RNA Sequencing data **3**
RNA STAR-aligned reads 3
RNA eQTL files 3
RNA expression cis QTL 4
Cis QTL nominal stats 5
Fine-mapped cis eQTLs 5
RNA sQTL files 8
RNA splicing cis QTL 8
Fine-mapped sQTL 8
Splicing junctions sQTL 8
RNA auxiliary files 9
RNA metadata metrics 9
RNA-SeQC 2 metrics 9
RNA RSEM 12
eQTL and sQTL covariates 12
Proteomic data **13**
Normalized protein expression (NPX) data 13
Proteomic raw NPX files 13
Normalized tsv & replicate removed tsv 14
Proteomics pQTL files 15
Proteomics cis pQTL 15
Fine-mapped pQTL 16
Proteomics auxiliary files 18
Olink analysis report 18
pQTL covariates 19

This article describes how the RNA Sequencing and Proteomics data are organized. For a getting started informational article, please see Getting started with multi-omic data. For information about the RNA Sequencing and Proteomic pipelines and Quality control, please see the All of Us Genomics & Multi-omics Quality Report.

The RNA-seq data consists of paired-end sequencing reads from 8,980 whole blood samples processed with Watchmaker RNA Library Prep kit with Polaris Depletion and includes STAR-aligned reads, RNA eQTL files, and RNA sQTL files. The BAM files are a mapped record of gene activity and the eQTL and sQTL files reveal specific genetic differences that actually influence those activities.

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.

The following is an overview of the RNA and proteomics deliverables.

RNA sequencing data - 8,980

- 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
    - eQTL and sQTL covariates

Proteomics data - 10,096

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

RNA Sequencing and Proteomics deliverables

RNA Sequencing Deliverables	
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-SeQC 2 RSEM results RNA metadata metrics eQTL and sQTL covariates
Proteomics deliverables	
Raw data	Normalized protein expression (NPX) data in tsv and Parquet format Additionally normalized NPX tsv
QTL files	Cis pQTL Fine-mapped pQTL
Auxiliary files	Project-specific Olink analysis report metrics pQTL covariates

RNA Sequencing data

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).

RNA cis eQTL file description

Field name	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
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
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 gene expression per allele copy
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)

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.

Cis eQTL nominal stats file description

Field name	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
af	Allele Frequency
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
slope_se	The standard error of the effect size (slope), measuring the statistical uncertainty of the estimate

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.

The following fields flag VEP-predicted consequences for the variant. See Ensembl's VEP consequence documentation for full definitions: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

- intron_variant, non_coding_transcript_variant, splice_region_variant, splice_polypyrimidine_tract_variant, splice_donor_region_variant, splice_donor_5th_base_variant, 3_prime_utr_variant, synonymous_variant, upstream_gene_variant, 5_prime_utr_variant, missense_variant, downstream_gene_variant, splice_donor_variant, nmd_transcript_variant, non_coding_transcript_exon_variant, inframe_insertion, stop_gained, coding_sequence_variant, intergenic_variant, frameshift_variant, inframe_deletion, start_lost, splice_acceptor_variant, mature_mirna_variant, stop_retained_variant, stop_lost

Fine-mapped cis eQTLs file description

Field name	Notes
chromosome	Chromosome of the variant
start	Start coordinate of the variant interval
end	End coordinate of the variant interval
width	Interval width
strand	Strand of the record (typically * for variants)
molecular_trait_id	ID of the phenotype/feature being fine-mapped (gene or transcript)
variant	Variant ID
position	Genomic coordinate of the variant
ref	Reference allele
alt	Alternate allele
cs_id	Identifier of the credible set the variant belongs to
cs_index	Numeric index of the credible set (SuSiE effect/L number) within the trait
region	The genomic locus or window evaluated
pip	Posterior Inclusion Probability: The probability that this specific variant is a true causal variant
z	Marginal z-score for the variant from the association test feeding SuSiE
cs_min_r2	Minimum pairwise LD (r^2) among variants in the credible set
cs_avg_r2	Mean pairwise LD (r^2) among variants in the credible set
cs_size	Number of variants in the credible set

posterior_mean	Posterior mean effect size for the variant
posterior_sd	Posterior standard deviation of the effect size
cs_log10bf	log10 Bayes factor for the credible set
group	Ancestry group the fine-mapping was run in (e.g., AFR, AMR, EUR)
ALT_FREQS	The frequency of the alternative allele in the study population
MAF	Minor allele frequency
AF_bin	Categorical MAF bin
gene_id	Ensembl gene ID for the associated gene
tss	Transcription start site coordinate of that gene
gene_type	Gencode gene biotype (protein_coding, lncRNA, etc.)
gene_name	Gene symbol
distTSS	Distance from variant to the gene's TSS
PIP_bin	Categorical bin of the PIP value
PIP_decile	Decile bucket (1–10) of the PIP value
V4	ENCODE SCREEN representative DNase hypersensitive sites accession
V5	ENCODE SCREEN cis-Regulatory Element (cCRE) accession
CTCF.only	Overlaps a CTCF-only candidate cCRE
CTCF.bound	Overlaps a CTCF-bound candidate cCRE
pELS	Overlaps a proximal enhancer-like signature candidate cCRE
DNase.H3K4me3	Overlaps a DNase + H3K4me3 (promoter-like, low CTCF) cCRE
dELS	Overlaps a distal enhancer-like signature cCRE
PLS	Overlaps a promoter-like signature cCRE
FANTOM5	Overlaps a FANTOM5 CAGE-defined enhancer/promoter
VEP consequence fields	Please see the VEP consequence reference page.
phylop	Conservation score at the variant position (higher = more conserved)
lof.pLI	gnomAD probability of loss-of-function intolerance for the gene (closer to 1 means the gene is highly constrained/essential)
Constrained	Boolean/categorical constraint flag indicating if the variant or gene region is under strong purifying selection (derived from gnomAD pLI/LOEUF)
Annotation	Annotation label from ENCODE/FANTOM5

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 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.

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.
leafcutter.bed.gz.tbi	The Tabix index file for leafcutter.bed.gz.

.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.

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).

The following fields are created by the RNA-SeQC 2 tool and are described within their documentation: https://github.com/getzlab/rnaseqc/blob/master/Metrics.md.

- Mapping Rate, Unique Rate of Mapped, Duplicate Rate of Mapped, Duplicate Rate of Mapped excluding Globins, Base Mismatch, End 1 Mapping Rate, End 2 Mapping Rate, End 1 Mismatch Rate, End 2 Mismatch Rate, Expression Profiling Efficiency, High Quality Rate, Exonic Rate, Intronic Rate, Intergenic Rate, Intragenic Rate, Ambiguous Alignment Rate, High Quality Exonic Rate, High Quality Intronic Rate, High Quality Intergenic Rate, High Quality Intragenic Rate, High Quality Ambiguous Alignment Rate, Discard Rate, rRNA Rate, End 1 Sense Rate, End 2 Sense Rate, Avg. Splits per Read, Raw Counts (a category, not a single field), Read Length, Genes Detected, Estimated Library Complexity, 3' Bias Mean, 3' Bias Median, 3' Bias Std Deviation, 3' Bias Median Absolute Deviation, 3' Bias 25th Percentile, 3' Bias 75th Percentile, Fragment Length Mean, Fragment Length Median, Fragment Length Std Deviation, Fragment Length Median Absolute Deviation, Median of Transcript Coverage Mean, Median of Transcript Coverage Std Deviation, Median of Transcript Coverage Coefficient of Variation, Median Exon CV, Exon CV MAD

RNA-SeQC 2 metrics file descriptors

Field name	Notes
sample_id	The unique identifier for the analyzed sample
RNA-SeQC 2 metrics fields	Please see the RNA-SeQC Output Metrics page.
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 the 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) - upstream of RNA-SeQC 2
Fragment GC Content Std	The standard deviation of fragment GC content across the sample - upstream of RNA-SeQC 2
Fragment GC Content Skewness	The asymmetry of the fragment GC content distribution around its mean - upstream of RNA-SeQC 2
Fragment GC Content Kurtosis	The sharpness/tailedness of the fragment GC content distribution - upstream of RNA-SeQC 2

RNA RSEM

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
rsem_genes_tpm.txt.gz	Contains gene-level normalized expression values in Transcripts Per Million (TPM)
rsem_transcripts_expected_count.txt.gz	Contains the fractional estimated fragment/read counts assigned to individual isoforms/transcripts rather than whole genes.
rsem_transcripts_isopct.txt.gz	Contains Isoform Percentage (IsoPct) values
rsem_transcripts_tpm.txt.gz	Contains transcript/isoform-level normalized expression values in TPM

eQTL and sQTL covariates

Phenotype and genetic principal components (PCs) are used to adjust for confounding variation in the eQTL, sQTL, and pQTL analyses. Covariate files are provided per analysis type in TSV format: the first column holds the covariate name (GeneticPC1, GeneticPC2, etc. for genotype PCs; PC1, PC2, etc. for phenotype PCs), and the remaining columns hold each sample's value for that covariate, one column per Research ID. The number of genetic and phenotype PCs included varies by analysis type and by ancestry group.

Covariate files are available for each genetic ancestry group (AFR, AMR, EAS, EUR, MID, SAS) individually, as well as a combined file across all ancestry groups.

Proteomic data

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

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 sample's specific well location on the laboratory plate
PlateID	The unique identifier for the multi-well plate
DataAnalysisRefID	A reference ID linking the data to a specific analysis batch
OlinkID	The unique internal identifier assigned by Olink to a specific protein assay
UniProt	The universal UniProt accession number for the target protein
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
Panel	The name of the specific Olink multiplex panel used
Block	Refers to a specific sub-panel or organizational block within larger high-throughput workflow
Count	The raw sequencing read counts
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 Control Normalization)
PCNormalizedNPX	NPX values from Plate Control 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 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

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. The NPX value is now normalized using a reference plate median approach. The PCNormalizedNPX remains the same as the original NPX value. See additional fields above for all descriptions.

Proteomics normalized and replicate-removed NPX files

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

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.

Proteomics cis pQTLs file description

Field name	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
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

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)

Fine-mapped 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.

The following fields flag VEP-predicted consequences for the variant. See Ensembl's VEP consequence documentation for full definitions: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

- intron_variant, non_coding_transcript_variant, splice_region_variant, splice_polypyrimidine_tract_variant, splice_donor_region_variant, splice_donor_5th_base_variant, 3_prime_utr_variant, synonymous_variant, upstream_gene_variant, 5_prime_utr_variant, missense_variant, downstream_gene_variant, splice_donor_variant, nmd_transcript_variant, non_coding_transcript_exon_variant, inframe_insertion, stop_gained, coding_sequence_variant, intergenic_variant, frameshift_variant, inframe_deletion,

start_lost, splice_acceptor_variant, mature_mirna_variant, stop_retained_variant, stop_lost

Fine-mapped pQTL fields

Field name	Notes
chromosome	Chromosome of the variant
start	Start coordinate of the variant interval
end	End coordinate of the variant interval
width	Interval width
strand	Strand of the record (typically * for variants)
molecular_trait_id	ID of the phenotype/feature being fine-mapped (gene or transcript)
variant	Variant ID
position	Genomic coordinate of the variant
ref	Reference allele
alt	Alternate allele
cs_id	Identifier of the credible set the variant belongs to
cs_index	Numeric index of the credible set (SuSiE effect/L number) within the trait
region	The genomic locus or window evaluated
pip	Posterior Inclusion Probability: The probability that this specific variant is a true causal variant
z	Marginal z-score for the variant from the association test feeding SuSiE
cs_min_r2	Minimum pairwise LD (r^2) among variants in the credible set
cs_avg_r2	Mean pairwise LD (r^2) among variants in the credible set
cs_size	Number of variants in the credible set
posterior_mean	Posterior mean effect size for the variant
posterior_sd	Posterior standard deviation of the effect size
cs_log10bf	log10 Bayes factor for the credible set
group	Ancestry group the fine-mapping was run in (e.g., AFR, AMR, EUR)
ALT_FREQS	The frequency of the alternative allele in the study population
MAF	Minor allele frequency
AF_bin	Categorical MAF bin
gene_id	Ensembl gene ID for the associated gene

tss	Transcription start site coordinate of that gene
gene_type	Gencode gene biotype (protein_coding, lncRNA, etc.)
gene_name	Gene symbol
distTSS	Distance from variant to the gene's TSS
PIP_bin	Categorical bin of the PIP value
PIP_decile	Decile bucket (1–10) of the PIP value
V4	ENCODE SCREEN representative DNase hypersensitive sites accession
V5	ENCODE SCREEN cis-Regulatory Element (cCRE) accession
CTCF.only	Overlaps a CTCF-only candidate cCRE
CTCF.bound	Overlaps a CTCF-bound candidate cCRE
pELS	Overlaps a proximal enhancer-like signature candidate cCRE
DNase.H3K4me3	Overlaps a DNase + H3K4me3 (promoter-like, low CTCF) cCRE
dELS	Overlaps a distal enhancer-like signature cCRE
PLS	Overlaps a promoter-like signature cCRE
FANTOM5	Overlaps a FANTOM5 CAGE-defined enhancer/promoter
VEP consequence fields	Please see the VEP consequence reference page.
phylop	Conservation score at the variant position (higher = more conserved)
lof.pLI	gnomAD probability of loss-of-function intolerance for the gene (closer to 1 means the gene is highly constrained/essential)
Constrained	Boolean/categorical constraint flag indicating if the variant or gene region is under strong purifying selection (derived from gnomAD pLI/LOEUF)
Annotation	Annotation label from ENCODE/FANTOM5

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.

pQTL covariates

Phenotype and genetic principal components (PCs) are used to adjust for confounding variation in the eQTL, sQTL, and pQTL analyses. Covariate files are provided per analysis type in TSV format: the first column holds the covariate name (GeneticPC1, GeneticPC2, etc. for genotype PCs; PC1, PC2, etc. for phenotype PCs), and the remaining columns hold each sample's value for that covariate, one column per Research ID. The number of genetic and phenotype PCs included varies by analysis type and by ancestry group.

Covariate files are available for each genetic ancestry group (AFR, AMR, EAS, EUR, MID, SAS) individually, as well as a combined file across all ancestry groups.