SUPPLEMENTARY INFORMATION TO

Multi-tissue integrative analysis of personal epigenomes

Table of Contents

List of supplementary files and corresponding section

phased_block_ind3.txt	S1.1.4
fithic2_out.tar.gz	S1.3.10
TopDomTADcalls.tar.gz	S1.3.10
Supp_data_proteomics.xlsx	S1.3.11
Supp_DE_genes.tsv	S1.4
hic_files.tar.gz	S2.3.2
Associated_AS_Disease_Genes.xlsx	S3.1.3
AS_highpower_set.tar.gz	S3.3
Supp_data_compatibility.xlsx	S4.3.3
Supp_Data_SVs_associated_with_eQTL.xlsx	S6.1
cCRE_Decoration.matrix	S7.2
active.combined_set.txt.zip	S7.2
bivalent.combined_set.txt.zip	S7.2
repressed.combined_set.txt.zip	S7.2
stringent.regions.MF.hg38.bed	S7.2
cCRE_histoneSignals_qnorm.tar.gz	S7.2
ENTEx_fully_repressed_regions_independent_of_cCREs.bed	S7.3
normalized_proteomics_RNA-seq.dat	S7.5
Similarity_of_functional_genomic_activities_of_cCREs.xlsx	S7.5
Tissue_Specificity.zip	S8.1
QTL_enrichment.zip	S9.1
GWAS_enrichment.zip	S9.2
ENTEx.Explorer.cCRE.Combined.zip	S10.6
ENTEx.Explorer.Expression.Combined.zip	S10.6
ENTEx.Proteomics.cCRE.Combined.zip	S10.6

Overview

This document provides a comprehensive and organized reference to all dataset, methods and analyses associated with the EN-TEx project. The structure of this document is presented in a parallel fashion to the main text. Results and figures in each section in the main text are documented in the relevant subsections titled "Supp. content to 'Main-text-section-title'" within this supplement. Supplement figures are numbered based on the secondary heading of their relevant supplement text section (e.g., Figure S1.1a-f are associated with Section S1.1). Unless specified, we refer to the supplement text in a general manner as "see supp." in the main text. However, we explicitly refer to the supplementary figures with exact figure numbers. All supplementary figures are attached at the end of this document. A list of supplementary figures could be found as part of the table of contents.

The large dataset produced under the EN-TEx project includes more than 25 different functional genomic assays sampled in more than 30 tissues of 4 individuals. These include genotyping, RNA-seq, transcription factor ChIP-seq, histone ChIP-seq, DNAse-seq, ATAC-seq, Hi-C and more. The EN-TEx dataset also includes processed data, such as allele specific (AS) heterozygous SNPs catalog and candidate cis-regulatory elements(cCRE) annotation. Together, this dataset provides a valuable resource for studies of gene regulation and precision medicine. However, due to the richness of the data, it is challenging and difficult to include all the details within the main text of this paper.

The data resource is organized in a hierarchical pyramid-like structure, where the raw data files are located at the base. On top of those are processed data, and finally the high-level summaries lie at the top. Specifically, the main text summarizes everything in a broad manner, providing a macroscopic view of this study. Raw data, the cornerstone of this study, are hosted online in the ENCODE portal. Links to metadata, bam files as well as other raw data could be found in the "Raw Data" link in the EN-TEx data portal website (ENTEx.encodeproject.org). The processed data files, located at the middle of the pyramid, are detailedly described in their corresponding sections in this document. When mentioned, these files are referred to as "File: file_name". All these files are hosted in the EN-TEx data portal website, with the exact same file names as in "File: file_name". Additionally, on the website, each file is followed with the supplement text section number that contains the description of that file.

S1. Supp. content to main text section “Personal genomes & matched data matrix”

S1.1. Personal Genome Construction

S1.1.1 sequencing of the personal genome

Figure S1.1 panel a summarizes the technologies used to sequence the whole genomes of the four individuals. We followed Zheng et al. (2016) \cite(*1*) to perform 10X linked-read sequencing. Protocols of Pacbio sequencing were from Eid et al. (2009) \cite(*2*) and Nattestad et al. (2018) \cite(*3*). Using the Illumina genome sequencing each of the 4 genomes were sequenced to a minimum of 60x coverage with an average read fragment length *m* = 350 bp. Assembled fragments using 10X Genomics combined with Illumina sequencing resulted in 35x coverage with a *m* = 117k bp. Genome sequencing using long-reads (*m* = 7.5k bp) resulted in a genome coverage of 55x. Hi-C experiments were carried out to complement the 10X experiments to verify the phasing of entire chromosomes.

S1.1.2. Personal Genome construction and variants calling

For individual 2 and 3, personal genomes were constructed from a combination of long-range Hi-C reads, 10x linked reads, and PacBio long reads using CrossStitch (Figure S1.1), a software pipeline developed by the Schatz lab (https://github.com/schatzlab/crossstitch). A previous study has shown that it is possible to phase SVs with variants identified from 10x linked reads \cite(*4*).

First, the following preprocessing steps were done:

1. Align all reads (HiC, 10X, PacBio) to the human reference (GRCh38).
2. Call small variants from the linked reads with LongRanger (ver. 2.1.2) (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger).
3. Phase small variants with HapCUT2 (ver. 1.1) \cite(*5*) using HiC and 10X data.
4. Call large structural variants with Sniffles (ver. 1.0.11) \cite(*6*) and pbsv (ver. 2.2.1) (https://github.com/PacificBiosciences/pbsv) and merge the callsets with SURVIVOR (ver. 1.0.6) \cite(*7*), discarding SVs which were only identified by pbsv.

Then, the CrossStitch software (commit 53f64af) performed the following steps to obtain a personal genome:

5. Refine structural variants with Iris (ver. 1.0).
6. Phase long reads using the phased small variants they overlap.
7. Phase large structural variants based on the phasing of the reads supporting them (Figure S1.1).
8. Splice the phased variants into two copies of each human chromosome to produce personal diploid chromosome sequences using vcf2diploid (ver. 1.0) \cite(*8*).

9. Assign one sequence of each chromosome to pseudo-haplotype 1 and the other to pseudo-haplotype 2.

Note that each chromosome was phased independently from other chromosomes, so that pseudo-haplotype 1 of one chromosome may correspond to pseudo-haplotype 2 of another chromosome. Unfortunately, the data available is insufficient to distinguish such cases and assemble full haplotypes.

For individual 1 and 4, due to insufficient long-read coverage, large structural variants were omitted, but the process was otherwise the same.

In all four samples, the use of 10x and Hi-C data resulted in chromosome-arm-length phase blocks for all autosomes (Fig. 1A and Figure S1.1). In addition, in both samples for which long reads were used, more than 90% of large indels were able to be confidently phased with CrossStitch.

A detailed overview of this method is illustrated in Figure S1.1.

For all 4 individuals, VCFs containing the SNVs and indels are accessible from the ENCODE portal \cite(*9*) (see Figure S1.1 for accession numbers)

S1.1.3. Refining Novel Insertion Sequences with Iris

Iris is a novel method for refining the breakpoints and sequence of insertion variants. Each of these calls, when taken directly from the variant caller, consists of an insertion sequence obtained from the alignment of a single representative read, and Iris improves upon this sequence by integrating all of the reads which support the variant's presence. It gathers the sequences of all of the reads listed in the RNAMES INFO field output by Sniffles, extracts the original insertion sequence with surrounding context from the reference genome, and uses the gathered reads to polish this sequence with racon (ver. 1.4.0) \cite(*10*). Then, this polished sequence is aligned back to the reference with minimap2 (ver. 2.17) \cite(*11*), and a refined insertion sequence is obtained. If no insertion is found from this alignment which has a length similar to that of the original variant call, Iris falls back on the original sequence to ensure it does not mask variants in more difficult-to-map regions.

We benchmarked the performance of Iris using data from HG002, a sample sequenced as part of the Genome in a Bottle release (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/). In this individual, we called structural variants separately using Oxford Nanopore (ONT) data and Pacbio Circular Consensus Sequencing (CCS) data, both sequenced to ~50x coverage using the ngmlr aligner\cite(*6*) and the sniffles variant caller. Because of the high accuracy of CCS reads, we used the insertion sequences obtained from these calls as a proxy for the ground truth to evaluate the accuracy of the ONT calls. We compared the CCS and ONT call sets before and after refining the ONT calls with Iris. In each comparison we evaluated all variant calls in the CCS dataset which had an ONT variant call within 10 kbp in both the refined and unrefined callsets. Among these 14,001 variants, we measured the average sequence similarity between the CCS call and the ONT call, with the similarity of two strings S and T measured as [1 - edit_distance(S, T)] / max[length(S), length(T)]. Using the unrefined calls, the average similarity was 0.854, while the refined calls gave an average similarity of 0.94, demonstrating the ability of Iris to obtain more accurate insertion breakpoints and sequences. Panel G in Figure S1.1 shows the distribution of sequence similarities before and after refinement.

Iris is available as a stand-alone method at the following link: https://github.com/mkirsche/Iris

S1.1.4. Assigning parental origin by imprinted genes

The list of known human imprinted genes was downloaded from the Imprinted Gene Database (geneimprint.com). For known imprinted genes that were detected ASE in tissues of individual 3, the haplotype-specific read counts were combined from these tissues and the potential parental origin of the haplotype blocks was determined based on the direction of the imbalance (haplotype 1 or haplotype 2) and the known expressed allele of the imprinted gene (maternal or paternal allele). The parental origin results of individual 3 are shown in Fig. 1A and available in File: phased_block_ind3.txt, where each line is a phased block. The first three columns are genomic coordinates of the phased block. The fourth and fifth columns are the parental origin of haplotype 1 and haplotype 2 respectively. 'NoInfo' indicates there were no imprinted genes in that phased block. 'Contradict' indicates there is at least one AS SNP-imprinted gene pair that has a different imbalance direction compared to other SNP-gene pairs, and thus reach contradictory conclusions of the same phased block. A similar approach can be used for the other EN-TEx individuals (See File: xxx)

S1.2. Analysis of Structural Variants

We focus our analysis on SVs that are larger than or equal to 50 bp, while the VCFs (see S1.1.2) also contain 7.6k and 7.3k smaller "SVs" for individual 2 and 3, respectively.

To analyze the sequence composition of the SVs found in individual 2 and individual 3, we used RepeatMasker (ver. 4.0.7, slow search mode) (http://www.repeatmasker.org) to classify the sequences that are inserted, deleted, or inverted.

We also estimated the allele frequencies of the SVs found in individual 2 and individual 3. For this purpose, we checked for overlaps in the location between the EN-TEx SVs and those reported by Audano et al. (2019) \cite(*12*). To increase the chance of finding an overlap between these two datasets, we used the confidence intervals of an EN-TEx SV's coordinates as the location of the SV. When an overlap is found, we further checked whether the two SVs are of the same type (e.g., both are deletions). If the two SVs are not the same type, we consider the two SVs to be different. Through this analysis, we matched SVs in Audano et al. (2019) with 65.9% and 63.4% of the SVs in individual 2 and individual 3, respectively, and assigned these EN-TEx SVs an allele frequency in European populations that is estimated by Audano et al. (2019) \cite(*12*). We performed a similar analysis by using more recent SVs called from long-read DNA sequencing data \cite(*13*) and gnomAD SVs \cite(*14*). 68.9% and 66.5% of the SVs in individual 2 and 3, respectively, overlap with the former dataset. Because gnomAD annotates SVs differently, we allow EN-TEx “INS” to match “INS”, “DUP”, “BND”, and “MCNV” in gnomAD, EN-TEx “DEL” to match gnomAD “DEL”, “BND”, and “MCNV”, and EN-TEx “INV” to match gnomAD “INV” and “BND”. In this way, we find a match for 61.4% and 60.3% of the SVs in individual 2 and individual 3, respectively.

To understand how SVs distribute in the genome, we generated a null expectation of SVs' distribution by shuffling the locations of SVs, using a method similar to that used in the 1000 Genome SV study \cite(*15*). Specifically, we put the SVs in random locations on the same chromosome while avoiding gaps in the assembly. A ratio of the number of unshuffled SVs intersecting a given genomic region over the number of shuffled SVs is calculated. We repeated the shuffling 1,000 times.

S1.3. Functional genomics data stack

The EN-TEx project includes more than 25 different functional genomic assays sampled in more than 30 tissues of 4 individuals (Fig 1C and Figure S1.3a). Additional information on the tissues and legends in Fig 1C could be found in Figure S1.3a. Below we describe the data processing pipeline for each experiment type.

S1.3.1. RNA sequencing

To survey transcriptomes across humans for 25 tissues samples sources from GTEx, we performed a variety of RNA-seq experiments in ENCODE phase III, which can be divided into three classes: (i) bulk RNA-seq surveys RNAs greater than 200 nt and comprises total RNA-seq, (ii) small RNA-seq surveys RNAs less than 200 nt; and (iii) microRNA-seq surveys microRNA levels by selecting for species less than 30 nt. Additional assay details, along with detailed experimental protocols, are available at the ENCODE Portal (https://www.encodeproject.org/data-standards/rna-seq/long-rnas/, https://www.encodeproject.org/data-standards/rna-seq/small-rnas/ and https://www.encodeproject.org/microrna/microrna-seq/). For all RNA-seq experiments, data quality is evaluated by calculating the number of aligned reads and replicate concordance. Details are included in the ENCODE pipelines.

S1.3.2. RAMPAGE

RAMPAGE captures 5′-complete cDNA to allow the identification and quantification of TSSs and transcript characterization. Production documents were generated for each experiment, and a representative experimental protocol is available at https://www.encodeproject.org/documents/0651efa6-7fd7-4b33-ab11-b05348c9f1c0/@@download/attachment/295491.pdf. Additional assay details are available at https://www.encodeproject.org/data-standards/rampage/. The ENCODE RAMPAGE pipeline is appropriate for libraries generated with RNAs longer than 200 nt, and it consumes reads in FASTQ format and produces alignments and normalized signals for both the + and − strands. Quality control is performed for the peaks, and the irreproducible discovery rate (IDR) is used to identify reproducible peaks between replicates. Data quality is evaluated by calculating read depth and replicate concordance.

S1.3.3. eCLIP

Enhanced crosslinking and immunoprecipitation (eCLIP) identifies transcriptome wide RBP occupancy sites. The experimental protocol is available at https://www.encodeproject.org/documents/842f7424-5396-424a-a1a3-3f18707c3222/@@download/attachment/eCLIP_SOP_v1.P_110915.pdf. Additional assay details are available at https://www.encodeproject.org/eclip/. We require all eCLIP antibodies to undergo primary and secondary characterizations. Detailed RBP antibody standards are available at https://www.encodeproject.org/documents/fb70e2e7-8a2d-425b-b2a0-9c39fa296816/@@download/attachment/ENCODE_Approved_Nov_2016_RBP_Antibody_Characterization_Guidelines.pdf. Data quality is evaluated by calculating the number of unique fragments, IDR, and the fraction of reads in peaks (FRiP).

S1.3.4. Histone ChIP–seq

Histone ChIP–seq surveys the interaction between DNA and histone proteins, selecting for a specific protein variant or post-translational modification through immunoprecipitation followed by sequencing. The experimental protocols are available at https://www.encodeproject.org/documents/be2a0f12-af38-430c8f2d-57953baab5f5/@@download/attachment/Epigenomics_Alternative_Mag_Bead_ChIP_Protocol_v1.1_exp.pdf. Additional assay details are available at https://www.encodeproject.org/chip-seq/histone/. We required all commercial histone antibodies to be validated by at least two independent methods, and antibodies need to be analysed independently. Histone mark antibody standards are available at https://www.encodeproject.org/documents/4bb40778-387a-47c4-ab24-cebe64ead5ae/@@download/attachment/ENCODE_Approved_Oct_2016_Histone_and_Chromatin_associated_Proteins_Antibody_Characterization_Guidelines.pdf. Data quality is evaluated by calculating read depth, non-redundant fraction (NRF) (that is, the number of distinctly uniquely mapping reads over the total number of reads), and PCR bottlenecking coefficients (PBC1 and PBC2).

S1.3.5. ChIP–seq of chromatin-associated proteins

ChIP–seq surveys the interaction between DNA and DNA regulatory proteins (CTCF, EP300 and PolII) through immunoprecipitation followed by sequencing. The protocol is available at https://www.encodeproject.org/documents/20ebf60b-4009-4a57-a540-8fd93407eccc/@@download/attachment/Epigenomics_CR_ChIP_Protocol_v1.0.pdf, https://www.encodeproject.org/documents/6ecd8240-a351-479b-9de6-f09ca3702ac3/@@download/attachment/ChIP-seq_Protocol_v011014.pdf, https://www.encodeproject.org/documents/a59e54bc-ec64-4401-8cf6-b60161e1eae9/@@download/attachment/EN-TEx%20ChIP-seq%20Protocol%20-%20Myers%20Lab.pdf, and https://www.encodeproject.org/documents/f2aa60f2-90a6-4e4b-863a-c6831be371a2/@@download/attachment/ChIP-Seq%20Biorupter%20Pico%20TruSeq%20protocol%20for%20Syapse-c5bdc444fe0511e69d6a06346f39f379.pdf. Additional assay details are available at https://www.encodeproject.org/chip-seq/transcription_factor/. Data quality is evaluated by calculating read depth, NRF, PCR bottlenecking coefficients (PBC1 and PBC2), replicate concordance using IDR, and FRiP.

S1.3.6. ATAC–seq

ATAC–seq surveys open chromatin regions through the insertion of primers into the genome via transposase followed by sequencing. Experimental protocols are available at https://www.encodeproject.org/documents/404ab3a6-4766-45ca-af80-878a344f07b6/@@download/attachment/ATAC-Seq%20protocol.pdf. Additional details can be found at https://www.encodeproject.org/atac-seq/. Data quality is evaluated by calculating the number of non-duplicate, non-mitochondrial aligned reads, alignment rate, IDR, NRF, PCR bottlenecking coefficients (PBC1 and PBC2), number of resulting peaks, fragment length distribution, FRiP, and TSS enrichment.

S1.3.7. DNase-seq

DNase-seq surveys open chromatin regions through genomic cleavage by endonuclease DNase I followed by sequencing. Experimental protocols are available at https://www.encodeproject.org/documents/c6ceebb6-9a7a-4277-b7be-4a3c1ce1cfc6/@@download/attachment/08112010_nuclei_isolation_human__tissue_V6_3.pdf. Additional details are available at https://www.encodeproject.org/data-standards/dnase-seq/. Data quality is evaluated by calculating the number of uniquely mapping reads, the fraction of mitochondrial reads, and the signal portion of tags (SPOT) score.

S1.3.8. WGBS

To map DNA methylation, WGBS uses bisulfite treatment to convert unmethylated cytosines into uracils, leaving methylated cytosines unchanged. Through sequencing and alignment to a transformed genome, CpG, CHG, and CHH methylation levels can be extracted. The experimental protocol is available at https://www.encodeproject.org/documents/9d9cbba0-5ebe-482b-9fa3-d93a968a7045/@@download/attachment/WGBS_V4_protocol.pdf. Additional details are available at https://www.encodeproject.org/data-standards/wgbs/. Data quality is evaluated by genomic coverage, C-to-T conversion rate, and correlation of CpG methylation levels between replicates.

S1.3.9. DNAme array

DNAme arrays measure methylation at CpGs. Like WGBS, DNA is treated with bisulfite to convert unmethylated cytosines to uracils. After amplification, DNA is hybridized to an array (Illumina Infinium Methylation EPIC BeadChip) with probes for both methylated and unmethylated states. Methylation is then quantified by comparing the signal between the two probes. All ENCODE uniform processing pipelines can be found at https://github.com/ENCODE-DCC.

S1.3.10. Hi-C

We generated high quality in situ Hi-C data from samples collected from the gastrocnemius medialis and transverse colon tissues of four different donors. The in situ Hi-C protocol was used to produce Hi-C libraries as described previously \cite(*16*), A detailed protocol document is provided with each dataset in the ENCODE data portal (https://www.encodeproject.org/documents/e1ef20c9-7539-40bc-bdbf-a4deab7f72c7/). Approximately 20 mg of tissue was used for each experiment, and the MboI restriction enzyme was used for restriction digests. All sequencing was performed on an Illumina 4000 platform. The data was processed twice, once utilizing a reference genome and once utilizing personal genomes that were constructed for each tissue of each individual (see S1.1).

Hi-C Data Processing details

(a) Creation of interaction matrices: The interaction matrices were generated using the Juicer pipeline \cite(*17*), an open source tool for analyzing large Hi-C libraries. We utilized BWA-MEM \cite(*18*) to align individual reads to the hg38 reference genome, which was obtained from the ENCODE data portal. For each paired-end read, the two individual sequences were first separately aligned to the reference genome before being paired based on their read names. Chimeric reads and PCR duplicates were removed prior to the creation of an interaction matrix for each tissue of each individual (Figure S1.3b). Figure S1.3c provides information on the number of reads and number of contacts per sample utilized to create the matrices.

(b) A/B compartments: Determination of the A and B compartments were done using the Juicer pipeline \cite(*17*) in 1MB resolution. In detail, the observed/expected interaction matrices were normalized using the Knight-Ruiz matrix-balancing (KR). A correlation matrix from these interaction matrices was calculated, with the first eigenvector of the matrix corresponding to A/B compartments. The negative values of the vector indicate the regions belonging to the A compartment, while the positive values of the vector correspond to the regions in the B compartment (Figure S1.3d-e).

(c) Significant Hi-C interactions: Significant intrachromosomal Hi-C interactions were identified with FitHiC2 (v2.0.7) \cite(*19, 20*). Preprocessing of EN-TEx Hi-C interaction matrices followed the author's instructions in the FitHiC2 GitHub repository's README (https://github.com/ay-lab/fithic). Matrices were binned at a resolution of 50 kb and bin biases were generated using the author's provided software (HiCKRy.py — with percentOfSparseToRemove set to 0.1). FitHiC2 output for each sample can be found in the compressed folder File: fithic2_out.tar.gz on the EN-TEx resource website. See Figure S1.3f for the number of all vs. significant interactions for each sample.

(d) TAD annotations: Topologically Associating Domains (TAD) were identified using TopDom (v0.9.0) \cite(*21*) with a window_size of 3. Before running TopDom, EN-TEx Hi-C libraries were binned at a resolution of 100 kb and normalized using the Knight-Ruiz matrix-balancing algorithm \cite(*22*) implemented by Juicer \cite(*17*). TopDom's window_size parameter was optimized for the known enrichment of CTCF motif directionality at TAD boundaries \cite(*16*) and visual consistency/fit with Hi-C interaction matrices. CTCF directionality was identified using paired EN-TEx ChIP-seq peak (narrowPeak format) files and the 'CTCF_known1' motif as described in Cameron et al. (2020) \cite(*23*). TAD calls across EN-TEx individuals for the same tissue remain consistent (based on the frequency of TADs for a given size) and show slight differences between the two tissues observed by Hi-C (Figure S1.3g), supporting their cell-type specific nature. TAD boundary similarity was calculated by overlapping TAD annotations between individuals/tissues with a buffer of three bins when considering two boundaries to be the same.

TAD annotations were generated by applying TopDom \cite(*21*) to the Hi-C data from the two tissues of all four donors. Annotations are available for download via the ENTEx.encodeproject.org portal. These files are found within a compressed folder named File: TopDomTADcalls.tar.gz

S1.3.11. Proteomics

LC-MS/MS Analysis

For 10 mg tissue, 200 µl lysis buffer (50 mM Tris-HCl pH8.5, 50 mM NaCl, 8 M urea, 4% SDS, and Halt protease inhibitor (Thermo)) was added. After the tissue was homogenised by a pestle/mortar, a Dounce homogeniser, or similar device, the sample was heated at 95°C for 10 min, and followed by probe sonication until viscosity was reduced. Sample was then centrifuged at 13,000 rpm for 15 min, and supernatant was collected. RNA was first extracted from samples (See RNA-seq method).

Protein concentration was measured by Pierce 660nm Protein Assay (Thermo). 100 µg proteins were taken for each sample, and volumes were equalised by 100 mM TEAB to 100 µl, reduced by 20 mM TCEP (Sigma) then alkylated by 40 mM iodoacetamide (Sigma). Proteins were purified by 20% TCA precipitation. 100 mM TEAB was added to the sample, followed by digestion with trypsin (MS grade, Thermo) at 37°C for 18 hours. The peptides were labelled by TMT10plex as per the manufacturer's instruction, labelled samples were pooled and SpeedVac dried. 300 µg peptides were fractionated on a U3000 HPLC system (Thermo Fisher) using an XBridge BEH C18 column (2.1 mm id x 15 cm, 130 Å, 3.5 µm, Waters) at pH 10, at 200 µl/min in 30 min linear gradient from 5 - 35% acetonitrile /NH_4OH. The fractions were collected every 30 sec into a 96-well plate, these were concatenated to 35 fractions and dried.

The peptides were resuspended in 0.5% formic acid (FA), 50% was injected for LC-MS/MS analysis on an Orbitrap Fusion Tribrid mass spectrometer coupled with U3000 RSLCnano UHPLC system (Thermo Fisher). The peptides were loaded onto a PepMap C18 trap (100 µm i.d. x 20 mm, 100 Å, 5 µm) for 10 min at 10 µl/min with 0.1% FA/H_2O, then separated on a PepMap C18 column (75 µm i.d. x 500 mm, 100 Å, 2 µm) at 300 nl/min and a linear gradient of 4-33.6% ACN/0.1% FA in 90 min /cycle at 120 min, or 4-32% ACN/0.1% FA in 150 min or 180 min with cycle time at 180 min or 210 min for each fraction. The data acquisition used the SPS10-MS3 method with Top Speed at 3s per cycle time. The full MS scans (m/z 380-1500) were acquired at 120,000 resolution at m/z 200, and the AGC was set at 400,000 with 50 ms maximum injection time. The most abundant multiply-charged ions (z = 2-6, above 5000 counts) were subjected to MS/MS fragmentation by CID (35% CE) and detected in ion trap for peptide identification. The isolation window by quadrupole was set m/z 1.0, and AGC at 10,000 with 35 ms maximum injection time. The dynamic exclusion window was set ±7 ppm with a duration at 60 sec. Following each MS2, the 10-notch MS3 was performed on the top 10 most abundant fragments isolated by Synchronous Precursor Selection (SPS). The precursors were fragmented by HCD at 60% CE then detected in Orbitrap at m/z 110-400 with 50,000 resolution for peptide quantification data. The AGC was set 50,000 with maximum injection time at 86 ms.

Personal Proteome Database

GENCODE v27 \cite(*24*) annotation was lifted over from GRCh38 to each EN-TEx donor's personal genomes, to generate 8 sets of GFF annotations. The GFFRead utility \cite(*25*) was used to extract the nucleic acid sequence for all protein coding transcripts. An in-house python script was then used to translate each protein coding transcript into its amino acid sequence. All protein sequences from the 8 genomes were combined with GENCODE v27 reference, redundant sequences were removed and each unique protein sequence was given a unique accession id that included the genomes which contain the protein. The final database contained 128,063 unique protein sequences, 82,136 (64%) from GENCODE reference and 45,927 (36%) unique to the EN-TEx donors. 6344 protein sequences from GENCODE (8% of reference proteome) did not match to any of the alleles in the 4 individuals. File: Supp_data_proteomics.xlsx on the entex resource website provides cross mapping of protein accessions to ENSEMBL transcript and gene ids. Decoy protein sequences were generated using the DecoyPYrat tool \cite(*26*).

MS Identification and Quantification

Spectra were processed using ProteomeDiscoverer v2.4 (Thermo Scientific) and searched against the personal proteome database using both Mascot v2.4 (Matrix Science) and SequestHT with target-decoy scoring evaluated using Percolator \cite(*27*). The precursor tolerance was set at 30ppm, the fragment tolerance set as 0.5 Da and spectra were matched with fully tryptic peptides with a maximum of 2 missed cleavages. Fixed modifications included: carbamidomethyl [C] and TMT6plex [N-Term]. Variable modifications included: TMT6plex [K], oxidation [M], carbamyl [K], methyl [DE], deamidation [NQ], acetyl [N-term]. The carbamyl and methyl modifications were included due to their high incidence after samples were exposed to high concentrations of urea during the RNA extraction process. Peptide results were initially filtered to a 1% FDR (0.01 q-value). The reporter ion quantifier node included a TMT-11-plex quantification method with an integration window tolerance of 15 ppm and integration method based on the most confident centroid peak at MS3 level. Protein quantification was performed using unique peptides only, with protein groups considered for peptide uniqueness. Peptides were quantified and normalised using the TMT isobaric tags. Peptide results from ProteomeDiscoverer were remapped to the protein database and marked as reference, genome, or allele specific(see File: Supp_data_proteomics.xlsx). Gene level quantification of proteins was conducted by summing normalised unambiguous peptide TMT tag intensities.

Peptide and Gene Identification and Quantification Results
At a 1% FDR we report 256,512 peptide to spectrum matches (PSMs), and 117,934 distinct peptide sequences (0.01 q-value at peptide level) of which 45,276 were quantified using TMT isobaric labels. Personal peptides were further filtered to unambiguously match one gene and have a Posterior Error Probability below 0. 699 of these peptides did not map to the reference genome, only matching personal protein sequences. 4489 peptides identified were not present in all 8 genomes across the 4 donors, 830 of these peptides were missing in 1 or more of the donors completely and 4334 were only present on a single allele in at least one of the donors. This corresponds to 13% coverage of the possible observable personal peptides across all protein coding genes in the personal genomes, and a 1% increase in the number of significant distinct peptide sequences quantified. (Figure S1.3h)

Gene quantification was conducted using only unambiguous peptides summing the peptide isobaric tag intensities. 9242 genes were quantified, 540 genes had non-reference peptides, 1333 genes had peptides not present in all 8 genomes (personal peptides), 518 genes had peptides absent in at least 1 donor and 1260 genes had peptides specific to a single allele in at least 1 donor.

RNA-seq comparison
For comparison between proteomics and RNA-seq abundances, a paired set of samples and confidently identified genes matching between the proteomic and RNA-seq datasets were extracted. In each dataset the values were normalised and then scaled to the maximum value across the samples/tissues. A Pearson correlation was then used to test the similarity between the two sets across the samples.

Novel Gene Discovery
All spectra were also processed via the ICR GENCODE OpenMS novel peptide discovery proteomics pipeline \cite(*28*) against a database containing GENCODE v27 reference proteins and a set of potential novel protein coding sequences including many unannotated PhyloCSF conserved regions \cite(*29*). Novel peptide results were filtered according to high stringency criteria \cite(*30*). This resulted in 291 novel peptides, these were further filtered to remove peptides that could be explained by semi-tryptic cleavage or single amino acid variants. The 27

remaining peptides were assessed validating 8 novel protein models, which have all now been annotated in the GENCODE reference set (Figure S1.3i).

All spectra, results and supporting files including the personal proteome database have been deposited in the PRIDE \cite(*31*) proteomic repository (https://www.ebi.ac.uk/pride/) under project accession: PXD022787

S1.4. Mapping Functional Genomics Data to the Personal Genomes

Typically, the analysis of NGS data makes use of a current version of the reference human genome; this includes SNV detection, DNA methylation, transcriptional analysis (RNA-seq), identification of TF binding sites and histone modifications (ChIP-seq), and analysis of 3D chromosomal looping interactions (Hi-C). Depending on the application, this process usually includes selection of either more lenient or more stringent mapping criteria, for example allowing reads to map to single or multiple genomic regions with varying numbers of permitted mismatches \cite(*32, 33*). Using the personal genome to map functional genomics reads becomes particularly important if the mapping requires more stringent criteria, which is necessary when the goal is precision and individualization.

We used DNA from transverse colon tissues to construct both haplotype sequences for each individual. Mapping sequences to the derived haplotypes, rather than to the reference genome, resulted in an overall improvement in mapping accuracy across different assays (RNA-seq, DNA-seq, Hi-C, and ChIP-seq). By applying conventional mapping criteria, we observed an increase in the number of mapped reads of about 0.5 to 1%. When we applied more stringent filtering criteria to select for high-quality, uniquely mapping sequences, we observed a much larger improvement, reaching an increase of 2-4% across assays over the four individuals (Fig. 1B). We also generated a list of genes differentially expressed between mapping to personal genome vs reference genome (see File: Supp_DE_genes.tsv). Figure S1.4 summarizes the numbers of reads and percentages for precision mapping across four individuals for DNAseq, CHIPseq, HiC and RNAseq. Mapping categories include mapping to haplotype 1 (Hap.1), haplotype 2 (Hap.2), union of Hap.1 and Hap.2 (Hap1&Hap2), reference (ref), intersection between Hap.1 and Hap.2 but not ref (Hap1&&Hap2¬Ref) and improvement as a measure of Gain=[(Hap.1 ∪ Hap.2)-Ref]/Ref.

For all assays, we excluded counting reads that mapped to X, Y and M chromosomes for all individuals. In general, to ensure high quality mapping we selected reads with at most 2 mismatches and unique mapping. We used raw reads from transverse colon, publicly available at Encode portal, with the exception of DNA-seq. For DNA-seq mapping, we used reads from blood samples that we obtained from GTEX to avoid any bias deriving from the construction of haplotypes using DNA sequences. For DNA-seq and RNA-seq mapping we used paired-end reads. For RNA-seq, to account for gene splicing, we used *.gtf files with transcript genomic coordinations and STAR Aligner v2.7. For DNA-seq, Hi-C and ChIPseq we used BWA v0.7.17 and selected reads with at most 2 mismatches and quality Q>30. For RNA-seq we used sequences with quality mapping Q=255.

S2. Supp. content to main text section "Measurement of allele-specific activity in diverse assays"

S2.1. Allele-specific Expression (ASE), Binding (ASB) and Chromatin Accessibility (ASCA)

S2.1.1. Overview

Allele-specific phenomena arise from differential activity and/or modifications between the two haplotypes of the same individual, which can lead to allele-specific RNA and protein expression in a specific tissue/cell type. Most frequently, measurements of allele-specific activity are performed using RNA-seq, ChIP-seq, ATAC-seq, or DNase-seq data \cite(*34-39*). The EN-TEx project simultaneously and uniformly characterizes almost all the functional genomics activity (i.e., expression, binding, methylation, etc) at every heterozygous locus of each individual.

Allele-specific expression and binding were measured with an extended version of the AlleleSeq pipeline. AlleleSeq \cite(*8*) was originally developed for the 2012 ENCODE rollout \cite(*8, 40*). It was subsequently refined for a number of applications, including the 1KGP functional interpretation group and for other projects \cite(*37, 38, 41*). Broadly, the pipeline incorporates personal variation, including large structural variants, which allows it to account for reference bias \cite(*8, 37, 42*) in a straightforward way. In addition, we have included additional filters to mitigate ambiguous mapping biases \cite(*37, 43*). In order to account for the overdispersed nature of the functional genomic readcount data, the significance of the allelic imbalance is assessed with the beta-binomial test \cite(*37*)(Figure S2.1a).

S2.1.2. Mapping

For each available replicate of the EN-TEx experiments, functional genomics reads were mapped to both personal haplotypes simultaneously using STAR-2.6.0c \cite(*44*). We required stringent mapping criteria, allowing maximum number of mismatches at 3% of the read length. For both the ChIP-seq and ATAC-seq datasets, mapping was performed forbidding spliced alignments. Adapters were also removed from the ATAC-seq reads with cutadapt \cite(*45*). For RNA-seq data we used Gencode v24 \cite(*24*) annotation converted to personal coordinates and the mapping was performed in the 2-pass mode to identify and incorporate novel junctions. Read duplicates were identified and removed from all alignments using picard (http://broadinstitute.github.io/picard/)

To visualize functional genomic reads on individual haplotypes (e.g., Fig. 5B), we use samtools (1.9) \cite(*46*) to extract haplotype-specific reads from the bam files generated by STAR from the last step. If an assay has multiple replicates, we merged all the bam files. The number of reads mapped to a given region in the personal genome was calculated by bedtools (2.29.2) \cite(*47*) and stored in bedgraphs, lifted over to the reference genome with UCSC liftOver \cite(*48*), and converted to bigwigs with bedgraphToBigWig (2.8) \cite(*49*). Figure S2.1b summarizes the pipeline to generate the haplotype-specific bigwigs. The bigwigs are displayed with Integrative Genomics Viewer (igv) \cite(*50*). See Figure S2.1c for accession numbers of data used to generate the signal tracks in Fig. 4-5.

(File: xxx)

S2.1.3. Filtering and assessment of read imbalance at heterozygos SNVs (hetSNVs)

The number of reads overlapping each hetSNV and carrying the corresponding alleles was calculated after filtering. The filtering included:

- potentially misphased loci
- reads bearing an incorrect allele
- hetSNVs located in potential CNV sites through assessment of surrounding read-depth (+/- 1kb)
- Sites with potential ambiguous mapping \cite(*37, 43*)
- Non-autosomal chromosomes: we generated separate call-sets that include chr X for the female individuals but (unless specified otherwise) all downstream analyses were performed on call-sets that only include autosomal loci.

We aggregated read counts from all replicates available for each experiment (sample). We then calculated the significance of the imbalance at each heterozygous loci as described previously \cite(*37*) and called ASE and ASB sites at FDR 10%. (Figure S2.1d-e).

(File: xxx)

We provide the read counts and p-values for all the ASE and ASB that are either significantly imbalanced or accessible (SNVs that have at least the minimum number of reads needed to be statistically detectable for allelic imbalance). Columns in the hetSNV files:

1) chr	: chromosome
2) ref_start	: GRCh38 locus start position (0-based)
3) ref_end	: GRCh38 locus end position (1-based)
4) ref_allele	: reference allele
5-6) hap1_allele/hap2_allele	: haplotype 1/2 allele
7) experiment_accession	: ENCODE experiment ID
8) donor	: ENTEx individual
9) tissue	: tissue
10) assay	: assay
11-14) cA/cC/cG/cT	: number of reads with A/C/G/T
15) ref_allele_ratio	: number of reads with reference / total number of reads
16) p_betabinom	: p values calculated from the beta-binomial test
17) imbalance significance	: '1' passes the FDR10% threshold, '0' not a significantly imbalanced site.

S2.2. Allele-specific Methylation (ASM)

We used WGS variant calls to determine the positions of heterozygous SNVs (hetSNVs) and identify all homozygous CpG positions in the genome of each donor. With such information, and with the fully processed tissue-specific WGBS aligned reads, an in-house script was then used to identify positions exhibiting significant allelic differences in CpG methylation. Our script counted the number of times a methylated or unmethylated homozygous CpG occurred in the same read as each of the two possible alleles at the hetSNV position for autosomal chromosomes. If the same read overlapped multiple CpGs, they were each considered as independent observations. Reads with a low-quality score (Phred < 20) on the SNP position, or with a base call that did not match either of the two alleles expected in that position based on the WGBS calls, were discarded. Due to the nature of bisulfite sequencing data, where cytosines may be observed as thymines due to bisulfite conversion, it was not possible to determine which allele the read came from in several cases. In such cases, the read was also discarded. If a low-quality score, or an unexpected base call, was observed on a CpG position for a particular read, that observation did not contribute to the final counts. The significance of the association between the allele at the hetSNV position and the methylation state of the CpGs in the 300bp surrounding region was assessed using Fisher's exact test. The 300bp windows surrounding the hetSNV position were chosen as the WGBS dataset was composed of paired-end 150bp reads. The test was only performed for hetSNV positions that showed a minimum of 6 observations of either a methylated or unmethylated CpG position for both alleles. Once the p-values were computed for all such hetSNV positions, the Benjamini-Hochberg procedure was used to control the false discovery rate for such associations. The difference in the level of methylation between alleles was also computed for each hetSNV. Any positions overlapping CNVs, or small INDELS, were discarded from further analyses. Finally, ASM calls were made by identifying the heterozygous SNP positions with false discovery rate (FDR) below a specified threshold (10%), and absolute difference in methylation between alleles above a minimum threshold of 10%.

(File: xxx)

S2.3. Hi-C - Allele-specific interactions

S2.3.1. Creation of haplotype specific interaction matrices

Each pair of the paired-end reads were aligned separately to both of the parental haplotypes using BWA-MEM \cite(*18*). Sequencing reads are then paired based on their read names. Each paired-end read is then assigned to either one or both of the parental haplotypes as follows: for each paired-end read, a score is assigned to each parental haplotype based on the number of mismatches of the mapping to that haplotype. Paired-end reads are then either assigned to haplotype 1 or haplotype 2 based on their corresponding score. In brief, pairs of reads are assigned to a haplotype if they map exclusively or with a better score to that haplotype. Additionally, pairs of reads that exclusively map to one of the haplotypes are also assigned to that haplotype. After every paired-end read is assigned to a parental haplotype, chimeric reads and PCR duplicates are removed and we generate an interaction matrix for each haplotype of each tissue of each individual (see Figure S2.3a for the pipeline and Figure S2.3b for the matrices).

S2.3.2. Allele-specific interactions

For each significant interaction captured by Fit-Hi-C, we found the number of reads that map to haplotype 1 and haplotype 2 using the haplotype specific interaction matrices. If there is a difference in the number of reads that mapped to a haplotype vs. the other one, we then calculated the p-value for the significance of the allelic imbalance using a binomial test. File: hic_files.tar.gz contains two folders: “ref” and “pgenome”. “ref” folder contains .hic files for each individual and tissue (each tissue and individual combination is a separate folder, totaling up to 8 folders); these files contain information of the genome-wide interaction matrices. The information can be extracted using juicer tools and the contact matrices can be visualized using juicebox (see Figure S2.3b for an example). “Pgenome” folder contains two sub folders: “hap1” and “hap2”. Each of these folders contain two .hic files for each chromosome of each individual and tissue. Chr*.hap*.hic files contain the Hi-C data for that chromosome in personal genome coordinates and Chr*.hap*2ref.hic files contain the Hi-C data for that chromosome in a reference genome coordinate (lifted over using personal genome chain files). Figure S2.3c shows the total number of raw allele specific interactions and significant allelic imbalances per sample (calculated using the binomial test described above).

S2.4. Proteomics - Allele specific Peptide (ASP) Analysis

The proteomics data were mapped at the gene level and filtered to a set containing one or more allele specific peptides in any donor. These fell into two categories: genes with allelic specific peptides for one allele only or those with peptides specific to both alleles. Both groups were considered for ASP ratios. The ASP ratios were calculated for each tissue and donor in which allelic peptides were quantified, based on the ratio of the summed peptide intensities of peptides specific to the two alleles. Individual ASPs were filtered to require a minimum of 3 distinct peptides unambiguously identifying a gene, an expression level for the tissue that was not less than 5 fold lower than the highest expressed tissue and an ASP ratio that was greater than 0.75. Figure S4.3c summarizes key numbers of genes with allelic peptides. See File: Supp_data_proteomics.xlsx for a full list of allelic peptides.

S3. Supp. content to main text section “Aggregation of allele-specific events, forming a catalog”

S3.1. Assessment of read imbalance at genomic elements

S3.1.1. Genes, binding peaks and cCREs

We extended our pipeline to measure allelic imbalance at genomic regions and elements of interest. To do so, we aggregated read counts from all hetSNVs within the relevant region and assessed the significance of imbalances between personal haplotypes for individual hetSNVs as described above. We provide a large catalog of genomic elements measured for allelic activity (e.g., ASE genes, ASB peaks, cCREs etc) with corresponding haplotype-specific assay read counts and significance of the imbalance (Figure S3.1a-b).

(File: xxx)

1) chr : chromosome
2) start : GRCh38 locus start position (0-based)
3) end : GRCh38 locus end position (1-based)
4) region_id : gene name (gencode v24) or peak/cCRE id
5-6) hap1_count/hap2_count : number of reads mapped to haplotype 1/2
7) experiment_accession : ENCODE experiment ID
8) donor : ENTEx individual
9) tissue : tissue
10) assay : assay
16) p_betabinom : p values calculated from the beta-binomial test
17) imbalance significance : ‘1’ passes the FDR10% threshold, ‘0’ not a significantly imbalanced site.

S3.1.2. Methylation

(Text)

S3.1.3. Correlation between AS genes and diseases

We compared the set of allele specific genes to a set of genes associated with certain diseases. The list of disease genes are genes known to be affected by disease-associated mutations and expressed in disease-related tissues \cite(*51*). For every tissue and individual, we noted the genes that were present in both the set of AS genes and the set of disease genes. The list of the overlapping genes and their associated diseases can be found in File: Associated_AS_Disease_Genes.xlsx

S3.2. Aggregation across tissues and assays

S3.2.1. Allele-specific expression and binding

We observed a large increase in detection power when we pooled reads for each hetSNV across all tissues in each individual. For each individual, we pooled reads for each hetSNV across all tissues. We calculated the significance of the imbalance at each hetSNV for the pooled call-set in the same manner as for individual tissues and called ASE and ASB sites at FDR 10%.

S3.2.2. Methylation

We aggregated the counts of methylated and unmethylated CpG position surrounding both alleles of each heterozygous SNV across tissues for each individual to assess the cross-tissue association between the allele at the hetSNV position and the methylation state of the CpGs. Significance of association is computed using Fisher's exact test and Benjamini-Hochberg procedure was used to control the false discovery rate. For aggregated observation, the test is only performed for hetSNV positions that showed a minimum observation of 12 methylated or unmethylated CpG positions for both alleles. ASM were called at FDR 10% and absolute methylation difference larger than 10%.

We also generated a combined ASM call set that includes cross-tissue counts of methylated and unmethylated CpG observations surrounding "accessible" hetSNVs for all four individuals. We defined accessible hetSNVs as those that pass the threshold of at least 12 observations of either methylated or unmethylated CpG positions in the surrounding area. We annotated the hetSNVs with associated genes, distance to gene, genomic region, and alternative allele frequency based on refGene and gnomAD 3.0 databases using ANNOVAR \cite(*52*). We also annotated the hetSNVs with cCREs from the ENCODE encyclopedia.

S3.3. High-confidence and high-power call-sets

We also developed a "high-confidence" call set requiring that at least one read from both alleles is detected in the functional genomics assay, thus accounting for potential false positive genotype calls.

In addition, we generated a "high-power" tissue-specific call-set (see File: AS_highpower_set.tar.gz) by allowing a more relaxed threshold (FDR 20%) for loci that were detected as significantly imbalanced after read pooling-based joint calling across all tissues (Figure S3.3).

(File: xxx)

S4. Supp. content to main text section "Mining the catalog"

S4.1. Integration with the ClinGen Allele Registry

The variants identified in all four EN-TEx individuals are registered to the ClinGen Allele Registry \cite(*53*) which provides unique variant identifiers for canonical alleles defined at the level of nucleic acid sequences or at the protein level. The unique identifier integrates different types of labels and definitions of the same allele across multiple databases including dbSNP, gnomAD, ClinVar, and ExAC. All variants are bulk registered in vcf format using API specified by Allele Registry documentation (http://reg.clinicalgenome.org/doc/AlleleRegistry_1.01.xx_api_v1.pdf). Query of variants can be done either programmatically via APIs or via search interface using any type of id associated with the variant. Metadata for allele(s) are available in machine readable form (JSON).

S4.2. Effect of Allele Specificity on Purifying Selection

In order to calculate the purifying selection on allele specific events, population scale variants from three cohorts were used. We used two measures of purifying selection and conservation for this analysis. The first is rare DAF, which is calculated as #rare/(#rare + #common) variants falling in a given allele specific region. In order to categorize variants as rare or common, ancestral alleles were used and a MAF of 0.05 was used. This is a commonly used metric for calculating selection in populations \cite(*37, 38, 41*). The second was phastCons, which measures the cross-species conservation \cite(*54*). All purifying selection analyses were performed for AS+ cCRE, AS- cCRE, ASB+ H3K27ac, ASB- H3K27ac regions and ASE+ and ASE- genes. The results can be seen in Figure S4.2.

S4.3. Allelic effect prediction with the BERT model

BERT is a natural language model based on Transformer neural network architecture. It has been widely applied to natural language processing due to its ability to incorporate long-range contextual information \cite(*55*). Thus, it could also be applied to extracting meaningful sequential patterns from genomic sequences, such as prediction of allelic effects of SNPs.

We extract the 128bp sequence upstream and downstream of the SNP in question as the input. The sequences are labeled as positive or negative based on their allelic effects. For balancing considerations, the negative set is randomly downsampled to the same size as the positive. The dataset is then split into training, cross-validation and testing set by 8:1:1.

We initialize the BERT model with the weights of the pre-trained DNABERT model \cite(*56*). A single-layer classifier is added on top of the output of DNABERT and the model is fine-tuned on the allelic effect datasets. For fine tuning, we selected from a range of hyperparameters (learning rate=1e-5, 5e-5; training epoch = 5, 10, 20). As the pre-trained DNABERT model has different versions with k-mer size 3~6, we report the model with the highest performance.

The model is first trained with only SNPs from donor individual 3. For many of the prediction tasks, the model achieved performance of f1 score > 0.7 and accuracy > 0.7 on the validation set, significantly higher than logistic regression on sequence embeddings (Figure S4.3; see below for more details). We then tested the model performance on validation sets composed of SNPs exclusive to the other three donors. Specifically, the validation sets for these three individuals have been randomly downsampled to the same size as the validation set for individual 3. The sampling is repeated 10 times and average results are reported. As expected, the performance is lower compared to individual 3.

For model interpretation, we used the method implemented by \cite(*56*), where the attention scores of the last layer for the first token are averaged over all 12 attention heads, and then regularized by k-mer coverage.

As a comparison, we use the dna2vec model released by \cite(*57*) to transform k-mers to continuous-valued vectors, preserving their contextual preference. Using the same training, test and validation data as above, we represented each input sequence as an average over the embedding of all its k-mers. We then trained a logistic regression classifier based on the average embedding vector. We tried embedding with k-mer size 3~8 and reported the one with the highest performance.

S4.4. Cross-assay Compatibility

S4.4.1. Compatible & incompatible: single chromatin mark vs gene expression

Using the methods described in previous sections, we identified promoters (± 2Kb from TSS) with allelic imbalance in the chromatin state measured by H3K27ac, H3K27me3, etc. We determined the compatibility between allelic promoter chromatin states and allelic gene expression in a straightforward way. The allele with more active promoter chromatin should have a higher expression level, otherwise the promoter and the gene are incompatible. Similarly, alleles with more repressed promoter chromatin are compatible with lower expression levels. We treated histone mark H3K27ac, H3K4me3 and H3K4me1, chromatin openness indicated by ATAC-seq or by DNase-seq, and the binding of EP300, POLR2A, POLR2AP, and CTCF, as marks of active chromatin. Histone mark H3K27me3 and H3K9me3, and CpG methylation were considered marks of repressed chromatin.

Because allelic gene expression and/or allelic chromatin state can be tissue-specific and/or individual-specific, we did not merge compatible (or incompatible) promoter-gene pairs that appear in multiple samples. Overall, the number of ASE genes compatible with at least one of the 13 marks is 35 per tissue per individual, while the number of ASE genes suitable for the compatibility analysis (i.e., the promoters of these genes are accessible for measuring the potential allelic chromatin state indicated by any of the 13 marks) is 226 per tissue per individual. See Figure S4.4a for more compatibility results.

We note that some assays were performed twice for a given tissue of a given individual. For example, the RNA-seq of individual 3's liver includes two experiments (ENCSR226KML and ENCSR504QMK), while there is only one H3K27ac ChIP-seq experiment for the same sample. In another example, there are two CTCF ChIP-seq experiments for individual 3's spleen (ENCSR756URL and ENCSR773JBP), while there is only one RNA-seq experiment for the same sample. In these cases, we combined ASE genes or ASB promoters that were called from either of the duplicated experiments, excluding those where the directions of the allelic imbalance are the opposite in the two experiments. The combined ASE genes and ASB promoters were analyzed for compatibility. File: Supp_data_compatibility.xlsx lists the compatibility of genes with AS expression in each tissue and individual.

To test the numbers of compatible versus incompatible promoter-gene pairs, we identified genes that have allele-specific expression in at least one tissue of at least one individual. We shuffled the gene-promoter relation for these genes and calculated the ratio of N_compatible versus N_incompatible. We repeated this process 1,000 times for each chromatin mark (after excluding replicates where N_incompatible is zero) to calculate a Z-score of the ratios shown in Fig. 3E.

We also checked the association between AS chromatin state of the regulatory sequences and the AS expression of the corresponding genes while ignoring the compatibility between the two. Figure S4.4b shows that genes with AS expression are more highly enriched near promoters with AS methylation and/or TF binding than near non-regulatory sequences with AS methylation.

S4.4.2. Compatibility with AS Proteomics

Of the high stringency ASP (allele specific peptides) set, 114 overlapped ASE events calculated from RNA-seq data, 58 showed compatibility and 56 showed incompatibility (Figure S4.4c). The z-score 0.26 of the ratio of the compatible to the incompatible (based on ASP/ASE pairs being randomized 1000 times) was not significant indicating the compatibility between the RNA-seq and protein level allele expression is near random. Although some of this incompatibility is likely down to technical issues, manual examination of the most biased ASPs overlapping ASEs, shows that the evidence for allelic specific protein expression is very compelling implicating post translational regulation (Figure S4.4d). For some of the incompatible cases there are clear biological reasons for the difference between ASP and ASE ratios such as frameshift variants. File: Supp_data_compatibility.xlsx shows details of the significant 114 ASPs overlapping ASEs.

S5. Supp. content to main text section “Examples of coordinated AS activity across assays”

We detected the inactive copy of X chromosome in the majority of tissues from female individuals. We found that both gene expression and active chromatin signal is significantly skewed toward one haplotype, while repressive chromatin signal is significantly skewed toward opposite haplotype in many tissues.

We then investigated the specific allelic coordination at the chromosome level using the X chromosome. X chromosome inactivation ensures that females have only one functional copy of the X chromosome, and occurs by random selection of the inactivated copy early in embryonic development. For the two female individuals, the EN-TEx data enabled comprehensive analysis of the allele-specific activity in 24 tissues. We identified the active copy of the X chromosome by examining the overall gene expression levels in 24 tissues of individual 3. We identified the active copy of the X chromosome by examining the overall gene expression levels. The gene expression, active histone marks, and repressive histone marks were coordinated in terms of their haplotype-specific activity. As shown in Fig. 4C, the gene expression values of all genes in the X chromosome were higher in haplotype 2 than those in haplotype 1 (see Figure S5.1a and Figure S5.1b). In accordance with this finding, enrichment of the active histone mark H3K27ac was also higher in haplotype 2 than in haplotype 1. Moreover, enrichment of the repressive histone mark H3K27me3 was imbalanced in the opposite direction (i.e., higher in haplotype 2). A similar coordination was observed using other allele-specific activity such as POL2R and CTCF binding (Figure S5.1a).

When focusing on specific genes, we found that the DHRSX gene located on the pseudo-autosomal region had balanced expression in both haplotypes, whereas the SLC25A5 gene located on the inactivated region showed a significant skew in gene expression towards the active haplotype in accordance with the chromosomal level imbalance. We also identified a known “escaper” gene, KDM6A \cite(*58*), that demonstrated balanced expression in both haplotypes while being located on the inactivated region of ChrX. The expressed haplotype of the allele-specific SLC25A5 gene in tibial nerve tissue showed significant allele-specific activity for the activating histone mark H3K27ac, while the “inactive” haplotype had significant allele-specific activity for the repressive histone mark H3K27me3 as well as DNA methylation. In addition, our analysis of haplotype-specific Hi-C data revealed an allele-specific skew in Hi-C interactions between another gene, XACT, and its potential distal regulatory element on the active haplotype of Chr X (see Figure S5.1c).

We found an allele-specific activity example of a less characterized locus. We detected allele-specific Hi-C interactions in the XACT locus (Figure S5.1c) on the active copy of chromosome X. We first determined the active copy of chromosome X by looking at the gene expression distribution on both haplotypes and found that haplotype 2 has more gene expression than haplotype 1. We then looked at the differential interaction of chromosome X by subtracting the Hi-C matrices of the haplotypes. We found that an interaction between the XACT locus and an region upstream of it is significantly elevated in the active haplotype. We also found that both XACT locus and the upstream region are bound to CTCF, which might be mediating the interaction. XACT is a long non-coding RNA found to be active in the active copy of chromosome X early in cell development. This CTCF mediated haplotype-specific interaction could play a role in activating the XACT locus established at early stages of cell development. While such observations are interesting, they are provisional on additional supportive data.

S6. Supp. content to main text section “Relating SVs to chromatin & expression”

S6.1. Associate SVs with eQTLs

We are interested in heterozygous SVs that potentially cause allelic gene expression and underlie the action of known eQTLs. To do this, we first identified eQTLs \cite(*59*) that are compatible with the allelic expression of the associated genes. For each ASE gene, we check if the two alleles at each associated eQTL locus have the expected regulatory effect. We used the compatible eQTLs associated with a given ASE gene to define a window, which spans from –10 kb of the compatible eQTL on the far 5' end to +10 kb of the compatible eQTL on the far 3' end. For a heterozygous SV that intersects with this window, we determined whether the SV and the compatible eQTLs may locate on the same linkage block by comparing their allele frequency and haplotype. Specifically, for each SVs identified in the last step, we identified all compatible eQTLs (with respect to the given ASE gene) that fall within +/- 10 kb of the SV. Suppose the SV is on haplotype 1, then we calculated the allele frequencies of the alleles of the compatible eQTLs on haplotype 1. Here, we used the allele frequency reported by the 1000 Genome project high-resolution data \cite(*60*) for the alleles at each compatible eQTL locus. We excluded eQTLs for which the derived alleles cannot be found in the 1000 Genome project. If at least half of the hap1-alleles of the compatible eQTLs within +/- 10 kb of the SV have similar allele frequencies as the SV's allele frequency (defined between 80% to 120% of the SV's allele frequency), then we consider the SV is potentially linked to the compatible eQTLs and may contribute to the allelic expression of the given gene. We listed SVs that meet this criteria, the associated ASE gene, and the compatible eQTLs +/- 10 Kb from the SVs in File: Supp_Data_SVs_associated_with_eQTL.xlsx.

We identified known eQTL-associated SVs (including SV-eQTLs) \cite(*15, 61*) in our list of potential eQTL-associated SVs. We consider that our SV hits a match if within +/- 100 bp of this SV a reported eQTL-associated SV was found and both SVs are associated with the same gene. We searched for matches in tissue-specific and non-tissue-specific ways. For individual 2, our list includes 193 SVs that are associated with eQTLs in at least one tissue, and 44 of them match known eQTL-associated SVs. The numbers are 174 and 33 for individual 3. Details of these results are listed in Supp_Data_SVs_associated_with_eQTL.xlsx. For comparison, we also calculated the fraction of known eQTL-associated SVs in our SVs that are close to genes with AS expression. We pooled genes that have AS expression in at least one tissue. Because GTEx eQTLs fall in +/- 1Mb from the TSS of genes \cite(*59*), we used the same window to look for SVs near the genes with AS expression, requiring the SVs to at least partially overlap with the windows. We further required SVs to be heterozygous, clearly phased, and relatively common (i.e., present in Audano et al. (2019) \cite(*12*)). We found 4415 SVs in individual 2 that meet these criteria, of which 560 match known eQTL-associated SVs. This fraction is significantly lower than the observed fraction of 44/193 (p = 4.6e-5, Chi-square test). For individual 3, the expected fraction is 594/3909, which is lower than the observed fraction of 33/174, but not significant (p = 0.18, Chi-square test).

S6.2. Aggregating the Impact of SVs on Neighboring Chromatin

Our goal is to calculate potential changes in the chromatin state in the neighborhood of SVs. Intuitively, this can be done by comparing the chromatin state between individuals with and without the given SVs. We excluded SVs that are closer than 5 kb from any other SVs in individual 2 or individual 3, and SVs that fall on the sex chromosomes. In the remaining SVs, we focused on heterozygous SVs that have relatively precise breakpoints. Specifically, we kept SVs where the total length of the start position's confidence interval and the end position's confidence interval is at most 50 bp. To minimize the influence of SVs on mapping sequence reads, we further excluded SVs for which the average mappability of a window +/- 500 bp of the SV is below 0.9. Because EN-TEx requires the length of a ChIP-seq read to be at least 50-bp, we used the 50-mer multi-reads Umap mappability \cite(*62*) when filtering SVs for the purpose of calculating potential disruption to chromatin openness (measured by ATAC-seq) and H3K27ac. We also excluded SVs that fall in blacklist regions that are known to give problematic ChIP-seq reads \cite(*63*). 1974 SVs in individual 2 and 2154 in individual 3 passed all the filters above.

For each SV that passed the above filters, we calculated the average chromatin state in the SV's flanking regions. We define that flanking regions of a SV as the -500 bp ~ -100 bp region and the 100 bp ~ 500 bp (Figure S6.2a) – the extra 100 bp upstream and downstream of the SV are extra buffer regions which should reduce the influence of SVs on mapping ChIP-seq reads. The average chromatin state is the fold-change over control of the number of ChIP-seq reads averaged over all base pairs of the flanking regions. Knowing the genomic coordinates of the flanking regions, we can calculate the average chromatin state in the corresponding regions in the other individual who does not carry the given SV (Figure S6.2a). If the average chromatin state in the flanking regions of a SV is lower than 70% of that in the wild type individual, we consider the given SV reduces the chromatin state in the neighbourhood. Because the chromatin state can be tissue specific, we treated the SV neighbourhoods as tissue-specific - even for the same SV (Figure S6.2a). We pooled the SV neighbourhoods from all tissues, and binned neighbourhoods based on their chromatin state in the wild type individual. We reported for each bin the fraction of neighbourhoods with reduced chromatin state due to the presence of the SVs.

In theory, we could also investigate the impact of SVs by comparing between the two haplotypes of the same individual. To do this, there needs to be heterozygous SNPs and/or indels present in the neighborhood of SVs in order to determine the chromatin states of each haplotype. However, we found many SVs don't have heterozygous SNPs accessible for determining the chromatin states in nearby regions.

S7. Supp. content to main text section "Decorating the ENCODE encyclopedia"

S7.1. Signal normalization method

In order to overcome batch effects, matrices of gene expression and histone marks' values were quantile-normalized across samples (tissues and donors). The choice of quantile normalization method was made after performing a benchmark of a number of normalization methods. The methods selected for the benchmarking are among the ones analyzed in a recent publication \cite(*64*): quantile normalization, smooth quantile normalization, upper-quartile normalization, variance stabilization normalization (VSN) and local regression normalization (two variants: LoessF and LoessCyc). These are normalization techniques widely applied also in other bioinformatics fields, such as microarray and proteomics analysis. The pilot analysis was performed independently for two cell lines, K562 and GM12878, for which different polyA+ RNA-seq evaluation datasets were produced by Wold, Gingeras and Graveley labs during the ENCODE phase 2. The benchmark consisted of three steps: i) for each method, we computed the distribution of Pearson's and Spearman's correlation coefficients across all genes between each pair of samples; ii) we then ranked the methods based on the mean of the distribution of all genes' variance across samples \cite(*65*), and on iii) the Relative Log Expression (RLE) distribution (distribution of $\log_2$ ratio for a given gene between one particular sample and the median across all samples), which should be close to 0 \cite(*66*). Overall, quantile and smooth quantile normalization techniques performed similarly between each other and better than the other methods. We thus opted for quantile normalization. In particular, for each of the histone modifications used in the decoration procedure below, we provide the quantile normalized fold change signals of cCREs across all the available tissues and individuals. The data file for each of the histone modifications is a data matrix, in which each row corresponds to a cCRE and each column corresponds to a tissue from an individual. As a result, the element in the matrix is the quantile normalized signal of the histone modification observed in the cCRE from a tissue. These files are available in a tar ball: cCRE_histoneSignals_qnorm.tar.gz.

S7.2. Decoration of Regulatory Annotations

We used the ChIP-seq datasets of both active and repressed marks to decorate (i.e., re-annotate) the cCREs from the Encyclopedia, which are based on a set of high-quality DHSs \cite(*67*). The ENCODE encyclopedia regulatory elements consist of 0.9 million cCREs averaging ~400bp. For each type of functional genomic data, we normalized the activity signals of the cCREs from all tissues and focused on the cCREs with relatively strong signals (Figure S7.2a). In the decoration, we considered three active marks (H3K27ac, H3K4me1, and H3K4me3) and three repressed marks (H3K27me3, H3K9me3, and DNA methylation). ChIP-seq datasets were uniformly processed using the ENCODE standard pipeline, including alignment, quality control, and peak calling. With the uniformly processed ChIP-seq datasets, the average epigenomic signals were calculated and normalized for a registry of cCREs from the Encyclopedia (Figure S7.2a). Namely, we first calculated the average fold-change against control, typically input DNA, for each cCRE. The average fold-change is quantile normalized independently across experiments but jointly between individuals and tissues. Finally, the scores for each experiment are scaled from 1 to 10. For a particular tissue type, we defined a set of cCREs for each epigenomic mark that are considered as "active" (i.e., thresholding the normalized and scaled quantile values of the cCREs). The thresholding value is calculated for each assay by maximizing the similarity -- the fraction of shared active cCRE -- between the four individuals across tissues. We used the average threshold score across the Transverse Colon, Spleen, and Esophagus since those were the most commonly comprehensive assays across individuals.

For each tissue, we then defined a set of active, repressed and bivalent cCREs based on their active and repressed epigenomic signals, respectively (Figure S7.2b; Figure S7.2c as an example from spleen). Briefly, the active cCREs show high activity for only active marks (i.e., H3K27ac, H3K4me1, and H3K4me3); the repressed cCREs show high activity for only repressed marks (i.e., H3K27me3, H3K9me3, and DNA methylation); and the bivalent cCREs show high activity for both the active as well as the repressed marks. The cCREs were then separated into distal and proximal ones according to their distance to TSSs (proximal as those within 2kb to annotated TSSs). We also intersect these cCREs with the CTCF binding sites from the matched tissue type to define CTCF+ and CTCF- cCREs. Finally, the active and repressed cCREs were further annotated using their allelic signature to identify a set of allelic-specific (AS+) and non-allelic-specific (AS-) ones, respectively. In the allelic-specific decoration, we used the allelic signature from the matched epigenomic marks to define the active/repressed AS+ and AS- cCREs. Any active/repressed cCREs intersecting with the AS+ cCREs were considered to be active/repressed AS+. The active/repressed AS+ from different individuals were pooled together to generate the set of active/repressed AS+ cCREs in the corresponding tissue. We found that the numbers of repressed cCREs are comparable to those of active cCREs in many tissue types, highlighting the necessity of decoration using the repressed markers (Figure S7.2d). Finally, we provided the cCRE decoration results in all the tissue types (see File: cCRE_Decoration.matrix or separately File: active.combined_set.txt.zip, File: bivalent.combined_set.txt.zip and File: repressed.combined_set.txt.zip) (Figure S7.2e).

In order to further subset cCREs, we created an annotation set that focuses on regions with high H3K27ac signals. We call this set the “stringent” annotation set. In order to create this stringent annotation, we intersected the cCRE regions with the top 1% of scored regions as prioritized by the H3K27ac feature from Matched Filter \cite(*68*). This stringent annotation was further used in other analysis, and labeled as “stringent” in the main manuscript and figures. A file containing these stringent regions (bed file) can be found at File: stringent.regions.MF.hg38.bed. These bed regions represent the highest scoring Matched Filter regions.

S7.3. Repressed regions

Gene activation and repression can be mediated through the combination of different histone marks. Historically, much effort has been devoted to elucidating how genes are activated; however, evidence is emerging to demonstrate the requirement of appropriate heterochromatin formation for the preservation of genome stability and the cell type-specific silencing of genes \cite(*69*). In mammalian genome, H3K9me3 and H3K27me3 are well-documented histone marks enriched for “constitutive” and “facultative” heterochromatin, respectively. For genomic regions not containing any active regulatory elements (cCREs), we have identified a set of elements that are marked by either H3K9me3 or H3K27me3 and do not have any active marks (H3K27ac, H3K36me3, H3K4me1 and H3K4me3) nor transcriptional activities as fully repressed in the EN-TEx tissues. Regions within ENCODE4 GRCh38 blacklist (ENCSR636HFF) and GENCODE gene list (GRCh38_v24) were removed. In summary, 45,207 non-overlapping elements of size no less than 200 bp (roughly approximate nucleosome size) are uniquely marked by H3K9me3, spanning 12,655,795 bp (less than 0.4%) of the reference genome, and 24,006 elements by H3K27me3, spanning 7,474,178 bp (less than 0.3%). Identified elements can be found in File: ENTEx_fully_repressed_regions_independent_of_cCREs.bed. As shown in the Figure S7.3, nearly 75% of these elements are specifically repressed in a certain tissue, and the rest also show some degree of tissue-specificity. Although it was known that H3K27me3-enriched facultative heterochromatin contains repressed genes in a cell type-specific manner whereas H3K9me3-enriched constitutive heterochromatin mainly occur constantly at the same gene-lacking regions in every cell type \cite(*70, 71*), observations also suggest that large domains of H3K9me2/3 form in a cell type-specific manner and can influence cell identity by silencing lineage-inappropriate genes and impeding the conversion of terminally differentiated cells into a different cell type, highlighting a role for H3K9me3 in cell type-specific gene regulation \cite(*69, 72-75*).

DNA methylation is one major contributor to gene repression, and has been reported to be interacting with H3K9me3 in chromatin repressive pathways \cite(*76*). We further analyzed the methylation rate of CpG sites within these repressed elements. Whole-genome shotgun bisulfite sequencing (WGBS) of CpG sites were available for 11 EN-TEx tissues that also have H3K9me3 and H3K27me3 ChIP-seq data. For the same tissue from different donors, we aggregated the CpG reads by taking the sum of reads from all donors, and a CpG site is considered to be methylated (meCpG) when the site is covered by at least 5x reads and the ratio of meCpG reads is at least 50%. The overall meCpG rate in each tissue was calculated and used as control to evaluate the meCpG rate in H3K9me3 and H3K27me3 marked elements. As shown in the Figure S7.3, H3K9me3 uniquely marked elements show significantly (t-test, pValue < 0.05) higher meCpG rate than elements uniquely marked by H3K27me3. Compared with the control, H3K9me3 marked regions seem to be hypermethylated, whereas H3K27me3 marked regions to be hypomethylated. This is consistent with the current understanding for constitutive heterochromatin and facultative heterochromatin, of which the former is defined by high levels of DNA methylation and H3K9me3 and the later displays DNA hypomethylation and high H3K27me3 \cite(*77*).

S7.4. Validating Annotations using three-dimensional genome organization

Chromosome compartments that are observed from PCA analysis on Hi-C correlation matrix give insight into activity level of the chromatin. Chromosomes are divided into two distinct compartments, A and B at megabase scale \cite(*78*). A compartment (positive values) corresponds to the active regions on the chromosome and B compartment (negative values) corresponds to the inactive regions. Chromatin interactions are constrained within the compartment types, e.g the loci in A compartment interact with the loci in the same compartment. Since A/B compartment assignments are proxies for the activity level of different loci, our tissue-specific regulatory element annotations can be validated by looking at their corresponding compartment in the tissue level Hi-C data. We showed that our annotated tissue-specific active regulatory elements are dominantly located in the active compartment of the chromosomes of corresponding tissues, with significantly higher number of regulatory elements per megabase observed in the positive compartment values when layered on to first-principal component of the Hi-C data.

We have assessed where the cCREs are located with respect to the chromatin compartments. For that, we first binned the genome into 1 MB consecutive bins. We then counted the total number of cCREs in each bin and divided that number by the total number of cCCREs in the genome. This gave us the cCRE density per 1 MB. We then plotted this density against the A/B compartment score obtained by the first eigenvector of the correlation matrix calculated from the Hi-C contact matrix. We did this analysis for the master cCRE list from ENCODE3, tissue-specific active cCRE list derived in this study, more restrictive tissue-specific active cCRE list derived in this study, and tissue-specific repressed cCRE list derived in this study. Below are the scatter plots for two tissues and 4 individuals (Figure S7.4).

S7.5. Variation Analysis of cCRE Activity

S7.5.1. Visualizing the variation of cCRE Activity with JIVE

To visualize the relationship among the functional genomic data across the tissues, we used a dimension-reduction approach, namely Joint and Individual Variance Explained (JIVE) \cite(*79*). For each functional genomic experiment of histone modifications, we calculated its signals at the cCREs using UCSC genome browser bigWig tools \cite(*49*). For proteomics and RNA-seq experiments, we simply used the normalized protein abundance and RNA abundance of each gene from the experiments. For each type of assay, we have a data matrix in which the columns are the tissues from the four individuals and the rows are cCREs or genes, and each element is the signal of the functional genomic activity measured by the assay. For each assay type, we quantile-normalized the signals. For the joint analysis of the different experimental assays, we combined those matrices by column to form a meta-matrix. In each separate data matrix, some columns in each data matrix are not shared by all the assays, and thus these columns are excluded from the meta-matrix.

To reduce computation burdens, we removed the rows that have low standard deviation. From this informative meta-matrix, we applied the JIVE algorithm to project the columns into a two-dimensional (2D) space (Fig. 6C). As expected, this projection used all the information of the matrix. In addition, from the matrix of each assay, the JIVE algorithm excluded the information that can be explained by the other matrices, and then projected the matrix containing the information unique to the assay into a 2D space (Fig. 6C). For example, in the 2D space of RNA-seq, the same tissues from different individuals are very well clustered together, and the different tissues are well separated. This tendency is weaker for other assays. Taken together, this observation indicates that RNA-seq likely captures the most unique signatures of different tissues.

S7.5.2. Using regression-based approach to quantify activity variation

With a linear regression approach, we used the explained variation of the regression to measure the similarity between two experiments. A larger explained variation of the regression indicated a higher similarity between the two experiments. To elaborate on the variation, we here use a concrete example: the H3K27ac signals of cCREs from the spleens of two individuals. In this example, each of these individuals had two technical replicates of the H3K27ac signals measured by ChIP-seq. In each replicate, the signal at a cCRE was the fold change of reads between the IP experiment and the control experiment. For each cCRE, we first calculated the percentage difference of the signals between the two replicates. We focused on the cCREs with differences smaller than a certain cutoff so that the signals of these selected cCREs in one replicate can be largely explained by their counterparts in the other replicate using linear regression (i.e., $R^2 > 0.95$). To compare the two individuals, we used the common set of the selected cCREs with low technical noise. For each of the two individuals, we averaged the signals of the two replicates for the common cCREs. Therefore, we generated two sets of cCREs with H3K27ac signals having little noise respectively for the two individuals. Again, using a simple linear regression, we calculated the variance in one of the sets explained by the other. A high value indicates that the two sets of H3K27ac signals are very similar in terms of a linear relationship. As an example, the explained variation between replicates and the explained variation between experiments for different types of histone modifications in spleen is demonstrated in Figure S7.5a.

The aforementioned calculation was used for all the available histone modifications and samples (examples shown in Figure S7.5b) as well as normalized protein and RNA abundances (Figure S7.5c). For each modification, we estimated the variance explained between individuals (i.e., the same tissues of different individuals) and between tissues (i.e., the different tissues of the same individual). In addition, we estimated the variance explained between two different histone modifications (i.e., the same tissue of an individual). For mass-spectrometry, to make the protein abundances of different genes comparable across different tissues, we normalized the protein abundances of each gene across tissues so that the highest and lowest protein abundances are one and zero, respectively. The mass-spectrometry we used pooled and labeled multiple samples together to determine protein abundances in a batch, resulting in little technical noises across the samples. To be comparable, we also normalized the RNA-seq data of the samples in the same way.

In general, histone modifications have high similarity between the same tissue of two individuals; as expected, this number is smaller when comparing different tissues of the same individuals (Figure S7.5b). The similarity between different types of functional genomic activities from the same tissue is extremely low (Figure S7.5b). For example, H3K27ac between individuals was very similar in spleen and in transverse colon. However, the H2K27ac similarity between the two tissues was substantially reduced (Figure S7.5b). In line with this disparity across tissues, the similarity between normalized gene expression and protein abundance also varied substantially across tissues. The lower similarity in prostate is consistent with previous observations \cite(*80*). The similarities between all the available histone modifications are reported in the File: Similarity_of_functional_genomic_activities_of_cCREs.xlsx. The normalized proteomics and RNA-seq data of genes are in the File: normalized_proteomics_RNA-seq.dat.

S8. Supp. content to main text section “Measuring tissue specificity”

In this section, we compared the tissue-specificity of genes, cCREs and epigenomic peaks in a systematic manner. We included protein-coding genes, non-coding genes, different catalogs of decorated cCREs, and diverse types of epigenetic marks.

S8.1. Tissue Specific Results

There are many methods for determining the tissue-specificity, most of which are based on continuous positive values \cite(*81*). Here we chose the simple method of tissue count to determine the tissue-specificity of genes/cCREs based on the thresholds \cite(*81*). We did this because we can consistently apply this method across different annotations including cCREs, genes and epigenomic peaks. Most of the other methods that are based on continuous positive values can be only applied on one annotation category (e.g., genes). Briefly, all the genes and cCREs were defined as active and inactive by thresholding their expression/activity level in a particular tissue type. The numbers of tissue types that these genes/cCREs active in were then summarized. For each gene/cCRE group, we then calculated the tissue-specificity score using the number genes/cCREs that are active in only one tissue type divided by the total number of genes/cCREs. The tissue-specificity scores range between 0 and 1, with higher scores indicating stronger tissue specificity.

For the genes, we included three gene types: protein-coding genes (from mass spectrometry and RNA sequencing technology), long noncoding RNAs and pseudogenes. To better estimate the expression level of pseudogenes, we applied our previously developed pipeline to quantify the expression level of pseudogenes, which can minimize the effects of multiple mapping bias in RNA-seq data \cite(*82*) (Figure S8.1c). We then applied this pipeline to all the three gene types, and defined a set of active genes in the tissues by thresholding the FPKM values (FPKM>1 for protein-coding genes; FPKM>0.5 for long noncoding RNAs and pseudogene) (Figure S8.1a). Over 40% and 35% of the detected pseudogenes and lncRNAs, respectively, were actively transcribed in a single tissue, confirming that non-coding RNAs exhibit higher tissue specificity than protein-coding genes \cite(*83, 84*). Of the pseudogenes demonstrating tissue specificity, a large fraction showed transcriptional activity only in testis (Figure S8.1b). For the cCREs, we used the decorated annotations in the tissues to calculate the tissue-specificity scores as described above (Figure S8.1d and Figure S8.1e). We also explored the tissue specificity of regulatory elements and epigenomic peaks (Fig. 7A). The epigenetic profiles analyzed, including H3K27ac or DNase, demonstrated tissue specificity, with the exception of DNA methylation, which exhibited strong ubiquity. An example of tissue-specificity of RAMPAGE data are shown in Figure S8.1f. The tissue specificity of the genes, cCREs and epigenomic peaks are in the File: Tissue_Specificity.zip.

S8.2. Tissue Specificity of Allele-Specific Binding and Expression

Similar to H3K27ac-ASB cCREs (Fig. 6), most ASE genes were detected in a single tissue (Figure S8.2a). For the ~20 genes that were detected ASE across all tissues, the allelic imbalance is in the same direction (Figure S8.2b). We further compare our pan-tissue H3K27ac-ASB and ASE genes with the housekeeping genes. Annotation results are shown in Figure S8.2c-d.

S8.3. The Effect of Tissue Specificity on Conservation

Tissue specificity influence on conservation is shown in Figure S8.3a. Candidates are separated in categories of active, bivalent, and repressive. Number of candidates, rare DAF, and the corresponding total SNP count (from gnomAD) are given as a function of increasing tissue specificity (shared tissue count). In order to select rare variants, a MAF of 0.05 was used.

Various decorations further subset categories and affect the conservation level. Specifically whether elements are distal or proximal as well as if they are CTCF bound or not. Conservation is shown for both phastCons (cross-species) and rare DAF (cross-population) in Figure S8.3b.

We show the conservation across active and repressed cCREs in both ubiquitous and tissue specific cases in Figure S8.3c. We include the results across 1KG, PCAWG, and gnomAD. Additionally, we also show an increase in conservation when filtering for high H3K27ac signals (using stringent definitions for active elements with Matched Filter \cite(*68*)). See above in the supplement text), which is supported by all three data sets.

S9. Supp. content to main text section "Relating encyclopedia decorations to QTLs & GWAS loci"

We utilized various methods to evaluate the regulatory impact of our cCRE decorations. QTL and GWAS SNPs are important functional genomic variants and are useful for interpreting the function of our decorations. We performed GWAS enrichment analysis using eQTL and GWAS SNPs to assess the disease-relevance of our cCRE decorations.

S9.1. QTL Enrichment Analysis

We estimated the QTL (eQTL and sQTL) enrichment in the cCREs by calculating an odds ratio (OR) score using the numbers of real QTL SNPs and the control SNPs located in the cCREs comparing to those in the baseline regions (Figure S9.1a).

$$\mathrm{OR} = \frac{a / b}{c / d}$$

in which a is the number of QTL SNPs in the cCREs; b is the number of control SNPs in the cCREs; c is the number of QTL SNPs in the baseline region; d is the number of control SNPs in the baseline region.

The eQTL and sQTL SNPs were downloaded from GTEx v8 \cite(*85*). The baseline regions are the union of all the functional and putative functional regions in the human genome, including CDS, UTRs, noncoding RNA genes, open chromatin regions, TF binding sites, active and repressed histone peaks from multiple tissue and cell types as well as evolutionary conserved regions \cite(*86*). The set of control SNPs were generated with the same number and same minor allele frequency distribution as the real QTLs, and this procedure is repeated 30 times to calculate standard deviation for the SNP enrichment. The results of the QTL enrichment are in the File: QTL_enrichment.zip.

We also compared the eQTL/sQTL enrichment in the regulatory elements from EN-TEx and those from Roadmap (Figure S9.1b and Figure S9.1c). First, we found that the distal regulatory elements from EN-TEx show stronger enrichment than the enhancer annotation from Roadmap. In addition, the active proximal regulatory elements from EN-TEx show stronger eQTL/sQTL enrichment than the TSS-associated annotations from Roadmap.

S9.2. GWAS Enrichment Analysis

We downloaded the GWAS tag SNPs from the GWAS Catalog \cite(*87*). We did several steps of quality control to generate a set of high-quality GWAS tag SNPs by removing some insignificant SNPs (p-values>5*10-8), low-confidence SNPs, and SNPs from non-European studies. We also removed all SNPs in the HLA locus (for hg38: chr6:29,723,339-33,087,199). Next, we extended the set of tag SNPs by including the SNPs in high linkage disequilibrium (LD scores>0.6) with the tag SNPs, which can generate more SNPs to increase the statistical power in the enrichment analysis. Some GWAS with very few LD-extend SNPs were removed. Finally, we have a clean dataset with ~70K unique tag SNPs from 1140 GWAS covering 717 unique traits.

We then applied the hypergeometric test to estimate the enrichment of the GWAS tag SNPs in the cCREs from a particular tissue type (Figure S9.2a).

$$P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

in which N is the total number of cCREs in the genome; K is the total number of cCREs that carry GWAS tag SNPs; n is the number of cCREs in a particular tissue type; and k is the number of cCREs in a particular tissue type that also carry GWAS tag SNPs. Notably, we extended the cCREs 500bp on both sides in the calculation (Figure S9.2c). The results of the GWAS tag SNP enrichment are in the File: GWAS_enrichment.zip.

For the active distal cCREs, we identified 141 GWAS that are enriched in at least one tissue type (Figure S9.2d). However, for the active proximal cCREs, we did not find any enriched GWAS in any tissue type. These results are consistent with previous studies that the causal GWAS SNPs are enriched in the enhancers instead of the near-gene promoters \cite(*88, 89*); and also suggest that the active distal cCREs from our decoration are indeed significantly enriched in enhancers as we observed in original Roadmap annotations (Figure S9.2e).

Stratified LDSC were also calculated for each tissue using 1000G LD Scores and GWAS summary statistics provided by Bulik-Sullivan, et al. It regresses chi-square statistics from the GWAS summary statistics with LD scores to estimate partitioned heritability in a disease-specific manner. The p-value indicates enrichment for a particular trait within an annotation.

In section Figure S9.2a, we show the p-value enrichment of each tissue with respect to various GWAS traits. Notably, distal active allele-specific regions experienced higher enrichment compared to distal active non-allele specific regions (Figure S9.2f), and both of types of regions all experienced higher enrichment compared to original Roadmap annotations (Figure S9.2b). For LDSC enrichment analysis of distal active elements in Coronary Artery (Figure S9.2b), we found stronger associations between allele-specific elements with respect to Celiac's disease, Neuroticism, and Type II Diabetes, which were elucidated in previous clinical studies \cite(*90-92*). These results demonstrate that allelic elements can significantly improve GWAS trait enrichment compared to the total set of elements across different traits as well as diverse tissue types, indicating that allelic elements are valuable for the interpretation of GWAS data and that they potentially help pinpoint small subsets of regulatory elements driving the trait in specific tissues.

S9.3. Providing evidence for the buffering hypothesis using AS cCREs and housekeeping genes

Genetic variants in cCREs can change functional signal and gene expression. For these changes to actually occur, the variants need to escape from buffering effects \cite(*38*). Such effects are strong in important genomic regions.We used AS as a proxy for escaping buffering. Based on our allelic decoration, we evaluated the allelic specificity of housekeeping genes expressed in EN-TEx tissues, shown in Figure S9.3a. For each tissue, expressed protein-coding genes were split into housekeeping genes and non-housekeeping genes according to Housekeeping and Reference Transcript Atlas (http://www.housekeeping.unicamp.br) \cite(*93*). Two-sided fisher exact test was performed to measure the enrichment of AS housekeeping genes. We found that, compared with non-housekeeping genes, the expression of housekeeping genes shows less allelic specificity, supporting the buffering hypothesis. We further examined the allelic specificity of proximal active (pAct) cCREs in a ± 10kb window centered on the transcription starting site (TSS, defined by gene starting site) of each housekeeping and non-housekeeping gene. The cCREs flanking housekeeping genes are significantly (Figure S9.3a, paired-tissue two-sided t-test, p-value < 2.2e-16) longer than cCREs flanking non-housekeeping genes. To control this factor, we split genes into 20 bins based on the total length of flanking cCREs. Within each bin, cCRE length remains similar (paired-tissue two-sided t-test, p-value > 0.05) between housekeeping and non-housekeeping genes. The bins having less than 30 housekeeping or non-housekeeping genes were removed from further analysis. The pAct cCREs flanking housekeeping genes are less likely AS than the ones flanking non-housekeeping genes (Two-sided t-test).

The buffering effect is likely due to redundant TFs. To test this, we counted the number of TF motifs that intersect with each CTCF+ and CTCF- cCRE in each tissue. For this calculation, we used the motifs of 206 TFs (CTCF excluded) from Cis-BP \cite(*94*). The total count of all TF motifs was compared between CTCF+ cCREs and CTCF- cCREs using two-sided t-test. As shown in Figure S9.3b, for both distal and proximal cCREs, CTCF+ cCREs have significantly (p-value < 0.05) more TF motifs than CTCF- cCREs.

S10. Additional information about the EN-TEx resource

S10.1. The EN-TEx Supplemental Data Repository

All processed data files are detailedly described in their corresponding sections in this document. When mentioned, these files are referred to as "File: file_name". All these files are hosted in the EN-TEx data portal website ENTEx.encodeproject.org, with the exact same file names as in "File: file_name". Additionally, on the website, each file is followed with the supplement text section number that contains the description of that file. Links to all the raw data of this project could be found in the EN-TEx data portal website as well.

S10.2. Open-consented of data

In concert with the GTEx project an IRB-approved consent for unrestricted access to data collected as part of the GTEx and EN-TEx project was written and given to the next-of-kin of each of the donors. The consent form allows for unrestricted use of the primary data and metadata collected from each donor. It was made clear that no identification of the donor or family constituted part of these data, it is within the realm of possibly that individual identification could be made. Specific details of the consent document is contained in here https://www.genome.gov/Pages/Research/ENCODE/GTEx_Consent_ENCODE_addendum_10-9-14.pdf

S10.3. The EN-TEx Chromosome-Level Data Visualization Tool

Because the EN-TEx data spams over a wide range of the human genome, it may be useful to visualize their distribution over each chromosome. Accordingly, we here present the EN-TEx Chromosome-Level Data Visualization Tool, which generates heat maps for data sets for all assays, individuals, and tissues present in the EN-TEx data catalog. The data, which are initially in BED format, are preprocessed with in-house Bash and Python scripts and converted to GRCh38 coordinates using liftOver \cite(*95*) prior to the generation of the plots using the R package chromoMap \cite(*96*). The EN-TEx Chromosome-Level Data Visualization Tool was also used to generate the plots present in Figure 5A of the main text.

The EN-TEx Chromosome-Level Data Visualization Tool can be accessed at ENTEx.gersteinlab.org. For each track, users are able to determine the data displayed by changing the individual, assay, ploidy, and color parameters, for up to 4 tracks per plot (Figure S10.3a). If no additional settings are selected in the “Advanced” tab, press submit, and the tool will generate heat maps for the data of each chromosome, at a fixed resolution of 2.5Mb. The plots produced are interactive: by hovering the mouse cursor over each of the bins produced, the user is able to get information about the data displayed in that specific bin. In the advanced tab, users are able to generate plots with custom chromosome and region selections. To view the data in only one chromosome, open the Advanced tab and select the chromosome of interest in the first dropdown menu. To visualize only a subset region of the chromosome, in the “Region” input text box, input the region in the format initial_position:final_position (e.g., if the user wishes to visualize data in between 1Mb and 2Mb, the user would input 1000000:2000000). Please note that the resolution of the data for subset regions of the chromosome will always be equal to the length of the inputted interval divided by a factor of one hundred (e.g. for the 1000000:2000000 interval, the resolution will be equal to 10kb). Moreover, users also have the option to visualize the data as heatmaps accompanied by either histograms or scatterplots. To do so, select the desired type of additional data representation in the "Plot type" dropdown at the end of the advanced section. A series of plots generated with this tool are shown in Figure S10.3b.

S10.4. Explorer Tool

The EN-TEx explorer tool, which can be installed as an offline executable or hosted on a website, allows for the interactive exploration of low-dimensional visualizations created by a data analysis pipeline (Figure S10.4). This pipeline performs dimensionality reduction on cCRE signals, genomic data, and proteomic data. Methods include principal component analysis (PCA), variational autoencoders (VAE), uniform manifold approximation and projection (UMAP) \cite(*97*), potential of heat diffusion for affinity-based transition embedding (PHATE)\cite(*98*), set intersection plots generated by user-specified thresholds (Sets), and t-Distributed Stochastic Neighbor Embedding (tSNE). The visualizations generally cluster samples from common tissues together. Through extensive precomputation, the tool allows users to interactively adjust analysis parameters, including scaling, normalization, feature subsetting, method-specific hyperparameters, the type of visualization used (ggplot2, plotly 2D, plotly 3D, boxplot, heatmap, UpsetR, Venn Diagram), and the appearance of the resulting figures. Users are able to save figures as images, download analysis results as excel spreadsheets(Figure S10.4), or bookmark their sessions as URLs that can be easily shared.

The results can be visualized as interactive scatter plots / heatmaps / boxplots, which can be indexed via session-specific bookmarks, or downloaded for further investigation. Dimensionality reduction techniques include principal component analysis (PCA), variational autoencoder (VAE), uniform manifold approximation and project (UMAP), t-distributed stochastic neighbor embedding (tSNE), and potential of heat-diffusion for affinity-based trajectory embedding (PHATE)

References

1. G. X. Zheng *et al.*, Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311 (2016).
2. S. Ardui, A. Ameur, J. R. Vermeesch, M. S. Hestand, Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* **46**, 2159-2168 (2018).
3. M. Nattestad *et al.*, Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* **28**, 1126-1135 (2018).
4. M. Cretu Stancu *et al.*, Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* **8**, 1326 (2017).
5. P. Edge, V. Bafna, V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-812 (2017).
6. F. J. Sedlazeck *et al.*, Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).
7. D. C. Jeffares *et al.*, Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**, 14061 (2017).
8. J. Rozowsky *et al.*, AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
9. J. Jou *et al.*, The ENCODE Portal as an Epigenomics Resource. *Curr Protoc Bioinformatics* **68**, e89 (2019).
10. R. Vaser, I. Sovic, N. Nagarajan, M. Sikic, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746 (2017).
11. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
12. P. A. Audano *et al.*, Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675 e619 (2019).
13. P. Ebert *et al.*, Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, (2021).
14. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
15. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
16. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
17. N. C. Durand *et al.*, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98 (2016).
18. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
19. A. Kaul, S. Bhattacharyya, F. Ay, Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc* **15**, 991-1012 (2020).
20. F. Ay, T. L. Bailey, W. S. Noble, Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* **24**, 999-1011 (2014).
21. H. Shin *et al.*, TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**, e70 (2016).
22. P. A. Knight, D. Ruiz, A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**, 1029-1047 (2012).
23. C. J. Cameron, J. Dostie, M. Blanchette, HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biol* **21**, 11 (2020).

24. A. Frankish *et al.*, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
25. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, (2020).
26. J. C. Wright, J. S. Choudhary, DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *J Proteomics Bioinform* **9**, 176-180 (2016).
27. M. Spivak, J. Weston, L. Bottou, L. Kall, W. S. Noble, Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J Proteome Res* **8**, 3737-3745 (2009).
28. H. Weisser, J. C. Wright, J. M. Mudge, P. Gutenbrunner, J. S. Choudhary, Flexible Data Analysis Pipeline for High-Confidence Proteogenomics. *J Proteome Res* **15**, 4686-4695 (2016).
29. J. M. Mudge *et al.*, Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res* **29**, 2073-2087 (2019).
30. J. C. Wright *et al.*, Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat Commun* **7**, 11778 (2016).
31. Y. Perez-Riverol *et al.*, The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**, D442-D450 (2019).
32. S. C. Munger *et al.*, RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics* **198**, 59-73 (2014).
33. S. Huang, J. Holt, C. Y. Kao, L. McMillan, W. Wang, A novel multi-alignment pipeline for high-throughput sequencing data. *Database (Oxford)* **2014**, (2014).
34. M. Pirinen *et al.*, Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31**, 2497-2504 (2015).
35. Y. Baran *et al.*, The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25**, 927-936 (2015).
36. M. T. Maurano *et al.*, Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* **47**, 1393-1401 (2015).
37. J. Chen *et al.*, A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7**, 11101 (2016).
38. V. Onuchic *et al.*, Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* **361**, (2018).
39. S. E. Castel *et al.*, A vast resource of allelic expression data spanning human tissues. *Genome Biol* **21**, 234 (2020).
40. M. B. Gerstein *et al.*, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
41. E. Khurana *et al.*, Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
42. J. F. Degner *et al.*, Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212 (2009).
43. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-1063 (2015).
44. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
45. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).
46. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
47. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

48. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144-161 (2013).

49. W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, D. Karolchik, BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207 (2010).

50. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).

51. L. Jiang *et al.*, A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283 e219 (2020).

52. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

53. P. Pawliczek *et al.*, ClinGen Allele Registry links information about genetic variants. *Hum Mutat* **39**, 1690-1701 (2018).

54. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).

55. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, in *NAACL-HLT*. (2019).

56. Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, (2021).

57. P. Ng, dna2vec: Consistent vector representations of variable-length k-mers. *ArXiv* **abs/1701.06279**, (2017).

58. Y. Itoh *et al.*, The X-linked histone demethylase Kdm6a in CD4+ T lymphocytes modulates autoimmunity. *J Clin Invest* **129**, 3852-3863 (2019).

59. G. T. Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

60. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

61. C. Chiang *et al.*, The impact of structural variation on human gene expression. *Nat Genet* **49**, 692-699 (2017).

62. M. Karimzadeh, C. Ernst, A. Kundaje, M. M. Hoffman, Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46**, e120 (2018).

63. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

64. T. Valikangas, T. Suomi, L. L. Elo, A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* **19**, 1-11 (2018).

65. L. Chen *et al.*, GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **6**, e4600 (2018).

66. B. A. Berghoff, T. Karlsson, T. Kallman, E. G. H. Wagner, M. G. Grabherr, RNA-sequence data normalization through in silico prediction of reference genes: the bacterial response to DNA damage as case study. *BioData Min* **10**, 30 (2017).

67. E. P. Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).

68. A. Sethi *et al.*, Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat Methods* **17**, 807-814 (2020).

69. J. S. Becker, D. Nicetto, K. S. Zaret, H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. *Trends Genet* **32**, 29-41 (2016).

70. G. Gerlitz, The Emerging Roles of Heterochromatin in Cell Migration. *Front Cell Dev Biol* **8**, 394 (2020).

71. N. Saksouk, E. Simboeck, J. Dejardin, Constitutive heterochromatin formation and transcription in mammals. *Epigenetics Chromatin* **8**, 3 (2015).

72. M. Ninova, K. Fejes Toth, A. A. Aravin, The control of gene expression and cell identity by H3K9 trimethylation. *Development* **146**, (2019).

73. D. Nicetto, K. S. Zaret, Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. *Curr Opin Genet Dev* **55**, 1-10 (2019).

74. J. S. Becker *et al.*, Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol Cell* **68**, 1023-1037 e1015 (2017).

75. L. Pace *et al.*, The epigenetic control of stemness in CD8(+) T cell fate commitment. *Science* **359**, 177-186 (2018).

76. J. Du, L. M. Johnson, S. E. Jacobsen, D. J. Patel, DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol* **16**, 519-532 (2015).

77. N. Saksouk *et al.*, Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation. *Mol Cell* **56**, 580-594 (2014).

78. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

79. K. H. Hellton, M. Thoresen, Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics* **17**, 537-548 (2016).

80. I. Kosti, N. Jain, D. Aran, A. J. Butte, M. Sirota, Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci Rep* **6**, 24799 (2016).

81. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**, 205-214 (2017).

82. C. Sisu *et al.*, Transcriptional activity and strain-specific history of mouse pseudogenes. *Nat Commun* **11**, 3695 (2020).

83. J. D. Ransohoff, Y. Wei, P. A. Khavari, The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol* **19**, 143-157 (2018).

84. A. Necsulea *et al.*, The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640 (2014).

85. G. T. Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

86. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235 (2015).

87. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).

88. L. Yao, Y. G. Tak, B. P. Berman, P. J. Farnham, Functional annotation of colon cancer risk SNPs. *Nat Commun* **5**, 5114 (2014).

89. S. Whalen, K. S. Pollard, Most chromatin interactions are not in linkage disequilibrium. *Genome Res* **29**, 334-343 (2019).

90. R. D. Gajulapalli, D. J. Pattanshetty, Risk of coronary artery disease in celiac disease population. *Saudi J Gastroenterol* **23**, 253-258 (2017).

91. A. Almas, J. Moller, R. Iqbal, Y. Forsell, Effect of neuroticism on risk of cardiovascular disease in depressed persons - a Swedish population-based cohort study. *BMC Cardiovasc Disord* **17**, 185 (2017).

92. R. Naito, T. Kasai, Coronary artery disease in type 2 diabetes mellitus: Recent treatment strategies and future perspectives. *World J Cardiol* **7**, 119-124 (2015).

93. B. W. Hounkpe, F. Chenou, F. de Lima, E. V. De Paula, HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* **49**, D947-D955 (2021).

94. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).

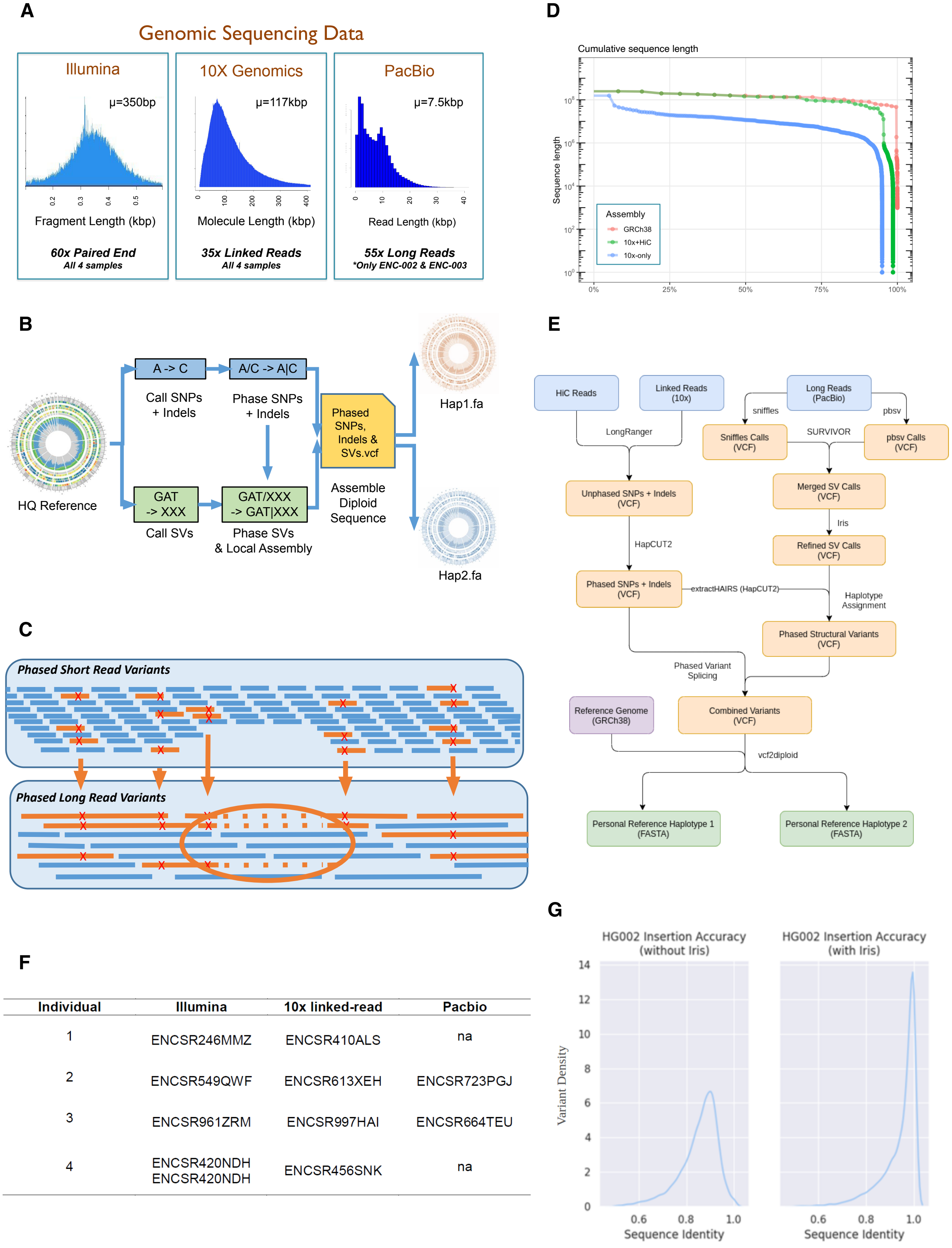
95. A. S. Hinrichs *et al.*, The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-598 (2006).

96. L. Anand, C. M. R. LÛpez, chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes. *bioRxiv*, (2019).

97. L. McInnes, J. Healy, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* **abs/1802.03426**, (2018).
98. K. R. Moon *et al.*, Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492 (2019).
99. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
100. Y. S. Son *et al.*, A SMN2 Splicing Modifier Rescues the Disease Phenotypes in an In Vitro Human Spinal Muscular Atrophy Model. *Stem Cells Dev* **28**, 438-453 (2019).
101. W. Muller-Felber *et al.*, Infants Diagnosed with Spinal Muscular Atrophy and 4 SMN2 Copies through Newborn Screening - Opportunity or Burden? *J Neuromuscul Dis* **7**, 109-117 (2020).
102. N. Mujahid *et al.*, A UV-Independent Topical Small-Molecule Approach for Melanin Production in Human Skin. *Cell Rep* **19**, 2177-2184 (2017).
103. J. Hansen *et al.*, De novo mutations in SIK1 cause a spectrum of developmental epilepsies. *Am J Hum Genet* **96**, 682-690 (2015).
104. C. Proschel *et al.*, Epilepsy-causing sequence variations in SIK1 disrupt synaptic activity response gene expression and affect neuronal morphology. *Eur J Hum Genet* **25**, 216-221 (2017).
105. F. Reese, A. Mortazavi, Swan: a library for the analysis and visualization of long-read transcriptomes. *Bioinformatics*, (2020).
106. D. Garrido-Martin, E. Palumbo, R. Guigo, A. Breschi, ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS Comput Biol* **14**, e1006360 (2018).
107. R. Beraldi *et al.*, Genetic ataxia telangiectasia porcine model phenocopies the multisystemic features of the human disease. *Biochim Biophys Acta Mol Basis Dis* **1863**, 2862-2870 (2017).
108. M. E. Levine *et al.*, An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* **10**, 573-591 (2018).
109. S. Horvath, Erratum to: DNA methylation age of human tissues and cell types. *Genome Biol* **16**, 96 (2015).
110. D. Szklarczyk *et al.*, STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).
111. W. Sungnak *et al.*, SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat Med* **26**, 681-687 (2020).

Supplemental figures

Figure S1. Supp. figures to main text section “Personal genomes & matched data matrix”



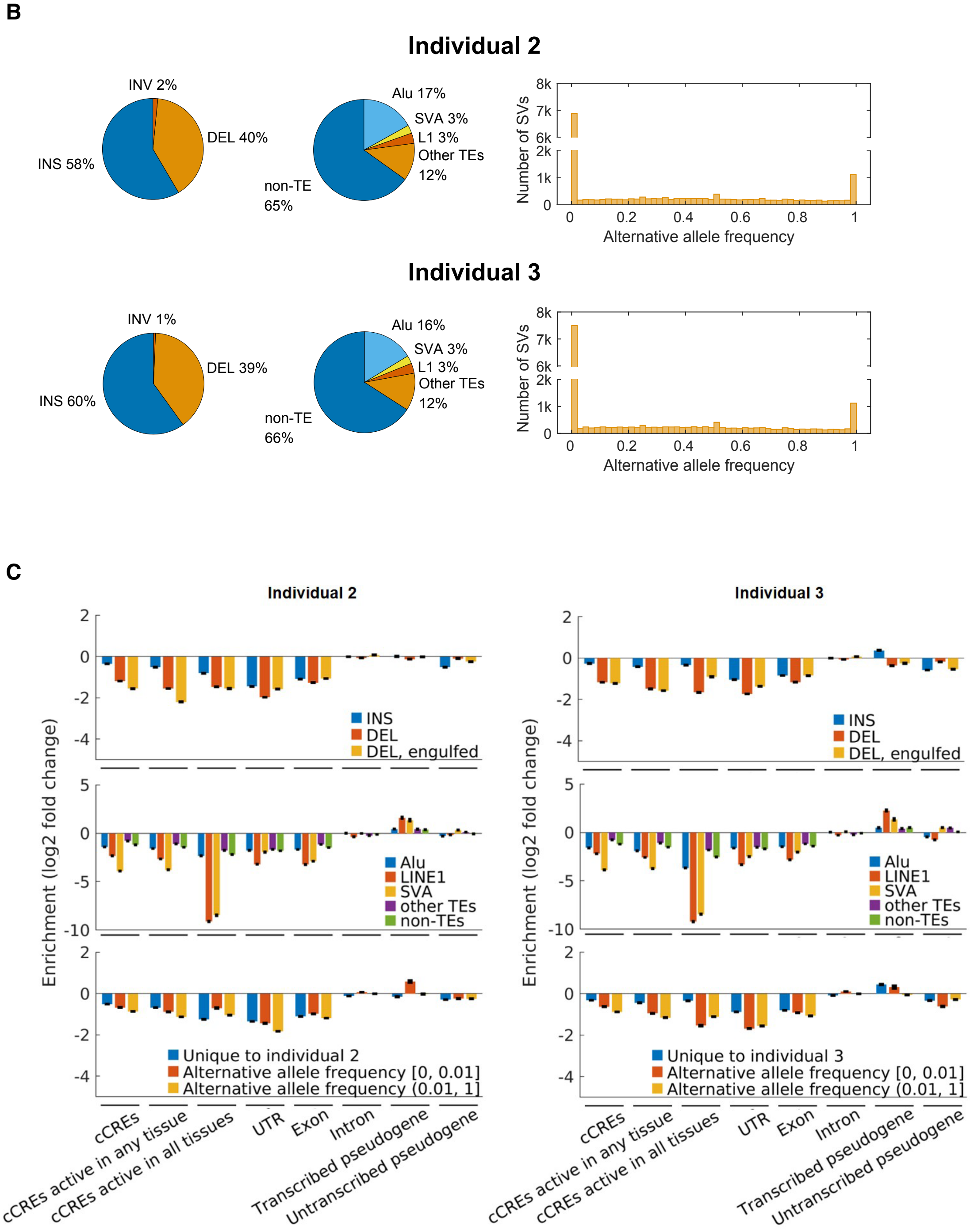
Individual	Illumina	10x linked-read	Pacbio
1	ENCSR246MMZ	ENCSR410ALS	na
2	ENCSR549QWF	ENCSR613XEH	ENCSR723PGJ
3	ENCSR961ZRM	ENCSR997HAI	ENCSR664TEU
4	ENCSR420NDH ENCSR420NDH	ENCSR456SNK	na

Figure S1.1. Personal genome construction

(A) Summary of whole genome sequencing. All four individuals were sequenced with regular Illumina short-reads and 10x linked-reads. Individual 2 and 3 were additionally sequenced with PacBio long-reads. The figure shows the sequencing depth and the distribution of read length under each platform. See Figure S1.1f for accession numbers of the relevant data. **(B)** An overview of the CrossStitch workflow. SNPs and small indels are called and phased, while unphased SV calls are obtained independently. Then, the phase blocks from the small variants are used to assign haplotypes to heterozygous SVs, and the phased variants are used to construct a phased personal genome assembly based on a high quality reference sequence. (Note that in this figure "mat" and "pat" are just schematic representations of hap1 and hap2 and do not reflect actual parent of origin assignments.) **(C)** SV Phasing with CrossStitch Phased small variants are used to assign a haplotype to each long read, and SVs are phased by observing the haplotypes of the long reads which indicate the presence of that variant. In this example, a deletion is phased by observing that all three of the long reads including that deletion have small variants which are unique to the orange haplotype. **(D)** Phase block length This figure shows the size of phase blocks in individual 2 obtained with HapCUT2 when performing small variant phasing with 10x reads only, as well as with a combination of 10x and Hi-C reads. When both data types are used, the contiguity of the phase blocks obtained is very similar to that of GRCh38. **(E)** Detailed CrossStitch Methods Summar. This diagram shows an overview of the CrossStitch methods with the specific software and data types used. **(F)** Accession numbers of whole-genome sequence data. All data could be downloaded from the ENCODE portal. **(G)** Refining Novel Insertion Sequences with Iris. This figure shows sequence similarity of ONT calls to CCS calls in the Genome-in-a-Bottle sample HG002, used to benchmark the performance of Iris. The sequence similarity between two sequences S and T is calculated as edit_distance(S, T) / [max(length(S), length(T))].

A

Individual	SNVs	indels	SVs
1	3,900,246	536,621	n.a.
2	3,878,924	545,419	17,649
3	4,023,587	577,594	18,542
4	3,952,264	556,055	n.a.



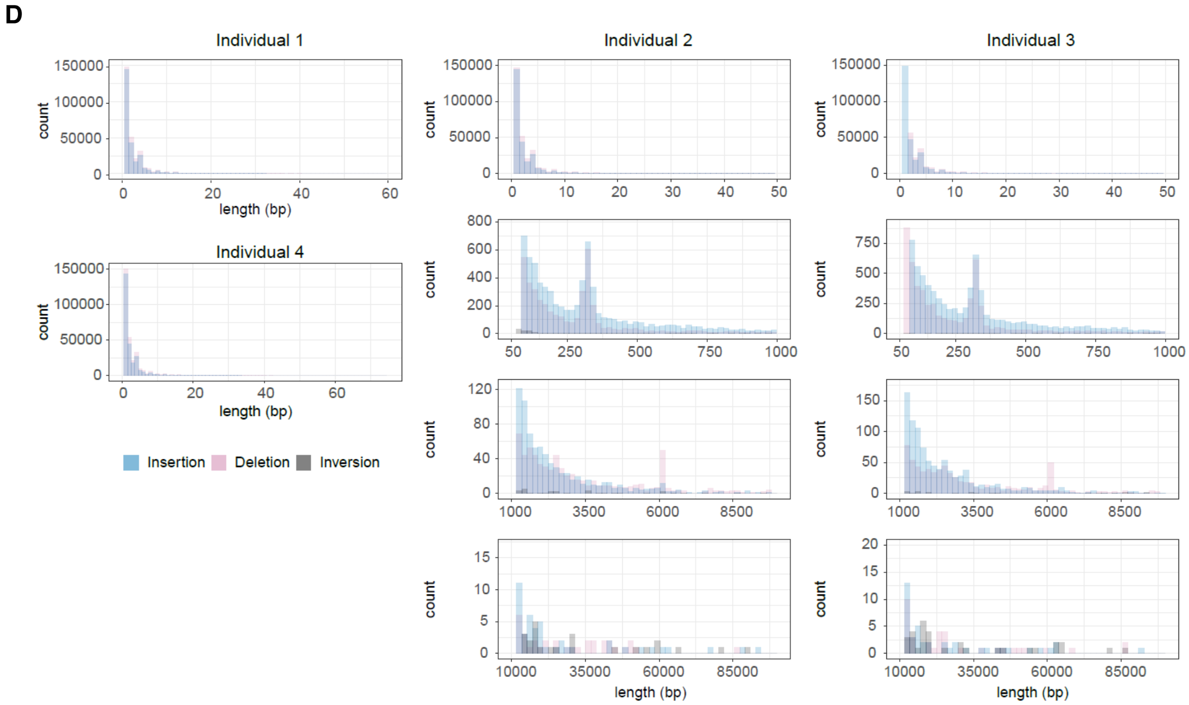
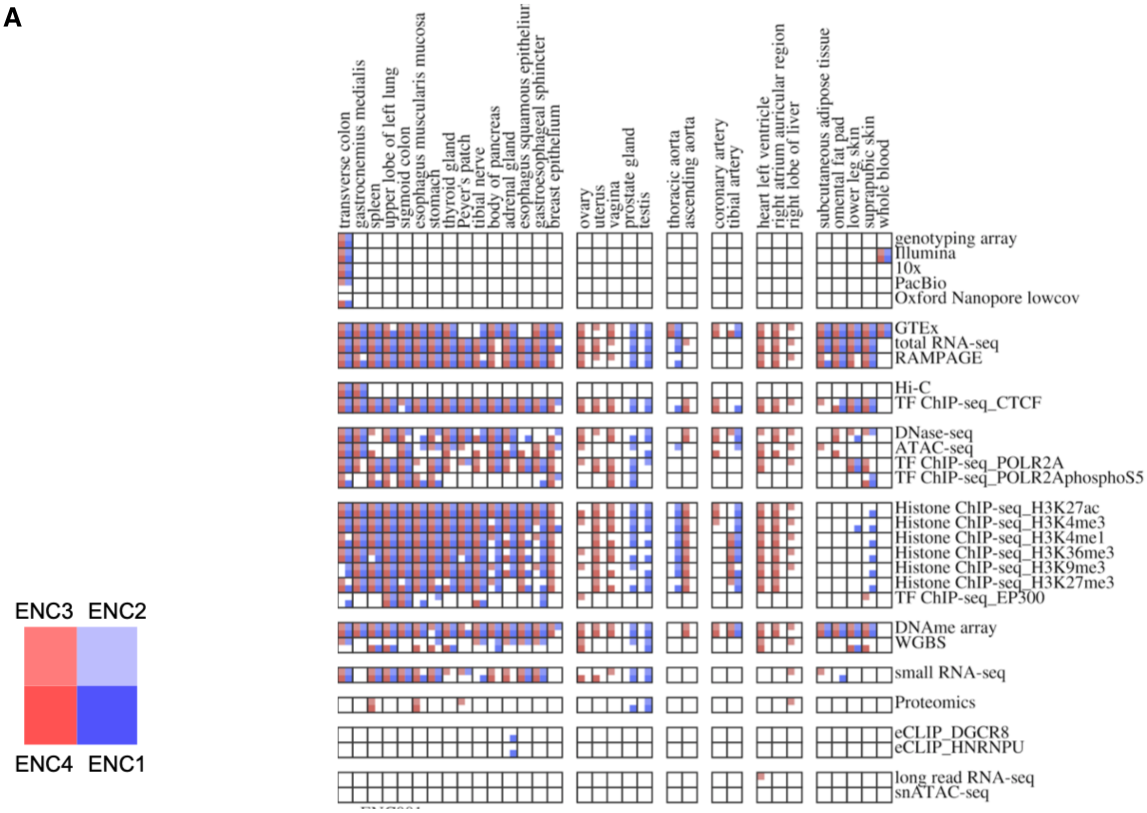


Figure S1.2. Analysis of SV

(A) Number of genomic variants in the four individuals. **(B)** Summary of SVs in individual 2 and 3. Left panels: the fractions of INS, DEL, and INV. Middle panels: the fractions of SVs involving transposable elements. Right panels: Allele frequencies of SVs in European population calculated by overlapping with Audano et al. (2019) \cite(*12*). SVs that have no overlap in Audano et al. are placed in the first bin. **(C)** Overlaps between SVs and functional genomic regions. We shuffle the locations of SVs (see Supplementary texts for details) to determine whether SVs are enriched or depleted in a given type of genomic regions. For DELs, we consider cases where a DEL partially overlaps with a given genomic region (DEL, partial) and cases where a DEL is engulfed by a given genomic region (DEL, engulfed). **(D)** Lengths of genomic variants in the four individuals. Each panel corresponds to an individual, showing the length distributions of SNVs, indels, and SVs (if available in the given individual).

A



B

ENTEx Tissue Names	Abbrev	Color Hex	GTEx Tissue Names
transverse colon	CLNTRN	#CC9955	Colon_Transverse
sigmoid colon	CLNSGM	#EEBB77	Colon_Sigmoid
upper lobe of left lung	LUNG	#99FF00	Lung
stomach	STMACH	#FFDD99	Stomach
spleen	SPLEEN	#778855	Spleen
gastrocnemius medialis	GASMED	#AAAAFF	Muscle_Skeletal
adrenal gland	ADRNLG	#33DD33	Adrenal_Gland
esophagus muscularis mucosa	ESPMSM	#BB9988	Esophagus_Muscularis
thyroid gland	THYROID	#006600	Thyroid
gastroesophageal sphincter	ESPGES	#8B7355	Esophagus_Gastroesophageal_Junction
tibial nerve	NERVET	#FFD700	Nerve_Tibial
body of pancreas	PNCREAS	#995522	Pancreas
esophagus squamous epithelium	ESPSQE	#552200	Esophagus_Mucosa
Peyer's patch	PEYERP	#555522	
breast epithelium	BREAST	#33CCCC	Breast_Mammary_Tissue
suprapubic skin	SKINNS	#0000FF	Skin_Not_Sun_Exposed_Suprapubic
prostate gland	PRSTTE	#DDDDDD	Prostate
heart left ventricle	HRTLV	#660099	Heart_Left_Ventricle
testis	TESTIS	#AAAAAA	Testis
vagina	VAGINA	#FF5599	Vagina
lower leg skin	SKINS	#7777FF	Skin_Sun_Exposed_Lower_leg
tibial artery	ARTTBL	#FF0000	Artery_Tibial
uterus	UTERUS	#FF66FF	Uterus
right atrium auricular region	HRTAA	#9900FF	Heart_Atrial_Appendage
ovary	OVARY	#FFAAFF	Ovary
omental fat pad	ADPVSC	#FFAA00	Adipose_Visceral_Omentum
subcutaneous adipose tissue	ADPSBQ	#FF6600	Adipose_Subcutaneous
ascending aorta	AORTASC	#FF5555	Artery_Aorta
right lobe of liver	LIVER	#AABB66	Liver
thoracic aorta	AORTTHO	#FF5555	Artery_Aorta
coronary artery	ARTCRN	#FFAA99	Artery_Coronary

Figure S1.3a. Functional genomics data

(A) Data matrix of the EN-TEx resource showing tissues vs. assays. Each square is partitioned into four, representing the four individuals. **(B)** Information of EN-TEx tissues

The table shows the full name, abbreviation and color code of EN-TEx tissues, as well as their matching relationship with GTEx tissues. This tissue color scheme is also used in other main and supplementary figures.

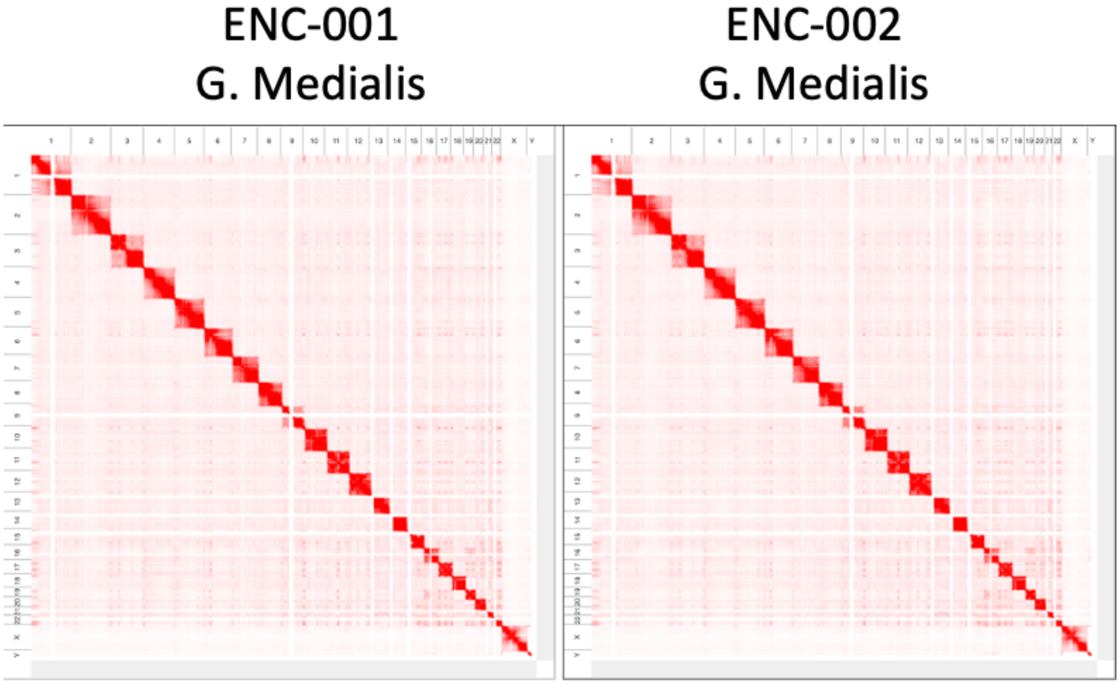
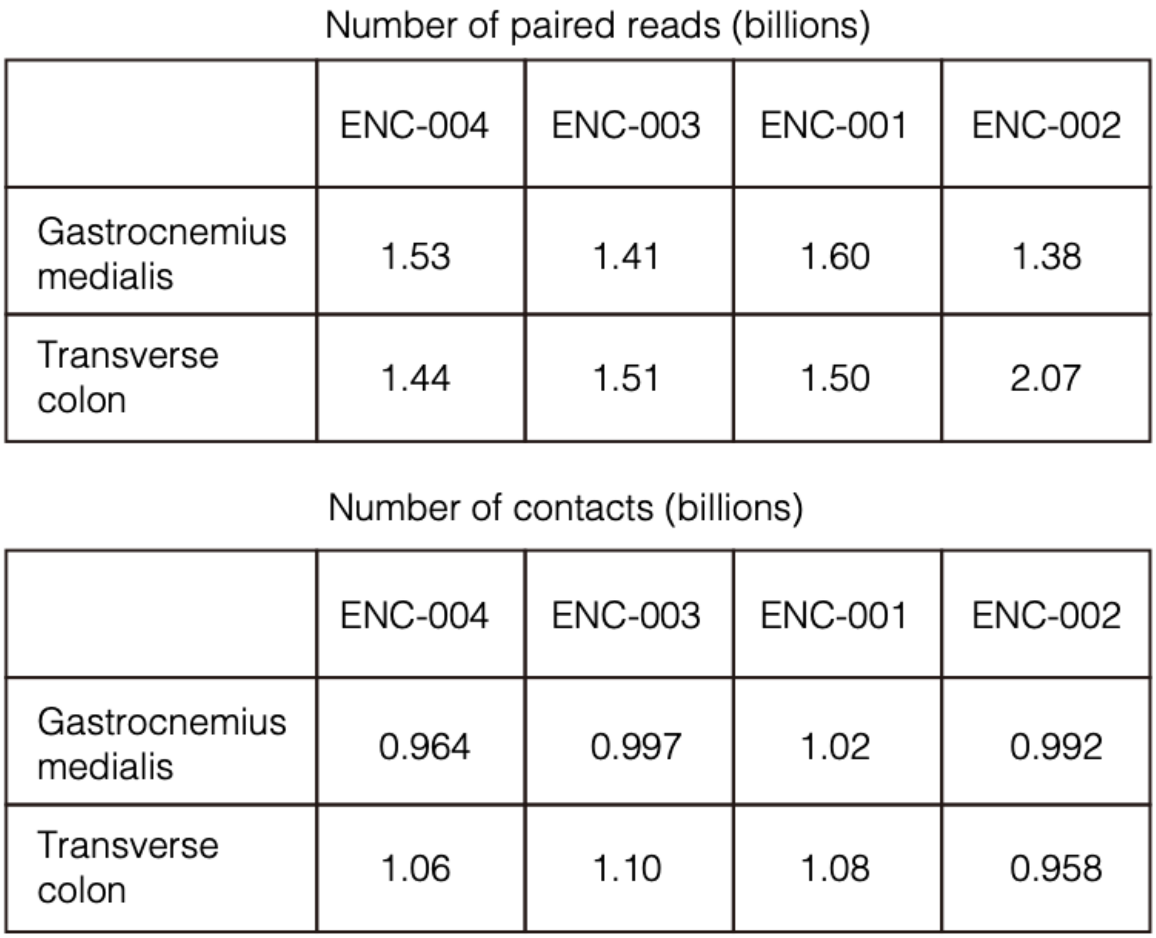


Figure S1.3b. Example of reference-aligned genome-wide Hi-C maps shown for two individuals for their skeletal muscle tissue



Number of paired reads (billions)

	ENC-004	ENC-003	ENC-001	ENC-002
Gastrocnemius medialis	1.53	1.41	1.60	1.38
Transverse colon	1.44	1.51	1.50	2.07

Number of contacts (billions)

	ENC-004	ENC-003	ENC-001	ENC-002
Gastrocnemius medialis	0.964	0.997	1.02	0.992
Transverse colon	1.06	1.10	1.08	0.958

Figure S1.3c. Number of paired reads and number of contacts from reference-aligned genome-wide Hi-C contact maps

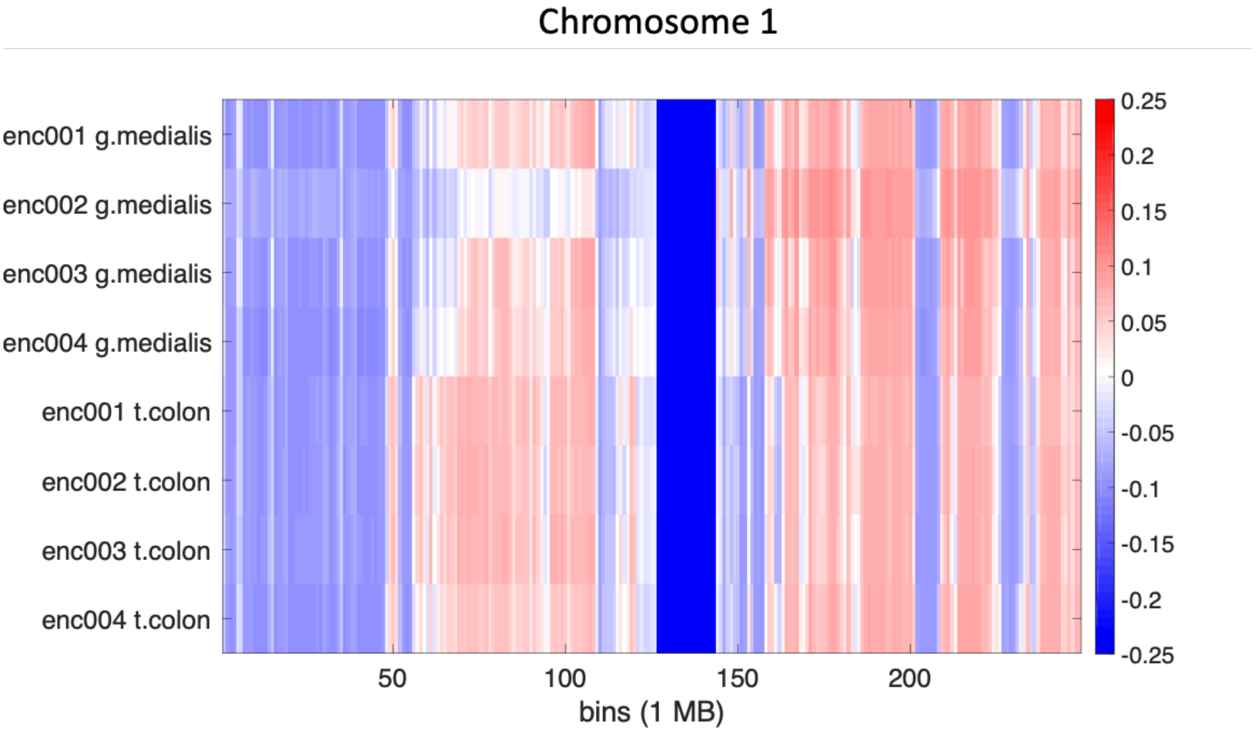


Figure S1.3d. A/B compartment annotation of 4 individuals and 2 tissues for Chromosome 1

Red means the 1 MB region is in A compartment and blue means it is in B compartment. Dark blue band corresponds to centromere.

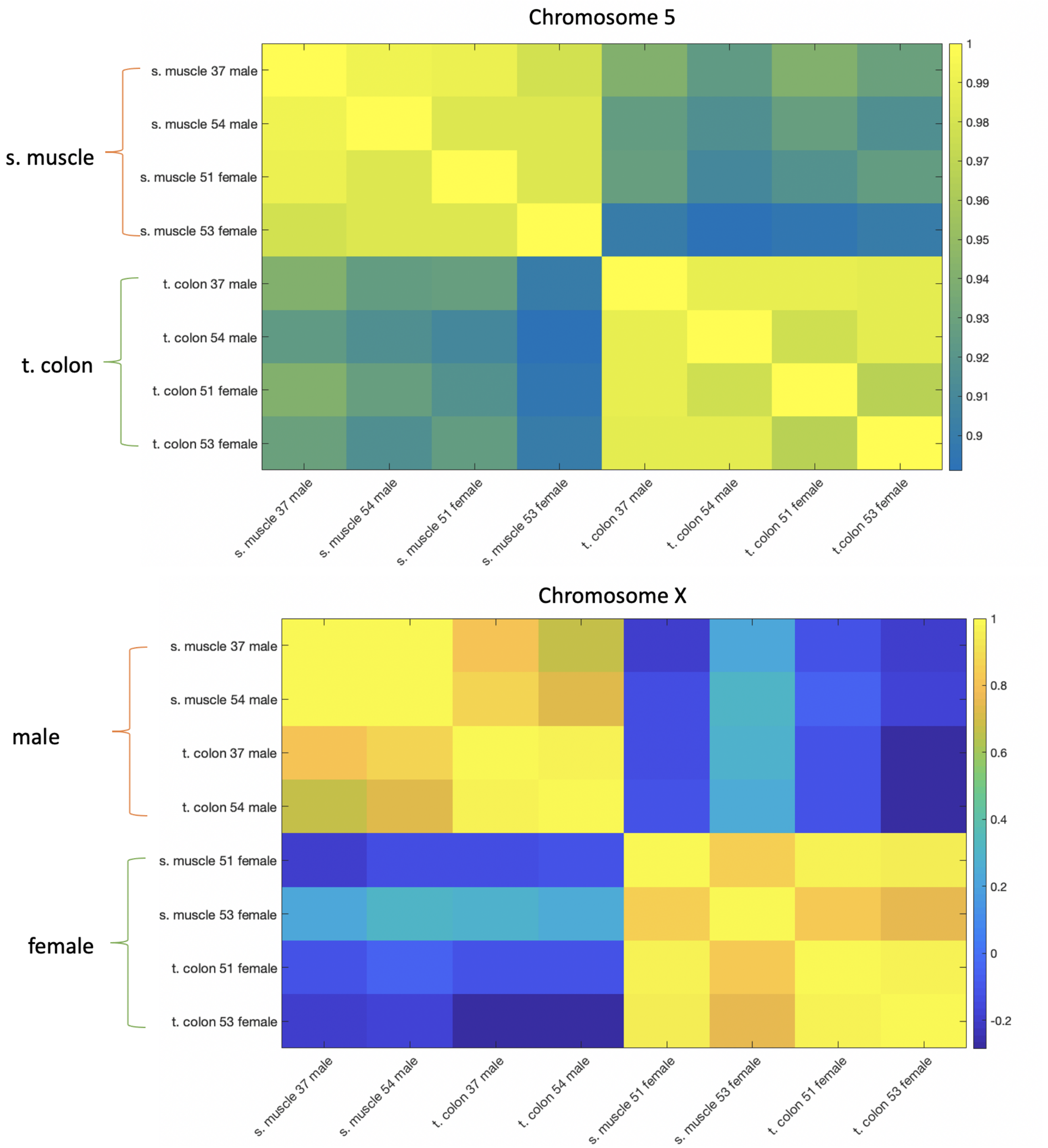


Figure S1.3e. A/B compartments cluster based on tissue in autosomes or sex in Chromosome X

		CLNTRN				GASMED			
Chrom	**Total**	**Indiv 1**	**Indiv 2**	**Indiv 3**	**Indiv 4**	**Indiv 1**	**Indiv 2**	**Indiv 3**	**Indiv 4**
chr1	12397710	6788490	6857968	6640741	6263058	6128125	6218242	6188268	6282354
chr10	3579150	2953439	2990590	2916989	2831756	2915114	2923222	2929788	2996873
chr11	3649051	2869655	2948299	2816640	2778830	2860206	2887171	2860242	2961438
chr12	3552445	2969628	3041869	2919601	2856463	2939877	2980095	2940746	3026456
chr13	2616328	1704423	1724468	1682424	1668618	1628849	1620200	1607366	1674749
chr14	2290870	1410898	1444375	1384357	1381694	1413111	1418356	1402899	1433712
chr15	2079780	1119023	1143407	1119432	1094365	1107323	1136572	1128846	1138837
chr16	1631721	945521	979481	945231	922589	908303	956847	932352	922697
chr17	1386945	1028955	1055611	1029448	1009073	1014825	1056832	1042864	1050488
chr18	1292028	1076987	1086267	1064598	1062482	1067970	1061511	1060048	1081935
chr19	687378	560395	570534	560516	559704	564088	575239	571257	578818
chr2	11729746	8394990	8525226	8309864	7938906	7840021	7789126	7833348	8259552
chr20	830116	679685	692916	663099	670959	674590	680375	673769	691748
chr21	436645	207828	213035	203748	205200	206996	207260	206634	212110
chr22	516636	217507	220358	218031	217149	216516	220675	219461	220895
chr3	7862595	6015900	6134904	5902231	5774444	5434165	5464917	5391741	5857473
chr4	7237110	5562209	5681942	5600166	5271711	5254951	5245142	5313566	5522909
chr5	6590265	5129851	5233827	5049672	4852197	4903566	4859322	4876589	5108741
chr6	5836236	4649176	4768735	4545579	4416778	4436206	4469419	4422407	4647679
chr7	5076891	3880470	3956534	3827833	3644663	3778164	3740318	3791757	3882221
chr8	4212253	3506901	3567642	3465779	3373636	3384990	3372726	3383295	3507176
chr9	3829528	1910384	1963561	1897205	1867856	1748547	1802543	1761180	1909601
chrX	4868760	2923158	4102701	4010558	2739157	2745686	3901268	3961831	2903581
chrY	654940	44238	224	319	42376	42181	636	338	43792

Figure S1.3f. Summary of significant interactions determined by FitHiC2
“Total” is the total number of intrachromosomal interactions for a given chromosome (e.g., chr1, chr2, …, chrY).

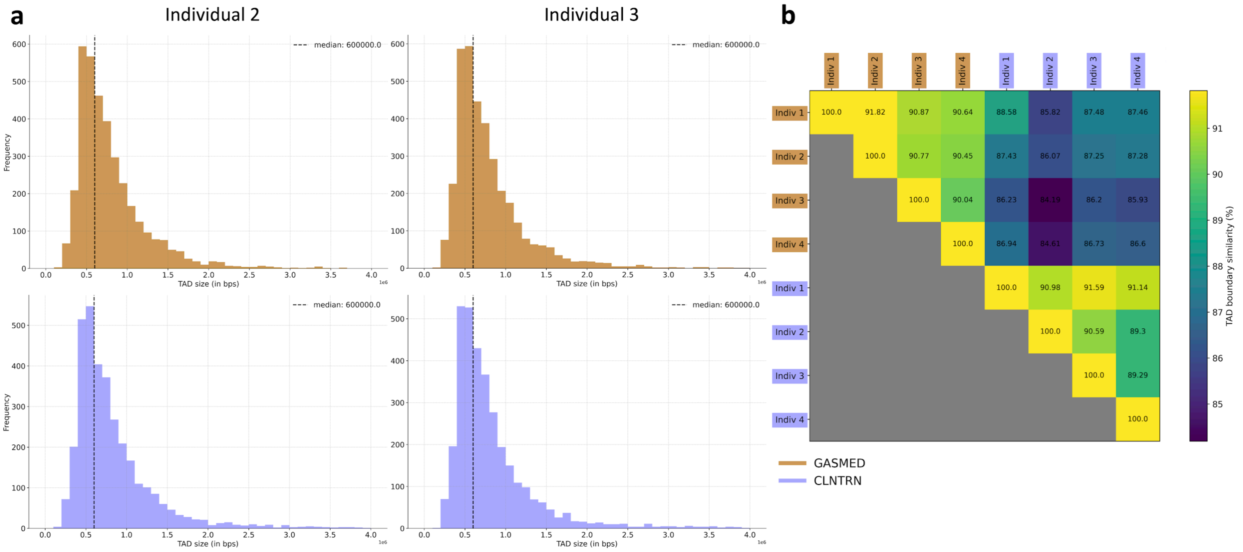


Figure S1.3g. Comparison of TopDom TAD calls for EN-TEx individuals and available Hi-C tissues

(A) TADs were shown to have a similar size distribution and median TAD size across individuals and tissues. The TAD-size distribution of individuals 2 (left) and 3 (right) for available Hi-C tissue types gastrocnemius medialis (GASMED - top) and transverse colon (CLNTRN - top) are shown. **(B)** Pair-wise comparison of TAD calls across all four individuals and available Hi-C tissue types. Within the same tissue, TAD calls were shown to be more similar (i.e., located at the same position along a chromosome) within the same tissue than across different tissues.

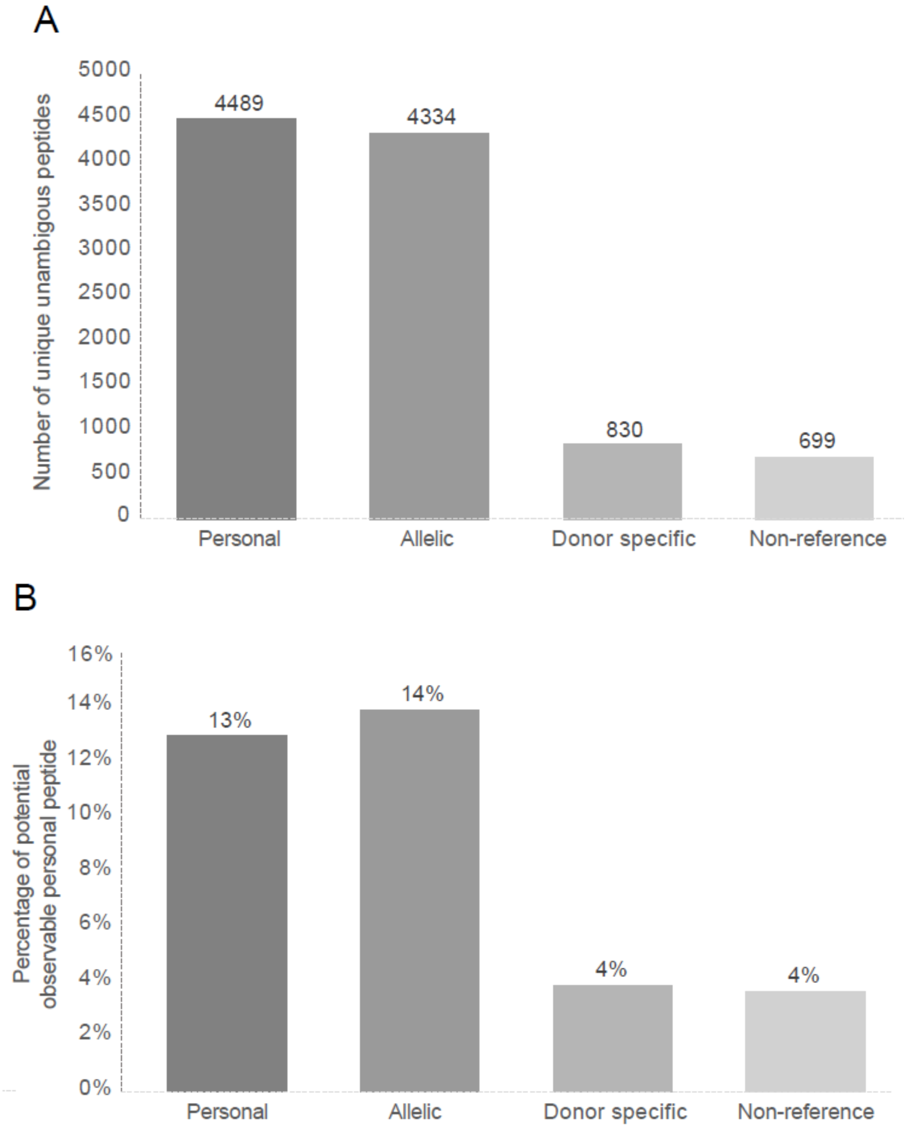


Figure S1.3h. Observed Personal Peptide Summary
(A) Total number of significantly identified unambiguous personal peptides (Filtered for 0.01 Posterior Error Probability and Unambiguous Gene Mapping). Personal category includes all types of personal peptide, Allelic Peptides are those that are specific to only one allele in at least one individual, Donor Specific are peptides that are completely absent in at least one of the four donors, and Non-Reference are peptides that do not match the reference genome. Due to TMT method there is a bias towards the most common peptide-form between the four donors (usually the reference peptide) as TMT boost signal for common peptides. **(B)** This shows the coverage of all potentially observable personal peptides calculated by an in silico tryptic digest. Although in silico peptides are filtered for unambiguity and are limited to amino acid length between 6 and 60, there will be a vast number of unobservable peptides due to MS incompatible charge states and chemical properties.

Peptide	Type	Gene	Ensemble id
VETAGSEPGDTEPJEJGGPGAEPEQK	NewModel	HYOU1	ENSG00000149428
RPESPGDAEAAAAAAPGAPGGR	NewModel	SNX25	ENSG00000109762
SHMMDVQQGSTQDSAJK	NewModel	PDIA4	ENSG00000155660
SQGVQPJPSQGGK	NewModel	FAM120A	ENSG00000048828
ASAAEGVGEPGASAGR	NewModel (nonATG)	WDR26	ENSG00000162923
HPKPEVJGSSADGAJJVSJDGJR	AddedModel	TNXB	ENSG00000168477
DSNQGJYGJSPEGVDR	AddedModel	TNXB	ENSG00000168477
SSJDTGSSJSTDR	AddedModel	IQSEC1	ENSG00000144711
SGASGASAAPAASAAAAJAPSATR	REFerror	CENPV	ENSG00000166582
QTFENQVNR	REFerror	POLR2A	ENSG00000181222
GGGSCVJCCGDJEATAJGR	REFerror	ZNF598	ENSG00000167962
VJWJDEJQQAVDEANVDEDR	REFerror	IQGAP2	ENSG00000145703
GPGGVWAAEAJSDAR	Multiple-Variants	SAA1	ENSG00000173432
JPQEQSQJPNPSEASTTFPESHJR	Multiple-Variants	IFI16	ENSG00000163565
GTJVTVSSASTK	IG Allelism		
VTVSSASTK	IG Allelism		
GTTVTVSSASTK	IG Allelism		
VDEYJAWQHTTJR	AltAssembly	GSTT1	ENSG00000277656
GQHJSDAFAQVNPJK	AltAssembly	GSTT1	ENSG00000277656
VEAAVGEDJFQEAHEVJJK	AltAssembly	GSTT1	ENSG00000277656
AJEMENSQJCK	Some Evidence	HNRNPA0	ENSG00000177733
AEATESAMER	Some Evidence	HNRNPA2B1	ENSG00000122566
GAGSMATGJGEPVYGJSEDEGESR	Weak Evidence	NEDD4L	ENSG00000049759
GSSPEAGAAAMAESJJJR	Weak Evidence	NPLOC4	ENSG00000182446
AJPGSSMADQAPFDTDVNTJTR	No Evidence	FBP1	ENSG00000165140
DTEQTJYQER	No Evidence	LAMB2	ENSG00000172037

Figure S1.3i. Novel Peptides

A

Precision Mapping Table

	RNA-seq	DNA-seq	ChIP-seq	HiC
Unique to reference	0.001	0.002	0.001	0.001
Unique to Haplotypes	0.020	0.045	0.016	0.024
Gain=((Hap1∪Hap2)-Ref)/Ref	0.019	0.045	0.015	0.023

B

DNAseq	ENC-001 37yo male	ENC-002 54 yo male	ENC-003 53 yo female	ENC-004 51 yo female	Average
Hap1&Hap2 (reads)	224,383,498	260,499,463	243,467,335	266,766,596	244,872,476
Ref (reads)	214,959,471	249,767,410	233,252,112	254,366,883	234,192,822
Ref only (%)	0.002	0.002	0.003	0.002	0.002
Hap1&&Hap2¬Ref (reads)	9,842,890.00	11,359,688.00	10,864,254.00	12,883,160.00	11,196768
Hap1&&Hap2¬Ref (%)	0.043784638	0.043502525	0.044504407	0.048206391	0.04549848
Improvement	0.043840948	0.042968188	0.043794772	0.048747356	0.04546103

HiC	ENC-001 37yo male	ENC-002 54 yo male	ENC-003 53 yo female	ENC-004 51 yo female	Average
Hap1&Hap2(reads)	185,631,684	663,535,209	217,727,831	181,462,412	312,089,284
Ref (reads)	181,352,202	649,127,067	212,497,875	177,218,164	305,048,827
Ref Only%	0.001	0.002	0.002	0.001	0.001
Hap1&&Hap2¬Ref (reads)	4,479,970	15,707,837	5,583,723	4,436,038	7,551,892
Hap1&&Hap2¬Ref (%)	0.024	0.024	0.026	0.024	0.024
Improvement	0.024	0.022	0.025	0.024	0.024

CHIPseq	ENC-001 37yo male	ENC-002 54 yo male	ENC-003 53 yo female	ENC-004 51 yo female	Average
Hap1&Hap2(reads)	16,523,306	30,068,339	12,834,649	7,886,359	16,828,163.250
Ref (reads)	16,271,254	29,616,292	12,637,708	7,770,250	16,573,876
Ref Only%	0.001	0.001	0.001	0.001	0.001
Hap1&&Hap2¬Ref (reads)	267,944	491,961	212,830	122,841	273,894
Hap1&&Hap2¬Ref (%)	0.016	0.016	0.017	0.016	0.016
Improvement	0.015	0.015	0.016	0.015	0.015

RNAseq	ENC-001 37yo male	ENC-002 54 yo male	ENC-003 53 yo female	ENC-004 51 yo female	Average
Hap.1(reads)	14,512,594	38,983,875	39,631,595	38,983,875	33,027,984.750
Hap.2(reads)	14,508,357	39,004,820	39,660,757	39,004,820	33,044,688.500
Total Reference(reads)	14,252,554	38,639,784	39,294,659	38,639,784	32,706,695.250
Ref only (reads)	14,053	40,126	58,062	40,126	38,091.750
Ref (%)	0.001	0.001	0.001	0.001	0.001
Hap1&&Hap2¬Ref (reads)	383,723	721,245	724,982	721,245	637,798.750
Hap1&&Hap2¬Ref (%)	0.026	0.018	0.018	0.018	0.020
Improvement	0.025	0.018	0.017	0.018	0.019

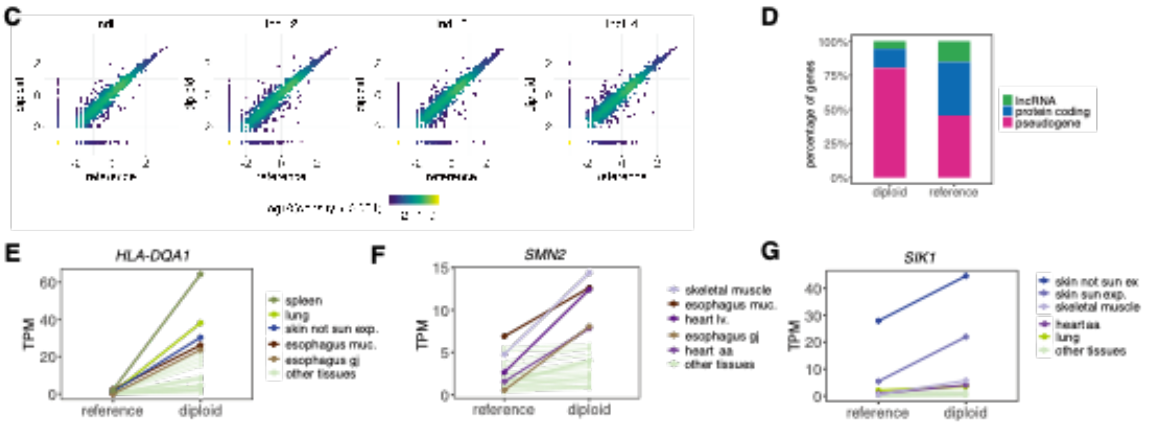


Figure S1.4 Mapping to personal genomes
(A) Summary of percentage with precision mapping. **(B)** Numbers of reads and percentage with precision mapping. More details of precision mapping is available in S1.4. **(C) - (G)** Comparing gene expression mapped to personal genome vs reference genome. **(C)** Scatterplots reporting, for each of the four individuals, gene expression quantifications obtained after mapping to the reference (x axis) and to the diploid genomes (y axis). Expression values are reported in log10(TPM + 0.001). Gene density is color-coded. **(D)** Differential gene expression analysis between quantifications obtained after mapping to the reference and to the diploid genomes. DE analysis was performed with DeSEq2 \cite(*99*), for each donor, on RNA-seq read counts across multiple tissues. Genes with adjusted p-value (Benjamini-Hochberg) < 0.1 and |log2 FC| > 1 were considered differentially expressed. See File: Supp_DE_genes.tsv for a full list of differentially expressed genes. The barplot depicts the gene type for genes upregulated when mapping to the diploid genome (left bar, n genes = 107), and for genes upregulated when mapping to the reference genome (right bar, n genes = 112). **(E) - (G)** examples of genes upregulated when mapping to the diploid genome (ind 3). HLA-DQA1 belongs to the HLA class II alpha chain paralogues. SMN2 belongs to the SMN complex and plays a role in pre-mRNA splicing. The gene is part of an inverted duplication on chromosome 5q13, a region prone to rearrangements and deletions. Mutations in this gene have been associated with spinal muscular atrophy \cite(*100, 101*). SIK1 encodes for a member of the salt-inducible kinase family, which has been associated with pigment gene expression \cite(*102*). Mutations in this gene have been associated with neurodevelopmental impairments \cite(*103, 104*).

Figure S2. Supp. figures to main text section “Measurement of allele-specific activity in diverse assays”

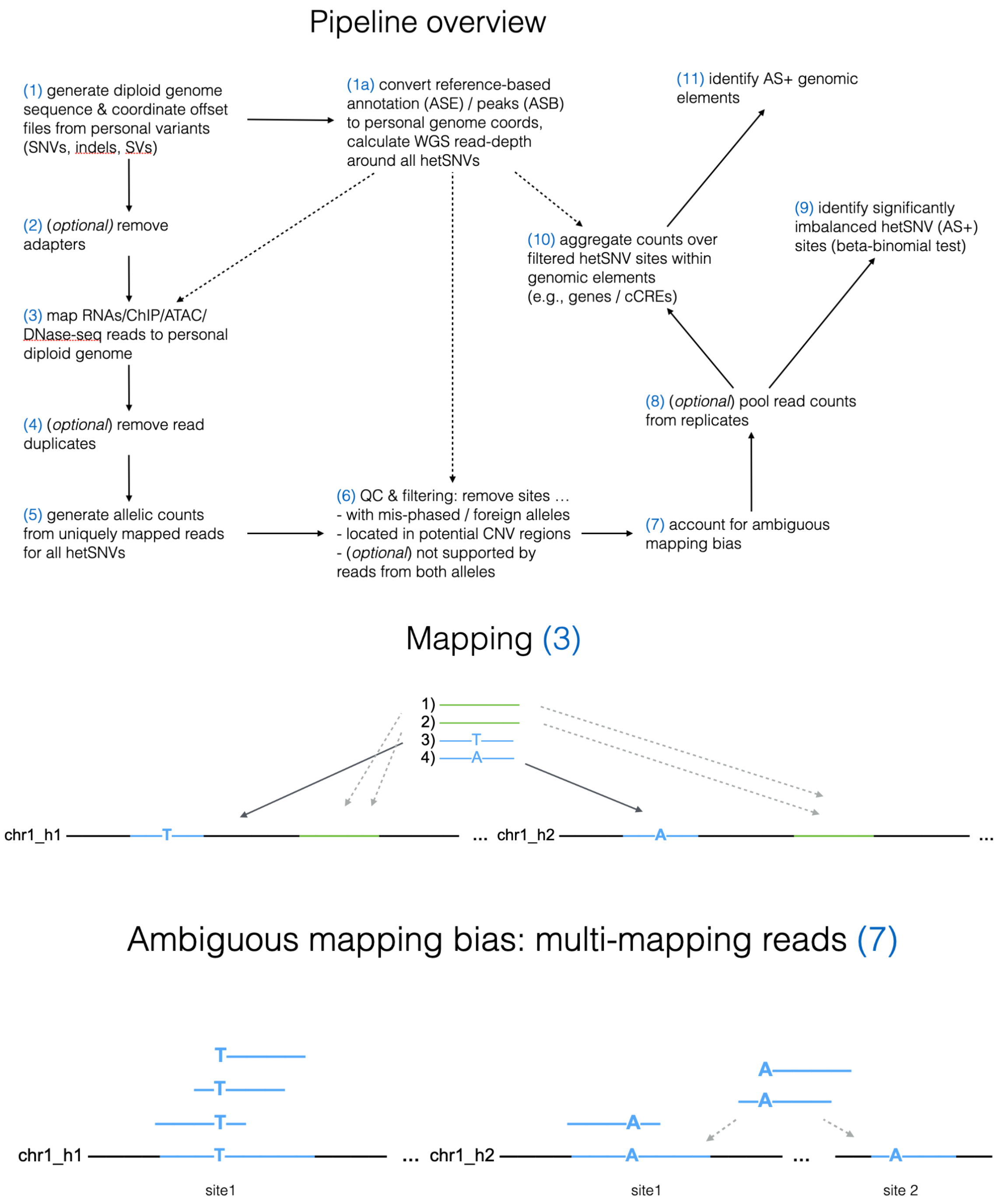
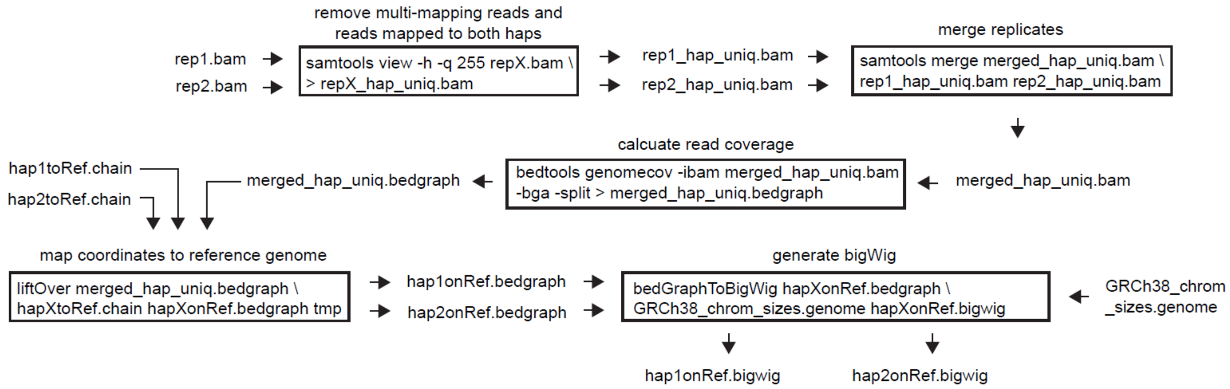

Pipeline overview
(1) generate diploid genome sequence & coordinate offset files from personal variants (SNVs, indels, SVs)
(1a) convert reference-based annotation (ASE) / peaks (ASB) to personal genome coords, calculate WGS read-depth around all hetSNVs
(11) identify AS+ genomic elements
(2) (optional) remove adapters
(10) aggregate counts over filtered hetSNV sites within genomic elements (e.g., genes / cCREs)
(9) identify significantly imbalanced hetSNV (AS+) sites (beta-binomial test)
(3) map RNAs/ChIP/ATAC/DNase-seq reads to personal diploid genome
(8) (optional) pool read counts from replicates
(4) (optional) remove read duplicates
(5) generate allelic counts from uniquely mapped reads for all hetSNVs
(6) QC & filtering: remove sites …
- with mis-phased / foreign alleles
- located in potential CNV regions
- (optional) not supported by reads from both alleles
(7) account for ambiguous mapping bias
Mapping (3)
1)
2)
3) T
4) A
chr1_h1 T
... chr1_h2 A
...
Ambiguous mapping bias: multi-mapping reads (7)
T
T
T
chr1_h1 T
... chr1_h2 A
A
A
A
...
A
site1
site1
site 2

Figure S2.1a. Workflow of the AlleleSeq pipeline



```
samtools view -h -q 255 repX.bam \
> repX_hap_uniq.bam
```

```
samtools merge merged_hap_uniq.bam \
rep1_hap_uniq.bam rep2_hap_uniq.bam
```

```
bedtools genomecov -ibam merged_hap_uniq.bam
-bga -split > merged_hap_uniq.bedgraph
```

```
liftOver merged_hap_uniq.bedgraph \
hapXtoRef.chain hapXonRef.bedgraph tmp
```

```
bedGraphToBigWig hapXonRef.bedgraph \
GRCh38_chrom_sizes.genome hapXonRef.bigwig
```

Figure S2.1b. Generating haplotype-specific signal tracks

The starting bam files are generated by step (3) in Figure S2.1a. Each box is a command to process files. For assays with replicates (as in this case), we pool all the replicates by merging the bam files related to each replicate. If there are no replicates, then “merge replicates” is skipped.

Figures	Source data
Fig.4A	RNA: ENCFF660SLV, ENCFF751QEC CTCF: ENCFF296YDQ, ENCFF255INZ H3K27ac: ENCFF184LPK, ENCFF789APL, ENCFF707VEV, ENCFF298AKE
Fig.4C	RNA: ENCFF281PBY, ENCFF760KXM H3K27ac: ENCFF699EFW, ENCFF075RQB H3K27me3: ENCFF888EIC, ENCFF595OTK, ENCFF011LXD, ENCFF626DTV
Fig.5B, Fig.S6.1bA	RNA: ENCFF719MSG, ENCFF120MML ENCFF337ZBN, ENCFF481IQE H3K27ac: ENCFF339ODV, ENCFF870TZH TF binding clusters: UCSC encRegTfbsClustered
Fig.5C, Fig.S6.1cA	RNA: ENCFF038JEE, ENCFF897TAN H3K27ac: ENCFF143SOY, ENCFF244ISL, ENCFF804MSF, ENCFF976BRQ TF binding clusters: UCSC encRegTfbsClustered
Fig.5D, Fig.S6.1dA	ind3 RNA: ENCFF534JLO, ind3 H3K27ac: ENCFF066DSD, ind3 CTCF: ENCFF417IMY ind2 RNA: ENCFF520WUA ind2 H3k27ac: ENCFF439NXI ind2 CTCF: ENCFF178GEC TF binding clusters: UCSC encRegTfbsClustered
Fig.5E. Fig.S6.1e Fig.S6.1f	ind3 RNA: ENCFF216VOH, ind3 H3K9me3: ENCFF423DVX, ind3 long-read RNA: ENCFF185VYD ind2 RNA: ENCFF187KAR ind2 H3k27ac: ENCFF095CZX ind2 long-read RNA: ENCFF912HPY
Fig.S6.1a	RNA: ENCFF326CGI, ENCFF663VCC H3K27ac: ENCFF935UTO, ENCFF653PKW, ENCFF235IVE, ENCFF226YFN ATAC: ENCFF591BAY, ENCFF332SCG, CTCF: ENCFF800GHL, ENCFF100YUK, ENCFF861WPS, ENCFF056JNV, ENCFF608GCT, ENCFF682AOT TF binding clusters: UCSC encRegTfbsClustered
Fig.S6.1cB	RNA: ENCFF122HNW, ENCFF069KBE, ENCFF483NBR, ENCFF226NNE H3K27ac: ENCFF459LBY, ENCFF949SUD, ENCFF481TGO, ENCFF359AHW, ENCFF252NKY, ENCFF920PYS, ENCFF384MQH, ENCFF270YMP, ENCFF605JUU, ENCFF264CZV, ENCFF867PQG, ENCFF003LQT, ENCFF219DYV, ENCFF113KFQ TF binding clusters: UCSC encRegTfbsClustered
Fig.S6.1cD	RNA: ENCFF411WXY, ENCFF543BVT, ENCFF072VKD, ENCFF484BLA, ENCFF086TFZ, ENCFF351OAS, H3K27ac: ENCFF214DHU, ENCFF209OKJ, ENCFF330KKH, ENCFF343NQH, ENCFF706KXN, ENCFF349JBL, ENCFF945XBP, ENCFF382QHO, ENCFF922CDY, ENCFF778KZF, ENCFF040XEO, ENCFF173QJF, ENCFF033YTT, ENCFF088QFN

Figure S2.1c. Data used to generate signal tracks

Data in blue are given as the accession numbers in the ENCODE portal. TF binding clusters are available at UCSC table browser.

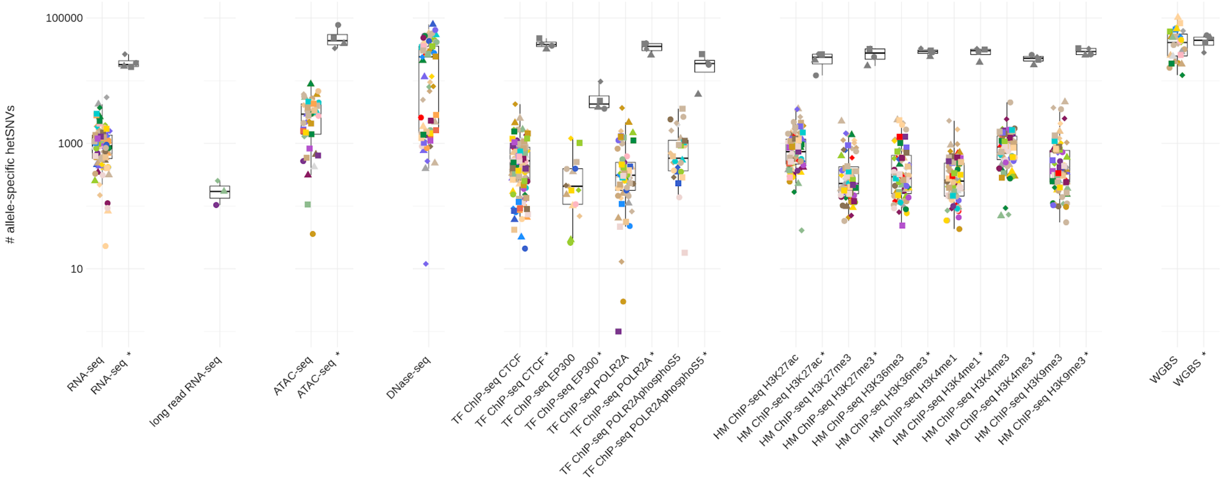


Figure S2.1d. Distribution of the numbers of hetSNVs associated with allele-specific behavior across different EN-TEx donors, tissues, and assays

Tissues are colored as in Figure 1. Call-sets based on pooled reads from all tissues for each donor and assay are shown in gray.

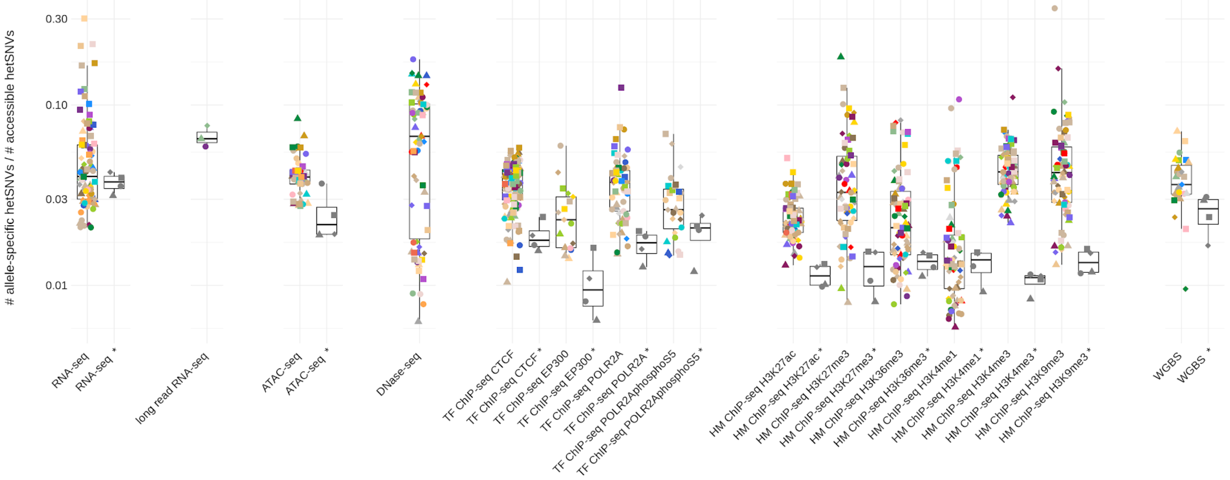


Figure S2.1e. Distribution of the fractions of the numbers of hetSNVs associated with allele-specific behavior relative to the number of accessible hetSNVs

Tissues are colored as in Figure 1. Call-sets based on pooled reads from all tissues for each donor and assay are shown in gray

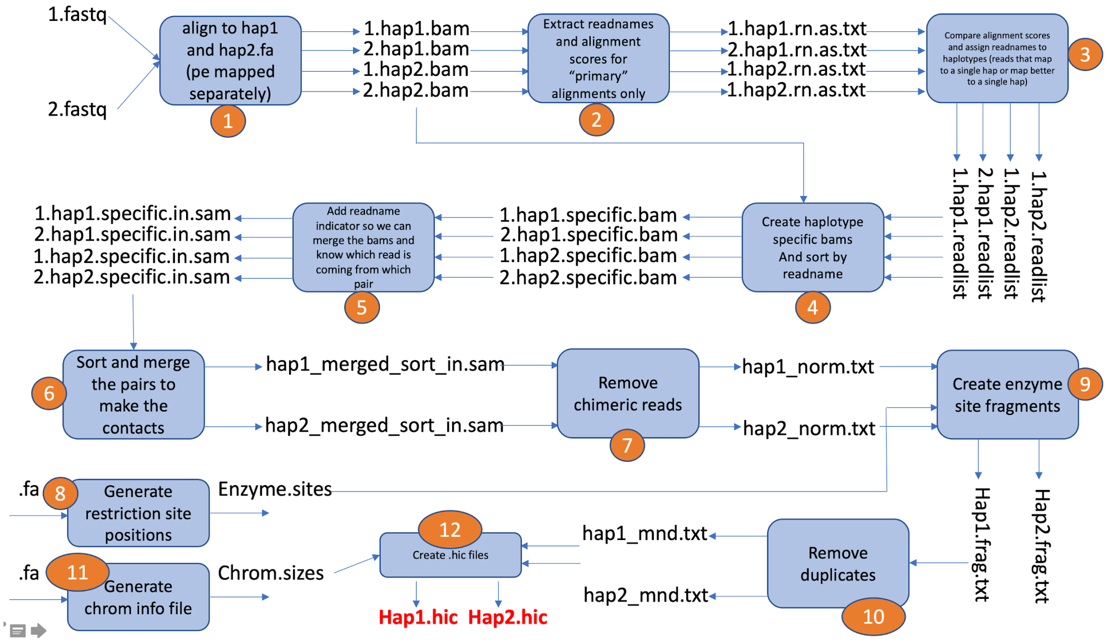
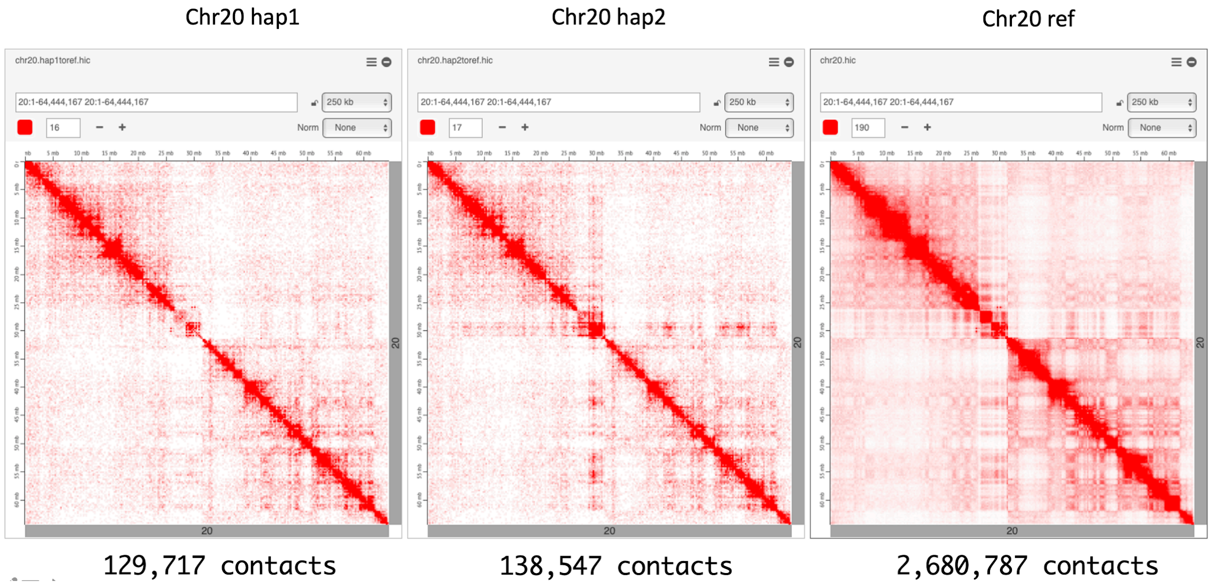


Figure S2.3a. The workflow for generation of haplotype-specific Hi-C contact maps


Chr20 hap1

chr20.hap1toref.hic

20:1-64,444,167 20:1-64,444,167

250 kb

16

Norm None

Chr20 hap2

chr20.hap2toref.hic

20:1-64,444,167 20:1-64,444,167

250 kb

17

Norm None

Chr20 ref

chr20.hic

20:1-64,444,167 20:1-64,444,167

250 kb

190

Norm None

129,717 contacts

138,547 contacts

2,680,787 contacts

Figure S2.3b. The haplotype-specific contact maps for Chr20 generated using the personal genome coordinates

Third map is the bulk Hi-C contact map of Chr20 generated using the reference genome.

individual/tissue	intra-chromosomal interactions	in hap1	in hap2	hap1 or hap2	significantly imbalanced
ind1 skeletal muscle	39,013,901.00	4,049,203.00	4,034,602.00	7,041,417.00	577,728.00
ind2 skeletal muscle	4,405,480.00	1,117,328.00	1,146,381.00	2,072,227.00	140,317.00
ind3 skeletal muscle	40,412,585.00	4,345,533.00	4,359,297.00	7,493,069.00	574,836.00
ind4 skeletal muscle	41,569,344.00	4,028,293.00	4,021,800.00	6,983,660.00	523,931.00
ind1 transverse colon	45,534,793.00	4,942,660.00	4,924,000.00	8,574,917.00	702,953.00
ind2 transverse colon	25,548,308.00	2,148,267.00	2,151,803.00	3,842,621.00	261,752.00
ind3 transverse colon	43,917,995.00	4,716,549.00	4,722,227.00	8,118,858.00	609,973.00
ind4 transverse colon	43,406,680.00	4,343,051.00	4,334,617.00	7,506,125.00	583,468.00

Figure S2.3c. The number of Hi-C contacts obtained from haplotype specific Hi-C contact maps

Figure S3. Supp. figures to main text section “Aggregation of allele-specific events, forming a catalog”

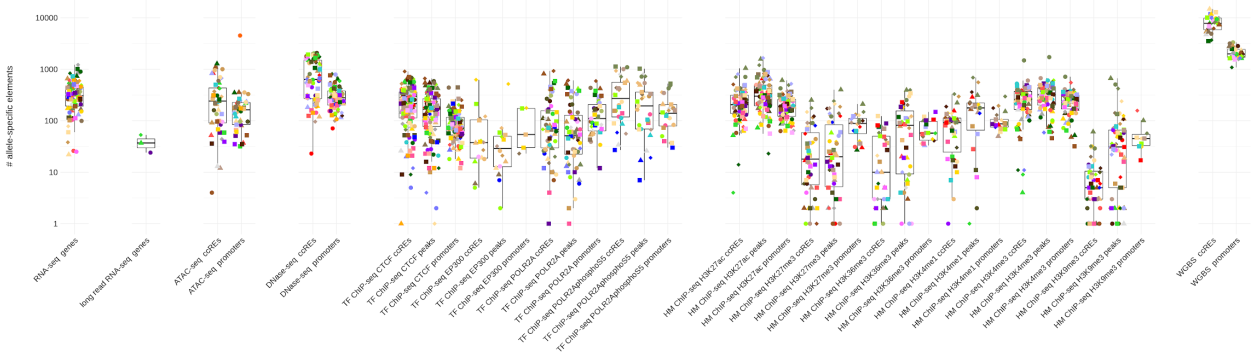


Figure S3.1a. Distribution of the numbers of genomic elements associated with allele-specific behavior across different EN-TEx donors, tissues, and assays

Tissues are colored as in Figure 1. Call-sets based on pooled reads from all tissues for each donor and assay are shown in gray.

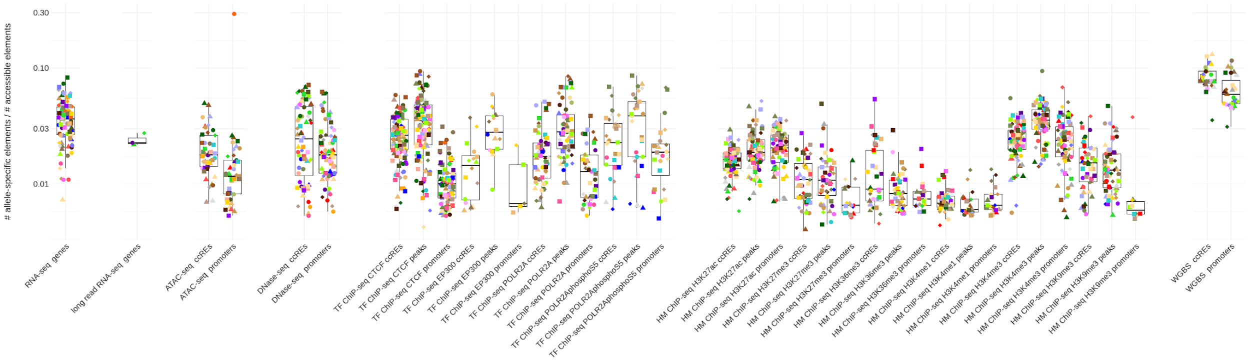


Figure S3.1b. Distribution of the fractions of the numbers of genomic elements associated with allele-specific behavior relative to the number of accessible hetSNVs
Tissues are colored as in Figure 1. Call-sets based on pooled reads from all tissues for each donor and assay are shown in gray.

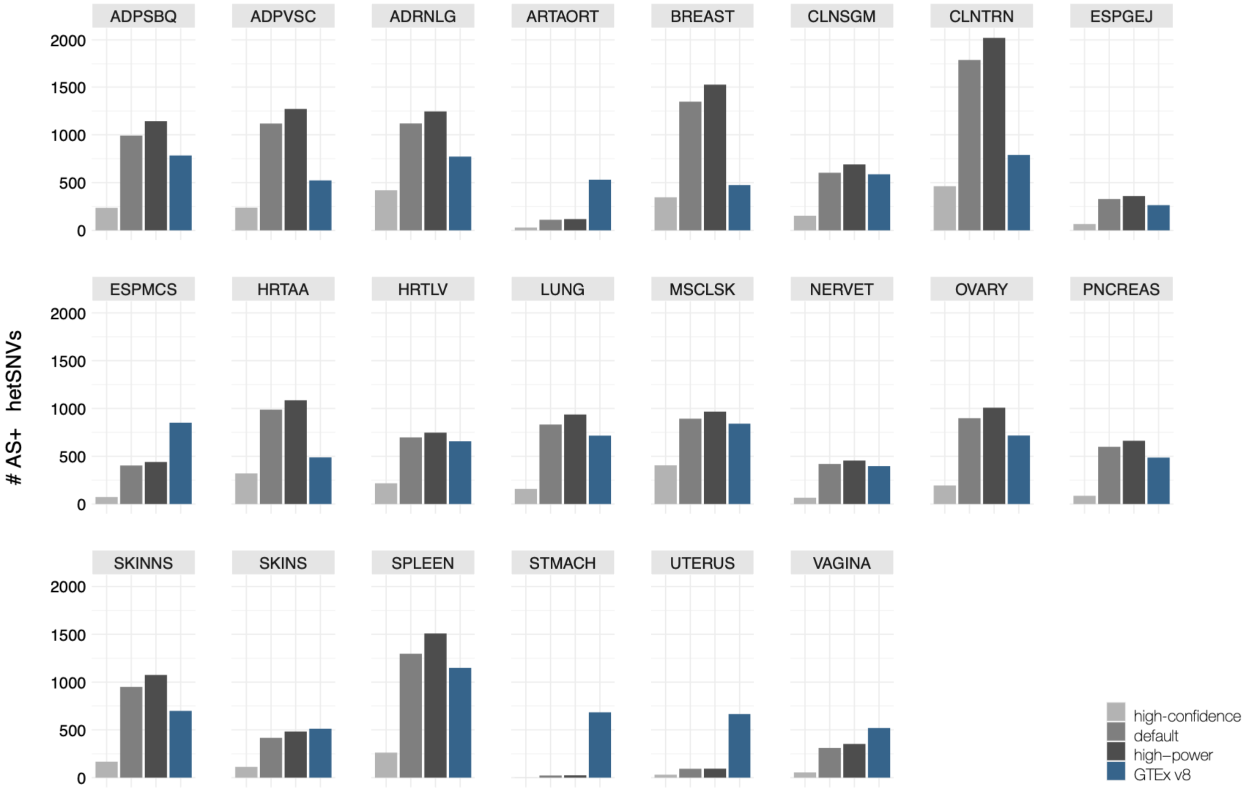


Figure S3.3. Numbers of AS+ hetSNVs detected from RNA-seq in different tissues of individual 3

To produce high-power tissue-specific call-sets, for each tissue we called ASE and ASB sites at a relaxed FDR threshold if the hetSNV was called AS in the pooled call-set, and at the usual 10% FDR otherwise. The “relaxed” FDR varied somewhat from tissue to tissue due to granularities in calculation but was at most 20%.Typically, the high-power call-sets produce 10-20% more AS hetSNVs than the original. File: AS_highpower_set.tar.gz

Figure S4. Supp. figures to main text section “Mining the catalog”

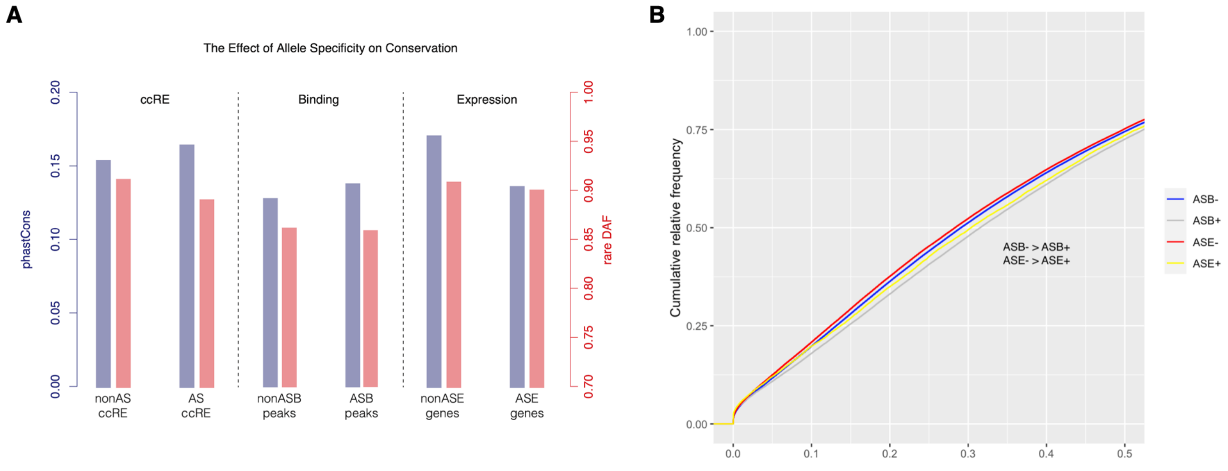


Figure S4.2. Conservation of AS

(A) shows the conservation of various AS annotations are calculated using phastCons and rare DAF (see supplemental text). Specifically we considered AS+/AS- ccREs, AS+/AS- binding peaks from H3K27ac, and AS+/AS- genes. An alternate way to see the same phenomena is to show the cumulative relative frequency of variants, shown in **(B)**. Here we can see that AS- events demonstrate stronger purifying selection as compared to AS+ events.

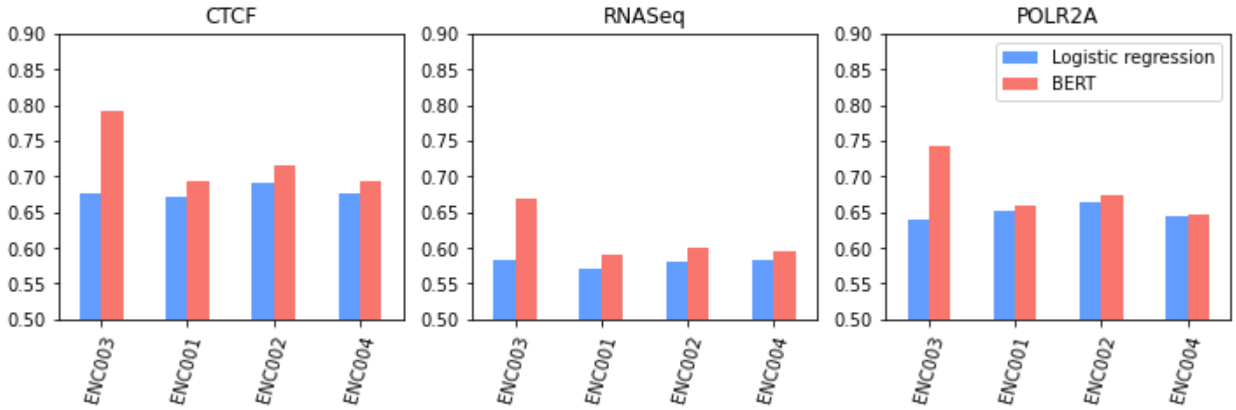
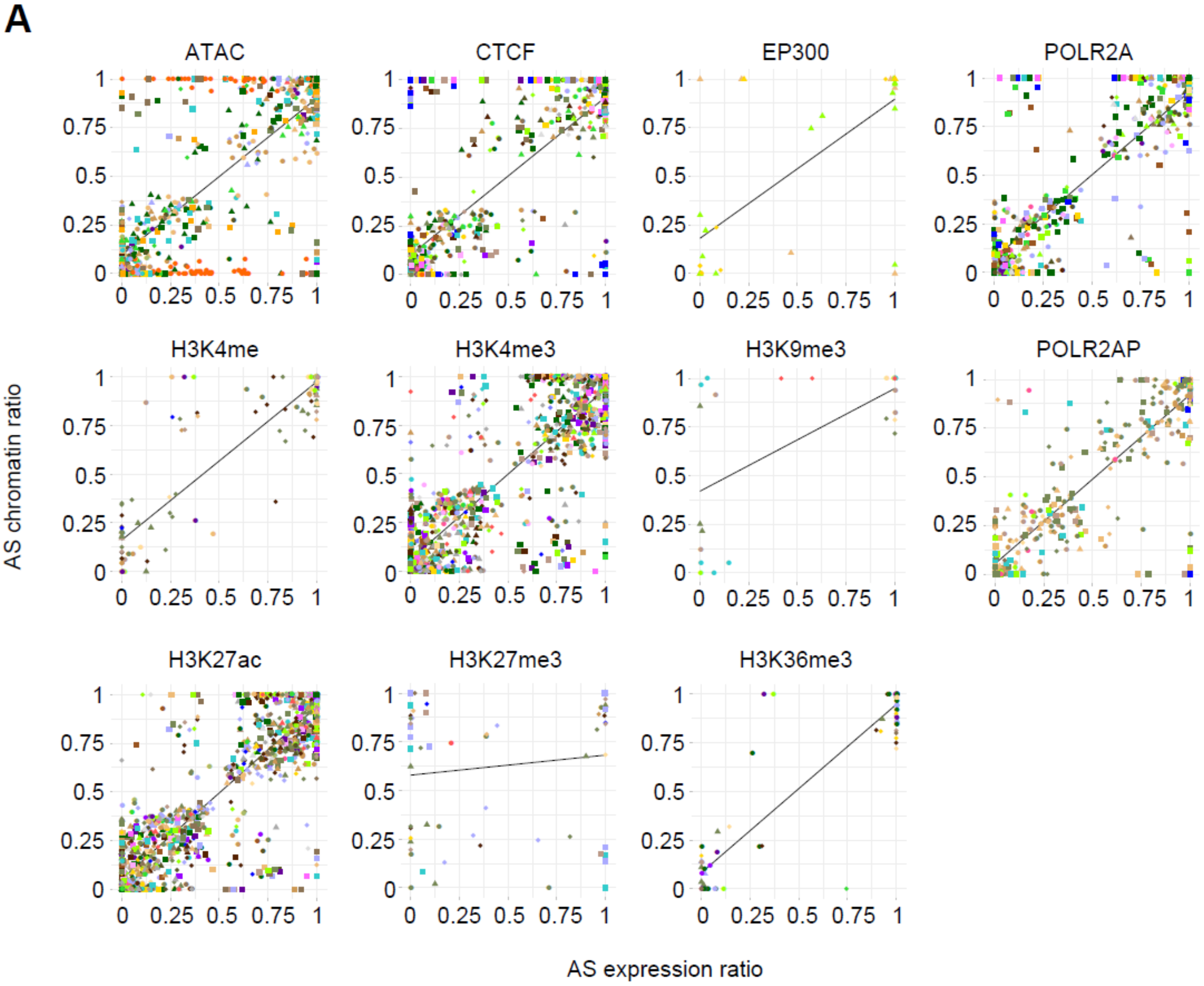


Figure S4.3. The performance of allelic effect prediction models trained
“Logistic regression” refers to simple logistic regression on the dna2vec embedding of the input sequence; “BERT” refers to the fine-tuned DNABERT model. Both models are trained on SNPs of individual 3 and the results on the validation sets from all four individuals are reported.


A
ATAC
CTCF
EP300
POLR2A
H3K4me
H3K4me3
H3K9me3
POLR2AP
H3K27ac
H3K27me3
H3K36me3
0
0.25
0.5
0.75
1
AS chromatin ratio
AS expression ratio

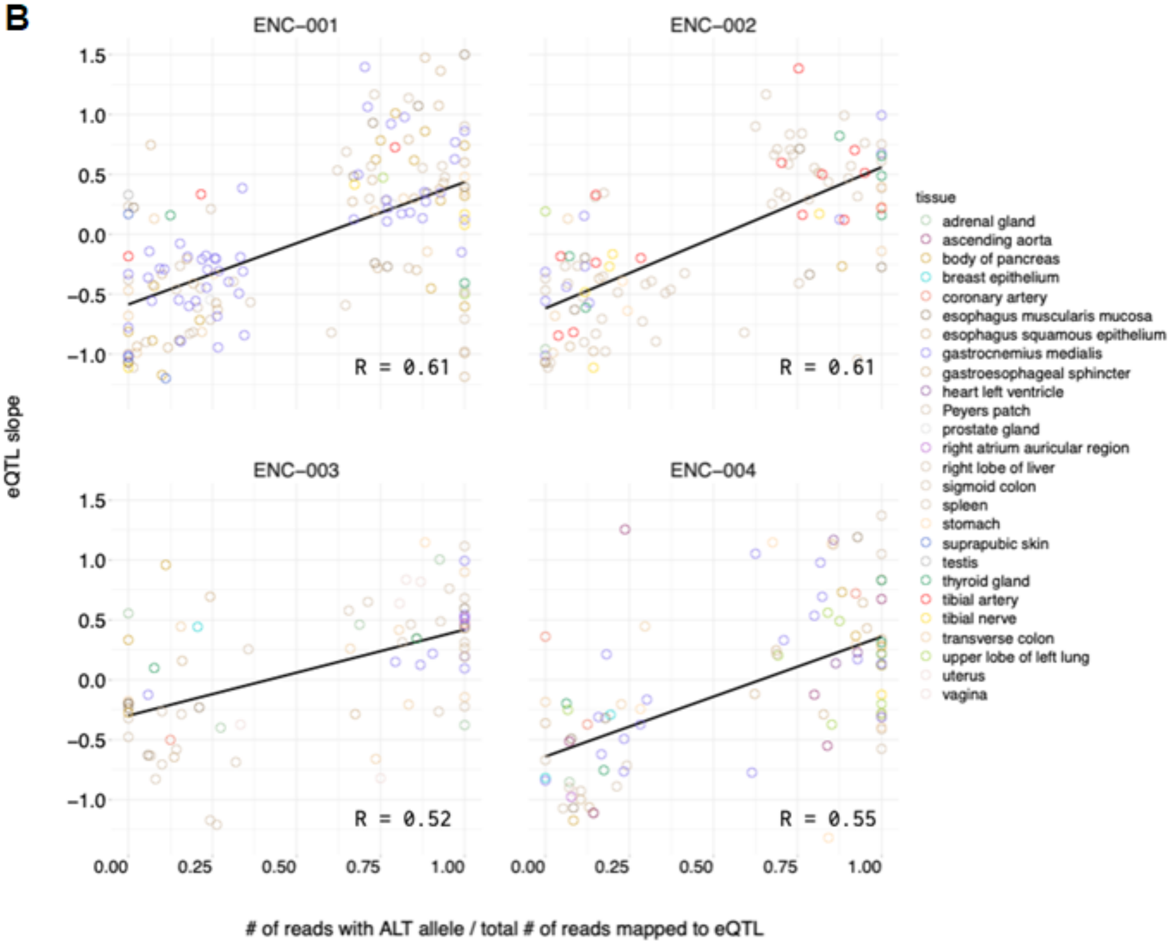


Figure S4.4a. Compatibility between allelic events.

(A) Compatibility between AS chromatin state of the promoters (+/- 2kb from the TSS) and the AS expression of the corresponding genes. AS chromatin ratio is the fraction of hap1 ChIP-seq reads among total reads. AS expression ratio is the fraction of hap1 RNA-seq reads among total reads. Each dot is a gene in a given tissue (marked by colors) and individual (marked by shape). See Fig. 1 for details of colors and shapes. **(B)** AS H3K27ac hetSNVs in ENTEx individuals and known GTEx eQTLs.

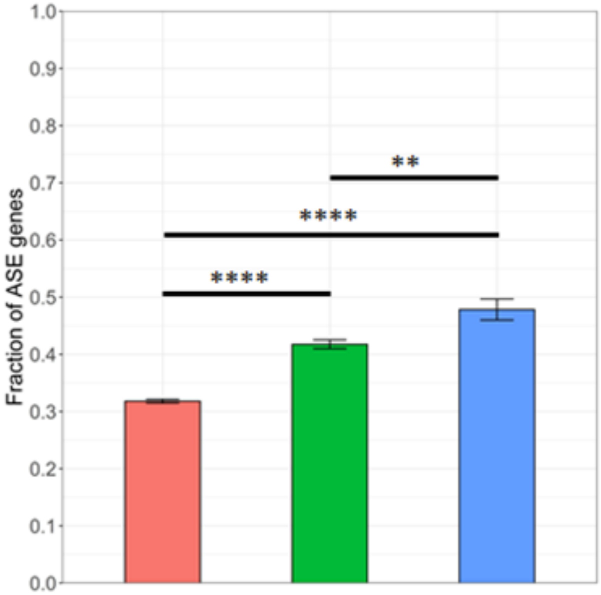


Figure S4.4b. Enrichment of ASE in genes with promoter ASM

Enrichment of ASE in genes with promoter ASM (allele-specific methylation), with (blue) or without (green) ASB of transcription, relative to genes associated with non-regulatory ASM variants (red). ** p<0.01, **** p<0.0001, χ2 test.

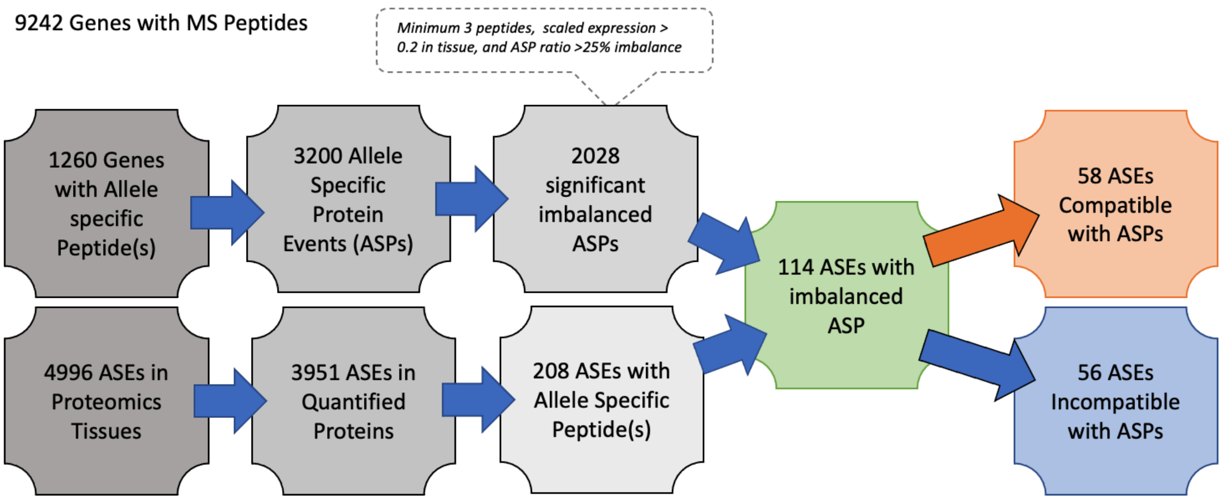

9242 Genes with MS Peptides
Minimum 3 peptides, scaled expression > 0.2 in tissue, and ASP ratio >25% imbalance
1260 Genes with Allele specific Peptide(s)
3200 Allele Specific Protein Events (ASPs)
2028 significant imbalanced ASPs
4996 ASEs in Proteomics Tissues
3951 ASEs in Quantified Proteins
208 ASEs with Allele Specific Peptide(s)
114 ASEs with imbalanced ASP
58 ASEs Compatible with ASPs
56 ASEs Incompatible with ASPs

Figure S4.4c. Flow Chart of Filtering and ASE/ASP Comparison
Proteomics data was mapped at the gene level and filtered for proteins containing allele specific peptides. ASPs were calculated for each tissue in which allelic peptides were quantified. The ASP ratio was calculated as the summed peptide intensity of the first allele divided by the total specific to either allele. ASPs were filtered by the number of peptides expression level and ASP ratio. A p-value was calculated at 0.7 using z-scores.

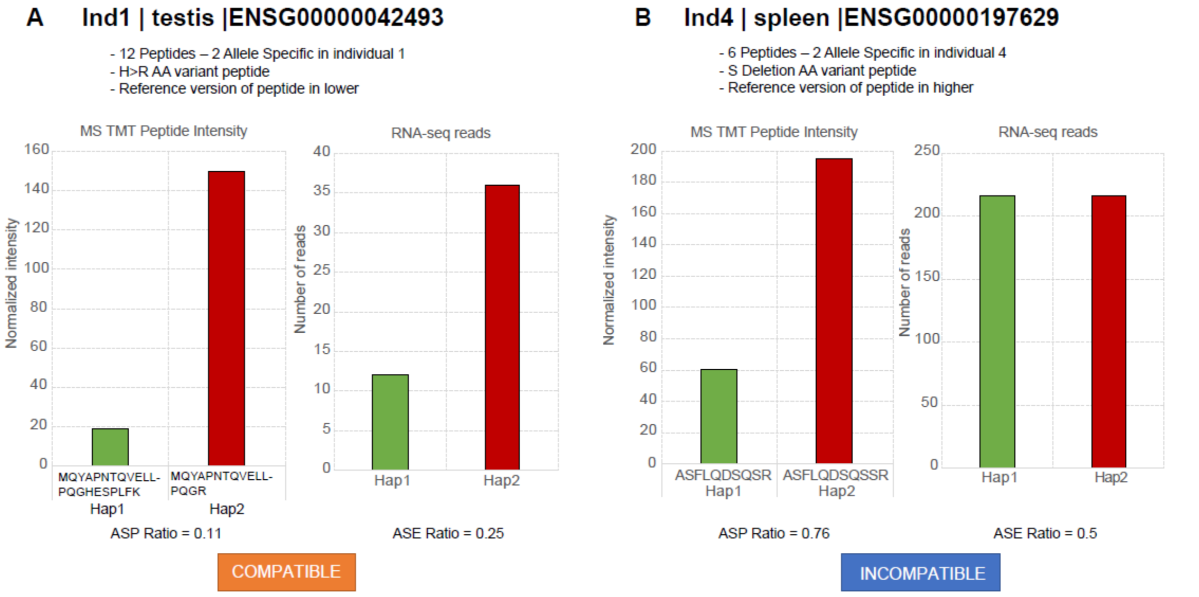


Figure S4.4d. Compatibility between AS mRNA and AS peptide.
(A) An example of a compatible ASP and ASE ratio, both the proteomics and transcriptomics agree that the second allele is expressed more highly. **(B)** An example of an incompatible ASP/ASE pairing the transcriptomics does not show any bias in the gene expression however, at the protein level the second allele is more highly expressed.

Figure S5. Supp. figures to main text section “Examples of coordinated AS activity across assays”

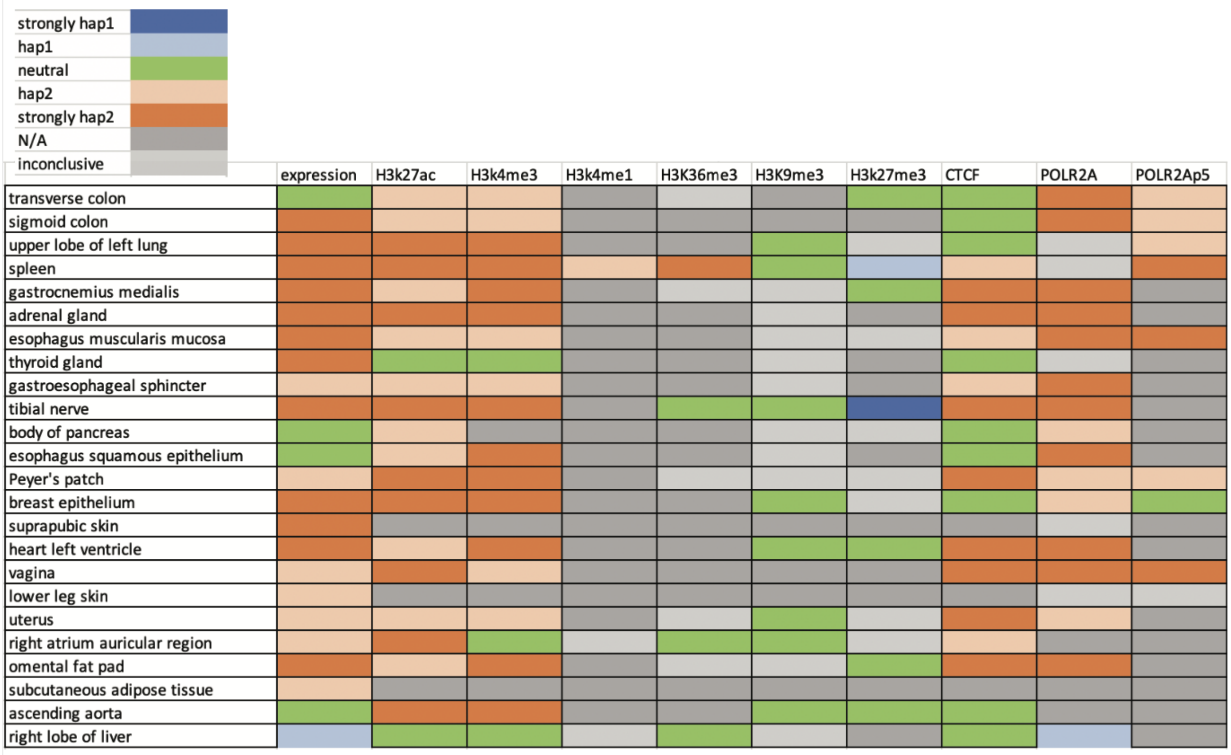
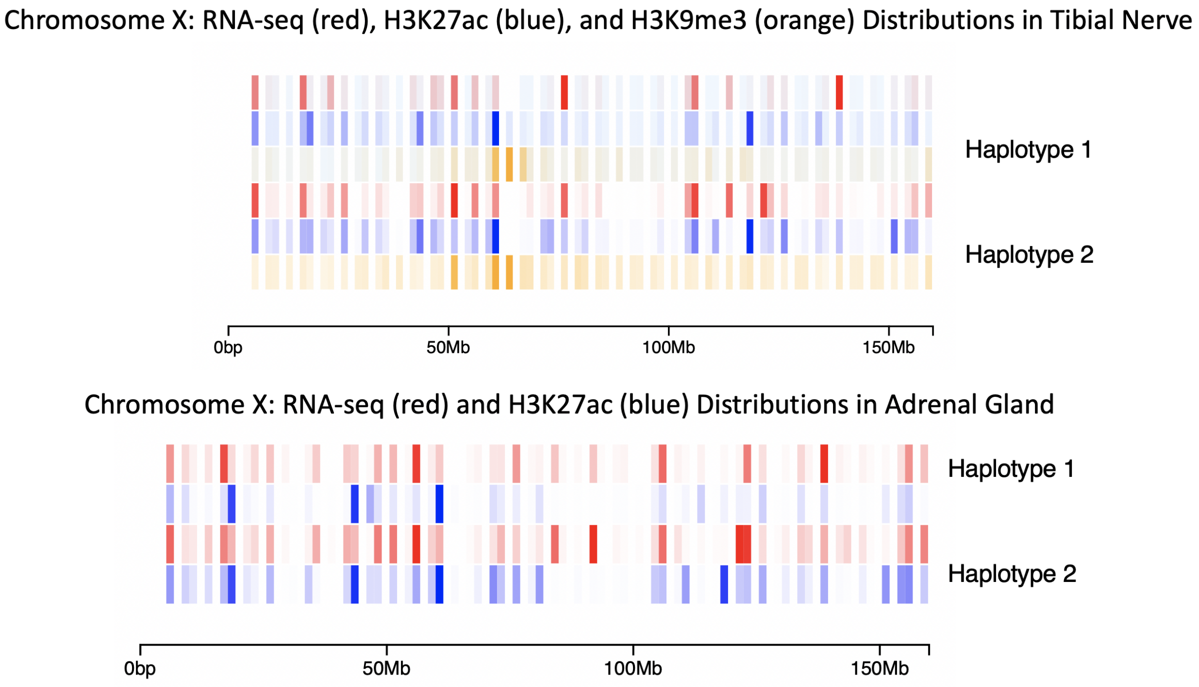


Figure S5.1a. Heatmap to show haplotype specificity of Chr X for all the assays and tissues of individual 3

Orange color indicates that there are more expression and binding peaks in haplotype 2 and blue color indicates that there are more expression and binding peaks in haplotype 1. Green means expression and binding are balanced between haplotypes. Light gray means the number of data points is small, therefore we cannot conclude which haplotype has more expression and binding; while dark gray means we do not have data for that assay and tissue.



Chromosome X: RNA-seq (red), H3K27ac (blue), and H3K9me3 (orange) Distributions in Tibial Nerve

Chromosome X: RNA-seq (red) and H3K27ac (blue) Distributions in Adrenal Gland

Figure S5.1b. Chromosome painting of ChrX using RNA-Seq and ChIP-Seq in both haplotypes of individual 3 in two tissues

This plot also depicts that the active haplotype is haplotype 2 in Chr X of individual 3 as there is more activity in haplotype 2.

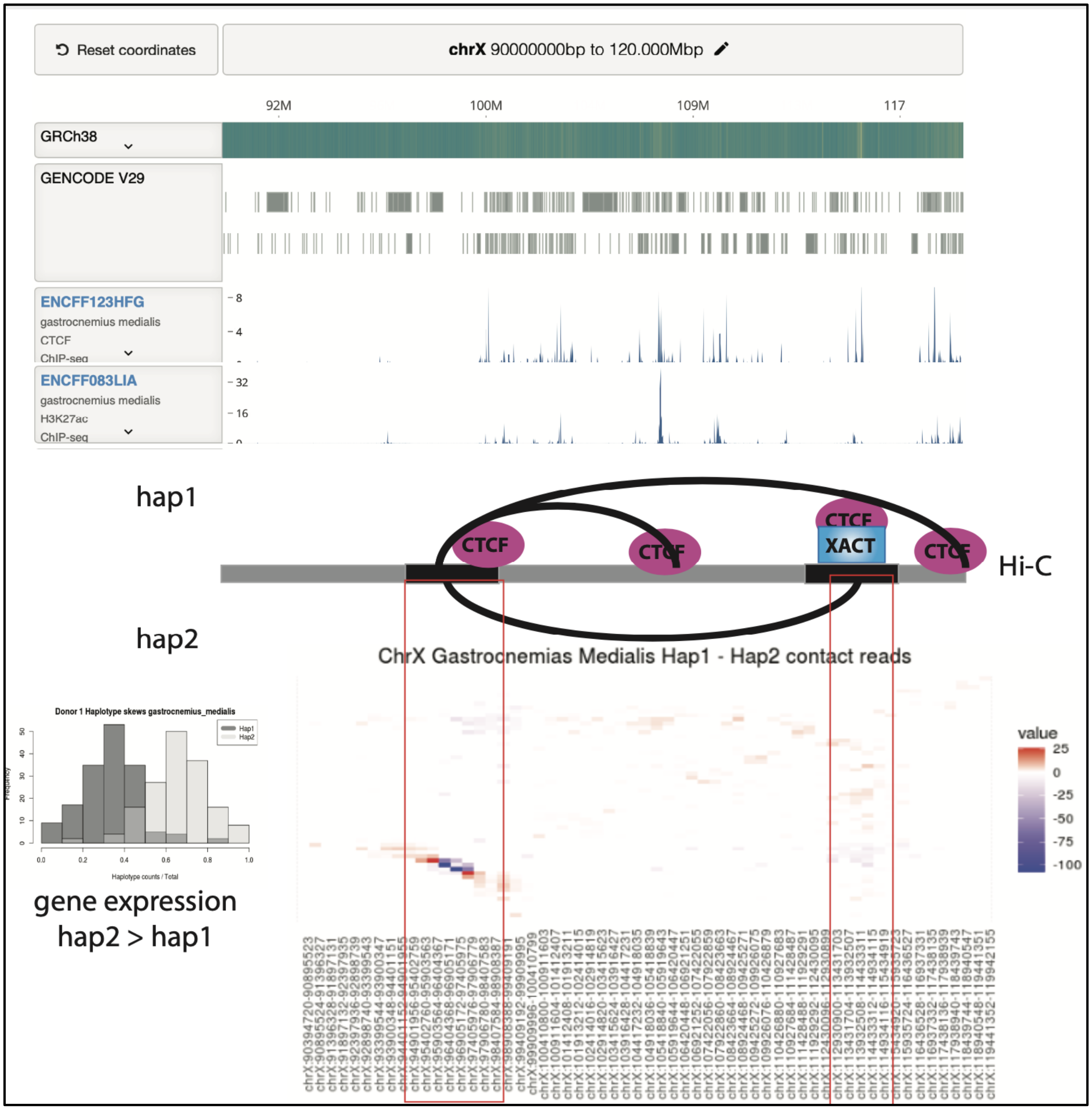


Figure S5.1c. XACT locus on ChrX is shown to have haplotype specific chromatin interactions with an upstream region.
In the signal tracks, both XACT and upstream locus are shown to have CTCF bound, which is also associated with H3K27ac signal. The heatmap shows the differential chromatin interactions from haplotype resolved Hi-C. The allele-specific Hi-C interaction with XACT locus and an upstream element is located on the active haplotype, which was characterized by the difference in the allele-specific gene expression values (histogram).

Figure S6. Supp. figures to main text section “Relating SVs to chromatin & expression”

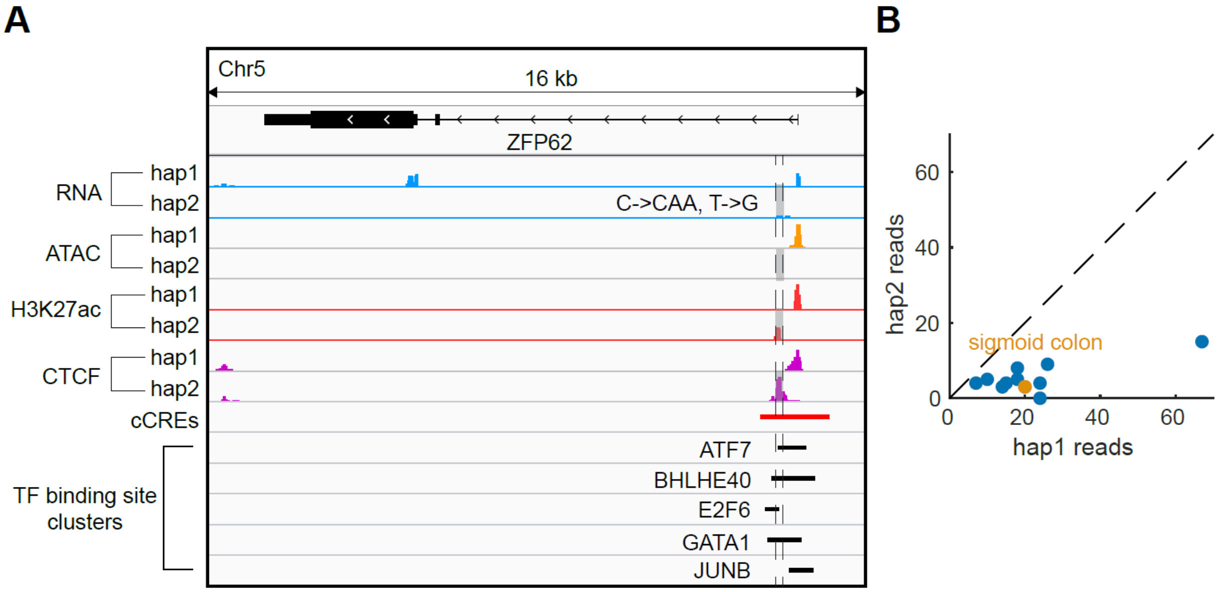


Figure S6.1a. An indel that potentially changes gene expression.
(A) In the sigmoid colon of individual 2, the gene ZFP62 has lower expression in haplotype 2. The TSS region of ZFP62 in hap2 shows lower chromatin accessibility and changes in the positions of H3K27ac and CTCF binding peaks, compared with the same region in hap1. In hap2, a 2-bp insertion and a SNV were found in a cCRE near the TSS of the gene (the two variants are very close and are shown together by a single grey box). These variants and nearby variants that cannot be phased (not shown) might affect the function of the cCRE. **(B)** The gene has lower hap2 expression in multiple tissues, suggesting a universal factor changing the expression between haplotypes.

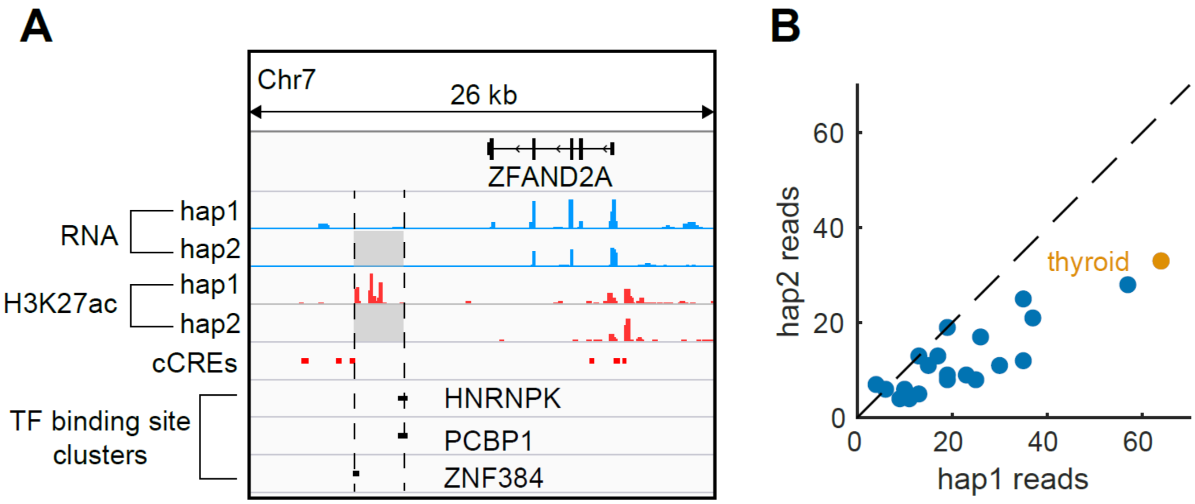


Figure S6.1b. Shadow figure associated with Fig. 5B

(A) Similar to Fig. 5B, the deletion in hap2 could disrupt cCREs identified in thyroid and the binding of several TFs. **(B)** ZFAND2A has lower hap2 expression among multiple tissues, suggesting that the deletion may affect the gene's expression globally.

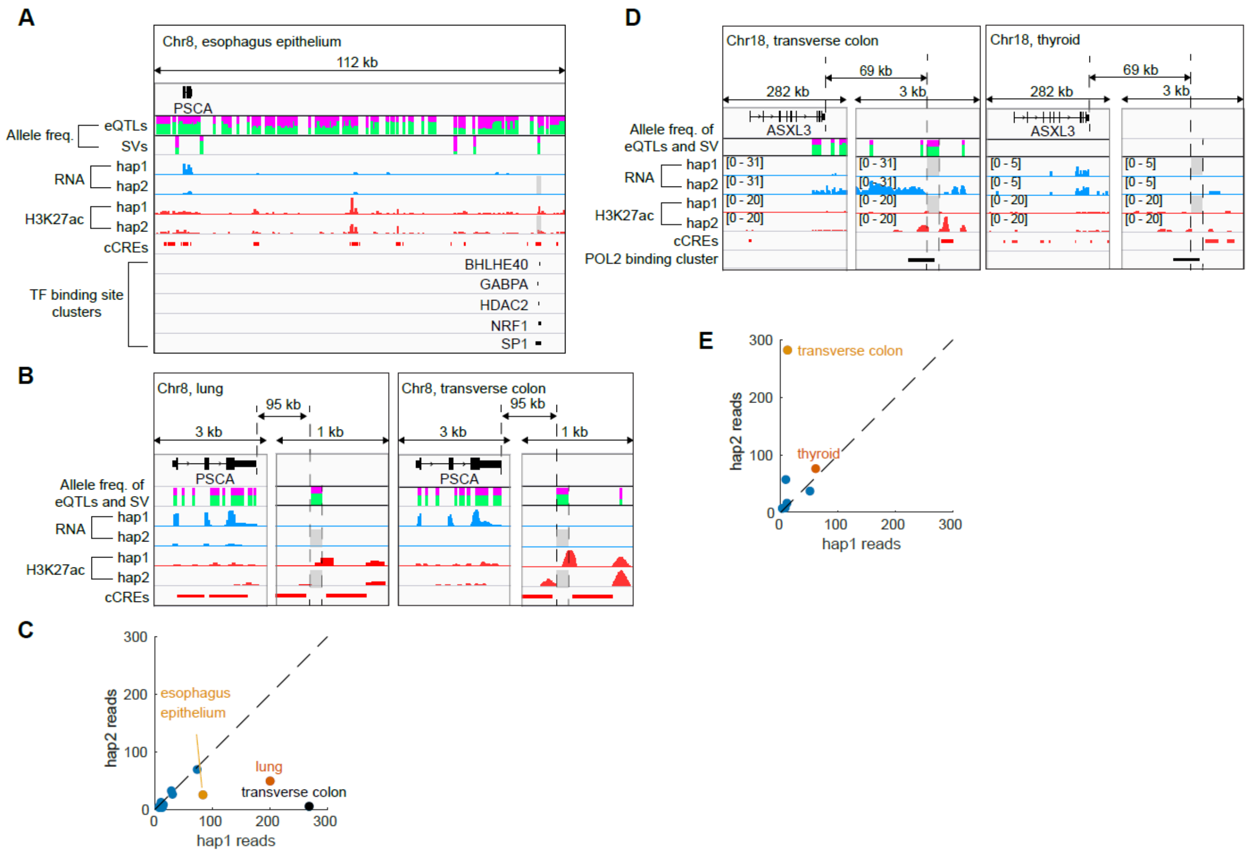


Figure S6.1c. SVs potentially linked to eQTLs.

Panels **(A)** - **(C)** are shadow figures of Fig. 5C. **(A)** The panel is the same as Fig. 5C, but shows a panoramic view near the gene PSCA, including additional eQTLs that are compatible with the AS expression of PSCA. The allele frequencies of the hap2 alleles at these eQTL sites are shown as the heights of the green bars. SVs near PSCA and their allele frequencies are also shown. The left four SVs are deletions in hap1, and rightmost SV is the hap2 deletion shown in Fig. 5C. cCREs and TF binding sites that could be potentially disrupted by the deletion of interest are shown. **(B)** PSCA also has lower expression from hap2 in the lung and the transverse colon of individual 3. In both tissues, the deletion has similar allele frequency with some of the tissue-specific eQTLs compatible with the AS expression of PSCA, and appears to remove a H3K27ac peak in hap2, potentially causing the reduced expression of PSCA. **(C)** Imbalance in the AS expression of PSCA appears to be restricted to three tissues shown in **(A)** and **(B)**. **(D)** Another example of a deletion that could be in linkage with compatible eQTLs of ASXL3. In the transverse colon of individual 2, ASXL3 has lower expression in hap1. The relevant deletion is in hap1 and appears to disrupt H3K27ac and cCREs near the gene. Note that the H3K27ac levels at this cCRE and the expression levels of PSCA are both lower in thyroid than in transverse colon, suggesting an association between the activity of this cCRE with the expression of PSCA. **(E)** Imbalance in the AS expression of ASXL3 appears tissue-specific.

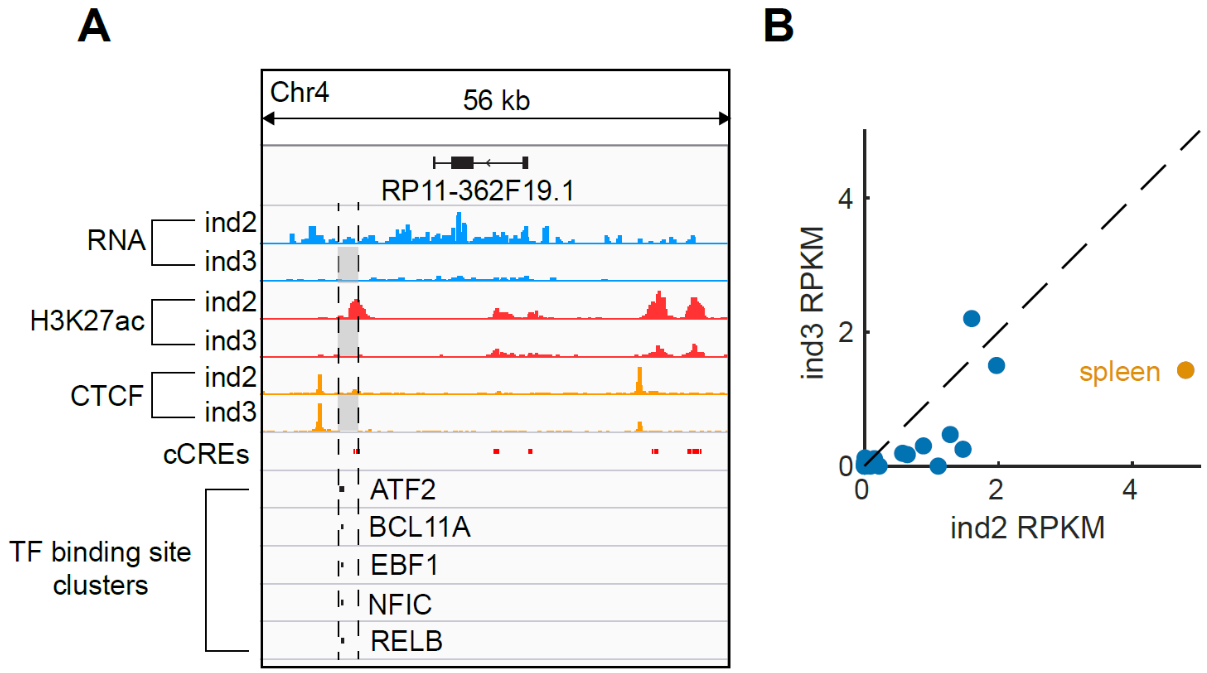


Figure S6.1d. Shadow figure associated with Fig. 5D
(A) Similar to Fig. 5B, the deletion in hap2 could disrupt spleen-specific cCREs and the binding of several TFs. **(B)** In multiple tissues, RP11-362F19.1 has lower expression in individual 3 than in individual 2, suggesting that the deletion may affect the gene's expression globally.

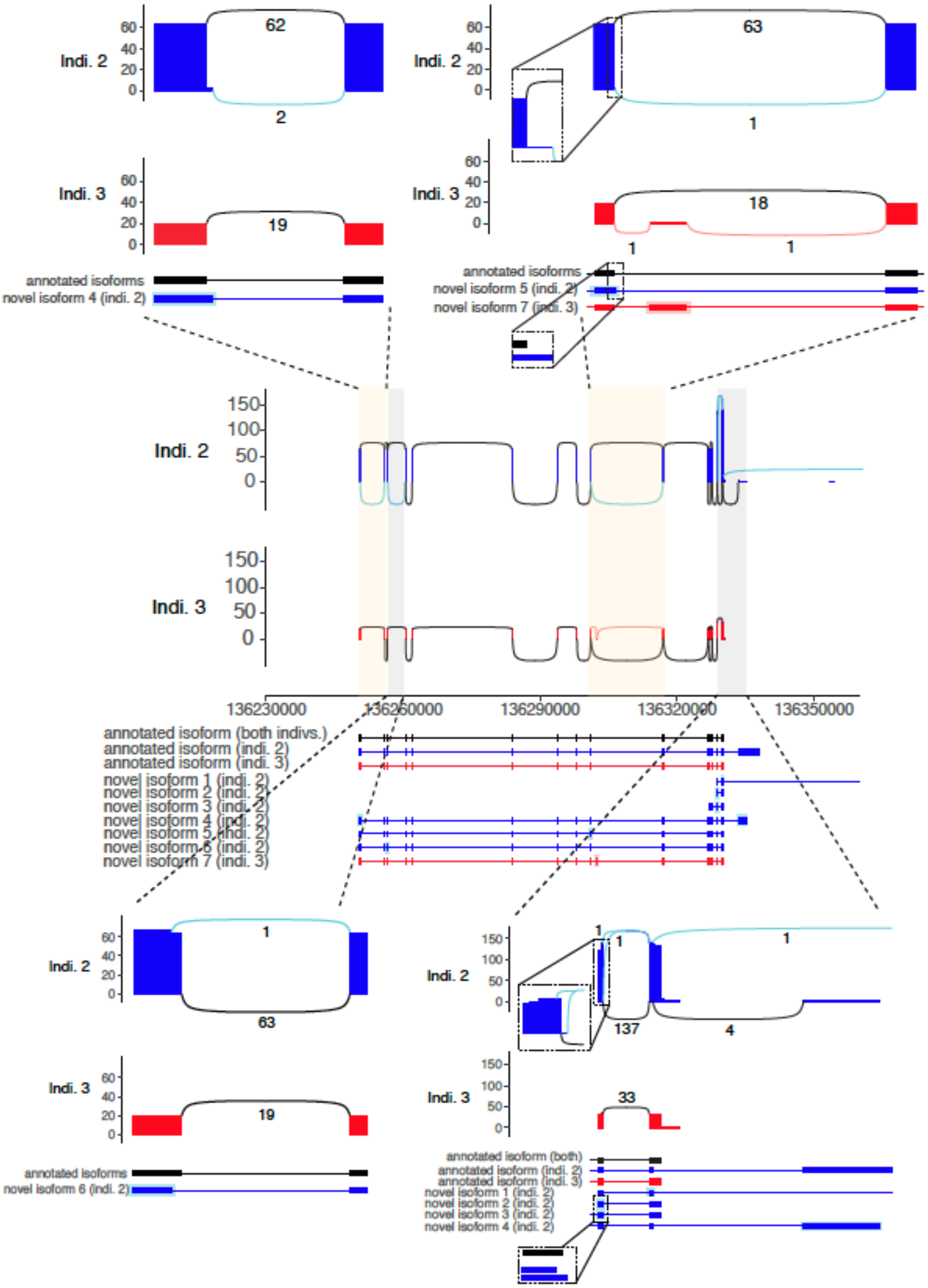


Figure S6.1e. Novel splicing variants of PCCB.

Shadow figure for Fig. 5E. Sashimi plot and exonic structure representation of the PCCB isoforms expressed in individuals 2 (blue) and 3 (red) in adrenal gland and heart left ventricle tissues, respectively. The central panel contains the whole gene's representation. In the sashimi plot, exons are represented by vertical lines either in blue (Ind. 2) or red (Ind. 3). Splicing connections of annotated isoforms are represented by black arcs, while novel connections observed in a specific individual are color-coded (zoom-ins into the specific regions are provided, as well as the number of reads supporting each connection). The exonic structure of annotated and novel isoforms is reported at the bottom. The black isoform is expressed in both individuals, while those expressed in only one individual are color-coded. Annotated and novel isoforms were retrieved, for each individual, using Swan \cite(*105*). Specifically, a swan gene report was generated for each individual by providing as input transcriptome annotation and quantification files available, from long-read RNA-seq experiments, on the ENCODE portal (https://www.encodeproject.org/). The plots were obtained using ggashimi \cite(*106*).

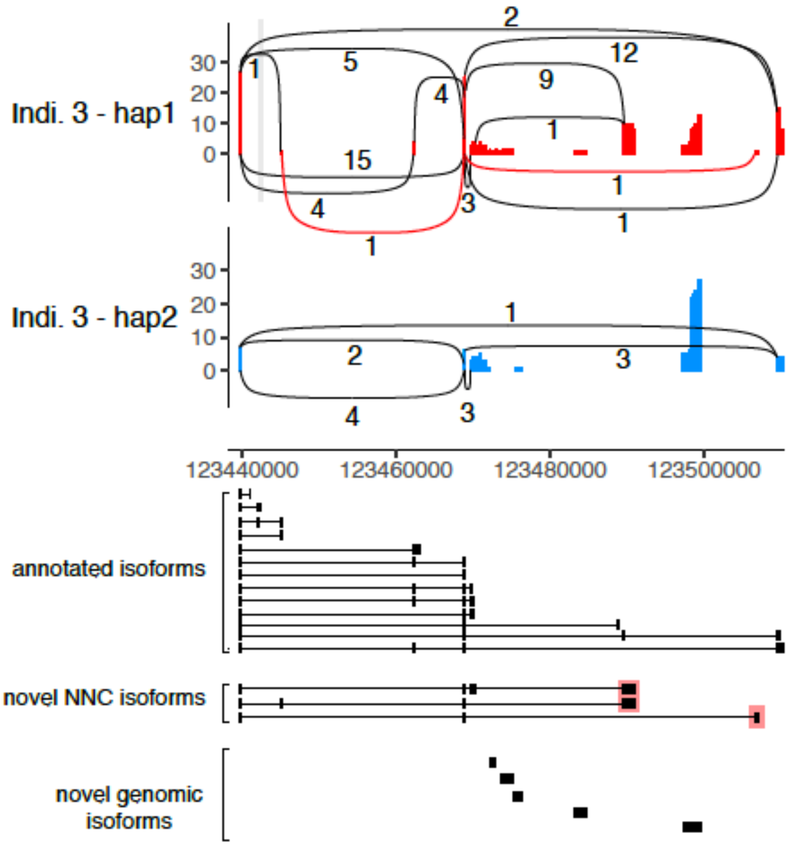


Figure S6.1f. Novel splicing variants of *TRDN-AS1*

Sashimi plot and exonic structure representation of the lncRNA TRDN-AS1 isoforms in individual 3 in the heart left ventricle. The gene carries a heterozygous deletion on haplotype 1 (highlighted in gray) and shows allelic-specific expression in the right atrium auricular region (hap1 being more expressed than hap2). For the sashimi representation, reads available from long RNA-seq experiments (see ENCODE portal) were phased to the two haplotypes using heterozygous SNVs that overlap the gene's exons. Reads phasing was performed with ASCIIGenome (https://github.com/dariober/ASCIIGenome/) \cite(*107*). Long-read RNA-seq reads show consistently higher expression of hap1 compared to hap2. Moreover, reads mapping to hap1 give rise to two novel splicing junctions (represented by red arcs) as well as two novel exons (highlighted in red in the exonic structure representation at the bottom). Annotated and novel isoforms were retrieved, for each individual, using Swan \cite(*105*). Specifically, a swan gene report was generated for individual 3 by providing as input transcriptome annotation and quantification files available, from long-read RNA-seq experiments, on the ENCODE portal. Only novel not in catalog (NNC) and genomic isoforms are shown. The plots were obtained using ggashimi \cite(*106*).

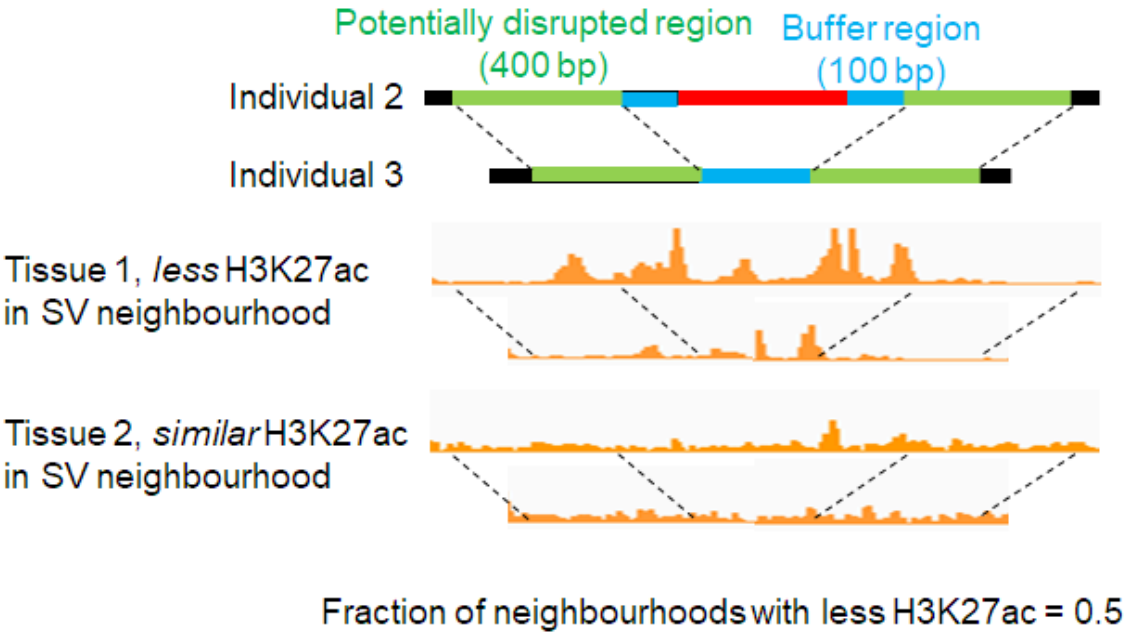

Potentially disrupted region
(400 bp)
Buffer region
(100 bp)
Individual 2
Individual 3
Tissue 1, *less* H3K27ac in SV neighbourhood
Tissue 2, *similar* H3K27ac in SV neighbourhood
Fraction of neighbourhoods with less H3K27ac = 0.5

Figure S6.2a. Calculate changes in the chromatin state in SV neighbourhood
We use H3K27ac level as an example. In this example, individual 3 carries a deletion (red bar) while individual 2 is wild-type at the same locus, therefore we will compare the chromatin states in the two green regions between the two individuals. In tissue 1, the H3K27ac level in the green region is lower in individual 3, but in tissue 2, the H3K27ac level is similar in both individuals. Therefore, only half of the neighbourhoods of this deletion show reduction in H3K27ac.

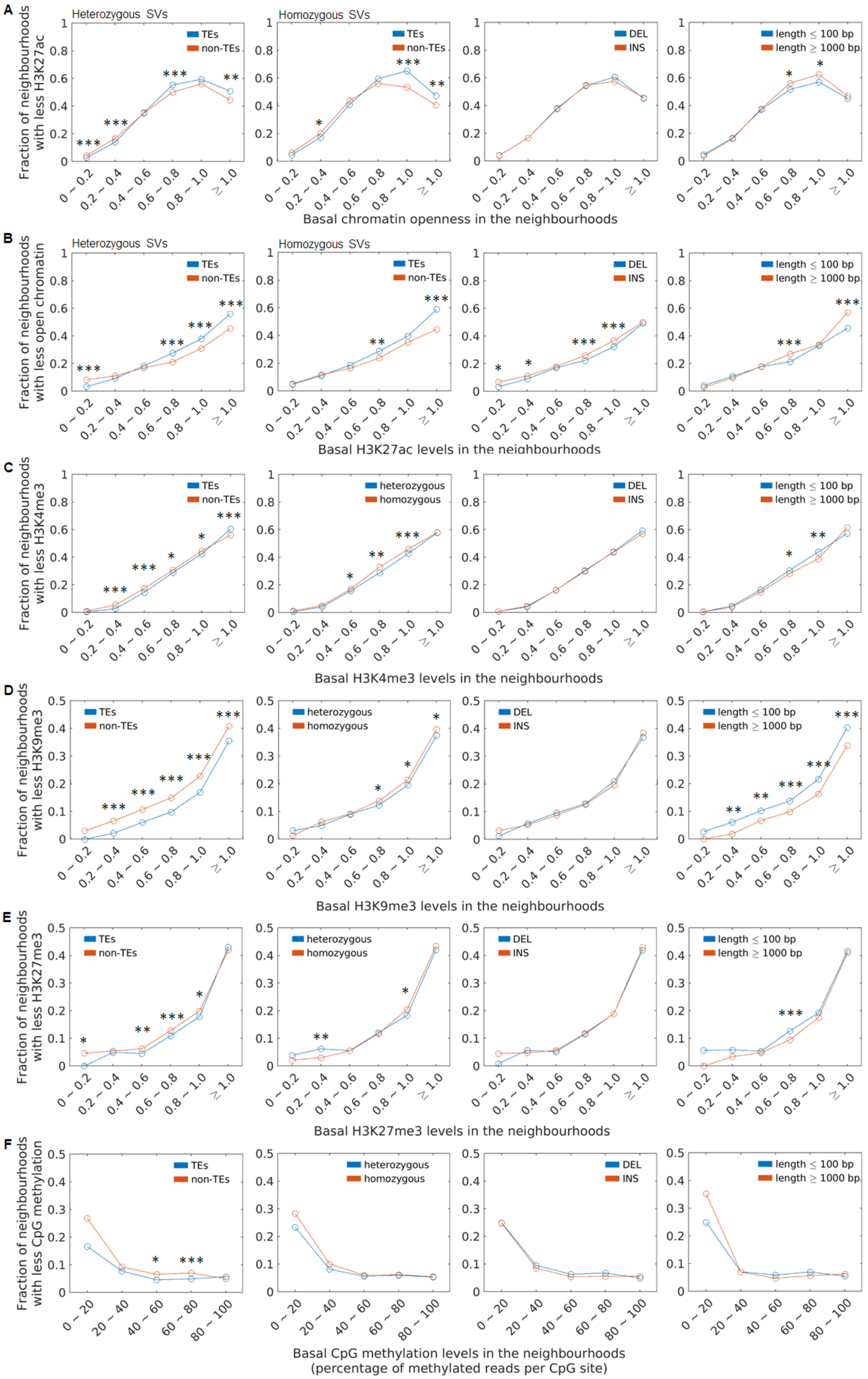

A
Heterozygous SVs
Homozygous SVs
TEs
non-TEs
DEL
INS
length ≤ 100 bp
length ≥ 1000 bp
Fraction of neighbourhoods with less H3K27ac
Basal chromatin openness in the neighbourhoods
B
Fraction of neighbourhoods with less open chromatin
Basal H3K27ac levels in the neighbourhoods
C
heterozygous
homozygous
Fraction of neighbourhoods with less H3K4me3
Basal H3K4me3 levels in the neighbourhoods
D
Fraction of neighbourhoods with less H3K9me3
Basal H3K9me3 levels in the neighbourhoods
E
Fraction of neighbourhoods with less H3K27me3
Basal H3K27me3 levels in the neighbourhoods
F
Fraction of neighbourhoods with less CpG methylation
Basal CpG methylation levels in the neighbourhoods (percentage of methylated reads per CpG site)
0 ~ 0.2
0.2 ~ 0.4
0.4 ~ 0.6
0.6 ~ 0.8
0.8 ~ 1.0
≥ 1.0
0 ~ 20
20 ~ 40
40 ~ 60
60 ~ 80
80 ~ 100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1
*
**
***

Figure S6.2b. Changes in the chromatin state of SV neighbourhoods
Similar to Fig. 5D, we investigate whether the presence of an SV may change the chromatin state of the nearby regions, and whether the changes are associated with the SV's length, genotype, type (INS or DEL), and/or association with TEs. Panel **(A)** - **(F)** correspond to H3K27ac, chromatin openness measured by ATAC-seq, H3K4me3, H3K9me3, H3k27me3, and CpG methylation, respectively. * p < 0.05, ** p < 0.01, *** p < 0.001, based on Chi-square test.

Figure S7. Supp. figures to main text section “Decorating the ENCODE encyclopedia”

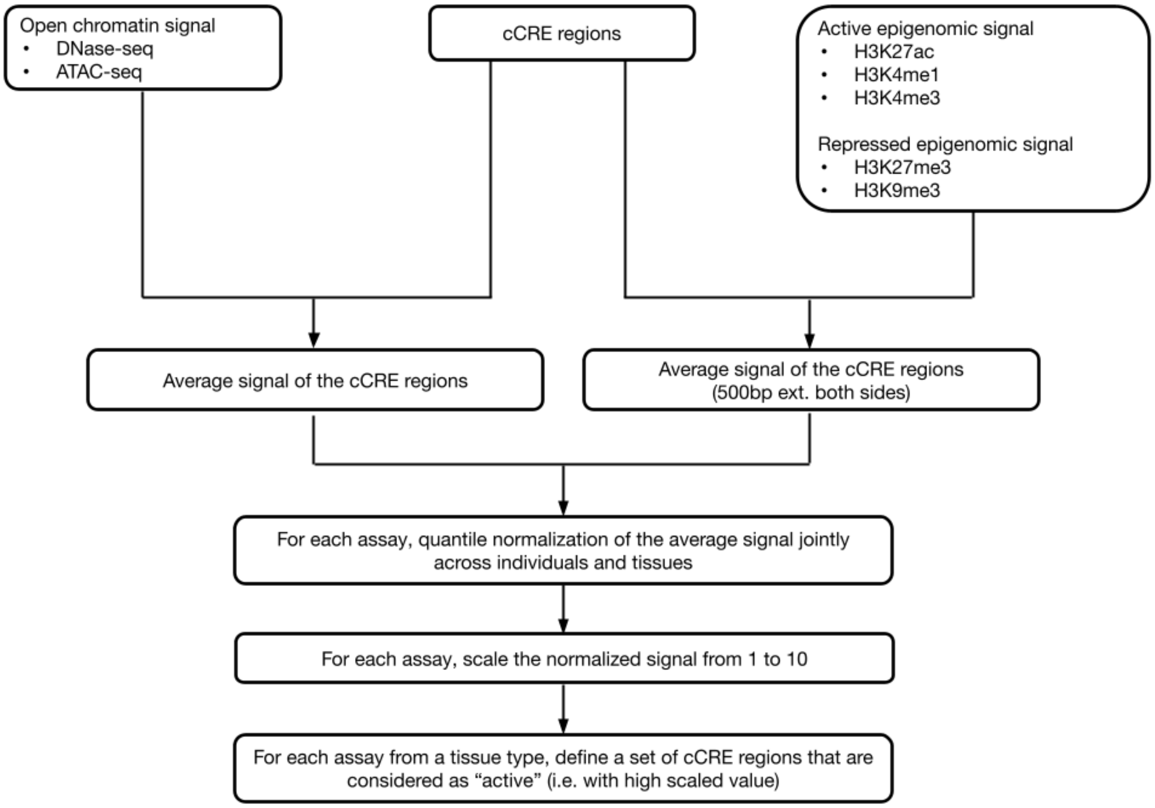

Open chromatin signal
• DNase-seq
• ATAC-seq
cCRE regions
Active epigenomic signal
• H3K27ac
• H3K4me1
• H3K4me3
Repressed epigenomic signal
• H3K27me3
• H3K9me3
Average signal of the cCRE regions
Average signal of the cCRE regions (500bp ext. both sides)
For each assay, quantile normalization of the average signal jointly across individuals and tissues
For each assay, scale the normalized signal from 1 to 10
For each assay from a tissue type, define a set of cCRE regions that are considered as "active" (i.e. with high scaled value)

Figure S7.2a. Data Preprocessing
We compute the average signal for each cCRE region using the datasets from DNase-seq, ATAC-seq, and 5 histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K27me3 and H3K9me3). For DNase-seq and ATAC-seq, the signals are averaged across the genomic positions of the cCRE regions. The signals of histone modifications are averaged across the genomic positions of the cCRE regions with a 500-bp extended region on each side. For each assay, we perform quantile normalization on the average signal from the cCRE regions jointly across all the biosamples. Then we scale the normalized signal from 1 to 10, and define a set of "active" cCREs for each assay from each tissue type.

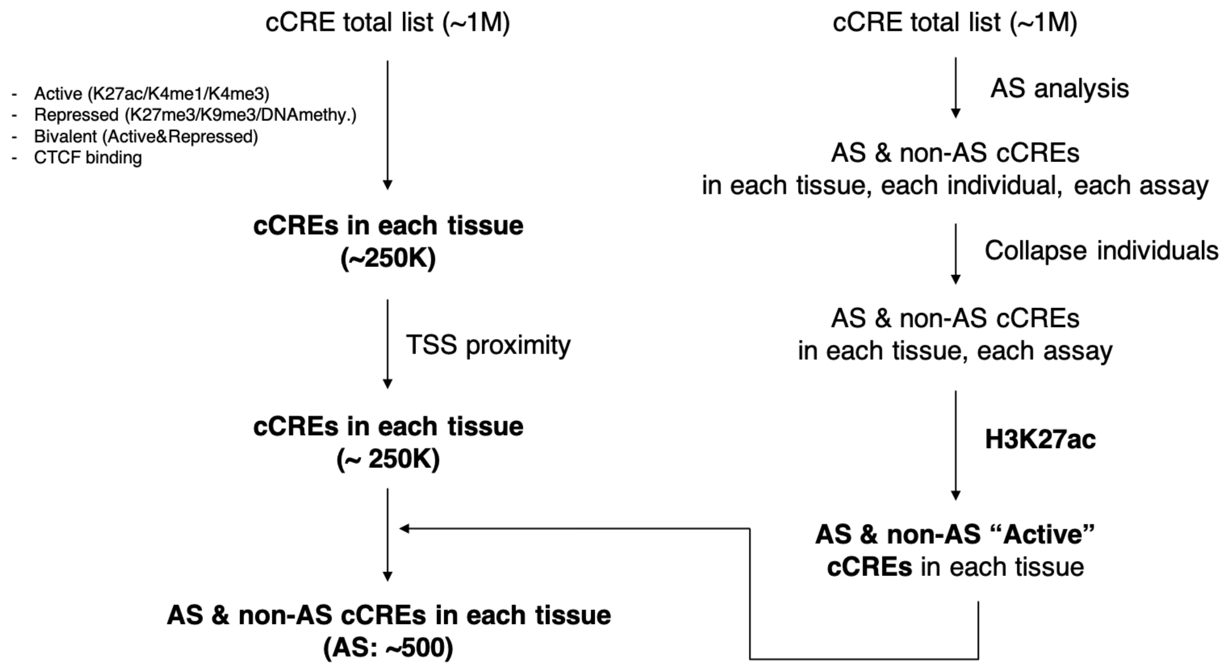

cCRE total list (~1M)
- Active (K27ac/K4me1/K4me3)
- Repressed (K27me3/K9me3/DNAmethy.)
- Bivalent (Active&Repressed)
- CTCF binding
cCREs in each tissue (~250K)
TSS proximity
cCREs in each tissue (~ 250K)
AS & non-AS cCREs in each tissue (AS: ~500)
cCRE total list (~1M)
AS analysis
AS & non-AS cCREs in each tissue, each individual, each assay
Collapse individuals
AS & non-AS cCREs in each tissue, each assay
H3K27ac
AS & non-AS "Active" cCREs in each tissue

Figure S7.2b. Framework of cCRE decoration

We decorate the cCREs from encyclopedia using the active and repressed histone modification signals and CTCF binding sites from tissues. The decorated the cCREs are then separated into proximal and distal ones based on their proximity to the annotated TSSs. At another layer, these cCRE subgroups are further annotated as allelic-specific and non-allelic-specific ones based on their allelic signature.

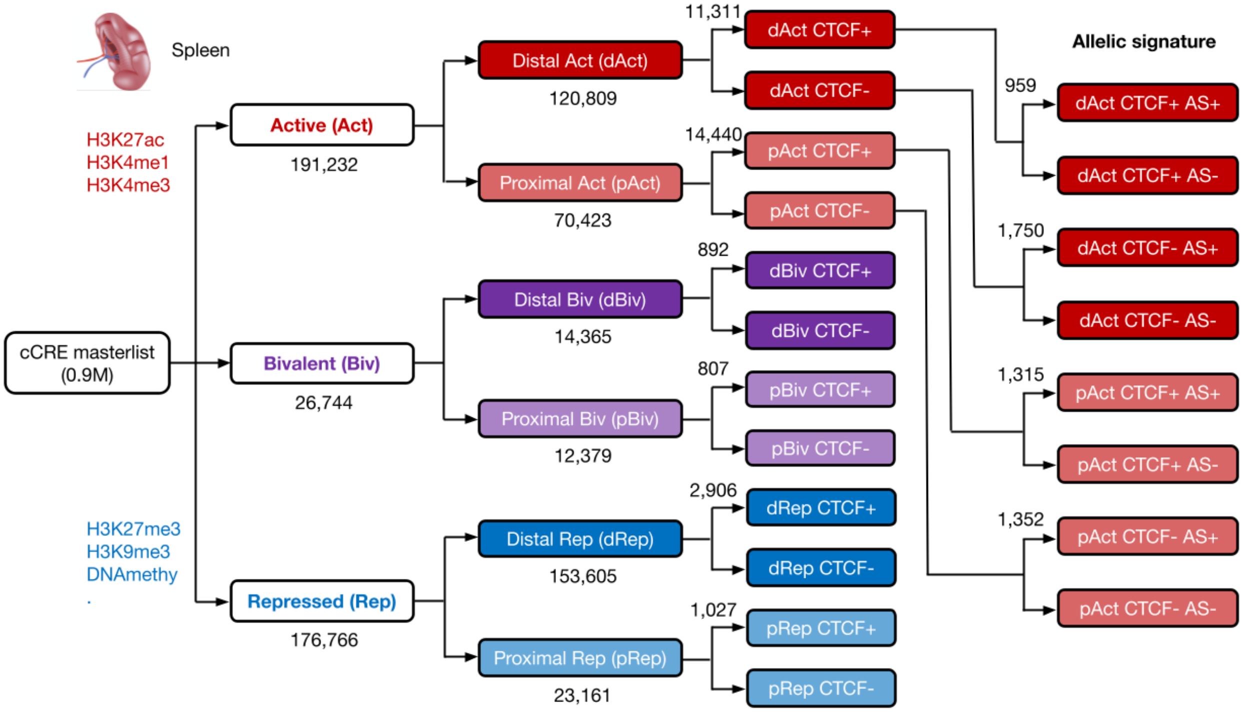


Figure S7.2c. Framework of cCRE decoration in spleen

This figure shows the workflow of cCRE decoration and the numbers of different subgroups of cCREs in the spleen. Note that we define a number of abbreviations for the various decorations. dACT: distal active; pACT: proximal active; dBiv: distal bivalent; pBiv: proximal bivalent; dRep: distal repressed; pRep: proximal repressed; CTCT+ and CTCF- indicates with and without CTCF binding respectively; AS+ and AS- indicates with and without allelic signature respectively.

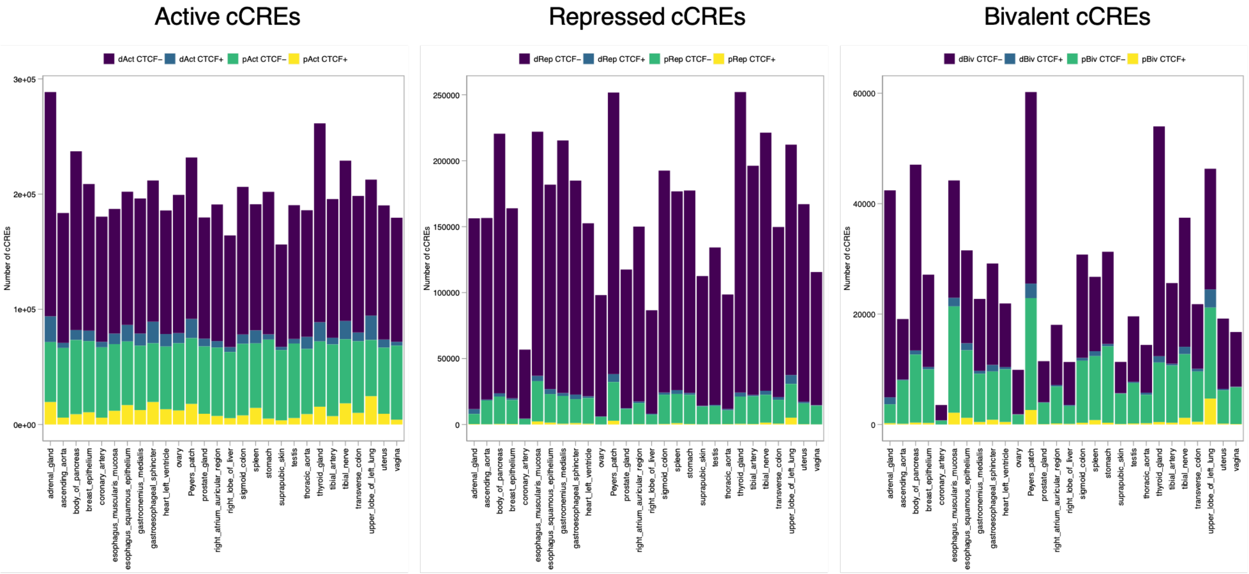
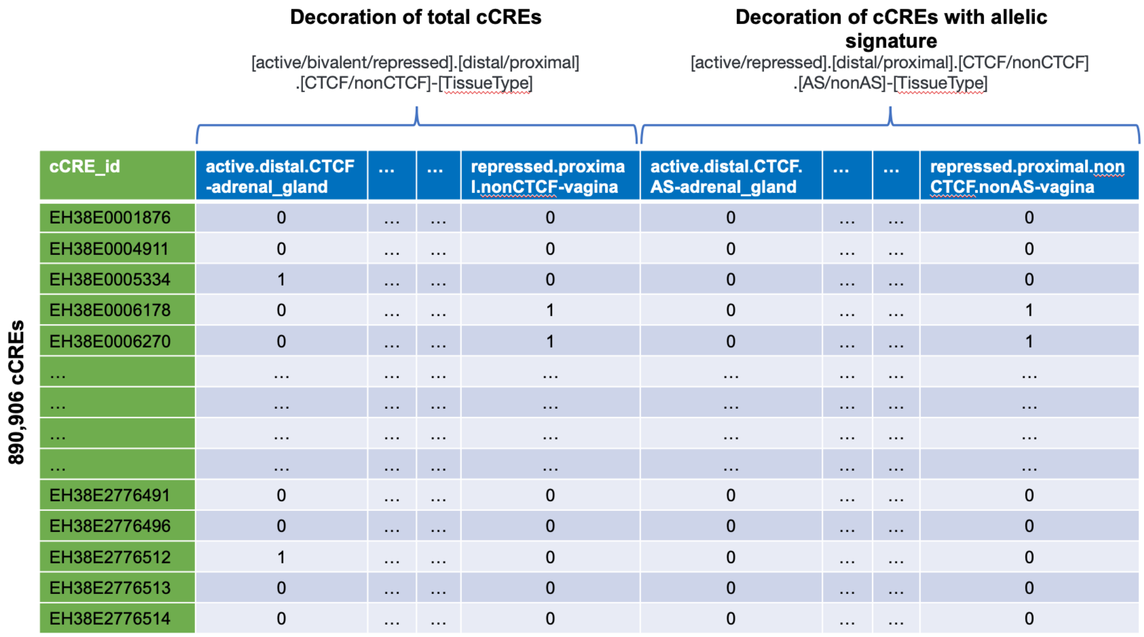


Figure S7.2d. Number of cCREs in Various Tissues
We show the number of different subgroups of decorated cCREs in each tissue type. In each panel, the colors indicate their TSS proximity (proximal vs. distal) and CTCF binding states (CTCF+ vs. CTCF-). Note that the different decoration terms are defined in Figure S7.2c.



Decoration of total cCREs
[active/bivalent/repressed].[distal/proximal].[CTCF/nonCTCF]-[TissueType]
Decoration of cCREs with allelic signature
[active/repressed].[distal/proximal].[CTCF/nonCTCF].[AS/nonAS]-[TissueType]
890,906 cCREs

cCRE_id	active.distal.CTCF-adrenal_gland	…	…	repressed.proximal.nonCTCF-vagina	active.distal.CTCF.AS-adrenal_gland	…	…	repressed.proximal.nonCTCF.nonAS-vagina
EH38E0001876	0	…	…	0	0	…	…	0
EH38E0004911	0	…	…	0	0	…	…	0
EH38E0005334	1	…	…	0	0	…	…	0
EH38E0006178	0	…	…	1	0	…	…	1
EH38E0006270	0	…	…	1	0	…	…	1
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
EH38E2776491	0	…	…	0	0	…	…	0
EH38E2776496	0	…	…	0	0	…	…	0
EH38E2776512	1	…	…	0	0	…	…	0
EH38E2776513	0	…	…	0	0	…	…	0
EH38E2776514	0	…	…	0	0	…	…	0

Figure S7.2e. cCRE Decoration Results Matrix

We generate the annotation matrix for all the decorated cCREs from tissue types. We used 1 and 0 to indicate that the cCREs are defined as “active” in terms of that cCRE nomenclature as shown. This corresponds to File: cCRE_decoration.matrix.

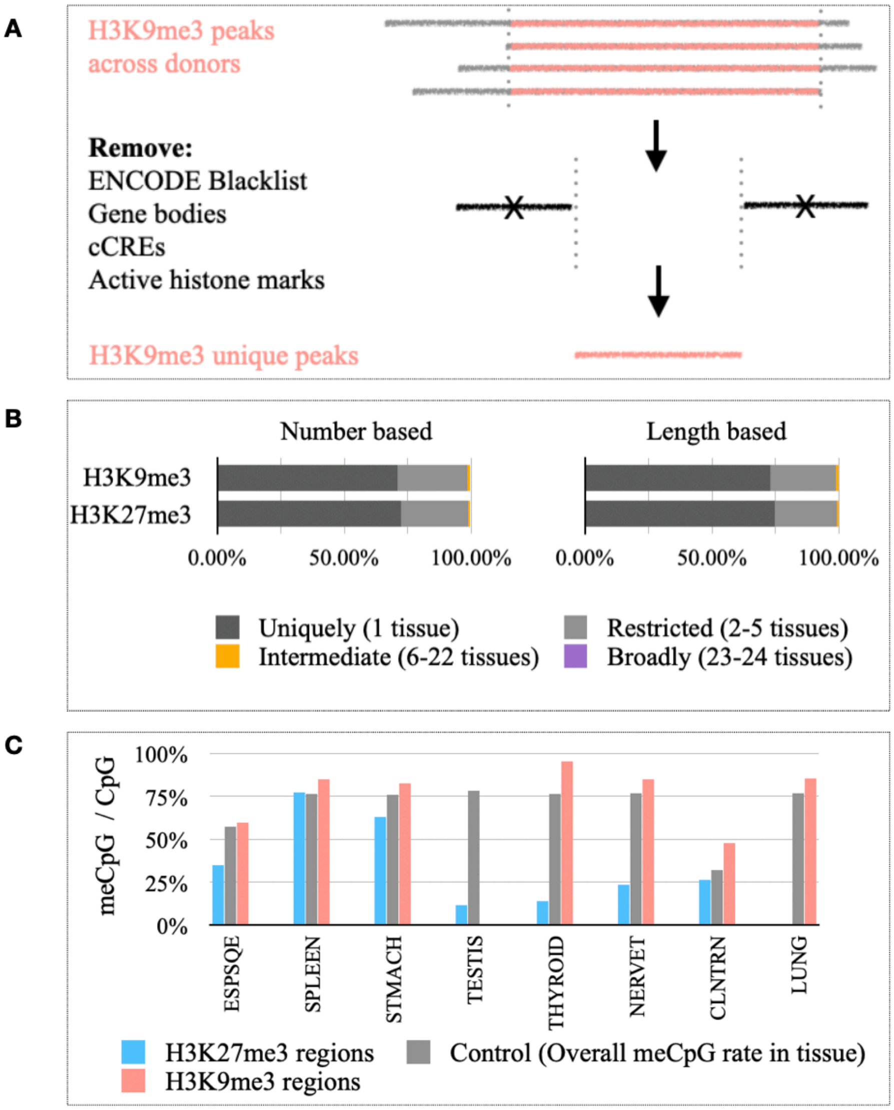


Figure S7.3. Identify fully repressed elements independent of cCREs

(A) For genomic regions outside cCREs and annotated genes, elements longer than 200bp that are uniquely marked by either H3K9me3 or H3K27me3 are defined as fully repressed. 45,207 (covering 12,655,795 bp) and 24,006 (covering 7,474,178 bp) non-overlapping elements are identified based on H3K9me3 and H3K27me3, respectively. Identified elements can be found in File: ENTEx_fully_repressed_regions_independent_of_cCREs.bed. **(B)** The majority of these elements are repressed in a tissue-specific manner. **(C)** For tissues with available datasets, DNA methylation within these elements was evaluated, and H3K9me3 marked elements show significantly (t-test, pValue<0.05) higher CpG methylation (meCpG) rate than elements marked uniquely by H3K27me3.

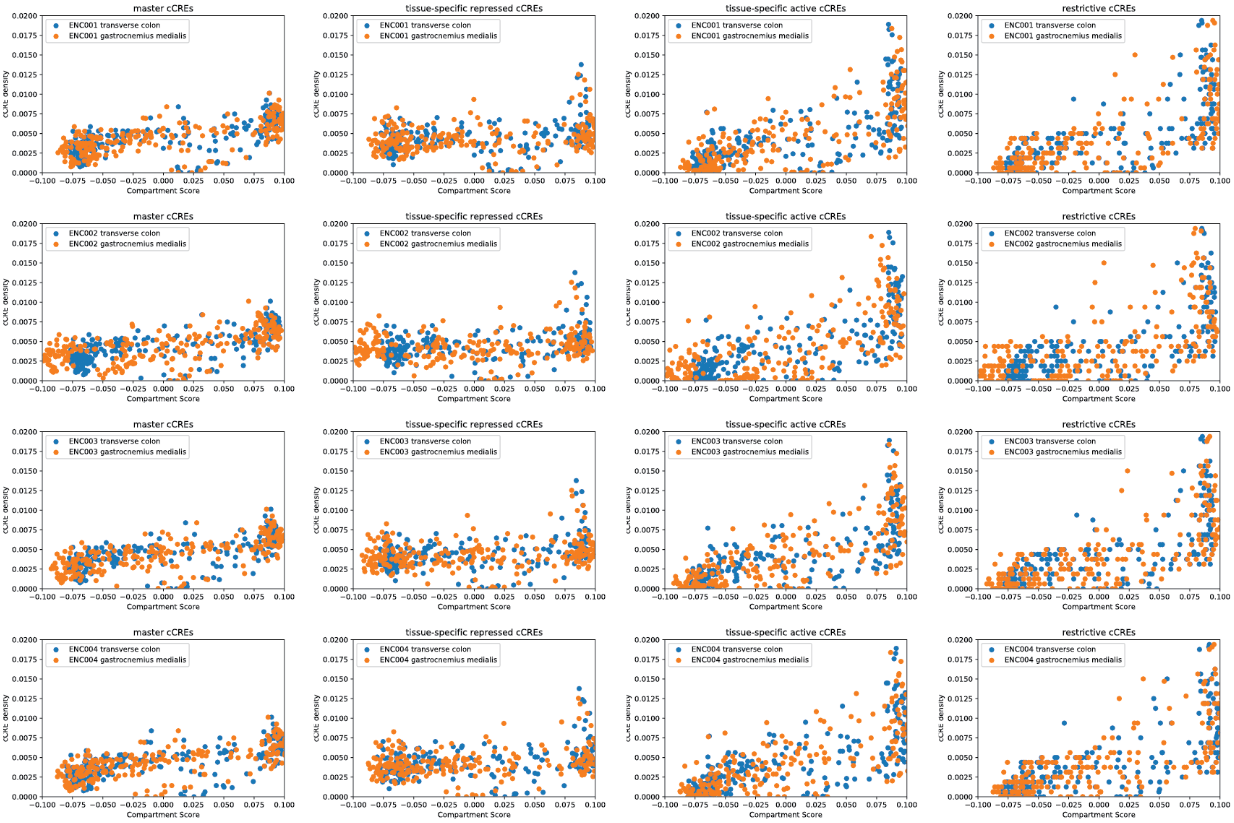


Figure S7.4. cCRE enrichment with respect to A/B compartments

These plots show the cCRE enrichment in the A vs. B compartment of two different tissues. We show this for the master cCRE list from ENCODE encyclopedia, tissue-specific active and repressed cCREs. As the tissue specificity increases, we see more cCRE enrichment in the active A compartment compared to the inactive B compartment.

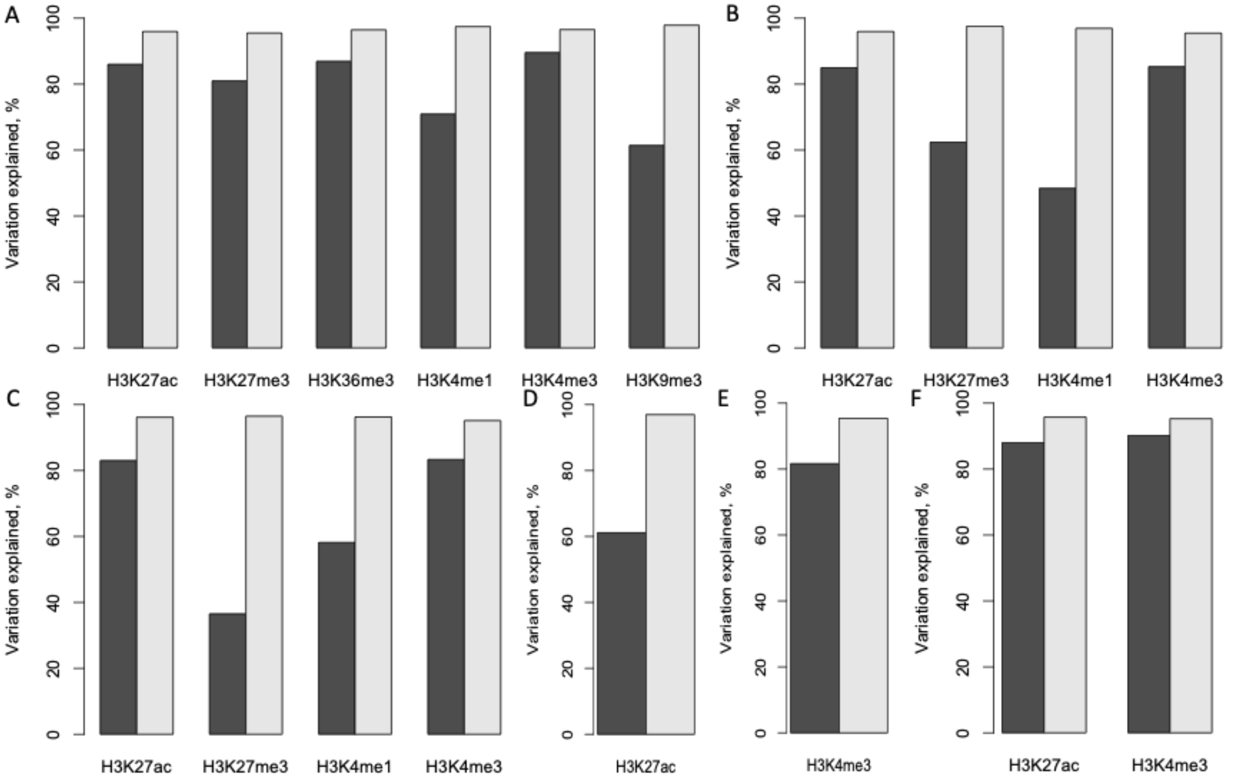
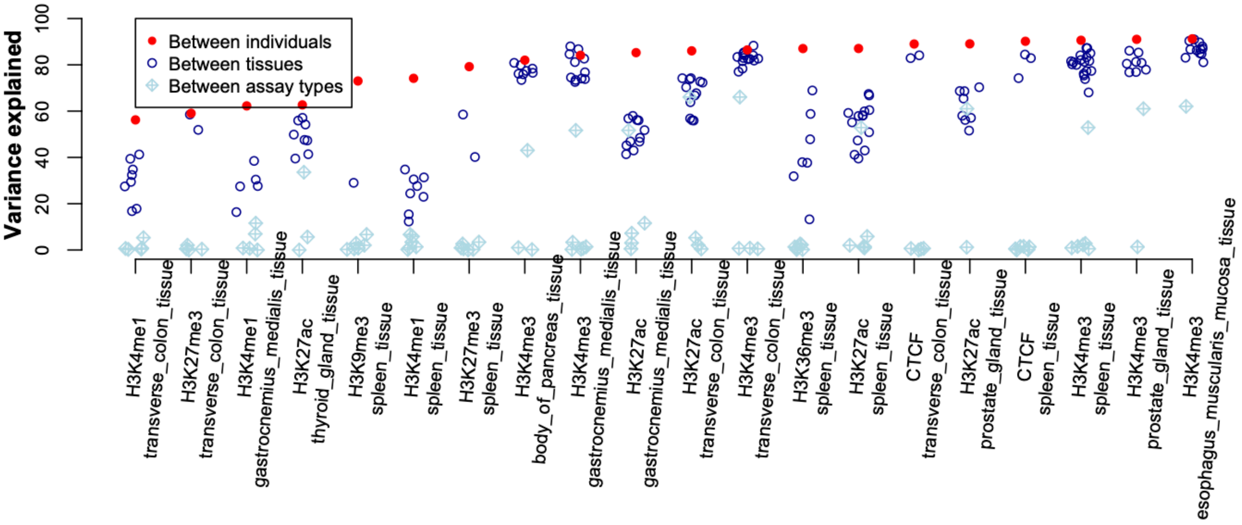
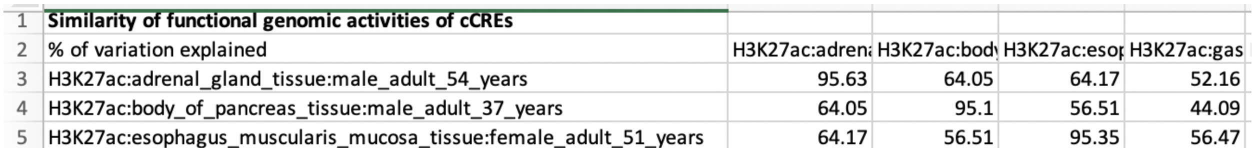


Figure S7.5a. Variation explained between two experiments corrected by replicates
To calculate the variation explained between experiments (e.g., the two H3K27ac ChIP-seq experiments of the spleens from two individuals), for each of the experiment, we identify the cCREs that renders the replicates of the experiment having high variation explained (> 95%). The intersect set of such consistent cCREs respectively from the two experiments are used to calculate the variation explained between the two experiments (black bars; e.g., 87% for the two H3K27ac). The average variation explained between the replicates respectively from the two experiments is indicated by the white bars (e.g., 96% for H3K27ac). The results in spleen, transverse colon, gastrocnemius medialis, thyroid gland, pancreas, and prostate gland are shown in **(A)** - **(F)**.





1	**Similarity of functional genomic activities of cCREs**				
2	% of variation explained	H3K27ac:adrena	H3K27ac:body	H3K27ac:esop	H3K27ac:gas
3	H3K27ac:adrenal_gland_tissue:male_adult_54_years	95.63	64.05	64.17	52.16
4	H3K27ac:body_of_pancreas_tissue:male_adult_37_years	64.05	95.1	56.51	44.09
5	H3K27ac:esophagus_muscularis_mucosa_tissue:female_adult_51_years	64.17	56.51	95.35	56.47

Figure S7.5b. Similarity between the signals of two functional genomic experiments
For each of the cCREs, the signal of a functional genomic experiment is measured by the average fold change over control across the cCRE region. For two experiments, linear regression is used for the cCREs with low technical noise between replicates. The variance of one experiment explained by the other is used to indicate the similarity between the experiments across the cCREs. The similarity between all possible pairs of experiments is reported in the supplementary table, namely Similarity of functional genomic activities of cCREs.

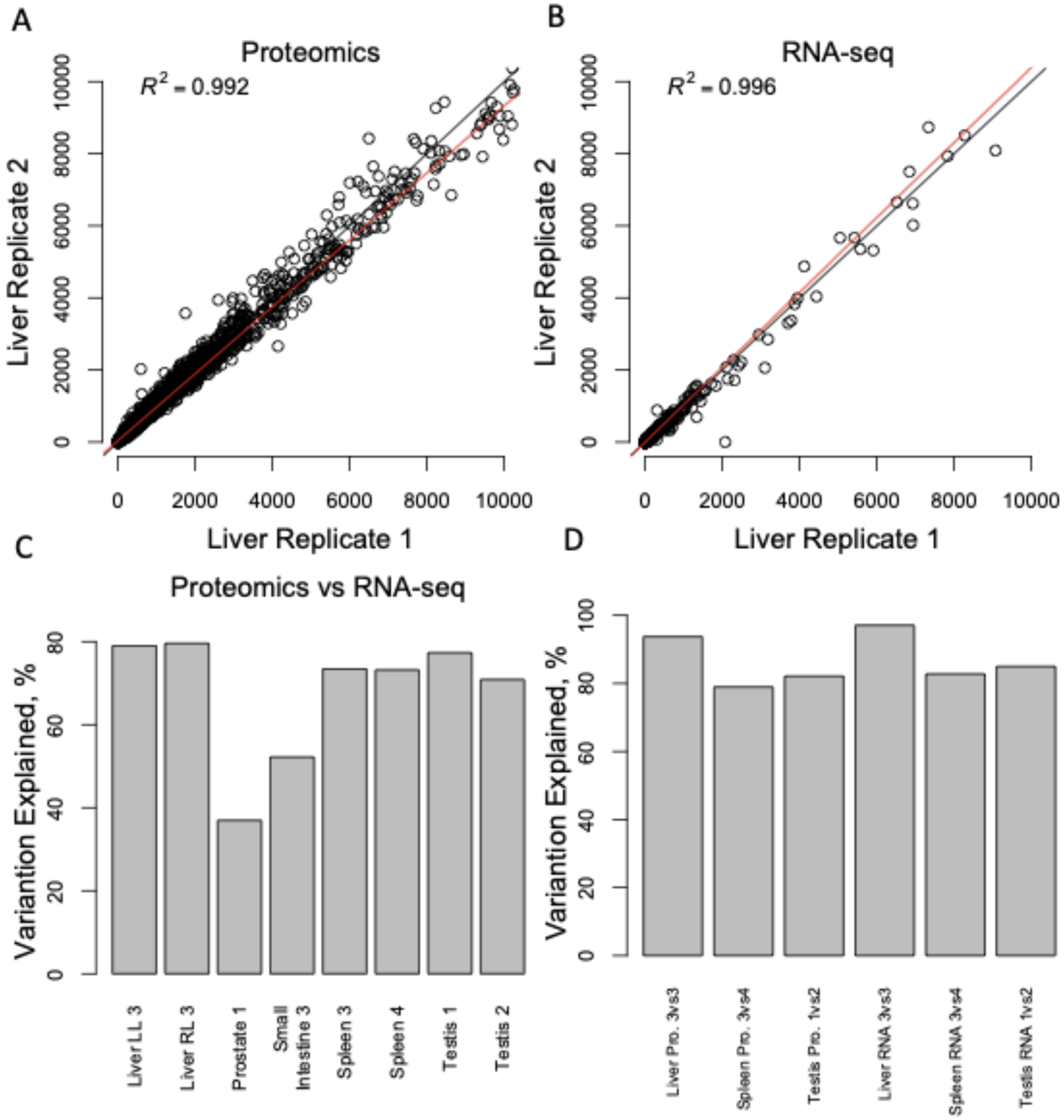


Figure S7.5c. Variation explained between proteomics data and RNA-seq data
(A) The normalized protein abundances are highly consistent between replicates. **(B)** This is also true for the normalized RNA abundances. **(C)** The variation explained between the normalized protein abundances and the normalized RNA abundances varies across tissues. suggesting that for some tissues protein abundances and RNA abundances have low consistency. LL indicates the left lobe of the liver, and the RL indicates the right lobe. The numbers in the labels indicate the donors. However, respectively for protein abundances and RNA abundances, the variation explained between donors is higher **(D)** than the variation explained between the normalized protein abundances and RNA abundances **(C)**. The normalized proteomics and RNA-seq data matrix used for panels **(C)** & **(D)** are normalized_proteomics_RNA-seq.dat.

Figure S8. Supp. figures to main text section “Measuring tissue specificity”

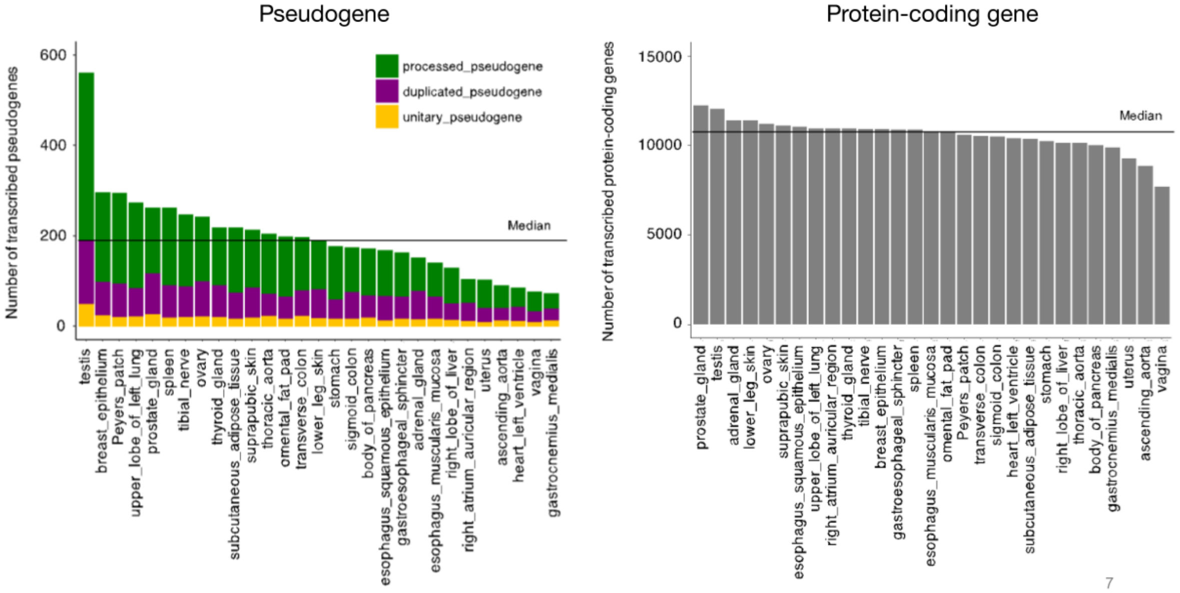


Figure S8.1a. The number of transcribed genes in tissues

We show the number of transcribed pseudogenes (left) and protein-coding genes (right) across all tissue types. The median of transcribed pseudogenes and protein-coding genes across the tissues is 200 and ~11K, respectively.

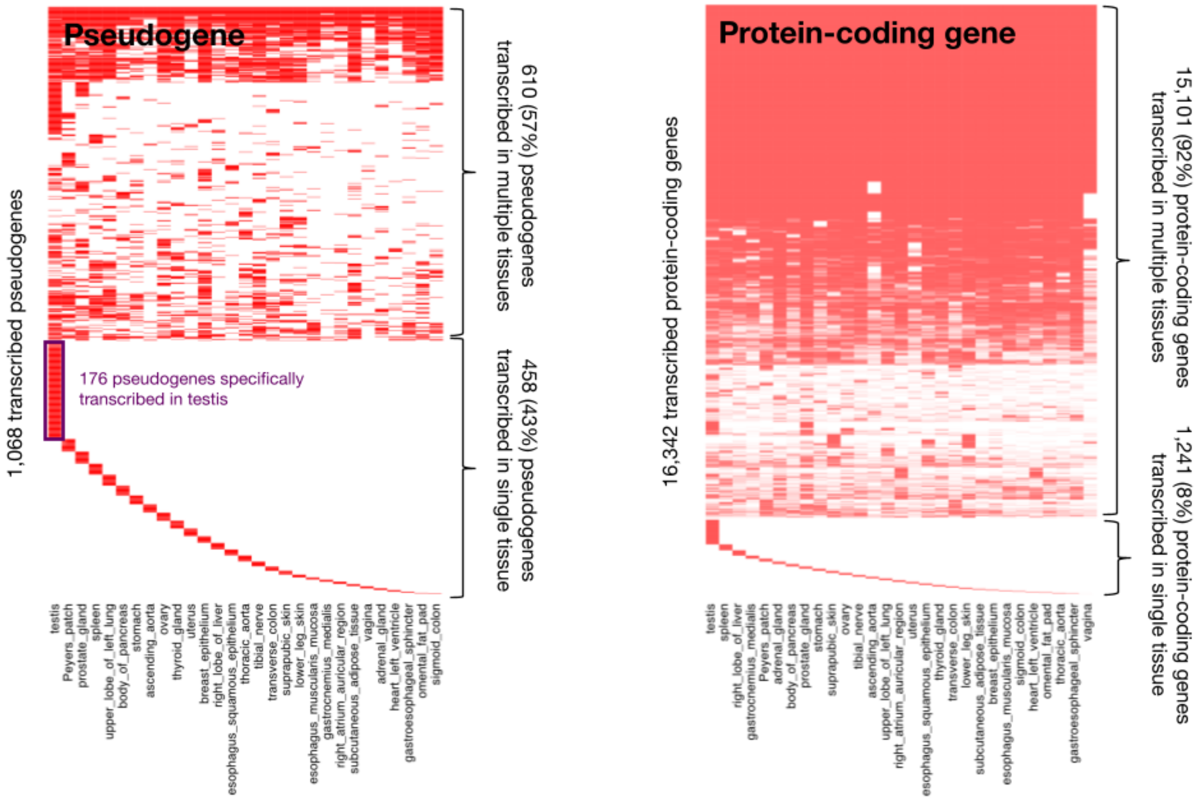


Figure S8.1b. Tissue-specificity of transcribed genes
The heatmaps show the activity of pseudogenes (left) and protein-coding genes (right) across tissue types. In each tissue, the pseudogenes/protein-coding genes are classified as actively transcribed (shown in red) or not based on their expression level.

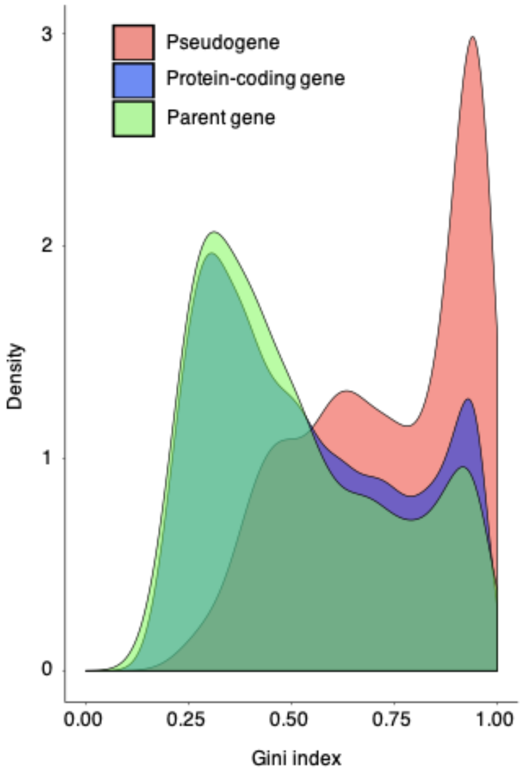


Figure S8.1c. Gini Index of Gene Expression Level Across Tissues

We apply the Gini index to quantify the tissue-specificity of protein-coding genes, pseudogenes and parent genes based on their expression level. The pseudogenes show higher Gini index than protein-coding genes, suggesting stronger tissue-specificity of pseudogenes. The Gini index distribution of pseudogenes is quite different from that from parent genes, confirming that the multi-mapping bias from quantification of the pseudogene expression level has been minimized.

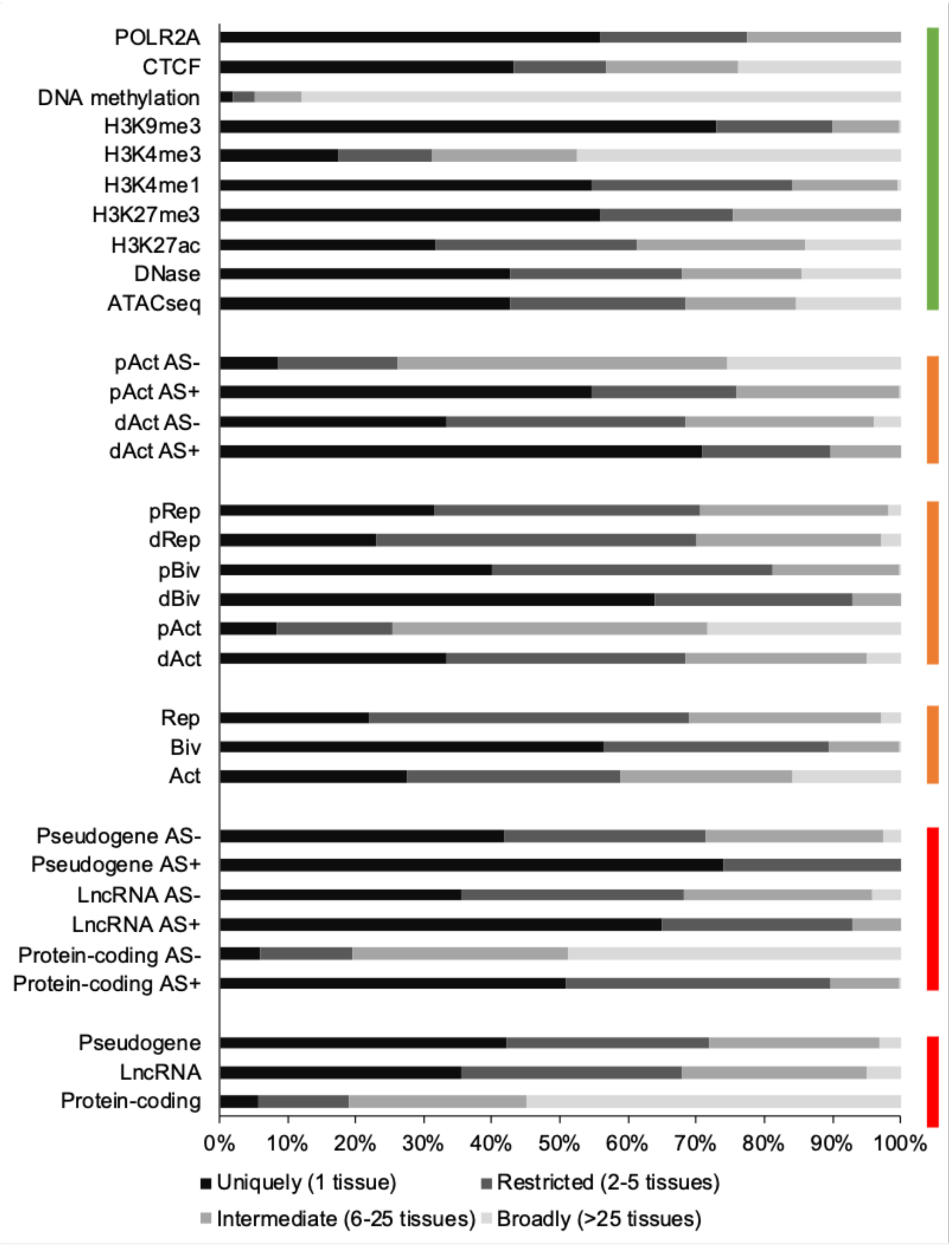


Figure S8.1d. Tissue-specificity of different subgroups of cCREs vs. genes and epigenomic peaks

We compare the tissue-specificity of protein-coding genes, non-coding genes, different subgroups of decorated cCREs and various epigenomic peaks. The uniqueness of their activity across tissue types are shown in different colors. Note that the different decoration terms are defined in Figure S7.2c.

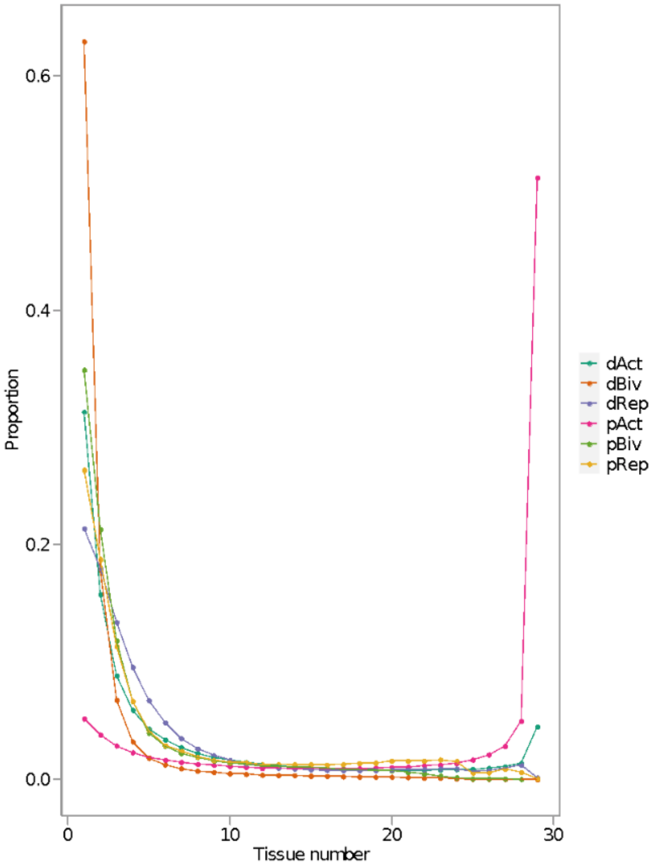


Figure S8.1e. Tissue-specificity of different subgroups of cCREs
For each cCRE subgroup, we show the proportion of the cCREs that are defined as “active” across the different numbers of tissue types ranging from one (i.e., high tissue-specificity) to all tissue types (i.e., low tissue-specificity). Note that the different decoration terms are defined in Fig. S7.2c.

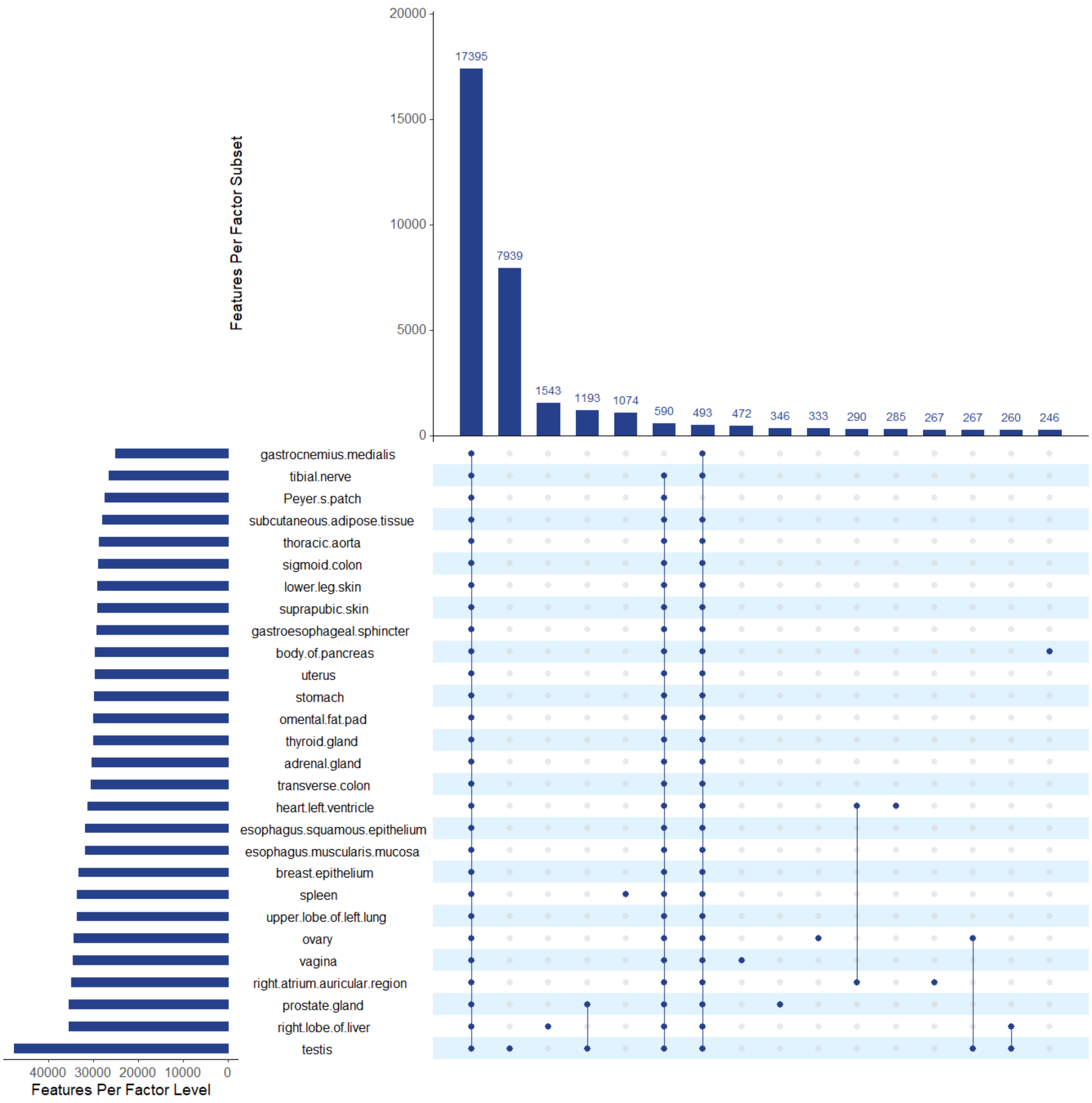


Figure S8.1f. Tissue-specificity of RAMPAGE Data at TSSs of Protein-Coding Genes
UpSet plot of counts of GENCODE TSSs of genes (vertical bars), measured using RAMPAGE data in combinations of tissues (sets of dots), sorted by the number of TSSs. Bars on the left correspond to the number of TSSs in each tissue. Ubiquitously expressed TSSs using RAMPAGE are the most abundant.

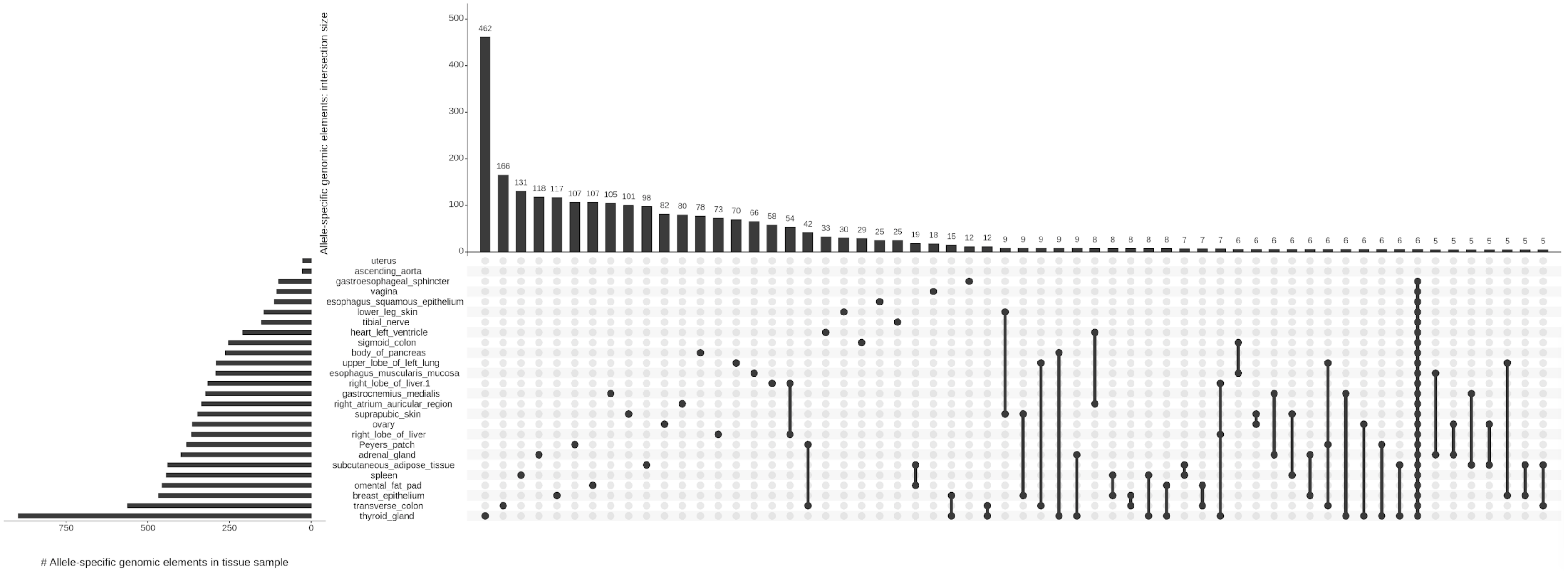

Allele-specific genomic elements: intersection size
0
100
200
300
400
500
462
166
131
118
117
107
107
105
101
98
82
80
78
73
70
66
58
54
42
33
30
29
25
25
19
18
15
12
12
9
9
9
9
9
8
8
8
8
8
7
7
7
6
6
6
6
6
6
6
6
6
6
6
6
5
5
5
5
5
5
5
uterus
ascending_aorta
gastroesophageal_sphincter
vagina
esophagus_squamous_epithelium
lower_leg_skin
tibial_nerve
heart_left_ventricle
sigmoid_colon
body_of_pancreas
upper_lobe_of_left_lung
esophagus_muscularis_mucosa
right_lobe_of_liver.1
gastrocnemius_medialis
right_atrium_auricular_region
suprapubic_skin
ovary
right_lobe_of_liver
Peyers_patch
adrenal_gland
subcutaneous_adipose_tissue
spleen
omental_fat_pad
breast_epithelium
transverse_colon
thyroid_gland
750
500
250
0
# Allele-specific genomic elements in tissue sample

Figure S8.2a. ASE genes across different tissues of individual 3.
Counts of genes (bars) called ASE in the combinations of tissues (sets of dots) with the largest number of ASE genes. Bars on the left correspond to the number of ASE genes in each tissue.

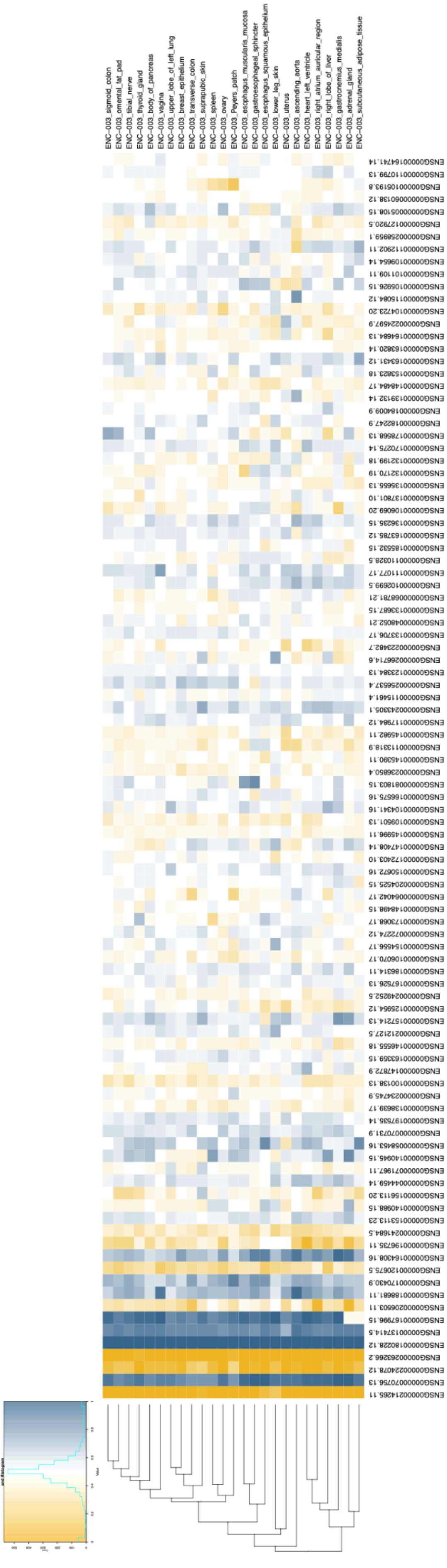
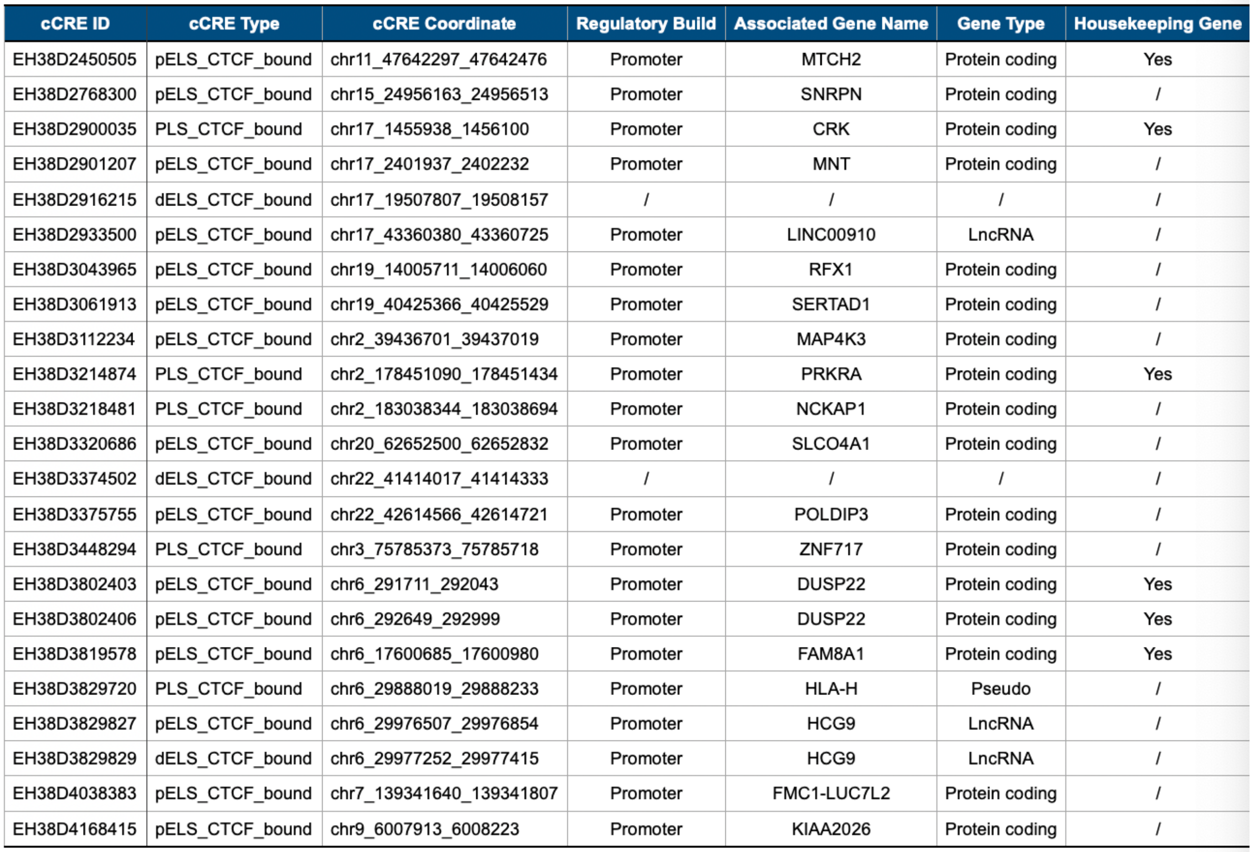

ENC–003_sigmoid_colon
ENC–003_omental_fat_pad
ENC–003_tibial_nerve
ENC–003_thyroid_gland
ENC–003_body_of_pancreas
ENC–003_vagina
ENC–003_upper_lobe_of_left_lung
ENC–003_breast_epithelium
ENC–003_transverse_colon
ENC–003_suprapubic_skin
ENC–003_spleen
ENC–003_ovary
ENC–003_Peyers_patch
ENC–003_esophagus_muscularis_mucosa
ENC–003_gastroesophageal_sphincter
ENC–003_esophagus_squamous_epithelium
ENC–003_lower_leg_skin
ENC–003_uterus
ENC–003_ascending_aorta
ENC–003_heart_left_ventricle
ENC–003_right_atrium_auricular_region
ENC–003_right_lobe_of_liver
ENC–003_gastrocnemius_medialis
ENC–003_adrenal_gland
ENC–003_subcutaneous_adipose_tissue
ENSG00000164741.14
ENSG00000110799.13
ENSG00000105193.8
ENSG00000060138.12
ENSG00000005108.15
ENSG00000127920.5
ENSG00000258959.1
ENSG00000112902.11
ENSG00000109654.14
ENSG00000101109.11
ENSG00000105926.15
ENSG00000115084.12
ENSG00000104723.20
ENSG00000224597.9
ENSG00000164684.13
ENSG00000163820.14
ENSG00000163431.12
ENSG00000153823.18
ENSG00000148484.17
ENSG00000139132.14
ENSG00000184009.9
ENSG00000182247.9
ENSG00000178568.13
ENSG00000170275.14
ENSG00000132199.18
ENSG00000132170.19
ENSG00000135655.13
ENSG00000137801.10
ENSG00000106069.20
ENSG00000136235.15
ENSG00000163785.12
ENSG00000185532.15
ENSG00000110328.5
ENSG00000111077.17
ENSG00000102699.5
ENSG00000068781.21
ENSG00000133687.15
ENSG00000048052.21
ENSG00000133706.17
ENSG00000223482.7
ENSG00000266714.6
ENSG00000123384.13
ENSG00000256537.4
ENSG00000115461.4
ENSG00000243305.1
ENSG00000117984.12
ENSG00000145982.11
ENSG00000113318.9
ENSG00000145390.11
ENSG00000236850.4
ENSG00000081803.15
ENSG00000166575.16
ENSG00000104341.16
ENSG00000109501.13
ENSG00000145996.11
ENSG00000147408.14
ENSG00000172403.10
ENSG00000150672.16
ENSG00000204525.15
ENSG00000064042.17
ENSG00000148498.15
ENSG00000173068.17
ENSG00000072274.12
ENSG00000154556.17
ENSG00000106070.17
ENSG00000186314.11
ENSG00000167526.13
ENSG00000249252.5
ENSG00000125954.12
ENSG00000157214.13
ENSG00000212127.5
ENSG00000146555.18
ENSG00000163359.15
ENSG00000147872.9
ENSG00000100138.13
ENSG00000234745.9
ENSG00000138639.17
ENSG00000197535.14
ENSG00000070731.9
ENSG00000058453.16
ENSG00000140945.15
ENSG00000071967.11
ENSG00000044459.14
ENSG00000156113.20
ENSG00000140988.15
ENSG00000153113.23
ENSG00000241684.5
ENSG00000196735.11
ENSG00000164308.16
ENSG00000120675.5
ENSG00000170430.9
ENSG00000188681.11
ENSG00000206503.11
ENSG00000167996.15
ENSG00000137414.5
ENSG00000180228.12
ENSG00000263266.2
ENSG00000224078.12
ENSG00000070756.13
ENSG00000214265.11
Color Key
and Histogram
Value
0
0.2
0.4
0.6
0.8
1
Count
0
100
200
300
400
500

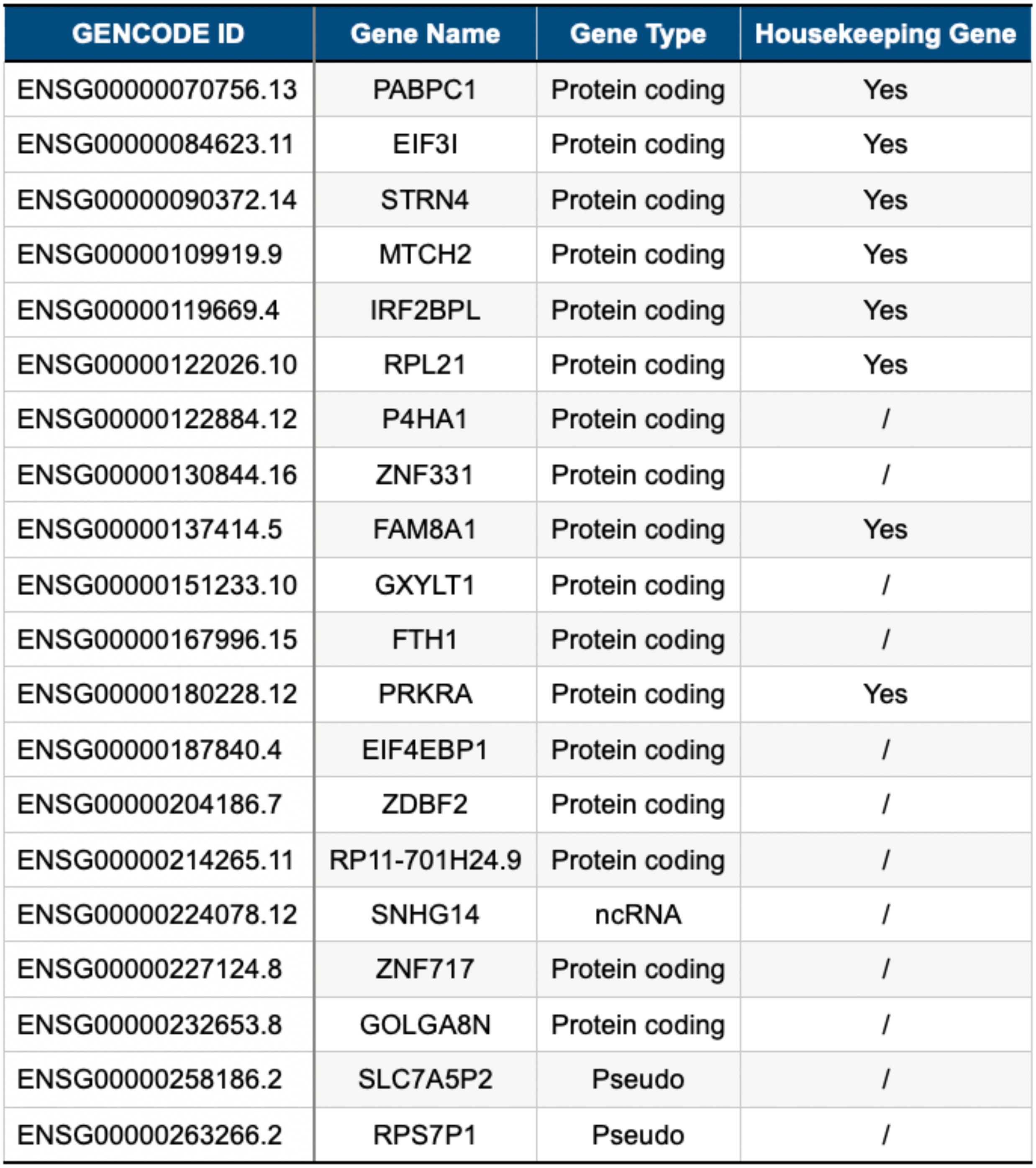
Figure S8.2b. Allelic ratios (haplotype 1 reads over the total number of reads) for gene expression of genes that are accessible across all tissues and allele-specific in at least one tissue of individual 3

This parallels the allelic ratios for H3K27ac in Fig. 7c and shows the same trend for expression as for histone modification.



cCRE ID	cCRE Type	cCRE Coordinate	Regulatory Build	Associated Gene Name	Gene Type	Housekeeping Gene
EH38D2450505	pELS_CTCF_bound	chr11_47642297_47642476	Promoter	MTCH2	Protein coding	Yes
EH38D2768300	pELS_CTCF_bound	chr15_24956163_24956513	Promoter	SNRPN	Protein coding	/
EH38D2900035	PLS_CTCF_bound	chr17_1455938_1456100	Promoter	CRK	Protein coding	Yes
EH38D2901207	pELS_CTCF_bound	chr17_2401937_2402232	Promoter	MNT	Protein coding	/
EH38D2916215	dELS_CTCF_bound	chr17_19507807_19508157	/	/	/	/
EH38D2933500	pELS_CTCF_bound	chr17_43360380_43360725	Promoter	LINC00910	LncRNA	/
EH38D3043965	pELS_CTCF_bound	chr19_14005711_14006060	Promoter	RFX1	Protein coding	/
EH38D3061913	pELS_CTCF_bound	chr19_40425366_40425529	Promoter	SERTAD1	Protein coding	/
EH38D3112234	pELS_CTCF_bound	chr2_39436701_39437019	Promoter	MAP4K3	Protein coding	/
EH38D3214874	PLS_CTCF_bound	chr2_178451090_178451434	Promoter	PRKRA	Protein coding	Yes
EH38D3218481	PLS_CTCF_bound	chr2_183038344_183038694	Promoter	NCKAP1	Protein coding	/
EH38D3320686	pELS_CTCF_bound	chr20_62652500_62652832	Promoter	SLCO4A1	Protein coding	/
EH38D3374502	dELS_CTCF_bound	chr22_41414017_41414333	/	/	/	/
EH38D3375755	pELS_CTCF_bound	chr22_42614566_42614721	Promoter	POLDIP3	Protein coding	/
EH38D3448294	PLS_CTCF_bound	chr3_75785373_75785718	Promoter	ZNF717	Protein coding	/
EH38D3802403	pELS_CTCF_bound	chr6_291711_292043	Promoter	DUSP22	Protein coding	Yes
EH38D3802406	pELS_CTCF_bound	chr6_292649_292999	Promoter	DUSP22	Protein coding	Yes
EH38D3819578	pELS_CTCF_bound	chr6_17600685_17600980	Promoter	FAM8A1	Protein coding	Yes
EH38D3829720	PLS_CTCF_bound	chr6_29888019_29888233	Promoter	HLA-H	Pseudo	/
EH38D3829827	pELS_CTCF_bound	chr6_29976507_29976854	Promoter	HCG9	LncRNA	/
EH38D3829829	dELS_CTCF_bound	chr6_29977252_29977415	Promoter	HCG9	LncRNA	/
EH38D4038383	pELS_CTCF_bound	chr7_139341640_139341807	Promoter	FMC1-LUC7L2	Protein coding	/
EH38D4168415	pELS_CTCF_bound	chr9_6007913_6008223	Promoter	KIAA2026	Protein coding	/

Figure S8.2c. Annotation of pan-tissue H3K27ac AS+ cCREs of individual 3.
Among the 23 H3K27ac AS+ cCREs that were detected across all available tissues of individual 3, 21 cCREs are within promoter regions of known genes including 6 promoters of housekeeping genes. Promoters and associated genes are based on Ensembl, and housekeeping genes are based on HRT Atlas \cite(*93*).



GENCODE ID	Gene Name	Gene Type	Housekeeping Gene
ENSG00000070756.13	PABPC1	Protein coding	Yes
ENSG00000084623.11	EIF3I	Protein coding	Yes
ENSG00000090372.14	STRN4	Protein coding	Yes
ENSG00000109919.9	MTCH2	Protein coding	Yes
ENSG00000119669.4	IRF2BPL	Protein coding	Yes
ENSG00000122026.10	RPL21	Protein coding	Yes
ENSG00000122884.12	P4HA1	Protein coding	/
ENSG00000130844.16	ZNF331	Protein coding	/
ENSG00000137414.5	FAM8A1	Protein coding	Yes
ENSG00000151233.10	GXYLT1	Protein coding	/
ENSG00000167996.15	FTH1	Protein coding	/
ENSG00000180228.12	PRKRA	Protein coding	Yes
ENSG00000187840.4	EIF4EBP1	Protein coding	/
ENSG00000204186.7	ZDBF2	Protein coding	/
ENSG00000214265.11	RP11-701H24.9	Protein coding	/
ENSG00000224078.12	SNHG14	ncRNA	/
ENSG00000227124.8	ZNF717	Protein coding	/
ENSG00000232653.8	GOLGA8N	Protein coding	/
ENSG00000258186.2	SLC7A5P2	Pseudo	/
ENSG00000263266.2	RPS7P1	Pseudo	/

Figure S8.2d. Annotation of pan-tissue ASE genes of individual 3.
Among the 20 ASE genes that were detected across at least 90% of available tissues of individual 3, 8 genes are annotated as housekeeping genes in HRT Atlas \cite(*93*).

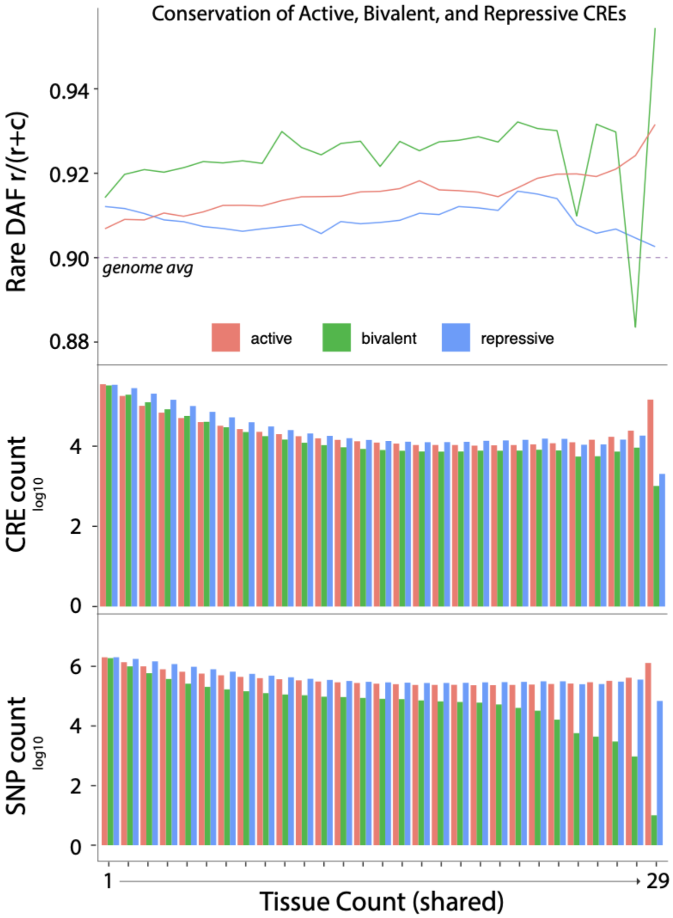


Figure S8.3a. Rare DAF (# rare variants / (# rare variants + # common variants)) for active, bivalent, and repressive cCREs in increasing tissue count

Total cCRE count and SNP count (taking into account all SNPS, common and rare) shown for tissue count as well.

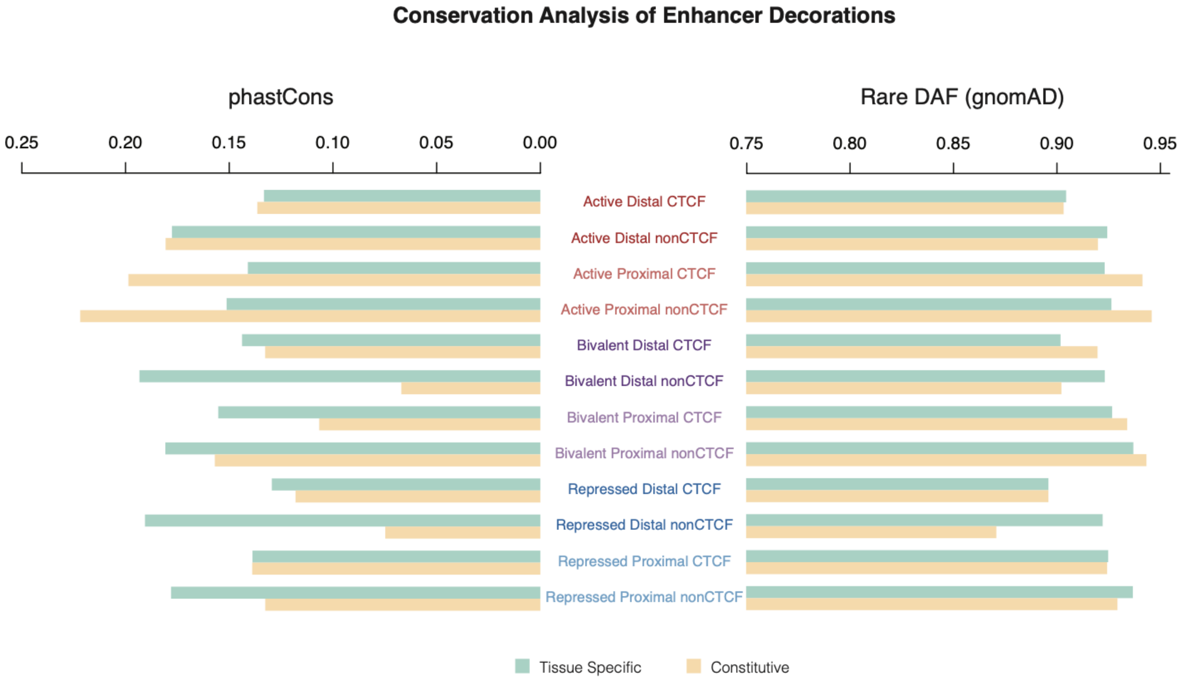


Figure S8.3b. Conservation of various cCRE decorations.
The conservation is calculated in terms of phastCons score and Rare DAF, based on the frequencies in the gnomAD database. The annotations are from Figure S7.2c.

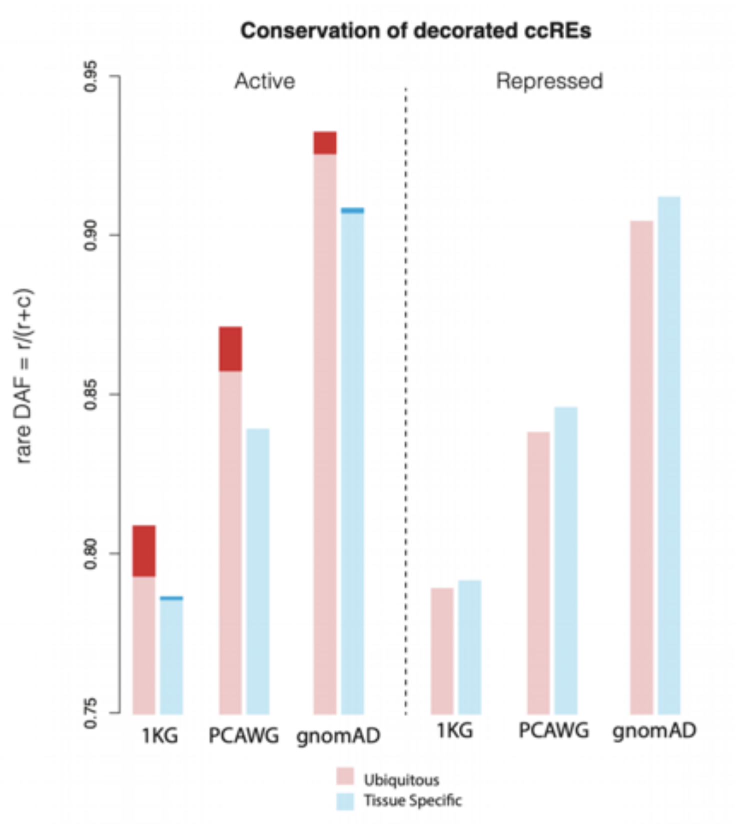


Figure S8.3c. Conservation of active and repressed cCREs for tissue specific and ubiquitous categories

Dark red shows an increase in conservation for more stringently defined cCREs (These are selected via top 1% of Matched Filter signals. More details in the Supplement text). The databases for this calculation are 1KG (1000 Genome, ref XXX), PCAWG (Pan cancer analysis working group, ref XXX, gnomaod ref XXX)

Figure S9. Supp. figures to main text section “Relating encyclopedia decorations to QTLs & GWAS loci”

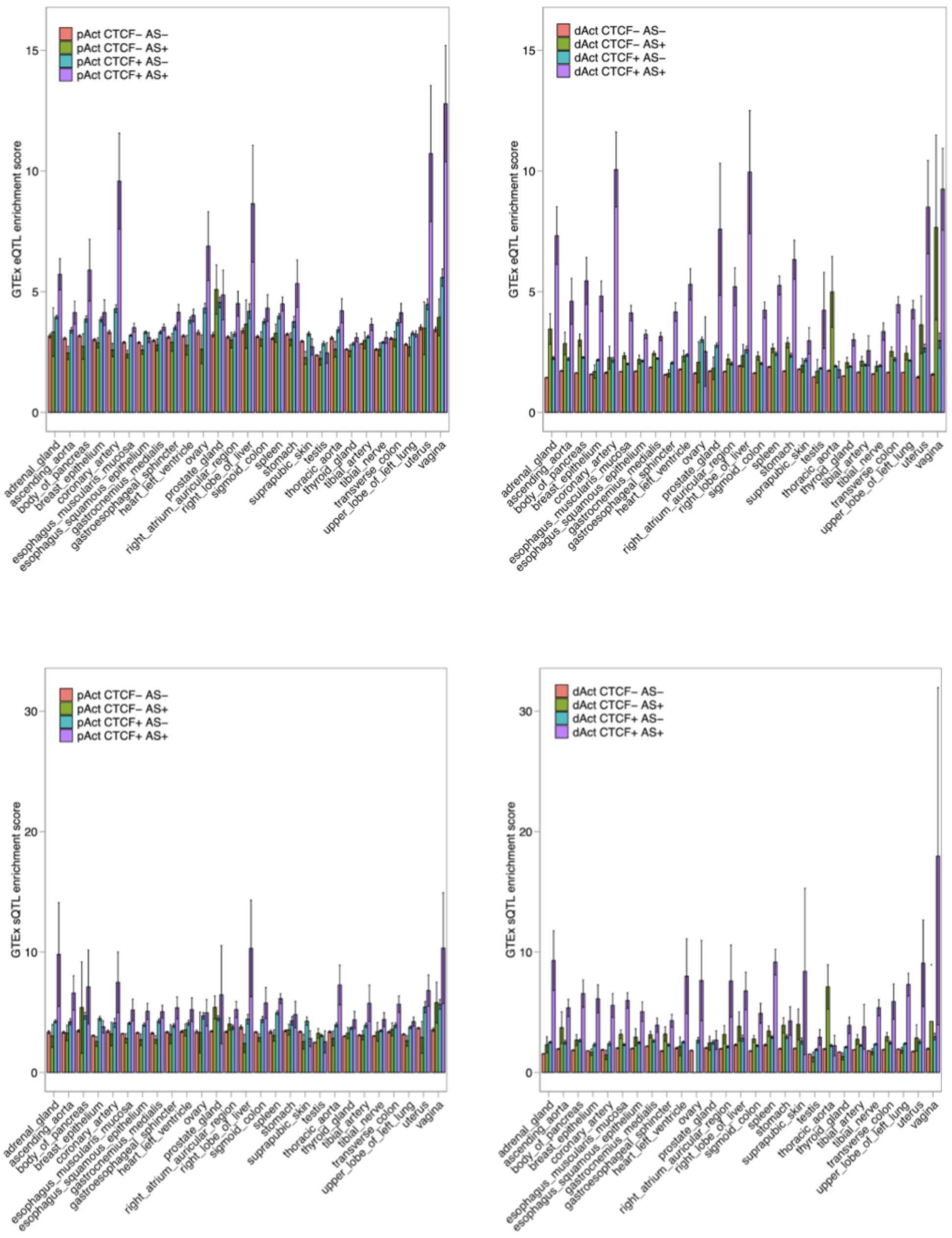
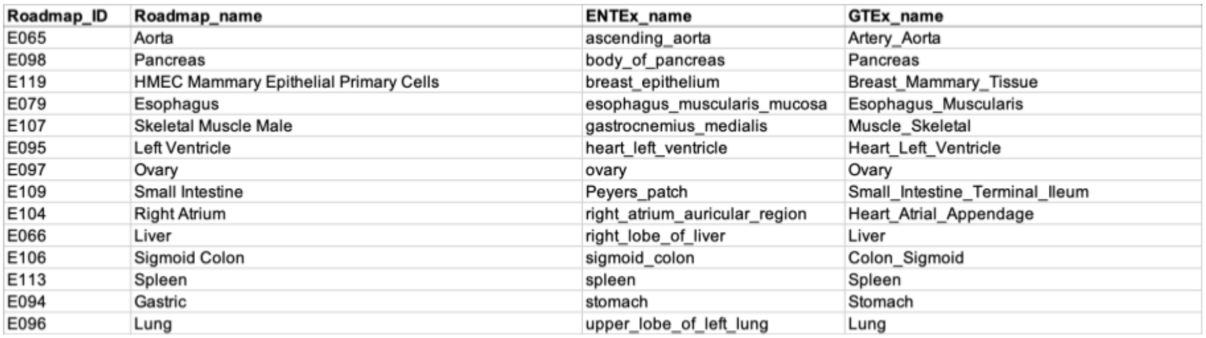


Figure S9.1a. eQTL and sQTL Enrichment in cCREs

We compute the odd ratios (ORs) to estimate the enrichment of the eQTL (upper panel) and sQTL (lower panel) SNPs identified from GTEx tissues in the cCREs from EN-TEx tissues. The ORs are calculated using the numbers of real QTL SNPs and the control SNPs located in the cCREs compared to those in the baseline regions. This procedure is repeated 30 times to calculate standard deviation, and the values are indicated by the whiskers. (See supplement text for details). In each panel, we show the QTL enrichment in the proximal active (left in each panel) and distal active (right in each panel) cCREs from each tissue type. In each figure, the cCREs are further separated into subgroups based on their CTCF binding and allelic-specific patterns. Note that the different decoration terms are defined in Figure S7.2c.



Roadmap_ID	Roadmap_name	ENTEx_name	GTEx_name
E065	Aorta	ascending_aorta	Artery_Aorta
E098	Pancreas	body_of_pancreas	Pancreas
E119	HMEC Mammary Epithelial Primary Cells	breast_epithelium	Breast_Mammary_Tissue
E079	Esophagus	esophagus_muscularis_mucosa	Esophagus_Muscularis
E107	Skeletal Muscle Male	gastrocnemius_medialis	Muscle_Skeletal
E095	Left Ventricle	heart_left_ventricle	Heart_Left_Ventricle
E097	Ovary	ovary	Ovary
E109	Small Intestine	Peyers_patch	Small_Intestine_Terminal_Ileum
E104	Right Atrium	right_atrium_auricular_region	Heart_Atrial_Appendage
E066	Liver	right_lobe_of_liver	Liver
E106	Sigmoid Colon	sigmoid_colon	Colon_Sigmoid
E113	Spleen	spleen	Spleen
E094	Gastric	stomach	Stomach
E096	Lung	upper_lobe_of_left_lung	Lung

Figure S9.1b. Roadmap Annotations
We select 14 tissue types that are matched across EN-TEx, GTEx and Roadmap projects to compare the QTL enrichment in the EN-TEx cCREs and Roadmap regulatory annotations. We use the 15-state Roadmap annotations in the analysis.

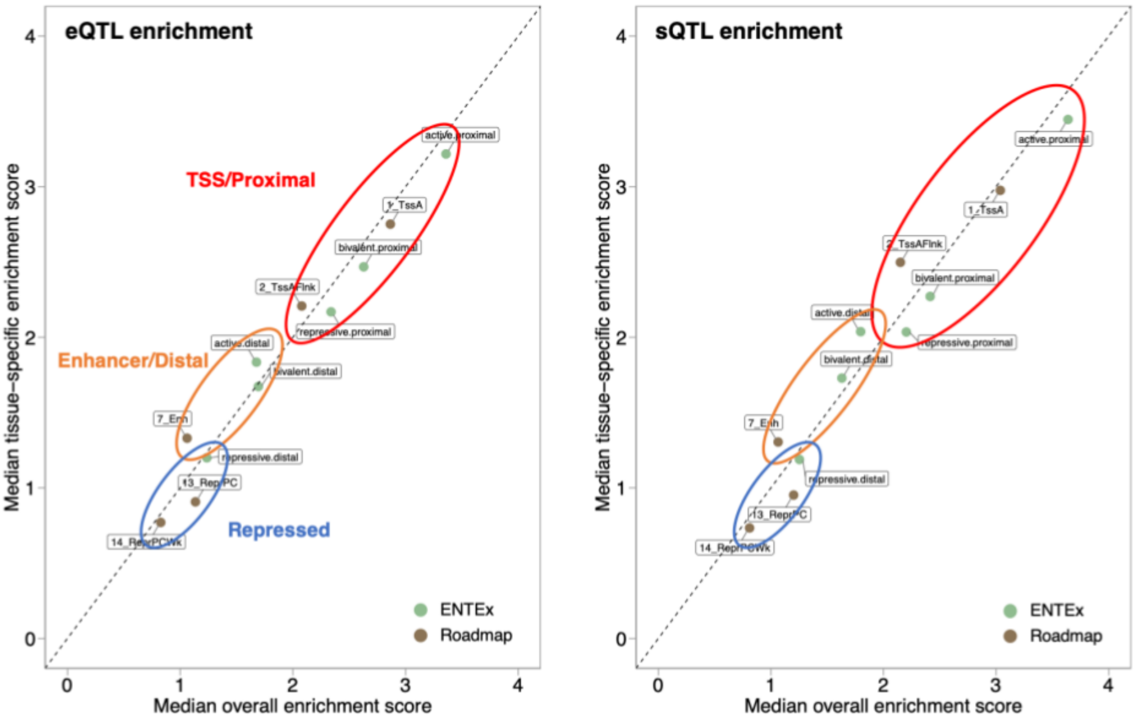


Figure S9.1c. QTL enrichment in cCREs: EN-TEx vs. Roadmap
We compare the enrichment of eQTL (left) and sQTL (right) SNPs in the TSS/proximal regions, enhancer/distal regions and repressed regions. For this calculation, we matched the annotations between ENTex and Roadmap as shown Fig S9.1b above.

A

histone mark	# of SNPs	FDR < 10%	OR > 1
H3K27ac	176,260	554	508
H3K4me3	64,650	86	84
H3K4me1	191,689	30	30
H3K36me3	232,610	18	18
H3K27me3	50,973	0	0
H3K9me3	50,236	0	0

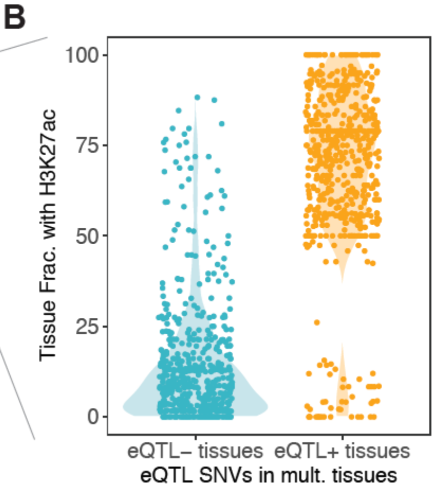


Figure S9.1d. H3K27ac marks loci associated with eQTL effect.

Of the 9,888,472 SNPs tested by the GTEx Consortium for eQTL effect in all of the 28 EN-TEx tissues, we identified 1,353,101 SNPs that are called as eQTLs in ≥ 5 tissues and that are not called as eQTLs in ≥ 5 other tissues. For each histone mark, we further subset for those SNPs marked in ≥ 10% of the ChIP-seq samples (table a "# of SNPs"). A Fisher test is thus performed for each histone mark and SNP, by comparing the proportion of tissues in which the SNP is marked being or not an eQTL. Only SNPs showing significant different proportions between the two groups of tissues are reported (Benjamini-Hochberg adjusted p-value < 0.1; table A "FDR < 10%"). Cases reporting an odds-ratio (OR) > 1 (table a, "OR > 1") correspond to SNPs being more frequently marked in the tissues in which they are called eQTLs. Violin plot B represents the 554 SNPs differentially marked by H3K27ac: a higher frequency of H3K27ac marking is observed for those tissues in which the SNP is an eQTL (orange), compared to tissues in which the SNP is not an eQTL (cyan).

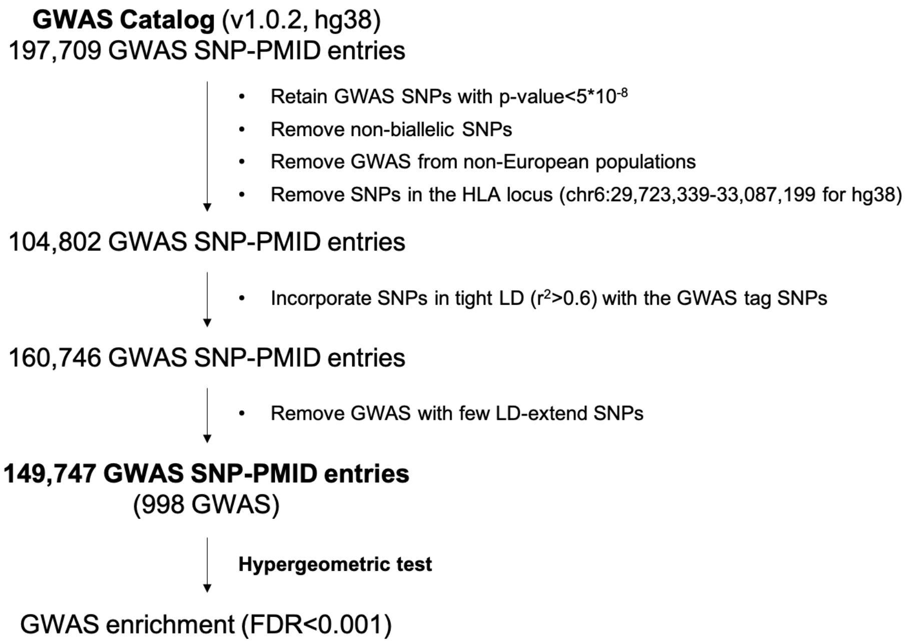

GWAS Catalog (v1.0.2, hg38)
197,709 GWAS SNP-PMID entries
• Retain GWAS SNPs with p-value<5*10^{-8}
• Remove non-biallelic SNPs
• Remove GWAS from non-European populations
• Remove SNPs in the HLA locus (chr6:29,723,339-33,087,199 for hg38)
104,802 GWAS SNP-PMID entries
• Incorporate SNPs in tight LD (r^2>0.6) with the GWAS tag SNPs
160,746 GWAS SNP-PMID entries
• Remove GWAS with few LD-extend SNPs
149,747 GWAS SNP-PMID entries
(998 GWAS)
Hypergeometric test
GWAS enrichment (FDR<0.001)

Figure S9.2a. Framework of GWAS enrichment analysis.

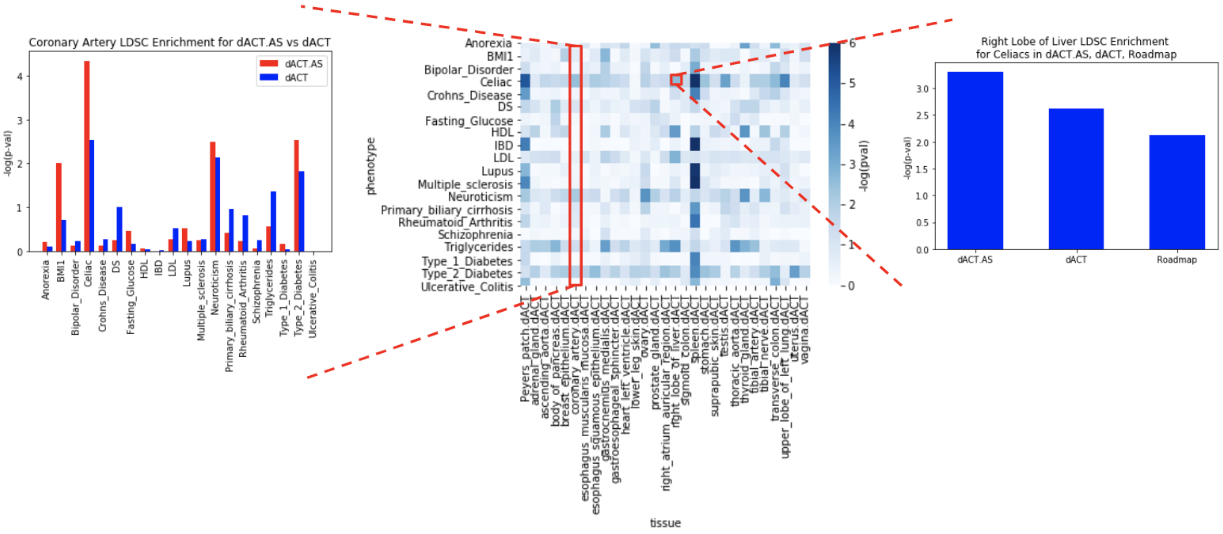


Figure S9.2b. Stratified LDSC enrichment: EN-TEx AS+ vs. AS- sv. Roadmap
This is a shadow figure for Fig. 8B in the main text. The central heatmap is the stratified LDSC enrichment of various GWAS traits over distal active elements of all EN-TEx tissues. In the left panel, we compare LDSC enrichment of distal active AS+, AS- over all traits for Coronary Artery. In the right panel, we compare LDSC enrichment of distal active AS+, AS-, and roadmap annotations in the Right Lobe of Liver.

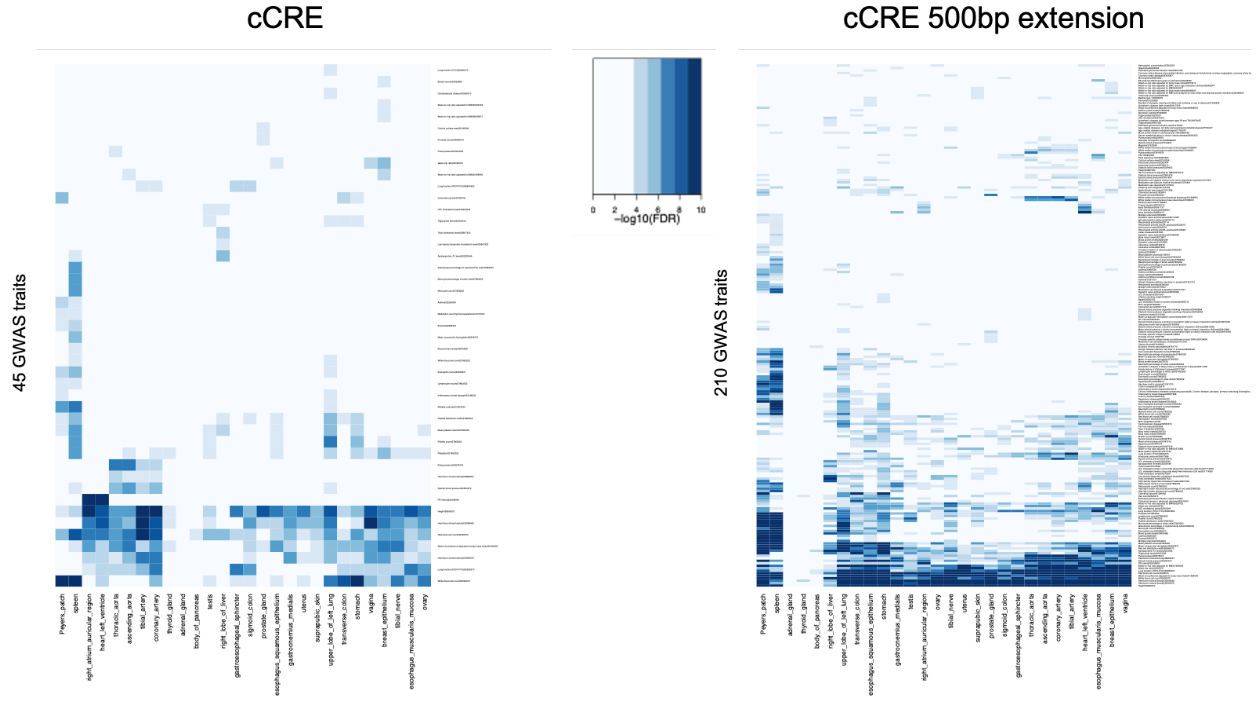


Figure S9.2c. GWAS enrichment: cCREs vs. cCREs with 500bp extension

We perform GWAS enrichment analysis on the original cCRE regions and the cCRE regions with 500bp extension on both sides. More significantly enriched GWAS traits can be identified on the cCRE regions with extension, suggesting the necessity to include the flanking regions in the GWAS enrichment analysis.

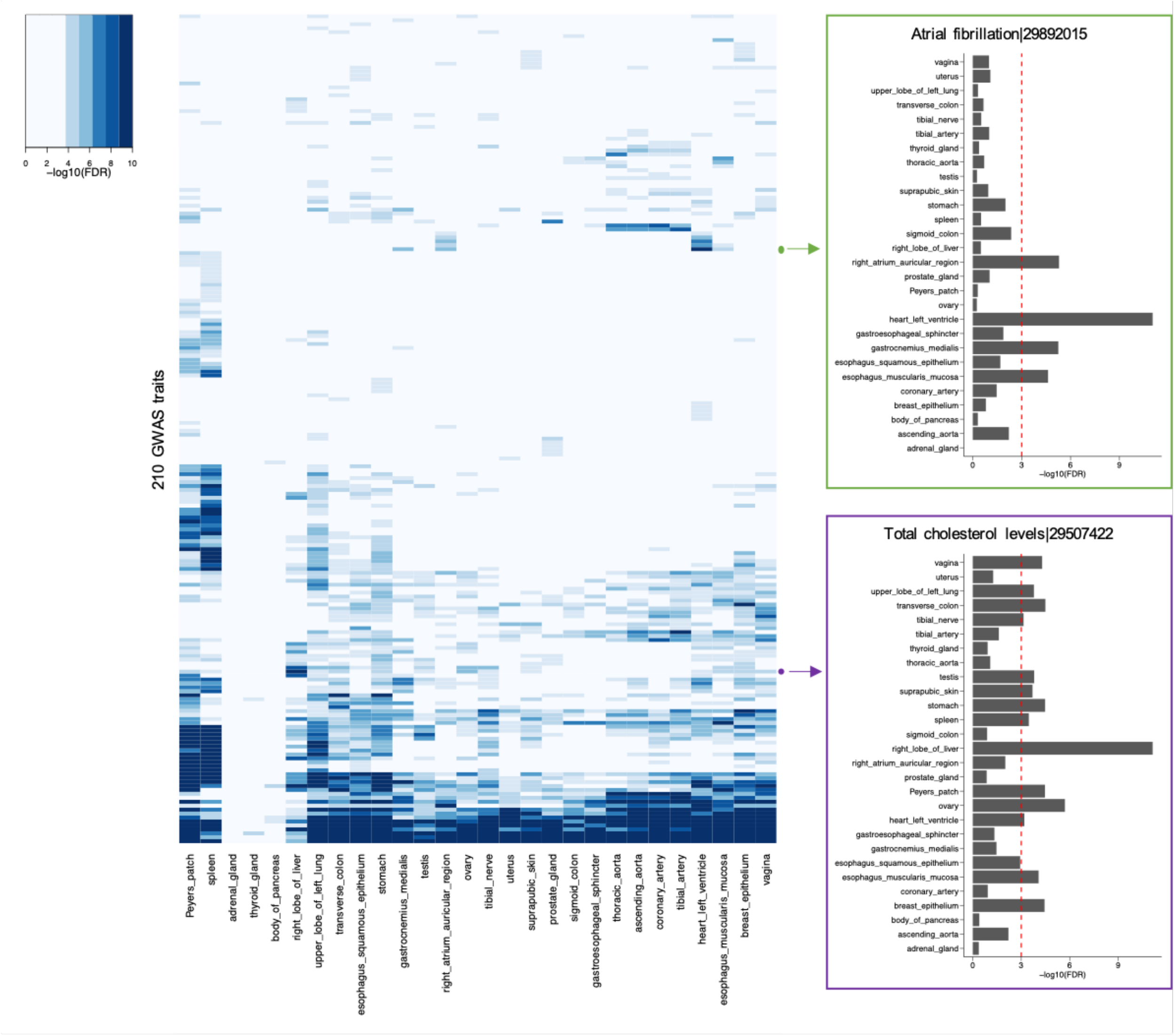
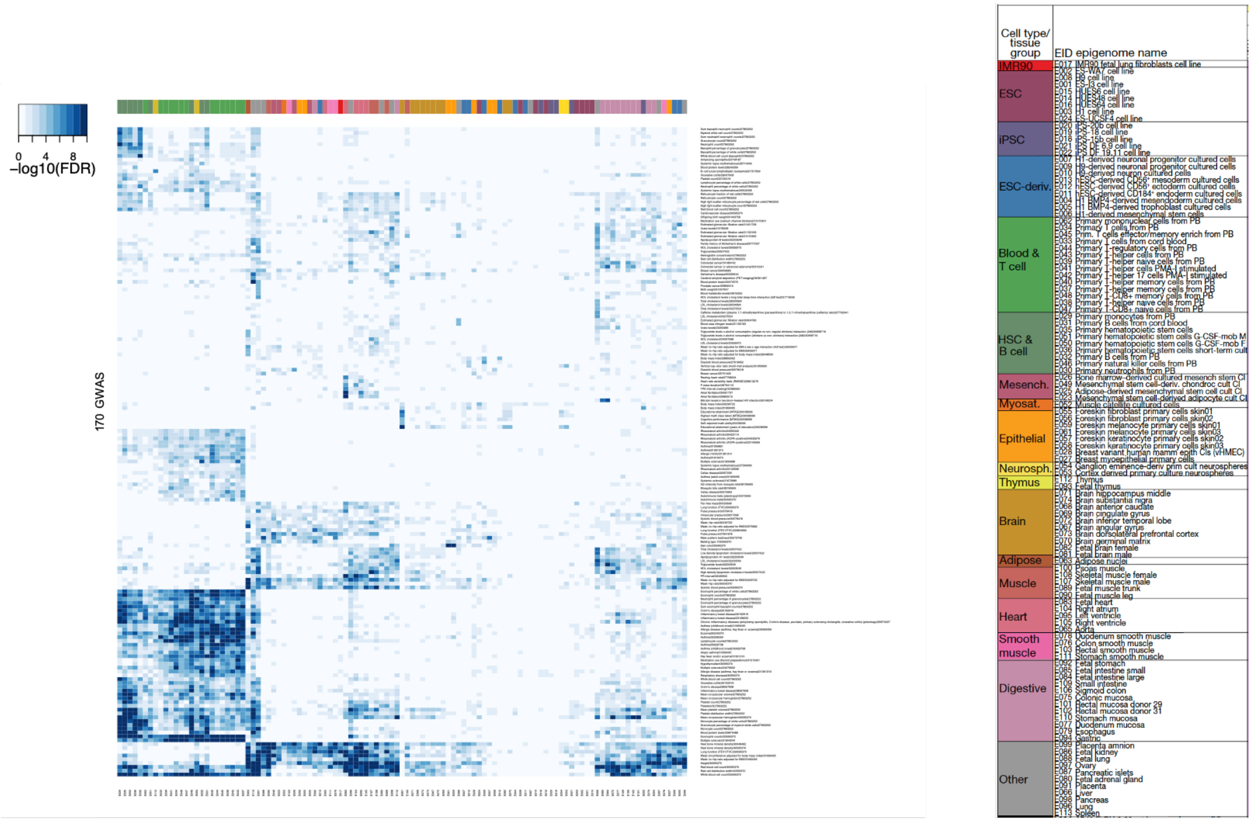


Figure S9.2d. GWAS enrichment: across tissues

We select two GWAS traits, atrial fibrillation and total cholesterol levels, to show their enrichment scores across all the tissue types.



Cell type/ tissue group	EID epigenome name
IMR90	E017 IMR90 fetal lung fibroblasts cell line
ESC	E002 ES-WA7 cell line E008 H9 cell line E001 ES-I3 cell line E015 HUES6 cell line E014 HUES48 cell line E016 HUES64 cell line E003 H1 cell line E024 ES-UCSF4 cell line
iPSC	E020 iPS-20b cell line E019 iPS-18 cell line E018 iPS-15b cell line E021 iPS DF 6.9 cell line E022 iPS DF 19.11 cell line
ESC-deriv.	E007 H1-derived neuronal progenitor cultured cells E009 H9-derived neuronal progenitor cultured cells E010 H9-derived neuron cultured cells E013 hESC-derived CD56⁺ mesoderm cultured cells E012 hESC-derived CD56⁺ ectoderm cultured cells E011 hESC-derived CD184⁺ endoderm cultured cells E004 H1 BMP4-derived mesendoderm cultured cells E005 H1 BMP4-derived trophoblast cultured cells E006 H1-derived mesenchymal stem cells
Blood & T cell	E062 Primary mononuclear cells from PB E034 Primary T cells from PB E045 Prim. T cells effector/memory enrich from PB E033 Primary T cells from cord blood E044 Primary T-regulatory cells from PB E043 Primary T-helper cells from PB E039 Primary T-helper naive cells from PB E041 Primary T-helper cells PMA-I stimulated E042 Primary T-helper 17 cells PMA-I stimulated E040 Primary T-helper memory cells from PB E037 Primary T-helper memory cells from PB E048 Primary T-CD8+ memory cells from PB E038 Primary T-helper naive cells from PB E047 Primary T-CD8+ naive cells from PB
HSC & B cell	E029 Primary monocytes from PB E031 Primary B cells from cord blood E035 Primary hematopoietic stem cells E051 Primary hematopoietic stem cells G-CSF-mob M E050 Primary hematopoietic stem cells G-CSF-mob F E036 Primary hematopoietic stem cells short-term cult E032 Primary B cells from PB E046 Primary natural killer cells from PB E030 Primary neutrophils from PB
Mesench.	E026 Bone marrow-derived cultured mesench stem Cl E049 Mesenchymal stem cell-deriv. chondroc cult Cl E025 Adipose-derived mesenchymal stem cell cult Cl E023 Mesenchymal stem cell-derived adipocyte cult Cl
Myosat.	E052 Muscle satellite cultured cells
Epithelial	E055 Foreskin fibroblast primary cells skin01 E056 Foreskin fibroblast primary cells skin02 E059 Foreskin melanocyte primary cells skin01 E061 Foreskin melanocyte primary cells skin03 E057 Foreskin keratinocyte primary cells skin02 E058 Foreskin keratinocyte primary cells skin03 E028 Breast variant human mamm epith Cls (vHMEC) E027 Breast myoepithelial primary cells
Neurosph.	E054 Ganglion eminence-deriv prim cult neurospheres E053 Cortex derived primary culture neurospheres
Thymus	E112 Thymus E093 Fetal thymus
Brain	E071 Brain hippocampus middle E074 Brain substantia nigra E068 Brain anterior caudate E069 Brain cingulate gyrus E072 Brain inferior temporal lobe E067 Brain angular gyrus E073 Brain dorsolateral prefrontal cortex E070 Brain germinal matrix E082 Fetal brain female E081 Fetal brain male
Adipose	E063 Adipose nuclei
Muscle	E100 Psoas muscle E108 Skeletal muscle female E107 Skeletal muscle male E089 Fetal muscle trunk E090 Fetal muscle leg
Heart	E083 Fetal heart E104 Right atrium E095 Left ventricle E105 Right ventricle E065 Aorta
Smooth muscle	E078 Duodenum smooth muscle E076 Colon smooth muscle E103 Rectal smooth muscle E111 Stomach smooth muscle
Digestive	E092 Fetal stomach E085 Fetal intestine small E084 Fetal intestine large E109 Small intestine E106 Sigmoid colon E075 Colonic mucosa E101 Rectal mucosa donor 29 E102 Rectal mucosa donor 31 E110 Stomach mucosa E077 Duodenum mucosa E079 Esophagus E094 Gastric
Other	E099 Placenta amnion E086 Fetal kidney E088 Fetal lung E097 Ovary E087 Pancreatic islets E080 Fetal adrenal gland E091 Placenta E066 Liver E098 Pancreas E096 Lung E113 Spleen

Figure S9.2e. GWAS enrichment: Roadmap

We perform GWAS enrichment analysis on the enhancer annotations from the 127 cell and tissue types from Roadmap Epigenomics Project. Tissue names are on the horizontal axis and traits are on the vertical. As is obvious, simple clustering of this matrix reveals a blocky structure with sets of traits associated with groups of tissues.

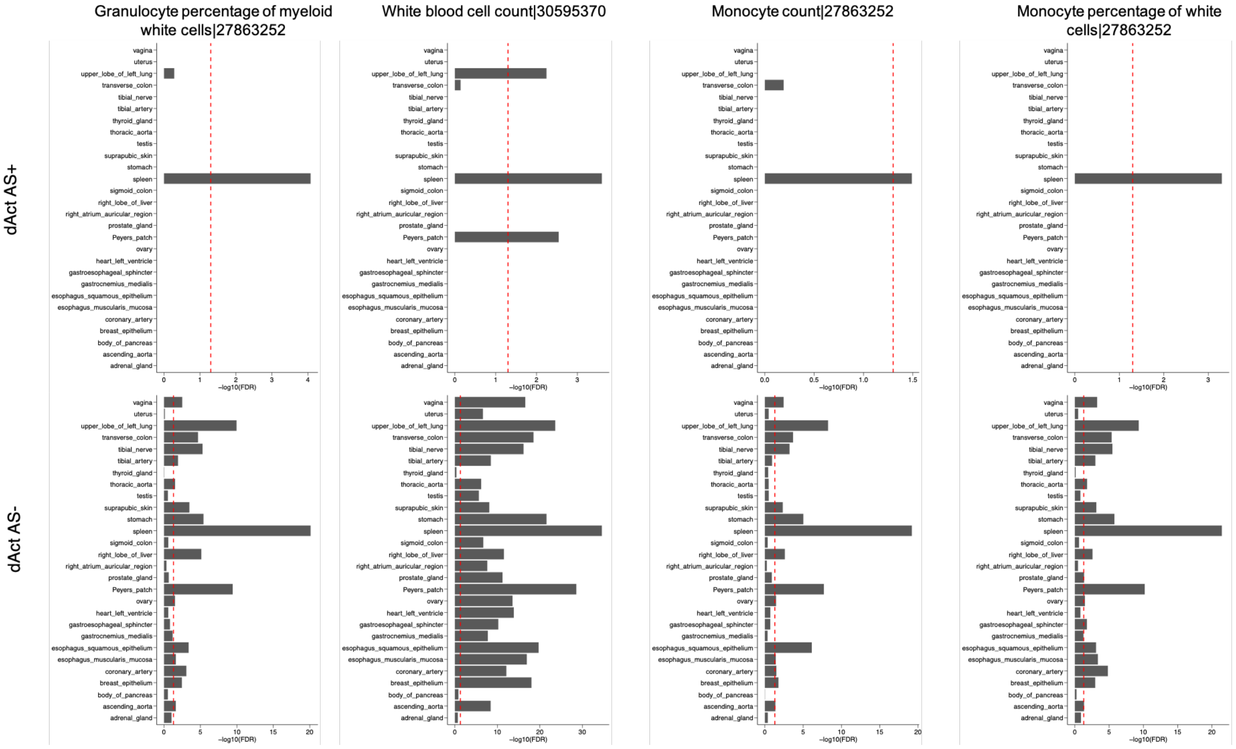
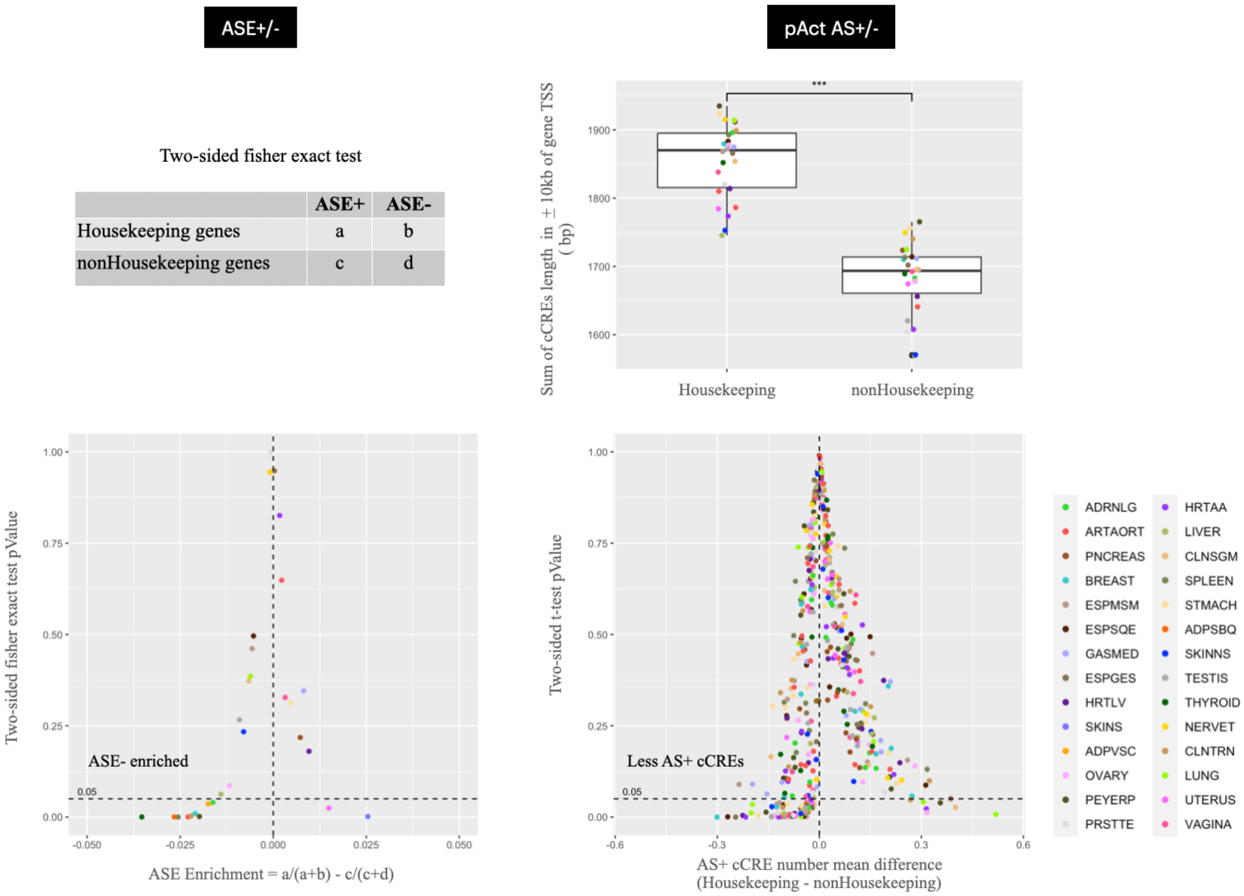


Figure S9.2f. GWAS enrichment: AS+ vs. AS- cCREs

We compare the GWAS enrichment scores on the distal active cCREs with (upper) and without (low) allelic-specific signature using the GWAS tag SNPs from blood-associated traits. Note that the different decoration terms are defined in Figure S7.2c.



	ASE+	ASE-
Housekeeping genes	a	b
nonHousekeeping genes	c	d

Figure S9.3a. Allelic specificity of housekeeping genes

Left: for each tissue, expressed protein-coding genes were split into housekeeping genes and non-housekeeping genes. Based on two-sided fisher exact tests, housekeeping genes are generally less allele-specifically expressed than non-housekeeping genes. Right: for each tissue, we examined the allelic specificity of pAct cCREs flanking the transcription starting site (TSS, defined by gene starting site) of housekeeping genes. To eliminate the bias caused by significantly different cCRE length flanking the genes, we split genes into 20 bins based on the total length of flanking cCREs. Within each bin, the number of pAct AS+ cCREs was compared between housekeeping and non-housekeeping genes, and pAct cCREs flanking housekeeping genes display relatively less allele specificity than the ones flanking non-housekeeping genes.

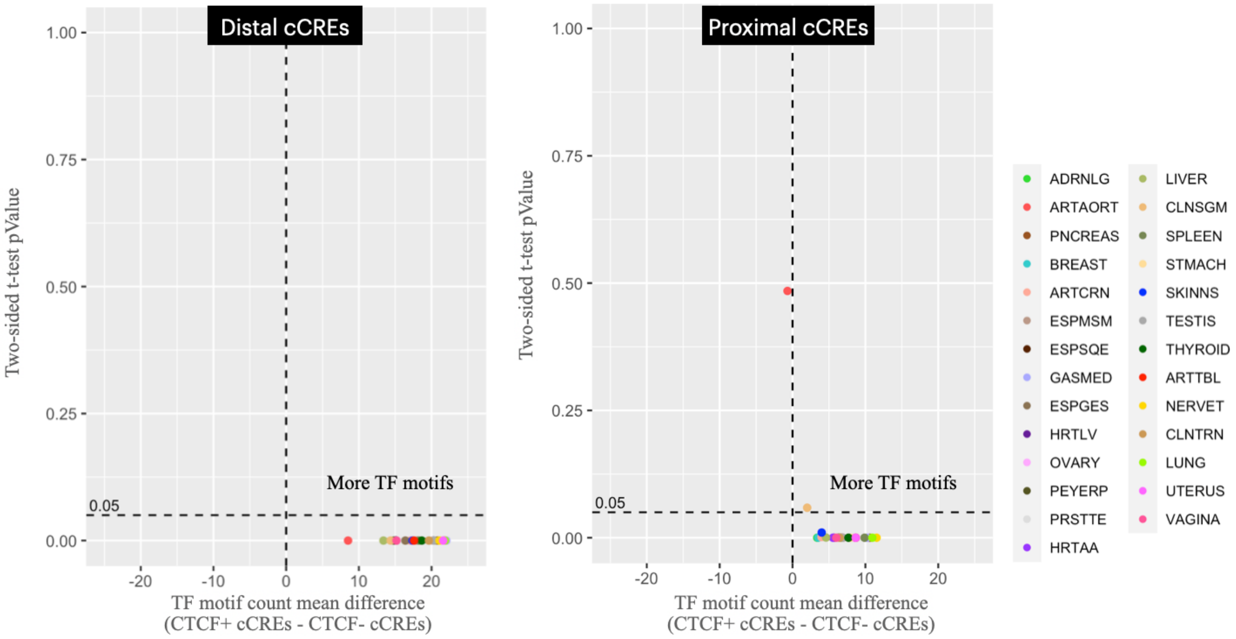
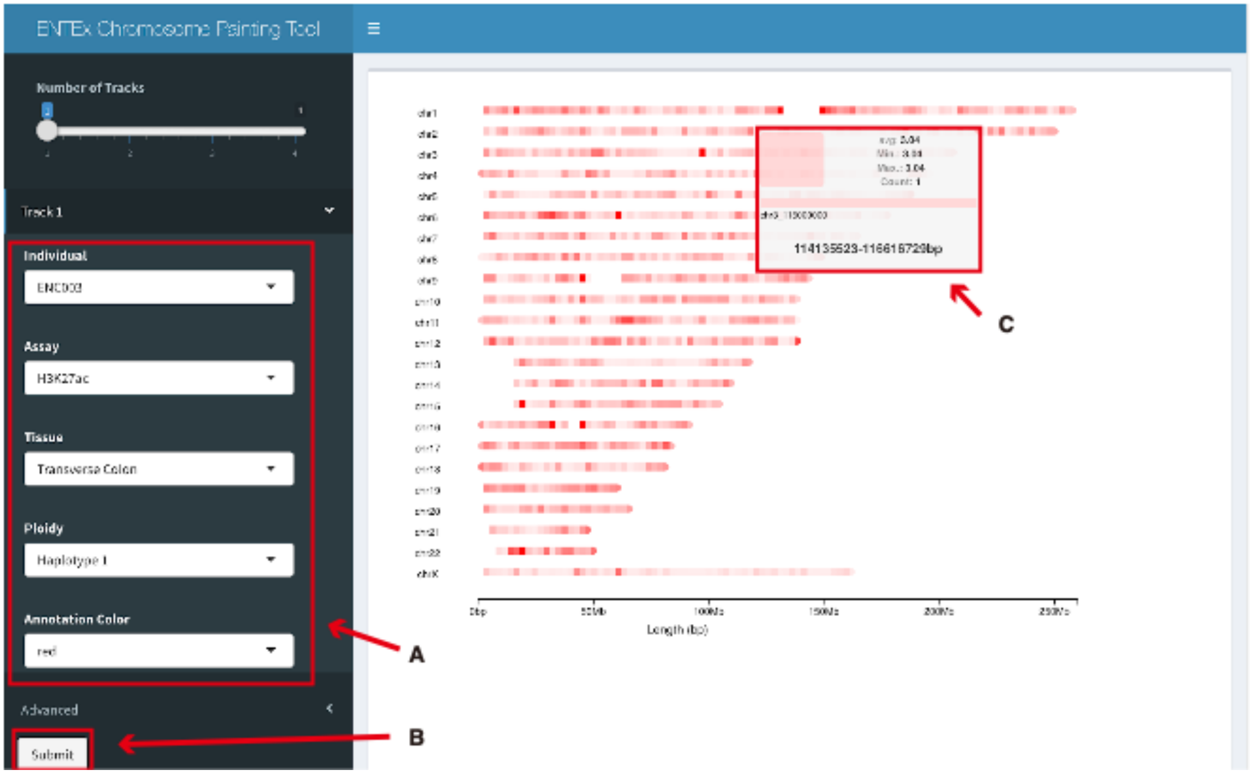


Figure S9.3b. Enrichment of TF motifs in CTCF+ cCREs

A list of 206 TF motifs (CTCF excluded) was used to count the total number of TF motifs that intersect with each CTCF+ and CTCF- cCRE in each tissue. For both distal and proximal cCREs, CTCF+ cCREs have significantly (paired-tissue two-sided t-test, p-value < 0.05) more TF motifs than CTCF- cCREs.

Figure S10. Supp. figures to supp. section “Additional information about the EN-TEx resource”



Individual: ENC003

Assay: H3K27ac

Tissue: Transverse Colon

Ploidy: Haplotype 1

Annotation Color: red

Figure S10.3a. Screenshot of EN-TEx Chromosome Painting Tool

(A) Parameters for data visualization of EN-TEx data. (B) Submit to generate visualization with the parameters. (C) Plots generated by Chromosome Painting Tool are interactive.

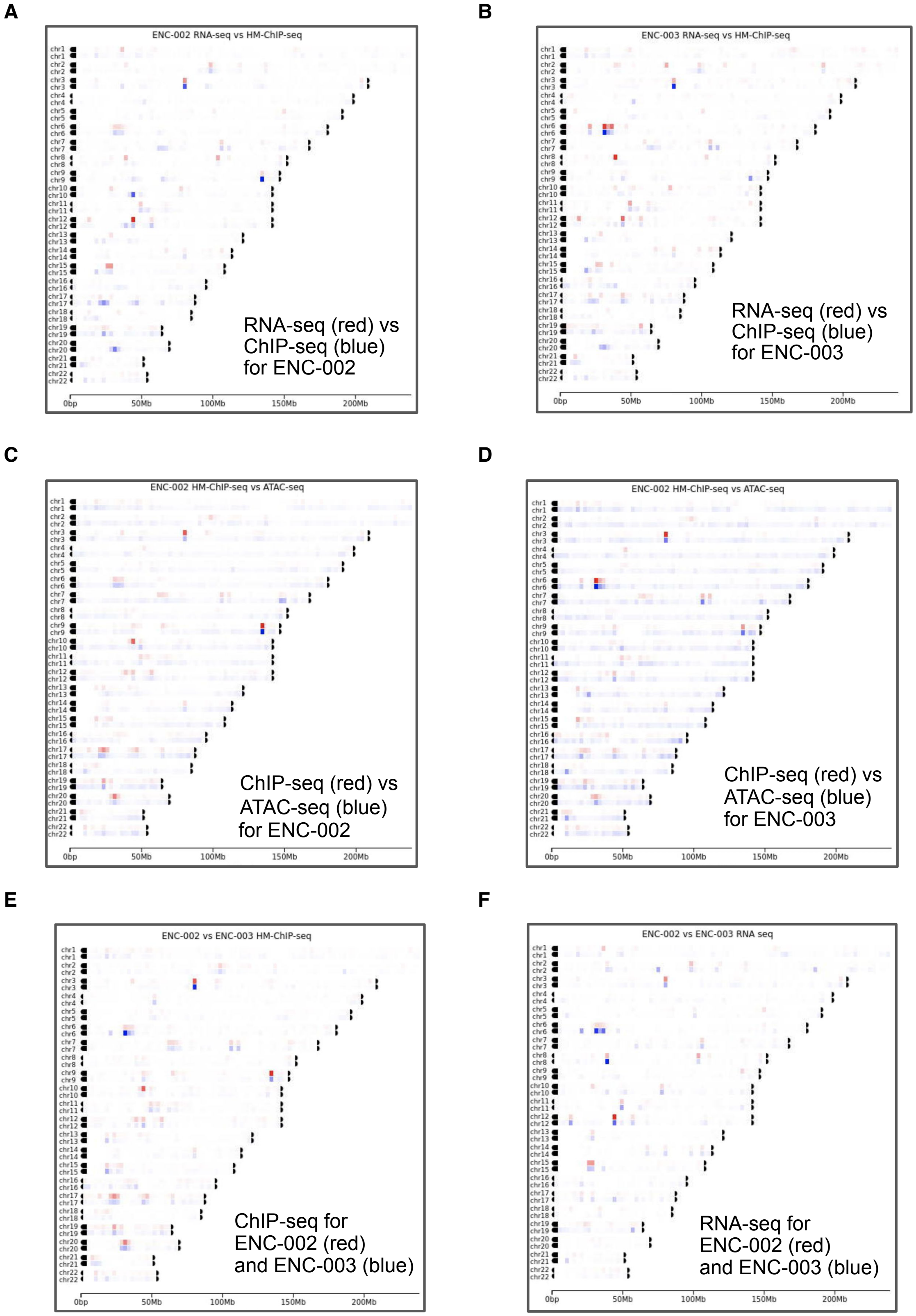
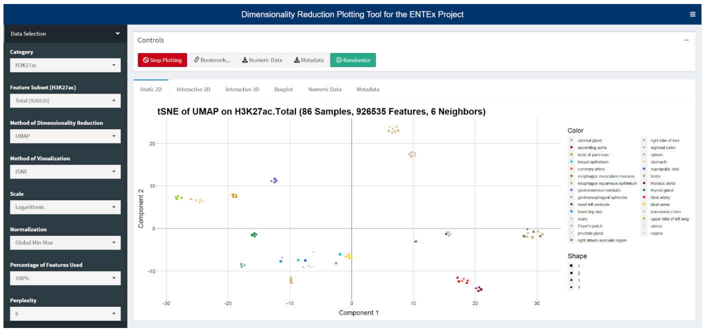

A
ENC-002 RNA-seq vs HM-ChIP-seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
RNA-seq (red) vs ChIP-seq (blue) for ENC-002
B
ENC-003 RNA-seq vs HM-ChIP-seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
RNA-seq (red) vs ChIP-seq (blue) for ENC-003
C
ENC-002 HM-ChIP-seq vs ATAC-seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
ChIP-seq (red) vs ATAC-seq (blue) for ENC-002
D
ENC-002 HM-ChIP-seq vs ATAC-seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
ChIP-seq (red) vs ATAC-seq (blue) for ENC-003
E
ENC-002 vs ENC-003 HM-ChIP-seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
ChIP-seq for ENC-002 (red) and ENC-003 (blue)
F
ENC-002 vs ENC-003 RNA seq
chr1 chr1 chr2 chr2 chr3 chr3 chr4 chr4 chr5 chr5 chr6 chr6 chr7 chr7 chr8 chr8 chr9 chr9 chr10 chr10 chr11 chr11 chr12 chr12 chr13 chr13 chr14 chr14 chr15 chr15 chr16 chr16 chr17 chr17 chr18 chr18 chr19 chr19 chr20 chr20 chr21 chr21 chr22 chr22
0bp 50Mb 100Mb 150Mb 200Mb
RNA-seq for ENC-002 (red) and ENC-003 (blue)

Figure S10.3b. Example of Chromomap
(A) - **(B)** RNA-seq (red) vs ChIP-seq (blue) for individual 2 and 3. **(C)** - **(D)** RNA-seq and ChIP-seq for individual 2 (red) and individual 3 (blue). **(E)** - **(F)** ChIP-seq (red) vs ATAC-seq (blue) for individual 2 and 3.

A



Dimensionality Reduction Plotting Tool for the ENTEx Project

Data Selection

Category: H3K27ac

Feature Subset (H3K27ac): Total (926535)

Method of Dimensionality Reduction: UMAP

Method of Visualization: tSNE

Scale: Logarithmic

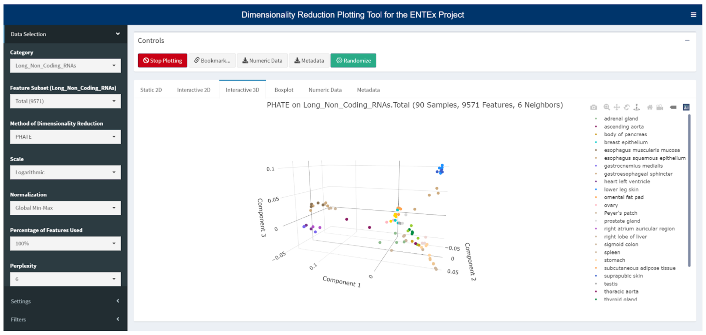
Normalization: Global Min-Max

Percentage of Features Used: 100%

Perplexity: 6

Controls

B



Dimensionality Reduction Plotting Tool for the ENTEx Project

Data Selection

Category: Long_Non_Coding_RNAs

Feature Subset (Long_Non_Coding_RNAs): Total (9571)

Method of Dimensionality Reduction: PHATE

Scale: Logarithmic

Normalization: Global Min-Max

Percentage of Features Used: 100%

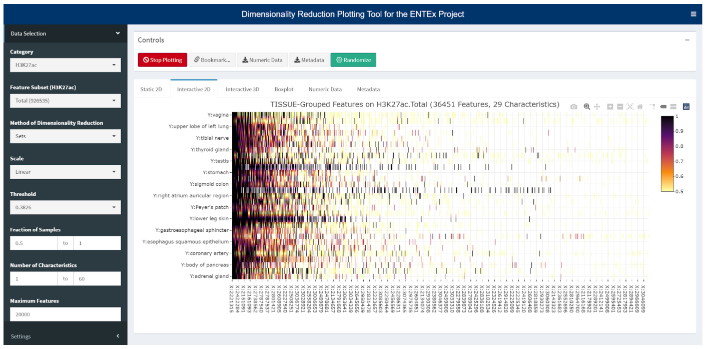
Perplexity: 6

Settings

Filters

Controls

C



Dimensionality Reduction Plotting Tool for the ENTEx Project

Data Selection

Category: H3K27ac

Feature Subset (H3K27ac): Total (926535)

Method of Dimensionality Reduction: Sets

Scale: Linear

Threshold: 0.3826

Fraction of Samples: 0.5 to 1

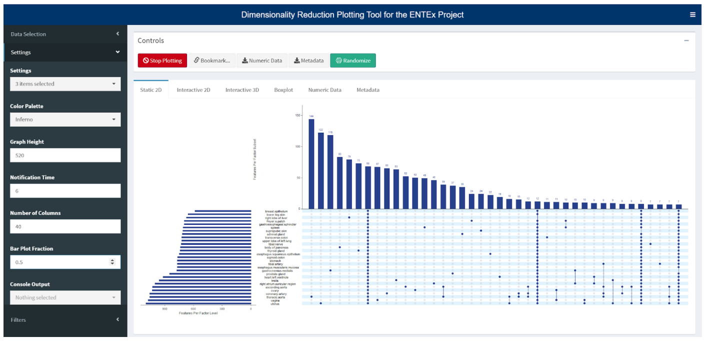
Number of Characteristics: 1 to 60

Maximum Features: 20000

Settings

Controls

D



Dimensionality Reduction Plotting Tool for the ENTEx Project

Data Selection

Settings

Settings: 3 items selected

Color Palette: Inferno

Graph Height: 520

Notification Time: 6

Number of Columns: 40

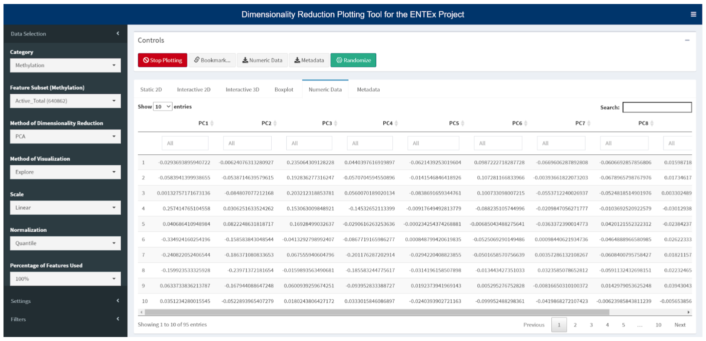
Bar Plot Fraction: 0.5

Console Output: Nothing selected

Filters

Controls

E



Dimensionality Reduction Plotting Tool for the ENTEx Project

Data Selection

Category: Methylation

Feature Subset (Methylation): Active_Total (640862)

Method of Dimensionality Reduction: PCA

Method of Visualization: Explore

Scale: Linear

Normalization: Quantile

Percentage of Features Used: 100%

Settings

Filters

Controls

Show 10 entries

Search:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
	All	All	All	All	All	All	All	All	All
1	-0.0293693895940722	-0.00624076313280927	0.235064309128228	0.0440397616919897	-0.0621439253019604	0.0987222718287728	-0.0669606287892808	-0.0606692857856806	0.01598718
2	-0.0583941399938655	-0.0538714639579615	0.192836277316247	-0.0570704594550896	-0.0141546846418926	0.107281166833966	-0.00393661822073203	-0.0678965798767976	0.01734617
3	0.00132757171673136	-0.084807077212168	0.203212318853781	0.0560070189020134	-0.0838691659344761	0.100733098007215	-0.0553712240026937	-0.0524818514901976	0.003302489
4	0.257414765104558	0.0306251633524262	0.153063009848921	-0.14532652113399	-0.00917649492813779	-0.088235105744996	-0.0209847056271777	-0.0103692520922579	-0.03012938
5	0.040686410948984	0.0822248631818717	0.16928499032637	-0.0290616263253636	-0.000234254374268881	-0.00685043488275641	-0.0363372390014773	0.0420121552322312	-0.02384237
6	-0.334924160254196	-0.158583843048544	-0.0413292798992407	-0.0867719165986277	0.000848799420619835	-0.0525069290149486	0.00098440621934736	-0.0464888966580985	0.02622333
7	-0.240822052406544	-0.186371080833653	0.067555940604796	-0.201176287202914	-0.0294220408823855	-0.0501658570756639	0.00357286132108267	-0.0608400795758427	0.01821157
8	-0.159923533325928	-0.23971372181654	-0.0159893563490681	-0.185583244775617	-0.0314196158507898	-0.013443427351033	0.0323585078652812	-0.0591132432698151	0.02232465
9	0.0633733836213787	-0.167944088647248	0.0600939259674251	-0.093952833388727	0.0192373941969143	0.005295276752828	-0.00816650310100372	0.0142979053625248	0.03943043
10	0.0351234280015545	-0.0522893965407279	0.0180243806427172	0.0333015846086897	-0.0240393902721163	-0.099952488298361	-0.0419868272107423	-0.00623985843811239	-0.005653856

Showing 1 to 10 of 95 entries

Previous 1 2 3 4 5 … 10 Next

Figure S10.4 Explorer Tool

(A) Dimensionality Reduction. The EN-TEx Explorer Tool allows for the generation of low-dimensional plots of several assays comprising cCREs, genomic expression, and proteomic expression. Data is primarily reduced to ten dimensions through principal component analysis (PCA), a variational autoencoder (VAE), uniform manifold approximation and projection (UMAP), or potential of heat-diffusion for affinity-based trajectory embedding (PHATE). Components of the result can be plotted against each other (ex: principal component 1 vs principal component 2 on a scatter plot), summarized based on the reduction method, or reduced further with t-distributed stochastic neighbor embedding (tSNE). It is also possible to rapidly view different configurations of preprocessing parameters (scaling, normalization, feature variance) or hyperparameters through extensive precomputation. **(B)** Interactive Reduction Interactive 2D and 3D visualizations are also included for intuitively exploring the data. **(C)** UpSetR Plots. UpSetR plots serve the purpose of visualizing the intersection of genes in various tissues, taking the place of the traditional Venn Diagram for larger set numbers. In the context of EN-TEx, these tools apply user-defined thresholds for each gene, consider the fraction of samples for which that gene is present in that tissue, and then calculate the UpSetR plot. **(D)** Heatmaps. Heatmaps, which can also have dendrograms applied, visualize the data that is aggregated in the UpSetR plot. **(E)** Downloading Explorer Tool Data. The numeric data and metadata for all results can be bookmarked or downloaded for rapid sharing or analysis.

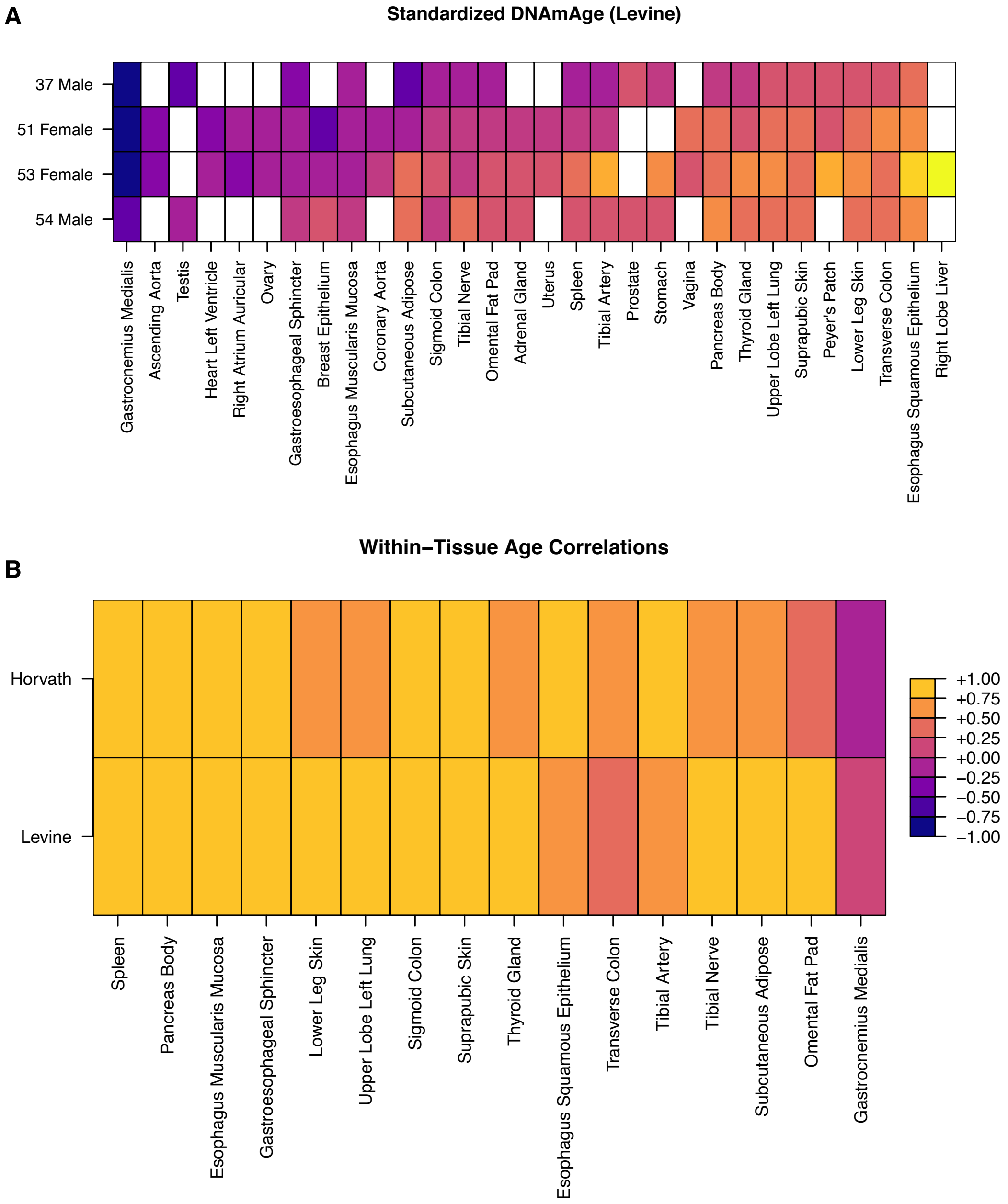


Figure S10.5. Predicting the ages of tissues from their DNA methylation.
The statistical model developed by Levine et al is used to predict the ages of the different tissues from the four individuals \cite(*108*). As a result, the different tissues of the same individuals have quite different predicted ages (A). However, for each tissue type, the predicted ages and the actual ages of the four individuals tend to be highly correlated (B), suggesting that the model is accurate for capturing the changes in tissues with the actual aging. The high correlation is also observed using other predictive models \cite(*109*). Taken together, these results suggest that the different tissues age with quite different speed.

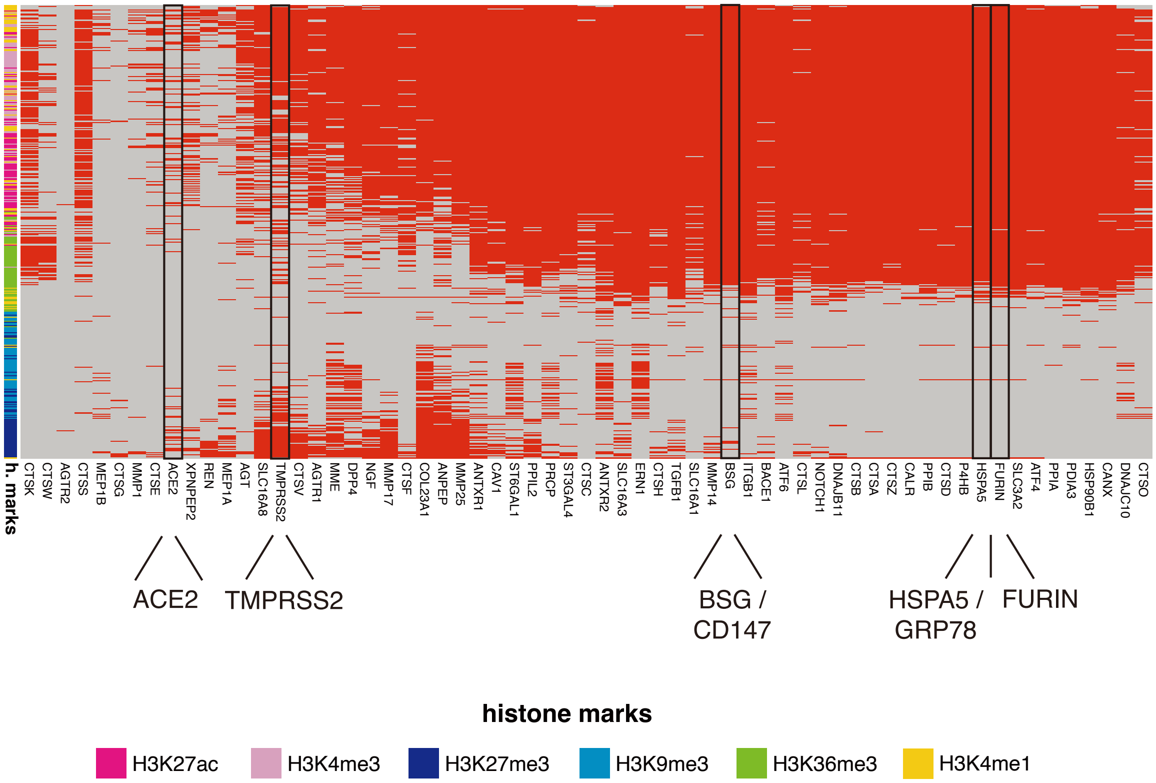


Figure S10.6. Histone ChIP-seq data for COVID19 related genes.
Chromating marking of COVID-19 related genes. The heatmap represents presence/absence (red/gray) patterns of ChIP-seq peaks for the six histone marks assayed across the EN-TEx tissues. The list of 63 genes includes ACE2, CD147, FURIN, GRP78, and their protein interactors as retrieved from STRING (https://string-db.org/cgi/input?sessionId=bDjsdV72Wbsr&input_page_show_search=off)\cite(*110*). Additional COVID-19/SARS-CoV-2 entry-associated genes proposed by the COVID19 Cell Atlas (https://www.covid19cellatlas.org/index.healthy.html), such as TMPRSS2, are also included \cite(*111*).