How the *All of Us* genomic and multi-omic data are organized

Introduction

The *All of Us* genomic and multi-omic data includes short read whole genome sequencing (srWGS) data, long read whole genome sequencing (lrWGS) data, microarray genotyping array ("array") data, RNA sequencing ("RNA-seq") data, and proteomics sequencing data. Researchers access these data through the Researcher Workbench (RW) Controlled Tier dataset (e.g. genomic data is not available through the Registered Tier). Bucket locations for accessing the data in analysis notebooks can be found in the Data Dictionary.

A summary of the file formats for each data type can be viewed in the list below.

In this article, we will summarize the genomic data formats and what information is available in each data type. In some cases, we will refer to other documentation when it describes the data format we deliver. This article assumes a general knowledge of genomics and bioinformatics. For a workspace on getting started with genomic data on the Researcher Workbench, please see the "How to Work with All of Us Genomic Data" notebook in the All of Us controlled Tier Featured Workspace. We also provide a detailed report on the quality of the genomic data with each release in the *All of Us* Genomic Data Quality Report available on the User Support Hub.

List of *All of Us* genomic data

Short-read whole genome sequencing (srWGS) - 535,662

- Sequencing reads in CRAM format, aligned to hg38/GRCh38
- SNP & Indel Variant data
    - Hail Variant Dataset (VDS): joint callset across the entire genome
    - Exome callset: Hail MT, VCF, PLINK bed, PGEN, BGEN, UCSC bed
    - ClinVar callset: Hail MT, VCF, PLINK bed, PGEN, BGEN, UCSC bed
    - ACAF threshold callset: Hail MT, VCF, PLINK bed, PGEN, BGEN, UCSC bed - variants that have population-specific AF > 1% or population-specific AC > 100
- Annotated variants: Variant Annotation Table
- Auxiliary Data
    - Genetic ancestry
    - Pharmacogenomics (star alleles)
    - Challenging medically relevant genes (CMRG) callset: VCF
    - Relatedness
    - Maximal set of unrelated samples
    - HLA variant calling
    - Flagged samples
    - Genomic QC values
    - Genomic metrics
    - Control samples

Short-read whole genome sequencing structural variants (SVs) - 96,405

- Joint-called SV VCF
- Sites-only SV VCF
- srWGS SV maximal set of unrelated samples
- Unrelated sites-only VCF
- srWGS SV samples with probable aneuploidies
    - Samples with probable mosaic autosomal aneuploidy
    - Samples with probable mosaic allosomal aneuploidy
    - Samples with probable germline allosomal aneuploidy
- Sample list

Genotyping array - 553,949

- Raw genotyping scanner data in IDAT format
- SNP & Indel variant data
    - Single-sample VCFs
    - Single-sample Hail MT
    - Single-sample PLINK bed

Long-read whole genome sequencing (lrWGS) - 14,521

- Cohorts grouped by sequencing facility and platform

- Sequencing reads in BAM format, aligned to grch38_noalt
    - Annotated with methylation signals
- *De novo assembly* in GFA and FASTA format for cohorts with PacBio data
- Variant data
    - Joint SNP & Indel variants for each cohort in GVCF & Hail MT formats
    - Single-sample SNP & Indel variants in GVCF format
    - Single-sample SVs from PBSV & Sniffles2
    - Single-sample PAV variants in VCF format for samples with PacBio data
- Auxiliary sample metrics for both reference versions

RNA sequencing data

- RNA STAR-aligned reads
- RNA expression QTL files
    - RNA expression cis QTL
    - Cis QTL nominal stats
    - Fine-mapped cis eQTLs
- RNA splicing QTL files
    - RNA splicing cis QTL
    - Fine-mapped sQTL
    - Splicing junctions sQTLs
- RNA auxiliary files
    - RNA metadata metrics
    - RNA-SeQC 2 metrics
    - RNA RSEM

Proteomics data

- Normalized protein expression (NPX files)
    - Parquet and tsv
    - Normalized and replicate removed tsvs
- Proteomics pQTL files
    - Cis pQTL files
- Olink analysis report

Short-Read Whole Genome Sequencing (srWGS) Data

The srWGS data includes raw data in CRAM format, variant data in a complete Variant Dataset (VDS), or as smaller callsets in VCF, Hail MT, BGEN, PGEN, and PLINK bed formats. We provide auxiliary data for srWGS data as annotated variants in the Variant Annotation Table (VAT), genetic ancestry, pharmacogenomics variant calls, relatedness, maximal set of unrelated samples, HLA variant calling, flagged samples, and genomic QC values.

Table 1. Short-read WGS deliverables

Deliverable	srWGS SNP & Indel
Reference version	hg38/GRCh38 reference: gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta
Raw data	CRAM files
Variant data	Joint-callset variant data for all samples: VDS Smaller callsets: ACAF threshold, exome, ClinVar - in VCF, Hail MT, BGEN, PGEN, PLINK bed, UCSC bed formats CMRG callset
Auxiliary files	Annotated variants: Variant Annotation Table Genetic ancestry Pharmacogenomics variant calls (star alleles) Relatedness Maximal set of unrelated samples HLA variant calling Flagged samples Genomic QC values srWGS Genomic metrics file Control samples

srWGS CRAM files

We provide raw data for srWGS samples in CRAM format, otherwise known as compressed SAM (sequence alignment map) format. The data are mapped to the hg38/GRCh38 reference. Refer to the *All of Us* Genomic Data Quality Report for more information on how variant calling was performed on these raw data files.

There is one CRAM file and one CRAM index file for each srWGS sample and the research ID appears in the file name. The path to each CRAM file is found in the manifest CSV file, which contains a row per sample of person_id,cram_uri,cram_index_uri

The raw data is more expensive to use because you must pay egress charges, which are the costs to retrieve the data from the cloud for analysis. We do not charge egress for variant data and so the raw data will be more expensive to use. Please see the Genomics FAQ for Recommendations for processing CRAMs with GATK on the Researcher Workbench.

srWGS SNP & Indel variant data

The srWGS SNP & Indel dataset is joint-called and delivered as a complete callset in VariantDataset (VDS) format, which is a Hail data storage format for large datasets. Hail MT, VCFs, and PLINK files are available for all samples over limited regions, including the exome, ClinVar variants, and common variants within each genomic ancestry group. For further information about the Hail MT, VCF, and PLINK files, please see Smaller Callsets for Analyzing Short Read WGS SNP & Indel Data with Hail MT, VCF, and PLINK.

Variant Dataset (VDS)

The Hail VariantDataset (VDS) is a data storage format we use for the *All of Us* srWGS SNP & Indel variant data. With one of the largest callsets in the world, the VDS helps to store variant data efficiently for all samples over the entire genome. The VDS is a sparse Hail data storage format that stores less data, but more information. As a comparison, the Hail MT is a dense variant storage format with every entry populated. For an overview of the VDS, check out 'The new VDS format for *All of Us* srWGS data' article.

If possible, we recommend that researchers use the smaller callsets for their analysis to save time and money. Most downstream analyses of the VDS involve filtering and converting the VDS into a VCF, Hail MT, or other dense format ("densifying"). We have performed this step already to cover most use cases with reduced srWGS SNP & Indel variant datasets in VCF, Hail MT, BGEN, and PLINK bed formats over commonly used areas of the genome (see Genomics FAQ: Smaller callsets for analyzing srWGS SNP & Indel data with Hail MT, VCF, and PLINK).

Instructions for densifying the VDS are available in the article 'The new VDS format for *All of Us* srWGS data' and the Manipulate Hail VariantDataset tutorial notebook.

In the following sections, we describe how the VDS stores variant data, reference data, and how to determine if a variant site is filtered.

Variant data

The VDS uses variant level row fields to store data for all samples, including the variant locus (`locus`), a list of alternate alleles (`alleles`), and site level filtering data (`filters`). Local fields store data that only apply to a single sample, including genotype metadata and genotype filtering. The local alleles (`LA`) array maps the alleles that appear in the individual sample to the list of alternate alleles (`alleles`), thus genotype metadata is only stored for samples with the genotype.

Some familiar annotations from a VCF or Hail MT are not present in the VDS, but can be rendered when densifying the VDS. The allele count for each alternate allele (AC), the total number of alleles at each site (AN), and the frequency of each alternate allele (AF) are also stored in the Variant Annotation Table (VAT) for all variants that pass filtering.

Tables 2-5 describe the fields in the *All of Us* VDS. Please see the Hail documentation for more information on the Hail data types.

Table 2. VDS column fields: stores sample name

VDS Field	Description	Hail data type
s	Research ID	str

Table 3. VDS row fields: stores variant data

VDS Field	Description	Hail data type	Example
locus	Positional data for the variant. Formatted as chromosome name and position separated by colon.	locus<GRCh38>	chr1:12807
alleles	List of alleles at a locus for all samples (otherwise known as global alleles). The first allele is the reference allele. All the alternate alleles are then listed in alphabetical order.	array<str>	[“C”, “T”]
filters	Site level filtering information. Hard threshold filters include EXCESS_ALLELES, NO_HQ_GENOTYPES, LowQual, and ExcessHet. If no filtering reason is provided or there is a PASS, then the site has passed filtering.	set<str>	{“LowQual”, “NO_HQ_GENOTYPES”}
as_vets	Variant Extract-Train-Score Filtering model information for this site. Does not contain information about whether or not the site was filtered. We recommend that most users ignore this field and look at filters for useful filtering information.	dict<str, struct { model: str, calibration_sensitivity: float64 }>	{"T":("INDEL",7.58e-01)}

Table 4. VDS entry fields: stores genotype level variant data

VDS Field	Description	Hail data type	Example
GQ	Genotype Quality. Follows VCF description.	int32	63
RGQ	Reference Genotype Quality. Follows VCF description.	int32	101
PS	Phase set - the set of phased genotypes to which this genotype belongs. The PS field contains an integer that represents the position of the first phased variant in the set. If the genotype is unphased, the corresponding PS field is ignored.	int64	26887031
LGT	Local genotype. The coordinates map to LA. LA always includes the reference allele so the call can be [0/1], [1/1], or [1/2].	call	[1/1]
LAD	Local allele depth, describes the allele depth for one sample. Maps to the alleles described in the local alleles (LA) array. See VCF description.	array<int32>	[0,8]
LA	Local alleles. The reference allele and allele(s) that appear in the sample are listed as coordinates mapping	array<int32>	[0,1]

	to the global `alleles` array. The reference coordinate is always included.		
FT	Boolean containing genotype level filtering. `True` for PASS, `False` for FAIL, and NA for (.). In most cases, NA should be treated as PASS. The filtering reason is not provided.	`bool`	`True`

Table 5. VDS global fields: filtering metadata for the entire callset

Note: These fields report metadata of the filtering model. See the row filter field `filters` or entry genotype field FT to see whether a variant did not meet the threshold reported in these fields.

VDS Field	Description	Hail data type
`truth_sensitivity_snp_threshold`	SNP sensitivity threshold	`float64`
`truth_sensistivity_indel_threshold`	Indel sensitivity threshold	`float64`

Reference data

The VDS also stores reference data for each sample as reference blocks in a separate component table `reference_data`. The row key is the `locus` and the `ref_allele` denotes the reference base at the genomic coordinate. Columns are keyed by the sample ID. No data at a particular location indicates that the sample has a variant call.

Table 6. VDS reference data column fields: stores sample name

VDS Field	Description	Hail data type
`s`	Research ID	`str`

Table 7. VDS reference data row fields: stores reference data

VDS Field	Description	Hail data type	Example
`locus`	Positional data for the variant. Formatted as chromosome name and position separated by colon.	`locus<GRCh38>`	`chr1:10029`
`ref_allele`	The reference allele at the genomic coordinate	`str`	“A”

Table 8. VDS reference data entry fields: stores reference blocks

VDS Field	Description	Hail data type	Example
`GQ`	Genotype Quality. Follows VCF description.	`int32`	`40`
`END`	Indicates the end of the reference block, which is the group of consecutive non-variant sites that have the same genotype quality. All coordinates between the start locus and the end coordinate are called as reference for the sample.	`int32`	`10036`

Filtering information

The variant filtering data is represented in two fields in the VDS, `filters` and the `FT` field (Table 3, Table 4). The `filters` array contains site level filters, including `EXCESS_ALLELES`, `NO_HQ_GENOTYPES`, `LowQual`, and `ExcessHet`. If no filtering reason is provided or the `filters` field contains `PASS`, then the site has passed filtering. The `FT` field contains genotype level filtering. The genotype level filtering reasons are not specified in the *All of Us* VDS, there will be a boolean describing the filtering status for the genotype. `True` is PASS and `False` is FAIL. If all genotypes fail at a site, the `True` or `False` boolean can also apply to the `filters` array. The variant filtering process is described in depth in the QC report. All filtered variants are soft filtered, which means the variants will be marked but not removed from the callset.

We provide a tutorial notebook for converting VDS to a Hail MT format, including code to transform the `FT` boolean `True` or `False` in the VDS to `PASS` or `FAIL` so that it is compatible for converting to a VCF.

srWGS SNP & Indel smaller callsets

We released the srWGS SNP and Indel callset in familiar data formats over limited genomic regions: VCF, Hail MT, BGEN, and PLINK bed formats. The smaller callsets, described in Smaller callsets for analyzing srWGS SNP & Indel data with Hail MT, VCF, and PLINK, cover regions of the genome that are popular for *All of Us* researchers: an Allele Count/Allele Frequency (ACAF) threshold callset, an exome callset, and a ClinVar callset. We recommend that you stick with these premade Hail smaller callsets instead of using the VDS, if possible, to save time and money.

The ACAF threshold callset contains variants that have a population-specific allele frequency (AF) greater than 1% or a population-specific allele count (AC) greater than 100 in any computed ancestry subpopulations. The exome callset contains variants that are within the exon regions of the Gencode v42 basic transcripts, with padding of 15 bases on either side of each exon. The ClinVar callset contains variants in ClinVar, regardless of pathogenicity.

The complete srWGS SNP and Indel callset across all sites is released as a VDS, which is a Hail sparse data format. We provide a tutorial notebook for converting VDS to a Hail MT format, though we recommend that you stick with the premade Hail MT, if possible, to save time and money.

Hail MT

We provide two Hail MTs for each smaller callset, both a multiallelic and multiallelic split Hail MT, resulting in six total Hail MT deliverables for the srWGS SNP and Indel callset. In the multiallelic split MT, sites with multiple alternate alleles will be split, so each row will only have one alternate allele. In the multiallelic MT, sites with multiple alternate alleles will be retained in the same row.

When using Hail MT files in the Researcher Workbench, read directly from the bucket location. Do not attempt to copy them locally.

The Hail MT follows Hail format specifications. For additional examples for what you may expect to see in the data, see the following VCF examples.

VCF

The srWGS limited callset VCFs are sorted and block compressed in bgz format (.vcf.bgz) with a local tabix index (.vcf.bgz.tbi). Each VCF is split into multiple non-overlapping sections of the genome by chromosome in separate files for usability (sharding).

Please note that we recommend using the FILTER column and the filter tag (FT) to determine the filtering status of a variant because the QUAL information is not available.

FORMAT fields (per sample-site):

- Genotype (GT) -- The GT field specifies the alleles carried by the sample, encoded as 0 for the reference (REF) allele, 1 for the first alternative (ALT) allele, 2 for the second ALT allele, etc. The allele values are separated by a / or |, depending on if the genotype is phased or not. The / separator is for unphased variants and the | separator is for phased variants (see PS field below). Since humans are diploid organisms, we expect two alleles (e.g. “0/1”). Please note that the GT calls for chrX and chrY variants may be reported as either haploid or diploid, even in the case of chrY and chrX in males.
- Allelic Depth (AD) -- Allelic depths for the reference allele and the alternate allele(s) present at this site. For more information about AD and which reads are counted, see this article on Allele Depth.
- Genotype Quality (GQ) -- The phred-scaled confidence that the called genotype is correct. A higher score indicates a higher confidence. For more information on GQ, please see the GQ documentation. For more information on interpreting phred-scaled values, please see Phred-scaled quality scores.
- Reference Genotype Quality (RGQ) -- The phred-scaled confidence that the reference genotypes are correct. A higher score indicates a higher confidence. For more information on RGQ, please see the GQ documentation, but note that RGQ applies to the reference, not the variant. For more information on interpreting phred-scaled values, please see Phred-scaled quality scores.
- Genotype Filter (FT) -- The srWGS SNP & Indel genotype-level filtering information. As part of our joint callset quality control processing, we run the Variant Extract-Train-Score (VETS) method, which is a genotype-level filtering algorithm. If the genotype passes, there will be no value in this field. If the genotype fails, the value will be `high_CALIBRATION_SENSITIVITY_SNP` or `high_CALIBRATION_SENSITIVITY_INDEL`. An example code snippet for filtering genotypes, in Hail, can be found in the Manipulating Hail Matrix Table tutorial notebook.
    - `high_CALIBRATION_SENSITIVITY_SNP`: Sample Genotype FT filter value indicating that the genotyped allele failed SNP model calibration sensitivity cutoff (0.997)

    - `high_CALIBRATION_SENSITIVITY_INDEL`: Sample Genotype FT filter value indicating that the genotyped allele failed INDEL model calibration sensitivity cutoff (0.99)
    - Add in something here about CALIBRATION SENSITIVITY. Info field of VCF. as_vets.get(alt).calibration_sensitivity
- Phase Set (PS) -- A phase set is defined as a set of phased genotypes to which this genotype belongs (See VCF 4.1 specifications). The PS value is an integer representing the position of the first phased variant in the set. It is not available for all variants. The first variant in the phase set will contain the PS identifier. If the genotype in the GT field is unphased, the corresponding PS field is ignored. The phasing data is from the DRAGEN 3.7.8 pipeline during the genotyping step, by comparing haplotypes and variants within an active variant region.
    - The PS field will appear in the VCF, Hail MT, and VDS and will not appear in PLINK data types. If using downstream tools from Hail or VCF that expect unphased data, then researchers need to perform a step to unphase the data.

INFO fields (per site):

Descriptions of the INFO fields can also be found in the header of the VCF.

- Allele Count (AC) -- the number of times we see each alternate allele for all samples. For example, a “1/1” genotype would count as 2 observations of the first alternate allele.
- Allele Number (AN) -- the total number of alleles seen. Usually, this will be the number of samples times two, since humans are diploid organisms. No-call genotypes (“./.”) are not counted towards AN.
- Allele Frequency (AF) -- the frequency of the alternate allele in the population that is the callset cohort. This is equivalent to AC/AN.
- QUAL approximation (QUALapprox) -- the sum of the phred-scaled homozygous reference probability values across all samples, which is a proxy for the site-level QUAL score, but without the SNP or indel heterozygosity applied as a per-site prior probability of variation.
- Allele-specific QUAL approximations (AS_QUALapprox) -- a per-allele, phred-scaled quality score derived from the sum of homozygous reference probability values across samples when each allele is considered in isolation. This is an approximation of the QUAL score for each allele. For more information on the QUAL score, see the VCF specification.

FILTER values (per site):

- QUAL score does not meet threshold (LowQual) -- sites with this filter have a posterior probability of being variant that is equal to or below the probability of being variant by chance, represented by the expected heterozygosity for humans (QUALapprox lower than 60 for SNPs; 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.

- No high-quality genotypes (NO_HQ_GENOTYPES) -- sites with this filter do not have any genotypes that are considered high quality (GQ>=20, DP>=10, and AB>=0.2 for heterozygotes)
    - Allele Balance (AB) is calculated for each heterozygous variant as the number of reads supporting the least-represented allele over the total number of read observations. In other words, min(allele depth)/(total depth) for diploid GTs.
- Excess Heterozygosity (ExcessHet) -- sites with this filter have more heterozygote genotypes than expected by chance under Hardy-Weinberg equilibrium. ExcessHet is a phred-scaled p-value. We cutoff anything more extreme than a z-score of -4.5 (p-value of 3.4e-06), which phred-scaled is 54.69
- Excess alleles (EXCESS_ALLELES) -- sites with this filter have an excess of alternate alleles, which our cutoff is 100. When a site has more than 100 alternate alleles, this filter will be present.

PLINK 1 binary biallelic genotype table (PLINK bed)

We provide PLINK 1.9 data (.bed / .bim .fam) for the srWGS SNP and Indel smaller callsets. The PLINK files are converted from the Hail MT using the export_plink command in Hail and contain all information in the Hail MT. PLINK file type information can be found at the PLINK site. The .bed file is the PLINK binary biallelic genotype table and contains genotype calls. The .bim file is the PLINK extended .map file, and is a text file containing variant information. The .fam file is a text file with sample information for each participant. Please refer to the published notebooks on how to use the PLINK 1.9 data.

PLINK 2 binary genotype table (PGEN)

We provide PLINK 2 data (.pgen, .psam, .pvar) for the srWGS SNP and Indel smaller callsets, sharded by chromosome. The binary .pgen file is the file containing genotype calls. It is accompanied by two text files: .pvar with variant information and .psam with sample information. Please see the PLINK documentation for more details. Tools such as SAIGE, Regenie, and PLINK2 can import PGEN files for analysis.

Binary GEN format (BGEN)

We have released the srWGS SNP and Indel smaller callsets in Binary GEN format (BGEN). The files are sharded by chromosome and only contain hard calls, which are calls with probability values of 0.0 or 1.0. Please see the BGEN documentation for more information about this format.

srWGS SNP & Indel smaller callset BED files

We provide the genomic territory, otherwise known as interval files, used to create the srWGS SNP and Indel smaller callsets as UCSC BED files. The BED files contain the genomic regions for the exome, ACAF threshold, and ClinVar callsets.

Challenging medically relevant genes (CMRG)

The CMRG callset is a separate callset for 30 protein-coding genes, including 7 challenging medically relevant genes (CMRG) such as KCNE1, CBS, and MAP2K3. These genes were impacted by falsely duplicated and collapsed errors in the GRCh38 reference genome as identified in a previous report in Science. We currently see reduced sensitivity in the srWGS SNP & Indel callsets for these genes.

To provide variant calls for these genes, we extracted reads from the CRAM files, we reconstructed BAM files using FixItFelix, called variants with DRAGEN-GATK, reblocked the VCFs, and performed joint calling with the GVS pipeline. We have provided USCS BED files containing the genomic regions that we called for the CMRG callset, available on the Researcher Workbench. The FixItFelix tool reconstructed the BAMs using a modified version of the hg38 reference with duplicate genes masked out. The modified version of the hg38 reference can be found at this link for download. Note: this callset has not been filtered and should not be intersected with other All of US callsets as it is called on a different reference.

Annotated variants - Variant Annotation Table (VAT)

The Variant Annotation Table (VAT) is a resource provided for all samples with srWGS SNP & Indel data. The VAT gives functional annotations for all passing variants. Variants must pass both site-level (`filters`) and genotype-level (FT) filtering. The Variant Annotation table contains site-level annotations such as allele counts for each alternate allele (AC), the total number of alleles at each site (AN), and the frequency of each alternate allele (AF). These site-level annotations are not in the VDS. Using the VAT in addition to the VDS can be used to determine variants of interest to your analysis. We provide the annotations as one single, merged tsv file (".tsv.bgz") which can be loaded into Hail. Please read the Variant Annotation Table article for more information.

srWGS auxiliary data

srWGS genetic ancestry

We provide genetic ancestry groupings for all samples with srWGS data as a .tsv file, sorted by research ID. 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) and 1000 Genomes Project training data (N=3,942) for all srWGS samples in *All of Us*. Please see the *All of Us* Genomic Data Quality Report Appendix G for more information.

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, the HGDP and 1000 Genomes: We use the following acronyms or terms to describe genetic similarity to a reference

population: 1KGP-HGDP-AFR-like (AFR or African); 1KGP-HGDP-AMR-like (AMR or Americas); 1KGP-HGDP-EAS-like (EAS or East Asian); 1KGP-HGDP-EUR-like (EUR or European); 1KGP-HGDP-MID-like (MID or Middle Eastern); 1KGP-HGDP-SAS-like (SAS or South Asian); and not belonging to one of the other ancestries or is an admixture (OTH or remaining individuals).

We provide the genetic ancestry groupings as a .tsv file along with a plot of the ancestry predictions (html file). The PCA analysis was performed using Hail's hwe_normalized_pca method. In order to allow researchers to reproduce these files and also apply our method for predicting genetic ancestry groupings on their own data, we also provide a set of files we used to predict genetic ancestry, described as follows:

- Loadings file: captures how each genetic variant contributes to the principal components (PCs). The file can be used to project an individual's genetic data on the same PCA space as the one used for the ancestry prediction. The loadings file is a Hail file type.
- Eigenvalues of the PCs: the eigenvalues represent the amount of genetic variation each PC explains.
- Classifier .pkl file: contains the trained ancestry prediction model.
- Training PCA: The genetic ancestry groupings of the training data (1000 Genomes and HGDP)
- Sites-only VCF: a sites-only VCF of the locations we used for training the ancestry predictions classifier (which is described as the HQ sites in the QC report, Appendix H). The VCF is block compressed and accompanied by a TBI index

Table 9. srWGS genetic ancestry TSV file description

Field Name	Key?	Type	Nullable?	Example Value	Notes
research_id	yes	String	No	1000055	This comes from sample metadata.
ancestry_pred	no	String	No	mid	The predicted ancestry for the sample, not including "other."
probabilities	no	Array[number]	No	[0.10, 0.99, 0.001, … 0.0]	Confidence of each output class (i.e. computed ancestry). Each will have a length equal to the number of possible computed ancestry labels minus one (6). Probabilities are listed in the order: AFR, AMR, EAS, EUR, MID, and SAS. The ancestry "Other" is computed separately based on the confidence of the other classes.
pca_features	no	Array[number]	No	[8.1232, 0.01234,	The principal

				3.1123, …, 0.00132]	components of the projection for the sample. Each value is an array with a length of 16.
ancestry_pred_other	no	String	No	oth	The predicted ancestry for the sample, including “other.”

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type. Arrays are possible.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

srWGS genetic admixture estimates

We provide genetic ancestry admixture estimates for all samples with srWGS data in .Q and FAM file formats. The analysis was performed with the Rye tool and the output file descriptions can be found in the Rye documentation.

The .Q file contains columns with the ancestry groups used in the training data and the rows are admixture estimates for each sample. The ancestry group labels that we use are 1KGP-HGDP-AFR-like (AFR), 1KGP-HGDP-AMR-like (AMR), 1KGP-HGDP-EAS-like (EAS), 1KGP-HGDP-EUR-like (EUR), 1KGP-HGDP-MID-like (MID), 1KGP-HGDP-SAS-like (SAS), and Remaining Individuals (OTH). We also provide the reference admixture estimates.

The .fam file contains the information for how each individual mapped to the training data.

Please note: The genetic admixture estimates for individuals with American ancestry may not be fully captured due to lack of appropriate samples from publicly available reference genome datasets (1KGP-HGDP in this case) to account for the full range of differences within this group. This inaccuracy may also exist for other global populations where there is limited reference data available such as the Middle Eastern group. Additionally, the ancestry proportion estimates for the *All of Us* participants in the 1KGP-HGDP-AMR-like genetic ancestry group is influenced by the presence of admixture within the genomes of 1KGP-HGDP-AMR individuals included in the reference datasets, affecting the accuracy. We advise caution when interpreting these estimates, as they may not fully capture the genetic richness within the Americas population.

srWGS pharmacogenomics data

A full description of the pharmacogenomics data is available in the article *All of Us* Pharmacogenomics (Star Allele) Calling.

The pharmacogenomics auxiliary dataset includes haplotype calls and predicted phenotypes for 18 genes relevant to human drug metabolism for all samples with srWGS data. Pharmacogenomic haplotype calling is also known as star allele calling. We provide variant data across 17 genes from PharmGKB Tier 1 and Tier2 lists that are supported by the tool Stargazer

and one gene called by Cyrius v1.1.1. Genes with strong validation data are provided in a set of "high concordance" outputs. Genes that play significant roles in drug metabolism but do not have convincing validation results are included in a set of "low concordance" outputs.

Star allele calls are provided as per-gene .tsv files. The .tsvs contain sample names and gene names and can be concatenated easily, but are provided separately for memory usage considerations.

We ran Stargazer 2.0.2 for all 18 genes other than CYP2D6. Stargazer output was post-processed to apply allele function definitions according to CPIC and improve phasing. Cyrius v1.1.1 was run on per-sample cram input to call CYP2D6 star alleles and gene copy number. Structural variation nomenclature was harmonized and phenotypes were applied using the cyp2d6_parser package.

High concordance genes: CYP2C_CLUSTER*, ABCG2, CACNA1S, CFTR, CYP2C9, CYP2D6, G6PD, NUDT15, RYR1, TPMT, VKORC1*

Low concordance genes: *CYP2B6, CYP2C19, CYP3A5, CYP4F2, DPYD, SLCO1B1, UGT1A1*

srWGS statistical phasing

We provide haplotype phasing data for all samples in the CDRv8 srWGS callset. The data is delivered as multi-sample VCFs, sharded by chromosome. Haplotype phasing is the estimation of haplotypes that are inherited from each parent. We use statistical methods to infer the sequence of alleles on each chromosome, following methods from the 2021 paper from Browning, Brian L. et al.: Fast two-stage phasing of large-scale sequence data.

To generate the phasing data, we performed focused QC on the srWGS VDS. We first removed variants with more than 31 alternate alleles. We then removed variants with an average sum of their AD less than 28, removed variants with a max AC less than 2 (to remove singletons and doubletons), removed variants with a mean GQ of less than 30, and variants with FILTER values of LowQual, NO_HQ_GENOTYPES, or ExcessHet.

We used Beagle 5.5 for phasing with a window parameter of 15, 20, or 40, depending on the chromosome. All other parameters were default. Genetic distances were interpolated from the HapMap genetic map.

srWGS HLA calling

We provide variant calls for the Human Leukocyte Antigen (HLA) complex for the srWGS data samples. The HLA complex is a highly polymorphic region of the human genome that plays a critical role in immune system regulation, antigen presentation, and transplant compatibility. The HLA variant calls are available in a cohort matrix.

The HLA pipeline uses the tools and is described in more detail in the *All of Us* HLA typing documentation (coming soon). The pipeline uses the tools HLA-HD, Polysolver, and Optitype to generate high-resolution HLA calls from CRAM files. The output is an output summary table with the fields sample_id, [Gene]_1, and [Gene]_2.

srWGS relatedness kinship scores

We calculate relatedness for all samples with srWGS data and report the kinship score of any pair with a score over 0.1. The kinship score is half of the fraction of the genetic material shared. (Parent-child or siblings will have a score of 0.25 while identical twins will have a score of 0.5). Please see the Hail pc_relate function documentation for more information, including interpretation.

We provide the kinship scores for pairwise samples with kinship scores above 0.1. We do not provide identity kinship scores (i.e. kinship of a sample with itself). Each pair will only appear once (in other words, {sample1, sample2, 0.25} is equivalent to {sample2, sample1, 0.25}).

Table 10. srWGS pairwise samples with a kinship score over 0.1 TSV file description

Field name	Type	Key?	Notes
i.s	string	yes	Sample ID of a sample in the pair
j.s	string	yes	Sample ID of the other sample in the pair
kin	float	no	Kinship score (0-0.5)

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type. Arrays are possible.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

srWGS SNP & Indel maximal set of unrelated samples

We provide a list of samples to prune in order to remove related samples from the srWGS SNP & Indel cohort. Relatedness is calculated as described in the kinship score description above. This will be the maximal independent set for related samples which minimizes the number of samples that need pruning.

Table 11. List of srWGS SNP & Indel related samples to prune TSV file description

Field name	Type	Key?	Notes
sample_id.s	string	Yes	Research ID of the sample

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type. Arrays are possible.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

HLA variant calling

We provide haplotype phasing data for all samples in the CDRv8 srWGS callset. The data is delivered as multi-sample VCFs, sharded by chromosome. Haplotype phasing is the estimation of haplotypes that are inherited from each parent. We use statistical methods to infer the sequence of alleles on each chromosome, following methods from the 2021 paper from Browning, Brian L. et al.: Fast two-stage phasing of large-scale sequence data.

To generate the phasing data, we performed focused QC on the srWGS VDS. We first removed variants with more than 31 alternate alleles. We then removed variants with an average sum of their AD less than 28, removed variants with a max AC less than 2 (to remove singletons and doubletons), removed variants with a mean GQ of less than 30, and variants with FILTER values of LowQual, NO_HQ_GENOTYPES, or ExcessHet.

We used Beagle 5.5 for phasing with a window parameter of 15, 20, or 40, depending on the chromosome. All other parameters were default. Genetic distances were interpolated from the HapMap genetic map.

Flagged srWGS samples

We provide a table listing samples that are flagged as part of the sample outlier QC for the srWGS SNP and Indel joint callset. This includes the specific residual tests that were failed.The schema is described in Table 12. The table will be released as a tsv.

Flagged sample tsv schema

- No fields can have a null value.
- Count fields do not include filtered variants.
- For all of the `fail_*` fields, a value of `true` *indicates that the sample is an outlier* and should be flagged.

Table 12. Flagged srWGS samples TSV file description

Field Name	Type	Key?	Example Value	Notes
s	int	yes	1000000	Research ID
ancestry_pred	string	no	eur	The predicted ancestry for the sample, not including “other.”
probabilities	array<float>	no	[0.10, 0.99, 0.001, … 0.0]	Confidence of each output class (i.e. computed ancestry). Each will have a length equal to the number of possible computed ancestry labels minus one (6). The ancestry “Other” is computed separately based on the confidence of the other classes.

pca_features	array<float>	no	[8.1232, 0.01234, 3.1123, …, 0.00132]	Each will have a length of 16.
ancestry_pred_other	string	no	oth	The predicted ancestry for the sample, including “other.”
snp_count	int	no	3910035	Number of SNPs called in this sample.
ins_del_ratio	float	no	0.98814	Ratio of insertion to deletion counts.
del_count	int	no	427102	
ins_count	float	no	456515	
snp_het_homvar_ratio	float	no	2.1119	
indel_het_homvar_ratio	float	no	2.3994	
ti_tv_ratio	float	no	1.9967	
singleton	int	no	15819	IMPORTANT: This is not the number of singletons in a sample. This field is a count of the number of variants not appearing in gnomAD 3.1.
fail_snp_count_residual	boolean	no	true	
fail_ins_del_ratio_residual	boolean	no	false	
fail_del_count_residual	boolean	no	true	
fail_ins_count_residual	boolean	no	false	
fail_snp_het_homvar_ratio_residual	boolean	no	true	
fail_indel_het_homvar_ratio_residual	boolean	no	false	
fail_ti_tv_ratio_residual	boolean	no	true	
fail_singleton_residual	boolean	no	false	
qc_metrics_filters	array<string>	no	["indel_het_homvar_ratio_residual","snp_count_residual"]	A list of each failed test. These will correspond to all fail_* fields with a value of “true.”

srWGS genomic QC values

We provide the QC testing values for all srWGS samples, which is a duplicate of the srWGS flagged sample schema, but for all srWGS samples. The table is released as a tsv and corresponds to the schema in Table 12.

srWGS genomic metrics

We provide a table with supplemental genomic QC metrics for each srWGS sample. The schema is described in Table 13. The table will be released as a tsv.

Genomic metrics tsv schema

- No fields can have a null value.
- No samples will be in the table if they do not pass the QC thresholds.

Table 13. Supplemental genomic metrics for each srWGS sample TSV file description

Field Name	Type	Key?	Example Value	Notes
research_id	int	yes	1000000	Unique identifier for each participant
sample_source	string	no	Whole Blood	Sample source (blood or saliva)
site_id	string	no	bi	The genome center (GC) where the sample was sequenced. This will be one of three values (bi = "Broad Institute", uw = "University of Washington", or bcm = "Baylor College of Medicine")
sex_at_birth	string	no	Female	Participant provided information for sex at birth
dragen_sex_ploidy	string	no	XX	Ploidy output from DRAGEN
mean_coverage	float	no	107.69	Mean number of overlapping reads at every targeted base of the genome (threshold ≥30x)
genome_coverage	float	no	97.61	Percent of bases with at least 20x coverage (threshold ≥90% at 20x)
aou_hdr_coverage	float	no	100	Percent of bases in the All of Us Hereditary Disease Risk gene (AoUHDR) with at least 20x coverage (threshold ≥95% at 20x)
dragen_contamination	float	no	0.003	Cross-individual contamination rate from DRAGEN
aligned_q30_bases	float	no	174329894399	Aligned Q30 bases from DRAGEN (threshold ≥8e10)
verify_bam_id2_conta	float	no	0.0000104116	Cross-individual contamination

mination				rate from VerifyBamID2
biosample_collection_date	string	no	2/11/2020	Date when biosamples were collected. This reflects the date the collection site finalized the order, which is generally close to, but may not exactly match, the actual time of collection.

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

srWGS control samples

We provide GVCF files for the eight public control samples that were used for the srWGS SNP and Indel sensitivity and precision evaluation (see Table F.1 in the *All of Us* Genomic Data Quality Report). The samples come from Genomes-in-a-Bottle (GiaB): from The International HapMap Project and Personal Genome Project. The samples were sequenced with the same protocol as the srWGS *All of Us* samples and are provided for researchers to use for their own QC processes and analyses. The data is provided in GVCF format along with their index files.

Structural variants (SVs) for srWGS data

Table 14. Short-read WGS SV deliverables

Deliverable	srWGS SNP & Indel
Reference version	hg38/GRCh38 reference: gs://genomics-public-data/references/hg38/v0/Homo_sapiens_assembly38.fasta
Raw data	CRAM files (same deliverable as srWGS SNP & Indels)
Variant data	Joint-called VCF Sites-only VCF Unrelated sites-only VCF
Auxiliary files	Ancestry and relatedness available for srWGS samples based on the srWGS SNP & Indel deliverables Maximal set of unrelated samples Samples with probable aneuploidies srWGS SV sample list

We provide structural variant (SV) calls for 97,061 participants with srWGS data. The SV dataset includes a standard VCF with genotypes, a sites-only VCF, a list of the maximal set of unrelated samples, a sites-only VCF containing annotations from the maximal set of unrelated samples, and lists of the samples with probable aneuploidies. Please read more information about the SV calls and pipeline in the *All of Us* Genomic Data Quality Report.

srWGS SV VCF

The SVs are joint-called and delivered as a joint VCF for all samples, a sites-only VCF, and a sites-only VCF with annotations for the maximal set of unrelated samples. The VCFs are sorted and block and block compressed (.vcf.gz) with a local tabix index (.vcf.gz.tbi).

The full VCF has genotypes for all 97,061 participants and is sharded by chromosome.

The GATK-SV team has documented the SV VCF format in an article on the GATK site: How to interpret SV VCFs. The format has many similarities to a short variant VCF but you will see some differences that are necessary to specify SV variant details. The header describes the data fields in the VCF.

The SV VCF is annotated with the GATK tool SVAnnotate. It adds the gene overlap and the predicted functional consequence. These annotations are added in the INFO field. The annotations produced by SVAnnotate are described in detail in the tool documentation. The GTF used for gene annotations was GENCODE v39.

Some of the most important fields in the VCF are described below:

- **CHROM:** The chromosome location of the start position of the SV
- **POS:** The start position of the SV
- **ID:** Unique identifier for the SV
- **REF:** Not commonly used in structural variant VCFs, commonly has an N
- **ALT:** Information about the SV type, descriptions of the SV types can be found in the header
- **FILTER:** Filtering information for the SV
    - HIGH_NCR: Unacceptably high rate of no-call GTs.
    - MULTIALLELIC: Multiallelic CNV site. This FILTER status does not mean that the site is not real, but it should be treated differently from a biallelic SV site.
    - UNRESOLVED: Variant is unresolved. There was some evidence for an SV at this site but it was not able to be resolved completely from the available evidence.
    - VARIABLE_ACROSS_BATCHES: Site appears at variable frequencies across batches. Likely reflects technical batch effects.
    - PASS: None of the above site-level filters were applied.
- **INFO:** Annotations describing the variant at the site level. The annotations are described in depth in the SV VCF header. Some of these annotations include:
    - END: End position of the structural variant
    - CHR2: Second chromosome for interchromosomal events
    - END2: Position of breakpoint on CHR2
    - ALGORITHMS: The original algorithm that called the SV (GATK-SV is an ensemble method)
    - SVLEN: SV length in base pairs
    - SVTYPE: SV type
    - CPX_TYPE: Subtype of complex rearrangement

    - CPX_INTERVALS: Details of complex rearrangement
- **FORMAT:** Annotations describing the variant at the genotype level (site- and sample-specific annotations). Depends on the SV type and the evidence categories that support the SV. All FORMAT annotations are described in the VCF header.

srWGS SV sites-only VCF

The sites-only VCF contains all of the sites and site-level annotations in the full VCF but no genotype information. It is useful as a smaller file when genotype information is not required. See the above information for SV VCF details.

srWGS SV maximal set of unrelated samples

We provide a list of samples to prune in order to remove related samples from the srWGS SV cohort. Relatedness is calculated as described in the kinship score description above. This will be the minimal list of related samples to prune in order to produce the maximal independent set of unrelated samples.

The samples are reported in a txt file as a list of research IDs. One research ID is listed per line and there is no header in the file.

srWGS SV unrelated sites-only VCF

We provide a sites-only VCF, containing no genotype information, with annotations for the maximal set of 93,360 unrelated samples. We removed from the complete VCF the 3,701 samples from the above list of samples to prune to obtain the maximal set of unrelated samples. Sites that were unique to the removed samples were removed from the VCF. We re-annotated allele frequencies in the VCF based on the remaining samples. This VCF is provided in order to save researchers computational time for analyses requiring unrelated samples.

srWGS SV samples with probable aneuploidies

We provide lists of samples with probable aneuploidies identified during srWGS SV ploidy estimation as tsv files. Ploidy estimation was performed across the srWGS SV samples using coverage estimations over binned regions of the genome as part of the GATK-SV pipeline. Details of this ploidy estimation process are described in the *All of Us* Genomic Data Quality Report.

There are three separate files for samples with probable aneuploidies: mosaic autosomal aneuploidy, mosaic allosomal aneuploidy, and germline allosomal aneuploidy.

srWGS SV samples with probable mosaic aneuploidies

We provide two files describing samples with probable mosaic aneuploidies. One is samples with mosaic autosomal aneuploidy and the second is samples with mosaic allosomal aneuploidy. Both files have the same format, described in Table 14. Note that fewer than 20

samples had more than one probable mosaic autosomal aneuploidy, so these samples appear once per affected chromosome.

Table 15. srWGS SV samples with probable mosaic aneuploidies TSV file description

Field name	Type	Key?	Notes
research_id	string	no	Research ID of the sample
chromosome	string	no	Chromosome for which the sample is predicted to have a mosaic aneuploidy, ie. chr8 or chrX
estimated_copy_ratio	float	no	Estimated copy ratio (see QC report Ploidy Estimation) for the chromosome with the probable mosaic aneuploidy
aneuploidy_type	string	no	Type of aneuploidy predicted. For the probable mosaic aneuploidies, the possible values are MOSAIC_GAIN or MOSAIC_LOSS

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

srWGS SV samples with probable germline allosomal aneuploidy

We provide one file describing samples with probable germline allosomal aneuploidies, described in Table 15.

Table 16. srWGS SV samples with probable germline allosomal aneuploidy TSV file description

Field name	Type	Key?	Notes
research_id	string	yes	Research ID of the sample
copy_number_chrX	integer	no	Estimated copy number for chrX, rounded to the nearest integer
copy_number_chrY	integer	no	Estimated copy number for chrY, rounded to the nearest integer
aneuploidy_type	string	no	Type of aneuploidy predicted. For the probable germline allosomal aneuploidies, the possible values are JACOBS, KLINEFELTER, and TRIPLE X (contains a space)

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

srWGS SV sample list

We provide a list file of all research_ids that have srWGS SV data. The file is a text file containing one research_id per line.

Genotyping Array ("Array") Data

The array data represents 447,278 participants and includes single sample VCFs, joint Hail MT files, joint PLINK files, and raw genotyping data in IDAT format.

Table 17. Short-read WGS deliverables

Deliverable	srWGS SNP & Indel
Reference version	hg38/GRCh38 reference Note: variants are called originally with hg19 reference but they are lifted over before release on RW
Raw data	IDAT files
Variant data	Single sample VCFs (all VCFs have the same variants) Hail MT (merged) PLINK bed files (merged)
Auxiliary files	Ancestry and relatedness available for array samples that have srWGS data

Array IDAT files

We provide IDAT files for all array samples. The IDAT file is a binary file containing raw BeadArray data directly from the scanner. There are two files for each sample, corresponding to the red and green intensity values. These values give information about specific nucleotides on the genome. You can read more about the steps to call variants from these IDAT files in the *All of Us* Genomic Data Quality Report.

For an in depth description and how to process these files, read more about the illuminaio tool.

Array variant data

The variant data for array samples is delivered in VCF, Hail MT, and PLINK format.

Array VCFs

We provide single-sample VCFs for all 447,278 participants with array data. The array VCFs are sorted and block compressed (vcf.gz) with local tabix index files (vcf.gz.tbi).

Array VCFs in the *All of Us* genomic dataset will contain the following:

Header

The header field of the VCF contains many attributes which generally describe the processing of the sample in the array. Many of these are specific to a single sample.

- arrayType - This contains the name of the genotyping array that was processed.
- autocallDate - The date that the genotyping array was processed by 'autocall' (aka gencall), the Illumina genotype calling software.
- autocallGender - The gender (sex) that autocall determined for the sample processed.
- autocallVersion - The version of the autocall/gencall software used.
- chipWellBarcode - The chip well barcode (a unique identifier for sample as processed on a specific location on the Illumina genotyping array).
- clusterFile - The cluster file used.
- extendedIlluminaManifestVersion - The version of the 'extended Illumina manifest' used by the VCF generation software.
- extendedManifestFile - The filename of the 'extended Illumina manifest' used by the VCF generation software.
- fingerprintGender - The gender (sex) determined using an orthogonal fingerprinting technology. This is populated by an optional parameter used by the VCF generation software.
- gtcCallRate - The gtc call rate of the sample processed. This value is generated by the autocall/gencall software and represents the fraction of callable loci that had valid calls.
- imagingDate - The date that the IDAT files (raw image scans) for the chip well barcode were created.
- manifestFile - The name of the Illumina manifest (.bpm) file used by the VCF generation software.
- sampleAlias - The name of the sample.

Note that there are many other attributes in the header (Biotin*, DNP*, Extension*, Hyb*, NP*, NSB*, Restore, String*, TargetRemoval) that are populated with Illumina control values. They are not described here.

INFO fields (per site)

These fields describe attributes specific to the probe on an array. The INFO specifier in the VCF header describes these fields. They are:

- AC - Allele Count in genotypes, for each ALT allele. A standard field, described in the VCF specification
- AF - Allele Frequency. A standard field, described in the VCF specification
- AN - Allele Number. A standard field, described in the VCF specification
- ALLELE_A - The A Allele, as annotated in the Illumina manifest (a *suffix indicates this is the reference allele)
- ALLELE_B - The B Allele, as annotated in the Illumina manifest (a *suffix indicates this is the reference allele)
- BEADSET_ID - The BeadSet ID. An Illumina identifier. Used for normalization.
- GC_SCORE - The Illumina GenTrain Score. A quality score describing the probe design

- ILLUMINA_BUILD - The Genome Build for the design probe sequence, as annotated in the Illumina manifest
- ILLUMINA_CHR - The chromosome of the design probe sequence, as annotated in the Illumina manifest.
- ILLUMINA_POS - The position of the design probe sequence (on ILLUMINA_CHR), as annotated in the Illumina manifest.
- ILLUMINA_STRAND - The strand for the design probe sequence, as annotated in the Illumina manifest.
- PROBE_A - The allele A probe sequence as annotated in the Illumina manifest.
- PROBE_B - The allele B probe sequence as annotated in the Illumina manifest. Note that this is only present on strand ambiguous SNPs.
- SOURCE - The probe source as annotated in the Illumina manifest.
- refSNP - The dbSNP rsId for this probe

FILTER values (per site):

There are several filters specific to genotyping array content. These are:

- DUPE - This filter is applied if there are multiple rows in the VCF for the same loci and alleles. That is, if there are two or more rows that share the same chromosome, position, ref allele and alternate alleles, all but one of them will have the 'DUPE' filter set.
- TRIALLELIC - This filter is applied if there is a site at which there are two alternate alleles and neither of them is the same as the reference allele.
- ZEROED_OUT_ASSAY - This filter is applied if the variant at the site was 'zeroed out' in the Illumina cluster file - this is typically done when the calls at the site are intentionally marked as unusual. Genotypes called sites that are 'zeroed out' will always be no-calls.

FORMAT fields (per sample-site)

These fields describe attributes specific to the sample genotyped on the array. The FORMAT specifier in the VCF header describes these fields. They are:

- GT - GENOTYPE. This field describes the genotype. It is a standard field, described in the VCF specification.
- IGC - Illumina GenCall Confidence Score. A measure of the call confidence.
- X - Raw X intensity as scanned from the original genotyping array
- Y - Raw Y intensity as scanned from the original genotyping array
- NORMX - Normalized X intensity
- NORMY - Normalized Y intensity
- R - Normalized R Value (one of the polar coordinates after the transformation of NORMX and NORMY)
- THETA - Normalized Theta value (one of the polar coordinates after the transformation of NORMX and NORMY)
- LRR - Log R Ratio
- BAF - B Allele Frequency

Array Hail MT

We have merged the array VCFs into a Hail MT with no additional processing across samples. Each column corresponds to the research ID of the sample and each row corresponds to the variant. Since the single sample array VCFs have identical sites and FILTER values, the FILTER field is populated with the value from a single sample VCF.

In conversion, we have dropped all of the 505 variants from alternate, unlocalized, and unplaced contigs (436 variants from ALT contigs (e.g. chr19_KI270866v1_alt), 72 from random contigs (e.g. chr1_KI270706v1_random), and 13 from chrUn (e.g. chrUn_KI270742v1). These variants are still in the compressed array VCFs. Please refer to the published Featured Workspaces on how we generated the Hail MT from the VCFs.

Array PLINK 1 binary biallelic genotype table (PLINK bed)

We provide PLINK 1.9 data (.bed / .bim / .fam) for array data, converted from the Hail MT using the export_plink command in Hail and contain all information in the Hail MT. PLINK file type information can be found within the PLINK documentation. The .bed file is the PLINK binary biallelic genotype table and contains genotype calls. The .bim file is the PLINK extended .map file, and is a text file containing variant information. The .fam file is a text file with sample information for each participant. Please refer to the published Featured Workspaces on how to use the PLINK files.

Long-Read Whole Genome Sequencing (lrWGS)

Table 18. Long-read WGS deliverables

Deliverable	lrWGS SNP & Indel
Reference version	T2Tv2.0 grch38_noalt
Raw data	BAM files for each reference version *De novo* assembly for PacBio cohorts: primary, alternate, and two chromosome copies in Graphical Fragment Assembly (GFA) format and FASTA format
Variant data	Joint SNP & Indel variants (GVCF & Hail MT format) Single-sample SNP & Indel variants (GVCF) Single-sample PBSV SVs (VCF) Single-sample Sniffles2 SVs (VCF & SNF) Single-sample PAV variants (VCF) - for PacBio cohorts
Auxiliary files	Ancestry and relatedness available for lrWGS samples based on the srWGS SNP & Indel deliverables lrWGS variant metrics lrWGS flagged samples

We provide lrWGS data representing 14,521 participants. These data are particularly useful for resolving complex genomic regions, structural variants, and phasing of alleles, to provide a

more comprehensive view of the genome. The lrWGS data represented multiple cohorts with different sequencing facilities and platforms (Table 16).

For each cohort, there are different data available depending on how we analyzed each sample (Table 17). There have been three data releases with lrWGS data. Within the various cohorts and data releases, there are varying deliverables for each set of samples. The CDRv9 cohort contains 14,433 samples with 13,530 participants. The CDRv8 cohort contains 1,773 participants with 1815 samples, because 42 participants were sequenced at two different facilities. The CDRv7 cohort contains 1027 samples and 1027 participants.

Table 19. Sample cohorts for all 14,521 participants with lrWGS data

Cohort name	Initial release version	Sequencing facility	Sequencing platform	Number of samples	Minimum coverage
HA_PacBio	V9	HA	PacBio Revio	1,210	Mid-pass (12x)
BI_PacBio	V9	BI	PacBio Sequel IIe, PacBio Revio	9,934	Mid-pass (12x)
BCM_PacBio	V9	BCM	PacBio Sequel IIe, PacBio Revio	748	High-pass (25x)
UW_PacBio	V9	UW	PacBio Sequel IIe, PacBio Revio	397	High-pass (25x)
BCM_ONT	V9	BCM	ONT R10.4 on PromethION	1,031	High-pass (25x)
JHU_ONT	V9	JHU	ONT R10.4 on PromethION	717	High-pass (25x)
UW_ONT	V9	UW	ONT R10.4 on PromethION	272	High-pass (25x)
UW_ONT_R9	V9	UW	ONT R9.4 on PromethION	124	High-pass (25x)
HA_Rev_mid	V8	HA	PacBio Revio	65	Mid-pass (12x)
BI_Seq_high	V8	BI	PacBio Sequel IIe	84	High-pass (25x)
BI_Seq_mid	V8	BI	PacBio Sequel IIe	198	Mid-pass (12x)
BI_Rev_mid	V8	BI	PacBio Revio	803	Mid-pass (12x)
BCM_Seq_high	V8	BCM	PacBio Sequel IIe	77	High-pass (25x)
BCM_Rev_high	V8	BCM	PacBio Revio	111	High-pass (25x)
BCM_ONT_high	V8	BCM	ONT R10.4 on PromethION	196	High-pass (25x)

JHU_ONT_high	V8	JHU	ONT R10.4 on PromethION	128	High-pass (25x)
UW_Seq_high	V8	UW	PacBio Sequel IIe	100	High-pass (25x)
UW_Rev_high	V8	UW	PacBio Revio	53	High-pass (25x)
HA_Seq_CDRv7	V7	HA	PacBio Sequel IIe and Sequel II	1027	Mid-pass (8x)
Total number of participants				14,521	
Total number of samples				17,275	

The data available for each CDRv9 cohort includes sequencing reads for grch38_noalt as a BAM file with methylation signals and various small variant and SV deliverables. Please see Table 17 for more information. For previous releases, see the section LrWGS data available for previous releases. Additionally, there are auxiliary metrics generated for each sequencing location. All samples sequenced at that sequencing location are represented on a per-sample level in the auxiliary metrics.

PacBio samples additionally have a de novo assembly available along with joint-called SNP and Indel variants and single-sample PAV variants. Samples from one sequencing center, UW, are aligned to T2Tv2.0.

For locations of the lrWGS files available, please see the lrWGS manifest and the CDR Directory Document.

Table 20. Data available for each lrWGS cohort

Cohort name	Sequencing reads	Variant data	Auxiliary data
All CDRv9 data	grch38_noalt BAM with methylation signals	Single sample SNP & indel variants (GVCF) Single sample PBSV SVs (VCF) Single sample Sniffles2 SVs (VCF) Single sample Sniffles2 SNF	Auxiliary metrics grch38_noalt
PacBio data	*All above data plus:* *De novo* assembly GFA files: primary *de novo* assembly, alternative *de novo* assembly, one *de novo* assembly for each chromosome copy FASTA: one for each GFA file	*All above data plus:* Joint-called SNP & Indel variants (GVCF & Hail MT) Joint-called SV data Single sample PAV variants (VCF)	

Data from UW	*First row of data plus:* T2Tv2.0 BAM with methylation signals		

lrWGS sequencing reads

Each sample in the lrWGS data is aligned to grch38_noalt in BAM format. Each BAM file is accompanied by an index BAI file. The samples from the UW sequencing center are also aligned to T2Tv2.0.

grch38_noalt corresponds to the GRCh38 reference with no alternate sequences. T2Tv2.0 in the CDRv8 release corresponds to the T2T-CHM13v2.0 reference with these modifications: 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).

Methylation signals

Methylation data are available in the long-read BAM files for both reference versions across all PacBio and ONT cohorts (Table 17). DNA methylation occurs in various forms, with 5-methylcytosine (5mC) being the predominant type in adult human genomes. This process involves enzymes adding a methyl group to a cytosine (C) at CpG sites—regions where a cytosine is immediately followed by a guanine (G). Methylation at these sites can influence gene transcription.

In the BAM files, methylation sites are annotated using MM and ML tags. The MM tag is binary, indicating whether a site is methylated. The ML tag provides a confidence score for the methylation status assigned by the caller. Note that while most reads include these methylation calls, some do not. To interpret the methylation data, see the PacBio BAM format specification and the SAM optional tags specifications. The documentation applies to both PacBio and ONT data.

lrWGS *de novo* assembly

Haplotype-resolved *de novo* assembly is available for all PacBio HiFi samples (Table 16) in Graphical Fragment Assembly (GFA) and FASTA format. Each *de novo* assembly includes a primary *de novo* assembly, an alternative *de novo* assembly, and two chromosome copies. The tool PAV is used to call variants from the PacBio GFA files.

GFA files

We release four Graphical Fragment Assembly (GFA) files for each PacBio sample sample, which are *de novo* graph-based assemblies. One GFA file is the primary assembly for the sample, another being the alternative assembly, and the other two GFA files are the chromosome copy assemblies. The GFA files describe the graph layouts of the contigs.

We use the tool hifiasm, which is a tool for generating haplotype-resolved de novo assemblies. Please check out the GFA specifications for more details about GFA format.

FASTA files

We provide four de novo assemblies as FASTA files for each PacBio HiFi lrWGS sample, matching the sequences from the GFA files. A FASTA file is a text file representation of genomic data. Each genomic sequence is described in two lines: the first line is a description line starting with a greater-than (">") symbol at the beginning and the second line contains the genomic sequence data as a string with the nucleotide sequence. Other than the first line of the FASTA file which is the description, these two lines representing genomic sequences are repeated in the file.

Long-read variant data

For a detailed description of the previous versions of lrWGS variant data, please see CDRv8 How the *All of Us* Genomic Data are Organized or the CDRv7 How the All of Us Genomic Data are Organized. There are featured workspaces containing joint callsets for BCM and ONT data, which will be posted at a later date.

lrWGS SNP & Indel GVCF

We perform SNP and Indel variant calling per-sample with DeepVariant. The single-sample variant data is released in GVCF format with accompanying GVCF TBI index files.

lrWGS joint callset Hail MT

We generate a lrWGS joint SNP & Indel callset by joining the single-sample GVCFs with GLNexus. The joint callsets are generated for all PacBio samples (Table 16 for the sample counts). The joint callset is available in Hail MT and GVCF format. The variants are hard-filtered with a QUAL cutoff of 40 (see the CDRv7 *All of Us* Genomic Data Quality Report for more information).

lrWGS structural variant VCFs

Structural variants are called from both PBSV and Sniffles2 for all lrWGS samples. Each lrWGS sample has a single VCF from each of the two variant callers, accompanied by TBI index files. Please see the headers of these VCF files for descriptions of the VCF fields. In addition, we output a Sniffles2 binary SNF file for use with Sniffles2's multi-sample SV calling mode.

lrWGS PAV phased variants

Variants from the tool PAV are provided in VCF format for each PacBio HiFi sample. The VCF files are accompanied by a TBI index. PAV variants are derived from the haplotype resolved

assembly (GFA files) generated by hifiasm. PAV-generated VCFs are phased. Please see the header of the PAV VCFs for a description of the VCF fields.

lrWGS auxiliary metrics

We provide two lrWGS variant metrics files, corresponding to each lrWGS reference, described in Table 18. To uniquely identify a sample, you need the combination of the sample_id, sequencing facility, and platform.

Table 21. lrWGS metrics file description

Field name	Type	Key?	Notes
sample_id	string	yes	Research ID of the sample
center	string	yes	The sequencing facility of the sample. Possible values are: BCM, BI, HA, JHU, UW.
platform	string	yes	Sequencing technology of the sample. Possible values are revio, sequel, ont
mosdepth_cov	float	no	Coverage from the mosdepth tool (See the QC report for a description)
aligned_frac_bases	float	no	Fraction of bases aligned to the reference
aligned_num_bases	float	no	Number of bases aligned to the reference
aligned_num_reads	float	no	Number of reads aligned to the reference
aligned_read_length_N50	float	no	N50 of the aligned reads length
aligned_read_length_median	float	no	Median length of the aligned reads
aligned_read_length_mean	float	no	Mean length of the aligned reads
aligned_read_length_stdev	float	no	Standard deviation of the aligned read length
average_identity	float	no	Mean percentage of matches to the reference per aligned read
median_identity	float	no	Median percentage of matches to the reference per aligned read
pbsv_nonBND_50bpSV_cnt	float	no	Number of SVs >= 50 bp called by PBSV (excluding break-end calls)
sniffles_nonBND_50bpSV_cnt	float	no	Number of SVs >= 50 bp called by Sniffles2 (excluding break-end calls)

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type.

- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

lrWGS flagged samples

As described in the QC doc, several lrWGS samples were flagged during the QC process, but not filtered. We release a separate file—a 4-column TSV—listing the samples that have been flagged. To uniquely identify a sample, you need the combination of the sample_id, sequencing facility, and platform.

Table 22. lrWGS flagged samples

Field name	Notes
sample_id	Research ID of the sample
sequencing_facility	The sequencing facility of the sample. Possible values are: BCM, BI, HA, JHU, UW.
platform	Sequencing technology of the sample. Possible values are revio, sequel, ont
reasons_for_flagging	The reasons for the sample to be flagged. There can be more than one reason for the sample to be flagged, separated by comma. No white spaces. Possible values: contamination_between_1_and_3_pct, coverage_slightly_below_target, diploid_assembly_length_anomaly, female_with_low_chrX_coverage, male_with_low_chrY_coverage, read_len_median_below_10kbp

lrWGS manifest

The location of each single sample file is listed in the lrWGS manifest file. This resource goes hand in hand with the Controlled CDR Directory Document, which lists the location of the manifest file and the paths for all joint callsts. Some samples will have two rows in the lrWGS manifest because they were sequenced at multiple sequencing facilities or on multiple platforms. To uniquely identify a sample, you need the combination of the sample_id, center, and platform.

Not all columns will be filled, depending on what data is available for the sample. See Table 17 for details.

Table 23. lrWGS manifest

Column name	Notes
research_id	Research ID of the sample
center	Sequencing facility of the sample. Possible values are: BCM, BI, HA, JHU, UW.

platform	Sequencing technology of the sample. Possible values are revio, sequel, ont
assembly_alternate_fa	*De novo* alternate assembly, in FASTA format. Only available for PacBio samples. CDRv8 files are block-gzipped and CDRv7 files are gzipped.
assembly_alternate_fa_gzi	*De novo* alternate assembly, FASTA index file. Only available for PacBio samples in CDRv8.
assembly_alternate_gfa	*De novo* alternate assembly, in GFA format. Only available for PacBio samples.
assembly_hap1_fa	*De novo* haplotype-resolved assembly for haplotype-1 (in no particular order), in FASTA format. Only available for PacBio samples. CDRv8 files are block-gzipped and CDRv7 files are gzipped.
assembly_hap1_fa_gzi	*De novo* haplotype-resolved assembly for haplotype-1 (in no particular order), FASTA index file. Only available for PacBio samples in CDRv8.
assembly_hap1_gfa	*De novo* haplotype-resolved assembly for haplotype-1 (in no particular order), in GFA format. Only available for PacBio samples.
assembly_hap2_fa	*De novo* haplotype-resolved assembly for haplotype-2 (in no particular order), in FASTA format. Only available for PacBio samples. CDRv8 files are block-gzipped and CDRv7 files are gzipped.
assembly_hap2_fa_gzi	*De novo* haplotype-resolved assembly for haplotype-2 (in no particular order), FASTA index file. Only available for PacBio samples in CDRv8.
assembly_hap2_gfa	*De novo* haplotype-resolved assembly for haplotype-2 (in no particular order), in GFA format. Only available for PacBio samples.
assembly_primary_fa	*De novo* primary assembly, in FASTA format. Only available for PacBio samples. CDRv8 files are block-gzipped and CDRv7 files are gzipped.
assembly_primary_fa_gzi	*De novo* primary assembly, FASTA index file. Only available for PacBio samples in CDRv8.
assembly_primary_gfa	*De novo* primary assembly, in GFA format. Only available for PacBio samples.

assembly_quast_report_html	HTML report for the primary, haplotype-1 and haplotype-2 assemblies generated by the QUAST program. Only available for CDRv7 PacBio samples.
assembly_quast_report_summary	A summary about the quality of the primary, haplotype-1 and haplotype-2 assemblies, reported by the QUAST program. Only available for PacBio samples.
chm13v2.0_bai	The accompanying index for the T2Tv2.0 BAM.
chm13v2.0_bam	T2Tv2.0 sequencing reads in BAM format
chm13v2.0_bam_pbi	The accompanying PBI index for the T2Tv2.0 BAM. Only available for CDRv7 samples.
chm13v2.0_deepvariant_phased_tbi	TBI index for the T2Tv2.0 PEPPER-Margin-DeepVariant phased VCF. Only available for CDRv7 samples.
chm13v2.0_deepvariant_phased_vcf	T2Tv2.0 PEPPER-Margin-DeepVariant phased single-sample VCF; a filter of QUAL<40 has been applied. Only available for CDRv7 samples.
chm13v2.0_deepvariant_tbi	TBI index for the T2Tv2.0 PEPPER-Margin-DeepVariant VCF. Only available for CDRv7 samples.
chm13v2.0_deepvariant_vcf	T2Tv2.0 PEPPER-Margin-DeepVariant single-sample VCF; a filter of QUAL<40 has been applied. Only available for CDRv7 samples.
chm13v2.0_dv_gtbi	TBI index for the DeepVariant T2Tv2.0 GVCF. Only available for CDRv8 samples.
chm13v2.0_dv_gvcf	T2Tv2.0 DeepVariant single-sample SNP & Indel GVCF. Only available for CDRv8 samples.
chm13v2.0_haplotagged_bai	T2Tv2.0 haplotagged BAM index. Only available for CDRv7 samples.
chm13v2.0_haplotagged_bam	T2Tv2.0 haplotagged BAM. Only available for CDRv7 samples.
chm13v2.0_pav_tbi	TBI index for the T2Tv2.0 PAV VCF. Only available for the PacBio samples.
chm13v2.0_pav_vcf	T2Tv2.0 PAV single-sample VCF. Only available for the PacBio samples.
chm13v2.0_pbsv_tbi	TBI index for the T2Tv2.0 PBSV SV single-sample VCF
chm13v2.0_pbsv_vcf	T2Tv2.0 PBSV SV single-sample VCF
chm13v2.0_sniffles_snf	T2Tv2.0 Sniffles2 single-sample SNF file
chm13v2.0_sniffles_tbi	TBI index for the T2Tv2.0 Sniffles2 VCF file

chm13v2.0_sniffles_vcf	T2Tv2.0 Sniffles2 single-sample VCF file
grch38_bai	The accompanying index for the grch38_noalt BAM.
grch38_bam	grch38_noalt sequencing reads in BAM format
grch38_bam_pbi	The accompanying PBI index for the grch38_noalt BAM. Only available for CDRv7 samples.
grch38_deepvariant_phased_tbi	TBI index for the grch38_noalt PEPPER-Margin-DeepVariant phased VCF. Only available for CDRv7 samples.
grch38_deepvariant_phased_vcf	grch38_noalt PEPPER-Margin-DeepVariant phased single-sample VCF; a filter of QUAL<40 has been applied. Only available for CDRv7 samples.
grch38_deepvariant_tbi	TBI index for the grch38_noalt PEPPER-Margin-DeepVariant VCF. Only available for CDRv7 samples.
grch38_deepvariant_vcf	grch38_noalt PEPPER-Margin-DeepVariant single-sample VCF; a filter of QUAL<40 has been applied. Only available for CDRv7 samples.
grch38_dv_gtbi	TBI index for the DeepVariant grch38_noalt GVCF. Only available for CDRv8 samples.
grch38_dv_gvcf	grch38_noalt DeepVariant single-sample SNP & Indel GVCF. Only available for CDRv8 samples.
grch38_haplotagged_bai	grch38_noalt haplotagged BAM index. Only available for CDRv7 samples.
grch38_haplotagged_bam	grch38_noalt haplotagged BAM. Only available for CDRv7 samples.
grch38_pav_tbi	TBI index for the grch38_noalt PAV VCF. Only available for the PacBio samples.
grch38_pav_vcf	grch38_noalt single-sample PAV VCF. Only available for the PacBio samples.
grch38_pbsv_tbi	TBI index for the grch38_noalt PBSV SV VCF
grch38_pbsv_vcf	grch38_noalt PBSV SV single-sample VCF
grch38_sniffles_snf	grch38_noalt Sniffles2 single-sample SNF file
grch38_sniffles_tbi	TBI index for the grch38_noalt Sniffles2 VCF file
grch38_sniffles_vcf	grch38_noalt single-sample Sniffles2 VCF file

LrWGS data available for previous releases

The general content in this report is focused on the CDRv9 lrWGS released data. For cohorts from CDRv7, please see descriptions in the CDRv7 version of the article How the *All of Us* Genomic Data are Organized. The files available for each CDRv7 sample are listed in Table 17. The major differences are as follows:

- CDRv7 has haplotagged BAM files available.
- In CDRv7, the single sample SNP & Indel variants were called with the PEPPER-MARGIN-DeepVariant pipeline.

For cohorts from CDRv8, please see the descriptions in the CDRv8 version of the article How the *All of Us* Genomic Data are Organized. The files available for each CDRv8 sample are listed in Table 17. The major differences are as follows:

- CDRv8 has methylation signals in the BAM files
- For PacBio, each BAM file is available for grch38_noalt and T2T
- The CDRv8 joint callsets are broken up into smaller cohorts.

Table 24. Data available for each CDRv7 and CDRv8 lrWGS cohort

Cohort name	Sequencing reads	Variant data	Auxiliary data
PacBio V8 cohorts	grch38_noalt BAM with methylation signals T2Tv2.0 BAM with methylation signals GFA files: primary *de novo* assembly, alternative *de novo* assembly, one *de novo* assembly for each chromosome copy FASTA: one for each GFA file	*One for each grch38_noalt & T2Tv2.0:* Joint-called SNP & Indel variants (GVCF & Hail MT) Single sample SNP & indel variants (GVCF) Single sample PBSV SVs (VCF) Single sample Sniffles2 SVs (VCF) Single sample Sniffles2 SNF Single sample PAV variants (VCF)	*Single sample data available in the the file per sequencing facility:* Auxiliary metrics grch38_noalt Auxiliary metrics T2Tv2.0
ONT V8 cohorts	grch38_noalt BAM with methylation signals T2Tv2.0 BAM with methylation signals	*One for each grch38_noalt & T2Tv2.0:* Joint-called SNP & Indel variants (GVCF & Hail MT) Single sample SNP & indel variants (GVCF) Single sample PBSV SVs (VCF) Single sample Sniffles2 SVs (VCF) Single sample Sniffles2 SNF	*Single sample data available in the the file per sequencing facility:* Auxiliary metrics grch38_noalt Auxiliary metrics T2Tv2.0
HA_Seq_CDRv7	grch38_noalt BAM: standard & haplotagged	*One for each grch38_noalt & T2Tv2.0:* Joint-called SNP & Indel	Auxiliary metrics grch38_noalt Auxiliary metrics

	T2Tv2.0 BAM: standard & haplotagged GFA files: primary *de novo* assembly, alternative *de novo* assembly, one *de novo* assembly for each chromosome copy FASTA: one for each GFA file	variants (VCF & Hail MT) Joint-called SVs (VCF): strict & lenient Single sample SNP & indel variants (VCF) Single sample SNP & Indel phased variants (VCF) Single sample PBSV SVs (VCF) Single sample Sniffles2 SVs (VCF) Single sample Sniffles2 SNF Single sample PAV variants (VCF)	T2Tv2.0

RNA Sequencing (RNA-seq)

The RNA-seq data consists of paired-end sequencing reads from 8,980 whole blood samples and includes STAR-aligned reads, RNA eQTL files, and RNA sQTL files.

Table 25. RNA sequencing deliverables

Deliverable	
Reference version	hg38 reference that excludes ALT, HLA, and decoy and the GENCODE v48 GTF
Raw data	BAM files containing STAR-aligned reads
QTL files	Cis expression QTL (eQTL) Cis eQTL nominal stats Fine-mapped cis eQTLs Cis splicing QTL (sQTL) Fine-mapped sQTL Splicing junctions sQTL
Auxiliary files	RNA sequencing metrics from RNA-SeQC2 RSEM results RNA metadata metrics

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.

BAM- mapped record of gene activity

eQTL and sQTL files reveal how specific genetic differences actually influence that activities

RNA STAR-aligned reads

RNA sequencing reads are delivered as a BAM file accompanied with an index BAI file. The reads have been aligned using the STAR aligner with the hg38 reference with GENCODE v48. The STAR (Spliced Transcripts Alignment to a Reference) aligner is specialized for RNA-seq

data because it can efficiently and accurately detect and map reads that span exon-intron junctions, which is essential for quantifying gene expression and studying splicing events.

RNA eQTL files

Expression Quantitative Trait Loci (eQTL) analysis is a downstream analysis from the STAR-aligned reads to map them to genetic variant data. The analysis connects genetic variants to their functional consequences with the expression data. eQTL analysis identifies genetic loci that are classified as cis-eQTLs when they are located near the target gene or trans-eQTLs when they are far away.

RNA expression cis QTL

TensorQTL was run in the cis permutations mode individually for each genetic ancestry group and the whole combined cohort. Each record in the output file typically includes identifying information for the associated genetic variant (e.g., chromosome, genomic position, and rsID) and the gene whose expression it affects. Essential statistical metrics contained in these files include the P-value quantifying the strength of the association, the effect size (or slope) indicating the direction and magnitude of the change in gene expression per allele copy, and a measure of multiple-testing correction, such as the False Discovery Rate (FDR).

Table 26. cis eQTL file description

Field name	Example value	Notes
phenotype_id		The identifier for the molecular trait being tested
num_var		The total number of genetic variants (SNPs/indels) tested within the defined cis-window for this specific phenotype
beta_shape1		The first shape parameter (alpha) of the Beta distribution fitted to the permutation results to model the null distribution.
beta_shape2		The second shape parameter (beta) of the Beta distribution fitted to the permutation results
true_df		The calculated true degrees of freedom for the model, which can be adjusted by the software to account for relatedness or population structure
pval_true_df		The p-value calculated using the true degrees of freedom before applying the Beta distribution approximation
variant_id		The identifier of the top variant (best hit) that has the strongest association (lowest p-value) with the phenotype within the cis-window
start_distance		The genomic distance (in base pairs) from the variant to the start coordinate of the phenotype feature (e.g., the transcription start site or gene boundary)
end_distance		The genomic distance (in base pairs) from the variant to the end coordinate of the phenotype feature

ma_samples		The number of samples in the analysis that carry at least one copy of the minor allele for the top variant
ma_count		The total count of the minor allele observed across the entire sample cohort for the top variant
af		Allele Frequency; the alternative allele frequency (or minor allele frequency) of the top variant in the dataset.
pval_nominal		The raw, uncorrected nominal p-value for the association between the top variant and the phenotype
slope		The effect size (beta coefficient) from the linear regression, indicating the direction and magnitude of change in protein expression per copy of the alternative allele
slope_se		The standard error of the effect size (slope), measuring the statistical uncertainty of the estimate
pval_perm		The empirical p-value calculated directly from the explicit permutations of the data
pval_beta		The permutation-adjusted p-value calculated via the fitted Beta distribution. This corrects for the number of variants tested per gene/protein and represents the gene-level significance
qval		The False Discovery Rate (FDR) corrected p-value (typically calculated using Storey's Q-value method) across all phenotypes to account for genome-wide multiple testing
pval_nominal_threshold		The nominal p-value threshold required for any variant within this phenotype's cis-window to be considered genome-wide significant (accounting for the local linkage disequilibrium structure)

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

Cis QTL nominal stats

For each cis eQTL, we provide nominal statistics. The nominal stats output represents the raw, uncorrected statistical results for every single variant-gene pair tested.

Table 27. cis eQTL nominal stats file description

Field name	Example value	Notes
phenotype_id		The identifier for the molecular trait being tested
variant_id		The identifier for the variant (usually chr:pos:ref:alt)
start_distance		The genomic distance (in base pairs) from the variant to the start coordinate of the phenotype feature (e.g., the transcription start site or gene boundary)
af		Allele Frequency; the alternative allele frequency (or minor allele frequency) of the top variant in the dataset.

ma_samples		The number of samples in the analysis that carry at least one copy of the minor allele for the top variant
ma_count		The total count of the minor allele observed across the entire sample cohort for the top variant
slope		The effect size (beta coefficient) from the linear regression, indicating the direction and magnitude of change in protein expression per copy of the alternative allele
slope_se		The standard error of the effect size (slope), measuring the statistical uncertainty of the estimate

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

Fine-mapped cis eQTLs

This aggregated summary file contains population-level statistics from a susieR (Sum of Single Effects) fine-mapping analysis. Designed to move from correlation to biological causation, it isolates the most likely causal variants driving gene expression within a cis window while accounting for linkage disequilibrium (LD) and multiple independent signals. Instead of raw p-values, the dataset provides site- and cluster-level metrics—specifically grouping highly correlated SNPs into 95% Credible Sets and assigning each a Posterior Inclusion Probability (PIP) to pinpoint the exact genetic changes altering expression.

Table 28. Fine-mapped cis eQTLs file description

Field name	Example value	Notes
chromosome		The chromosome where the variant or molecular trait is located
start		The starting genomic coordinate of the feature
end		The ending genomic coordinate of the feature
width		The genomic span/length of the feature
strand		The genomic strand of the molecular trait/gene.
molecular_trait_id		The unique identifier for the tested molecular trait (e.g., eQTL phenotype ID, transcript ID, or quantification ID)
variant		The unique identifier for the genetic variant (typically chr:pos:ref:alt or an rsID).
position		The specific genomic coordinate (BP) of the variant
ref		The reference allele in the human genome assembly
alt		The alternative (effect) allele tested
cs_ics_index		The index or indicator for whether a variant belongs to a specific Credible Set (CS) or Independent Credible Set (ICS) in SuSiE

region		The genomic locus or window evaluated during the finemapping process
pip		Posterior Inclusion Probability: The probability that this specific variant is a true causal variant
z		The Z-score of the variant from the univariate association test
cs_min_r2		The minimum pairwise LD between any two variants within the same Credible Set. Measures CS purity
cs_avg_r2		The average pairwise LD among all variants within the Credible Set.
cs_size		The total number of variants included in the Credible Set
posterior_mean		The estimated posterior effect size (beta) for the variant from SuSiE
posterior_sd		The standard deviation of the posterior effect size estimate
cs_log10bf		The log10 Bayes Factor for the Credible Set, indicating the strength of evidence for a causal signal in that set
group		Group assignment identifier (often used if multi-tissue or multi-condition mapping was performed)
ALT_FREQS		The frequency of the alternative allele in the study population
MAF		Minor Allele Frequency: The frequency of the less common allele in the population.
AF_bin		Categorical binning of the allele frequency (e.g., rare, low-frequency, common)
gene_id		The Ensembl or NCBI identifier for the target gene (e.g., ENSG...)
tss		The genomic coordinate of the gene's Transcription Start Site
gene_type		The biotype of the gene (e.g., protein_coding, lncRNA, pseudogene)
gene_name		The HGNC gene symbol (e.g., GAPDH, TP53)
distTSS		The distance (in base pairs) from the variant to the gene's Transcription Start Site (TSS)
PIP_bin		Categorical binning of the PIP values for stratified downstream analyses
PIP_decile		The decile ranking ($1\text{–}10$) of the variant based on its PIP
V4 / V5		Generic column placeholders, often representing unlabelled custom annotations or raw genomic coordinates from a BED file merge
CTCF.only		ENCODE/Screen annotation indicating the presence of a CTCF-binding site without other enhancer/promoter marks.
CTCF.bound		Binary/score indicator showing if the variant falls within an experimentally validated, protein-bound CTCF site
pELS		ENCODE annotation: proximal Enhancer-Like Signature
DNase.H3K4me3		Epigenomic mark indicating open chromatin (DNase) paired with

		promoter-associated histone modifications
dELS		ENCODE annotation: distal Enhancer-Like Signature
PLS		ENCODE annotation: Promoter-Like Signature.
FANTOM5		Binary or score column indicating overlap with robustly defined CAGE promoters/enhancers from the FANTOM5 consortium
intron_variant		VEP Consequence: Variant falls within an intron of a transcript
non_coding_transcript_variant		VEP Consequence: Variant falls within a transcript designated as non-coding
splice_region_variant		VEP Consequence: Variant falls within 1-3 bases of the exon or 3-8 bases of the intron.
splice_polypyrimidine_tract_variant		VEP Consequence: Variant alters the polypyrimidine tract in a splice region
splice_donor_region_variant		VEP Consequence: Variant occurs in the region surrounding a splice donor site
splice_donor_5th_base_variant		VEP Consequence: Specific variant mapping to the 5 position of a splice donor site.
3_prime_utr_variant		VEP Consequence: Variant falls within the 3' Untranslated Region
synonymous_variant		VEP Consequence: A coding variant that does not alter the resulting amino acid
upstream_gene_variant		VEP Consequence: Variant is located 5' (upstream) of the gene
5_prime_utr_variant		VEP Consequence: Variant falls within the 5' Untranslated Region
missense_variant		VEP Consequence: A coding variant that changes an amino acid
downstream_gene_variant		VEP Consequence: Variant is located $3'$ (downstream) of the gene
splice_donor_variant		VEP Consequence: Variant maps directly to the highly conserved splice donor consensus sequence
nmd_transcript_variant		VEP Consequence: Variant falls in a transcript predicted to undergo Nonsense-Mediated Decay
non_coding_transcript_exon_variant		VEP Consequence: Variant alters an exon of a non-coding RNA transcript
inframe_insertion		VEP Consequence: An insertion that preserves the triplet reading frame
stop_gained		VEP Consequence: A variant that introduces a premature stop codon (nonsense mutation)
coding_sequence_variant		VEP Consequence: Variant falls within the coding sequence
intergenic_variant		VEP Consequence: Variant lies in a genomic region between distinct

		genes
frameshift_variant		VEP Consequence: An indel that disrupts the triplet reading frame
inframe_deletion		VEP Consequence: A deletion that removes whole codons without shifting the reading frame
start_lost		VEP Consequence: Variant alters the start codon
splice_acceptor_variant		VEP Consequence: Variant maps directly to the highly conserved splice acceptor consensus sequence
mature_mirna_variant		VEP Consequence: Variant alters a mature miRNA sequence
stop_retained_variant		VEP Consequence: Variant alters the stop codon sequence but still codes for a stop
stop_lost		VEP Consequence: Variant alters the stop codon, causing elongation of the protein sequence
phylop		Conservation score; measures evolutionary conservation across species at the variant base (higher scores = more conserved)
lof.pLI		Probability of Loss-of-Function Intolerance score for the target gene (closer to 1 means the gene is highly constrained/essential)
Constrained		Metric indicating if the variant or gene region is under strong purifying selection
Annotation		A summary string or final tier assignment combining functional and structural annotations for easy filtering

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

RNA sQTL files

Splicing Quantitative Trait Loci (sQTL) analysis is a downstream analysis from the STAR-aligned reads to map them to genetic variant data. The analysis connects genetic variants to their functional consequences with alternative splicing data (such as intron excision ratios or isoform percentages). sQTL analysis identifies genetic loci that are classified as cis-sQTLs when they are located near the target splicing event.

RNA splicing cis QTL

TensorQTL was run in cis permutations mode for each genetic ancestry group and the combined cohort. Each record links a genetic variant (chromosome, position, rsID) to a specific splicing event (intron junction or cluster) and its parent gene. Key metrics include the p-value, the effect size (slope) representing the change in splicing ratio per allele copy, and an FDR-corrected significance measure for multiple testing.

See expression cis QTL.

Fine-mapped sQTL

This summary file details susieR fine-mapping for cis-sQTLs, isolating the most likely causal variants driving alternative splicing within a cis window. By accounting for linkage disequilibrium (LD) and multiple independent signals, it moves beyond correlation to pinpoint the exact genetic changes altering splicing. Instead of raw p-values, it groups correlated variants into 95% Credible Sets and assigns each a Posterior Inclusion Probability (PIP) indicating its likelihood of altering intron excision or isoform usage.

See fine-mapped cis eQTL.

Splicing junctions sQTL

In a LeafCutter cis-sQTL analysis, the output quantifies how genetic variants influence alternative splicing by measuring relative intron excision ratios rather than full transcript expression. LeafCutter groups overlapping introns into distinct clusters, and each phenotype is formatted as a specific junction string identifying its genomic location (e.g., chr:start:end:clu_ID_strand). The statistical output—such as the effect size (slope) and p-value—reflects how a given allele changes the usage of that specific intron junction relative to all other alternative splicing pathways within the same cluster.

Table 29. splicing junctions sQTL output descriptions

Field name	Notes
leafcutter.bed.gz	A standard, tab-delimited BED file (specifically in BED12+ format) for the molecular phenotype matrix. It contains calculated intron excision ratios. The first 4 columns specify the genomic location of the splicing cluster, and subsequent columns contain the phenotype values for each sample.
leafcutter.bed.gz.tbi	The Tabix index file for leafcutter.bed.gz. This index allows QTL mapping software to perform lightning-fast, random-access lookups of specific genomic coordinates without needing to read the entire compressed BED file into memory
.leafcutter.bed.parquet	A Parquet format version of your phenotype BED matrix
leafcutter.phenotype_groups.txt	A two-column mapping file that groups individual intron excision events into their overarching splicing clusters or genes. QTL mapping software requires this file during permutation/cis-window tests so it can properly correct for the dependency and multiple testing wrapper around multiple introns that all belong to the exact same splicing cluster.

RNA auxiliary files

RNA metadata metrics

The initial RNA alignment metrics are provided per sample and include the Research ID, alignment rate percent, RQS quality score, percent mRNA bases, the percent ribosomal bases,

number of aligned read pairs, the mean insert size, and the Dragen software version used for alignment.

RNA-SeQC 2 metrics

This metrics file from RNA-SeQC 2 contains critical Quality Control (QC) statistics grouped by alignment quality, genomic localization (where the reads mapped), strand specificity, and library/transcript preparation bias (like 3' degradation).

Table 30. RNA-SeQC2 metrics file descriptors

Field name	Example value	Notes
sample_id		The unique identifier for the analyzed sample
Mapping Rate		The fraction of total reads that successfully aligned to the reference genome (Mapped Reads / Total Reads)
Unique Rate of Mapped		The fraction of mapped reads that aligned to a single unique position in the genome
Duplicate Rate of Mapped		The fraction of mapped reads flagged as PCR or optical duplicates
Duplicate Rate of Mapped, excluding Globins		The duplicate rate calculated after removing reads mapping to globin genes (highly useful for whole-blood RNA-seq)
Base Mismatch		The overall alignment mismatch rate across all mapped bases
End 1 Mapping Rate		The mapping rate specifically for Read 1 (Forward read) in paired-end sequencing
End 2 Mapping Rate		The mapping rate specifically for Read 2 (Reverse read) in paired-end sequencing
End 1 Mismatch Rate		The base mismatch rate specifically calculated across Read 1 alignments
End 2 Mismatch Rate		The base mismatch rate specifically calculated across Read 2 alignments
Expression Profiling Efficiency		The ultimate benchmark metric: the fraction of total reads that successfully map to exons and are actually usable for quantifying gene expression
High Quality Rate		The fraction of reads that passed vendor quality controls and met mapping quality thresholds (High Quality Reads / Total Reads)
Exonic Rate		The fraction of mapped reads that overlap a known exon
Intronic Rate		The fraction of mapped reads that map entirely within an intron
Intergenic Rate		The fraction of mapped reads that map to genomic space outside of any known gene boundaries
Intragenic Rate		The sum of exonic and intronic rates (total reads mapping anywhere inside gene boundaries)

Ambiguous Alignment Rate		The fraction of reads that map to a region where multiple overlapping gene annotations exist on the same strand, making gene assignment ambiguous
High Quality Exonic Rate		The fraction of high-quality reads that map to exons
High Quality Intronic Rate		The fraction of high-quality reads that map to introns
High Quality Intergenic Rate		The fraction of high-quality reads that map to intergenic regions
High Quality Intragenic Rate		The fraction of high-quality reads that map within gene boundaries
High Quality Ambiguous Alignment Rate		The fraction of high-quality reads with ambiguous gene assignments
Discard Rate		The fraction of reads discarded due to failing vendor QC, low mapping quality, or being too short
rRNA Rate		The fraction of reads mapping to ribosomal RNA (an indicator of the performance of poly-A selection or rRNA depletion protocols)
End 1 Sense Rate		The fraction of Read 1 alignments that match the coding/sense strand of the annotated gene
End 2 Sense Rate		The fraction of Read 2 alignments that match the coding/sense strand of the annotated gene
Avg. Splits per Read		The average number of split-alignments per read, typically corresponding to reads crossing exon-exon splice junctions
Total Alignments		The absolute total number of alignments recorded in the BAM file (can be higher than total reads if multi-mapping is allowed)
Alternative Alignments		The count of secondary/alternative alignments for multi-mapping reads
Supplementary Alignments		The count of chimeric or split-read alignments flagged as supplementary by the aligner (e.g., STAR)
Total Reads		The total number of raw sequencing reads processed
Chimeric Fragments		The absolute count of paired-end fragments where the two ends map to entirely different chromosomes or far-apart loci
Chimeric Alignment Rate		The fraction of fragments flagged as chimeric (potential fusion genes or artifacts)
End 1 Antisense		The absolute count of Read 1 alignments mapping to the antisense strand
End 2 Antisense		The absolute count of Read 2 alignments mapping to the antisense strand
End 1 Bases		The total number of nucleotide bases sequenced across all Read 1s
End 2 Bases		The total number of nucleotide bases sequenced across all Read 2s

End 1 Mapped Reads		The absolute count of successfully mapped Read 1s
End 2 Mapped Reads		The absolute count of successfully mapped Read 2s
End 1 Mismatches		The total count of mismatched bases found in Read 1 alignments
End 2 Mismatches		The total count of mismatched bases found in Read 2 alignments
End 1 Sense		The absolute count of Read 1 alignments mapping to the sense strand
End 2 Sense		The absolute count of Read 2 alignments mapping to the sense strand
Exonic Reads		The absolute number of reads mapping to exons
Failed Vendor QC		The count of reads flagged as failing the sequencer's internal quality filters
High Quality Reads		The absolute count of reads passing both vendor QC and alignment quality thresholds
Intergenic Reads		The absolute number of reads mapping to intergenic regions
Intragenic Reads		The absolute number of reads mapping inside gene regions (exons + introns)
Ambiguous Reads		The absolute count of reads that could belong to multiple overlapping genes
Intronic Reads		The absolute number of reads mapping to introns
Low Mapping Quality		The count of reads discarded because their mapping quality score (MAPQ) fell below the threshold
Low Quality Reads		The count of reads failing basic quality checks
Mapped Duplicate Reads		The absolute count of mapped reads flagged as PCR/optical duplicates
Mapped Reads		The total absolute number of reads that aligned anywhere in the genome
Mapped Unique Reads		The total absolute number of reads that aligned to exactly one genomic position
Mismatched Bases		The absolute count of individual nucleotide mismatches across all alignments
Non-Globin Reads		The absolute count of reads remaining after subtracting those that mapped to hemoglobin genes
Non-Globin Duplicate Reads		The count of duplicate reads among the non-globin alignments.
Reads used for Intron/Exon counts		The final subset of high-quality, uniquely mapped reads that were successfully utilized to build your expression matrices
rRNA Reads		The absolute count of reads aligning to ribosomal RNA sequences
Total Bases		The total number of raw nucleotide bases sequenced (Total Reads times Read Length)

Total Mapped Pairs		The absolute number of paired-end fragments where both read ends successfully mapped
Unique Mapping, Vendor QC Passed Reads		The absolute count of cleanly aligned, non-duplicate, unique reads passing basic filters
Unpaired Reads		The number of reads where only one end of a paired-end block successfully mapped
Read Length		The nominal length (in base pairs) of individual sequencing reads
Genes Detected		The total count of distinct genes that achieved a non-zero expression count in this sample
Estimated Library Complexity		An algorithmic estimate of the total number of unique cDNA fragments in the library before PCR amplification, based on duplication rates
Genes used in 3' bias		The number of highly expressed, sufficiently long genes evaluated to calculate transcript coverage bias
Mean 3' bias		The average 3' bias across evaluated transcripts (1 indicates perfect uniform coverage; values closer to 0 indicate severe 3' degradation/bias)
Median 3' bias		The median value of the 3' bias calculation across evaluated transcripts
3' bias Std		The standard deviation of the 3' bias measurements across evaluated genes
3' bias MAD_Std		The Median Absolute Deviation (MAD) of the 3' bias scaled to act like a standard deviation (more robust against outliers)
3' Bias, 25th Percentile		The 25th percentile value of 3' bias across transcripts
3' Bias, 75th Percentile		The 75th percentile value of 3' bias across transcripts
Median of Avg Transcript Coverage		The median value of depth/coverage uniformity tracked across the body of all calculated transcripts
Median of Transcript Coverage Std		The median standard deviation of coverage depth across transcripts
Median of Transcript Coverage CV		The median Coefficient of Variation (CV = SD / Mean) of transcript coverage. Lower values indicate more uniform coverage
Median Exon CV		The median Coefficient of Variation of coverage measured across individual exons.
Exon CV MAD		The Median Absolute Deviation of the Exon CV
Fragment GC Content Mean		The mean GC percentage of sequenced fragments (checks for GC-bias during library prep/PCR)
Fragment GC Content Std		The standard deviation of fragment GC content across the sample
Fragment GC Content		The asymmetry of the fragment GC content distribution around its mean

Skewness		
Fragment GC Content Kurtosis		The sharpness/tailedness of the fragment GC content distribution

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

RNA RSEM

Table 31. RNA RSEM output descriptions

Field name	Notes
rsem_genes_expected_count.txt.gz	Contains the estimated number of fragments/reads that originated from each gene. Because RSEM uses an expectation-maximization algorithm to fractionally assign multi-mapping reads to their most likely source, these values are decimals (fractions) rather than integers.
rsem_genes_tpm.txt.gz	Contains gene-level normalized expression values in Transcripts Per Million (TPM). This represents the relative abundance of the gene, normalized for both transcript length and sequencing depth. The sum of all TPM values within a single sample equals 1,000,000
rsem_transcripts_expected_count.txt.gz	Contains the fractional estimated fragment/read counts assigned to individual isoforms/transcripts rather than whole genes. This is highly useful for detecting alternative splicing or differential isoform usage.
rsem_transcripts_isopct.txt.gz	Contains Isoform Percentage (IsoPct) values. This represents the percentage of a gene's total expression that is contributed by this specific transcript/isoform (Transcript TPM / Gene TPM times 100%). It isolated splicing dynamics from overall gene expression changes.
rsem_transcripts_tpm.txt.gz	Contains transcript/isoform-level normalized expression values in TPM. It tracks the absolute relative abundance of specific transcript structures across your samples.

Proteomic data

The Proteomics dataset consists of Olink normalized protein expression data from 10,096 samples. These 10,096 blood samples represent 9,969 participants because they include 127 technical replicates used for downstream normalizations.

Table 32. Proteomic sequencing deliverables

Deliverable	
Reference version	N/A
Raw data	Normalized protein expression (NPX) data in tsv and Parquet format Additionally normalized NPX tsv

QTL files	Cis pQTL
Auxiliary files	Project-specific Olink analysis report metrics

Normalized protein expression (NPX) data

We provide NPX data in Parquet and tsv format. NPX values are derived from raw data counts after a subtraction and normalization pipeline, to minimize technical intra- and inter-plate variation. The $\log_2$ scale values demonstrate the biological protein expression for differential expression analysis.

The NPX tsv and Parquets have the same information, but are in different formats depending on your downstream usage. Parquet format is a columnar binary storage format for data retrieval and file storage efficiency. Parquet format can be used downstream with cloud pipelines and use tools within Python, R, or Spark.

Proteomic raw NPX files

Table 33. Proteomics NPX raw parquet & tsv file description

Field name	Notes
SampleID	The unique identifier for each individual sample in the study. For All of Us, we concatenated the PlateID with the ResearchID to create a unique sample identifier.
SampleType	Categorizes the sample (e.g., Sample, Control, Negative Control, Calibrator)
WellID	The specific well location on the laboratory plate (e.g., A1, G12) where the sample was run
PlateID	The unique identifier for the specific physical run or multi-well plate
DataAnalysisRefID	A reference ID linking the data to a specific analysis batch or software run configuration
OlinkID	The unique internal identifier assigned by Olink to a specific protein assay
UniProt	The universal UniProt accession number for the target protein, used to cross-reference global biological databases
Assay	The gene symbol or common name of the specific protein target being measured (e.g., IL6, TNF)
AssayType	Indicates whether the assay is a standard protein target or an internal control (e.g., Extension Control, Detection Control)
Panel	The name of the specific Olink multiplex panel used (e.g., Inflammation, Cardiometabolic, Explore 384)
Block	Refers to a specific sub-panel or organizational block within larger high-throughput workflows like Olink Explore

Count	The raw sequencing read counts (for NGS-based readouts like Olink Explore) or raw signal intensity
ExtNPX	"Extended NPX" — an intermediate normalized value before the final study-wide or plate-scale normalization is applied
NPX	Normalized Protein Expression. Olink's proprietary relative quantification unit
Normalization	The specific normalization method applied to the dataset (e.g., Intensity Normalization or Plate Distribution Normalization)
PCNormalizedNPX	NPX values that have undergone Principal Component (PC) normalization to adjust for systematic batch effects or technical noise
AssayQC	The quality control status of the specific protein assay (e.g., PASS or WARN)
SampleQC	The quality control status of the individual sample (e.g., PASS or FAIL), determining if the sample layout met standard metrics
ExploreVersion	The software or platform version of the Olink Explore pipeline used to generate the data
IntraCV	Intra-assay Coefficient of Variation; measures the variation or precision between replicates within the same run
InterCV	Inter-assay Coefficient of Variation; measures the variation or reproducibility between different runs or plates
SampleBlockQCWarn	A warning flag indicating that a specific sample/block combination is close to, but hasn't completely failed, quality thresholds
SampleBlockQCFail	A failure flag indicating that the specific sample within that technical block failed quality control and should likely be excluded
BlockQCFail	A failure flag indicating that the entire technical run block failed QC metrics
AssayQCWarn	A warning flag specifically indicating that an assay's performance drifted slightly during the run

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

Normalized tsv & replicate removed tsv

The normalized tsv and replicate-removed tsv NPX files are downstream from the above NPX files and include further normalization. See the fields above for all descriptions.

Table 34. Proteomics NPX raw parquet & tsv file description New fields

Field name	Notes
Fields from above file	
ResearchID	An alternative or blinded subject identifier used to map clinical metadata to the SampleID
Project	The overall study name or internal project code assigned to this specific batch of

	data
SoftwareVersion	The specific version number of the Olink analysis software used to generate the output file.
SoftwareName	The name of the software application used for data processing (e.g., Olink NPX Signature or Olink Insight)
PanelDataArchiveVersion	The version of the Olink panel definition/library file used to decode the assay layouts and data
PreProcessingVersion	The specific version of the pre-processing algorithm applied to the raw data
PreProcessingSoftware	The software engine used to execute the initial raw data conversion and pre-processing
InstrumentType	The hardware system used to read out the assay (e.g., NovaSeq, NextSeq, or Signature Q100)
ReplicateType	Tracks the type of replicate sample used for validation (e.g., Technical to measure run precision vs. Bridge to normalize across different plates/batches)
Adj_factor	Adjustment factor; the specific mathematical value or constant subtracted/added during bridging or normalization to harmonize data across plates

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

Proteomics pQTL files

Proteomics pQTL files integrate NPX data with genomic profiles to identify genetic variants that regulate protein abundance. These files map specific alleles to changes in protein levels across a cohort, providing statistical metrics like p-values and effect sizes to pinpoint the genetic drivers of proteomic variation.

Proteomics cis pQTL

A proteomics cis-pQTL analysis identifies genetic variants located near a protein-coding gene that regulate a specific protein's abundance, measured via Olink. The output pairs a variant with a target protein, providing a p-value and effect size (slope) to show how an allele alters protein levels.

Table 35. Fine-mapped cis eQTLs file description

Field name	Example value	Notes
phenotype_id		The identifier for the molecular trait being tested
num_var		The total number of genetic variants (SNPs/indels) tested within the defined cis-window for this specific phenotype
beta_shape1		The first shape parameter (alpha) of the Beta distribution fitted to the

		permutation results to model the null distribution
beta_shape2		The second shape parameter (beta) of the Beta distribution fitted to the permutation results
true_df		The calculated true degrees of freedom for the model, which can be adjusted by the software to account for relatedness or population structure
pval_true_df		The p-value calculated using the true degrees of freedom before applying the Beta distribution approximation
variant_id		The identifier of the top variant (best hit) that has the strongest association (lowest p-value) with the phenotype within the cis-window
start_distance		The genomic distance (in base pairs) from the variant to the start coordinate of the phenotype feature (e.g., the transcription start site or gene boundary)
end_distance		The genomic distance (in base pairs) from the variant to the end coordinate of the phenotype feature
ma_samples		The number of samples in the analysis that carry at least one copy of the minor allele for the top variant
ma_count		The total count of the minor allele observed across the entire sample cohort for the top variant
af		Allele Frequency; the alternative allele frequency (or minor allele frequency) of the top variant in the dataset
pval_nominal		The raw, uncorrected nominal p-value for the association between the top variant and the phenotype
slope		The effect size (beta coefficient) from the linear regression, indicating the direction and magnitude of change in protein expression per copy of the alternative allele
slope_se		The standard error of the effect size (slope), measuring the statistical uncertainty of the estimate
pval_perm		The empirical p-value calculated directly from the explicit permutations of the data
pval_beta		The permutation-adjusted p-value calculated via the fitted Beta distribution. This corrects for the number of variants tested per gene/protein and represents the gene-level significance
qval		The False Discovery Rate (FDR) corrected p-value (typically calculated using Storey's Q-value method) across all phenotypes to account for genome-wide multiple testing
pval_nominal_threshold		The nominal p-value threshold required for any variant within this phenotype's cis-window to be considered genome-wide significant (accounting for the local linkage disequilibrium structure)

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Notes --** Any other relevant information.

Finemapped pQTL

The fine-mapping dataset is delivered as an aggregated summary-level file containing population-level statistics from a susieR analysis. To preserve participant privacy while maintaining granular genetic data, individual-level genotypes are omitted. The file contains both site-level and cluster-level fields.

They are:

- PIP (Posterior Inclusion Probability) -- A variant-level field grading the statistical probability that a specific genetic variant is truly causal for the targeted trait.

CS (Credible Set) -- A defined group of highly correlated variants where the model calculates a high confidence that at least one variant within the set is the true causal driver.

Effect Size & Standard Error -- Metadata fields describing the estimated directional impact and statistical variance of each evaluated variant.

LD (Linkage Disequilibrium) Matrix -- A correlation matrix detailing the local genetic architecture and pairwise correlation structure across the analyzed genomic region.

Table 36. susieR fine-mapping file description

Field name	Type	Key?	Notes
sample_id	string	yes	Research ID of the sample
center	string	yes	The sequencing facility of the sample. Possible values are: BCM, BI, HA, JHU, UW.
platform	string	yes	Sequencing technology of the sample. Possible values are revio, sequel, ont
mosdepth_cov	float	no	Coverage from the mosdepth tool (See the QC report for a description)
aligned_frac_bases	float	no	Fraction of bases aligned to the reference
aligned_num_bases	float	no	Number of bases aligned to the reference
aligned_num_reads	float	no	Number of reads aligned to the reference

Column Explanations:

- **Field name --** The name of the field. In tsv files, this will appear on the first row of the file.
- **Type --** Data type.
- **Key? --** Whether this field makes up a unique key for the row. Note that all key fields together make a unique key for the row.
- **Notes --** Any other relevant information.

Proteomics auxiliary files

Olink analysis report

We provide Olink PDF analysis reports for each project-specific metadata, going along with the 61 project parquets and tsvs. They contain fundamental run metrics, such as total sample counts and the specific software versions utilized during Olink data generation.

Frequently Asked Questions (FAQs) Regarding the Genomic Data Organization

1. Which variants in the VDS are included in the VAT?

   Variants included in the VAT must meet the following criteria:
   - Sites that pass the 'filters' field
   - Sites with 50 or fewer alternative alleles (for CDRv7)
   - Variants from these sites that pass the ‘FT' field and can be annotated by Nirvana from the VDS.

   Passing the 'FT' field means that at least one call for the variant has passed the 'FT' filter.

   Note: The cutoff for alternative alleles in CDRv7 is 50, though with other releases, this number can change.

2. Does the *All of Us* genomic dataset have Whole Exome Sequencing (WES) data?

   No, the *All of Us* genomic dataset has Whole Genome Sequencing (WGS) data and not WES data. WES data only contains sequencing data for the protein-coding regions of the genome, known as exons, whereas WGS data sequences the entire genome. If you are only interested in the exome, we recommend that you use the exome smaller callset, which provides the variants within the exome.

3. Where can I find the research ID in the srWGS CRAM and array IDAT files?

   The research ID is in the file names of the CRAM and IDAT files. To correlate research IDs between the variant files and the raw data files, use the research IDs in the file name of the raw data files (CRAM and IDATs).

4. Where is the gene name (rsID) stored for each variant?

   The rsID is a reference identifier from dbSNP for each variant and is stored in the Variant Annotation Table (VAT). If you have a rsID of interest, you can use the VAT to determine the genomic coordinates of the variant for analysis in the Hail MT, VCFs, or PLINK formats.