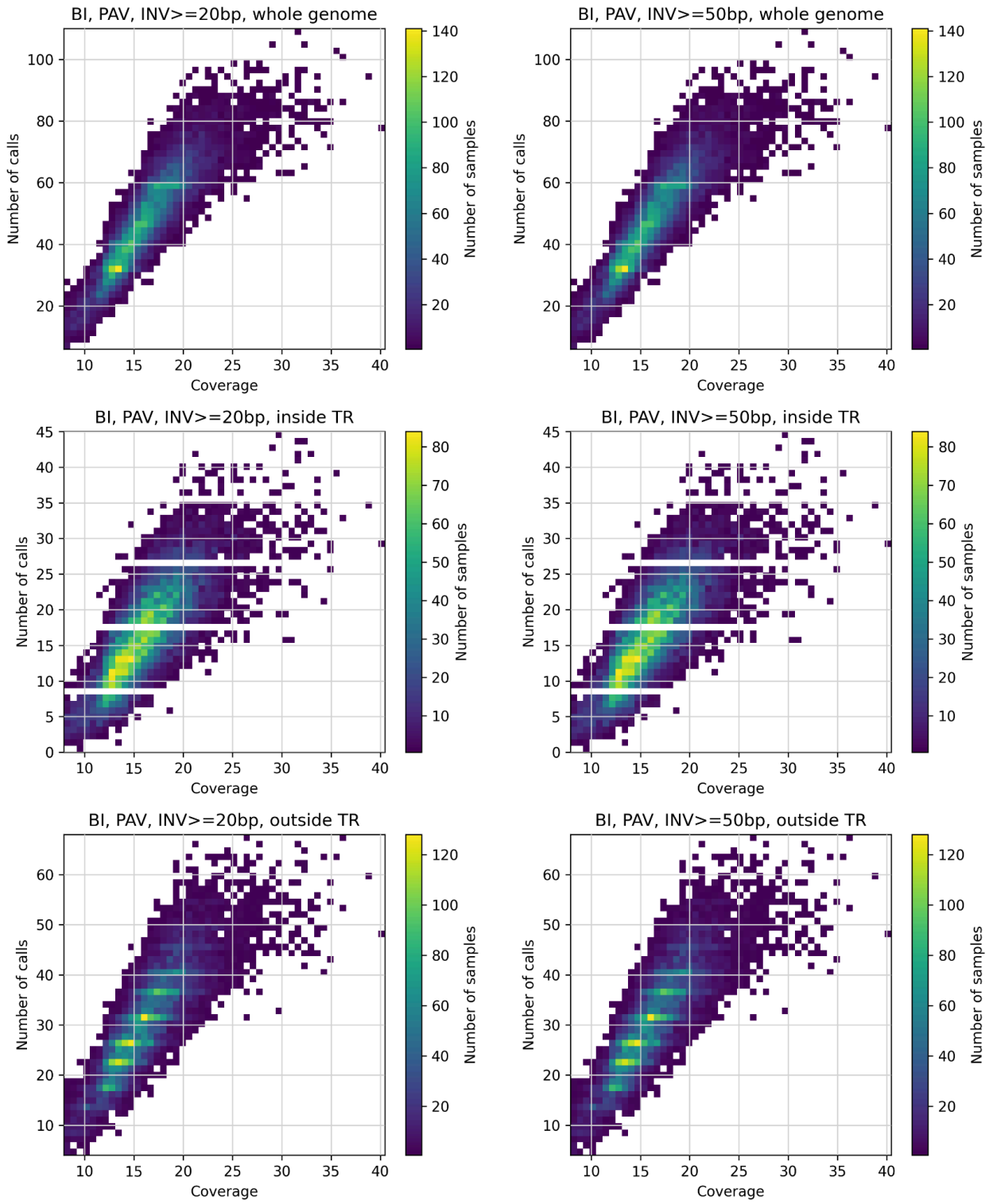


Figure P.4 -- BI_PacBio cohort, PAV inversions.

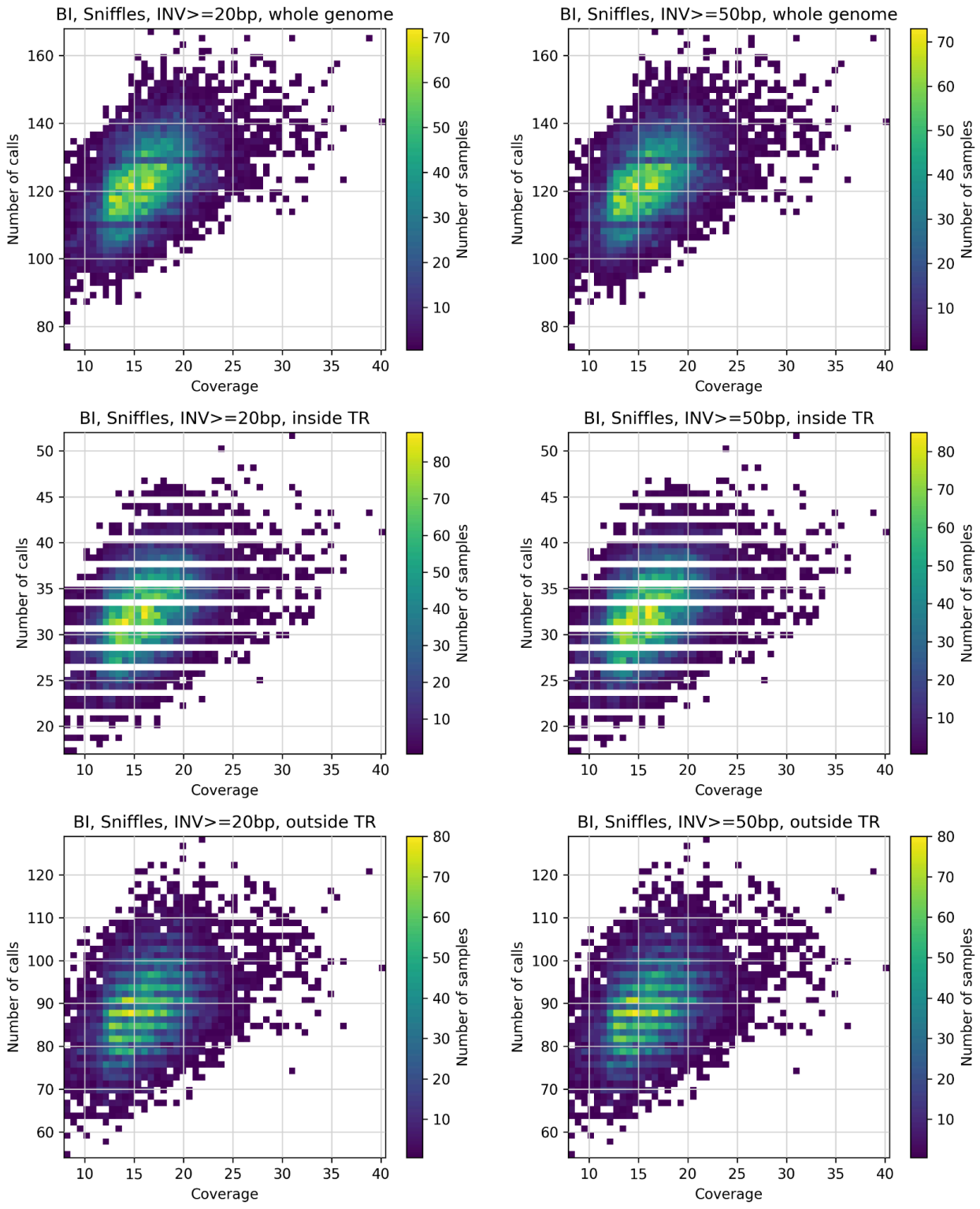


Figure P.5 -- BI_PacBio cohort, Sniffles inversions.

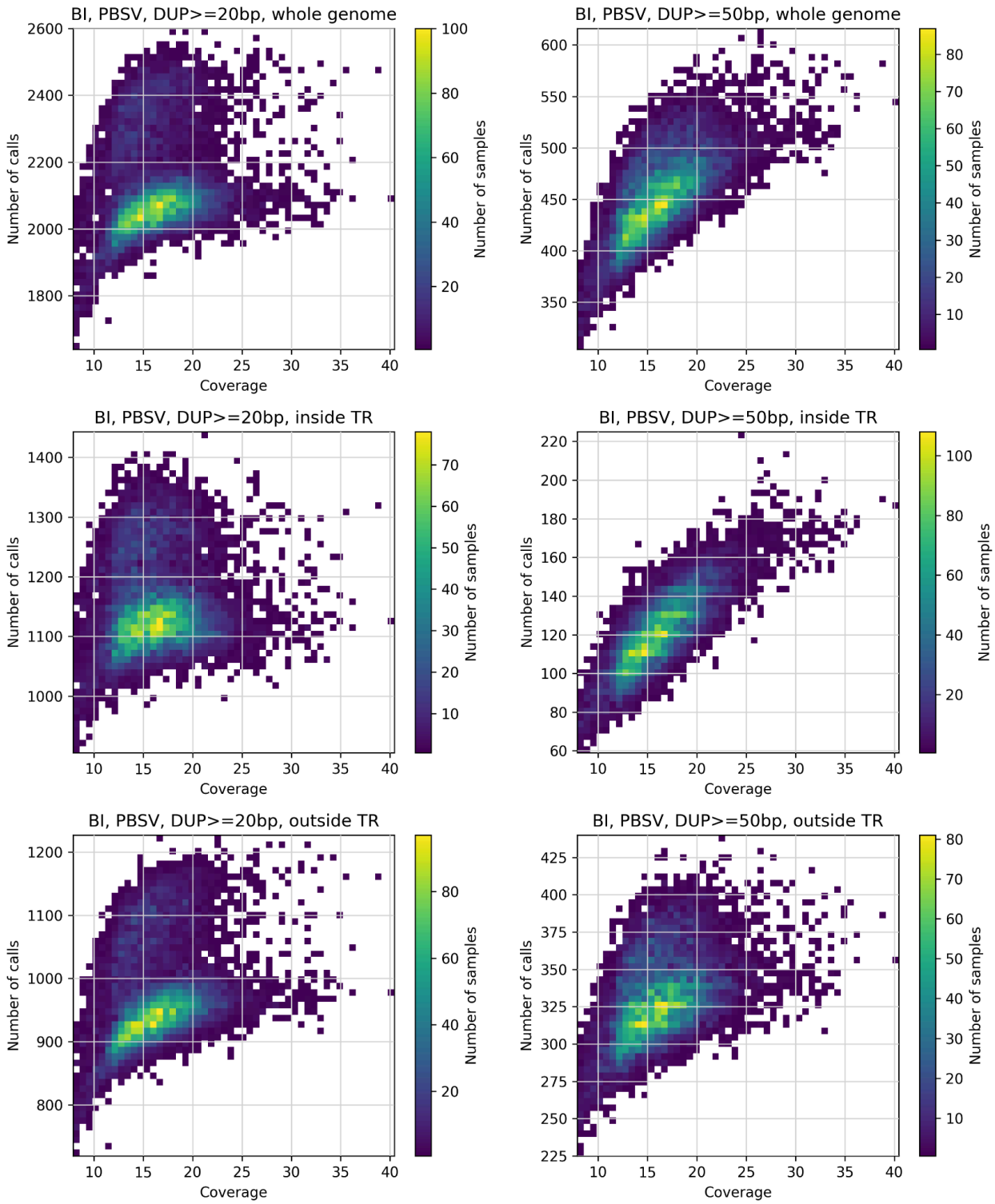


Figure P.6 -- BI_PacBio cohort, PBSV duplications.

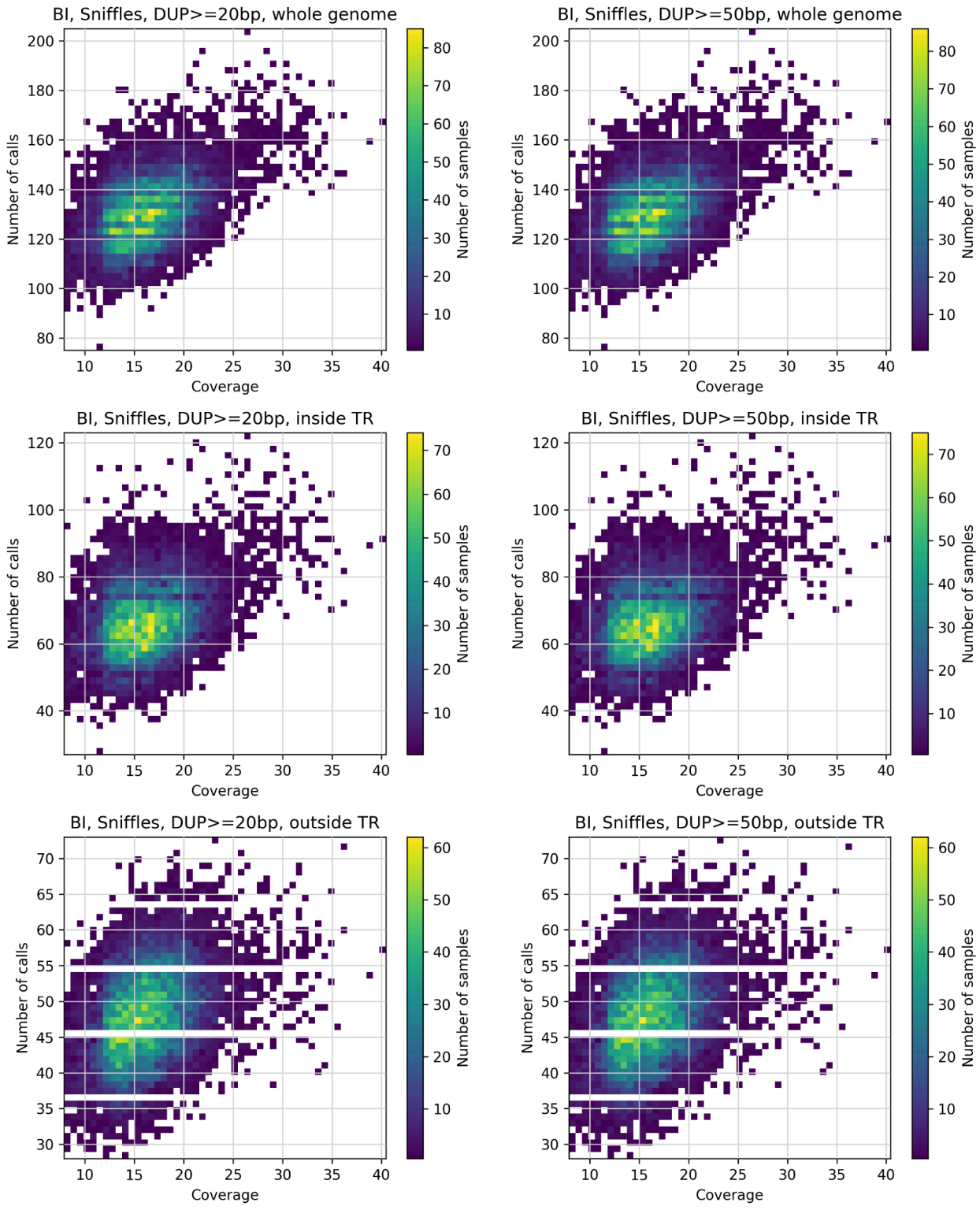


Figure P.7 -- BI_PacBio cohort, Sniffles duplications.

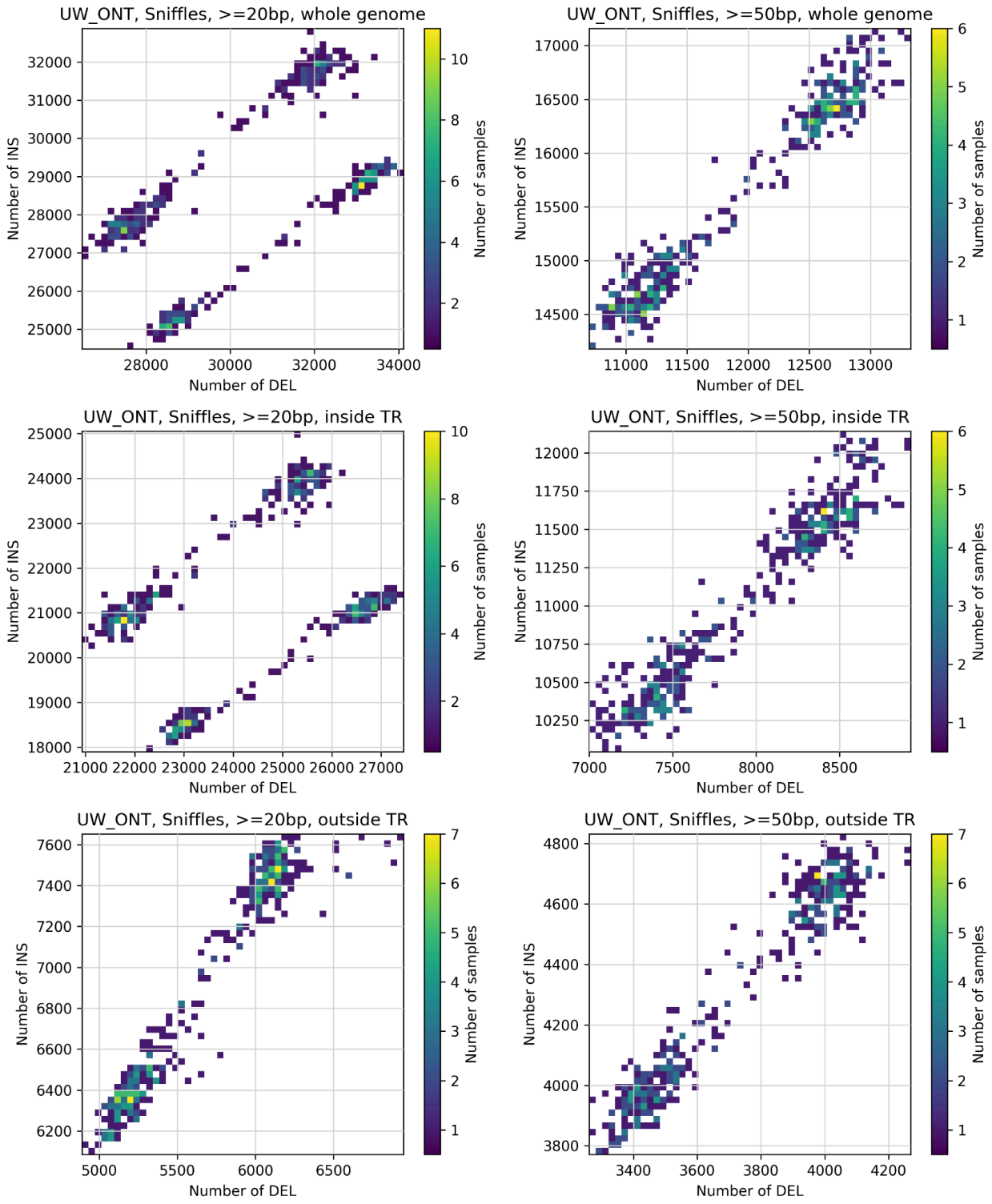


Figure P.8 -- UW cohort (ONT R9 and R10), Sniffles deletions and insertions.

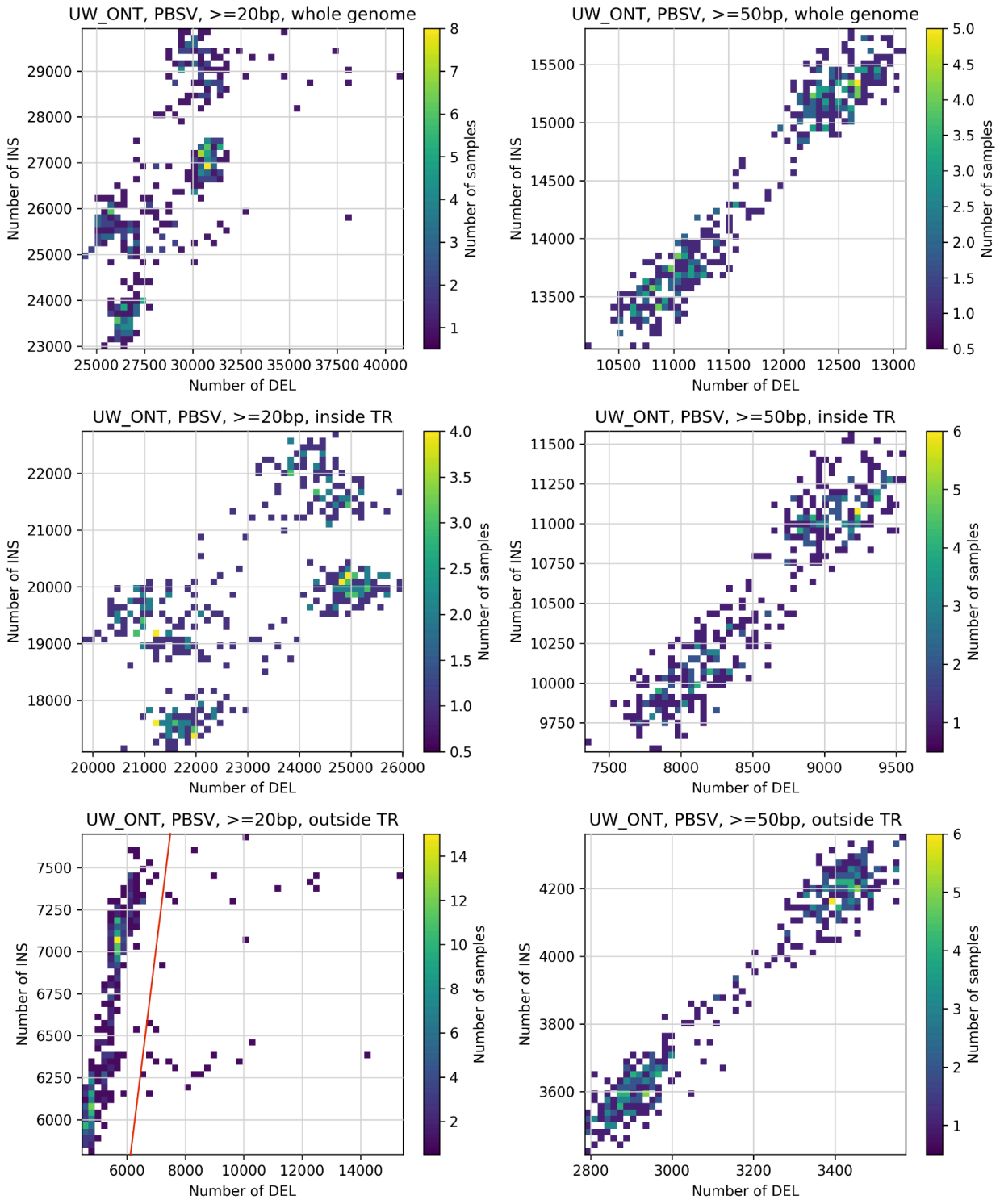


Figure P.9 -- UW cohort (ONT R9 and R10), PBSV deletions and insertions. The 28 samples to the right of the red line are flagged as outliers.

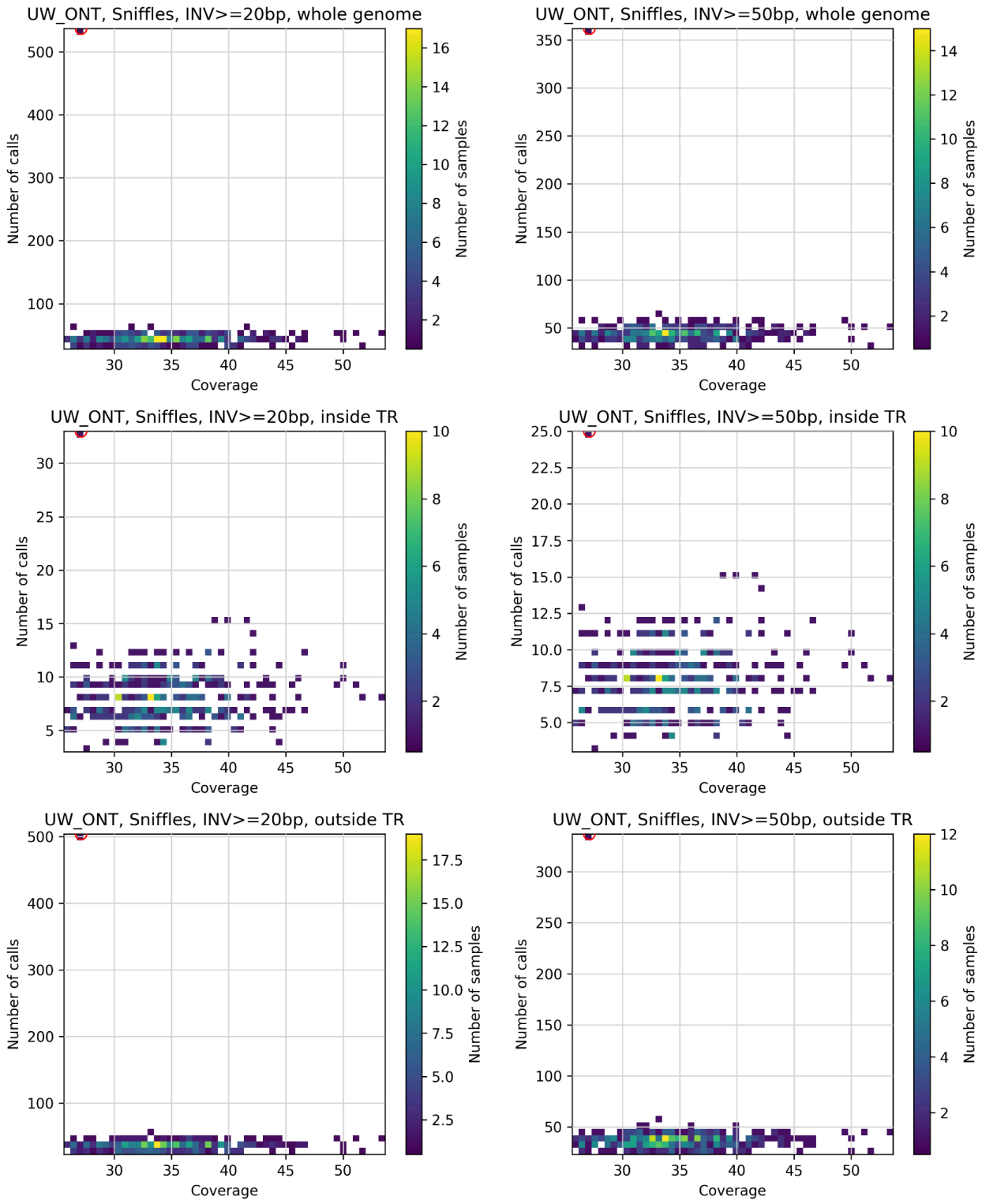


Figure P.10 -- UW cohort (ONT R9 and R10), Sniffles inversions. The outlier is circled red.

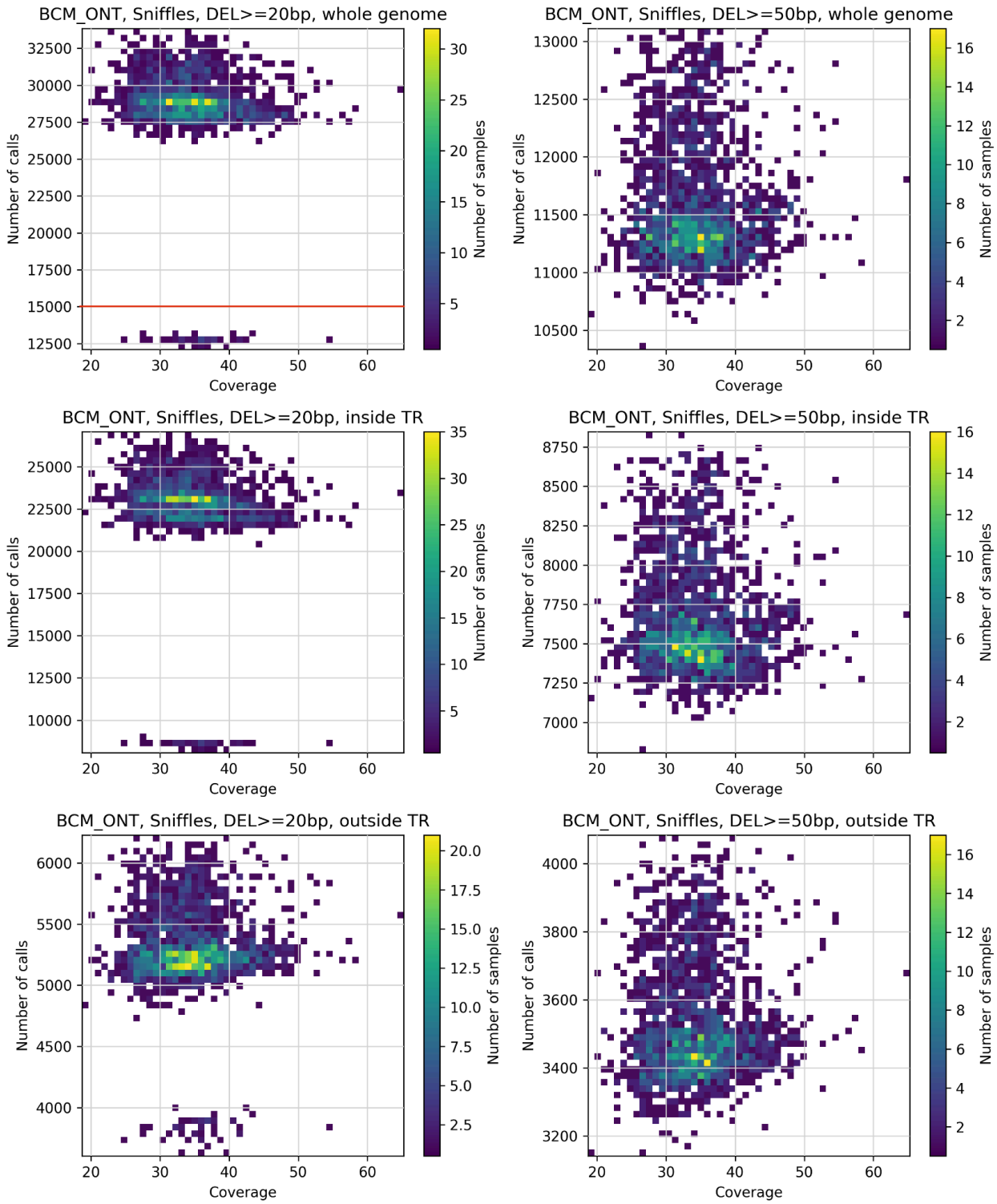


Figure P.11 -- BCM_ONT and JHU_ONT cohorts shown together, Sniffles deletions. The 50 samples below the red line (all BCM) are flagged as outliers.

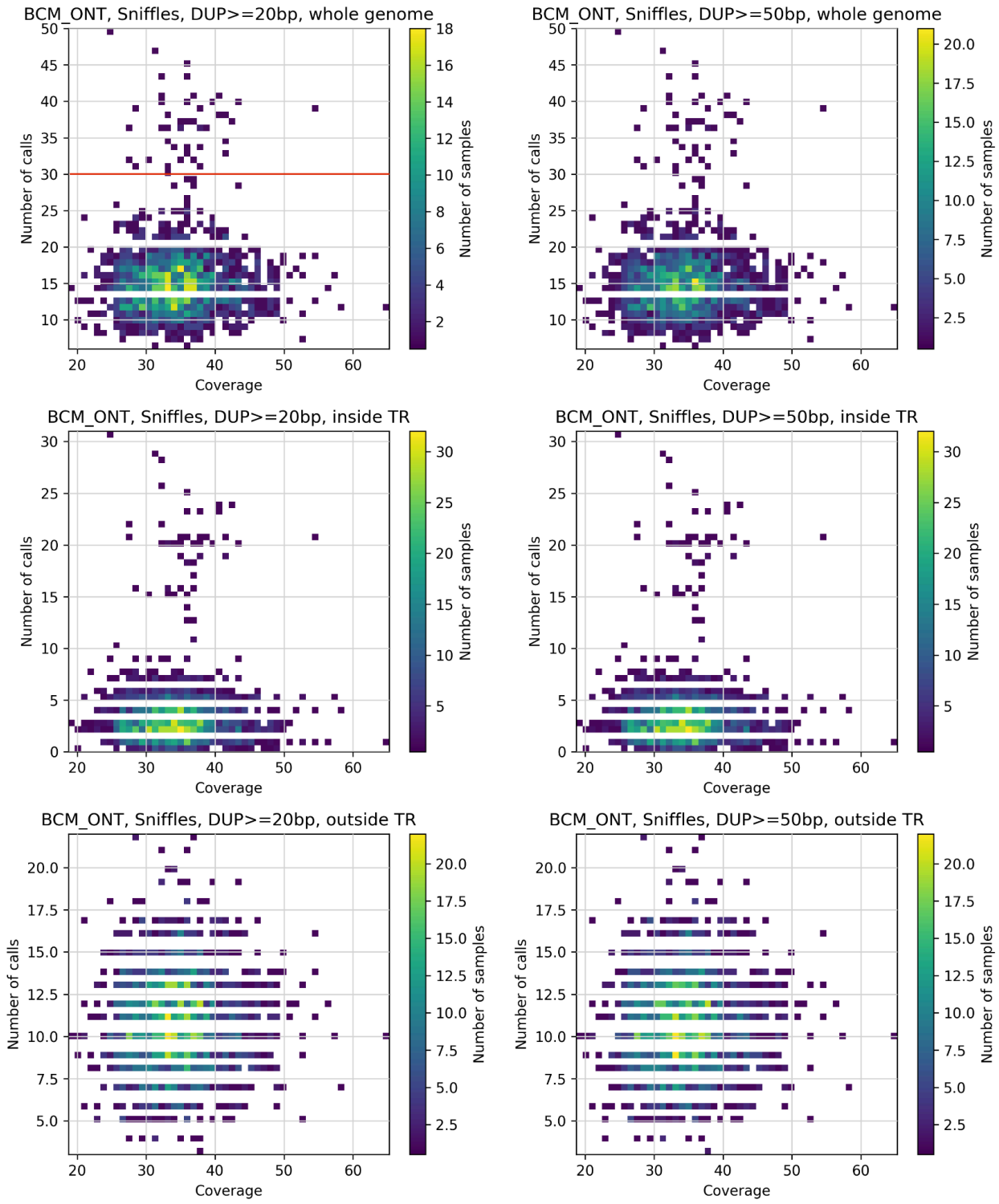


Figure P.12 -- BCM_ONT and JHU_ONT cohorts shown together, Sniffles duplications. The 45 samples above the red line (all BCM) are flagged as outliers.

Appendix Q: RNA sequencing processing overview

RNA alignment and QC

The genome centers established a consistent processing protocol by utilizing a modified version of the GTEx RNA-seq alignment pipeline available in the WARP repository. The RNA seq analysis workflow ensures that read extraction, alignment, and quantification are performed uniformly across all samples (Table Q.1).

To summarize the RNA alignment pipeline, reads were extracted from input CRAMs using the SamToFastq module and aligned to the reference genome using STAR v2.7.11b. WASP filtering was enabled and the pipeline used an hg38 reference that excludes ALT, HLA, and decoy and the GENCODE v48 GTF, linked here.

Table Q.1 -- Overview of RNA Alignment and QC

Processing Step	Software	Notes
Converting cram reads to FASTQ	Picard Tools SamToFastq v2.27.1	
Aligning reads with STAR	STAR v2.7.11b	Reference: Hg38 with GENCODE v48 GTF
Marking duplicates	Picard Tools MarkDuplicates v2.27.1	
Raw count quantification	RNASeQC-2 v2.4.3	Reference: Collapsed v48 GTF
Mapping to transcriptome and gene quantification	RSEM v1.3.1	Reference: Hg38 with v48 GTF

Raw read count quantification with RNA-SeQC-2

Raw read count quantification was used to determine the number of sequencing reads mapping to specific genomic features such as genes and exons. The RNA-SeQC2 tool analyzes the coordinate-sorted, duplicate-marked BAM files that are produced during the STAR alignment step. The analysis uses a collapsed version of the v48 GENCODE GTF, following GTEx instructions.

The results of RNA-SeQC2 processing are delivered as a collection of aggregated data files that provide a detailed overview of the transcriptomic profile for each of the 8,980 samples. These results are primarily organized into GCT files, specifically providing raw read counts for exons and genes, alongside normalized Gene TPM values for expression comparisons. Additionally, a comprehensive metrics text file is produced for the entire dataset, offering quantitative insights into the technical performance and integrity of the sequencing run.

Overall, the RNA seq data showed a median exonic rate of 57.2%, with reads mapping predominantly to exonic and intronic regions and only a small fraction mapping intergenically

(Figure Q.1). These results are consistent with good overall RNA-seq library quality. The distribution of mean 3′ bias was relatively narrow (median 0.468; IQR 0.442–0.508), indicating broadly consistent transcript-level read distribution and no evidence of widespread severe 3’ bias consistent with RNA degradation.

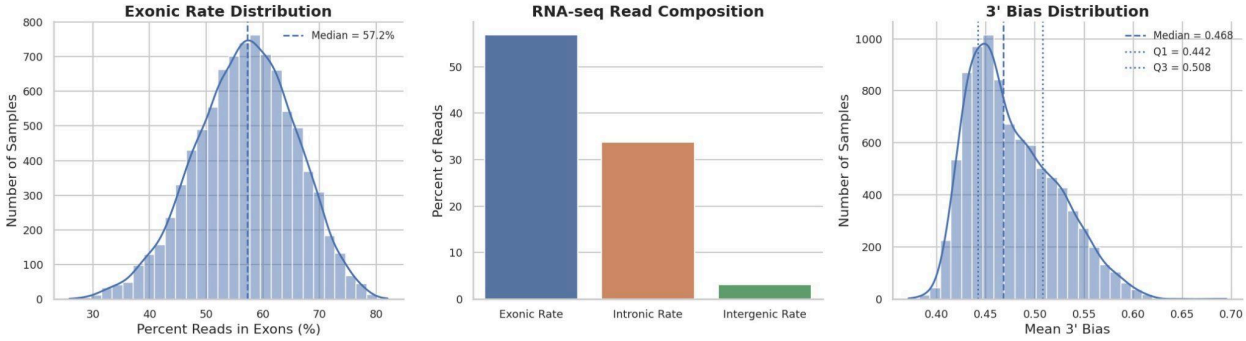


Figure Q.1 -- RNA-seq quality control metrics across samples. Distribution of exonic read fraction (left), overall read composition across exonic, intronic, and intergenic regions (middle), and mean 3′ bias (right).

Transcriptome alignment with RSEM

RSEM was used to align reads to the transcriptome using a hg38 reference with GENCODE v48 reference for isoform level quantification.

Expression Quantitative Trait Loci (eQTL) pipeline

Sample lists were created for each genetic ancestry group and for the combined dataset. Prior to downstream analysis, the distribution of RNA Quality Scores was assessed to ensure no genetic ancestry-related bias (Figure Q.2). RQS is similarly distributed across ancestry groups, indicating low ancestry-related RNA quality bias. For each subgroup, the RNA-SeQC2 gene reads were used to prepare a bed file for downstream analysis.

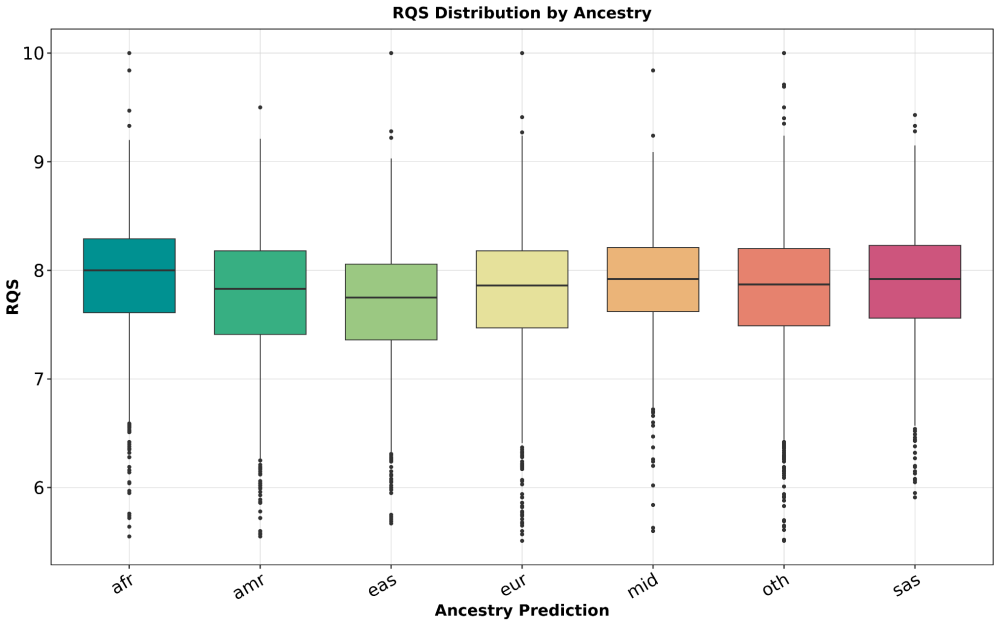


Figure Q.2 -- RNA Quality Scores by genetic ancestry. . Ancestry groups for eQTL analysis.

Expression normalization

Gene-level RNA-seq count data were provided as input in GCT format along with a corresponding gene annotation file (GTF) and the predefined list of samples to include. Transcription start site (TSS) coordinates for each gene were extracted from the GTF based on gene strand.

Raw count data were transposed to a sample-by-gene matrix and filtered to retain genes with counts greater than 6 in at least 20% of samples. Filtered counts were then normalized using the trimmed mean of M-values (TMM) method implemented in edgeR, followed by conversion to counts per million (CPM).

To stabilize variance and approximate normality, expression values for each gene were rank-based inverse normal transformed across samples. The normalized expression matrix was then merged with gene-level TSS coordinates to generate a BED-formatted file containing genomic positions and normalized expression values for each gene across samples.

The final output consists of a gzipped BED file (.expression.bed.gz) with columns for chromosome, TSS-based genomic coordinates, gene identifier, and normalized expression values per sample, suitable for downstream eQTL analysis.

Phenotype PC calculation

The covariate preparation step aggregates genotype and phenotype PCs with other technical factors to generate the structured matrices required for robust association mapping.

Phenotype PCs for covariates were calculated using principal component analysis (PCA) on the sample-by-gene expression matrix using PCAtools v2.22.4. The number of components to retain was determined using the Gavish–Donoho method for optimal hard thresholding, which estimates the effective dimensionality of the data based on the singular value spectrum. The selected number of PCs was then extracted from the rotated PCA matrix, with each component representing a major axis of variation in gene expression across samples.

The resulting phenotype PCs were output as a tab-delimited file (_phenotype_PCs.tsv), with rows corresponding to samples and columns representing the retained principal components, for use as covariates in downstream QTL analyses.

Genotype PC Calculation

Joint srWGS variants were filtered to biallelic variants, samples with missing genotypes and allele counts were removed, an AC threshold of 5 and a call rate of 95% was applied, and the VCFs were converted to PLINK for downstream QTL analysis. Genetic PCs were calculated using SNPRelate [73] after pruning with SNPs with an LD threshold = 0.2 and MAF = 0.01.

QTL Analysis with TensorQTL

TensorQTL cis permutations are performed to identify genetic associations. The phenotype and genotype PCs from the previous steps were merged into a single covariates file to use for eQTL analysis with TensorQTL. The exact version can be found in the public docker: gcr.io/broad-cga-francois-gtex/tensorqtl@sha256:f6efb9e592eb32c46cb75070be2769b34381d60cbb2709d2885771324abfe32a. TensorQTLwas run in the cis permutations mode individually for each genetic ancestry group and the whole combined cohort, using an FDR = 1 parameter.

TensorQTL was performed in the cis nominal mode. In cis nominal mode, linear regression was applied to test associations between variants and phenotypes within a 1 Mb cis window, adjusting for genetic principal components (PCs) as covariates. Default parameters were used. This mode generates nominal p-values and effect sizes for all variant–phenotype pairs, enabling downstream analyses such as fine-mapping and visualization.

As seen in Figure XX, the density of significant eQTL variants (q-value < 0.05) is enriched near the transcription start sites, consistent with expected cis-regulatory effects (Figure Q.3, panel A). We also see a typical inverse relationship between minor allele frequency (MAF) and the effect size, demonstrated by the variant-gene pairs in Figure Q.3, panel B.

There is an expected allele frequency spectrum of detected associations (Figure Q.3, panel C).

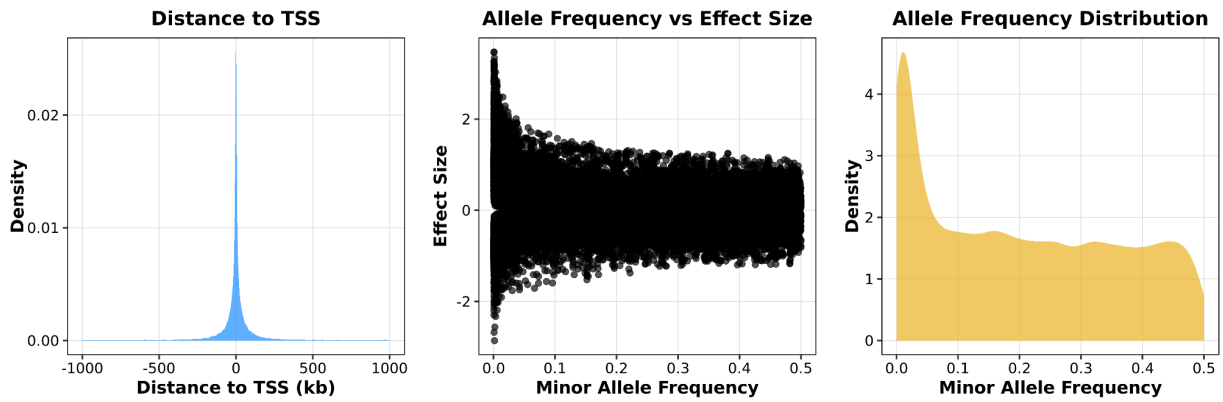


Figure Q.3 -- Summary of cis-eQTL characteristics among significant associations. (A) Distribution of distances between significant eQTL variants (q-value < 0.05) and transcription start sites (TSS), shown in kilobases (kb). (B) Relationship between minor allele frequency (MAF) and effect size (slope) for significant eQTLs. (C) Distribution of MAF for significant eQTLs.

Finemapping

Following cis permutations, the results are filtered to a recalculated False Discovery Rate (FDR) of 0.05.

To further identify specific genetic causal variants, Bayesian finemapping with Sum of Single Effects model with susieR was applied to significant phenotypes with an FDR > 0.05. For each phenotype, variants were extracted within a cis-window of ±1 Mb around the phenotype position. Expression values were transformed using a rank-based inverse normal transformation. Both phenotype and genotype matrices were residualized with respect to the covariate matrix, including an intercept term, using linear projection. Variants with no alternate alleles were removed and missing genotype dosages were imputed to the variant mean. The susieR software was run with a maximum of 10 single-effect components (L = 10), with residual variance and prior variance estimated from the data (estimate_residual_variance = TRUE, estimate_prior_variance = TRUE), scaled_prior_variance = 0.1, and univariate z-scores computed (compute_univariate_zscore = TRUE). Credible set purity was assessed using a minimum absolute correlation threshold of 0.5 (min_abs_corr = 0.5). Downstream reporting retained non-overlapping credible sets and excluded low-purity credible sets with minimum absolute correlation < 0.5. After susieR finemapping, the AF was calculated for the eQTLs and the all the finemapping data was aggregated into an annotated TSV.

Splicing Quantitative Trait Loci (sQTL)

Unless noted below, the process for performing sQTL was identical to the methods described for eQTLs. Splice junctions were extracted from duplicate-marked RNA-seq BAM files using a LeafCutter-based analysis. For each sample, we first restricted reads to high-confidence alignments by retaining only uniquely mapped reads (mapping quality = 255) and excluding

reads flagged by WASP to mitigate allelic mapping bias. Filtered reads were written to a temporary BAM file and indexed using samtools v1.19.2.

Splice junctions were then identified using regtools v0.5.2 (junctions extract), requiring a minimum anchor length of 8 bp, a minimum intron length of 50 bp, and a maximum intron length of 500 kb. Strand specificity was specified per dataset as applicable. The resulting junction counts were output in compressed text format for downstream clustering and splicing quantification with LeafCutter.

Intron clustering and splicing phenotype generation

Per-sample splice junction files generated with regtools were jointly clustered across samples using LeafCutter v0.2.9. Junctions were grouped into intron clusters using the leafcutter_cluster_regtools.py workflow with a minimum of 30 reads per cluster, a minimum junction-to-cluster read ratio of 0.001, and a maximum intron length of 500 kb. This produced per-intron count tables across all samples for downstream splicing quantification.

Intron clusters were then mapped to genes by intersecting intron boundaries with exon annotations. Under a strict assignment scheme, clusters were linked only to genes for which both splice sites matched annotated exon boundaries. Introns with insufficient representation or low information content were removed prior to QTL analysis, including introns with zero usage in more than 50% of samples, low variability across individuals, or localization to chromosome Y.

Filtered intron usage counts were normalized for QTL analysis using the LeafCutter phenotype preparation workflow, including quantile normalization and generation of chromosome-level phenotype files that were merged into a genome-wide BED file. Splicing phenotypes were then reassigned from sample IDs to participant IDs and annotated with transcription start site coordinates from the collapsed gene annotation to support cis-sQTL mapping. A phenotype group file linking introns to their corresponding genes was also generated for downstream association testing.

Phenotype PC Calculation

Splicing phenotypes were derived from LeafCutter intron usage BED files and subset to the analysis cohort. Intron usage values were normalized using rank-based inverse normal transformation across samples. The normalized matrix was combined with genomic coordinates and phenotype identifiers, sorted by genomic position and gene annotation, and written as a compressed BED file for downstream sQTL mapping. Phenotype PCs were calculated from the BED using the methods described in the eQTL section.

Genotype PC Calculation

Genotype PCs were created using the methods described in the eQTL Genotype PC calculation section.

QTL Analysis with TensortQTL

TensortQTL cis permutations were performed on the normalized BED with an FDR = 1. Merged PCs were used as covariates and phenotype groups were provided in a separate phenotypes group file.

As seen in Figure Q.4, the density of significant sQTL variants (q-value < 0.05) is enriched near the TSS, consistent with expected cis-regulatory effects. (Figure Q.4, panel A). We also see a typical inverse relationship between minor allele frequency (MAF) and the effect size, demonstrated by the variant-gene pairs in Figure Q.4, panel B.

There is an expected allele frequency spectrum of detected associations (Figure Q.4, panel C).

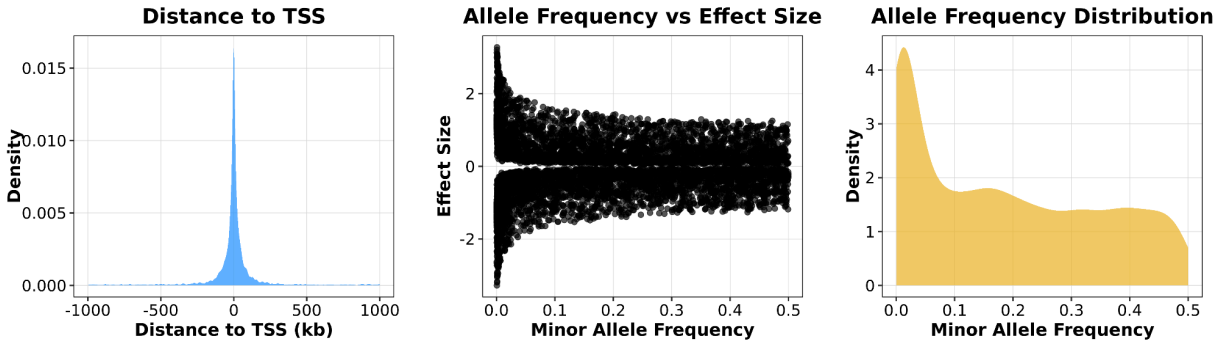


Figure Q.4 -- Summary of cis-sQTL characteristics among significant associations. (A) Distribution of distances between significant sQTL variants (q-value < 0.05) and transcription start sites (TSS), shown in kilobases (kb). (B) Relationship between minor allele frequency (MAF) and effect size (slope) for significant sQTLs. (C) Distribution of MAF for significant sQTLs.

Finemapped sQTLs

Finemapping with SusieR to generate finemapped sQTLs was performed as described in the eQTL section.

Appendix R: Proteomics pipeline processing overview

The proteomics processing workflow is designed to generate high-resolution protein expression profiles from 10,170 blood plasma samples using the Olink Explore HT platform. These samples were systematically organized into 61 individual Olink projects, each consisting of two plates, and were processed across two primary batches. To address potential batch effects and ensure longitudinal consistency, the workflow incorporated approximately 127 replicate samples, including 25 specific bridge samples designated for software-based bridging analysis. Sample identification is maintained through a unique SampleID that combines the research ID with the Olink PlateID, allowing for clear differentiation of replicates within the dataset.

As described in Table 5, the primary output of this processing pipeline is the Normalized Protein Expression (NPX) value, which is provided in both Parquet and TSV formats for each project. These files are accompanied by an Olink Analysis Report PDF that documents the project-specific metadata, sample counts, and software versions utilized during the run. For researchers conducting protein quantitative trait loci (pQTL) analysis, the data is further integrated with genomic information and provided in protein quantitative trait loci (pQTL) data.

The data was normalized using a reference-median based intensity normalization, technical replicates and outliers were removed, rank-based inverse normal transformation was performed, and genetic and phenotypic principal components that can be used as covariates for downstream cis qTL analysis with TensorQTL were calculated. We used SusieR frameworks for Bayesian fine-mapping and association testing across both cis regions of the genome.

Table R.1 -- Proteomics processing outline

Step	Method	Notes
Olink reference median normalization	olink_normalization(reference median)	Used to remove batch effects across plates
Technical replicate removal	Random selection	
Outlier removal	Connectivity and Z-score score calculation; outliers removed with X score below -3.	
Normalization	Rank Norm from RNOmni package	Rank-based inverse normal transformation
Phenotype PCs	PCATools	GavishDonoho
Genotype PCs	SNPRelate after pruning with SNPs with an LD threshold = 0.2 and MAF = 0.01	
QTL cis permutations	Tensorqtl --mode cis	

Reference median-based normalization

To remove technical artifacts, Olink NPX values were normalized using a reference median-based approach. Control samples were excluded prior to normalization. A single plate was selected as a reference, and assay-specific median NPX values were calculated from this plate. NPX values from all other plates were then adjusted to these reference medians using the Olink normalization procedure, harmonizing measurements across plates while preserving relative biological variation.

The batch effects before and after reference median normalization are observed in Figure R.1. Batch effects in Olink proteomic data were assessed by quantifying the proportion of variance in protein expression attributable to plate using the plate-control normalized values (Figure R.1, panel A) and reference median based normalized values (Figure R.1, panel B). For each protein, a linear model (NPX ~ PlateID) was fit across samples after restricting to assay measurements and excluding control samples. The coefficient of determination (R^2) from each model was used as a measure of plate-associated variance, and the distribution of R^2 values across proteins is shown. The model of batch effects (Panel B) and the distribution of NPX values after reference median normalization (Figure R.1, panel C) indicate the reference median-based normalization removed the batch effects.

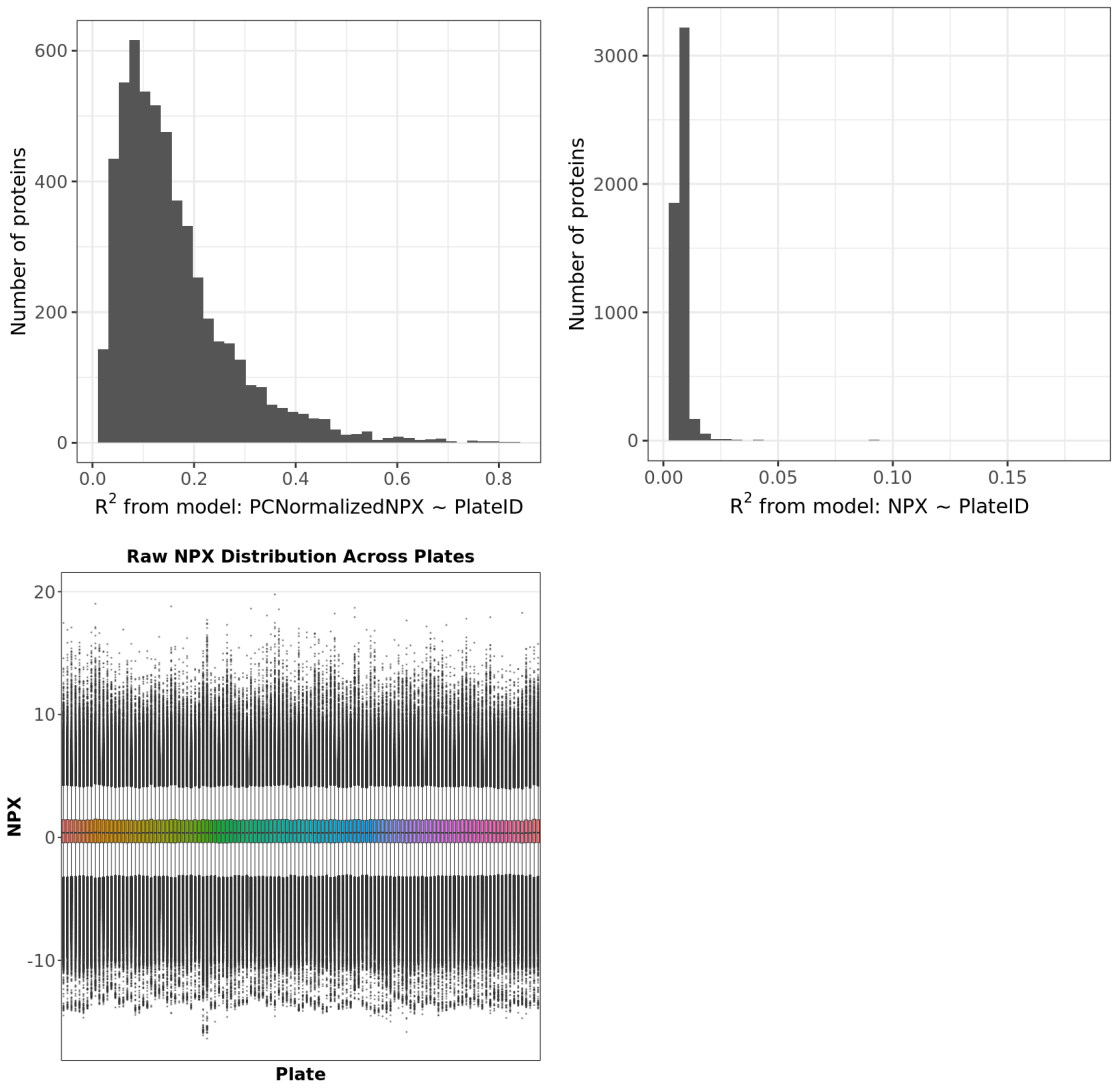


Figure R.1 -- Batch Effects before and after reference median-based normalization. (A) Batch effects by plate prior to reference median normalization. (B) Batch effects by plate after reference median normalization. (C) Distribution of NPX values after reference median normalization.

Replicate removal

To avoid pseudoreplication, samples with technical replicates were reduced to a single observation per participant. For each individual with multiple samples, one sample was randomly selected and retained, while all others were removed. Participants without replicates were retained in full.

Outlier removal, normalization and phenotype PC generation

Sample-level proteomic quality control was performed using a network connectivity approach applied to the reference median normalized Olink NPX matrix. Briefly, pairwise biweight midcorrelations were calculated across all samples, and the resulting adjacency matrix was

used to estimate sample connectivity. Connectivity values were converted to z scores, and samples with connectivity z scores less than -3 were designated as outliers and removed prior to downstream analysis. Following outlier exclusion, samples were stratified by ancestry and protein measurements were rank-based inverse normal transformed within each group. These transformed data were used to generate ancestry-specific phenotype BED files for QTL analysis. Principal component analysis was then performed on the transformed proteomic matrix, and the number of retained components was determined using the Gavish-Donoho criterion. The selected proteomic principal components were exported as covariates for downstream pQTL analyses.

Genotype PC calculation, cis QTL, and finemapping

Genotype PCs, cis qtls, and finemapping with susieR were calculated as described previously for eQTL and sQTL, to generate proteomics cis QTL. Proteomics fine-mapped pQTLs will be released in a future dataset release.

Appendix S: Saliva and blood batch effect analysis

Introduction

Beginning in the CDRv8 genomic data release, we included both saliva and blood samples whereas in previous data releases, we only included blood samples. We performed an analysis to determine blood and saliva batch effects.

We did not find appreciable batch effects between blood and saliva samples in SNP variants. However, we identified significant differences in indel counts, with a mean genome-wide difference of less than 0.4%, largely attributable to low complexity regions (LCR). Please note that variant calls are less reliable at LCR sites in all variant datasets.

To address this issue, we suggest that researchers include the sample source (blood or saliva) as a factor in their analysis when studying indels or regions where we see appreciable batch effects. We provide the sample source, for all srWGS samples, in the genomics metrics tsv file. The path for the file is available in the Data Dictionary [1].

Methods

We measured the batch effect size using Cohen's d [74], which is a measure of the differences of the means between two groups. We defined an "appreciable batch effect" as any value of d greater than 0.5 in magnitude, which is the convention for a medium-size effect [74].

We sampled 167,861,000 sites from 2181 blood and 2217 saliva samples from the CDRv6 srWGS SNP and indel callset. We compared the blood and saliva samples using multiple variant metrics across various genome regions. All genome centers are represented in our analysis.

We examined four variant metrics: SNP count, indel count, SNP transition to transversion (Ti/Tv) ratio, and Insertion/Deletion (Ins/Del) ratio. The genome regions we compared are the whole genome, the whole genome excluding low complexity regions (WG excl. LC), the Genome in a Bottle high confidence calling regions (GiaB), high GC content (GC > 0.85), low GC content (GC < 0.25), American College of Medical Genetics (ACMG)'s list of 59 genes (ACMG59), low mappability, segmental duplication regions (SegDupe), and tandem repeat regions (Table S.1).

Table S.1 -- Regions used in the batch effect analysis

Region	Description
Whole genome	All variant calls that are in the CDRv6 genomic dataset
Whole genome excluding low complexity regions (WG excl. LC)	Low complexity regions removed from the whole genome
Genome in a Bottle (GiaB) high	National Institute of Standards and Technology (NIST)

confidence regions	benchmark high confidence variant call regions provided by the Genome in a Bottle consortium
GC (GC > 0.85)	High guanine (G) and cytosine (C) content in a genomic region
AT-rich genome region (GC < 0.25)	Region with low GC content
American College of Medical Genetics (ACMG)'s list of 59 genes (ACMG59)	A list of 59 genes that the ACMG recommends be actively searched for pathogenic or likely pathogenic variants during clinical genomic sequencing
Low mappability	Low mappability regions are regions of the genome where sequences cannot be mapped with as high confidence as other regions of the genome. These regions are generally less reliable.
Segmental duplication regions (SegDupe)	Regions that contain sequences with copies in other areas of a genome. These regions are prone to mapping issues.
Tandem repeat regions	Regions that contain sequences with adjacent copies. These regions are prone to mapping issues.

We focused our study on three *All of Us* genetic ancestry groups: 1KGP-HGDP-AFR-like, 1KGP-HGDP-AMR-like, and 1KGP-HGDP-EUR-like. We did not study the other genetic ancestry groups, including 1KGP-HGDP-EAS-like, 1KGP-HGDP-MID-like, and 1KGP-HGDP-SAS-like, due to having too few samples. We also did not include the remaining samples (OTH), because they are too heterogeneous, which would skew to smaller effect sizes. The *All of Us* genetic ancestry groups are inferred by measuring the relative genetic similarity of each participant to global reference populations from harmonized continental metadata labels from the Human Genome Diversity Project (HGDP) and 1000 Genomes Project training data, as described in Appendix G.

Results

As a control, we randomized the batch labels to see what differences might appear by chance alone. In this test, the largest observed difference was d = 0.35 across all genomic regions, which serves as a baseline for interpreting batch effects in the real data.

We summarized the metrics where we found appreciable batch effects across the tested genetic ancestry groups in Table S.2.

The SNP count and SNP ti/tv ratio metrics across the entire genome do not indicate appreciable batch effects between blood and saliva samples. We do detect differences in the indel count and ins/del ratios in some regions indicating appreciable batch effects. The regions where we saw the differences are the whole genome, GC < 0.25, low mappability, SegDupe, and tandem repeat regions. The regions with no significant differences indicating batch effects are the WG excl. LC, GiaB, GC > 0.85, and ACMG59.

Our results indicate that the batch effects within the whole genome region are driven by other regions, specifically the low complexity regions, as when they are removed, the WG excl. LC demonstrates no significant batch effects.

Table S.2 -- Presence of appreciable batch effects across each region and metric

Region	SNP count	Indel count	SNP Ti/Tv ratio	Ins/Del ratio
Whole genome	No	Yes	No	Yes
WG excl. LC	No	No	No	No
GiaB	No	No	No	No
GC > 0.85	No	No	No	No
GC < 0.25	No	Yes	No	Yes
ACMG59	No	No	No	No
Low mappability	No	Yes	No	Yes
SegDupe	No	Yes	No	Yes
Tandem repeat regions	No	Yes	No	Yes

Whole genome results

The following sections contain the results for each metric for the whole genome region.

SNP count

In examining the SNP counts for the whole genome, we see that across the genetic ancestry groups, there are no significant batch effects (Table S.3). We have included the SNP count figures for 1KGP-HGDP-AFR-like and 1KGP-HGDP-EUR-like (Figure S.1).

Table S.3 -- Whole genome SNP count differences (Cohen's d)

All of Us genetic ancestry group	Cohen's d
1KGP-HGDP-AFR-like	d = 0.01
1KGP-HGDP-AMR-like	d = 0.15
1KGP-HGDP-EUR-like	d = 0.05

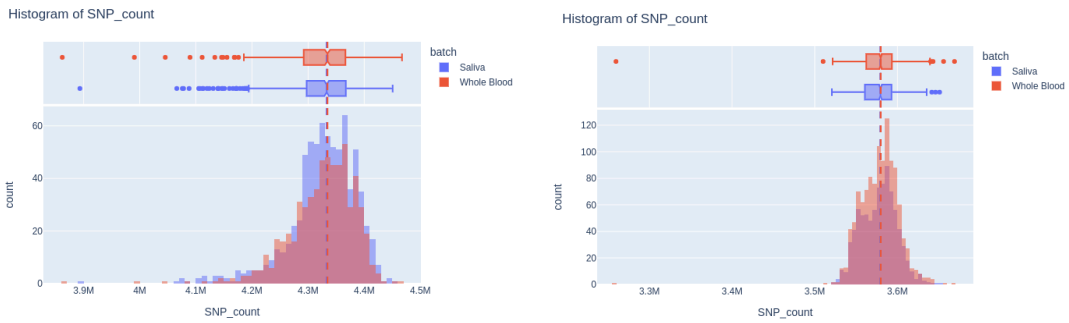


Figure S.1. SNP counts across the whole genome for the 1KGP-HGDP-AFR-like genetic ancestry group (d = 0.01, left) and the 1KGP-HGDP-EUR-like genetic ancestry group (d = 0.05, right).

Indel count

For the entire genome, the mean differences for indel counts are near a medium-size batch effect between blood and saliva samples (Table S.4). We have included the indel count figures for 1KGP-HGDP-AFR-like and 1KGP-HGDP-EUR-like (Figure S.2).

Table S.4 -- Whole genome indel count differences (Cohen's d)

All of Us genetic ancestry group	Cohen's d
1KGP-HGDP-AFR-like	d = 0.14
1KGP-HGDP-AMR-like	d = 0.24
1KGP-HGDP-EUR-like	d = 0.5

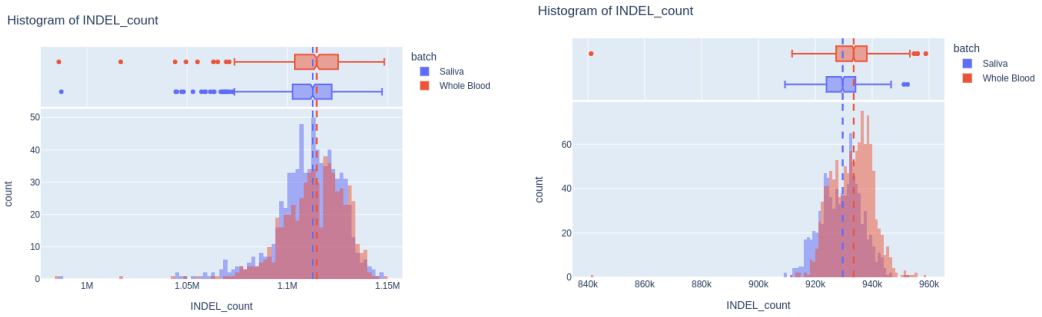


Figure S.2. Indel counts across the whole genome for the 1KGP-HGDP-AFR-like genetic ancestry group (d = 0.14, left) and the 1KGP-HGDP-EUR-like genetic ancestry group (d = 0.5, right).

SNP Ti/Tv ratio

In examining the SNP ti/tv ratios for the whole genome region, we see that the values do not indicate a batch effect between blood and saliva samples (Table S.5). We have included the SNP count figures for 1KGP-HGDP-AFR-like and 1KGP-HGDP-EUR-like (Figure S.3).

Table S.5 -- Whole genome SNP ti/tv ratio differences (Cohen's d)

All of Us genetic ancestry group	Cohen's d
1KGP-HGDP-AFR-like	d = 0.07
1KGP-HGDP-AMR-like	d = 0.13
1KGP-HGDP-EUR-like	d = 0.04

Figure S.3. SNP ti/tv ratios across the whole genome for the 1KGP-HGDP-AFR-like genetic ancestry group (d = 0.07, left) and the 1KGP-HGDP-EUR-like genetic ancestry group (d = 0.04, right).

Indel Ins/Del ratio

For the entire genome, the mean differences for the indel ins/del ratio are nearing values that indicate a batch effect between saliva and blood samples (Table S.6). We have included the indel count figures for 1KGP-HGDP-AFR-like and 1KGP-HGDP-EUR-like (Figure S.4).

Table S.6 -- Whole genome indel ins/del ratio differences (Cohen's d)

All of Us genetic ancestry group	Cohen's d
1KGP-HGDP-AFR-like	d = 0.34
1KGP-HGDP-AMR-like	d = 0.18
1KGP-HGDP-EUR-like	d = 0.68

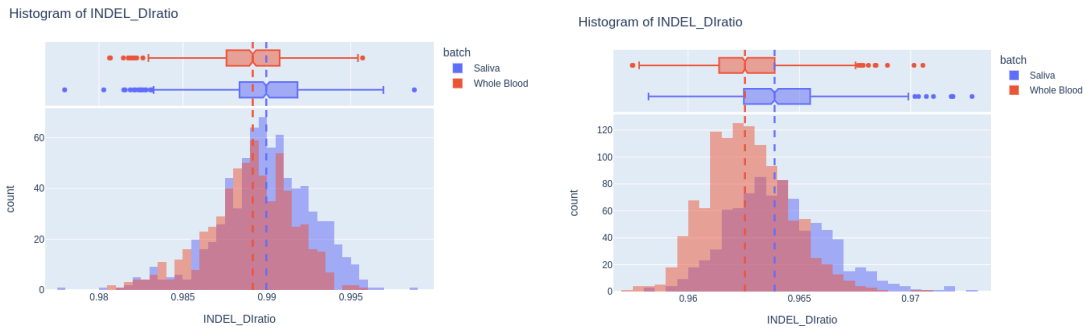


Figure S.4. Indel ins/del ratios across the whole genome for the 1KGP-HGDP-AFR-like genetic ancestry group (d = 0.34, left) and the 1KGP-HGDP-EUR-like genetic ancestry group (d = 0.68, right).

Whole genome minus low complexity (WG excl. LC)

When we remove the low complexity regions from the whole genome, for all of our statistics, we see no significant differences in the means that indicate batch effects between blood and saliva samples (Table S.2). This means that even in the indel counts and indel ins/del ratios, there are not significant differences indicating appreciable batch effects (Table S.7). We included the indel count figure for 1KGP-HGDP-AFR-like and 1KGP-HGDP-EUR-like (Figure S.5).

Other regions that indicated no appreciable batch effects for all metrics were GiaB, GC > 0.85, and ACMG59.

Table S.7 -- WG excl. LC indel count differences (Cohen's d)

All of Us genetic ancestry group	Indel count
1KGP-HGDP-AFR-like	d = 0.03
1KGP-HGDP-AMR-like	d = 0.2
1KGP-HGDP-EUR-like	d = 0.19

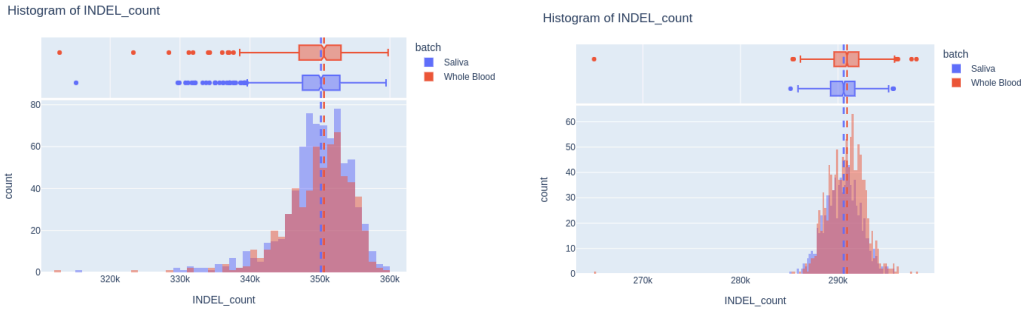


Figure S.5. Indel counts for the WG excl. LC region for the 1KGP-HGDP-AFR-like genetic ancestry group (d = 0.03, left) and 1KGP-HGDP-EUR-like genetic ancestry group (d = 0.19, right).

Conclusion

While there are no appreciable significant differences between blood and saliva samples in the SNP calls across the entire genome, we see some differences in indel calling as seen in the indel counts and indel ins/del ratio. The batch effects primarily impact low complexity regions, seen with the significant differences in the whole genome, GC < 0.25, for low mappability, SegDupe, and Tandem repeat regions. We did not see the same impacts in the whole genome once we removed low complexity regions (WG excl. LC), in GiaB, GC > 0.85, or the ACMG59 region.

We observe batch effects in regions where differences between saliva and blood sequencing are expected. Saliva DNA is often less pure and more fragmented, which amplifies mapping challenges and variant-calling issues in these hard-to-call regions. Shorter DNA fragments particularly complicate indel calling, as they make it more difficult to accurately determine the size and sequence of the variant.

We would predict that srWGS SV calls are more impacted, but the srWGS SV pipeline breaks the samples into batches and controls for these effects.

Remediation

We recommend that researchers take into consideration the sample source of saliva or blood if they are including indels in their research within regions where we see batch effects (Table S.2). This includes the smaller callsets such as the ACAF Threshold, ClinVar, and the exome. The sample source variable is located in the genomic metrics file in the column sample_source and can be used as a confounding variable in association testing.

Additionally, to control for the differences between blood and saliva samples, you can remove the regions containing differences indicating batch effects, such as low complexity regions.

Even with the observed differences between blood and saliva samples, the inclusion of saliva samples significantly broadens the participant base, providing a valuable resource for ongoing research.

Appendix T: Allele Frequencies of All of Us CDRv8 compared to gnomAD v3

We performed additional quality analyses beyond our standard to verify our methods and provide researchers with deeper insight into the *All of Us* genomic data. We compared the variants in the *All of Us* dataset to the variants in gnomAD, assessing the concordance of the allele frequencies. The allele frequency (AF) of a variant is equal to the number of times a certain allele is recorded divided by the number of total nucleotides sequenced at that location. The gnomAD v3 and *All of Us* CDRv8 datasets share around 100 million variants, and these variants have AF values that are almost identical. These patterns persist in the complete dataset and within the shared genetic ancestry groups (1KGP-HGDP-AFR-like, 1KGP-HGDP-AMR-like, 1KGP-HGDP-EAS-like, 1KGP-HGDP-EUR-like, 1KGP-HGDP-SAS-like, and Remaining Individuals (OTH)). The variants used in this analysis were taken from the *All of Us* Variant Annotation Table.

We found that there was a positive linear trend when comparing the AF values between *All of Us* and gnomAD (Figure T.1), indicating that the AF values were nearly identical for the same variants. We analyzed the Root Mean Square Error (RMSE), the Pearson correlation coefficient, and the extreme-difference between the two datasets to further analyze the similarity of the *All of Us* CDRv8 dataset to gnomAD v3 (Table T.1). The RMSE of 0.012 indicates the AF values had low variation between the two datasets. The Pearson correlation coefficient of 0.995 indicates the linear correlation between the two datasets. To calculate the extreme-difference, we counted the number of variants with an *All of Us* AF greater than 0.1 and a gnomAD AF less than or equal to 0.01, dividing by the total number of variants, and multiplying by 100%. We found the extreme-difference to be 0.021%. The values in the *All of Us* and gnomAD genetic ancestry groups also were consistent with these trends (Table T.1).

We only included variants with a high call rate in this analysis because the majority of variants with a sizable discrepancy in AF were not frequently sampled in our datasets. We measured the call rate using the allele number (AN), a count of the total number of alleles that were genotyped at the site. If all samples were genotyped, the AN will be the number of samples times 2. We filtered variants with a gnomAD AF of 0, or an AN in either dataset less than one-fifth of the maximum AN of the dataset (gnomAD: 30,463, *All of Us*: 165,932).

The minimal differences in AF values between the two datasets can likely be explained by the use of different variant calling methods in the two datasets and because these two datasets include individuals with different genetic ancestry groups.

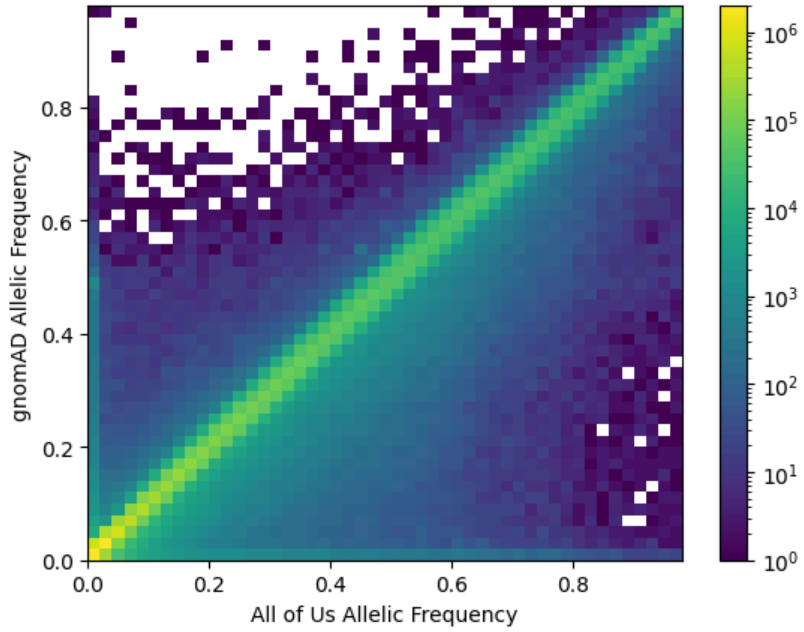


Figure T.1 -- A two-dimensional histogram of each variant and its corresponding AF in the *All of Us* CDRv8 and gnomAD v3 datasets, showing a linear correlation. We excluded variants with an AN in either dataset less than one-fifth of the maximum AN of the dataset (gnomAD: 30,463, *All of Us*: 165,932).

Table T.1 -- Comparison statistics for the entire callset and each genetic ancestry group, excluding sites with a low call rate.

Analysis group	RMSE	Pearson correlation coefficient	% of variants with an extreme difference in AF	Number of variants
Complete Dataset	0.012	0.995	0.021%	105,332,875
1KGP-HGDP-AFR-like (AFR)	0.013	0.996	0.027%	85,439,341
1KGP-HGDP-AMR-like (AMR)	0.016	0.994	0.024%	73,004,766
1KGP-HGDP-EAS-like (EAS)	0.023	0.995	0.058%	26,376,040
1KGP-HGDP-EUR-like (EUR)	0.013	0.996	0.026%	77,675,174
1KGP-HGDP-SAS-like (SAS)	0.020	0.995	0.049%	31,292,104
Remaining individuals (OTH)	0.016	0.996	0.024%	53,024,695

Appendix U: Sequencer (NovaSeq 6000 vs NovaSeqX) batch effect analysis

Introduction

Beginning in the CDRv9 genomic data release, we added srWGS samples sequenced with the NovaSeqX (NVX) whereas in previous data releases, we only included samples sequenced with the NovaSeq6000 (NS6000). We performed an analysis to determine batch effects between the two sequencers. In CDRv9, 27,749 of the 535,662 samples (5%) were sequenced with the NVX. All other samples (95%) were sequenced with the NovaSeq6000.

We performed a batch effect analysis, following similar methods to the Saliva and Blood batch effect analysis (Appendix S). We measured the batch effect size using Cohen's d [74], with an "appreciable batch effect" as any value of d greater than 0.5 in magnitude, which is the convention for a medium-size effect [74].

We examined four variant metrics to compare the sequencer data: SNP count, indel count, SNP transition to transversion (Ti/Tv) ratio, and Insertion/Deletion (Ins/Del) ratio (Table U.2). The genome regions we compared are the whole genome, the whole genome excluding low complexity regions (WG excl. LC), the Genome in a Bottle high confidence calling regions (GiaB), high GC content (GC > 0.85), AT-rich genome region(GC < 0.25), American College of Medical Genetics (ACMG)'s list of 59 genes (ACMG59), low mappability, the exome, the regions that meet the AC > 100 or AF > 1% cutoff (ACAF Threshold), and ACAF threshold excluding low complexity regions (Table U.1).

Table U.1 -- Regions used in the batch effect analysis

Region	Description
Whole genome excluding low complexity regions (WG excl. LC)	Low complexity regions removed from the whole genome
Genome in a Bottle (GiaB) high confidence regions	National Institute of Standards and Technology (NIST) benchmark high confidence variant call regions provided by the Genome in a Bottle consortium
GC (GC > 0.85)	High guanine (G) and cytosine (C) content in a genomic region
AT-rich genome region (GC < 0.25)	Region with low GC content
American College of Medical Genetics (ACMG)'s list of 59 genes (ACMG59)	A list of 59 genes that the ACMG recommends be actively searched for pathogenic or likely pathogenic variants during clinical genomic sequencing
Low mappability	Low mappability regions are regions of the genome where sequences cannot be mapped with as high confidence as other regions of the genome. These regions are generally less reliable.

AoU exome (smaller callset)	Region where we defined exome (see Smaller Callsets for Analyzing Short Read WGS SNP & Indel Data with Hail MT, VCF, and PLINK)
AoU ACAF Threshold (smaller callset)	Region where a SNP/Indel met the ACAF Threshold (see Smaller Callsets for Analyzing Short Read WGS SNP & Indel Data with Hail MT, VCF, and PLINK)
AoU ACAF Threshold (smaller callset) excl. low-mappability or GC extreme regions	Region where a SNP/Indel met the ACAF Threshold (see Smaller Callsets for Analyzing Short Read WGS SNP & Indel Data with Hail MT, VCF, and PLINK) minus the GC>0.85, low mappability, and GC<0.25 regions.

Results

The results from the batch effect analysis are demonstrated in Table U.2. While there are generally no batch effects across the whole genome excluding LC regions, the GiaB high confidence regions, the exome, and the ACAF Threshold smaller callset, we did see appreciable batch effects in the ACAF threshold Ins/Del ratio, the GC > 0.85, the GC < 0.25, and the low mappability regions.

This is consistent with the batch effects primarily affecting complex and hard-to-call genomic regions. These challenging regions inherently amplify mapping discrepancies and variant-calling inconsistencies between the two systems. Furthermore, shorter DNA fragments compound these alignment difficulties, making it more challenging to consistently determine the size and sequence of indels when transitioning across these sequencing platforms.

Please note that the Ins/Del ratio was small (d = 0.55), which was driven by GC extreme regions.

Remediation

We recommend that researchers take into consideration the sequencer version if they are including regions where we see batch effects (Table U.2). This includes the smaller callsets such as ACAF Threshold and ClinVar. The sequencer version is located in the genomic metrics file in the column sequencer and can be used as a confounding variable in association testing.

Additionally, to control for the differences between sequencers, you can remove the regions containing differences indicating batch effects, such as low complexity regions.

Even with the observed differences between these two sequencers, the inclusion of all sequenced samples significantly broadens the participant base, providing a valuable resource for ongoing research.

Table U.2 -- Presence of appreciable batch effects across each region and metric

Region	SNP count	Indel count	SNP Ti/Tv ratio	Ins/Del ratio

GiaB	No	No	No	No
WG excl. LC	No	No	No	No
AoU exome (smaller callset)	No	No	No	No
AoU ACAF Threshold (smaller callset) excl. low-mappability or GC extreme regions	No	No	No	No
AoU ACAF Threshold (smaller callset)	No	No	No	Yes*
GC > 0.85	Yes	Yes	Yes	Yes
GC < 0.25	Yes	Yes	Yes	Yes
ACMG59	No	No	No	No
Low mappability	No	Yes	Yes	Yes

* Small effect in the ACAF Threshold Ins/Del ratio of d = 0.55, driven by GC extreme regions.