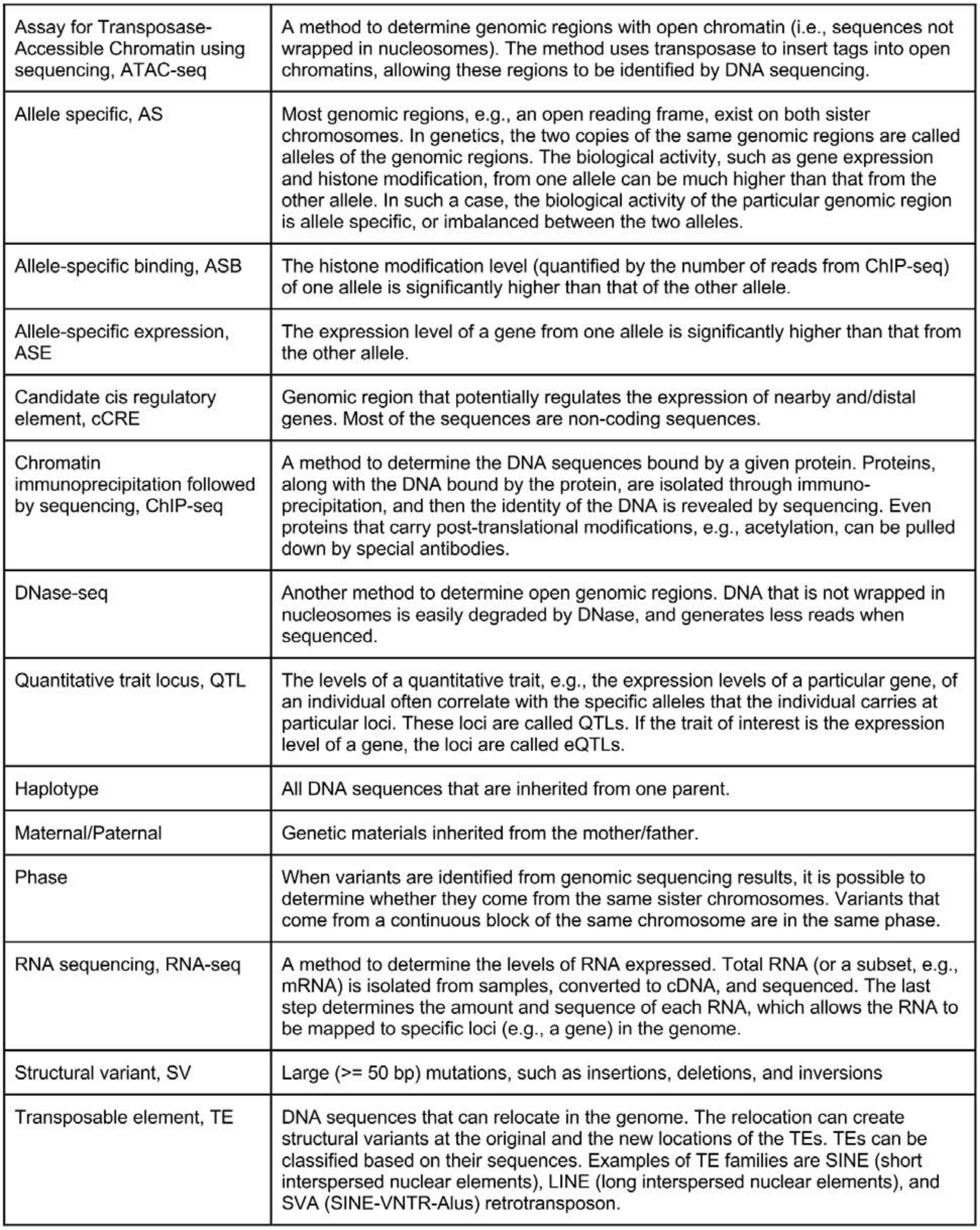
Document S1: Data S1 - Data S35

Additional figures for “The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models.”

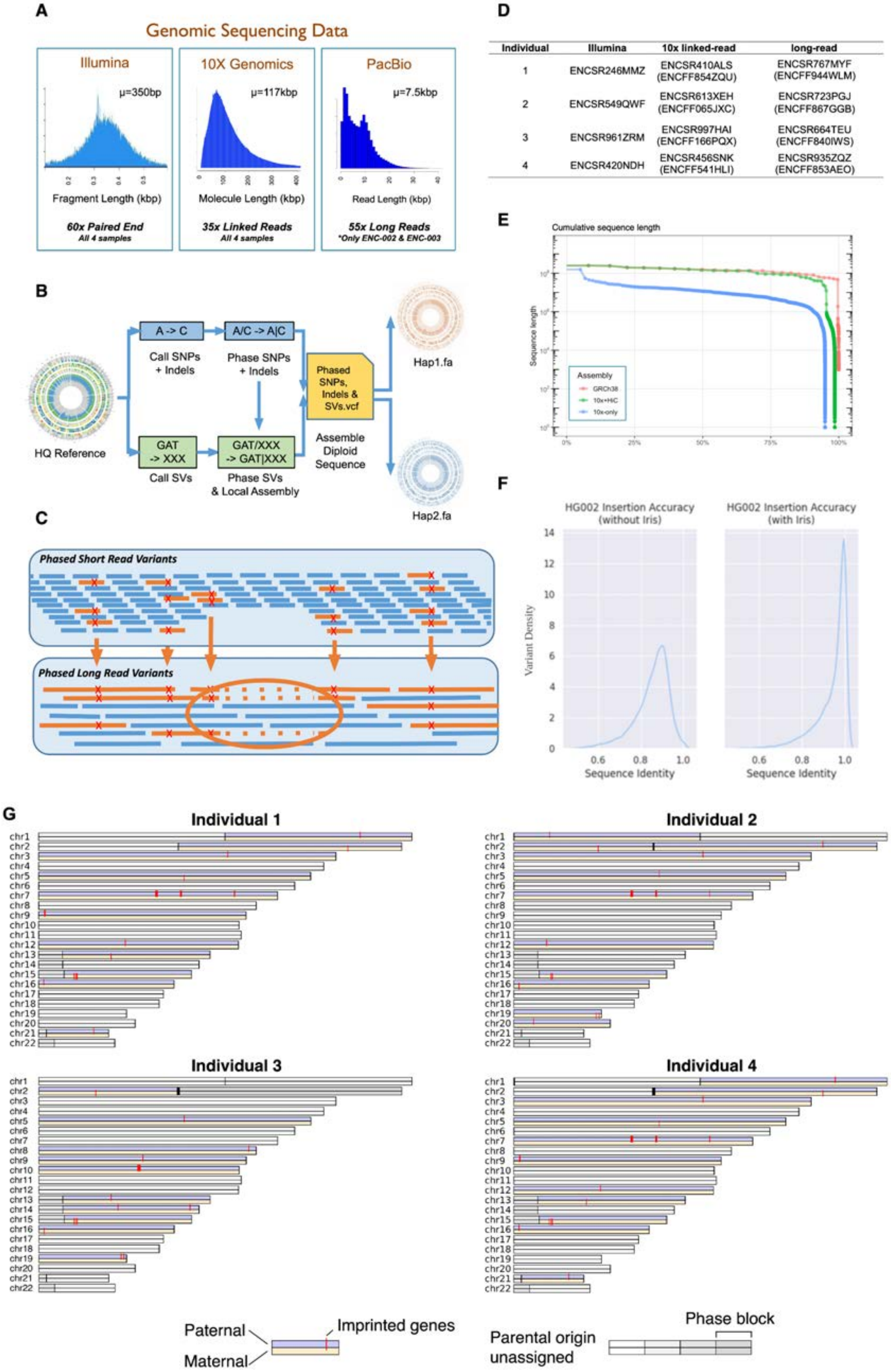
Table of contents



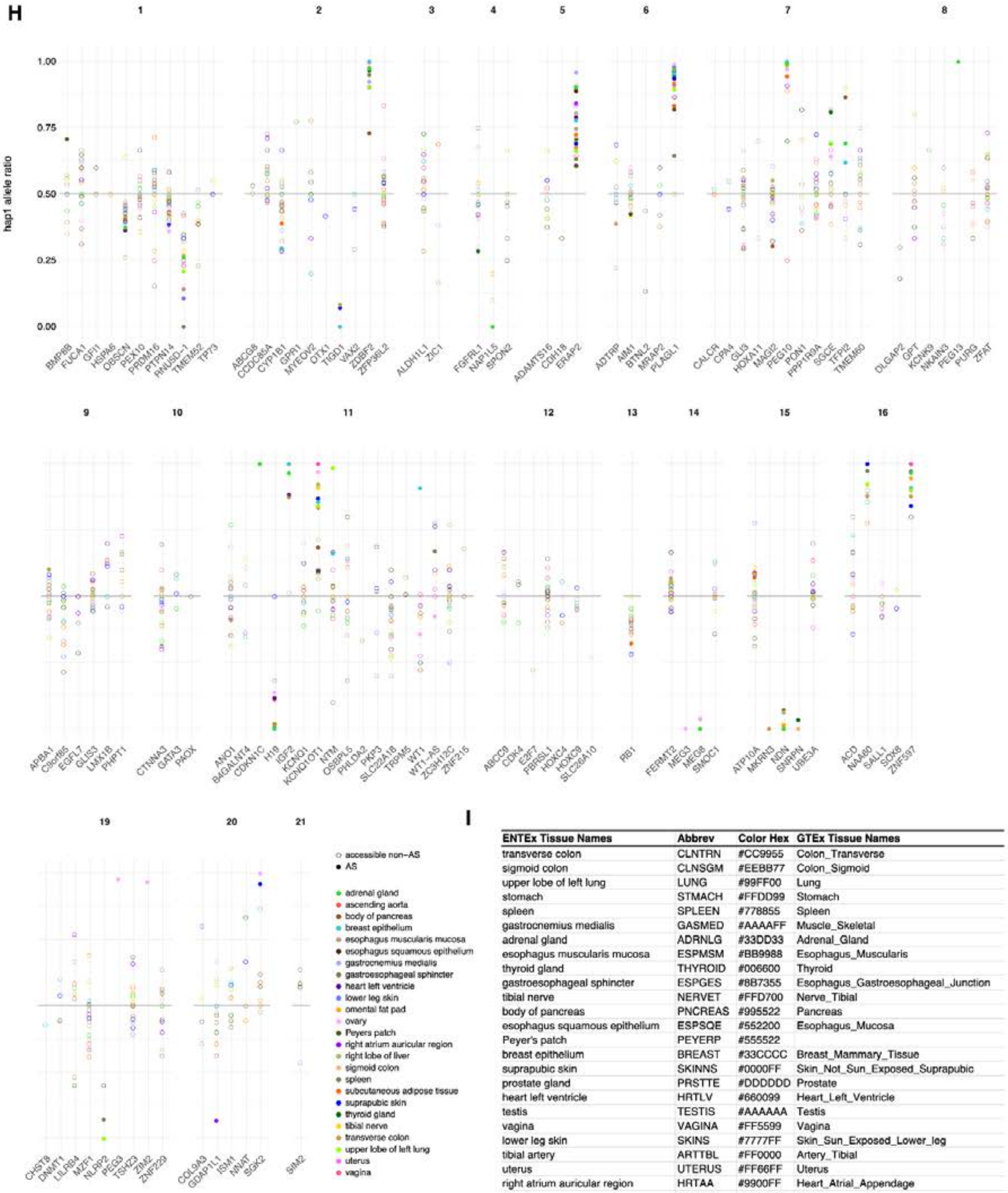
Assay for Transposase-Accessible Chromatin using sequencing, ATAC-seq	A method to determine genomic regions with open chromatin (i.e., sequences not wrapped in nucleosomes). The method uses transposase to insert tags into open chromatins, allowing these regions to be identified by DNA sequencing.
Allele specific, AS	Most genomic regions, e.g., an open reading frame, exist on both sister chromosomes. In genetics, the two copies of the same genomic regions are called alleles of the genomic regions. The biological activity, such as gene expression and histone modification, from one allele can be much higher than that from the other allele. In such a case, the biological activity of the particular genomic region is allele specific, or imbalanced between the two alleles.
Allele-specific binding, ASB	The histone modification level (quantified by the number of reads from ChIP-seq) of one allele is significantly higher than that of the other allele.
Allele-specific expression, ASE	The expression level of a gene from one allele is significantly higher than that from the other allele.
Candidate cis regulatory element, cCRE	Genomic region that potentially regulates the expression of nearby and/distal genes. Most of the sequences are non-coding sequences.
Chromatin immunoprecipitation followed by sequencing, ChIP-seq	A method to determine the DNA sequences bound by a given protein. Proteins, along with the DNA bound by the protein, are isolated through immuno-precipitation, and then the identity of the DNA is revealed by sequencing. Even proteins that carry post-translational modifications, e.g., acetylation, can be pulled down by special antibodies.
DNase-seq	Another method to determine open genomic regions. DNA that is not wrapped in nucleosomes is easily degraded by DNase, and generates less reads when sequenced.
Quantitative trait locus, QTL	The levels of a quantitative trait, e.g., the expression levels of a particular gene, of an individual often correlate with the specific alleles that the individual carries at particular loci. These loci are called QTLs. If the trait of interest is the expression level of a gene, the loci are called eQTLs.
Haplotype	All DNA sequences that are inherited from one parent.
Maternal/Paternal	Genetic materials inherited from the mother/father.
Phase	When variants are identified from genomic sequencing results, it is possible to determine whether they come from the same sister chromosomes. Variants that come from a continuous block of the same chromosome are in the same phase.
RNA sequencing, RNA-seq	A method to determine the levels of RNA expressed. Total RNA (or a subset, e.g., mRNA) is isolated from samples, converted to cDNA, and sequenced. The last step determines the amount and sequence of each RNA, which allows the RNA to be mapped to specific loci (e.g., a gene) in the genome.
Structural variant, SV	Large (>= 50 bp) mutations, such as insertions, deletions, and inversions
Transposable element, TE	DNA sequences that can relocate in the genome. The relocation can create structural variants at the original and the new locations of the TEs. TEs can be classified based on their sequences. Examples of TE families are SINE (short interspersed nuclear elements), LINE (long interspersed nuclear elements), and SVA (SINE-VNTR-Alus) retrotransposon.

Data S1. Glossary

Terms and acronyms that are frequently used in the manuscript.



Individual	Illumina	10x linked-read	long-read
1	ENCSR246MMZ	ENCSR410ALS (ENCFF854ZQU)	ENCSR767MYF (ENCFF944WLM)
2	ENCSR549QWF	ENCSR613XEH (ENCFF065JXC)	ENCSR723PGJ (ENCFF867GGB)
3	ENCSR961ZRM	ENCSR997HAI (ENCFF166PQX)	ENCSR664TEU (ENCFF840IWS)
4	ENCSR420NDH	ENCSR456SNK (ENCFF541HLI)	ENCSR935ZQZ (ENCFF853AEO)



I

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

Data S2. Personal genome construction, related to Figure 1, Figure S1, and STAR Methods “Personal Genome” Section

(A) Summary of whole-genome sequencing (WGS). All four individuals were sequenced with regular Illumina short reads, 10x linked reads, and long-reads (PacBio and Oxford Nanopore). Figure shows the sequencing depth and read-length distribution under each platform.

(B) 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.

(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 when all three of the long reads, including that deletion, have small variants that are unique to the orange haplotype.

(D) Accession numbers of the WGS data. Accession numbers in parentheses are for VCF files. For individuals 2 and 3 the accession numbers for the personal genome assemblies are ENCFF032RPN and ENCFF836JIE respectively. For individuals 1 and 4 the accession numbers for the personal genome assemblies are ENCFF477YTR and ENCFF132WPC respectively (these correspond to the SVs in Fig. S1D). Earlier assemblies for individuals 1 and 4 were used for some of the analyses in this paper (accessions ENCFF498DUG and ENCFF578XWE). These earlier assemblies for individuals 1 and 4 do not include the long-read Oxford Nanopore sequencing.

(E) Phase block length. This figure shows the size of the 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.

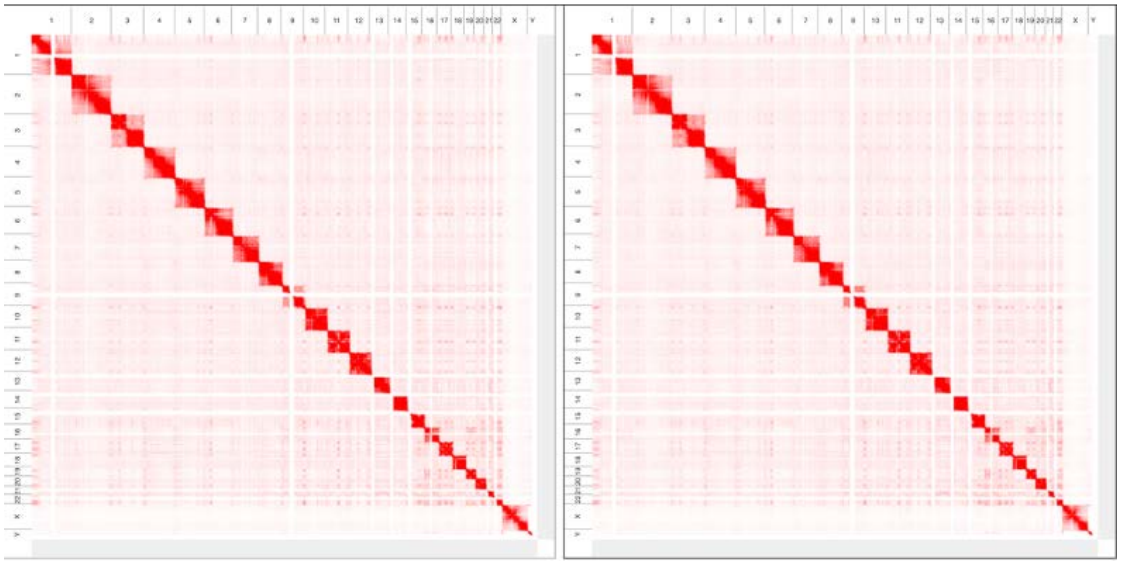
(F) Refining novel insertion sequences with Iris. Figure shows the sequence similarity of ONT calls to CCS calls in the Genome-in-a-Bottle sample HG002, used to benchmark the performance of Iris. The sequence similarity between two sequences S and T is calculated as edit_distance(S, T) / [max(length(S), length(T))].

(G) Phase blocks of personal genomes with parental origins. The parental origin of each phase block was determined based on the consistency between the direction of AS gene expression and the direction of known imprinted genes.

(H) Consistency of AS expression (ASE) imbalance direction in known imprinted genes across tissue samples (individual 3). Figure shows the fraction of reads preferentially mapping to haplotype 1 (hap1) in known imprinted genes. For most genes, the direction of the significantly imbalanced genes (filled circles) is consistent across samples from different tissues.

(I) Information on EN-TEx tissues. Table shows the full name, abbreviation, and color code of the EN-TEx tissues, as well as their matching relationship with GTEx tissues. This tissue color scheme is used in panel (H), Figure 1, and other main or supplementary figures, unless otherwise noted.

A


ENC-001
G. Medialis

ENC-002
G. Medialis

B

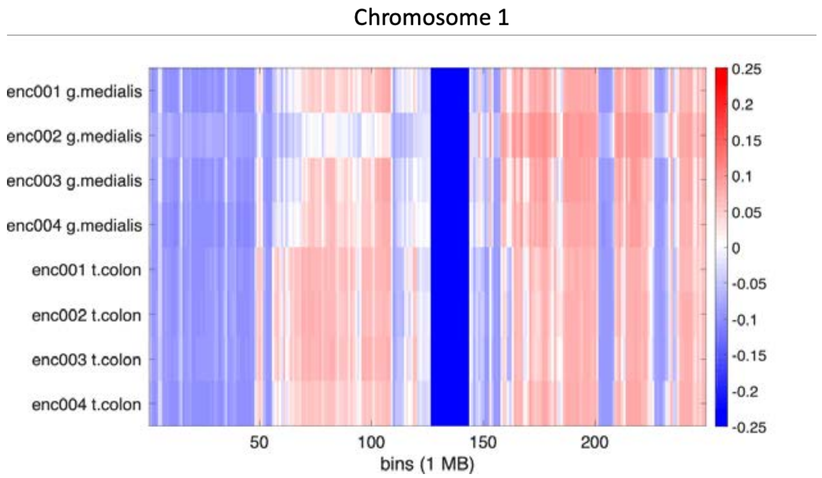
Number of paired reads (billions)

	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

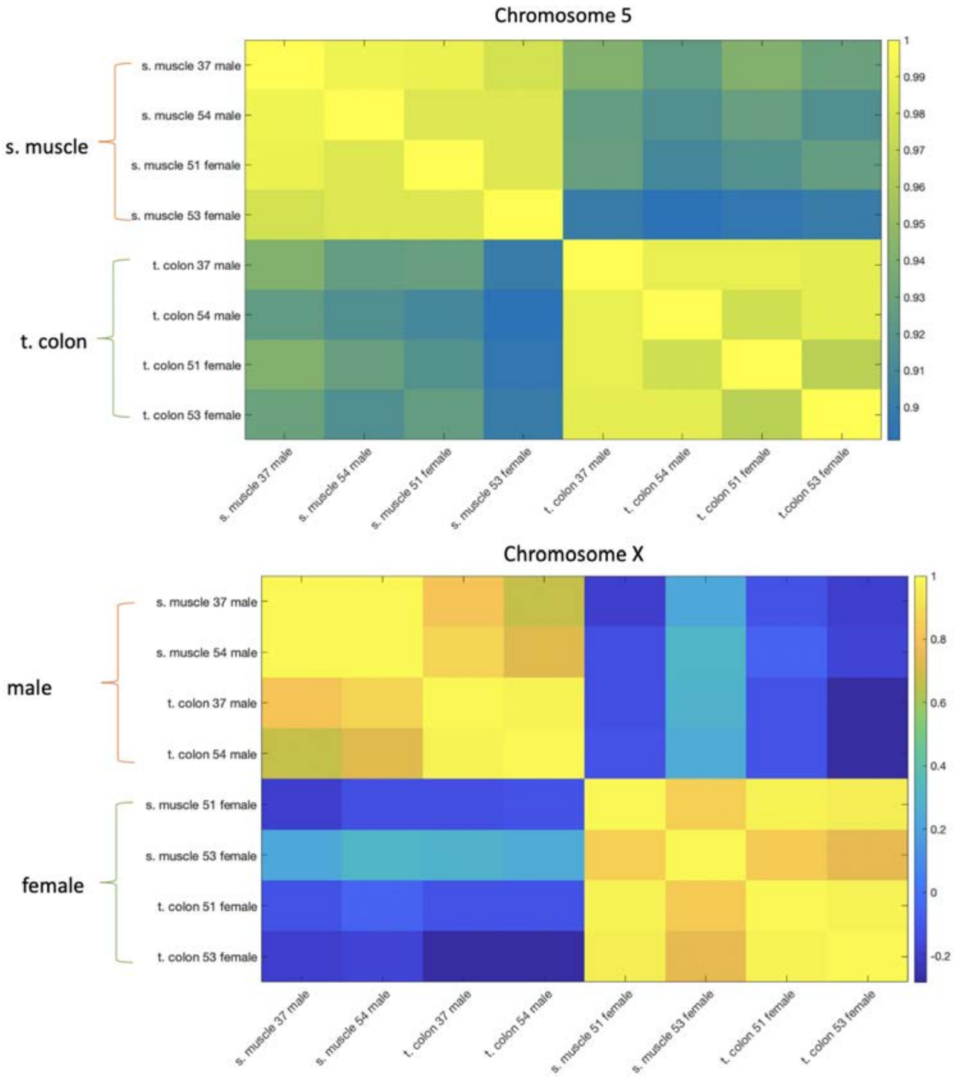
Number of contacts (billions)

	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

C


Chromosome 1
enc001 g.medialis
enc002 g.medialis
enc003 g.medialis
enc004 g.medialis
enc001 t.colon
enc002 t.colon
enc003 t.colon
enc004 t.colon
50
100
150
200
bins (1 MB)
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25

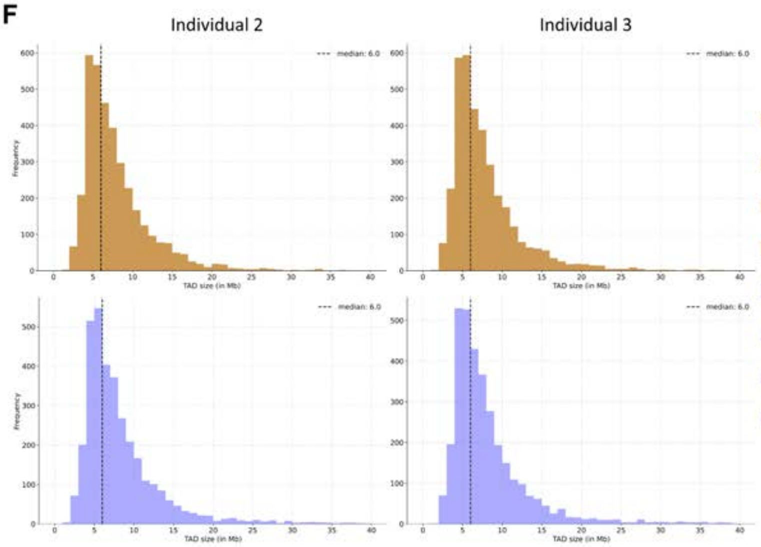
D


Chromosome 5
s. muscle
t. colon
s. muscle 37 male
s. muscle 54 male
s. muscle 51 female
s. muscle 53 female
t. colon 37 male
t. colon 54 male
t. colon 51 female
t. colon 53 female
s. muscle 37 male
s. muscle 54 male
s. muscle 51 female
s. muscle 53 female
t. colon 37 male
t. colon 54 male
t. colon 51 female
t.colon 53 female
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
Chromosome X
male
female
s. muscle 37 male
s. muscle 54 male
t. colon 37 male
t. colon 54 male
s. muscle 51 female
s. muscle 53 female
t. colon 51 female
t. colon 53 female
s. muscle 37 male
s. muscle 54 male
t. colon 37 male
t. colon 54 male
s. muscle 51 female
s. muscle 53 female
t. colon 51 female
t. colon 53 female
1
0.8
0.6
0.4
0.2
0
-0.2

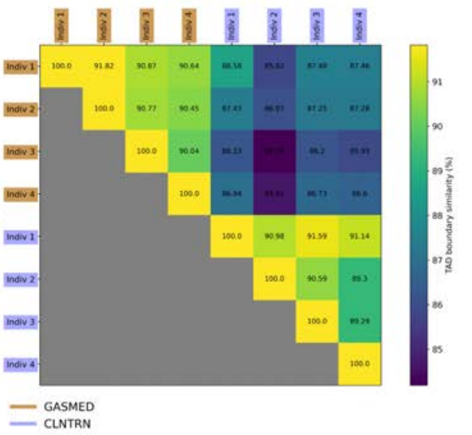
E

		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

F



G

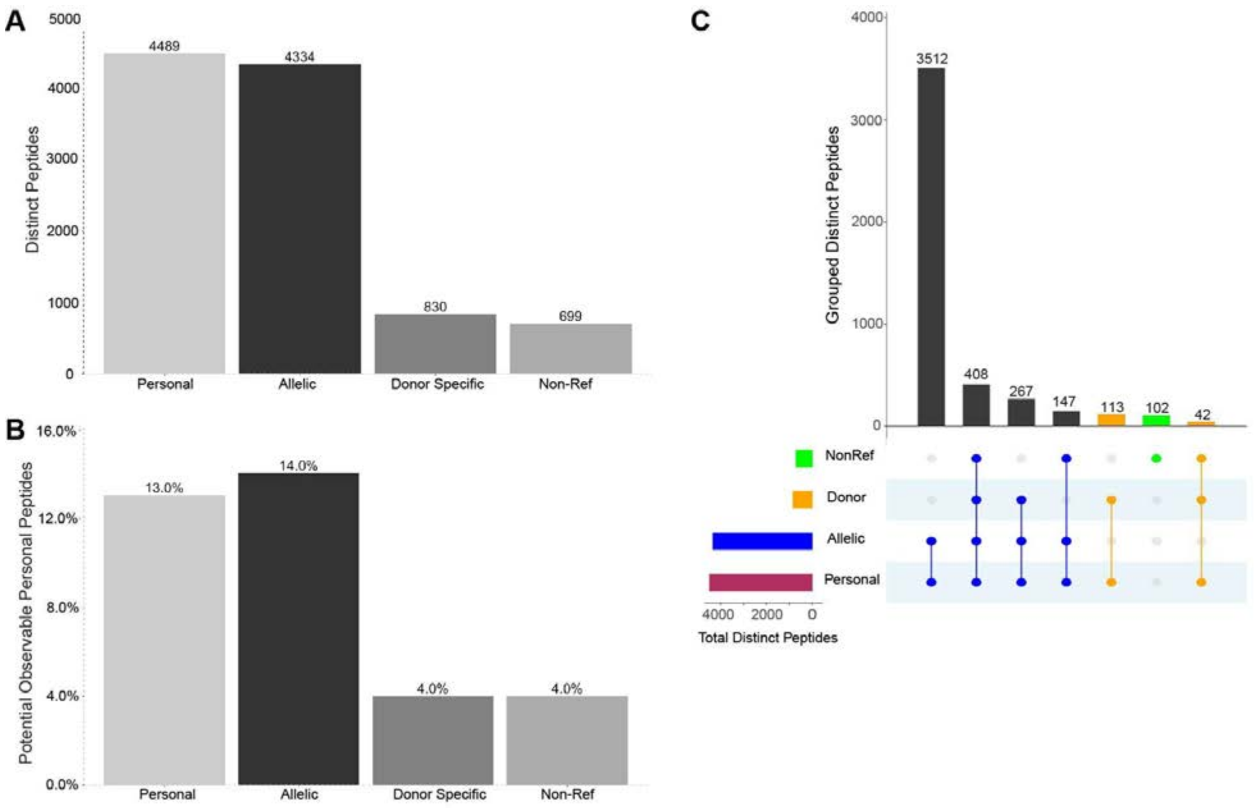


Data S3. Personal Hi-C, related to STAR Methods "Data Stack" Section

(A) Example of reference-aligned genome-wide Hi-C maps for the skeletal muscle tissue of two individuals.

(B) Number of paired reads and number of contacts from reference-aligned genome-wide Hi-C contact maps.

(C) A/B compartment annotation of four individuals and two tissues for chromosome 1. Red indicates that the 1 MB region is in the A compartment, whereas blue indicates that the region is in the B compartment. The dark blue band corresponds to the centromere.
(D) A/B compartments cluster based on tissue in autosomes or sex in chromosome X (chrX). In the figure, 37 male, 54 male, 51 female, and 53 female correspond to individuals 1, 2, 3, and 4, respectively.
(E) Summary of significant interactions determined by FitHiC2. The "Total" column presents the total number of intrachromosomal interactions for a given chromosome (e.g., chr1, chr2, …, chrY).
(F) - (G) Comparison of TopDom TAD calls for EN-TEx individuals and available Hi-C tissues. (F) TADs 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 - bottom) are shown. (G) Pair-wise comparison of TAD calls across all four individuals and available Hi-C tissue types. TAD calls were more similar (i.e., located at the same position along a chromosome) for the same tissues from different individuals than between different tissues.



D

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

Data S4. Personal proteomics, related to STAR Methods "Data Stack" Section

(A) Total number of significantly identified unambiguous personal peptides (filtered for 0.01 posterior error probability and unambiguous gene mapping). The personal category includes all types of personal peptides; allele-specific ("allelic") peptides are those that are specific to only one allele in at least one individual, donor-specific peptides are those that are completely absent in at least one of the four donors, and non-reference (non-ref) peptides are those that do not match the reference genome. Due to the use of the tandem mass tag (TMT) method, there is a bias towards the most common peptide from among the four donors (usually the reference peptide), as TMT boosts the signal for common peptides.

(B) 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 lengths of 6-60, there will be a vast number of unobservable peptides due to mass spectrometry (MS)-incompatible charge states and chemical properties.

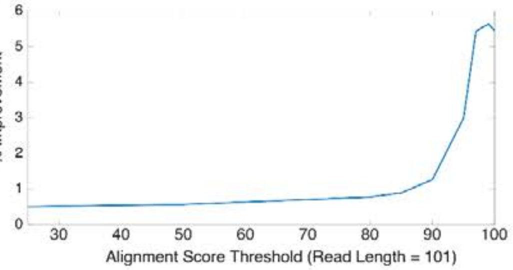
(C) UpSet plot showing the overlap in distinct peptides identified in the MS experiments among the four personal peptide categories. The majority of the peptides are considered personal because they are either donor-specific, AS, or both. There are a small number of non-reference peptides that are neither AS nor donor-specific; these are variant peptides that are common across all donors and alleles.

(D) Novel peptides. These are novel non-variant peptides identified with very high confidence that do not match known protein annotations in the GENCODE reference at the time of the experiment. These peptides were manually curated by GENCODE and annotated in combination with orthogonal evidence. Column 1 is the sequence of each novel peptide. Column 2 is the type of annotational outcome: NewModel, a new protein was annotated in the GENCODE reference (nonATG means the new model did not have a canonical ATG TSS); AddedModel, an existing protein annotation was adjusted or the peptide provided additional support for an annotation change already in progress; REFerror, although peptides matched genuine unannotated proteins, an underlying error in the reference genome GRCh38 sequence assembly meant that they could not be added to the current GENCODE annotation; Multiple-Variants, the peptide could be potentially explained by complex variants; IG Allelism, peptides matched in highly variable genomic regions; AltAssembly, peptides matched alternative genome assemblies; Some Evidence, peptides had some orthogonal evidence but fell below the current criteria for annotation of the new protein model; Weak Evidence, very little orthogonal evidence to support change in annotation; No Evidence, no support for a change in annotation. The affected genes and their Ensembl IDs are listed in columns 3 and 4. Additional information, e.g., the number of spectral reads matching the peptide, are provided in ancillary files in the EN-TEx portal.

A Precision Mapping Table

	Unique to Reference	Unique to Haplotypes	Gain= ((Hap1 ∪ Hap2)-Ref)/Ref
RNAseq	0.001	0.020	0.019
DNAseq	0.002	0.045	0.045
CHIPseq	0.001	0.016	0.015
Hi-C	0.001	0.024	0.023

C



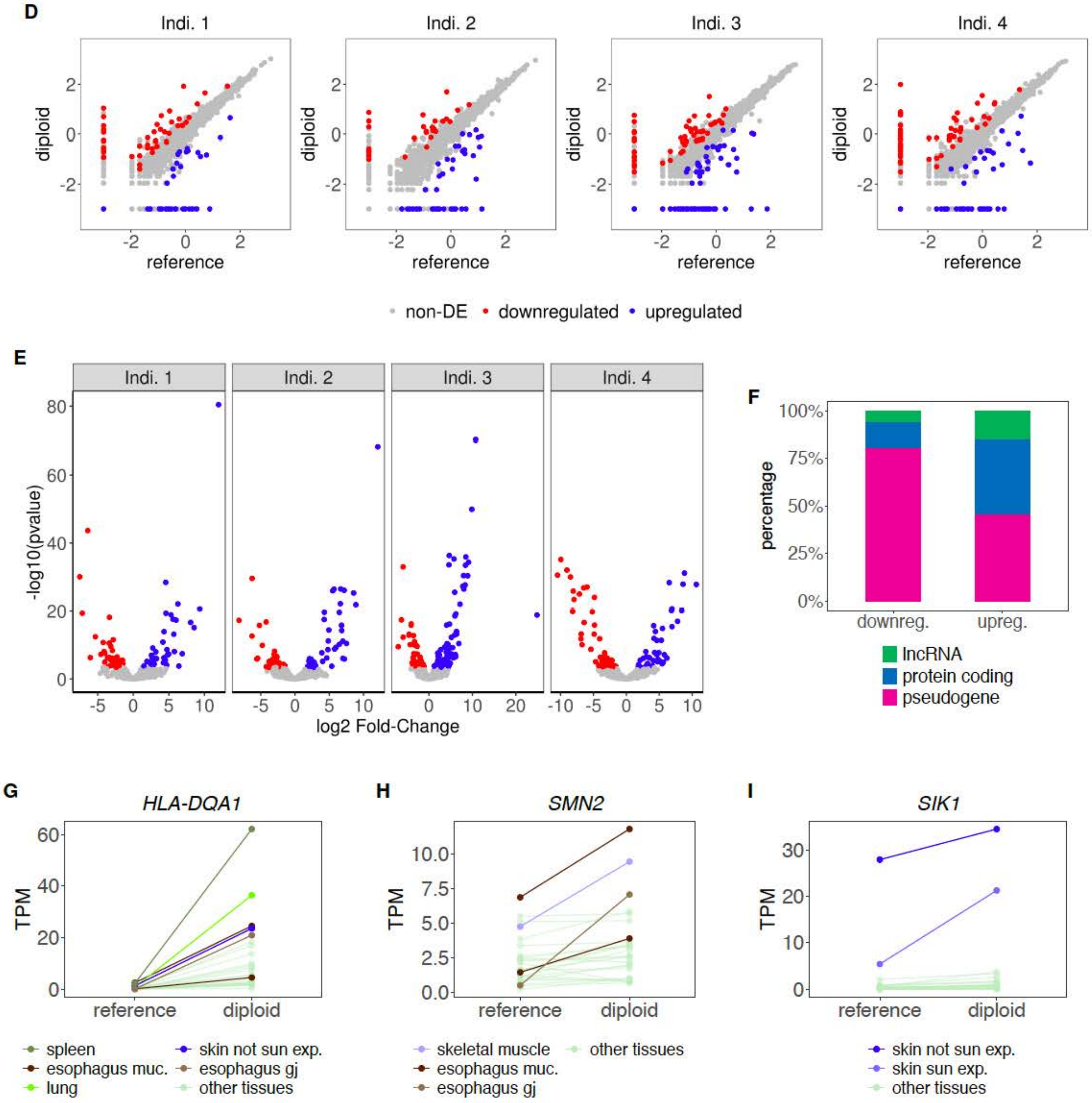
B

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 (%)	0.2	0.2	0.3	0.2	0.2
Hap1&&Hap2¬Ref (reads)	9,842,890	11,359,688	10,864,254	12,883,160	11,196,768
Hap1&&Hap2¬Ref (%)	4.4	4.4	4.5	4.8	4.5
Improvement	0.044	0.043	0.044	0.049	0.045

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 (%)	0.1	0.2	0.2	0.1	0.1
Hap1&&Hap2¬Ref (reads)	4,479,970	15,707,837	5,583,723	4,436,038	7,551,892
Hap1&&Hap2¬Ref (%)	2.4	2.4	2.6	2.4	2.4
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 (%)	0.1	0.1	0.1	0.1	0.1
Hap1&&Hap2¬Ref (reads)	267,944	491,961	212,830	122,841	273,894
Hap1&&Hap2¬Ref (%)	1.6	1.6	1.7	1.6	1.6
Improvement	0.015	0.015	0.016	0.015	0.015

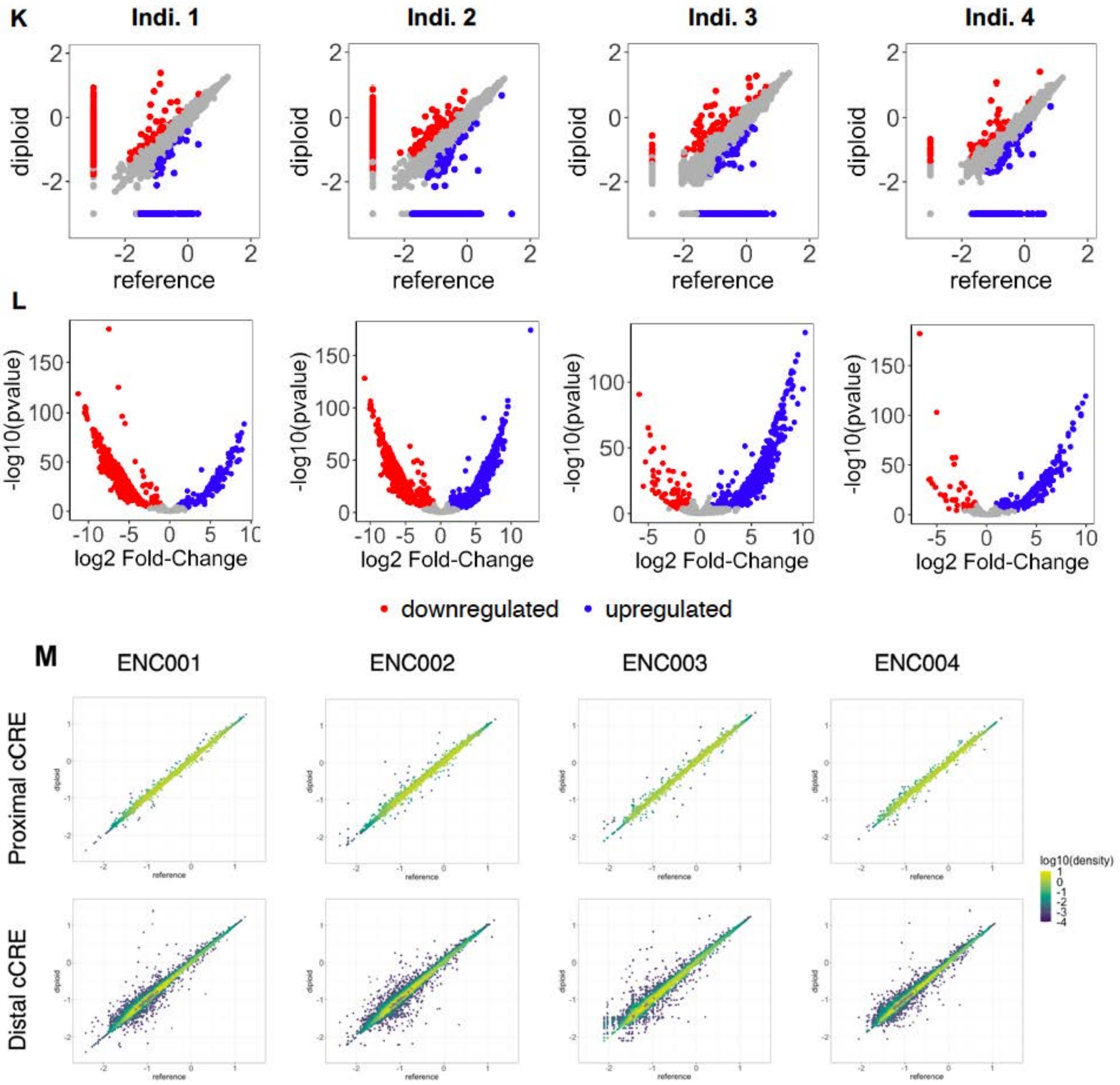
RNAseq	**ENC-001 37yo male**	**ENC-002 54 yo male**	**ENC-003 53 yo female**	**ENC-004 51 yo female**	**Average**
Hap.1(reads)	14,512,594	38,983,875	39,631,595	38,983,875	33,027,984.750
Hap.2(reads)	14,508,357	39,004,820	39,660,757	39,004,820	33,044,688.500
Reference (reads)	14,252,554	38,639,784	39,294,659	38,639,784	32,706,695.250
Ref (reads)	14,053	40,126	58,062	40,126	38,091.750
Ref only(%)	0.1	0.1	0.1	0.1	0.1
Hap1&&Hap2¬Ref (reads)	383,723	721,245	724,982	721,245	637,798.750
Hap1&&Hap2¬Ref (%)	2.6	1.8	1.8	1.8	2.0
Improvement	0.025	0.018	0.017	0.018	0.019

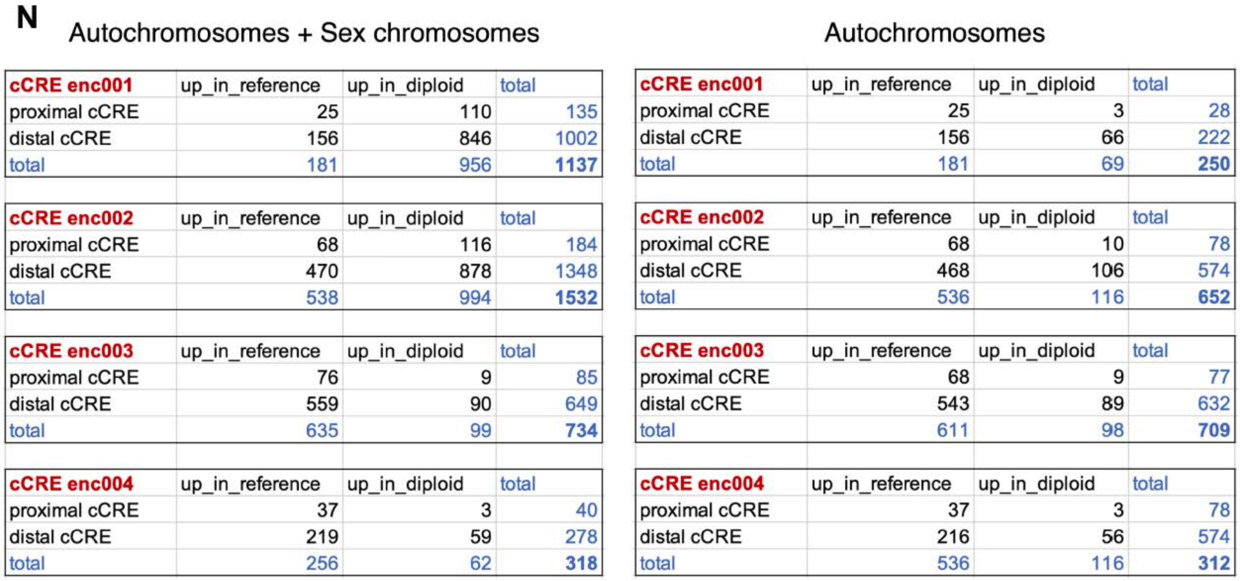


J

Experiment	# Downregulated	# Upregulated
enc001-ENCSR276MMH-adrenal_gland	5	9
enc001-ENCSR429EWK-thoracic_aorta	14	18
enc001-ENCSR532LJV-thyroid_gland	20	12
enc001-ENCSR853BNH-gastrocnemius_medialis	15	16
enc002-ENCSR023ZXN-thyroid_gland	26	27
enc002-ENCSR94PZB-adrenal_gland	6	15
enc003-ENCSR504QMK-ENCSR226KML-liver	14	21
total	59 (25 new)	53 (18 new)

Experiment	# Downregulated	# Upregulated
ENCSR000AED/AEF/AEG/AEH/COQ-GM12878	43 (34 new)	46 (31 new)


K
Indi. 1
Indi. 2
Indi. 3
Indi. 4
diploid
reference
2
0
-2
L
-log10(pvalue)
log2 Fold-Change
150
100
50
0
-10
-5
0
5
10
downregulated
upregulated
M
ENC001
ENC002
ENC003
ENC004
Proximal cCRE
Distal cCRE
diploid
reference
log10(density)
1
0
-1
-2
-3
-4



N

Autochromosomes + Sex chromosomes

cCRE enc001	up_in_reference	up_in_diploid	total
proximal cCRE	25	110	135
distal cCRE	156	846	1002
total	181	956	**1137**

cCRE enc002	up_in_reference	up_in_diploid	total
proximal cCRE	68	116	184
distal cCRE	470	878	1348
total	538	994	**1532**

cCRE enc003	up_in_reference	up_in_diploid	total
proximal cCRE	76	9	85
distal cCRE	559	90	649
total	635	99	**734**

cCRE enc004	up_in_reference	up_in_diploid	total
proximal cCRE	37	3	40
distal cCRE	219	59	278
total	256	62	**318**

Autochromosomes

cCRE enc001	up_in_reference	up_in_diploid	total
proximal cCRE	25	3	28
distal cCRE	156	66	222
total	181	69	**250**

cCRE enc002	up_in_reference	up_in_diploid	total
proximal cCRE	68	10	78
distal cCRE	468	106	574
total	536	116	**652**

cCRE enc003	up_in_reference	up_in_diploid	total
proximal cCRE	68	9	77
distal cCRE	543	89	632
total	611	98	**709**

cCRE enc004	up_in_reference	up_in_diploid	total
proximal cCRE	37	3	78
distal cCRE	216	56	574
total	536	116	**312**

Data S5. Mapping to personal genomes, related to Figure S2 and STAR Methods “Reference Comparison” Section

(A) Summary of percentages with precision mapping.

(B) Using DNA from transverse colon tissues, we constructed both haplotype sequences for each individual. Here, we show the summary statistics when comparing the mapping efficiency across different assays (DNA-seq, Hi-C, ChIP-seq, and RNA-seq) between mapping to haplotypes and to the reference genome for each individual. For the mapping of raw reads, stringent filtering criteria were applied (2 mismatches, Qm = 255 and Q > 30). Improvement was calculated as ((Haplotype1 U Haplotype 2)-Reference)/Reference.

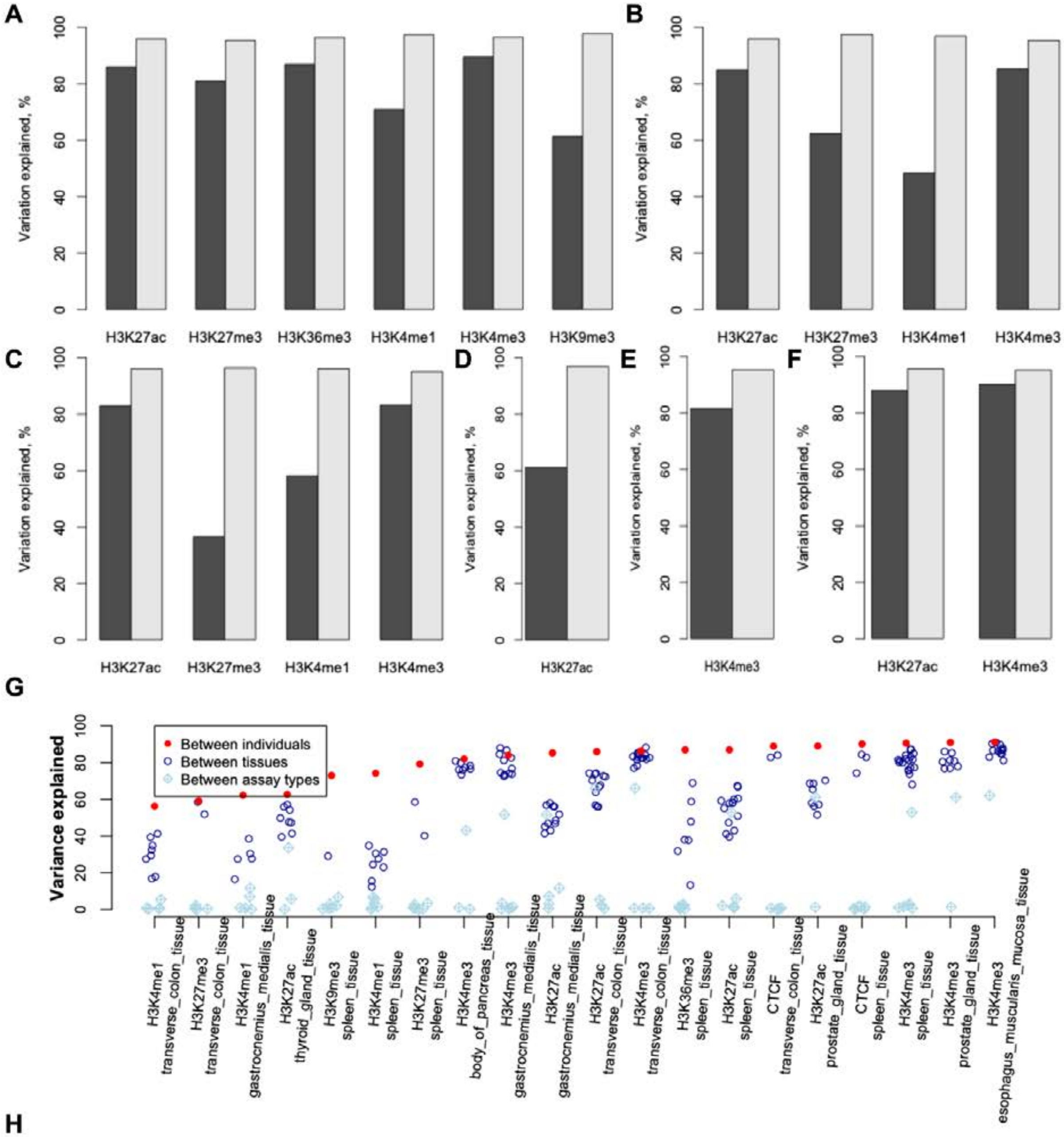
(C) Change in improvement as a function of alignment score threshold for Hi-C data. More stringent criteria yield greater improvement. However, in the alignment score thresholds that are commonly used in pipelines (e.g., 80-100 for a 101 bp read length), we observed improvements from 1.5% to 4%.

(D) Scatterplots reporting gene expression quantifications obtained after mapping to the reference genome (x-axis) and the diploid genomes (y-axis) for each of the four individuals. Expression values are reported as log10(TPM + 0.001) and correspond to the median value across tissues. Genes that are significantly differentially expressed between mapping to reference and personal genomes are color-coded (DESeq2, adjusted p-value Benjamini–Hochberg < 0.1 and |log2 FC| > 1; see STAR Methods “Reference Comparison” Section). Downregulated genes (red) are more expressed when mapped to the personal genome; upregulated genes (blue) are more expressed when mapped to the reference genome. Genes that are not differentially expressed are shown in gray.

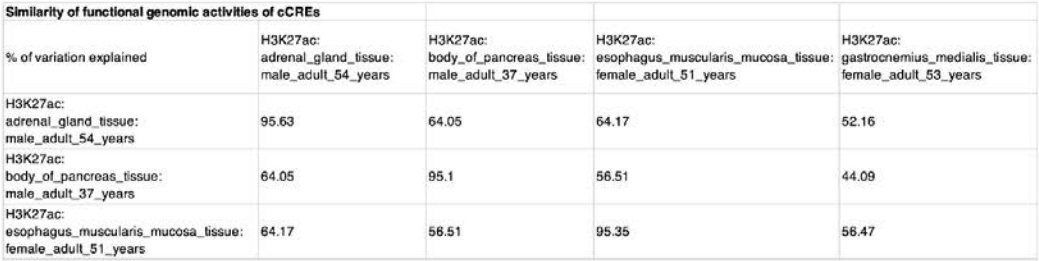
(E) Volcano plots reporting, for every gene, the log2(fold-change) and the -log10(p-value) obtained from the DGE analysis with DESeq2 in each of the four individuals. The same color schema as in (D) is applied.

(F) Barplot showing the gene type for the union of downregulated (n = 100) and upregulated (n = 112) genes across the four individuals.

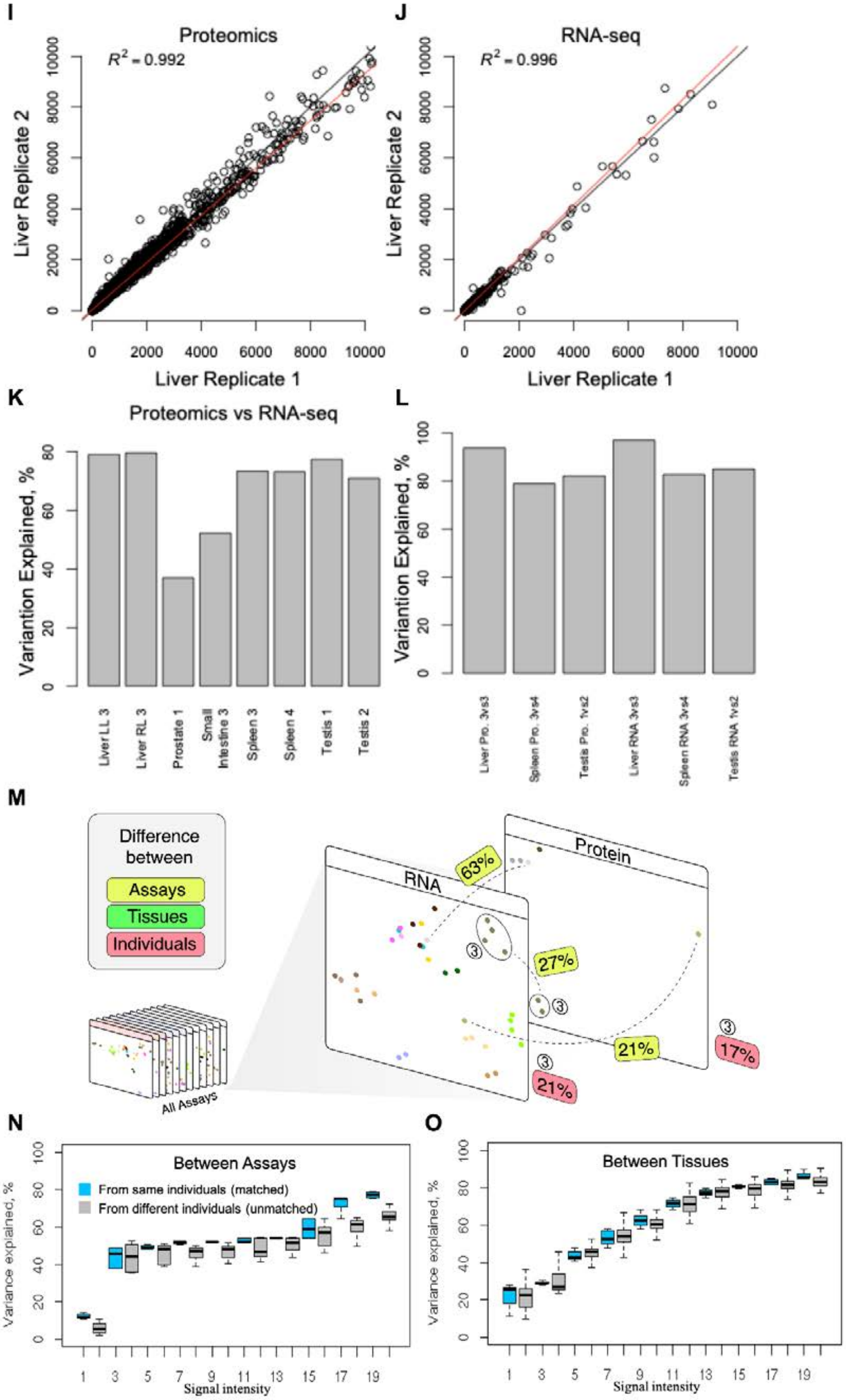
(G) - (I) Examples of downregulated genes (gene expression levels correspond to individual 3). *HLA-DQA1* belongs to the HLA class II alpha-chain paralogues. *SMN2* belongs to the SMN complex and plays a role in pre-mRNA splicing. This gene is part of an inverted duplication on chromosome 5q13, a region prone to rearrangements and deletions. Mutations in this gene have been associated with spinal muscular atrophy [1,2]. *SIK1* encodes a member of the salt-inducible kinase family, which has been associated with pigment gene expression [3]. Mutations in this gene have been associated with neurodevelopmental impairments [4,5].
(J) Upper panel: numbers of downregulated and upregulated genes identified by DESeq2 DGE analysis in each experiment with two available replicates. ENCODE experiment identifiers are reported in the first column. Lower panel: numbers of downregulated and upregulated genes identified by DESeq2 DGE analysis in the GM12878 cell line. This analysis employed ten replicates (five polyA+ RNA-seq experiments with two biological replicates each). The ENCODE experiment identifiers are: ENCSR000AED, ENCSR000AEF, ENCSR000AEG, ENCSR000AEH, and ENCSR000COQ.
(K) Differential cCRE activity. Scatterplots report H3K27ac activity obtained after mapping to the reference genome (x-axis) and the diploid genomes (y-axis) for each of the four individuals. The H3K27ac signals of the cCREs are log-transformed and correspond to the median value across tissues. cCREs that are significantly differentially marked between mapping to the reference and personal genomes are color-coded (DESeq2, adjusted p-value Benjamini–Hochberg < 0.1 and |log2 FC| > 1; see STAR Methods “Reference Comparison” Section). Downregulated cCREs (red) are more marked when mapped to the personal genome; upregulated cCREs (blue) are more marked when mapped to the reference genome. cCREs that are not differentially marked are shown in gray.
(L) Volcano plots reporting, for every cCRE, the log2(fold-change) and the -log10(p-value) obtained from the DESeq2 analysis in each of the four individuals. The same color schema as in (K) is applied.
(M) Scatterplots reporting quantifications of proximal and distal cCRE H3K27ac activity obtained after mapping to the reference genome (x-axis) and the diploid genomes (y-axis) for each of the four individuals. H3K27ac activity values are log-transformed and correspond to the median value across tissues. The density of cCREs in each scatterplot is color-coded. Note that distal cCREs appear to be more differentially marked between the reference and personal genomes, compared to proximal cCREs.
(N) The left and right panels show the numbers of differentially marked cCREs (total, proximal, and distal) located on all chromosomes and only auto-chromosomes, respectively.



H



Similarity of functional genomic activities of cCREs				
% of variation explained	H3K27ac: adrenal_gland_tissue: male_adult_54_years	H3K27ac: body_of_pancreas_tissue: male_adult_37_years	H3K27ac: esophagus_muscularis_mucosa_tissue: female_adult_51_years	H3K27ac: gastrocnemius_medialis_tissue: female_adult_53_years
H3K27ac: adrenal_gland_tissue: male_adult_54_years	95.63	64.05	64.17	52.16
H3K27ac: body_of_pancreas_tissue: male_adult_37_years	64.05	95.1	56.51	44.09
H3K27ac: esophagus_muscularis_mucosa_tissue: female_adult_51_years	64.17	56.51	95.35	56.47


I
Proteomics
$R^2 = 0.992$
Liver Replicate 2
Liver Replicate 1
0
2000
4000
6000
8000
10000
J
RNA-seq
$R^2 = 0.996$
Liver Replicate 2
Liver Replicate 1
K
Proteomics vs RNA-seq
Variantion Explained, %
0
20
40
60
80
Liver LL 3
Liver RL 3
Prostate 1
Small Intestine 3
Spleen 3
Spleen 4
Testis 1
Testis 2
L
Variantion Explained, %
0
20
40
60
80
100
Liver Pro. 3vs3
Spleen Pro. 3vs4
Testis Pro. 1vs2
Liver RNA 3vs3
Spleen RNA 3vs4
Testis RNA 1vs2
M
Difference between
Assays
Tissues
Individuals
RNA
Protein
63%
27%
21%
21%
17%
3
All Assays
N
Between Assays
From same individuals (matched)
From different individuals (unmatched)
Variance explained, %
Signal intensity
1
3
5
7
9
11
13
15
17
19
O
Between Tissues
Variance explained, %
Signal intensity

Data S6. Variation explained or similarity between experiments, related to Figure S2 and STAR Methods “Variation Analysis” Section

(A) - (F) 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 experiment we identified the cCREs that drive high variation explained (> 95%) in the replicates of the experiment. The intersecting set of cCREs from the two experiments was used to calculate the variation explained between the two experiments (black bars; e.g., 87% for the two H3K27ac experiments). The average variation explained between the replicates 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).

(G) Similarity between the signals of two functional genomic experiments. For each cCRE, the signal of a functional genomic experiment was measured by the average fold-change over control across the cCRE region. For two experiments, linear regression was 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.

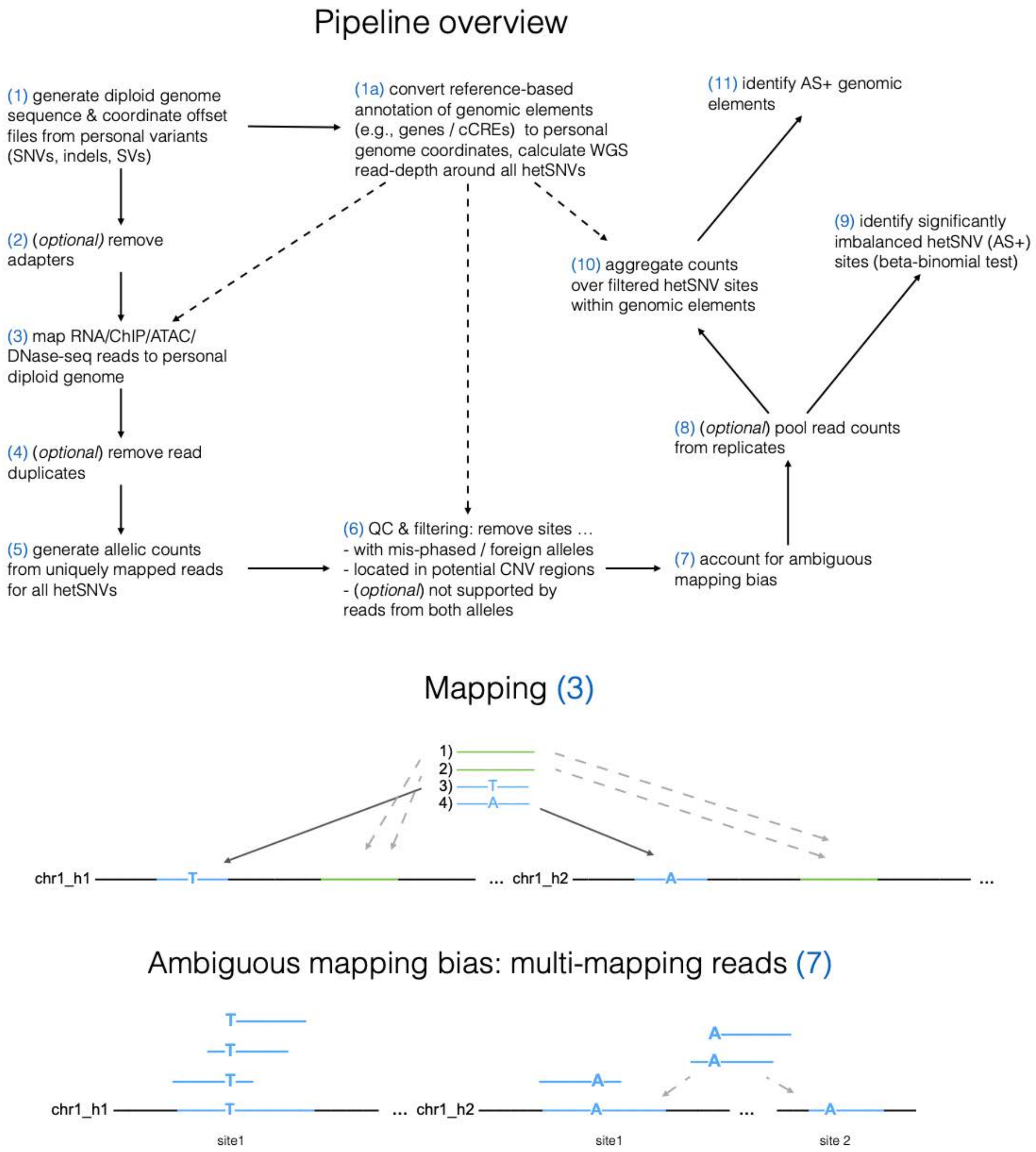
(H) The similarity between all possible pairs of experiments.

(I) - (L) Comparison between RNA-seq and proteomics data. (I) The normalized protein abundances are highly consistent between replicates. (J) This consistency is also observed for the normalized RNA abundances. (K) 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 RL indicates the right lobe. The numbers in the labels indicate the donors. (L) The variation explained between donors for protein abundances and RNA abundances, which was higher than those in (K). The normalized proteomics and RNA-seq data matrix used for panels (K) and (L) is available in the ancillary files.

(M) Consistently analyzing functional genomic data across individuals, tissues, and assays. Similar to Figure S2C, we used JIVE to project the RNA-seq and proteomics data to a 2D space and compare the variation of the data. For example, the difference of RNA-seq data in spleen between different individuals is 21%.

(N) - (O) Comparing the explained variances calculated using matched and unmatched data. (N) The explained variance is calculated between two types of histone modifications (between assays) using matched data (blue), i.e., the two assays are generated from the same tissue of an individual, or unmatched data (gray), i.e., the two assays are generated from the same tissues of different individuals. (O) The explained variance is calculated between the same histone modifications generated from two different tissues (between tissues). The tissues that are from the same individuals are referred to as matched.

A



Pipeline overview

(1) generate diploid genome sequence & coordinate offset files from personal variants (SNVs, indels, SVs)

(1a) convert reference-based annotation of genomic elements (e.g., genes / cCREs) to personal genome coordinates, calculate WGS read-depth around all hetSNVs

(2) (*optional)* remove adapters

(3) map RNA/ChIP/ATAC/DNase-seq reads to personal diploid genome

(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

(8) (*optional*) pool read counts from replicates

(9) identify significantly imbalanced hetSNV (AS+) sites (beta-binomial test)

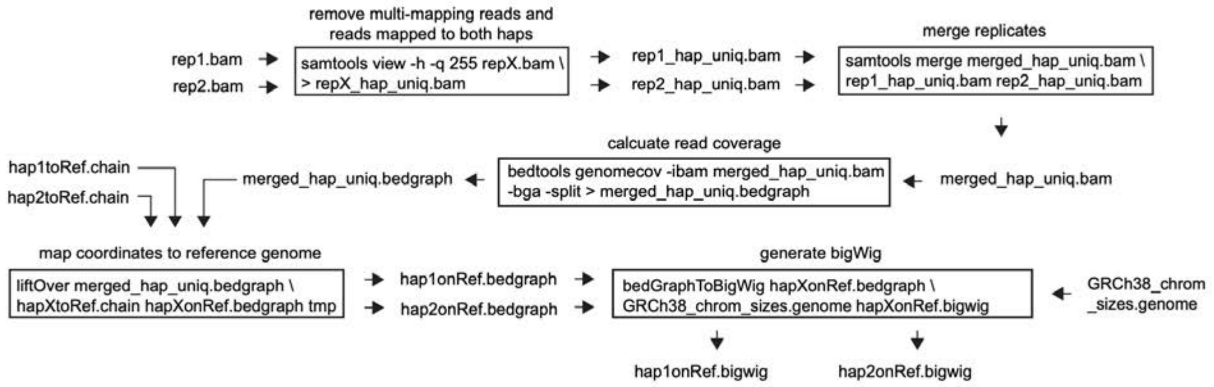
(10) aggregate counts over filtered hetSNV sites within genomic elements

(11) identify AS+ genomic elements

Mapping (3)

Ambiguous mapping bias: multi-mapping reads (7)

B



```
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
```

C

Fig. S4A	RNA: ENCFF660SLV, ENCFF751QEC CTCF: ENCFF296YDQ, ENCFF255INZ H3K27ac: ENCFF184LPK, ENCFF789APL, ENCFF707VEV, ENCFF298AKE
Fig. 2A Data S14H	ind3 RNA: ENCFF281PBY, ENCFF760KXM ind3 H3K27ac: ENCFF699EFW, ENCFF075RQB ind3 H3K27me3: ENCFF888EIC, ENCFF595OTK, ENCFF011LXD, ENCFF626DTV ind4 RNA: ENCFF711JSM, ENCFF355UJC, ENCFF415QZI, ENCFF912BRJ ind4 H3K27ac: ENCFF089KJG, ENCFF908MFI, ENCFF912LCB, ENCFF431PJB ind4 H3K27me3: ENCFF417VAA, ENCFF876PUF, ENCFF992SRG, ENCFF463QBH
Fig. 2D Data S17C	RNA: ENCFF719MSG, ENCFF120MML, ENCFF337ZBN, ENCFF481IQE H3K27ac: ENCFF339ODV, ENCFF870TZH TF binding clusters: UCSC encRegTfbsClustered
Fig. 2E Data S17G	RNA: ENCFF038JEE, ENCFF897TAN H3K27ac: ENCFF143SOY, ENCFF244ISL, ENCFF804MSF, ENCFF976BRQ TF binding clusters: UCSC encRegTfbsClustered
Fig. S4C Data S17E	ind3 RNA: ENCFF534JLO ind3 H3K27ac: ENCFF066DSD ind3 CTCF: ENCFF417IMY ind2 RNA: ENCFF232DNA ind2 H3K27ac: ENCFF439NXI ind2 CTCF: ENCFF178GEC TF binding clusters: UCSC encRegTfbsClustered
Fig. S4B Data S17K Data S17L	ind3 RNA: ENCFF216VOH ind3 H3K9me3: ENCFF423DVX ind3 long-read RNA: ENCFF185VYD ind2 RNA: ENCFF187KAR ind2 H3K27ac: ENCFF095CZX ind2 long-read RNA: ENCFF912HPY
Data S17A	RNA: ENCFF326CGI, ENCFF663VCC H3K27ac: ENCFF935UTO, ENCFF653PKW, ENCFF235IVE, ENCFF226YFN ATAC: ENCFF591BAY, ENCFF332SCG CTCF: ENCFF800GHL, ENCFF100YUK, ENCFF861WPS, ENCFF056JNV, ENCFF608GCT, ENCFF682AOT TF binding clusters: UCSC encRegTfbsClustered
Data S17H	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
Data S17I	RNA: ENCFF411WXY, ENCFF543BVT, ENCFF072VKD, ENCFF484BLA, ENCFF086TFZ, ENCFF351OAS H3K27ac: ENCFF214DHU, ENCFF209OKJ, ENCFF330KKH, ENCFF343NQH, ENCFF706KXN, ENCFF349JBL, ENCFF945XBP, ENCFF382QHO, ENCFF922CDY, ENCFF778KZF, ENCFF040XEO, ENCFF173QJF, ENCFF033YTT, ENCFF088QFN
Data S17J	ind1 RNA: ENCFF751LMQ, ENCFF188LMP, ENCFF151GUG, ENCFF628TMU ind1 H3K27ac: ENCFF259FFL, ENCFF570KZO, ENCFF176GZS ind4 RNA: ENCFF711JSM, ENCFF355UJC, ENCFF415QZI, ENCFF912BRJ ind4 H3K27ac: ENCFF930NCY, ENCFF089KJG, ENCFF908MFI, ENCFF912LCB, ENCFF431PJB
Data S17M	ind1 RNA: ENCFF150OQH, ENCFF460GHU, ENCFF661PVB, ENCFF152GUK ind2 RNA: ENCFF900BSQ, ENCFF184XUR ind3 RNA: ENCFF667RSP, ENCFF891NXQ
Data S17N	ind1 RNA: ENCFF060XKP ind1 H3K27ac: ENCFF665GNN ind2 RNA: ENCFF251QRT ind2 H3K27ac: ENCFF439ASM

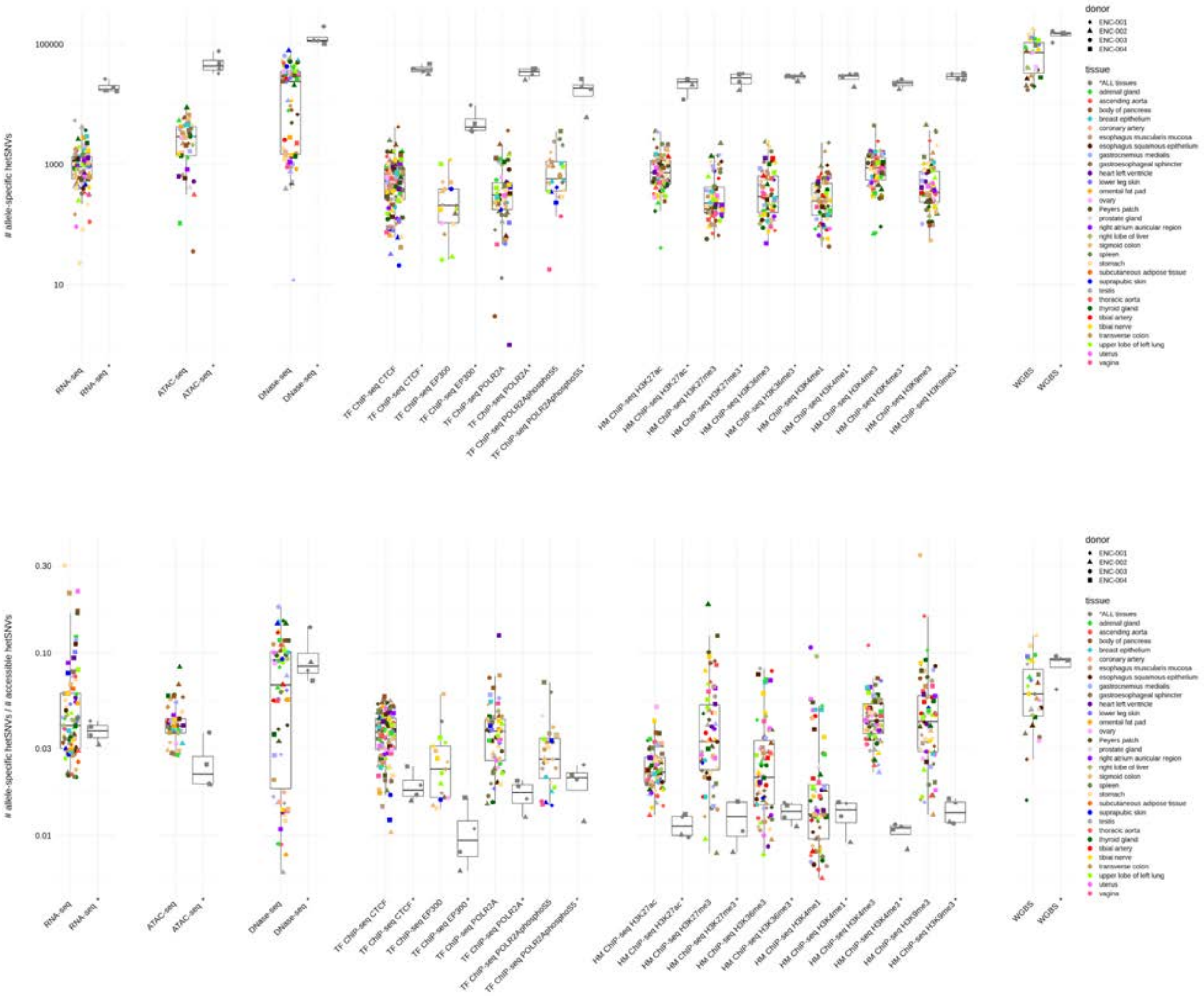
Data S7. AlleleSeq2 and haplotype-specific signal tracks, related to STAR Methods "AS Calling" Section

(A) Workflow of the AlleleSeq2 pipeline. We used genomic variants to construct the sequence and coordinates for each haplotype, and then mapped functional genomics assay reads to each haplotype by using the phased hetSNVs in a read (middle panel). This approach generated BAM files that contain both reads that are mapped uniquely to a region in a haplotype and reads that are mapped to multiple regions (within a haplotype or between two haplotypes, bottom panel). Since it is not possible to unambiguously identify the origin of the reads that multi-map within the haplotypes, we made the conservative assumption that all of these reads originate from the heterozygous locus and, unless the direction of the bias changed towards the opposite allele, we adjusted the allele counts including the multi-mapping reads. We pooled all of the reads for assays with replicates. We identified hetSNVs with allelic imbalance by performing a beta-binomial test on the allelic reads. To determine whether a genomic region has an allelic imbalance in RNA-seq, ChIP-seq, or ATAC-seq, we summed the AS reads from all hetSNVs within the region and performed a beta-binomial test.

(B) Workflow for generating haplotype-specific signal tracks. Each box is a command to process files. The input BAM files are generated by step (3) in panel (A), containing reads that are uniquely mapped to each haplotype and reads with ambiguous mapping. The first box removes the reads with ambiguous mapping. In this example, the assay has two replicates; therefore, we merge the two BAM files of unique reads (box 2). If an assay has no replicates, then we skip the "merge replicates" step. The read coverage of each chromosome in each haplotype is then calculated and stored in bedGraph files (box 3). Note that the coordinates of a given genomic region are in the personal genome; therefore, the two haplotypes can give different coordinates even for the same gene. To compare the same region between the two haplotypes, we convert the coordinates in each haplotype from the personal genome to the reference genome (box 4). We also convert the read coverage from bedGraph to bigWig, which can be plotted in the IGV Genome Viewer (box 5). A script that generates the haplotype-specific read coverage from BAM files is provided at https://github.com/gersteinlab/AlleleSeq2. An example of the intermediate files (except for the BAM files) in generating haplotype-specific signal tracks is available on the EN-TEx portal.

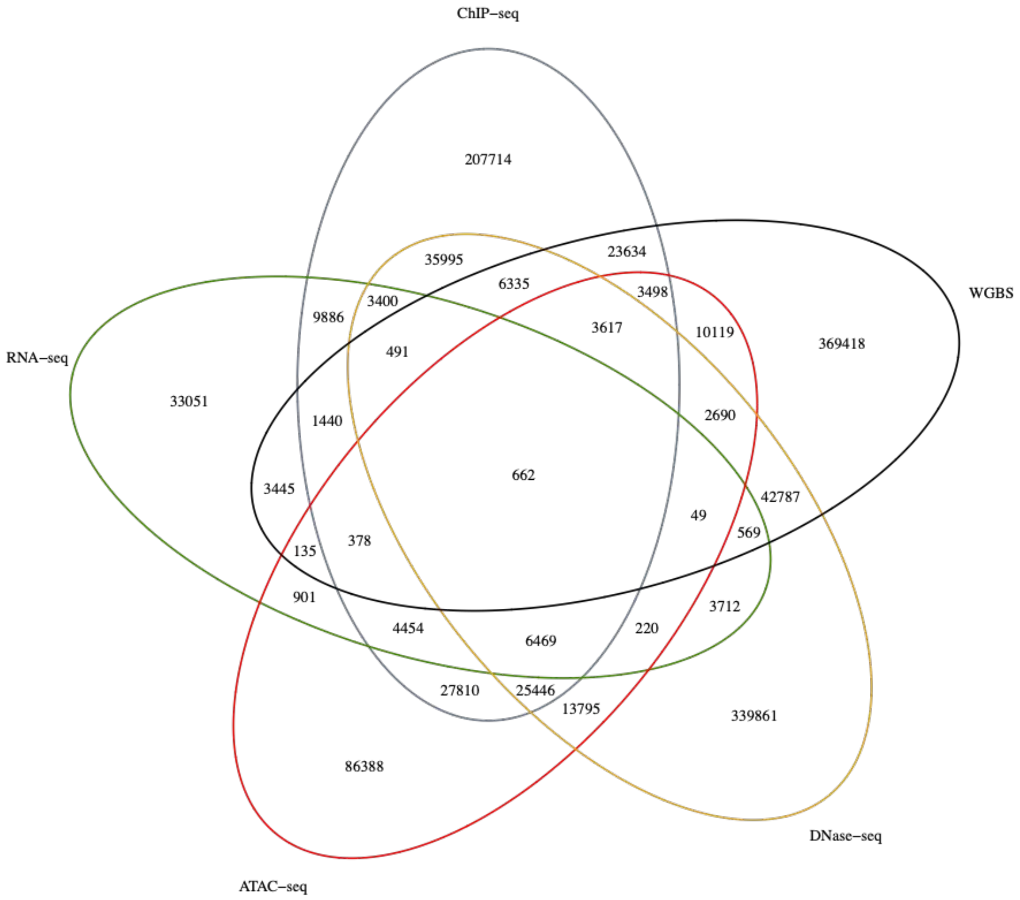
(C) Data used to generate signal tracks. Data in blue are given as the accession numbers in the ENCODE portal. TF binding clusters are available via the UCSC Table Browser.

A



# allele-specific hetSNVs
100000
1000
10
RNA-seq
RNA-seq *
ATAC-seq
ATAC-seq *
DNase-seq
DNase-seq *
TF ChIP-seq CTCF
TF ChIP-seq CTCF *
TF ChIP-seq EP300
TF ChIP-seq EP300 *
TF ChIP-seq POLR2A
TF ChIP-seq POLR2A *
TF ChIP-seq POLR2Aphospho5S
TF ChIP-seq POLR2Aphospho5S *
HM ChIP-seq H3K27ac
HM ChIP-seq H3K27ac *
HM ChIP-seq H3K27me3
HM ChIP-seq H3K27me3 *
HM ChIP-seq H3K36me3
HM ChIP-seq H3K36me3 *
HM ChIP-seq H3K4me1
HM ChIP-seq H3K4me1 *
HM ChIP-seq H3K4me3
HM ChIP-seq H3K4me3 *
HM ChIP-seq H3K9me3
HM ChIP-seq H3K9me3 *
WGBS
WGBS *
donor
ENC-001
ENC-002
ENC-003
ENC-004
tissue
*ALL tissues
adrenal gland
ascending aorta
body of pancreas
breast epithelium
coronary artery
esophagus muscularis mucosa
esophagus squamous epithelium
gastrocnemius medialis
gastroesophageal sphincter
heart left ventricle
lower leg skin
omental fat pad
ovary
Peyers patch
prostate gland
right atrium auricular region
right lobe of liver
sigmoid colon
spleen
stomach
subcutaneous adipose tissue
suprapubic skin
testis
thoracic aorta
thyroid gland
tibial artery
tibial nerve
transverse colon
upper lobe of left lung
uterus
vagina
# allele-specific hetSNVs / # accessible hetSNVs
0.30
0.10
0.03
0.01
RNA-seq
RNA-seq *
ATAC-seq
ATAC-seq *
DNase-seq
DNase-seq *
TF ChIP-seq CTCF
TF ChIP-seq CTCF *
TF ChIP-seq EP300
TF ChIP-seq EP300 *
TF ChIP-seq POLR2A
TF ChIP-seq POLR2A *
TF ChIP-seq POLR2Aphospho5S
TF ChIP-seq POLR2Aphospho5S *
HM ChIP-seq H3K27ac
HM ChIP-seq H3K27ac *
HM ChIP-seq H3K27me3
HM ChIP-seq H3K27me3 *
HM ChIP-seq H3K36me3
HM ChIP-seq H3K36me3 *
HM ChIP-seq H3K4me1
HM ChIP-seq H3K4me1 *
HM ChIP-seq H3K4me3
HM ChIP-seq H3K4me3 *
HM ChIP-seq H3K9me3
HM ChIP-seq H3K9me3 *
WGBS
WGBS *
donor
ENC-001
ENC-002
ENC-003
ENC-004
tissue
*ALL tissues
adrenal gland
ascending aorta
body of pancreas
breast epithelium
coronary artery
esophagus muscularis mucosa
esophagus squamous epithelium
gastrocnemius medialis
gastroesophageal sphincter
heart left ventricle
lower leg skin
omental fat pad
ovary
Peyers patch
prostate gland
right atrium auricular region
right lobe of liver
sigmoid colon
spleen
stomach
subcutaneous adipose tissue
suprapubic skin
testis
thoracic aorta
thyroid gland
tibial artery
tibial nerve
transverse colon
upper lobe of left lung
uterus
vagina

B



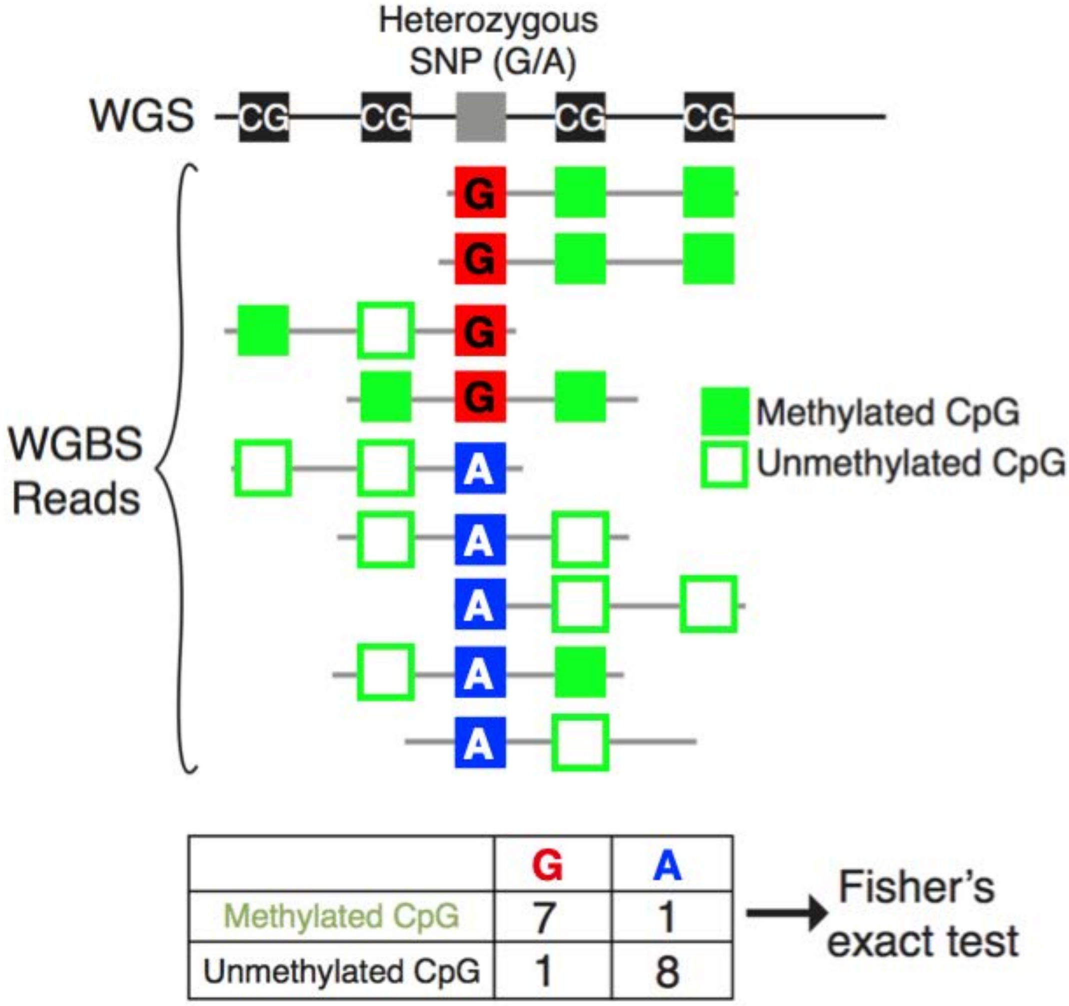
C

	all (avg.) variants in 4 individuals	unique variants in 4 individuals	unique/all
hetSNVs	9483K (2371K)	6023K	0.64
AS SNVs	1459K (365K)	1268K	0.87 (0.93)
ASM SNVs	516K (129K)	469K	0.91 (0.97)
ASE SNVs	79K (19.8K)	69K	0.87 (1.0)
ASB SNVs	412K (103)	361K	0.88 (0.98)
AS accessibility SNVs	682K (170.5K)	620K	0.91 (0.97)

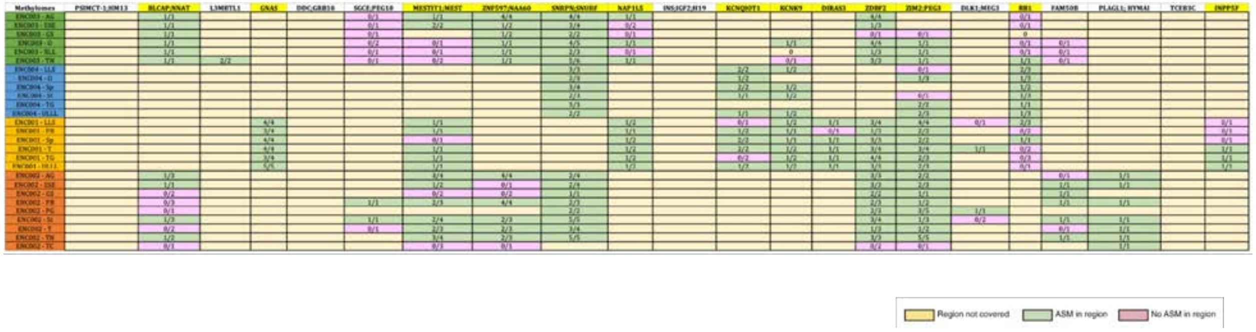
Data S8. AS SNVs, related to Figure S3 and STAR Methods "AS Calling" Section
(A) Distribution (top) and fraction (bottom) of the number of hetSNVs associated with AS behavior across different EN-TEx donors, tissues, and assays. The fraction is the number of hetSNVs associated with AS behavior relative to the number of accessible hetSNVs. Call sets based on pooled reads from all tissues for each donor and assay are shown in gray. An average of 820 AS events were detected in the RNA/ChIP/ATAC-seq samples (median 517, IQR 251-1,030; ~3.8% of the total number of accessible sites).

(B) Venn diagram of hetSNVs showing AS activity in different assays. The numbers are pooled from all four individuals. Heterozygous SNVs that appear in multiple individuals but have the same coordinates are collapsed into one unique heterozygous SNV.
(C) The total numbers of AS hetSNVs (i.e., no collapsing; numbers in parentheses are averages for an individual), the numbers of unique AS hetSNVs, and the unique-to-all ratios. For a comparison with the unique-to-all ratio of the AS SNVs, we randomly sampled the same number of heterozygous SNVs from the four individuals and calculated the unique-to-all ratio (numbers in parentheses). The smaller unique-to-all ratios of the AS SNVs suggest that these SNVs tend to be more common than random heterozygous SNVs. AS SNVs are those that show AS activity in any of the assays in (B). AS methylation (ASM) SNVs show AS activity in WGBS data, ASE SNVs in RNA-seq data, AS binding (ASB) SNVs in ChIP-seq data, and AS accessibility SNVs in ATAC-seq or DNase-seq data.

A



B



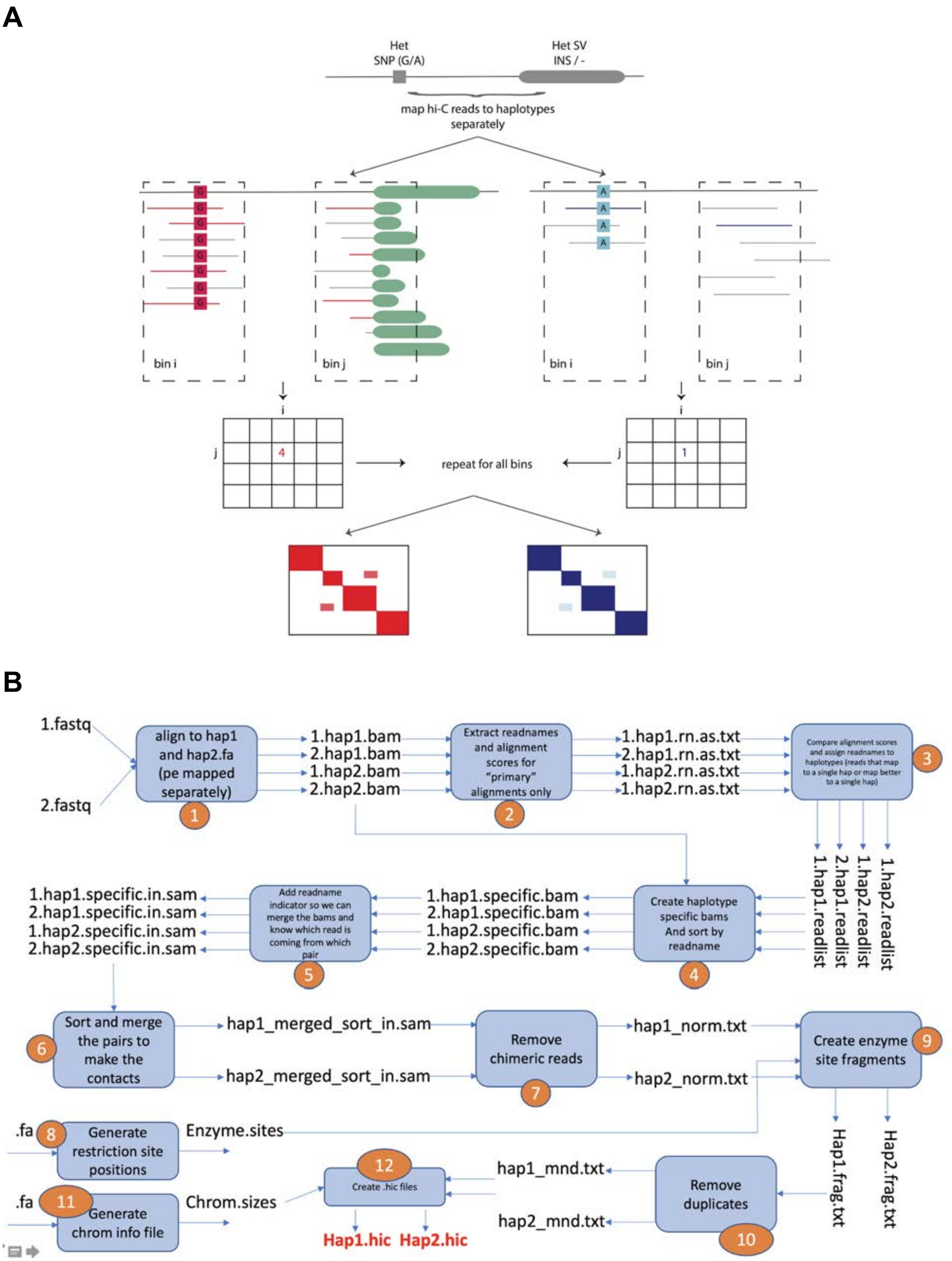
Methylomes	PSIMCT-1;HM13	BLCAP;NNAT	L3MBTL1	GNAS	DDC;GRB10	SGCE;PEG10	MESTIT1;MEST	ZNF597;NAA60	SNRPN;SNURF	NAP1L5	INS;IGF2;H19	KCNQ1OT1	KCNK9	DIRAS3	ZDBF2	ZIM2;PEG3	DLK1;MEG3	RB1	FAM50B	PLAGL1; HYMAI	TCEB3C	INPP5F
ENC003 - AG		1/1				0/1	1/1	4/4	4/4	1/1					4/4			0/1				
ENC003 - ESE		1/1				0/1	2/2	1/2	3/4	0/2					1/3			0/1				
ENC003 - GS		1/1				0/1		1/2	2/2	0/1					0/1	0/1		0				
ENC003 - G		1/1				0/2	0/1	1/1	4/5	1/1			1/1		4/4	1/1		0/1	0/1			
ENC003 - SLL		1/1				0/1	0/1	1/1	2/3	0/1			0		1/3	1/1		0/1	0/1			
ENC003 - TW		1/1	2/2			0/1	0/2	1/1	5/6	1/1			0/1		3/3	1/1		1/1	0/1			
ENC004 - LLE									3/3			2/2	1/2			0/1		2/3				
ENC004 - G									2/3			1/2				1/3		1/3				
ENC004 - Sp									3/4			2/2	1/2					1/2				
ENC004 - St									2/3			1/1	1/2			0/1		1/3				
ENC004 - TG									3/3							2/2		1/1				
ENC004 - ULLL									2/2			1/1	1/2			2/3		1/3				
ENC001 - LLE				4/4			1/1			1/2		0/1	1/2	1/1	3/4	4/4	0/1	2/3				0/1
ENC001 - FR				3/4			1/1			1/1		1/2	1/1	0/1	1/3	2/3		0/2				0/1
ENC001 - Sp				4/4			0/1			1/2		2/2	1/1	1/1	3/3	2/2		1/1				0/1
ENC001 - T				4/4			1/1			1/2		2/2	1/2	1/1	3/4	3/4	1/1	0/2				1/1
ENC001 - TG				3/4			1/1			1/2		0/2	1/2	1/1	4/4	2/3		0/3				1/1
ENC001 - ULLL				5/5			1/1			1/2		1/2	1/2	1/1	3/3	2/3		0/1				1/1
ENC002 - AG		1/3					3/4	4/4	2/4						3/3	2/2			0/1	1/1		
ENC002 - ESE		1/1					1/2	0/1	2/4						3/2	2/3			1/1	1/1		
ENC002 - GS		0/2					0/2	0/2	1/1						2/2	1/1			1/1			
ENC002 - FR		0/3				1/1	2/3	4/4	2/3						2/3	1/2			1/1	1/1		
ENC002 - PG		0/1							2/2						2/3	3/5	1/1					
ENC002 - St		1/3				1/1	2/4	2/3	5/5						3/4	1/3	0/2		1/1	1/1		
ENC002 - T		0/2				0/1	2/3	2/3	3/4						1/3	1/2			0/1	1/1		
ENC002 - TN		1/2					3/4	2/3	5/5						3/3	5/5			1/1	1/1		
ENC002 - TC		0/1					0/3	0/1							0/2	0/1				1/1		

🟨 Region not covered 🟩 ASM in region 🟥 No ASM in region

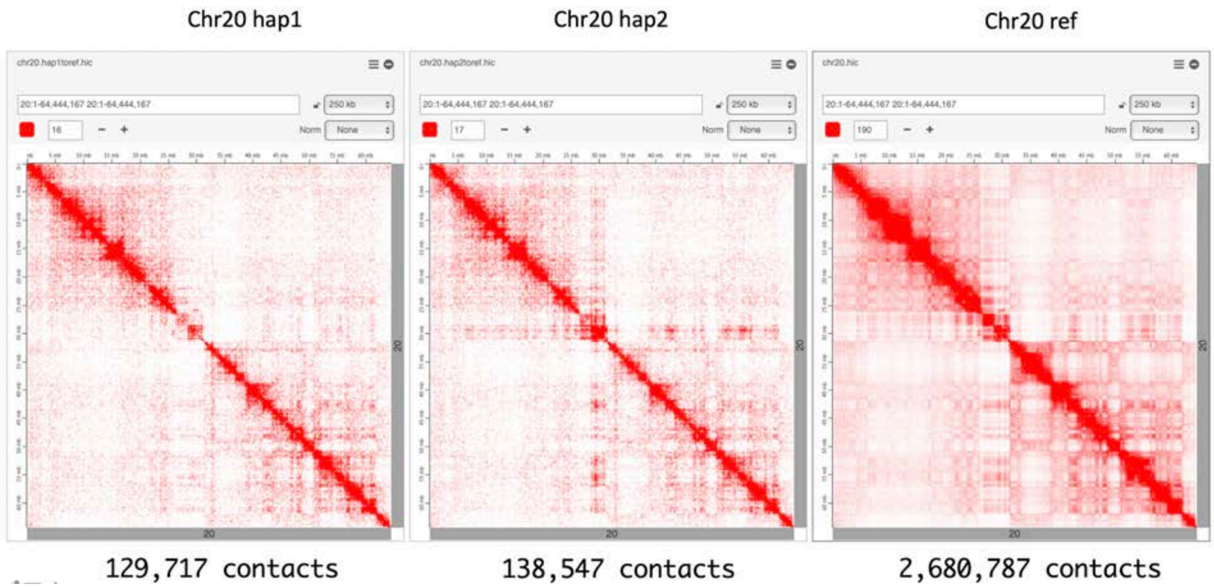
Data S9. Allele specific methylation, related to STAR Methods "AS Calling" Section
(A) Schematic showing how AS methylation is calculated. We can determine ASM by identifying AS methylated CpG sites near tag hetSNVs using the statistical test above. Since methylated cytosines are sequenced differently from unmethylated cytosines, we use a two-by-two

contingency test (Fisher’s exact test) in order to identify AS methylated CpGs in the vicinity of a tag hetSNV [6].

(B) ASM calls in known imprinting control regions. Number of ASM calls made out of the total number of accessible hetSNVs that overlap with an imprinting control region (ICR) [7] for each sample. Green cells represent ICRs that overlap with at least one ASM call in that sample. Red cells represent ICRs that overlap with at least one accessible hetSNV. No hetSNVs with a significant imbalance were observed. Yellow cells represent ICRs that did not overlap with any accessible hetSNVs in the sample.


A
Het
SNP (G/A)
Het SV
INS / -
map hi-C reads to haplotypes
separately
G
A
bin i
bin j
bin i
bin j
i
j
4
repeat for all bins
i
j
1
B
1.fastq
2.fastq
align to hap1
and hap2.fa
(pe mapped
separately)
1
1.hap1.bam
2.hap1.bam
1.hap2.bam
2.hap2.bam
Extract readnames
and alignment
scores for
"primary"
alignments only
2
1.hap1.rn.as.txt
2.hap1.rn.as.txt
1.hap2.rn.as.txt
1.hap2.rn.as.txt
Compare alignment scores and assign readnames to haplotypes (reads that map to a single hap or map better to a single hap)
3
1.hap1.readlist
2.hap1.readlist
1.hap2.readlist
1.hap2.readlist
Create haplotype
specific bams
And sort by
readname
4
1.hap1.specific.bam
2.hap1.specific.bam
1.hap2.specific.bam
2.hap2.specific.bam
Add readname indicator so we can merge the bams and know which read is coming from which pair
5
1.hap1.specific.in.sam
2.hap1.specific.in.sam
1.hap2.specific.in.sam
2.hap2.specific.in.sam
Sort and merge
the pairs to
make the
contacts
6
hap1_merged_sort_in.sam
hap2_merged_sort_in.sam
Remove
chimeric reads
7
hap1_norm.txt
hap2_norm.txt
Create enzyme
site fragments
9
Hap1.frag.txt
Hap2.frag.txt
.fa
8
Generate
restriction site
positions
Enzyme.sites
.fa
11
Generate
chrom info file
Chrom.sizes
12
Create .hic files
Hap1.hic
Hap2.hic
hap1_mnd.txt
hap2_mnd.txt
Remove
duplicates
10

C



chr20.hap1toref.hic
20:1-64,444,167 20:1-64,444,167 250 kb
16 Norm None

chr20.hap2toref.hic
20:1-64,444,167 20:1-64,444,167 250 kb
17 Norm None

chr20.hic
20:1-64,444,167 20:1-64,444,167 250 kb
190 Norm None

D

Individual/Tissue	Intra-chromosomal Interactions	Hap1	Hap2	Hap1 or Hap2	Significantly Imbalanced
ind1 skeletal muscle	39,013,901	4,049,203	4,034,602	7,041,417	577,728
ind2 skeletal muscle	4,405,480	1,117,328	1,146,381	2,072,227	140,317
ind3 skeletal muscle	40,412,585	4,345,533	4,359,297	7,493,069	574,836
ind4 skeletal muscle	41,569,344	4,028,293	4,021,800	6,983,660	523,931
ind1 transverse colon	45,534,793	4,942,660	4,924,000	8,574,917	702,953
ind2 transverse colon	25,548,308	2,148,267	2,151,803	3,842,621	261,752
ind3 transverse colon	43,917,995	4,716,549	4,722,227	8,118,858	609,973
ind4 transverse colon	43,406,680	4,343,051	4,334,617	7,506,125	583,468

Data S10. Haplotype-specific Hi-C, related to STAR Methods “AS Calling” Section

(A) Schematic showing the overall methodology for determining haplotype-specific 3D contact interactions using Hi-C paired-end reads.

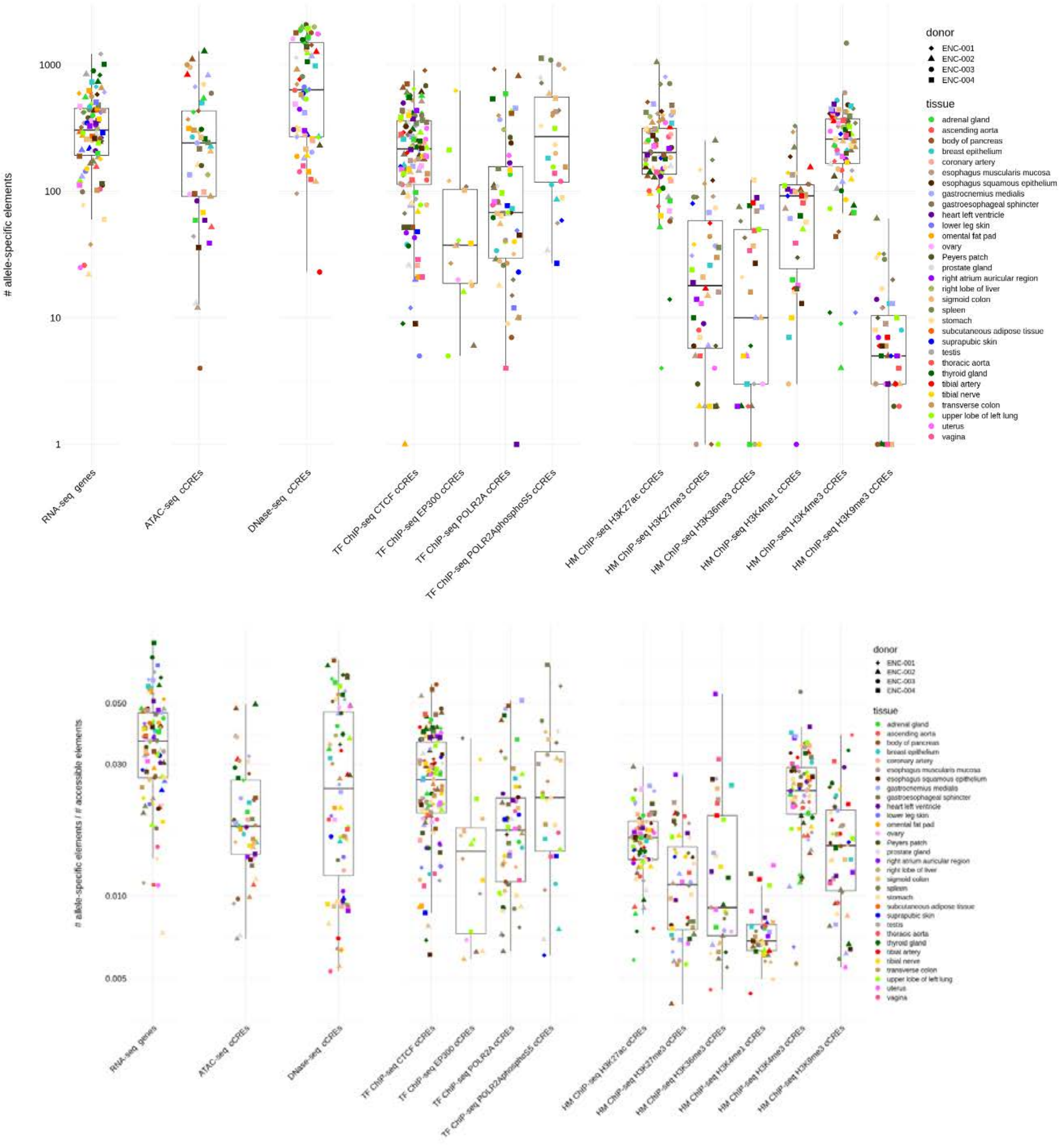
(B) Workflow for the generation of haplotype-specific Hi-C contact maps.

(C) Haplotype-specific contact maps for Chr20 generated using the personal genome coordinates. The third map is the bulk Hi-C contact map of Chr20 generated using the reference genome.

(D) Number of Hi-C contacts obtained from haplotype-specific Hi-C contact maps

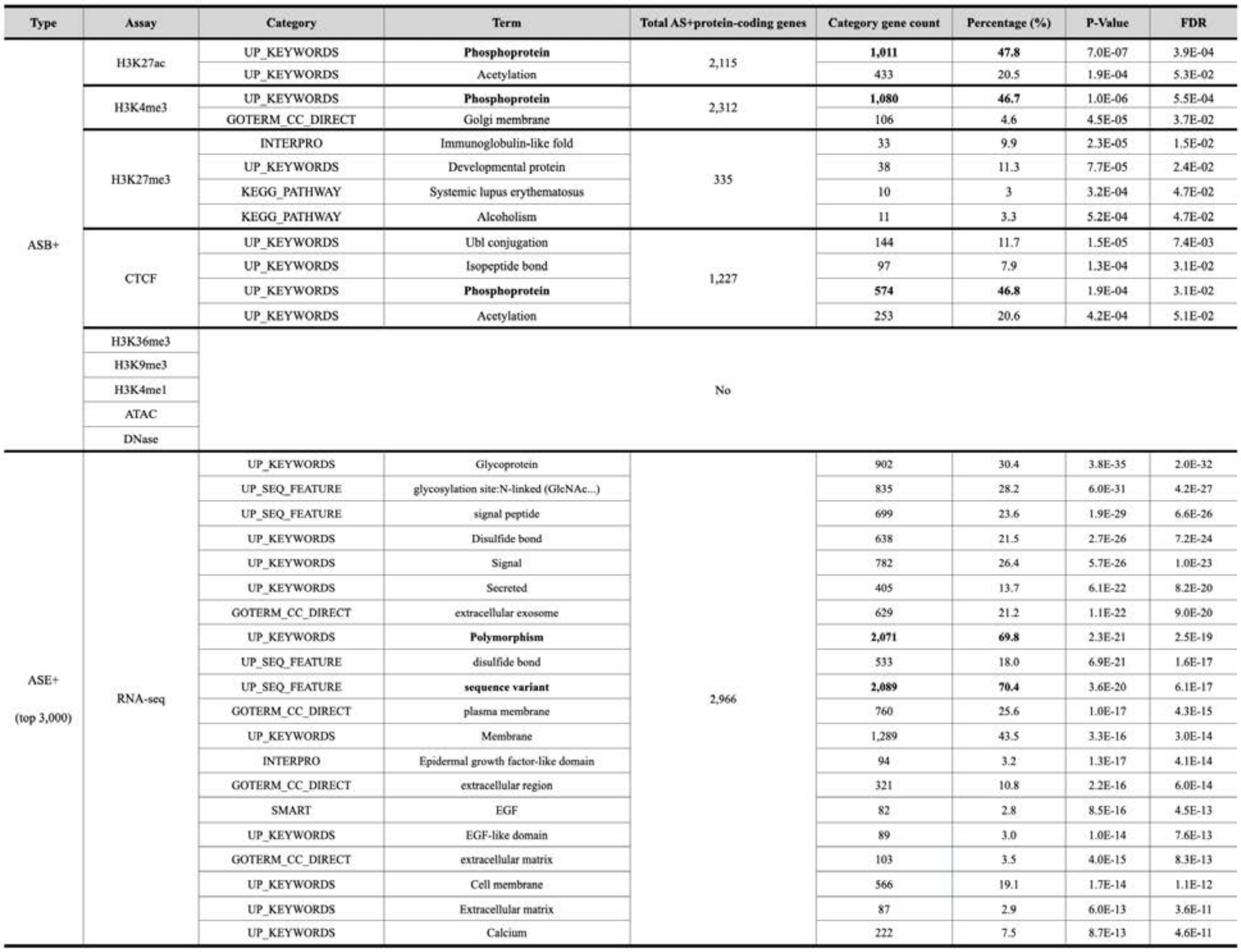
Of the average 6,454,111 interactions per sample, 496,859 showed significant AS behavior.

A



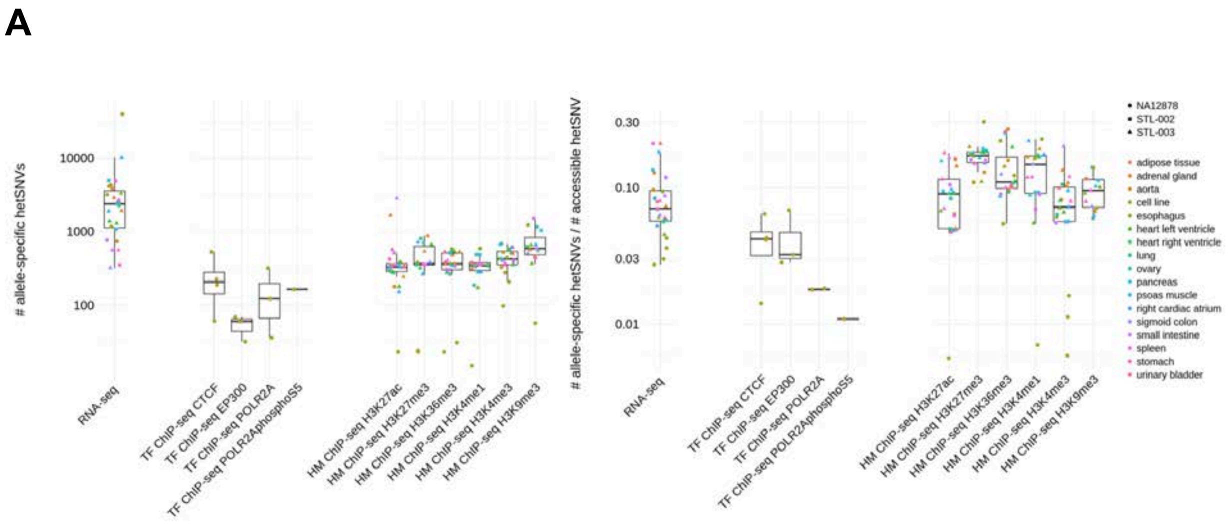
# allele-specific elements
1000
100
10
1
RNA-seq genes
ATAC-seq cCREs
DNase-seq cCREs
TF ChIP-seq CTCF cCREs
TF ChIP-seq EP300 cCREs
TF ChIP-seq POLR2A cCREs
TF ChIP-seq POLR2AphosphoS5 cCREs
HM ChIP-seq H3K27ac cCREs
HM ChIP-seq H3K27me3 cCREs
HM ChIP-seq H3K36me3 cCREs
HM ChIP-seq H3K4me1 cCREs
HM ChIP-seq H3K4me3 cCREs
HM ChIP-seq H3K9me3 cCREs
donor
ENC-001
ENC-002
ENC-003
ENC-004
tissue
adrenal gland
ascending aorta
body of pancreas
breast epithelium
coronary artery
esophagus muscularis mucosa
esophagus squamous epithelium
gastrocnemius medialis
gastroesophageal sphincter
heart left ventricle
lower leg skin
omental fat pad
ovary
Peyers patch
prostate gland
right atrium auricular region
right lobe of liver
sigmoid colon
spleen
stomach
subcutaneous adipose tissue
suprapubic skin
testis
thoracic aorta
thyroid gland
tibial artery
tibial nerve
transverse colon
upper lobe of left lung
uterus
vagina
# allele-specific elements / # accessible elements
0.050
0.030
0.010
0.005
RNA-seq genes
ATAC-seq cCREs
DNase-seq cCREs
TF ChIP-seq CTCF cCREs
TF ChIP-seq EP300 cCREs
TF ChIP-seq POLR2A cCREs
TF ChIP-seq POLR2AphosphoS5 cCREs
HM ChIP-seq H3K27ac cCREs
HM ChIP-seq H3K27me3 cCREs
HM ChIP-seq H3K36me3 cCREs
HM ChIP-seq H3K4me1 cCREs
HM ChIP-seq H3K4me3 cCREs
HM ChIP-seq H3K9me3 cCREs
donor
ENC-001
ENC-002
ENC-003
ENC-004
tissue
adrenal gland
ascending aorta
body of pancreas
breast epithelium
coronary artery
esophagus muscularis mucosa
esophagus squamous epithelium
gastrocnemius medialis
gastroesophageal sphincter
heart left ventricle
lower leg skin
omental fat pad
ovary
Peyers patch
prostate gland
right atrium auricular region
right lobe of liver
sigmoid colon
spleen
stomach
subcutaneous adipose tissue
suprapubic skin
testis
thoracic aorta
thyroid gland
tibial artery
tibial nerve
transverse colon
upper lobe of left lung
uterus
vagina

B

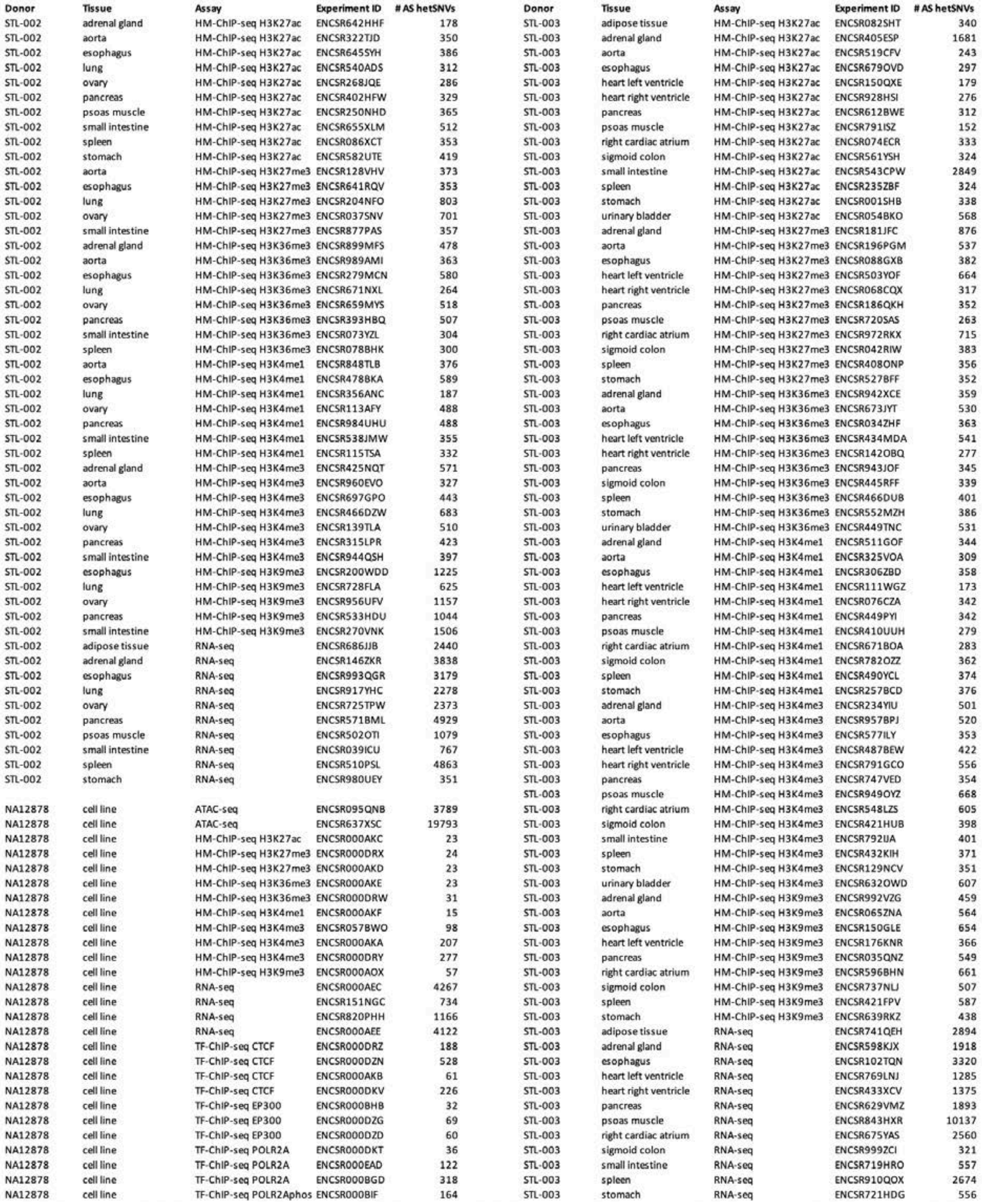


Type	Assay	Category	Term	Total AS+protein-coding genes	Category gene count	Percentage (%)	P-Value	FDR
ASB+	H3K27ac	UP_KEYWORDS	**Phosphoprotein**	2,115	**1,011**	**47.8**	7.0E-07	3.9E-04
		UP_KEYWORDS	Acetylation		433	20.5	1.9E-04	5.3E-02
	H3K4me3	UP_KEYWORDS	**Phosphoprotein**	2,312	**1,080**	**46.7**	1.0E-06	5.5E-04
		GOTERM_CC_DIRECT	Golgi membrane		106	4.6	4.5E-05	3.7E-02
	H3K27me3	INTERPRO	Immunoglobulin-like fold	335	33	9.9	2.3E-05	1.5E-02
		UP_KEYWORDS	Developmental protein		38	11.3	7.7E-05	2.4E-02
		KEGG_PATHWAY	Systemic lupus erythematosus		10	3	3.2E-04	4.7E-02
		KEGG_PATHWAY	Alcoholism		11	3.3	5.2E-04	4.7E-02
	CTCF	UP_KEYWORDS	Ubl conjugation	1,227	144	11.7	1.5E-05	7.4E-03
		UP_KEYWORDS	Isopeptide bond		97	7.9	1.3E-04	3.1E-02
		UP_KEYWORDS	**Phosphoprotein**		**574**	**46.8**	1.9E-04	3.1E-02
		UP_KEYWORDS	Acetylation		253	20.6	4.2E-04	5.1E-02
	H3K36me3	No						
	H3K9me3							
	H3K4me1							
	ATAC							
	DNase							
ASE+ (top 3,000)	RNA-seq	UP_KEYWORDS	Glycoprotein	2,966	902	30.4	3.8E-35	2.0E-32
		UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)		835	28.2	6.0E-31	4.2E-27
		UP_SEQ_FEATURE	signal peptide		699	23.6	1.9E-29	6.6E-26
		UP_KEYWORDS	Disulfide bond		638	21.5	2.7E-26	7.2E-24
		UP_KEYWORDS	Signal		782	26.4	5.7E-26	1.0E-23
		UP_KEYWORDS	Secreted		405	13.7	6.1E-22	8.2E-20
		GOTERM_CC_DIRECT	extracellular exosome		629	21.2	1.1E-22	9.0E-20
		UP_KEYWORDS	**Polymorphism**		**2,071**	**69.8**	2.3E-21	2.5E-19
		UP_SEQ_FEATURE	disulfide bond		533	18.0	6.9E-21	1.6E-17
		UP_SEQ_FEATURE	**sequence variant**		**2,089**	**70.4**	3.6E-20	6.1E-17
		GOTERM_CC_DIRECT	plasma membrane		760	25.6	1.0E-17	4.3E-15
		UP_KEYWORDS	Membrane		1,289	43.5	3.3E-16	3.0E-14
		INTERPRO	Epidermal growth factor-like domain		94	3.2	1.3E-17	4.1E-14
		GOTERM_CC_DIRECT	extracellular region		321	10.8	2.2E-16	6.0E-14
		SMART	EGF		82	2.8	8.5E-16	4.5E-13
		UP_KEYWORDS	EGF-like domain		89	3.0	1.0E-14	7.6E-13
		GOTERM_CC_DIRECT	extracellular matrix		103	3.5	4.0E-15	8.3E-13
		UP_KEYWORDS	Cell membrane		566	19.1	1.7E-14	1.1E-12
		UP_KEYWORDS	Extracellular matrix		87	2.9	6.0E-13	3.6E-11
		UP_KEYWORDS	Calcium		222	7.5	8.7E-13	4.6E-11

Data S11. AS elements, related to Figure S3 and STAR Methods "AS Elements" Section
(A) Distribution (top) and fractions (bottom) of the number of genomic elements (genes and cCREs) associated with AS behavior across different EN-TEx donors, tissues, and assays. On average, for each individual and tissue in the RNA/ChIP/ATAC-seq samples, 193 cCREs (median 123, IQR 32-284) and 351 genes (median 205, IQR 193-452) showed a significant AS imbalance per assay.
(B) Gene ontology enrichment analysis of AS genes. Functional annotation of EN-TEx AS protein-coding genes detected from different assays. Analysis was performed using DAVID Bioinformatics Resources 6.8. For each assay, the background list includes all protein-coding genes with accessible promoters for ASB or with accessible expressions for ASE. For ASE analysis, since DAVID has a 3,000 gene limit, the top 3,000 mostly ASE protein-coding genes were selected for the enrichment analysis, and the top 20 enriched terms are shown in the table. Terms with the largest number of genes are highlighted in bold.


A
# allele-specific hetSNVs
10000
1000
100
RNA-seq
TF ChIP-seq CTCF
TF ChIP-seq EP300
TF ChIP-seq POLR2A
TF ChIP-seq POLR2AphosphoS5
HM ChIP-seq H3K27ac
HM ChIP-seq H3K27me3
HM ChIP-seq H3K36me3
HM ChIP-seq H3K4me1
HM ChIP-seq H3K4me3
HM ChIP-seq H3K9me3
# allele-specific hetSNVs / # accessible hetSNV
0.30
0.10
0.03
0.01
RNA-seq
TF ChIP-seq CTCF
TF ChIP-seq EP300
TF ChIP-seq POLR2A
TF ChIP-seq POLR2AphosphoS5
HM ChIP-seq H3K27ac
HM ChIP-seq H3K27me3
HM ChIP-seq H3K36me3
HM ChIP-seq H3K4me1
HM ChIP-seq H3K4me3
HM ChIP-seq H3K9me3
● NA12878
■ STL-002
▲ STL-003
● adipose tissue
● adrenal gland
● aorta
● cell line
● esophagus
● heart left ventricle
● heart right ventricle
● lung
● ovary
● pancreas
● psoas muscle
● right cardiac atrium
● sigmoid colon
● small intestine
● spleen
● stomach
● urinary bladder

B



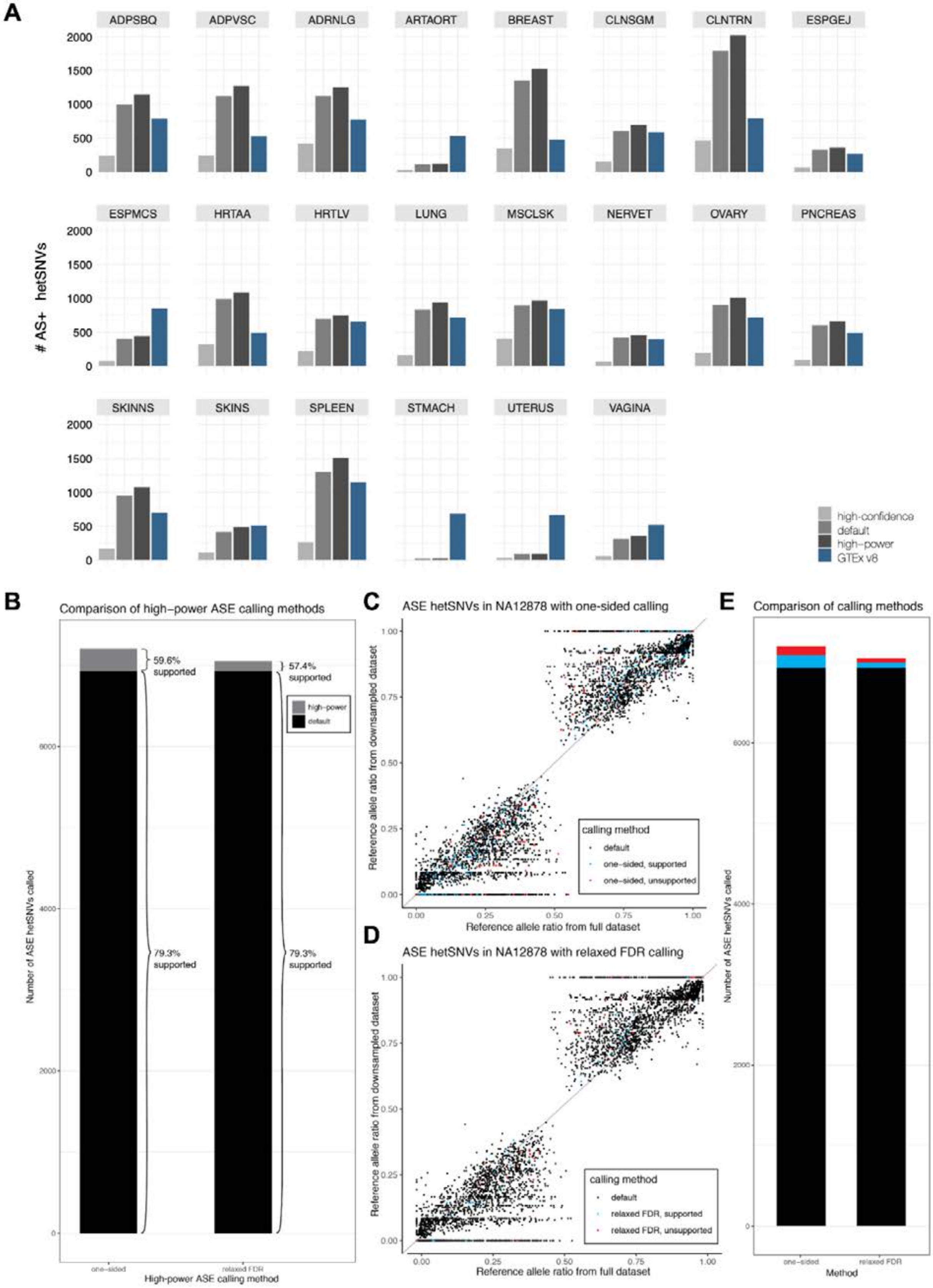
Donor	Tissue	Assay	Experiment ID	# AS hetSNVs
STL-002	adrenal gland	HM-ChIP-seq H3K27ac	ENCSR642HHF	178
STL-002	aorta	HM-ChIP-seq H3K27ac	ENCSR322TJD	350
STL-002	esophagus	HM-ChIP-seq H3K27ac	ENCSR645SYH	386
STL-002	lung	HM-ChIP-seq H3K27ac	ENCSR540ADS	312
STL-002	ovary	HM-ChIP-seq H3K27ac	ENCSR268JQE	286
STL-002	pancreas	HM-ChIP-seq H3K27ac	ENCSR402HFW	329
STL-002	psoas muscle	HM-ChIP-seq H3K27ac	ENCSR250NHD	365
STL-002	small intestine	HM-ChIP-seq H3K27ac	ENCSR655XLM	512
STL-002	spleen	HM-ChIP-seq H3K27ac	ENCSR086XCT	353
STL-002	stomach	HM-ChIP-seq H3K27ac	ENCSR582UTE	419
STL-002	aorta	HM-ChIP-seq H3K27me3	ENCSR128VHV	373
STL-002	esophagus	HM-ChIP-seq H3K27me3	ENCSR641RQV	353
STL-002	lung	HM-ChIP-seq H3K27me3	ENCSR204NFO	803
STL-002	ovary	HM-ChIP-seq H3K27me3	ENCSR037SNV	701
STL-002	small intestine	HM-ChIP-seq H3K27me3	ENCSR877PAS	357
STL-002	adrenal gland	HM-ChIP-seq H3K36me3	ENCSR899MFS	478
STL-002	aorta	HM-ChIP-seq H3K36me3	ENCSR989AMI	363
STL-002	esophagus	HM-ChIP-seq H3K36me3	ENCSR279MCN	580
STL-002	lung	HM-ChIP-seq H3K36me3	ENCSR671NXL	264
STL-002	ovary	HM-ChIP-seq H3K36me3	ENCSR659MYS	518
STL-002	pancreas	HM-ChIP-seq H3K36me3	ENCSR393HBQ	507
STL-002	small intestine	HM-ChIP-seq H3K36me3	ENCSR073YZL	304
STL-002	spleen	HM-ChIP-seq H3K36me3	ENCSR078BHK	300
STL-002	aorta	HM-ChIP-seq H3K4me1	ENCSR848TLB	376
STL-002	esophagus	HM-ChIP-seq H3K4me1	ENCSR478BKA	589
STL-002	lung	HM-ChIP-seq H3K4me1	ENCSR356ANC	187
STL-002	ovary	HM-ChIP-seq H3K4me1	ENCSR113AFY	488
STL-002	pancreas	HM-ChIP-seq H3K4me1	ENCSR984UHU	488
STL-002	small intestine	HM-ChIP-seq H3K4me1	ENCSR538JMW	355
STL-002	spleen	HM-ChIP-seq H3K4me1	ENCSR115TSA	332
STL-002	adrenal gland	HM-ChIP-seq H3K4me3	ENCSR425NQT	571
STL-002	aorta	HM-ChIP-seq H3K4me3	ENCSR960EVO	327
STL-002	esophagus	HM-ChIP-seq H3K4me3	ENCSR697GPO	443
STL-002	lung	HM-ChIP-seq H3K4me3	ENCSR466DZW	683
STL-002	ovary	HM-ChIP-seq H3K4me3	ENCSR139TLA	510
STL-002	pancreas	HM-ChIP-seq H3K4me3	ENCSR315LPR	423
STL-002	small intestine	HM-ChIP-seq H3K4me3	ENCSR944QSH	397
STL-002	esophagus	HM-ChIP-seq H3K9me3	ENCSR200WDD	1225
STL-002	lung	HM-ChIP-seq H3K9me3	ENCSR728FLA	625
STL-002	ovary	HM-ChIP-seq H3K9me3	ENCSR956UFV	1157
STL-002	pancreas	HM-ChIP-seq H3K9me3	ENCSR533HDU	1044
STL-002	small intestine	HM-ChIP-seq H3K9me3	ENCSR270VNK	1506
STL-002	adipose tissue	RNA-seq	ENCSR686JJB	2440
STL-002	adrenal gland	RNA-seq	ENCSR146ZKR	3838
STL-002	esophagus	RNA-seq	ENCSR993QGR	3179
STL-002	lung	RNA-seq	ENCSR917YHC	2278
STL-002	ovary	RNA-seq	ENCSR725TPW	2373
STL-002	pancreas	RNA-seq	ENCSR571BML	4929
STL-002	psoas muscle	RNA-seq	ENCSR502OTI	1079
STL-002	small intestine	RNA-seq	ENCSR039ICU	767
STL-002	spleen	RNA-seq	ENCSR510PSL	4863
STL-002	stomach	RNA-seq	ENCSR980UEY	351
NA12878	cell line	ATAC-seq	ENCSR095QNB	3789
NA12878	cell line	ATAC-seq	ENCSR637XSC	19793
NA12878	cell line	HM-ChIP-seq H3K27ac	ENCSR000AKC	23
NA12878	cell line	HM-ChIP-seq H3K27me3	ENCSR000DRX	24
NA12878	cell line	HM-ChIP-seq H3K27me3	ENCSR000AKD	23
NA12878	cell line	HM-ChIP-seq H3K36me3	ENCSR000AKE	23
NA12878	cell line	HM-ChIP-seq H3K36me3	ENCSR000DRW	31
NA12878	cell line	HM-ChIP-seq H3K4me1	ENCSR000AKF	15
NA12878	cell line	HM-ChIP-seq H3K4me3	ENCSR057BWO	98
NA12878	cell line	HM-ChIP-seq H3K4me3	ENCSR000AKA	207
NA12878	cell line	HM-ChIP-seq H3K4me3	ENCSR000DRY	277
NA12878	cell line	HM-ChIP-seq H3K9me3	ENCSR000AOX	57
NA12878	cell line	RNA-seq	ENCSR000AEC	4267
NA12878	cell line	RNA-seq	ENCSR151NGC	734
NA12878	cell line	RNA-seq	ENCSR820PHH	1166
NA12878	cell line	RNA-seq	ENCSR000AEE	4122
NA12878	cell line	TF-ChIP-seq CTCF	ENCSR000DRZ	188
NA12878	cell line	TF-ChIP-seq CTCF	ENCSR000DZN	528
NA12878	cell line	TF-ChIP-seq CTCF	ENCSR000AKB	61
NA12878	cell line	TF-ChIP-seq CTCF	ENCSR000DKV	226
NA12878	cell line	TF-ChIP-seq EP300	ENCSR000BHB	32
NA12878	cell line	TF-ChIP-seq EP300	ENCSR000DZG	69
NA12878	cell line	TF-ChIP-seq EP300	ENCSR000DZD	60
NA12878	cell line	TF-ChIP-seq POLR2A	ENCSR000DKT	36
NA12878	cell line	TF-ChIP-seq POLR2A	ENCSR000EAD	122
NA12878	cell line	TF-ChIP-seq POLR2A	ENCSR000BGD	318
NA12878	cell line	TF-ChIP-seq POLR2Aphos	ENCSR000BIF	164

Donor	Tissue	Assay	Experiment ID	# AS hetSNVs
STL-003	adipose tissue	HM-ChIP-seq H3K27ac	ENCSR082SHT	340
STL-003	adrenal gland	HM-ChIP-seq H3K27ac	ENCSR405ESP	1681
STL-003	aorta	HM-ChIP-seq H3K27ac	ENCSR519CFV	243
STL-003	esophagus	HM-ChIP-seq H3K27ac	ENCSR679OVD	297
STL-003	heart left ventricle	HM-ChIP-seq H3K27ac	ENCSR150QXE	179
STL-003	heart right ventricle	HM-ChIP-seq H3K27ac	ENCSR928HSI	276
STL-003	pancreas	HM-ChIP-seq H3K27ac	ENCSR612BWE	312
STL-003	psoas muscle	HM-ChIP-seq H3K27ac	ENCSR791ISZ	152
STL-003	right cardiac atrium	HM-ChIP-seq H3K27ac	ENCSR074ECR	333
STL-003	sigmoid colon	HM-ChIP-seq H3K27ac	ENCSR561YSH	324
STL-003	small intestine	HM-ChIP-seq H3K27ac	ENCSR543CPW	2849
STL-003	spleen	HM-ChIP-seq H3K27ac	ENCSR235ZBF	324
STL-003	stomach	HM-ChIP-seq H3K27ac	ENCSR001SHB	338
STL-003	urinary bladder	HM-ChIP-seq H3K27ac	ENCSR054BKO	568
STL-003	adrenal gland	HM-ChIP-seq H3K27me3	ENCSR181JFC	876
STL-003	aorta	HM-ChIP-seq H3K27me3	ENCSR196PGM	537
STL-003	esophagus	HM-ChIP-seq H3K27me3	ENCSR088GXB	382
STL-003	heart left ventricle	HM-ChIP-seq H3K27me3	ENCSR503YOF	664
STL-003	heart right ventricle	HM-ChIP-seq H3K27me3	ENCSR068CQX	317
STL-003	pancreas	HM-ChIP-seq H3K27me3	ENCSR186QKH	352
STL-003	psoas muscle	HM-ChIP-seq H3K27me3	ENCSR720SAS	263
STL-003	right cardiac atrium	HM-ChIP-seq H3K27me3	ENCSR972RKX	715
STL-003	sigmoid colon	HM-ChIP-seq H3K27me3	ENCSR042RIW	383
STL-003	spleen	HM-ChIP-seq H3K27me3	ENCSR408ONP	356
STL-003	stomach	HM-ChIP-seq H3K27me3	ENCSR527BFF	352
STL-003	adrenal gland	HM-ChIP-seq H3K36me3	ENCSR942XCE	359
STL-003	aorta	HM-ChIP-seq H3K36me3	ENCSR673JYT	530
STL-003	esophagus	HM-ChIP-seq H3K36me3	ENCSR034ZHF	363
STL-003	heart left ventricle	HM-ChIP-seq H3K36me3	ENCSR434MDA	541
STL-003	heart right ventricle	HM-ChIP-seq H3K36me3	ENCSR142OBQ	277
STL-003	pancreas	HM-ChIP-seq H3K36me3	ENCSR943JOF	345
STL-003	sigmoid colon	HM-ChIP-seq H3K36me3	ENCSR445RFF	339
STL-003	spleen	HM-ChIP-seq H3K36me3	ENCSR466DUB	401
STL-003	stomach	HM-ChIP-seq H3K36me3	ENCSR552MZH	386
STL-003	urinary bladder	HM-ChIP-seq H3K36me3	ENCSR449TNC	531
STL-003	adrenal gland	HM-ChIP-seq H3K4me1	ENCSR511GOF	344
STL-003	aorta	HM-ChIP-seq H3K4me1	ENCSR325VOA	309
STL-003	esophagus	HM-ChIP-seq H3K4me1	ENCSR306ZBD	358
STL-003	heart left ventricle	HM-ChIP-seq H3K4me1	ENCSR111WGZ	173
STL-003	heart right ventricle	HM-ChIP-seq H3K4me1	ENCSR076CZA	342
STL-003	pancreas	HM-ChIP-seq H3K4me1	ENCSR449PYI	342
STL-003	psoas muscle	HM-ChIP-seq H3K4me1	ENCSR410UUH	279
STL-003	right cardiac atrium	HM-ChIP-seq H3K4me1	ENCSR671BOA	283
STL-003	sigmoid colon	HM-ChIP-seq H3K4me1	ENCSR782OZZ	362
STL-003	spleen	HM-ChIP-seq H3K4me1	ENCSR490YCL	374
STL-003	stomach	HM-ChIP-seq H3K4me1	ENCSR257BCD	376
STL-003	adrenal gland	HM-ChIP-seq H3K4me3	ENCSR234YIU	501
STL-003	aorta	HM-ChIP-seq H3K4me3	ENCSR957BPJ	520
STL-003	esophagus	HM-ChIP-seq H3K4me3	ENCSR577ILY	353
STL-003	heart left ventricle	HM-ChIP-seq H3K4me3	ENCSR487BEW	422
STL-003	heart right ventricle	HM-ChIP-seq H3K4me3	ENCSR791GCO	556
STL-003	pancreas	HM-ChIP-seq H3K4me3	ENCSR747VED	354
STL-003	psoas muscle	HM-ChIP-seq H3K4me3	ENCSR949OYZ	668
STL-003	right cardiac atrium	HM-ChIP-seq H3K4me3	ENCSR548LZS	605
STL-003	sigmoid colon	HM-ChIP-seq H3K4me3	ENCSR421HUB	398
STL-003	small intestine	HM-ChIP-seq H3K4me3	ENCSR792IJA	401
STL-003	spleen	HM-ChIP-seq H3K4me3	ENCSR432KIH	371
STL-003	stomach	HM-ChIP-seq H3K4me3	ENCSR129NCV	351
STL-003	urinary bladder	HM-ChIP-seq H3K4me3	ENCSR632OWD	607
STL-003	adrenal gland	HM-ChIP-seq H3K9me3	ENCSR992VZG	459
STL-003	aorta	HM-ChIP-seq H3K9me3	ENCSR065ZNA	564
STL-003	esophagus	HM-ChIP-seq H3K9me3	ENCSR150GLE	654
STL-003	heart left ventricle	HM-ChIP-seq H3K9me3	ENCSR176KNR	366
STL-003	pancreas	HM-ChIP-seq H3K9me3	ENCSR035QNZ	549
STL-003	right cardiac atrium	HM-ChIP-seq H3K9me3	ENCSR596BHN	661
STL-003	sigmoid colon	HM-ChIP-seq H3K9me3	ENCSR737NLJ	507
STL-003	spleen	HM-ChIP-seq H3K9me3	ENCSR421FPV	587
STL-003	stomach	HM-ChIP-seq H3K9me3	ENCSR639RKZ	438
STL-003	adipose tissue	RNA-seq	ENCSR741QEH	2894
STL-003	adrenal gland	RNA-seq	ENCSR598KJX	1918
STL-003	esophagus	RNA-seq	ENCSR102TQN	3320
STL-003	heart left ventricle	RNA-seq	ENCSR769LNJ	1285
STL-003	heart right ventricle	RNA-seq	ENCSR433XCV	1375
STL-003	pancreas	RNA-seq	ENCSR629VMZ	1893
STL-003	psoas muscle	RNA-seq	ENCSR843HXR	10137
STL-003	right cardiac atrium	RNA-seq	ENCSR675YAS	2560
STL-003	sigmoid colon	RNA-seq	ENCSR999ZCI	321
STL-003	small intestine	RNA-seq	ENCSR719HRO	557
STL-003	spleen	RNA-seq	ENCSR910QOX	2674
STL-003	stomach	RNA-seq	ENCSR721HDG	556

Data S12. Construction of a validation dataset from AS events in non-EN-TEx datasets, related to STAR Methods “AS Catalog” Section

(A) Distribution and fraction of the number of hetSNVs associated with AS behavior detected in NA12878 and Roadmap individuals STL002 and STL003.

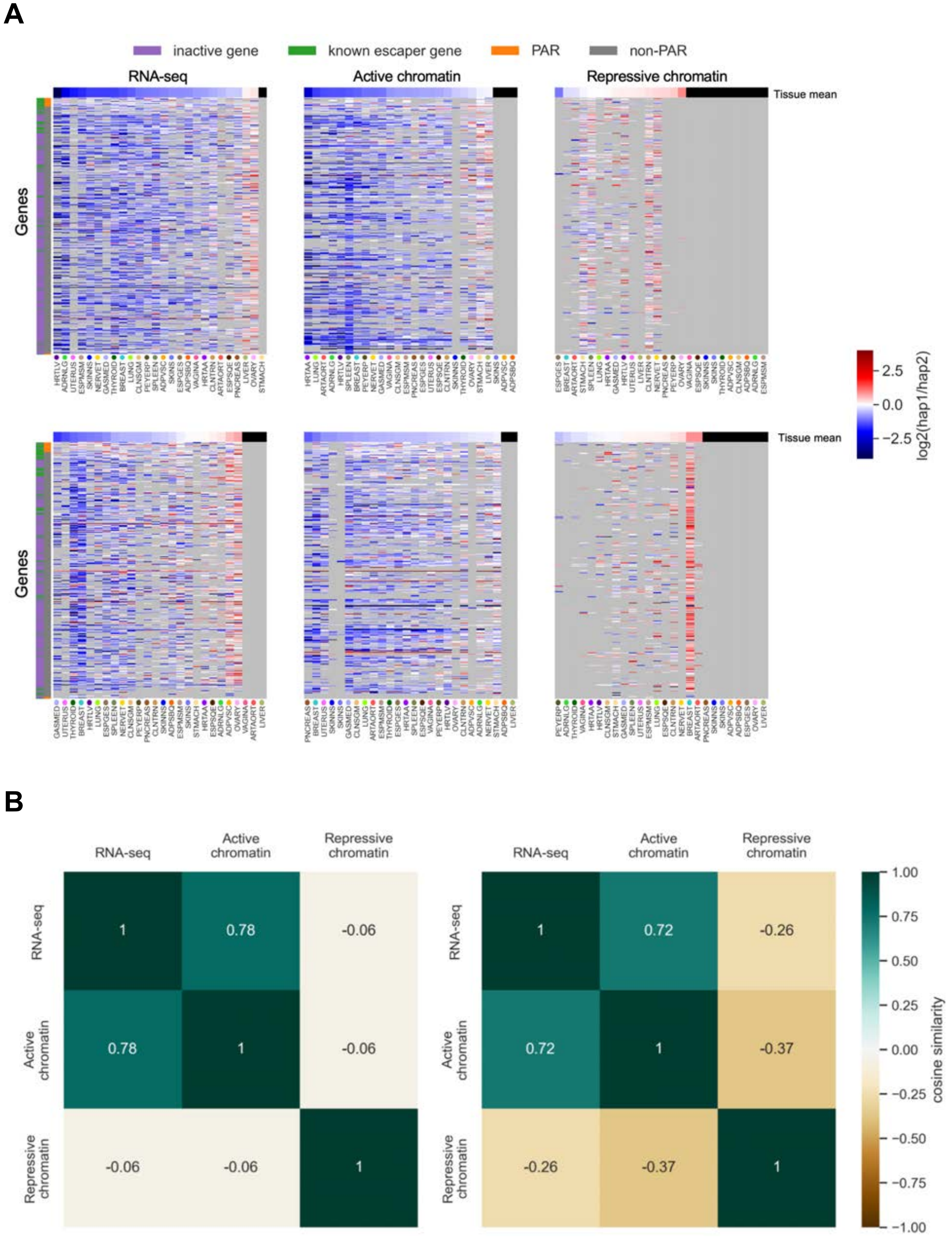
(B) Datasets used for calling AS events in the Roadmap individuals.


A
ADPSBQ
ADPVSC
ADRNLG
ARTAORT
BREAST
CLNSGM
CLNTRN
ESPGEJ
ESPMCS
HRTAA
HRTLV
LUNG
MSCLSK
NERVET
OVARY
PNCREAS
SKINNS
SKINS
SPLEEN
STMACH
UTERUS
VAGINA
# AS+ hetSNVs
0
500
1000
1500
2000
high-confidence
default
high-power
GTEx v8
B
Comparison of high-power ASE calling methods
59.6% supported
57.4% supported
79.3% supported
79.3% supported
high-power
default
Number of ASE hetSNVs called
0
2000
4000
6000
one-sided
relaxed FDR
High-power ASE calling method
C
ASE hetSNVs in NA12878 with one-sided calling
Reference allele ratio from downsampled dataset
Reference allele ratio from full dataset
0.00
0.25
0.50
0.75
1.00
calling method
default
one-sided, supported
one-sided, unsupported
D
ASE hetSNVs in NA12878 with relaxed FDR calling
Reference allele ratio from downsampled dataset
Reference allele ratio from full dataset
0.00
0.25
0.50
0.75
1.00
calling method
default
relaxed FDR, supported
relaxed FDR, unsupported
E
Comparison of calling methods
Number of ASE hetSNVs called
0
2000
4000
6000
one-sided
relaxed FDR
Method

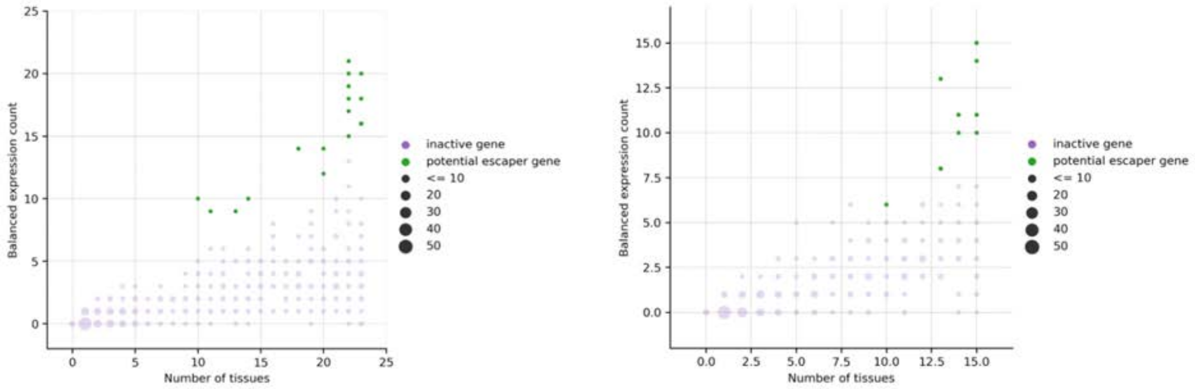
Data S13. High-power AS calling, related to STAR Methods “AS Catalog” Section

(A) Numbers of AS hetSNVs detected from RNA-seq in different tissues from individual 3. To produce high-power tissue-specific call sets, we called ASE and ASB sites for each tissue at a relaxed FDR threshold if the hetSNV was called AS in the pooled call set; otherwise, the sites were called at the usual 10% FDR. The “relaxed” FDR varied somewhat from tissue to tissue due to granularities in calculation but did not exceed 20%. Typically, high-power call sets produce 10%–20% more AS hetSNVs than typical call sets.

(B) - (E) Validation of high-power AS calling methods. To increase the detection power of ASE hetSNVs in datasets with fewer reads, we tested two high-power calling methods that selectively impose less-stringent tests on hetSNVs, which have been shown to have AS behavior in other experiments. The first method uses a one-sided beta-binomial test as its less-stringent test, while the second uses a two-sided beta-binomial test with a relaxed FDR of 20%. All hetSNVs that do not have prior evidence of being AS are evaluated with the standard two-sided beta-binomial test with an FDR of 10%. We validated both methods by testing on a deeply sequenced RNA-seq dataset from the GM12878 cell line and simulating a shallower experiment by downsampling this dataset by a factor of 4. (B) and (E) Using the default ASE calling method, 6,927 ASE hetSNVs were identified in the downsampled dataset. Of these, 79.3% were supported by the full RNA-seq dataset – that is, they were also called ASE in the full dataset. One-sided testing identified 275 additional ASE hetSNVs, and relaxed FDR testing identified 122 additional ASE hetSNVs. Both methods are enriched for supported hetSNVs as compared to the full pool of hetSNVs (59.6% and 57.4%, respectively), though they have a higher error rate than the default ASE calling method in this respect. (C) - (D) We also show a comparison of reference allele ratios of ASE hetSNVs under different calling methods. Overall, the ASE hetSNVs added by both one-sided and relaxed FDR calling display similar reference allele ratios to ASE hetSNVs identified by default calling.


A
inactive gene
known escaper gene
PAR
non-PAR
RNA-seq
Active chromatin
Repressive chromatin
Tissue mean
Genes
HRTLV ADRNLG UTERUS ESPMSM SKINNS NERVET GASMED THYROID BREAST LUNG CLNSGM PEYERP SPLEEN ADPVSC SKINS ESPGES ADPSBQ VAGINA HRTAA CLNTRN ARTAORT ESPSQE PNCREAS LIVER OVARY STMACH
HRTAA LUNG ARTAORT ADRNLG HRTLV SPLEEN BREAST PEYERP NERVET GASMED VAGINA CLNSGM ESPMSM PNCREAS ESPGES UTERUS ESPSQE CLNTRN SKINNS THYROID OVARY STMACH LIVER SKINS ADPVSC ADPSBQ
ESPGES BREAST ARTAORT STMACH SPLEEN LUNG HRTAA GASMED HRTLV UTERUS LIVER CLNTRN NERVET PNCREAS PEYERP OVARY VAGINA ESPSQE SKINNS SKINS THYROID ADPVSC CLNSGM ADPSBQ ADRNLG ESPMSM
2.5
0.0
−2.5
log2(hap1/hap2)
Tissue mean
Genes
GASMED UTERUS THYROID BREAST HRTLV LUNG ESPGES SPLEEN NERVET CLNSGM PEYERP PNCREAS CLNTRN SKINNS ADPSBQ ESPMSM SKINS STMACH HRTAA ESPSQE ADRNLG ADPVSC OVARY VAGINA ARTAORT LIVER
PNCREAS BREAST UTERUS SKINNS SKINS GASMED CLNSGM LUNG ARTAORT ESPMSM THYROID ESPGES HRTAA SPLEEN ESPSQE VAGINA PEYERP HRTLV OVARY CLNTRN ADPVSC ADRNLG NERVET STMACH ADPSBQ LIVER
PEYERP ADRNLG THYROID VAGINA HRTAA HRTLV CLNSGM STMACH GASMED SPLEEN UTERUS ESPMSM LUNG ESPSQE CLNTRN NERVET BREAST ARTAORT PNCREAS SKINNS SKINS ADPVSC ADPSBQ ESPGES OVARY LIVER
B
RNA-seq
Active chromatin
Repressive chromatin
1 0.78 -0.06
0.78 1 -0.06
-0.06 -0.06 1
RNA-seq
Active chromatin
Repressive chromatin
1 0.72 -0.26
0.72 1 -0.37
-0.26 -0.37 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
cosine similarity

C

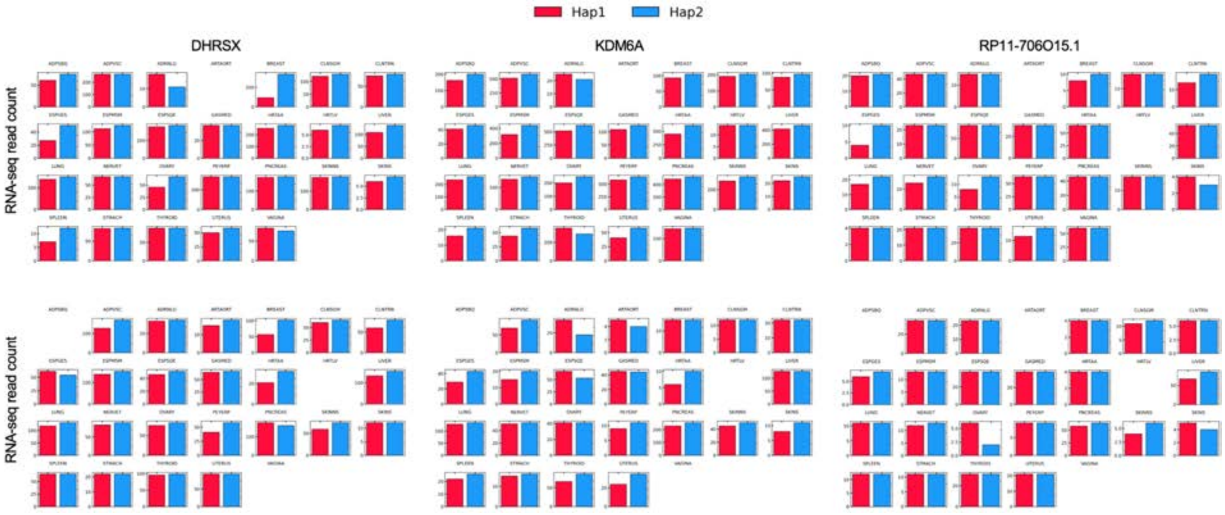


Gene	PAR 1	PAR 2	Ref
AKAP17A	yes	no	2
ARSD	no	no	2, 6
ASMTL	yes	no	2
CA5BP1	no	no	2
DHRSX	yes	no	2
DIAPH2-AS1	no	no	5
EIF1AX	no	no	2, 6
GTPBP6	yes	no	2
IL3RA	yes	no	2
KDM6A	no	no	2,6
PUDP	no	no	3
RP11-706O15.1	no	no	2, 6
WASH6P	no	yes	4
ZFX	no	no	2, 6

Gene	PAR 1	PAR 2	Ref
AKAP17A	yes	no	2
DHRSX	yes	no	2
GTPBP6	yes	no	2
KDM5C	no	no	1, 2, 6
KDM6A	no	no	2, 6
MXRA5	no	no	2, 6
RP11-706O15.1	no	no	2, 6
SLC25A6	yes	no	2
TRAPPC2	no	no	2, 6
WASH6P	no	yes	4

🟨 escaper genes found in both ENC-003 and ENC-004

D



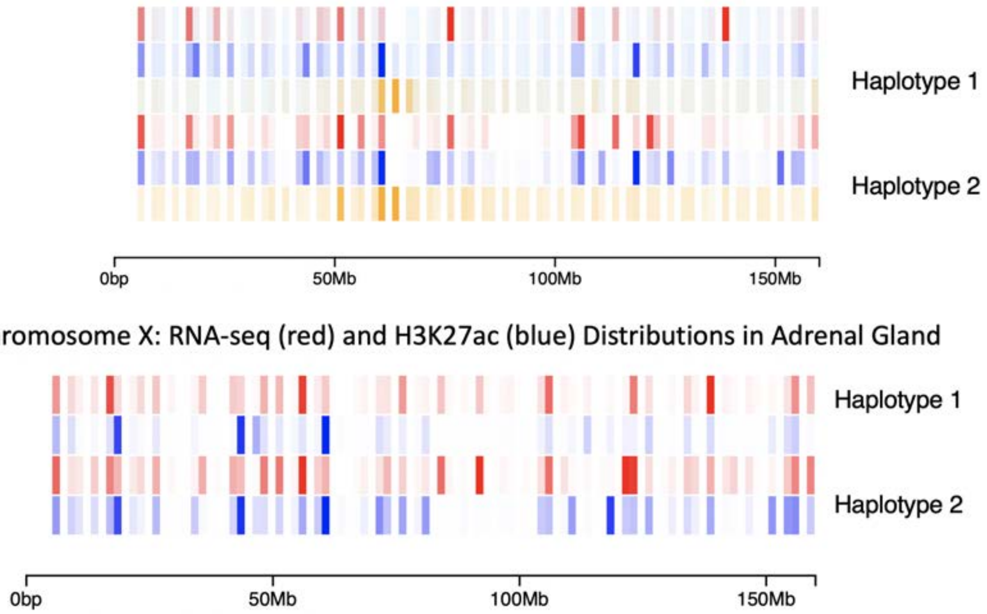
E

strongly hap1	🟦
hap1	🔹
neutral	🟩
hap2	🔸
strongly hap2	🟧
N/A	⬛
inconclusive	⬜

	expression	H3k27ac	H3k4me3	H3k4me1	H3K36me3	H3K9me3	H3k27me3	CTCF	POLR2A	POLR2Ap5
transverse colon	🟩	🔸	🔸	⬛	⬜	⬛	🟩	🟩	🟧	🔸
sigmoid colon	🟧	🔸	🔸	⬛	⬛	⬛	⬛	🟩	🟧	🔸
upper lobe of left lung	🟧	🟧	🟧	⬛	⬛	🟩	⬜	🟩	⬜	🔸
spleen	🟧	🟧	🟧	🔸	🟧	🟩	🔹	🔸	⬜	🟧
gastrocnemius medialis	🟧	🔸	🟧	⬛	⬜	⬜	🟩	🟧	🟧	⬛
adrenal gland	🟧	🟧	🟧	⬛	⬛	⬜	⬛	🟧	🟧	⬛
esophagus muscularis mucosa	🟧	🔸	🔸	⬛	⬛	⬜	⬜	🔸	🟧	🟧
thyroid gland	🟧	🟩	🟩	⬛	⬛	⬜	⬛	🟩	⬜	⬛
gastroesophageal sphincter	🔸	🔸	🔸	⬛	⬛	⬜	⬛	🔸	🟧	⬛
tibial nerve	🟧	🟧	🟧	⬛	🟩	🟩	🟦	🟧	🟧	⬛
body of pancreas	🟩	🔸	⬛	⬛	⬛	⬜	⬜	🟩	🔸	⬛
esophagus squamous epithelium	🟩	🔸	🟧	⬛	⬛	⬜	⬛	🟩	🟧	⬛
Peyer's patch	🔸	🟧	🟧	⬛	⬜	⬜	⬜	🟧	🔸	🔸
breast epithelium	🟧	🟧	🟧	⬛	⬛	🟩	⬜	🟩	🔸	🟩
suprapubic skin	🟧	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬜	⬛
heart left ventricle	🟧	🔸	🟧	⬛	⬛	🟩	🟩	🟧	🟧	⬛
vagina	🔸	🟧	🔸	⬛	⬛	⬛	⬛	🟧	🟧	🟧
lower leg skin	🔸	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬜	⬜
uterus	🔸	🔸	🔸	⬛	⬜	🟩	⬜	🟧	🔸	⬛
right atrium auricular region	🔸	🟧	🟩	⬜	🟩	🟩	⬜	🔸	⬛	⬛
omental fat pad	🟧	🔸	🟧	⬛	⬜	⬜	🟩	🟧	🟧	⬛
subcutaneous adipose tissue	🔸	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛
ascending aorta	🟩	🟧	🟧	⬛	⬛	🟩	🟩	🟩	⬛	⬛
right lobe of liver	🔹	🟩	🟩	⬜	🟩	⬜	⬛	🟩	🔹	⬛

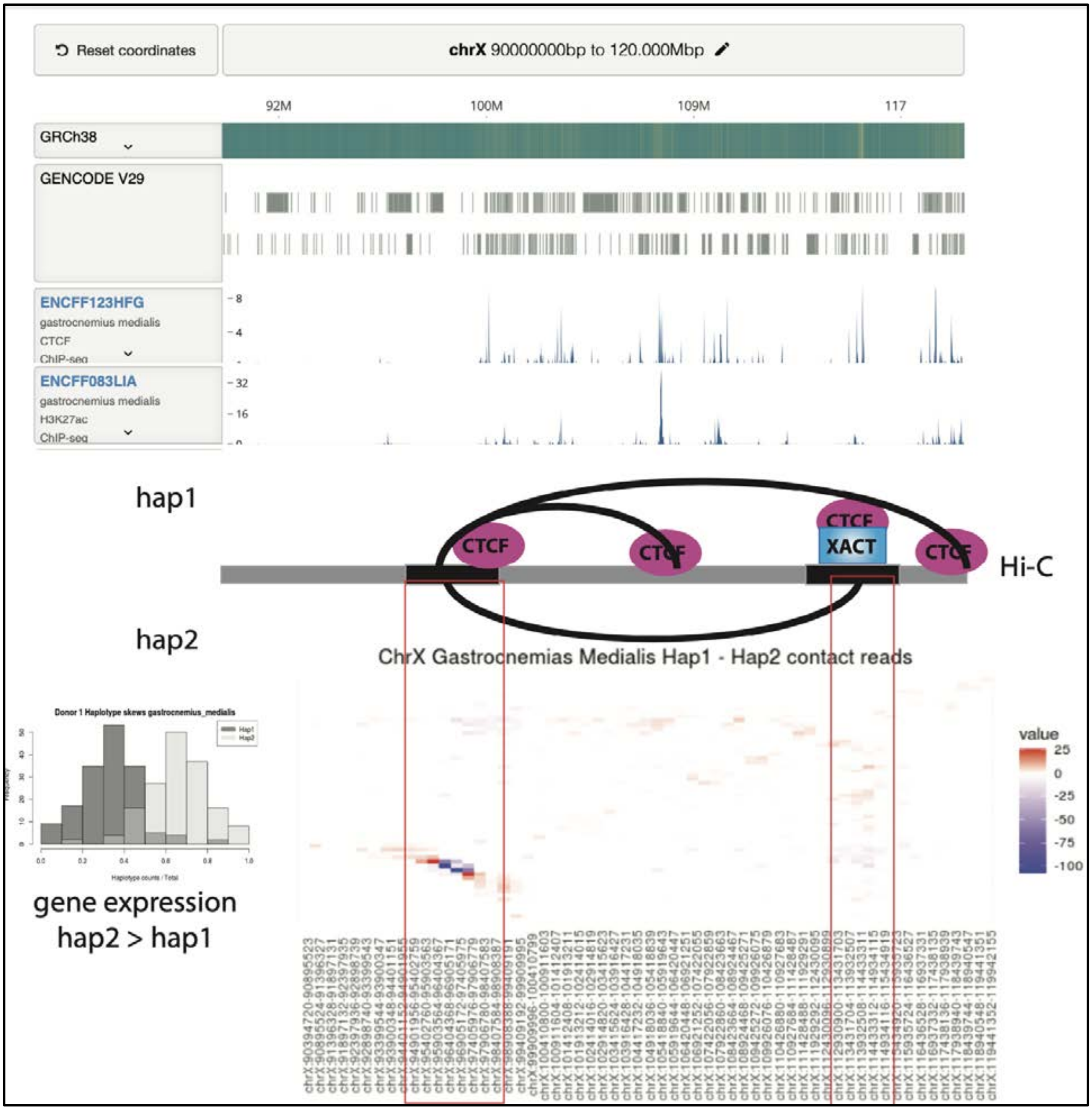
F

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

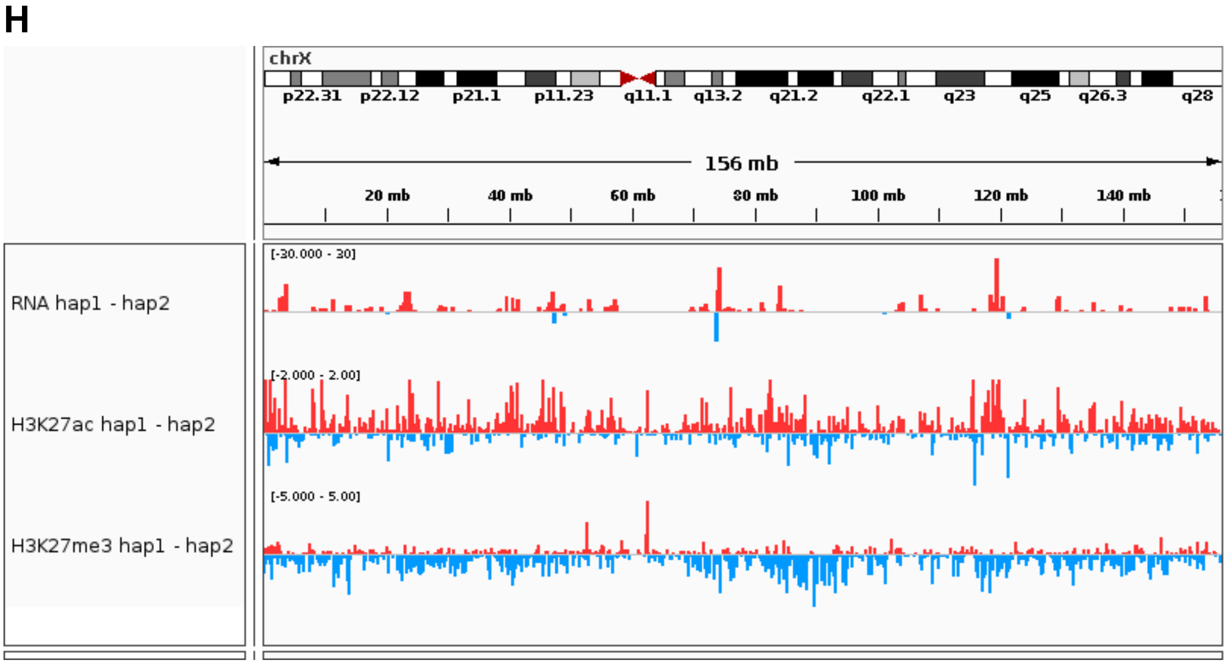


Chromosome X: RNA-seq (red) and H3K27ac (blue) Distributions in Adrenal Gland

G



Reset coordinates
chrX 90000000bp to 120.000Mbp
92M
100M
109M
117
GRCh38
GENCODE V29
ENCFF123HFG
gastrocnemius medialis
CTCF
ChIP-seq
8
4
ENCFF083LIA
gastrocnemius medialis
H3K27ac
ChIP-seq
32
16
0
hap1
CTCF
CTCF
CTCF
XACT
CTCF
Hi-C
hap2
ChrX Gastrocnemias Medialis Hap1 - Hap2 contact reads
Donor 1 Haplotype skews gastrocnemius_medialis
Hap1
Hap2
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
Haplotype counts / Total
gene expression
hap2 > hap1
value
25
0
-25
-50
-75
-100
chrX:90394720-90895523
chrX:90895524-91396327
chrX:91396328-91897131
chrX:91897132-92397935
chrX:92397936-92898739
chrX:92898740-93399543
chrX:93399544-93900347
chrX:93900348-94401151
chrX:94401152-94901955
chrX:94901956-95402759
chrX:95402760-95903563
chrX:95903564-96404367
chrX:96404368-96905171
chrX:96905172-97405975
chrX:97405976-97906779
chrX:97906780-98407583
chrX:98407584-98908387
chrX:98908388-99409191
chrX:99409192-99909995
chrX:99909996-100410799
chrX:100410800-100911603
chrX:100911604-101412407
chrX:101412408-101913211
chrX:101913212-102414015
chrX:102414016-102914819
chrX:102914820-103415623
chrX:103415624-103916427
chrX:103916428-104417231
chrX:104417232-104918035
chrX:104918036-105418839
chrX:105418840-105919643
chrX:105919644-106420447
chrX:106420448-106921251
chrX:106921252-107422055
chrX:107422056-107922859
chrX:107922860-108423663
chrX:108423664-108924467
chrX:108924468-109425271
chrX:109425272-109926075
chrX:109926076-110426879
chrX:110426880-110927683
chrX:110927684-111428487
chrX:111428488-111929291
chrX:111929292-112430095
chrX:112430096-112930899
chrX:112930900-113431703
chrX:113431704-113932507
chrX:113932508-114433311
chrX:114433312-114934115
chrX:114934116-115434919
chrX:115434920-115935723
chrX:115935724-116436527
chrX:116436528-116937331
chrX:116937332-117438135
chrX:117438136-117938939
chrX:117938940-118439743
chrX:118439744-118940547
chrX:118940548-119441351
chrX:119441352-119942155



Data S14. Coordinated allele-specific activity on X Chromosome across assays, related to Figure 2 and STAR Methods "AS Examples" Section

(A) Haplotype-specific gene expression and chromatin assays on chrX. Fold-change values (log2[haplotype1/haplotype2]) for phased RNA-seq read counts (left column), active chromatin marks (middle), and repressive chromatin marks (right) are shown above for EN-TEx individuals ENC-003 (top row) and ENC-004 (bottom row). Heatmap columns are the observed EN-TEx tissues and rows are the GENCODE v24 genes. Chromatin marks are the mean signal from a +/- 10 Kb region flanking genes. Active and repressive marks are pools of active (CTCF, EP300, H3K27ac, H3K4me3, H3K36me3, H3K4me1, POLR2A) and repressive (H3K27me3, H3K9me3) chromatin assays, respectively. Light gray cells of a heatmap represent missing or no signal for the given data type. Far left bars represent the inactive genes (purple), known escaper genes (green), pseudoautosomal regions (PAR [orange]), and non-PAR (dark gray). Top bars are the mean tissue (i.e., column) value for each data type's fold-change values (black is missing/no data). Heatmap columns are sorted by tissue means. Overall, phased RNA-seq read counts and active chromatin marks show a higher signal on the same allele. It is unclear whether repressive chromatin marks are more pervasive on the opposite allele due to the sparsity of the data (see data matrix in Figure S1A).

(B) Similarity of haplotype-specific gene expression and chromatin assays on chrX. Cosine similarity between the mean tissue values in Data S14A for individuals ENC-003 (*left*) and ENC-004 (*right*). Phased RNA-seq read counts are shown to be more similar to active chromatin marks than repressive chromatin marks for both individuals.

(C) Identifying potential X-chromosome inactivation (XCI) escaper genes. GENCODE v24 genes on chrX were classified as either inactive (purple) or potential escaper (green) genes based on the RNA-seq fold-change values shown in Data S14A for individuals ENC-003 (left column) and ENC-004 (right column). Escaper genes were identified as genes showing balanced expression (haplotype ratio within 30%) in a majority of their expressed tissues. To support pan-tissue chromatin analysis and avoid spurious observations from lowly expression genes, the analysis was limited to genes expressed in eight or more tissues. Identified individual-specific, potential escaper genes are listed below each scatterplot. All potential escaper genes have previously been found to escape XCI per the provided references: 1. Mugford et al., 2014 [8]; 2. Tukiainen et al., 2017 [9]; 3. Garieri et al., 2018 [10]; 4. Zhang et al., 2020 [11]; 5. Zito et al., 2021 [12]; and 6. Werner et al., 2022 [13].

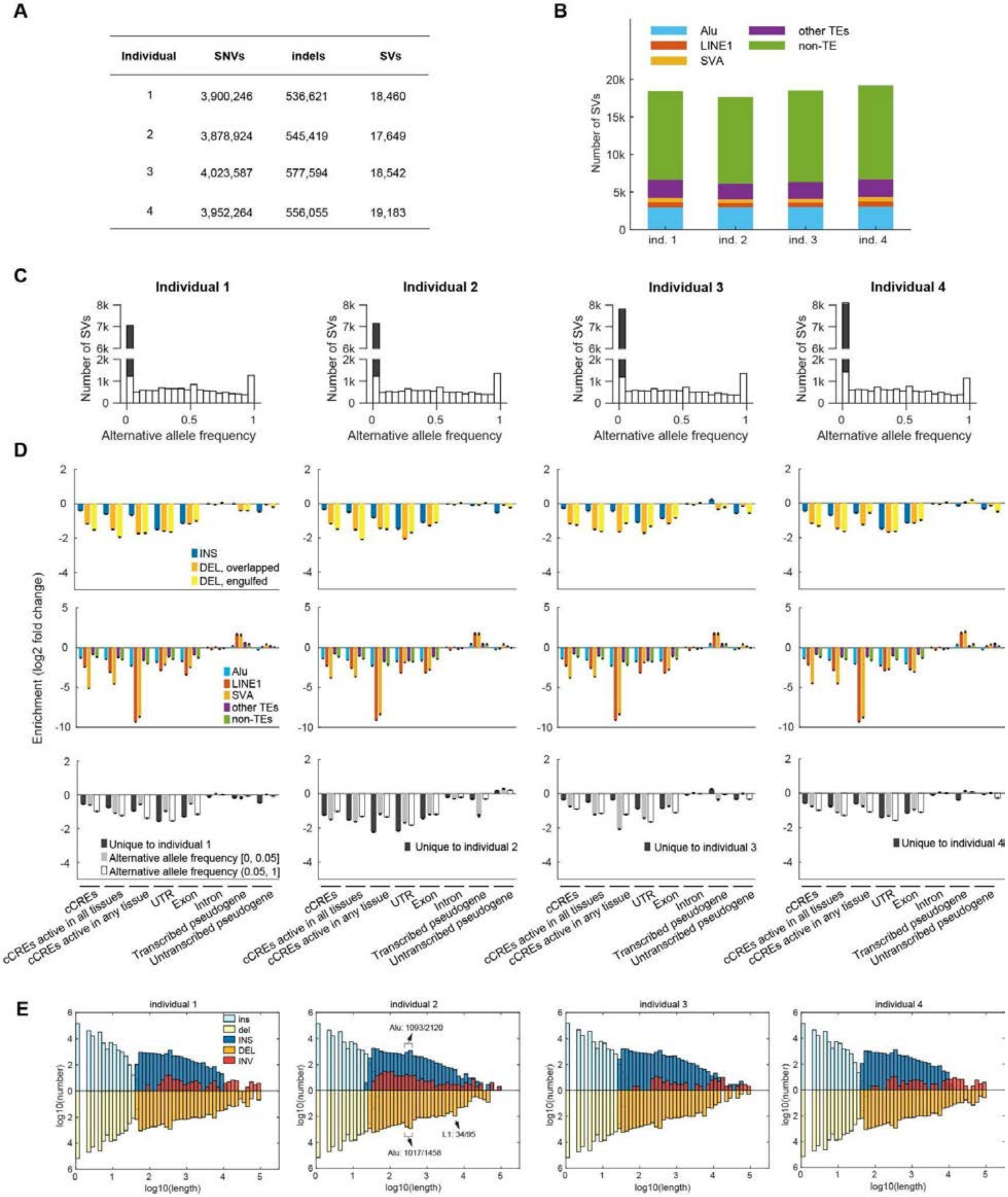
(D) Example potential X-chromosome inactivation escaper genes. Three previously known escaper genes (DHRSX, KDM6A, and RP11-706O15.1 - left, middle, and right columns) identified from EN-TEx phased RNA-seq data analysis for both individuals ENC-003 (top row) and ENC-004 (bottom row). Barplots represent the (im)balanced expression of a gene using hap1 (red) and hap2 (blue) RNA-seq read counts for each individual in a given tissue. Undrawn bar plots represent missing/no data for that individual and tissue combination. Shown escaper genes demonstrate balanced expression across tissues for each individual.

(E) Heatmap to show haplotype specificity of chrX for all assays and tissues from individual 3 We detected that ~21% of the accessible chrX genes have significantly imbalanced expression levels between the two haplotypes. Orange squares indicate more expression and binding peaks in hap2, whereas blue squares indicate more expression and binding peaks in hap1. Green squares indicate that the expression and binding are balanced between the two haplotypes. Light gray squares indicate that the number of data points is small and, consequently, we cannot conclude which haplotype has more expression and binding. Dark gray squares indicate that data are not available for a given assay and tissue. Our findings (Panel A-E) are consistent with a recent study [13] that demonstrated X-inactivation is shared across many tissues using GTEx and EN-TEx data. This study suggested that X-inactivation is completed before the germ-layer specification. Therefore, any skew in selecting which X-chromosome is activated propagates to the ectoderm, endoderm, and mesoderm, resulting in observations of the same skew across many tissues. The way in which cells activate which X chromosome is a random process that follows a probabilistic distribution before the specification of the germ layer. Developed tissues are much more likely to have cells with the same activated X-chromosome [13]. Our findings here are also consistent with this study regarding which tissues show bias towards which haplotype. For example, Werner et al. [13] show that the liver and ovary are ranked 46th and 42nd (out of 46 GTex tissues) in terms of their skew being the same direction as other tissues.

(F) Chromosome painting of chrX using RNA-seq and ChIP-seq in both haplotypes of individual 3 in two tissues. This plot shows that the active haplotype is hap2 in chrX of individual 3, as there is more activity in hap2.

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

(H) Coordinated AS activity in chrX. Similar to Figure 2A, we show the differences in the levels of gene expression, H3K27ac, and H3K27me3 between the two X chromosomes in the thyroid gland of individual 4. The high RNA expression levels from hap1 indicates that this chrX is active. Note the higher H3K27ac levels and lower H3K27me3 levels in this chrX. Note, only individual 3 was shown in Figure 2A.



Individual	SNVs	indels	SVs
1	3,900,246	536,621	18,460
2	3,878,924	545,419	17,649
3	4,023,587	577,594	18,542
4	3,952,264	556,055	19,183

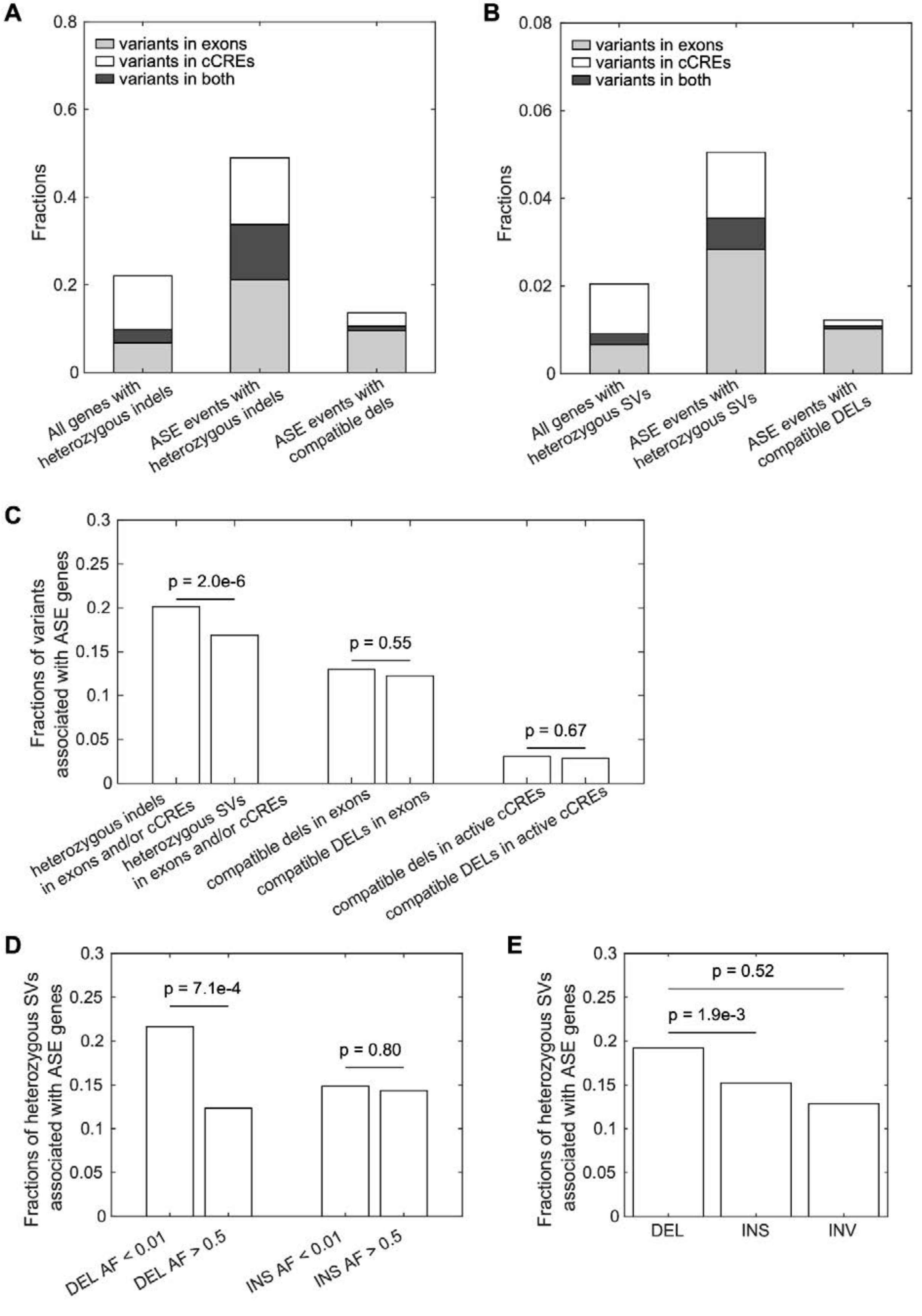
Data S15. Analysis of SVs, related to Figure 1 and STAR Methods "SVs" Section

(A) Number of genomic variants in the four individuals.

(B) Number of SVs associated with transposable elements.

(C) Allele frequencies of SVs in the European population calculated by overlapping with the results from Audano et al. [14]. SVs that have no overlap with the results from Audano et al. are shaded in the first bin.

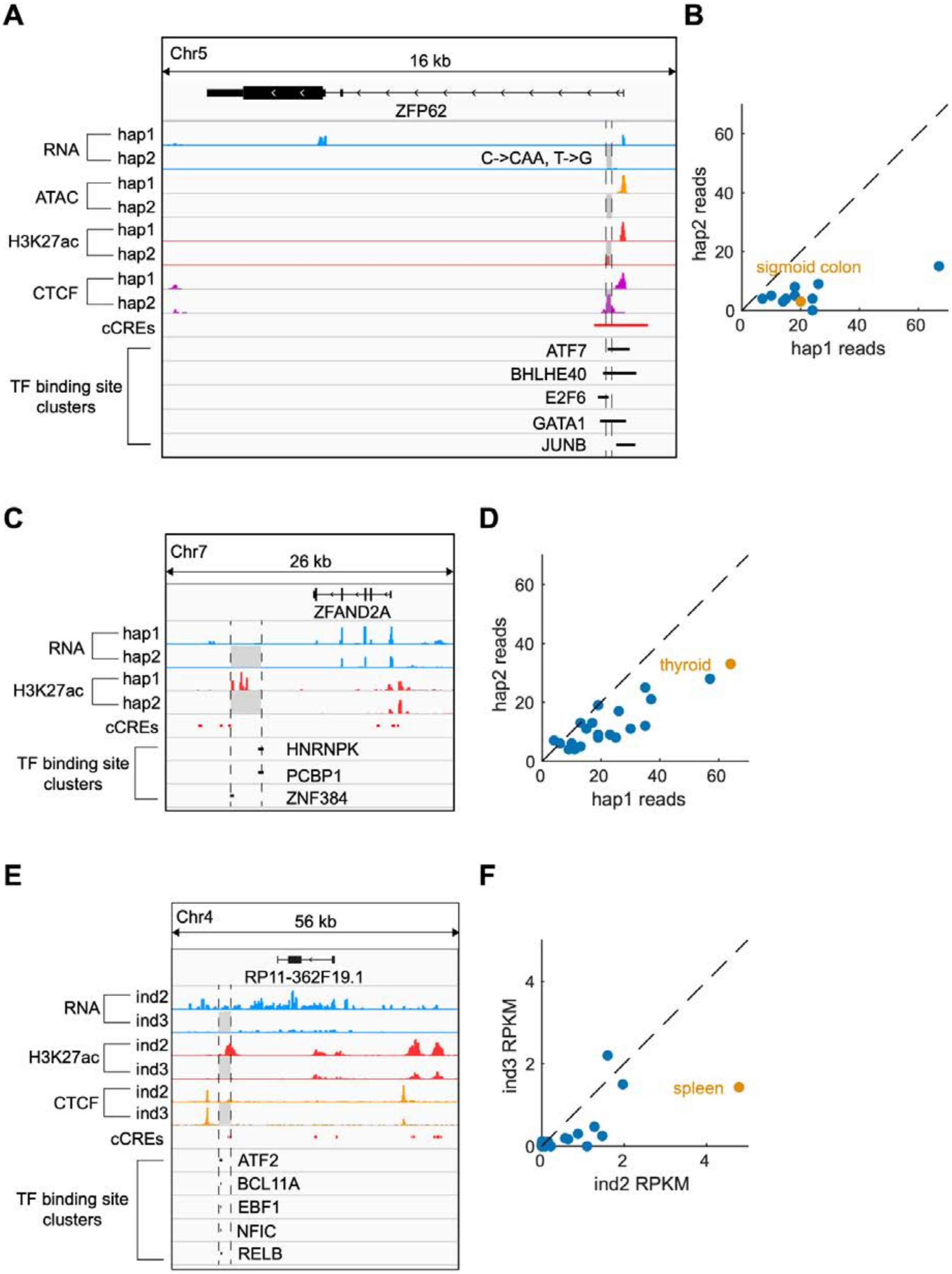
(D) Overlaps between SVs and functional genomic regions. We shuffled the locations of the SVs (see STAR Methods “SVs” Section) to determine whether SVs are enriched or depleted in a given type of genomic region. For DELs, we consider cases in which a DEL partially overlaps with a given genomic region (DEL, partial) and cases in which a DEL is engulfed by a given genomic region (DEL, engulfed).
(E) Lengths of indels and SVs in the four individuals. The peaks around 10^2.5 bp and 10^3.7 bp are due to Alu and LINE1. In individual 2, we show the fractions of SVs associated with Alu and LINE1 in the corresponding bins. Note that these fractions are much higher than those in (B).

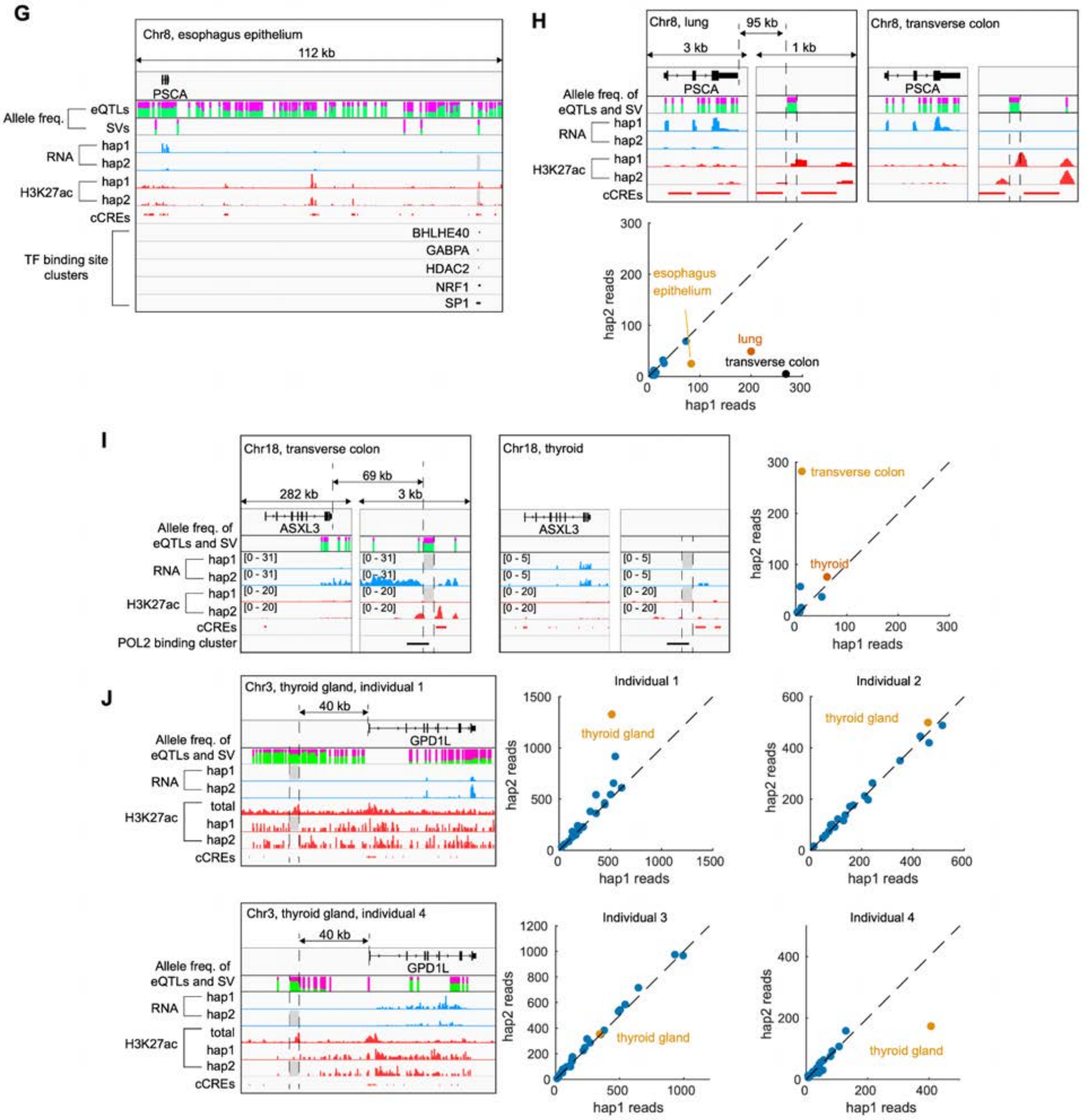
A
0.8
0.6
0.4
0.2
0
Fractions
variants in exons
variants in cCREs
variants in both
All genes with heterozygous indels
ASE events with heterozygous indels
ASE events with compatible dels
B
0.08
0.06
0.04
0.02
0
Fractions
variants in exons
variants in cCREs
variants in both
All genes with heterozygous SVs
ASE events with heterozygous SVs
ASE events with compatible DELs
C
0.3
0.25
0.2
0.15
0.1
0.05
0
Fractions of variants associated with ASE genes
p = 2.0e-6
p = 0.55
p = 0.67
heterozygous indels in exons and/or cCREs
heterozygous SVs in exons and/or cCREs
compatible dels in exons
compatible DELs in exons
compatible dels in active cCREs
compatible DELs in active cCREs
D
0.3
0.25
0.2
0.15
0.1
0.05
0
Fractions of heterozygous SVs associated with ASE genes
p = 7.1e-4
p = 0.80
DEL AF < 0.01
DEL AF > 0.5
INS AF < 0.01
INS AF > 0.5
E
0.3
0.25
0.2
0.15
0.1
0.05
0
Fractions of heterozygous SVs associated with ASE genes
p = 0.52
p = 1.9e-3
DEL
INS
INV

Data S16. Association between ASE events and indels or SVs, related to Figure 2 and STAR Methods “SVs” Section

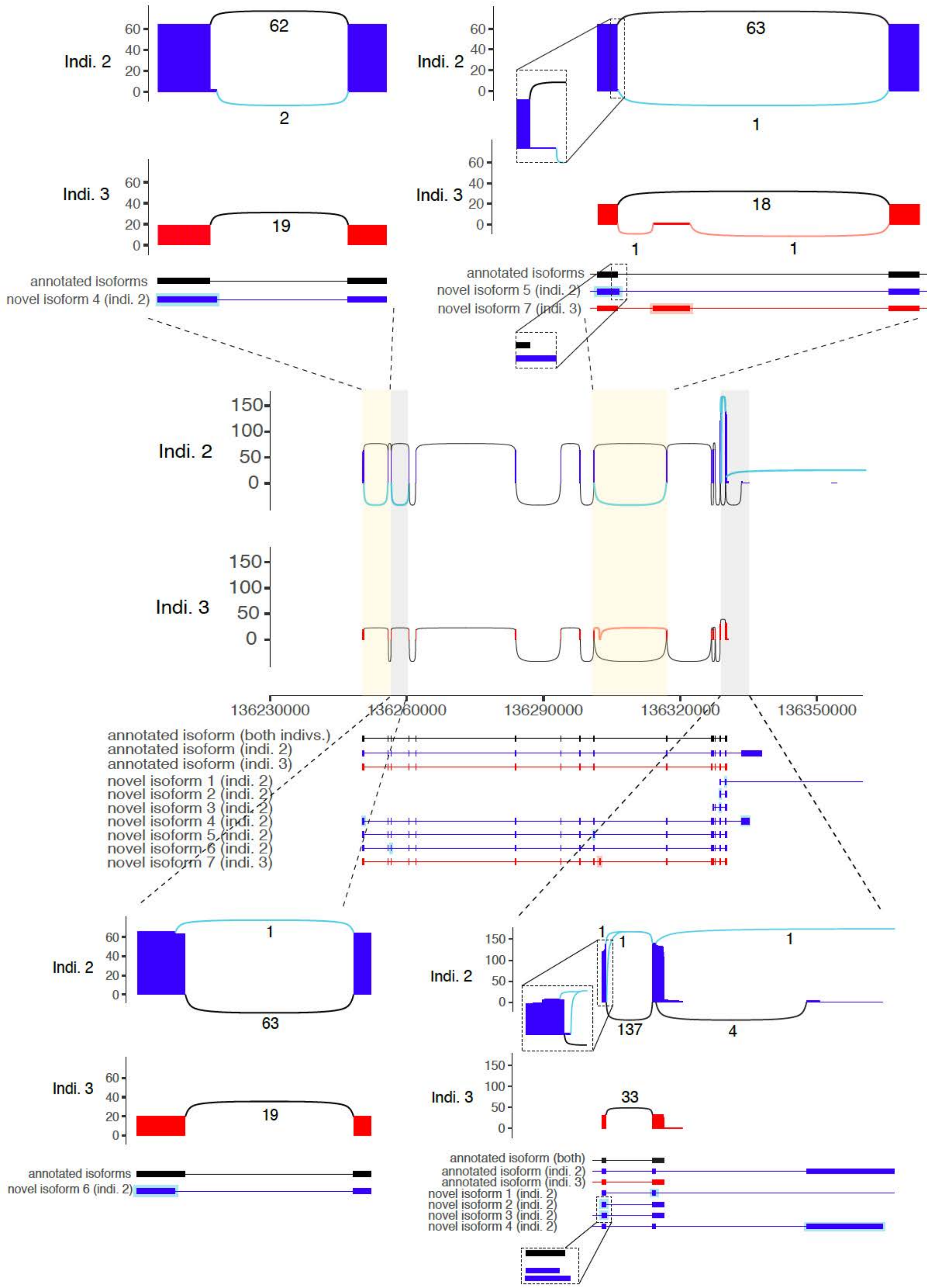
(A) - (B) ASE events associated with indels and SVs. For an ASE event in a given tissue of a given individual, we looked for heterozygous indels and SVs that intersect with the exons of the ASE gene and/or with cCREs within +/- 10 Kb of the gene’s TSS. For comparison, we also show the fractions of genes (ASE or not) whose exons and/or nearby cCREs intersect with a heterozygous indel and SV. If the heterozygous small deletion and DEL had clear genotypes, we further evaluated whether they are compatible with the ASE event, i.e., the presence of the variants in exons and/or the tissue-specific active cCREs (STAR Methods “SVs” Section) should reduce gene expression. Since the exact breakpoints of SVs are often uncertain and SVs may disrupt nearby regions, we expanded the location of each SV by 100 bp upstream and 100 bp downstream when intersecting it with exons and cCREs. (A) The fractions of ASE events associated with indels and compatible deletions. (B) The fractions of ASE events associated with SVs and compatible DELs. In both panels, we pooled the ASE events and ASE events with associated variants from all tissues of all four individuals before calculating the fractions. Specifically in (B), we found 42 ASE events that are associated with compatible DELs in the tissue-specific active cCREs, 323 associated with compatible DELs in exons, and 22 associated with compatible DELs in both.

(C) - (E) Indels and SVs associated with ASE. Similar to (A) and (B), we looked for heterozygous indels and SVs that intersect with at least one of two genomic regions: an exon, and cCREs that are within +/- 10 Kb of a TSS. Among these variants, we calculated the fractions of those where the associated gene shows ASE in at least one tissue of the individual who carries the variants. We expanded the location of the SV by 100 bp upstream and 100 bp downstream before intersecting with exons and cCREs. (C) The fractions of indels and SVs associated with ASE. For heterozygous deletions and DELs that have clear genotypes, we further evaluated whether they are compatible with the associated ASE (panels A and B). The fractions of compatible variants among those that intersect with an exon or any tissue-specific active cCREs are shown in separate groups. (D) The fractions of rare (allele frequency (AF) < 0.01) and common (AF > 0.05) SVs that are associated with ASE among those that intersect with an exon and/or cCREs. Because the AF of INVs is below 0.01 in each of the four individuals, we could not compare rare vs. common INVs. (E) The fractions of DELs, INSs, and INVs that are associated with ASE among each type of SV that intersect with an exon and/or cCREs. In all panels, we pooled variants of interest from all tissues of all four individuals before calculating the fractions. Differences between fractions were tested via the $\chi 2$ test.

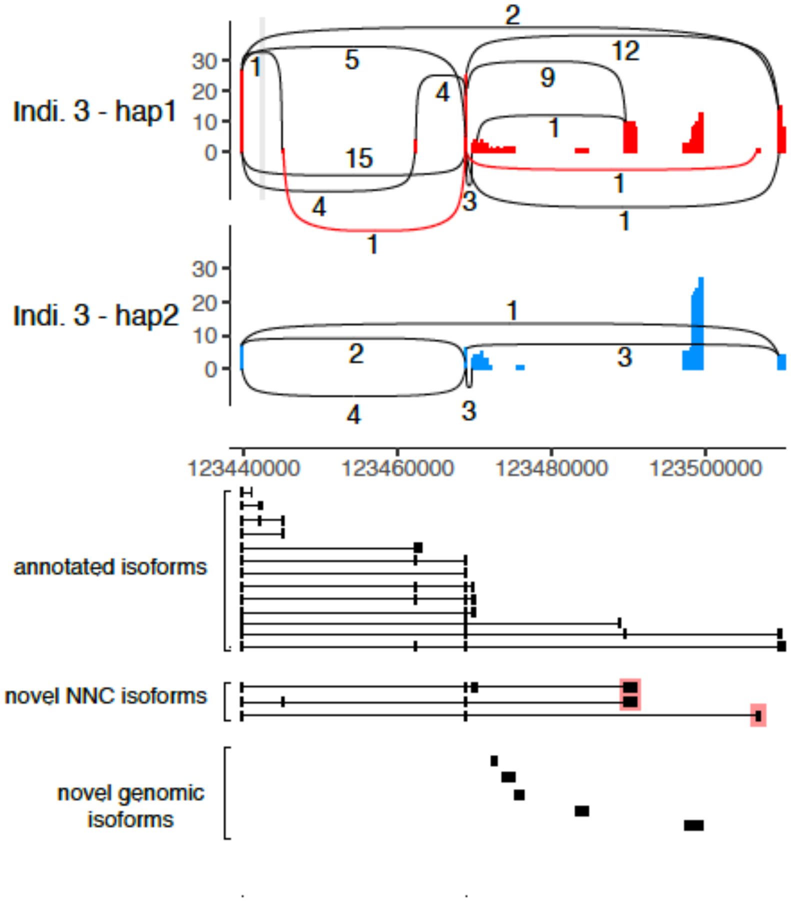

A
Chr5
16 kb
ZFP62
RNA
hap1
hap2
C->CAA, T->G
ATAC
hap1
hap2
H3K27ac
hap1
hap2
CTCF
hap1
hap2
cCREs
TF binding site clusters
ATF7
BHLHE40
E2F6
GATA1
JUNB
B
hap2 reads
hap1 reads
sigmoid colon
0
20
40
60
C
Chr7
26 kb
ZFAND2A
RNA
hap1
hap2
H3K27ac
hap1
hap2
cCREs
TF binding site clusters
HNRNPK
PCBP1
ZNF384
D
hap2 reads
hap1 reads
thyroid
0
20
40
60
E
Chr4
56 kb
RP11-362F19.1
RNA
ind2
ind3
H3K27ac
ind2
ind3
CTCF
ind2
ind3
cCREs
TF binding site clusters
ATF2
BCL11A
EBF1
NFIC
RELB
F
ind3 RPKM
ind2 RPKM
spleen
0
2
4


G
Chr8, esophagus epithelium
112 kb
PSCA
Allele freq.
eQTLs
SVs
RNA
hap1
hap2
H3K27ac
hap1
hap2
cCREs
TF binding site clusters
BHLHE40
GABPA
HDAC2
NRF1
SP1
H
Chr8, lung
95 kb
3 kb
1 kb
PSCA
Allele freq. of eQTLs and SV
RNA
hap1
hap2
H3K27ac
hap1
hap2
cCREs
Chr8, transverse colon
PSCA
hap2 reads
hap1 reads
0
100
200
300
esophagus epithelium
lung
transverse colon
I
Chr18, transverse colon
69 kb
282 kb
3 kb
ASXL3
Allele freq. of eQTLs and SV
RNA
hap1
hap2
H3K27ac
hap1
hap2
cCREs
POL2 binding cluster
[0 - 31]
[0 - 20]
Chr18, thyroid
ASXL3
[0 - 5]
[0 - 20]
hap2 reads
hap1 reads
transverse colon
thyroid
J
Chr3, thyroid gland, individual 1
40 kb
GPD1L
Allele freq. of eQTLs and SV
RNA
hap1
hap2
H3K27ac
total
hap1
hap2
cCREs
Individual 1
thyroid gland
0
500
1000
1500
hap2 reads
hap1 reads
Individual 2
thyroid gland
0
200
400
600
Chr3, thyroid gland, individual 4
40 kb
GPD1L
Individual 3
thyroid gland
0
200
400
600
800
1000
1200
500
Individual 4
thyroid gland
0
200
400

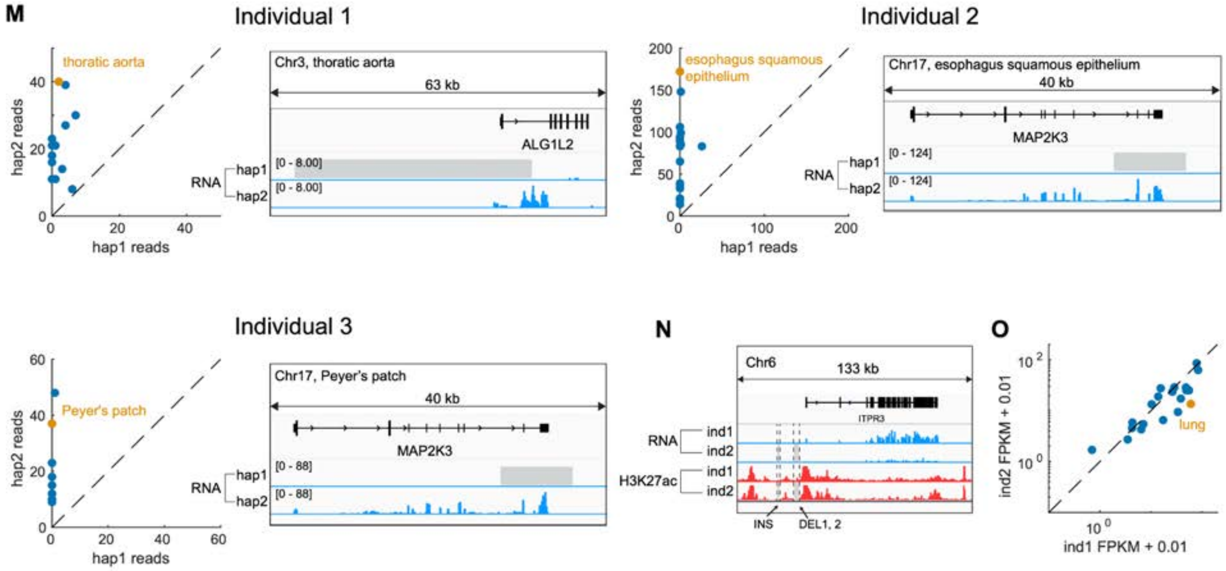
K


Indi. 2
60
40
20
0
62
2
Indi. 3
19
annotated isoforms
novel isoform 4 (indi. 2)
63
1
18
1
1
annotated isoforms
novel isoform 5 (indi. 2)
novel isoform 7 (indi. 3)
Indi. 2
150
100
50
0
Indi. 3
136230000
136260000
136290000
136320000
136350000
annotated isoform (both indivs.)
annotated isoform (indi. 2)
annotated isoform (indi. 3)
novel isoform 1 (indi. 2)
novel isoform 2 (indi. 2)
novel isoform 3 (indi. 2)
novel isoform 4 (indi. 2)
novel isoform 5 (indi. 2)
novel isoform 6 (indi. 2)
novel isoform 7 (indi. 3)
1
63
19
annotated isoforms
novel isoform 6 (indi. 2)
1
1
1
137
4
33
annotated isoform (both)
annotated isoform (indi. 2)
annotated isoform (indi. 3)
novel isoform 1 (indi. 2)
novel isoform 2 (indi. 2)
novel isoform 3 (indi. 2)
novel isoform 4 (indi. 2)

L



Indi. 3 - hap1
30
20
10
0
2
1
5
12
9
4
1
15
4
3
1
1
1
Indi. 3 - hap2
30
20
10
0
1
2
3
4
3
123440000
123460000
123480000
123500000
annotated isoforms
novel NNC isoforms
novel genomic isoforms



Data S17. Examples of SVs associated with changes in gene expression, related to Figure 2, Figure S4, and STAR Methods "SVs" Section

(A) - (B) An indel that potentially changes gene expression. (A) In the sigmoid colon of individual 2, the gene *ZFP62* has lower expression in hap2. 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 an 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 gray 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.

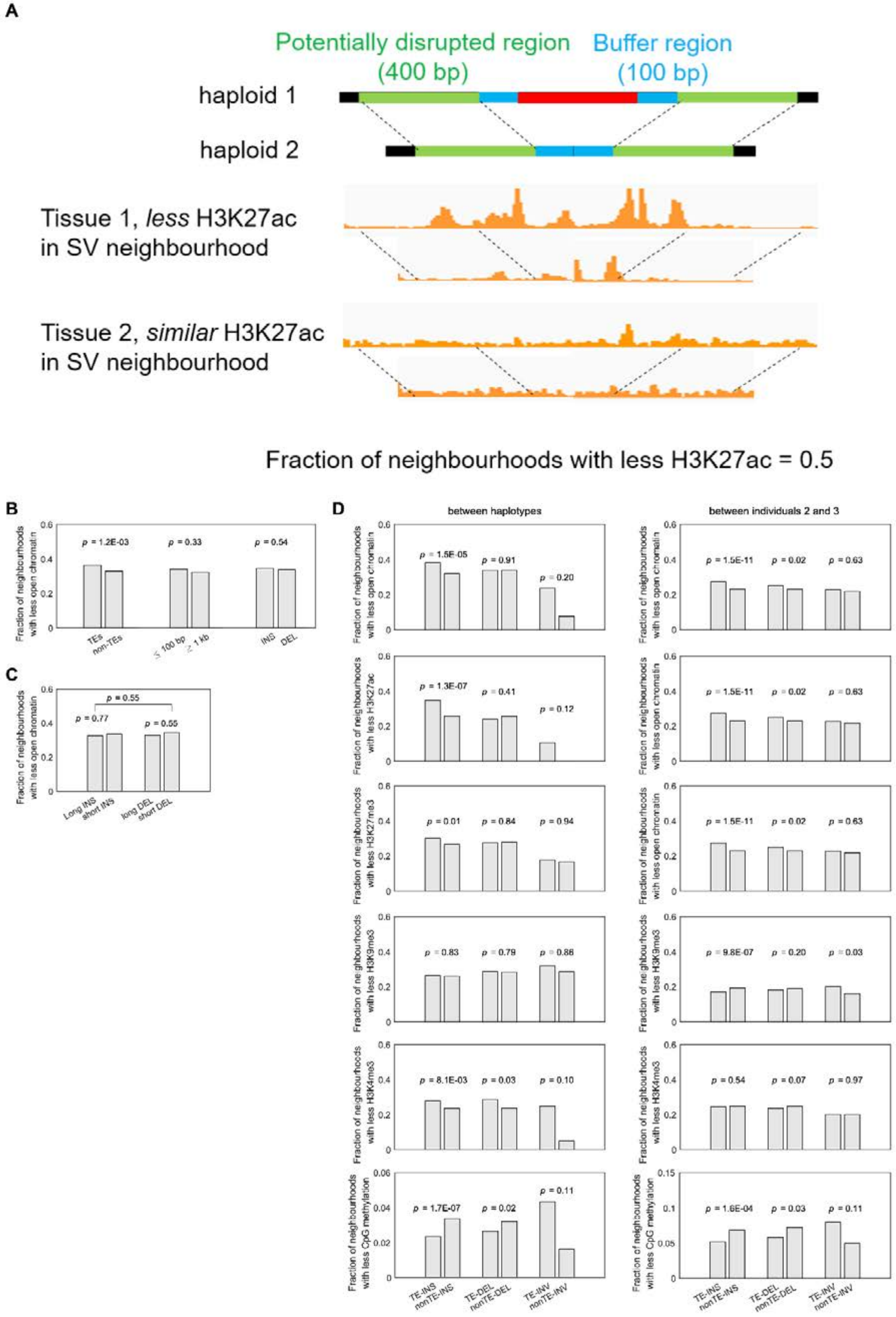
(C) - (D) Shadow figure associated with Figure 2D. (C) Similar to Figure 2D, the deletion in hap2 can disrupt cCREs identified in the thyroid and the binding of several TFs. (D) ZFAND2A has lower hap2 expression among multiple tissues, suggesting that the deletion may have a global effect on the expression of this gene. Note that this is consistent with the results in Chiang et al. [15]. Note that the p-value of the AS expression in Figure 2D was based on beta-binomial distribution.

(E) - (F) Shadow figure associated with Figure S4C. (E) Similar to Figure S4C, the deletion in hap2 can disrupt spleen-specific cCREs and the binding of several TFs. (F) In multiple tissues, *RP11-362F19.1* has lower expression in individual 3 than in individual 2, suggesting that the deletion may have a global effect on the expression of this gene.

(G) - (J) SVs potentially linked to eQTLs. Panels (G) and (H) are shadow figures of Figure 2E. (G) This panel is the same as Figure 2E, but shows a panoramic view near the gene *PSCA*, including additional eQTLs that are compatible with the ASE of *PSCA*. The allele frequencies of the hap2 alleles at these eQTL sites are shown as the heights of the green bars. Note, the height of a green bar plus its corresponding magenta bar sum to 1. Similar results were observed in two other tissues from individual 3. SVs near *PSCA* and their allele frequencies are also shown. The left four SVs are deletions in hap1, and the rightmost SV is the hap2 deletion shown in Figure 4E. cCREs and TF binding sites that can potentially be disrupted by the deletion of interest are

shown. (H) *PSCA* also has a lower expression of hap2 in the lung and transverse colon of individual 3. In both tissues, the deletion has an allele frequency similar to that of some of the tissue-specific eQTLs compatible with the ASE of *PSCA*; moreover, this deletion appears to remove an H3K27ac peak in hap2, potentially causing the reduced expression of *PSCA*. Imbalance in the ASE of *PSCA* appears to be restricted to three tissues shown in (G) and (H). (I) Another example of a deletion that may be linked with compatible eQTLs of *ASXL3*. In the transverse colon of individual 2, *ASXL3* has lower expression in hap1. The relevant deletion occurs 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 the thyroid than in transverse colon, suggesting an association between the activity of this cCRE with *PSCA* expression. Imbalance in the ASE of *ASXL3* appears to be tissue specific. (J) A known SV-eQTL of *GPD1L* [15] in the thyroid gland. Individuals 1 and 4 are heterozygous for this deletion, but the former has it on hap1 and the latter has it on hap2. As shown in the signal tracks, hap1 of individual 1 and hap2 of individual 2 show lower *GPD1L* expression than the other haplotype in the respective individual. There appears to be an active enhancer 40 kb upstream of *GPD1L*, as indicated by the total H3K27ac ChIP-seq signal (fold-change of the total reads from both haplotypes over the control) and by the locations of the active cCREs in the thyroid gland. This enhancer is removed by the deletion, potentially reducing the expression levels of *GPD1L* in the corresponding haplotype. The effect of the deletion is not obvious from the haplotype-specific H3K27ac ChIP-seq reads. This is likely because the region does not have enough SNVs, which are required to map ChIP-seq reads to both haplotypes. We note that individuals 2 and 3 are homozygous for this deletion, potentially explaining the lack of ASE of *GPD1L* in these two individuals. Note that the p-value of the AS expression in Figure 2E was based on beta-binomial distribution.
(K) Novel splicing variants of *PCCB.* Shadow figure for Figure S4B. 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 provides a representation of the whole gene. 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 (magnifications of specific regions are provided, as well as the number of reads supporting each connection). The exonic structures of annotated and novel isoforms are 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 [16]. Specifically, a Swan gene report was generated for each individual by inputting transcriptome annotation and quantification files available, from long-read RNA-seq experiments, in the ENCODE portal (https://www.encodeproject.org/). These plots were obtained using ggashimi [17].
(L) Novel splicing variants of *TRDN-AS1.* Sashimi plot and exonic structure representation of the lncRNA *TRDN-AS1* isoforms in individual 3 in heart left ventricle. This gene carries a heterozygous deletion on hap1 (highlighted in gray) and shows ASE in the right atrium auricular region (with hap1 being more highly expressed than hap2). For the sashimi representation, reads available from long-read RNA-seq experiments (see the ENCODE portal) were phased to the two haplotypes using heterozygous SNVs that overlap with the gene's exons. Read phasing was performed with ASCIIGenome (https://github.com/dariober/ASCIIGenome/) [18]. Long-read

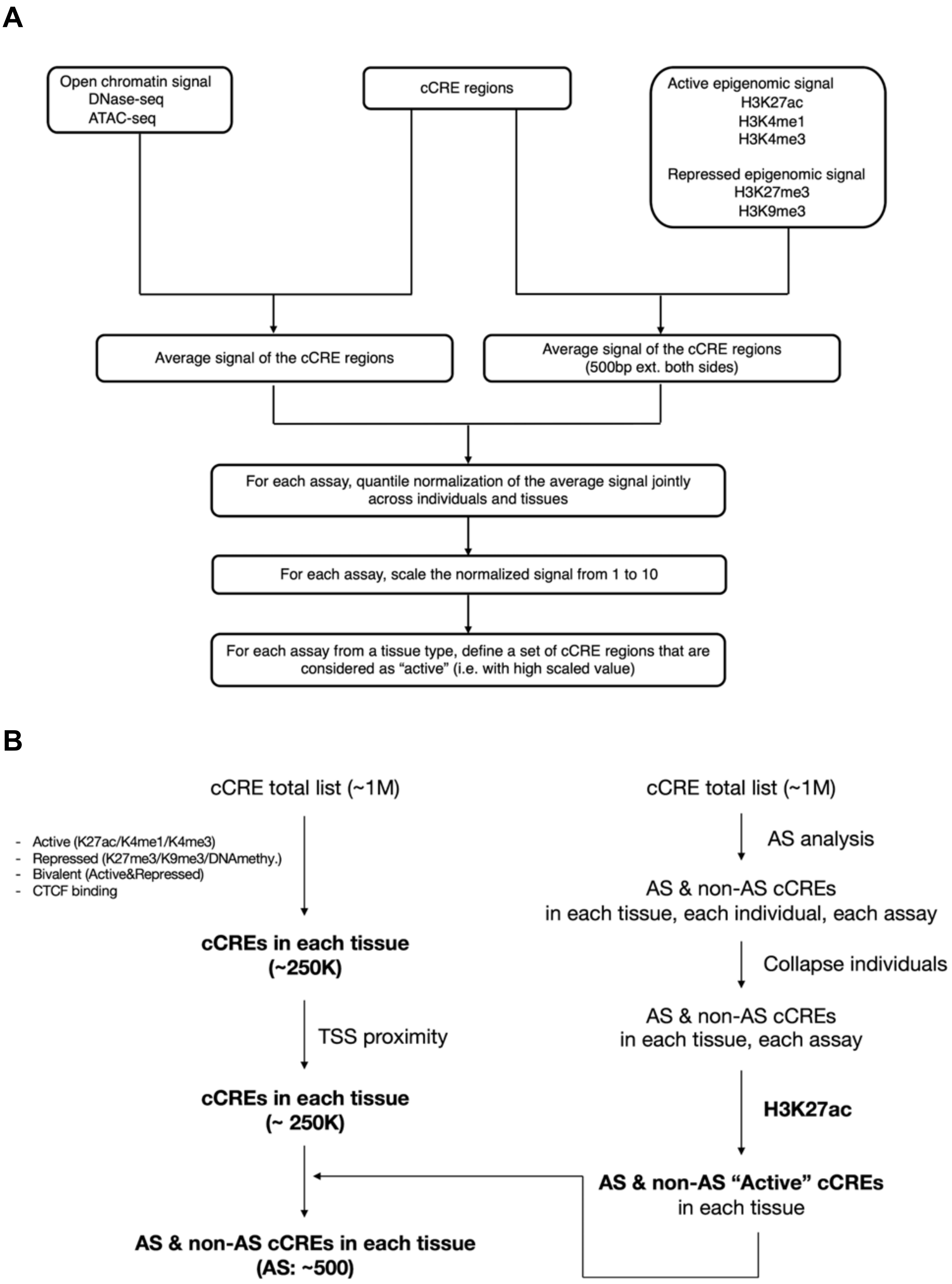
RNA-seq reads show consistently higher expression of hap1 compared with hap2. Moreover, reads mapping to hap1 give rise to two novel splicing junctions (represented by red arcs) and two novel exons (highlighted in red in the exonic structure representation at the bottom). Annotated and novel isoforms were retrieved for each individual using Swan [16]. A Swan gene report was generated for individual 3 by inputting transcriptome annotation and quantification files available from long-read RNA-seq experiments in the ENCODE portal. Only novel “not in catalog” (NNC) and genomic isoforms are shown. These plots were obtained using ggashimi [17].
(M) Null alleles potentially caused by deletion of entire exons. These examples show genes whose expression comes almost exclusively from hap2 in multiple tissues of a given individual. Further analysis revealed that each example contains a deletion in hap1 that removes multiple exons of the given gene in the given individual. We did not find null alleles associated with SV deletions in individual 4.
(N) - (O) Deletions in individual 2 but not individual 1. (N) One homozygous deletion and one heterozygous deletion upstream of *ITPR3* were found in individual 2 (the two variants are very close and are shown together by a single gray box), but are missing in individual 1. The deletions knock out part of the H3K27ac peak near the TSS of *ITPR3*, potentially reducing the gene’s expression in the lung of individual 2 compared with individual 1. (O) Across multiple tissues, the expression levels of *ITPR3* appear to be lower in individual 2 than individual 1.

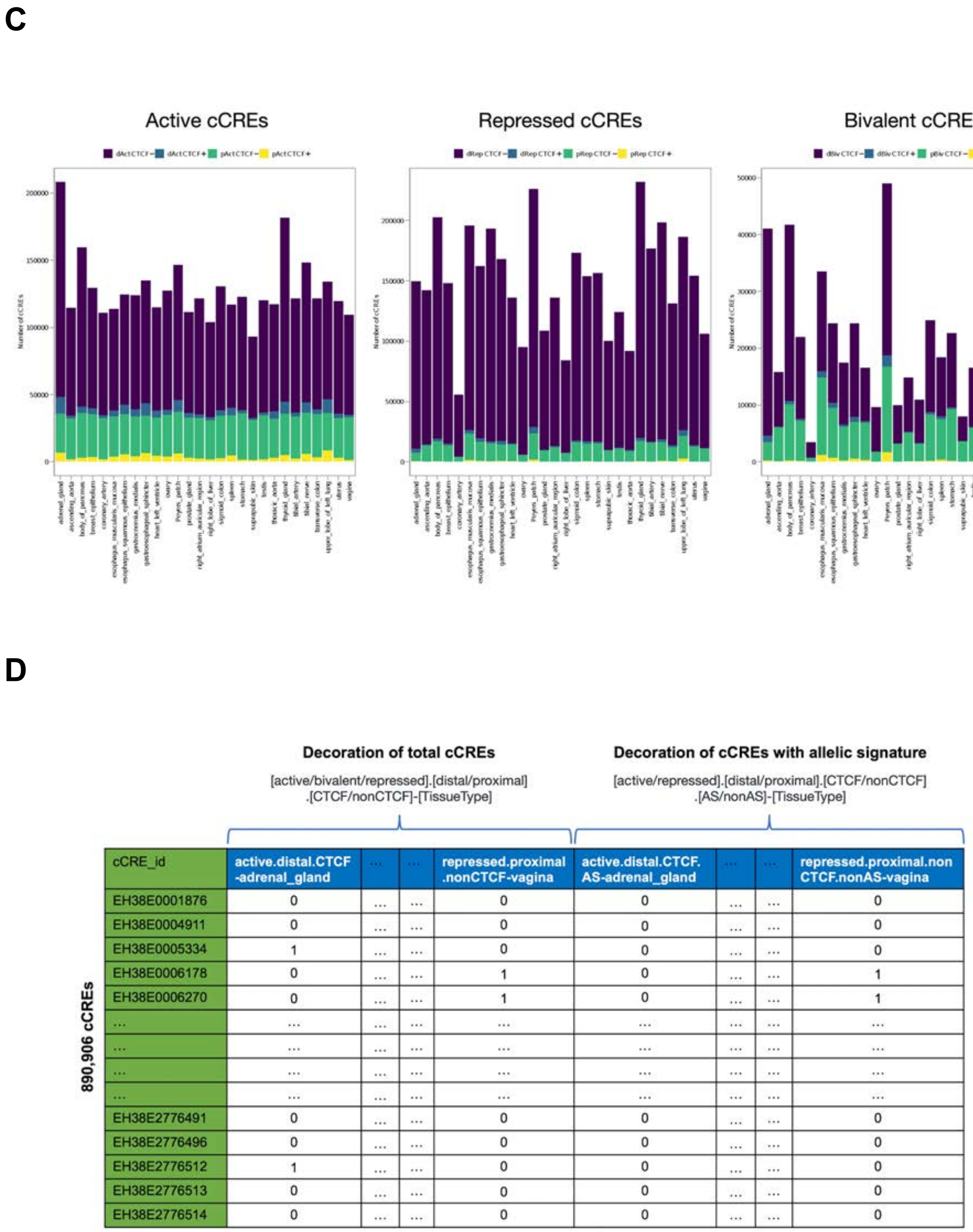

A
Potentially disrupted region
(400 bp)
Buffer region
(100 bp)
haploid 1
haploid 2
Tissue 1, *less* H3K27ac
in SV neighbourhood
Tissue 2, *similar* H3K27ac
in SV neighbourhood
Fraction of neighbourhoods with less H3K27ac = 0.5
B
Fraction of neighbourhoods with less open chromatin
p = 1.2E-03
p = 0.33
p = 0.54
TEs
non-TEs
≤ 100 bp
≥ 1 kb
INS
DEL
C
Fraction of neighbourhoods with less open chromatin
p = 0.55
p = 0.77
p = 0.55
Long INS
short INS
long DEL
short DEL
D
between haplotypes
between individuals 2 and 3
Fraction of neighbourhoods with less open chromatin
p = 1.5E-05
p = 0.91
p = 0.20
p = 1.5E-11
p = 0.02
p = 0.63
Fraction of neighbourhoods with less H3K27ac
p = 1.3E-07
p = 0.41
p = 0.12
Fraction of neighbourhoods with less open chromatin
p = 1.5E-11
p = 0.02
p = 0.63
Fraction of neighbourhoods with less H3K27me3
p = 0.01
p = 0.84
p = 0.94
Fraction of neighbourhoods with less open chromatin
p = 1.5E-11
p = 0.02
p = 0.63
Fraction of neighbourhoods with less H3K9me3
p = 0.83
p = 0.79
p = 0.86
p = 9.8E-07
p = 0.20
p = 0.03
Fraction of neighbourhoods with less H3K4me3
p = 8.1E-03
p = 0.03
p = 0.10
p = 0.54
p = 0.07
p = 0.97
Fraction of neighbourhoods with less CpG methylation
p = 1.7E-07
p = 0.02
p = 0.11
p = 1.6E-04
p = 0.03
p = 0.11
TE-INS
nonTE-INS
TE-DEL
nonTE-DEL
TE-INV
nonTE-INV
0
0.2
0.4
0.6
0.02
0.04
0.06
0.05
0.1
0.15

Data S18. Potential perturbations of SVs to the chromatin states of neighboring regions, related to Figure 2 and STAR Methods "SVs" Section

(A) Calculating changes in the chromatin state in the SV neighborhood. Using H3K27ac levels as an example, haploid 1 of individual 1 carries a deletion (red bar) while haploid 2 is wild type at the same locus; therefore, we compared the chromatin states in the two green regions between the two haploids. In tissue 1, the H3K27ac levels in the green region are lower in haploid 1, whereas the H3K27ac levels in tissue 2 are similar in both haploids. Therefore, only half of the neighborhoods of this deletion show a reduction in H3K27ac levels. Similar analyses can be performed between two individuals by substituting the two haploids with two individuals.

(B) - (D) Changes in the chromatin state of SV neighborhoods. Similar to Figure 2F, we investigated whether the presence of an SV changes the chromatin state of nearby regions and whether these changes are associated with different characteristics of the SVs. The genomic regions neighboring the TE insertions show reduced chromatin accessibility more often than those of the non-TE insertions. This difference is not observed between TE deletions and non-TE deletions. The change in accessibility is determined by comparing the accessibility (from ATAC-seq) between the two haplotypes of each individual, taking the comparison of the haplotype without the SV as a reference (panel A and STAR Methods "SVs"). P-values are based on the Chi-squared test. (B) The reduction in the chromatin openness near SVs does not differ by SV length or SV type, (C) nor does it differ between long (> 1 Kb) and short (< 100 bp) SV insertions. (D) Changes in other chromatin states near SVs. Left panels: changes in the chromatin states near heterozygous SVs in all four individuals. The changes were calculated by comparing the chromatin states between two haploids of the same individual. Right panels: changes in the chromatin states near SVs that are only present in either individual 2 or 3 (but not both). The changes were calculated by comparing the chromatin states between two individuals. p-values of the difference in fractions were calculated by the $\chi 2$ test.


A
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)
B
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



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

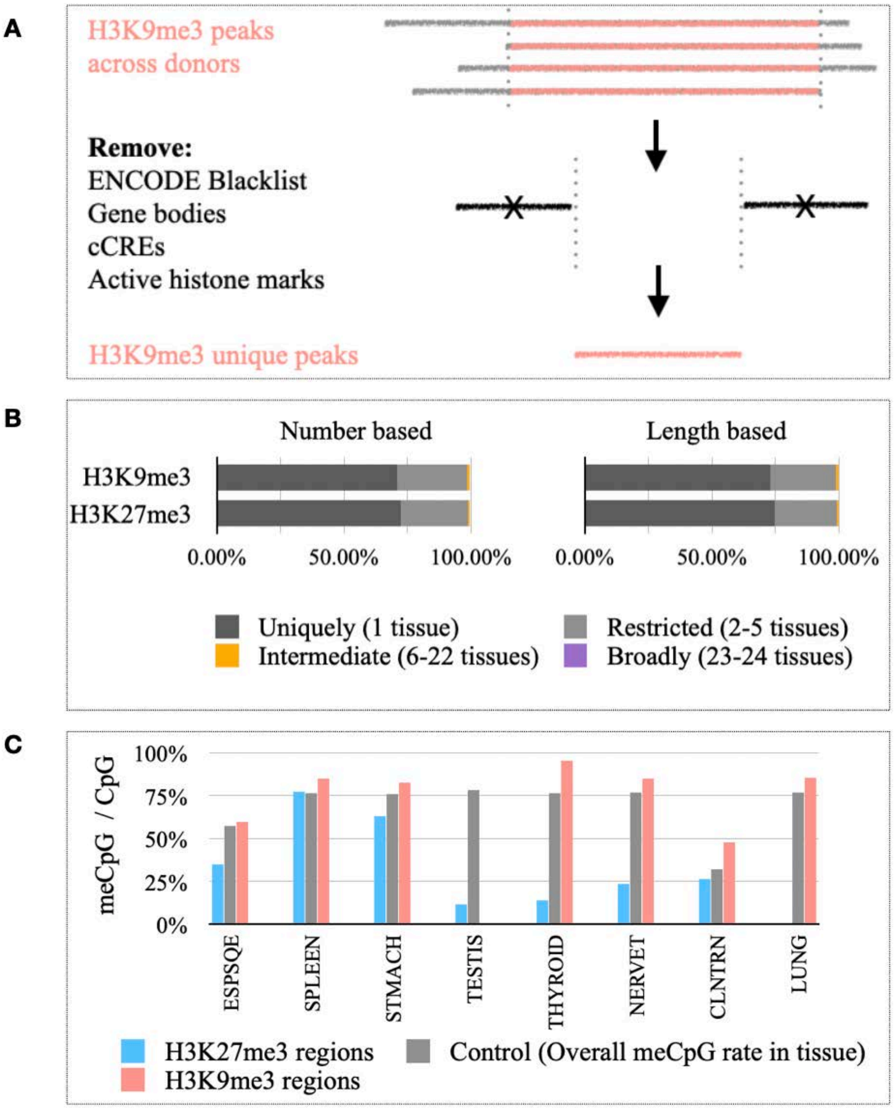
Data S19. cCRE decoration, related to Figure 3, Figure S5, and STAR Methods “Decoration Process” Section

(A) Data preprocessing. We computed the average signal for each cCRE region using the datasets from DNase-seq, ATAC-seq, and five histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K27me3, and H3K9me3). For DNase-seq and ATAC-seq, the signals were averaged across the genomic positions of the cCRE regions. The signals of histone modifications were averaged across the genomic positions of the cCRE regions with a 500 bp extended region on each side. For each assay, we performed quantile normalization on the average signal from the cCRE regions jointly across all of the biosamples. Then, we scaled the normalized signal from 1 to 10, and defined a set of “active” cCREs for each assay from each tissue type.

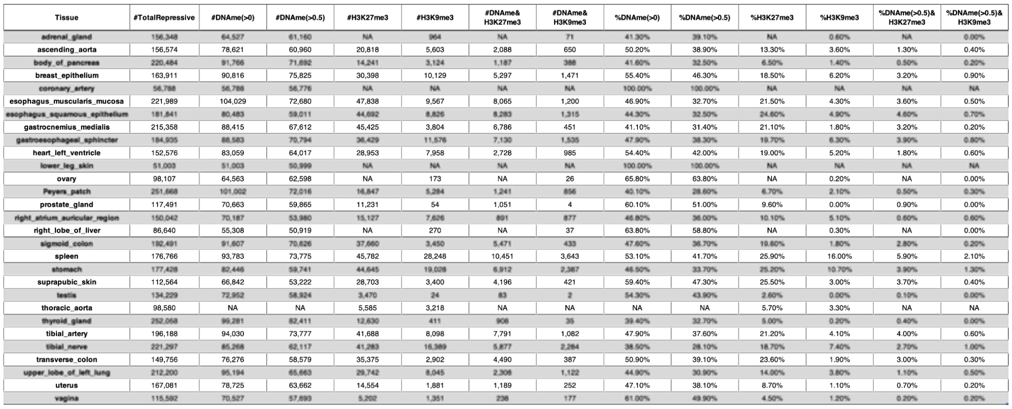
(B) Framework of cCRE decoration. We decorated the cCREs from the encyclopedia using the active and repressed histone modification signals and CTCF binding sites from tissues. The decorated cCREs were then separated into proximal and distal groups based on their proximity to the annotated TSSs. At another layer, these cCRE subgroups were further annotated as AS and non-AS based on their allelic signature.

(C) Number of cCREs in various tissues. This figure shows the number of different subgroups of decorated cCREs in each tissue type. In each panel, the colors indicate the TSS proximity (proximal vs. distal) and CTCF binding state (CTCF+ vs. CTCF-). Note that the decoration terms are defined in Figure S5A.

(D) cCRE decoration results matrix. We generated an annotation matrix for all the decorated cCREs from each tissue type.

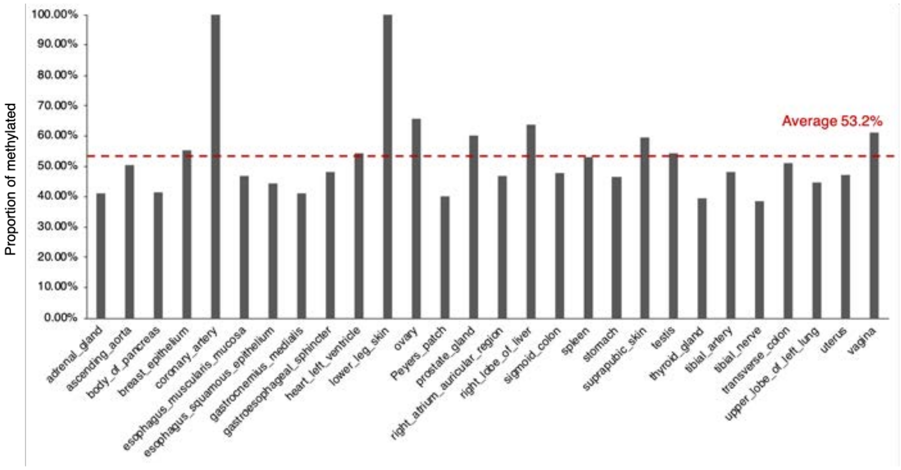

A
H3K9me3 peaks
across donors
Remove:
ENCODE Blacklist
Gene bodies
cCREs
Active histone marks
H3K9me3 unique peaks
B
Number based
Length based
H3K9me3
H3K27me3
0.00%
50.00%
100.00%
0.00%
50.00%
100.00%
Uniquely (1 tissue)
Restricted (2-5 tissues)
Intermediate (6-22 tissues)
Broadly (23-24 tissues)
C
meCpG / CpG
100%
75%
50%
25%
0%
ESPSQE
SPLEEN
STMACH
TESTIS
THYROID
NERVET
CLNTRN
LUNG
H3K27me3 regions
Control (Overall meCpG rate in tissue)
H3K9me3 regions

D



Tissue	#TotalRepressive	#DNAme(>0)	#DNAme(>0.5)	#H3K27me3	#H3K9me3	#DNAme& H3K27me3	#DNAme& H3K9me3	%DNAme(>0)	%DNAme(>0.5)	%H3K27me3	%H3K9me3	%DNAme(>0.5)& H3K27me3	%DNAme(>0.5)& H3K9me3
adrenal_gland	156,348	64,527	61,160	NA	964	NA	71	41.30%	39.10%	NA	0.60%	NA	0.00%
ascending_aorta	156,574	78,621	60,960	20,818	5,603	2,088	650	50.20%	38.90%	13.30%	3.60%	1.30%	0.40%
body_of_pancreas	220,484	91,766	71,692	14,241	3,124	1,187	388	41.60%	32.50%	6.50%	1.40%	0.50%	0.20%
breast_epithelium	163,911	90,816	75,825	30,398	10,129	5,297	1,471	55.40%	46.30%	18.50%	6.20%	3.20%	0.90%
coronary_artery	56,788	56,788	56,776	NA	NA	NA	NA	100.00%	100.00%	NA	NA	NA	NA
esophagus_muscularis_mucosa	221,989	104,029	72,680	47,838	9,567	8,065	1,200	46.90%	32.70%	21.50%	4.30%	3.60%	0.50%
esophagus_squamous_epithelium	181,841	80,483	59,011	44,692	8,826	8,283	1,315	44.30%	32.50%	24.60%	4.90%	4.60%	0.70%
gastrocnemius_medialis	215,358	88,415	67,612	45,425	3,804	6,786	451	41.10%	31.40%	21.10%	1.80%	3.20%	0.20%
gastroesophageal_sphincter	184,935	88,583	70,794	36,429	11,576	7,130	1,535	47.90%	38.30%	19.70%	6.30%	3.90%	0.80%
heart_left_ventricle	152,576	83,059	64,017	28,953	7,958	2,728	985	54.40%	42.00%	19.00%	5.20%	1.80%	0.60%
lower_leg_skin	51,003	51,003	50,999	NA	NA	NA	NA	100.00%	100.00%	NA	NA	NA	NA
ovary	98,107	64,563	62,598	NA	173	NA	26	65.80%	63.80%	NA	0.20%	NA	0.00%
Peyers_patch	251,668	101,002	72,016	16,847	5,284	1,241	856	40.10%	28.60%	6.70%	2.10%	0.50%	0.30%
prostate_gland	117,491	70,663	59,865	11,231	54	1,051	4	60.10%	51.00%	9.60%	0.00%	0.90%	0.00%
right_atrium_auricular_region	150,042	70,187	53,980	15,127	7,626	891	877	46.80%	36.00%	10.10%	5.10%	0.60%	0.60%
right_lobe_of_liver	86,640	55,308	50,919	NA	270	NA	37	63.80%	58.80%	NA	0.30%	NA	0.00%
sigmoid_colon	192,491	91,607	70,626	37,660	3,450	5,471	433	47.60%	36.70%	19.60%	1.80%	2.80%	0.20%
spleen	176,766	93,783	73,775	45,782	28,248	10,451	3,643	53.10%	41.70%	25.90%	16.00%	5.90%	2.10%
stomach	177,428	82,446	59,741	44,645	19,028	6,912	2,387	46.50%	33.70%	25.20%	10.70%	3.90%	1.30%
suprapubic_skin	112,564	66,842	53,222	28,703	3,400	4,196	421	59.40%	47.30%	25.50%	3.00%	3.70%	0.40%
testis	134,229	72,952	58,924	3,470	24	83	2	54.30%	43.90%	2.60%	0.00%	0.10%	0.00%
thoracic_aorta	98,580	NA	NA	5,585	3,218	NA	NA	NA	NA	5.70%	3.30%	NA	NA
thyroid_gland	252,058	99,281	82,411	12,630	411	908	35	39.40%	32.70%	5.00%	0.20%	0.40%	0.00%
tibial_artery	196,188	94,030	73,777	41,688	8,098	7,791	1,082	47.90%	37.60%	21.20%	4.10%	4.00%	0.60%
tibial_nerve	221,297	85,268	62,117	41,283	16,389	5,877	2,284	38.50%	28.10%	18.70%	7.40%	2.70%	1.00%
transverse_colon	149,756	76,276	58,579	35,375	2,902	4,490	387	50.90%	39.10%	23.60%	1.90%	3.00%	0.30%
upper_lobe_of_left_lung	212,200	95,194	65,663	29,742	8,045	2,308	1,122	44.90%	30.90%	14.00%	3.80%	1.10%	0.50%
uterus	167,081	78,725	63,662	14,554	1,881	1,189	252	47.10%	38.10%	8.70%	1.10%	0.70%	0.20%
vagina	115,592	70,527	57,693	5,202	1,351	238	177	61.00%	49.90%	4.50%	1.20%	0.20%	0.20%

E



Data S20. Identifying repressed elements, related to Figure 3 and STAR Methods "Decoration Process" Section

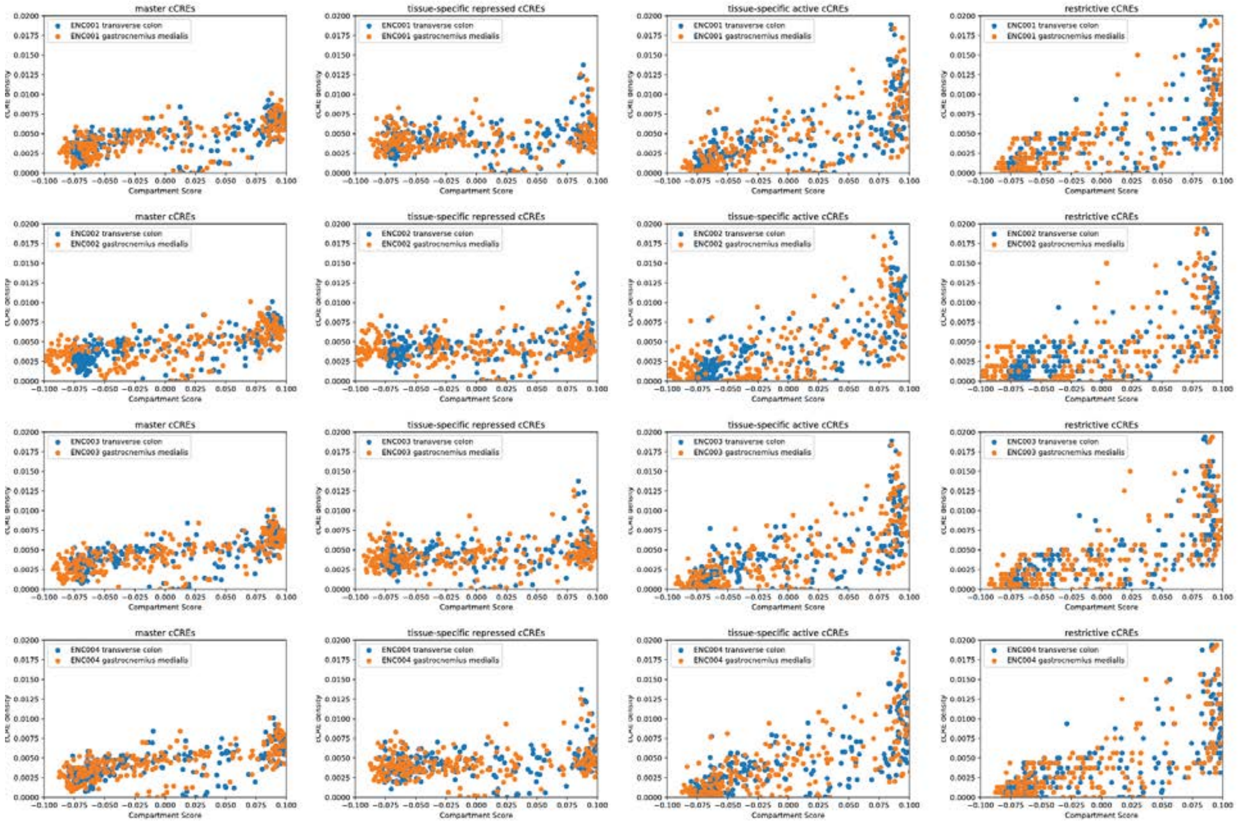
(A) For genomic regions outside of cCREs and annotated genes, elements longer than 200 bp that are uniquely marked by either H3K9me3 or H3K27me3 were defined as fully repressed. A total of 45,207 (covering 12,655,795 bp) and 24,006 (covering 7,474,178 bp) non-overlapping elements were identified based on H3K9me3 and H3K27me3, respectively.

(B) The majority of these elements were repressed in a tissue-specific manner.

(C) For tissues with available datasets, DNA methylation within these elements was evaluated, and H3K9me3-marked elements showed a significantly (t-test, p-value < 0.05) higher CpG methylation (meCpG) rate than elements marked uniquely by H3K27me3.

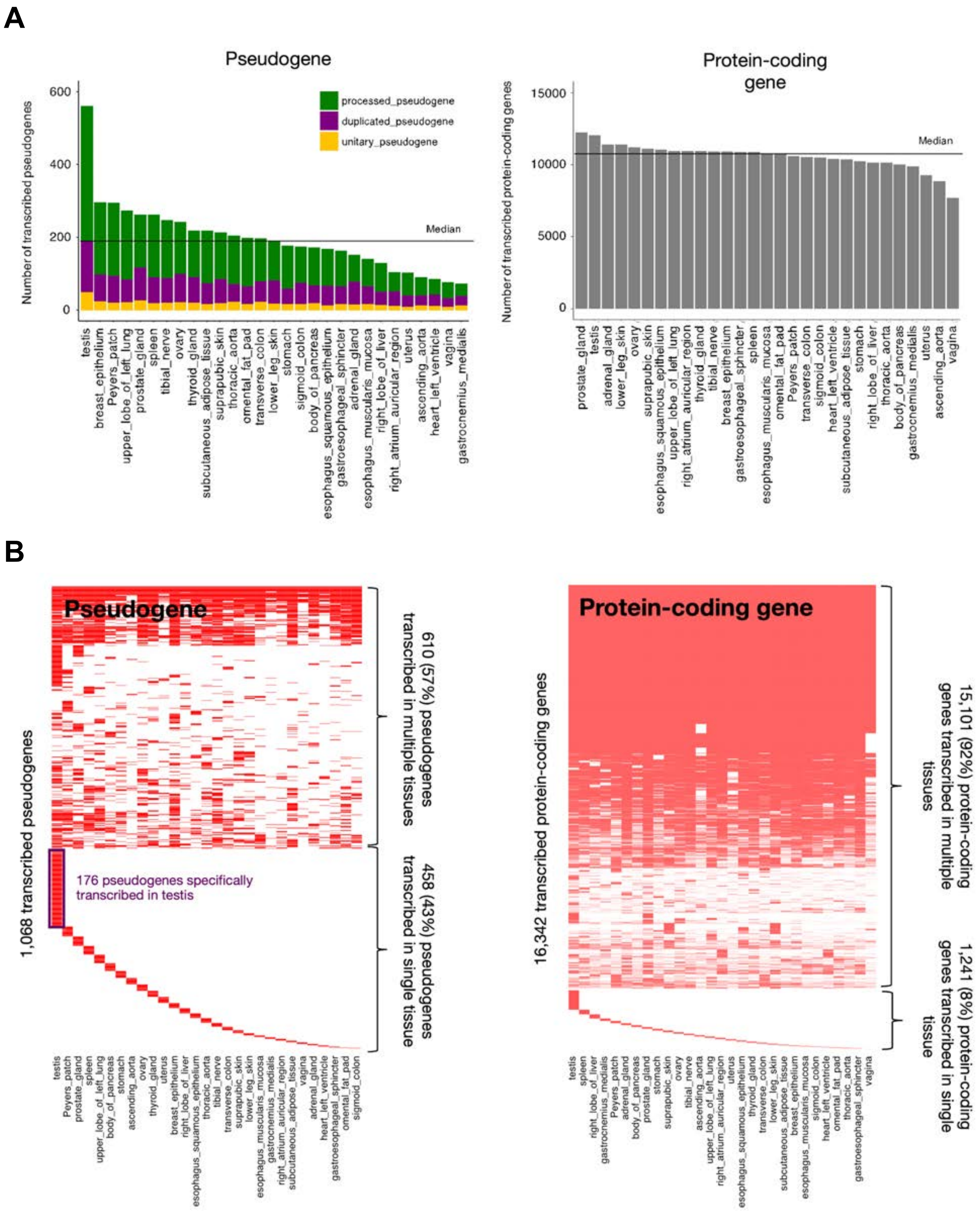
(D) Number and proportion of repressed cCREs that overlap with the repressive histone marks and/or DNA methylation in each tissue type.

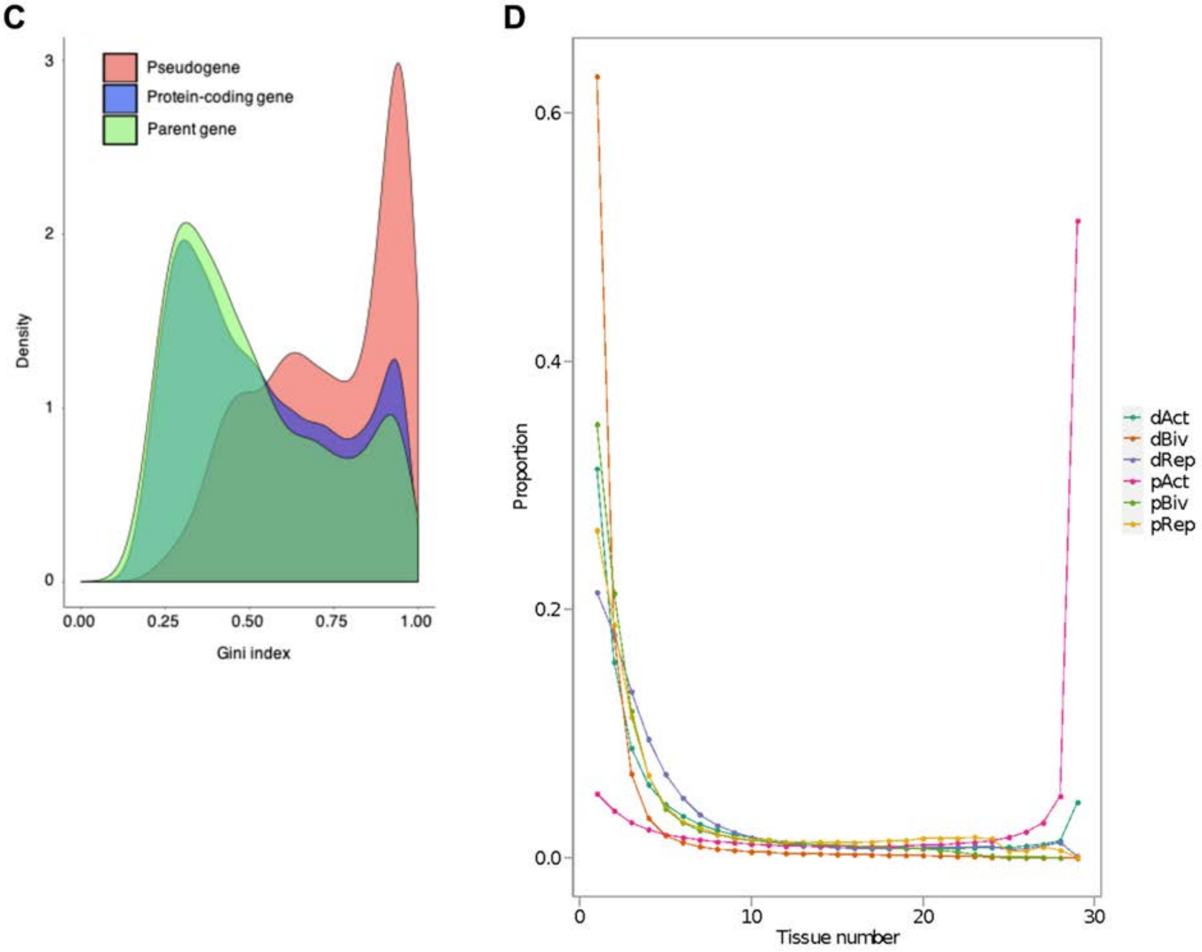
(E) The proportion of repressed cCREs that overlap with DNA methylation in each tissue type.



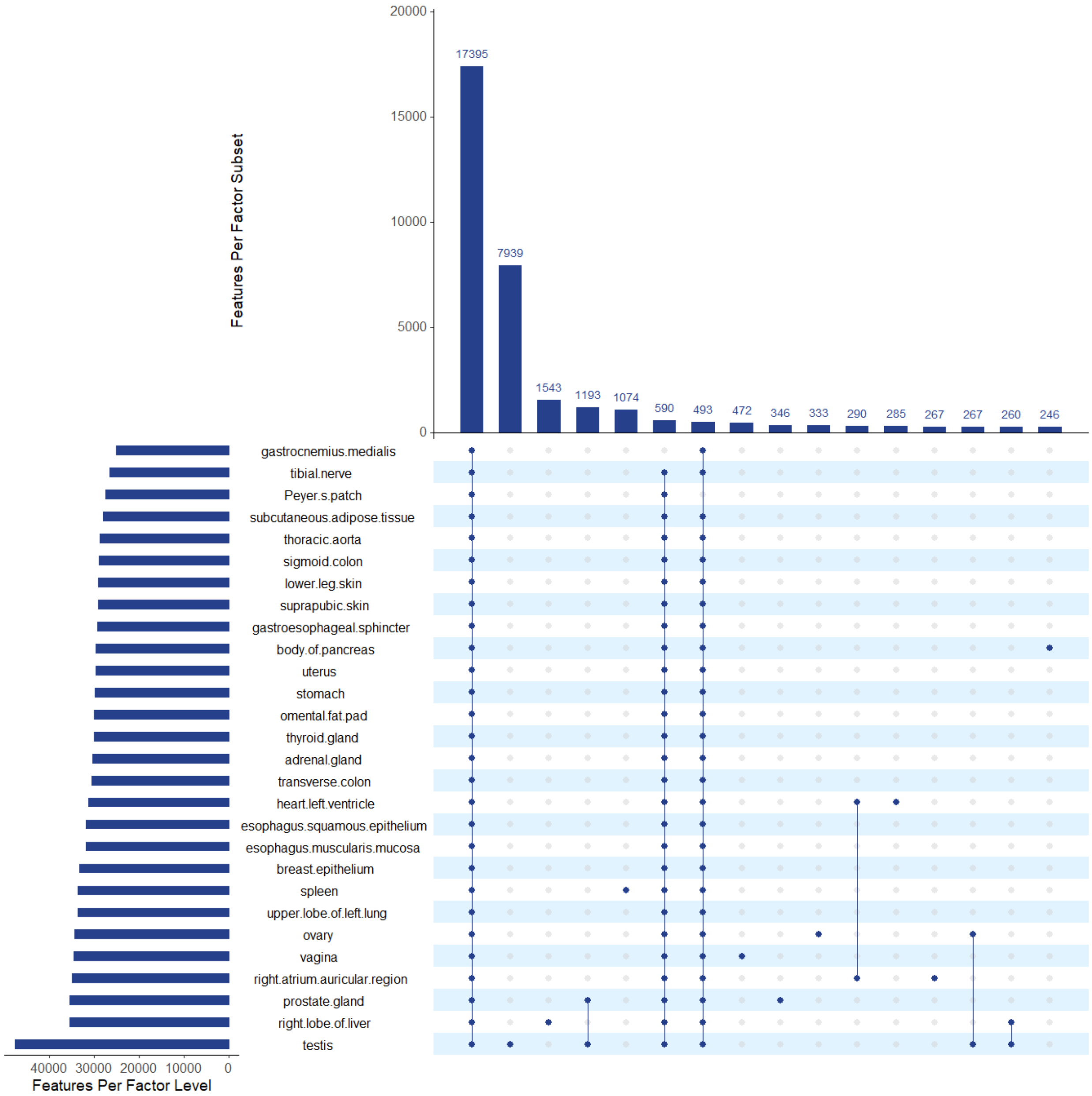
Data S21. cCRE enrichment with respect to A/B compartments "Decoration Process" Section

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, including both tissue-specific active and repressed cCREs. As the tissue specificity increases, the cCRE enrichment in the active A compartment increases compared with the inactive B compartment.

A
Pseudogene
Number of transcribed pseudogenes
600
400
200
0
processed_pseudogene
duplicated_pseudogene
unitary_pseudogene
Median
testis
breast_epithelium
Peyers_patch
upper_lobe_of_left_lung
prostate_gland
spleen
tibial_nerve
ovary
thyroid_gland
subcutaneous_adipose_tissue
suprapubic_skin
thoracic_aorta
omental_fat_pad
transverse_colon
lower_leg_skin
stomach
sigmoid_colon
body_of_pancreas
esophagus_squamous_epithelium
gastroesophageal_sphincter
adrenal_gland
esophagus_muscularis_mucosa
right_lobe_of_liver
right_atrium_auricular_region
uterus
ascending_aorta
heart_left_ventricle
vagina
gastrocnemius_medialis
Protein-coding gene
Number of transcribed protein-coding genes
15000
10000
5000
0
Median
prostate_gland
testis
adrenal_gland
lower_leg_skin
ovary
suprapubic_skin
esophagus_squamous_epithelium
upper_lobe_of_left_lung
right_atrium_auricular_region
thyroid_gland
tibial_nerve
breast_epithelium
gastroesophageal_sphincter
spleen
esophagus_muscularis_mucosa
omental_fat_pad
Peyers_patch
transverse_colon
sigmoid_colon
heart_left_ventricle
subcutaneous_adipose_tissue
stomach
right_lobe_of_liver
thoracic_aorta
body_of_pancreas
gastrocnemius_medialis
uterus
ascending_aorta
vagina
B
Pseudogene
1,068 transcribed pseudogenes
610 (57%) pseudogenes transcribed in multiple tissues
458 (43%) pseudogenes transcribed in single tissue
176 pseudogenes specifically transcribed in testis
testis
Peyers_patch
prostate_gland
spleen
upper_lobe_of_left_lung
body_of_pancreas
stomach
ascending_aorta
ovary
thyroid_gland
uterus
breast_epithelium
right_lobe_of_liver
esophagus_squamous_epithelium
thoracic_aorta
tibial_nerve
transverse_colon
suprapubic_skin
lower_leg_skin
esophagus_muscularis_mucosa
gastrocnemius_medialis
right_atrium_auricular_region
subcutaneous_adipose_tissue
vagina
adrenal_gland
heart_left_ventricle
gastroesophageal_sphincter
omental_fat_pad
sigmoid_colon
Protein-coding gene
16,342 transcribed protein-coding genes
15,101 (92%) protein-coding genes transcribed in multiple tissues
1,241 (8%) protein-coding genes transcribed in single tissue
testis
spleen
right_lobe_of_liver
gastrocnemius_medialis
Peyers_patch
adrenal_gland
body_of_pancreas
prostate_gland
stomach
suprapubic_skin
ovary
tibial_nerve
ascending_aorta
upper_lobe_of_left_lung
right_atrium_auricular_region
uterus
esophagus_squamous_epithelium
thyroid_gland
transverse_colon
lower_leg_skin
subcutaneous_adipose_tissue
breast_epithelium
esophagus_muscularis_mucosa
sigmoid_colon
heart_left_ventricle
omental_fat_pad
thoracic_aorta
gastroesophageal_sphincter
vagina

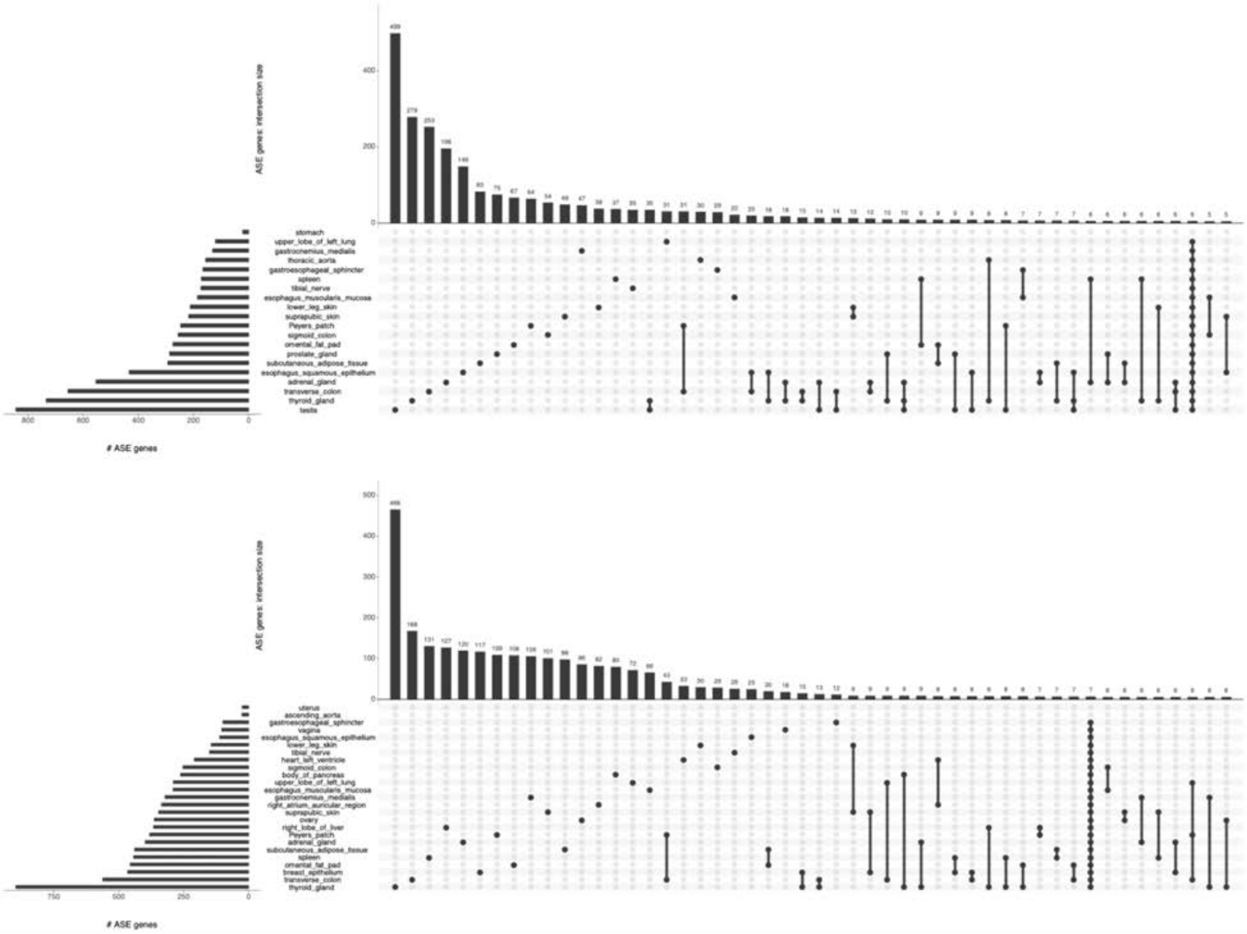

C
Pseudogene
Protein-coding gene
Parent gene
Density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Gini index
D
Proportion
0.6
0.4
0.2
0.0
0
10
20
30
Tissue number
dAct
dBiv
dRep
pAct
pBiv
pRep

E



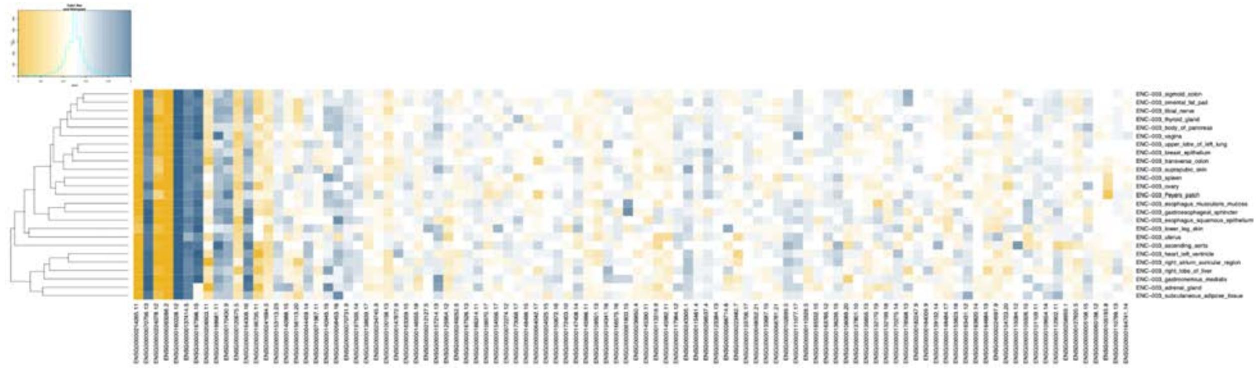
Features Per Factor Subset
20000
15000
10000
5000
0
17395
7939
1543
1193
1074
590
493
472
346
333
290
285
267
267
260
246
gastrocnemius.medialis
tibial.nerve
Peyer.s.patch
subcutaneous.adipose.tissue
thoracic.aorta
sigmoid.colon
lower.leg.skin
suprapubic.skin
gastroesophageal.sphincter
body.of.pancreas
uterus
stomach
omental.fat.pad
thyroid.gland
adrenal.gland
transverse.colon
heart.left.ventricle
esophagus.squamous.epithelium
esophagus.muscularis.mucosa
breast.epithelium
spleen
upper.lobe.of.left.lung
ovary
vagina
right.atrium.auricular.region
prostate.gland
right.lobe.of.liver
testis
40000
30000
20000
10000
0
Features Per Factor Level

F



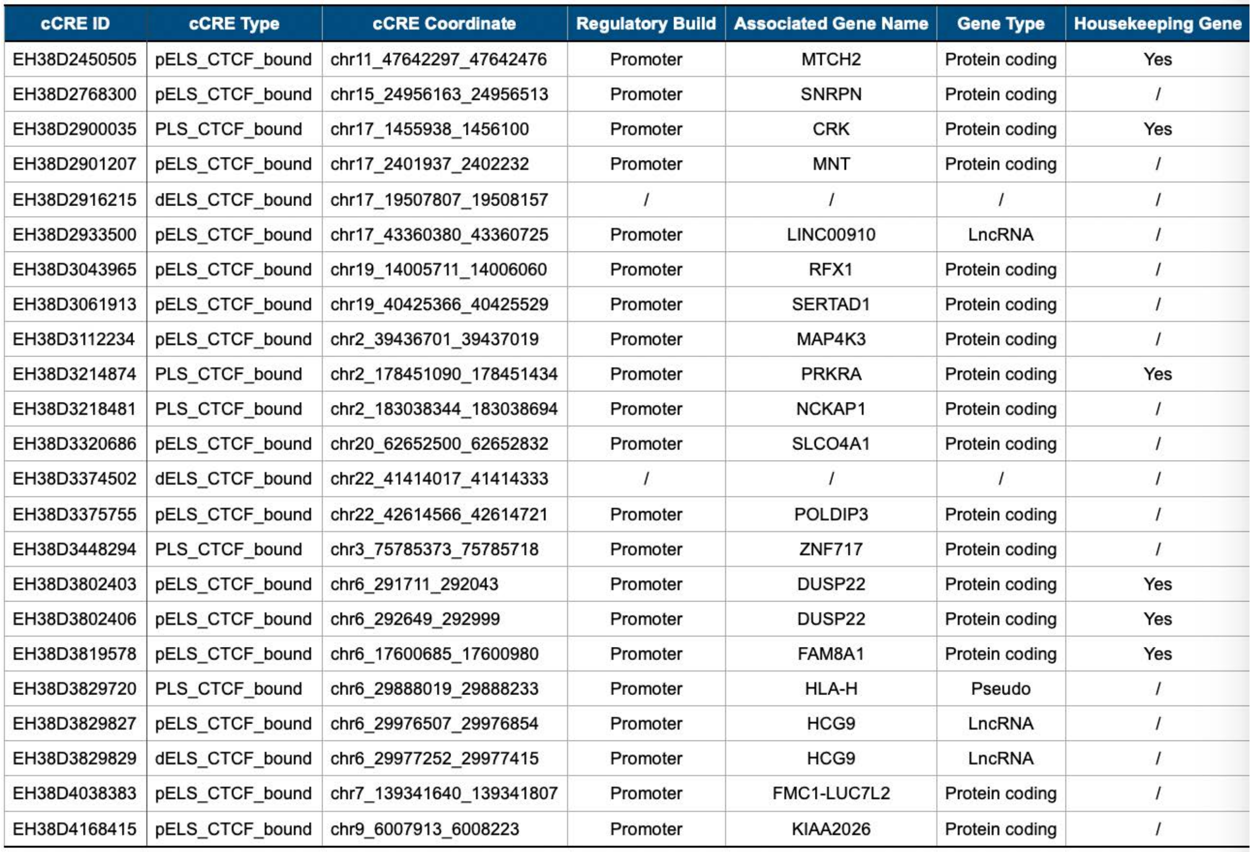
ASE genes: intersection size
0
200
400
499
279
253
196
149
stomach
upper_lobe_of_left_lung
gastrocnemius_medialis
thoracic_aorta
gastroesophageal_sphincter
spleen
tibial_nerve
esophagus_muscularis_mucosa
lower_leg_skin
suprapubic_skin
Peyers_patch
sigmoid_colon
omental_fat_pad
prostate_gland
subcutaneous_adipose_tissue
esophagus_squamous_epithelium
adrenal_gland
transverse_colon
thyroid_gland
testis
800
600
400
200
0
# ASE genes
ASE genes: intersection size
0
100
200
300
400
500
466
168
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
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
# ASE genes

G



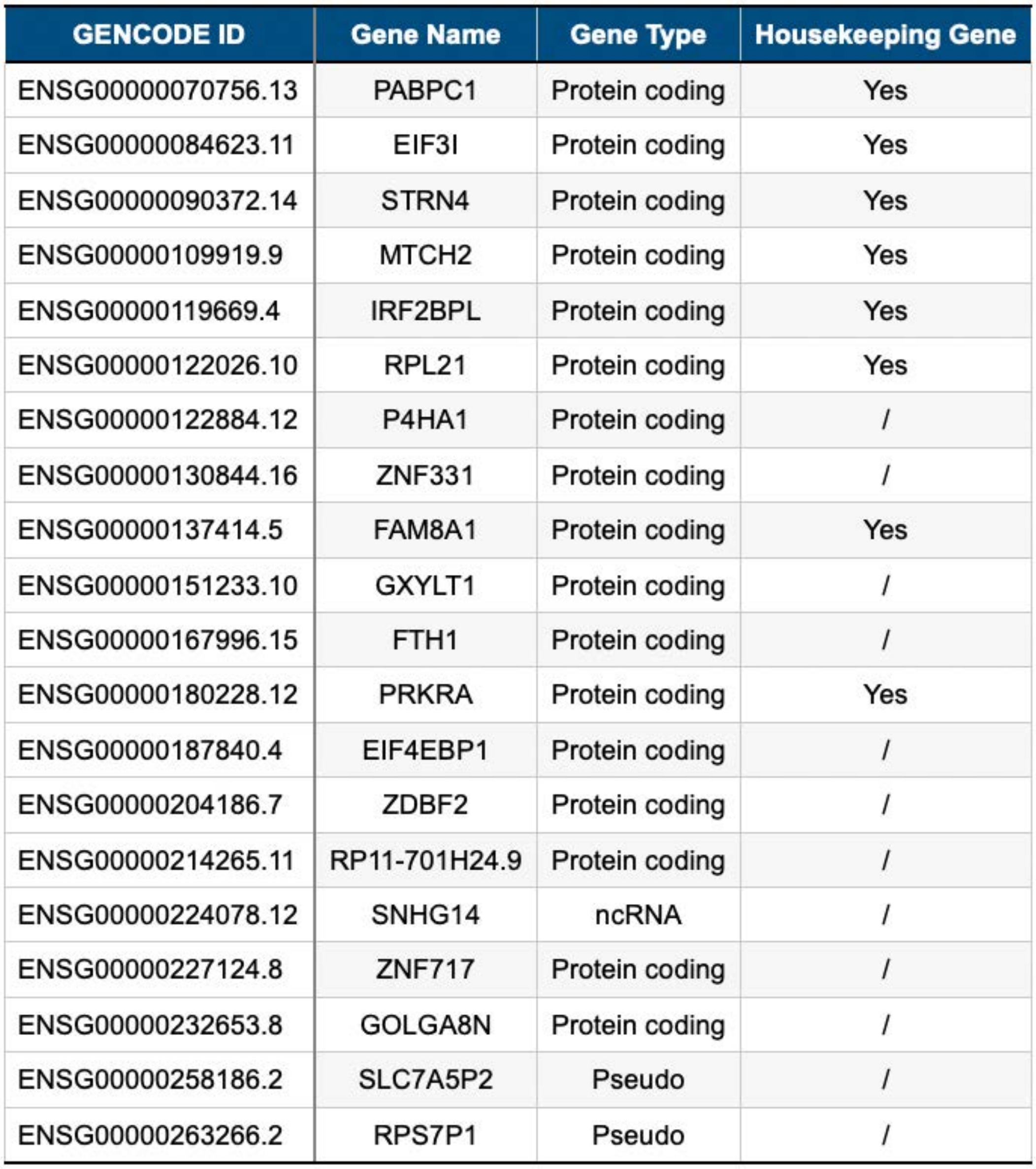
ENC-003_sigmoid_colon
ENC-003_omental_fat_pad
ENC-003_tibial_nerve
ENC-003_thyroid_gland
ENC-003_body_of_pancreas
ENC-003_vagina
ENC-003_upper_lobe_of_left_lung
ENC-003_breast_epithelium
ENC-003_transverse_colon
ENC-003_suprapubic_skin
ENC-003_spleen
ENC-003_ovary
ENC-003_Peyers_patch
ENC-003_esophagus_muscularis_mucosa
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

H

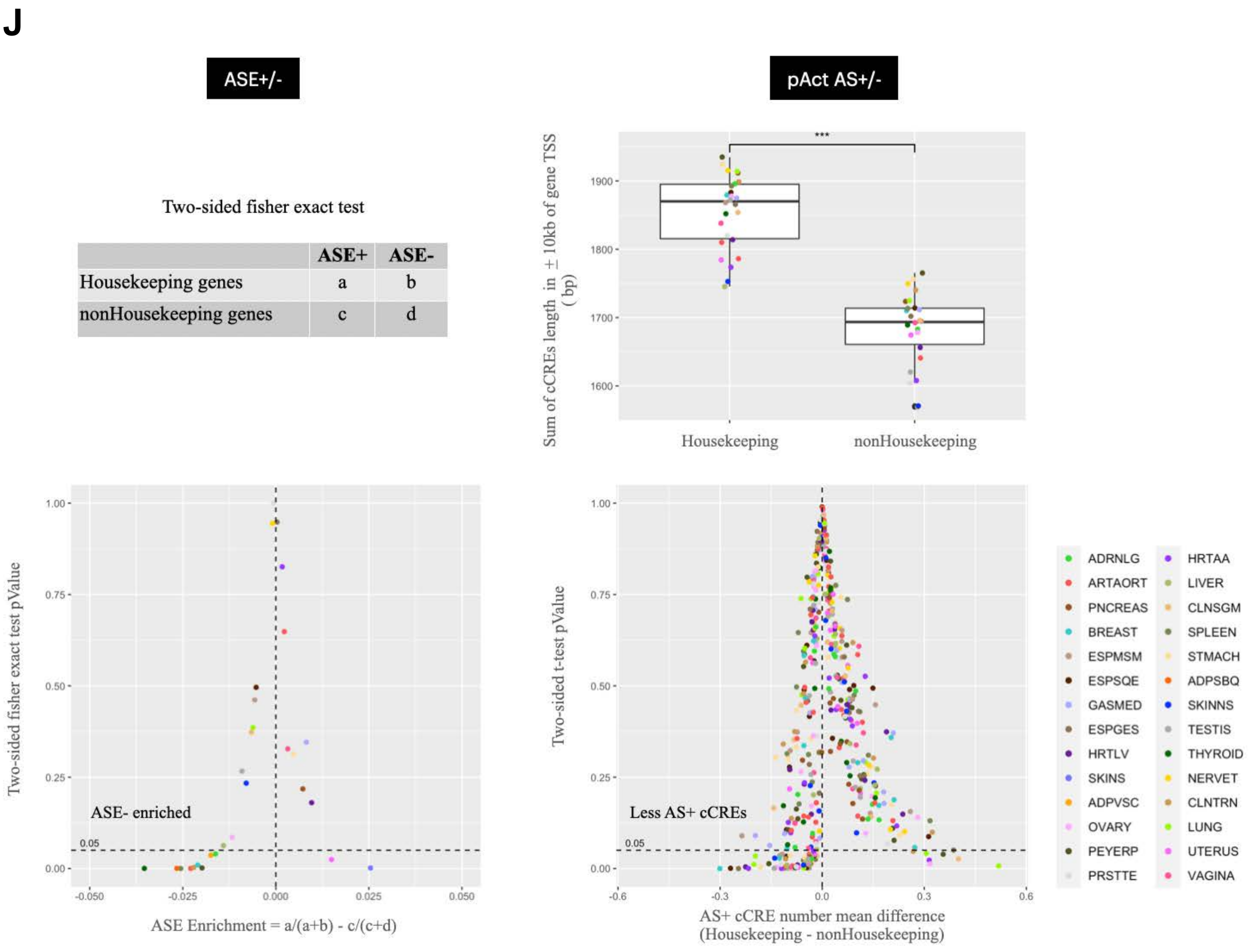


cCRE ID	cCRE Type	cCRE Coordinate	Regulatory Build	Associated Gene Name	Gene Type	Housekeeping Gene
EH38D2450505	pELS_CTCF_bound	chr11_47642297_47642476	Promoter	MTCH2	Protein coding	Yes
EH38D2768300	pELS_CTCF_bound	chr15_24956163_24956513	Promoter	SNRPN	Protein coding	/
EH38D2900035	PLS_CTCF_bound	chr17_1455938_1456100	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	/

I



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	/



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

Data S22. Tissue specificity of AS events, related to Figure 3, Figure S5, and STAR Methods "Tissue Specificity" Section

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

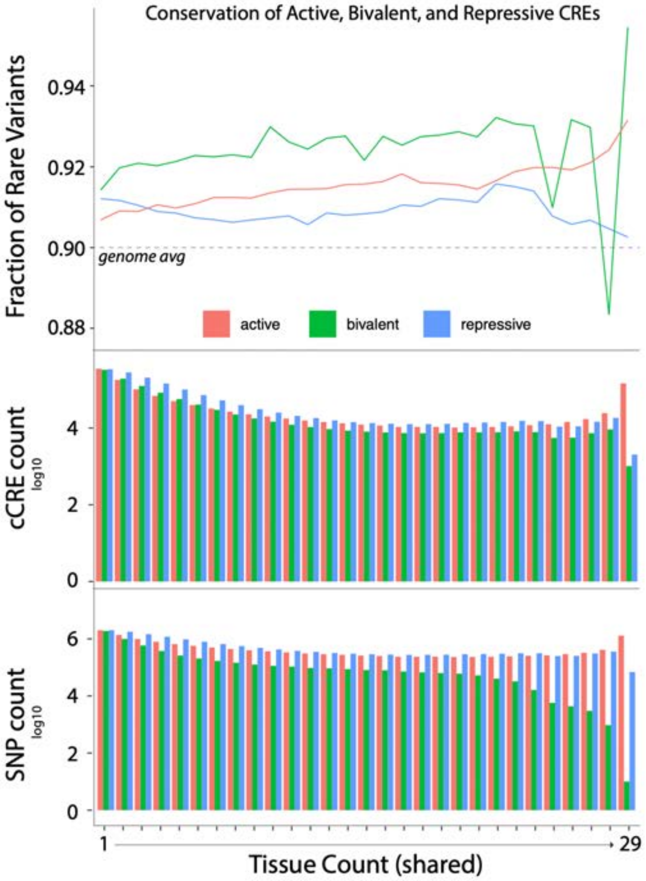
(B) Tissue specificity of transcribed genes. The heatmaps show the activity of pseudogenes (left) and protein-coding genes (right) across tissue types. In each tissue, the pseudogenes/protein-coding genes are classified as actively transcribed (shown in red) or not based on their expression level.

(C) Gini index of gene expression level across tissues. We applied 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 indexes than the protein-coding genes, suggesting stronger tissue specificity of pseudogenes. The Gini index distribution of the pseudogenes is quite different from that of the parent genes, confirming that the multi-mapping bias from quantification of the pseudogene expression level has been minimized.

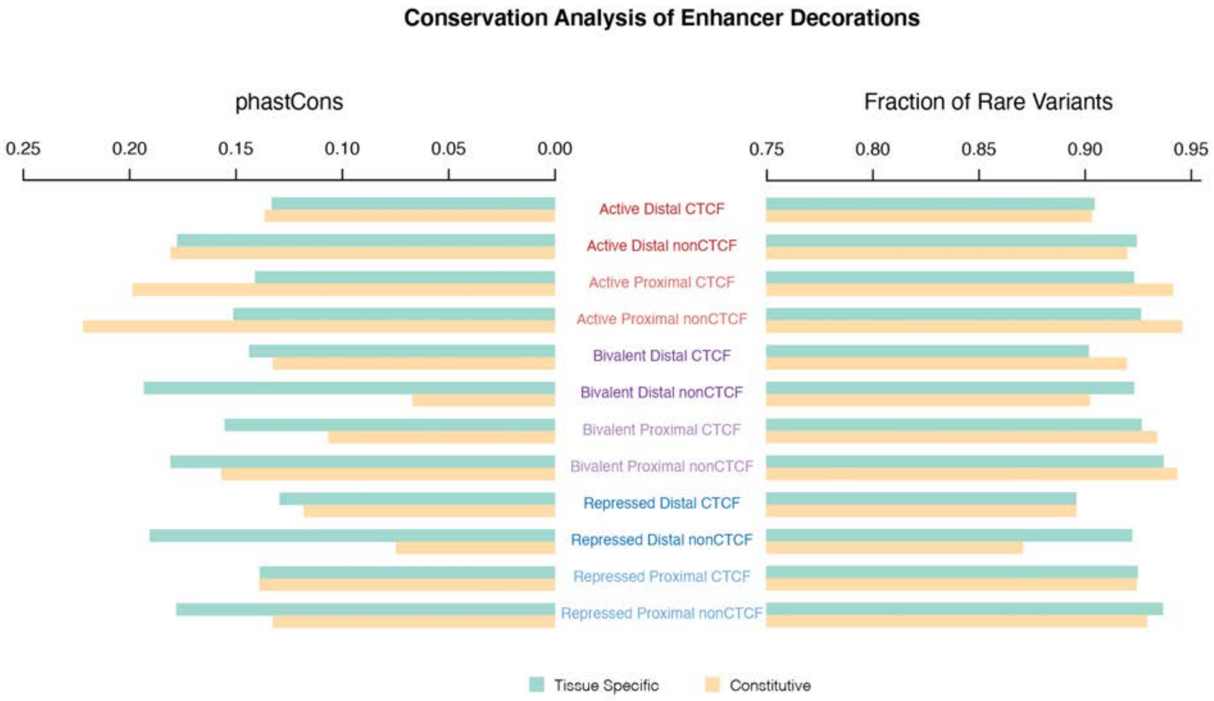
(D) Tissue specificity of different subgroups of cCREs. For each cCRE subgroup, we show the proportion of the cCREs that are defined as "active" across the different numbers of tissue types ranging from one (i.e., high tissue specificity) to all tissue types (i.e., low tissue specificity). Note that the decoration terms are defined in Figure S5A.

(E) Tissue specificity of RAMPAGE data at TSSs of protein-coding genes. This figure shows an 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.
(F) ASE genes across different tissues of individual 2 (top) and 3 (bottom). Counts of genes (bars) with detected ASE in the combinations of tissues (sets of dots) with the largest number of common AS genes. Bars on the left correspond to the number of AS genes in each tissue.
(G) Hap1 allele ratios (number of hap1 reads over the total number of reads) for expression of genes that are accessible across all tissues and AS in at least one tissue of individual 3. This figure parallels the allelic ratios for H3K27ac in Figure 3F and shows the same trend for expression as for histone modification. Note in Figure 3F, we did not find any significant bias in GO enrichment or in the chromosomal distribution of genes targeted by cCREs that flip imbalance direction across tissues. The allelic imbalance is measured by the fraction of unique reads mapped to each haplotype.
(H) 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 six promoters of housekeeping genes. Promoters and associated genes are based on Ensembl, and housekeeping genes are based on the HRT Atlas [19].
(I) 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, eight genes are annotated as housekeeping genes in the HRT Atlas.
(J) Allelic specificity of housekeeping genes. Left: for each tissue, expressed protein-coding genes were split into housekeeping genes and non-housekeeping genes. Based on the two-sided Fisher's exact test, housekeeping genes are generally expressed in less of an AS fashion than non-housekeeping genes. Right: for each tissue, we examined the allele specificity of pAct cCREs flanking the TSS (defined by the gene starting site) of housekeeping genes. To eliminate the bias caused by significantly different cCRE lengths flanking the genes, we split genes into 20 bins based on the total length of the flanking cCREs. Within each bin, the number of pAct AS cCREs was compared between the housekeeping and non-housekeeping genes. pAct cCREs flanking the housekeeping genes display relatively less allele specificity than the ones flanking non-housekeeping genes.

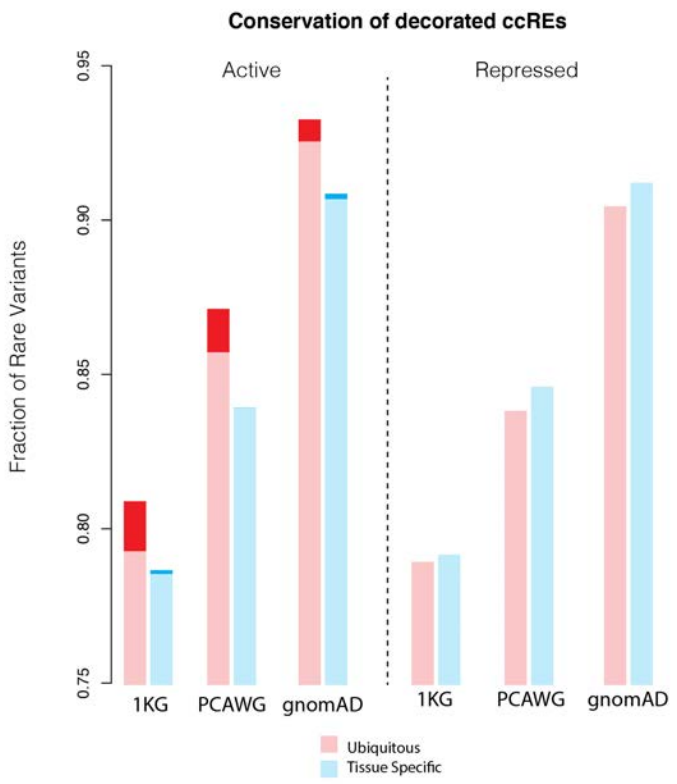
A

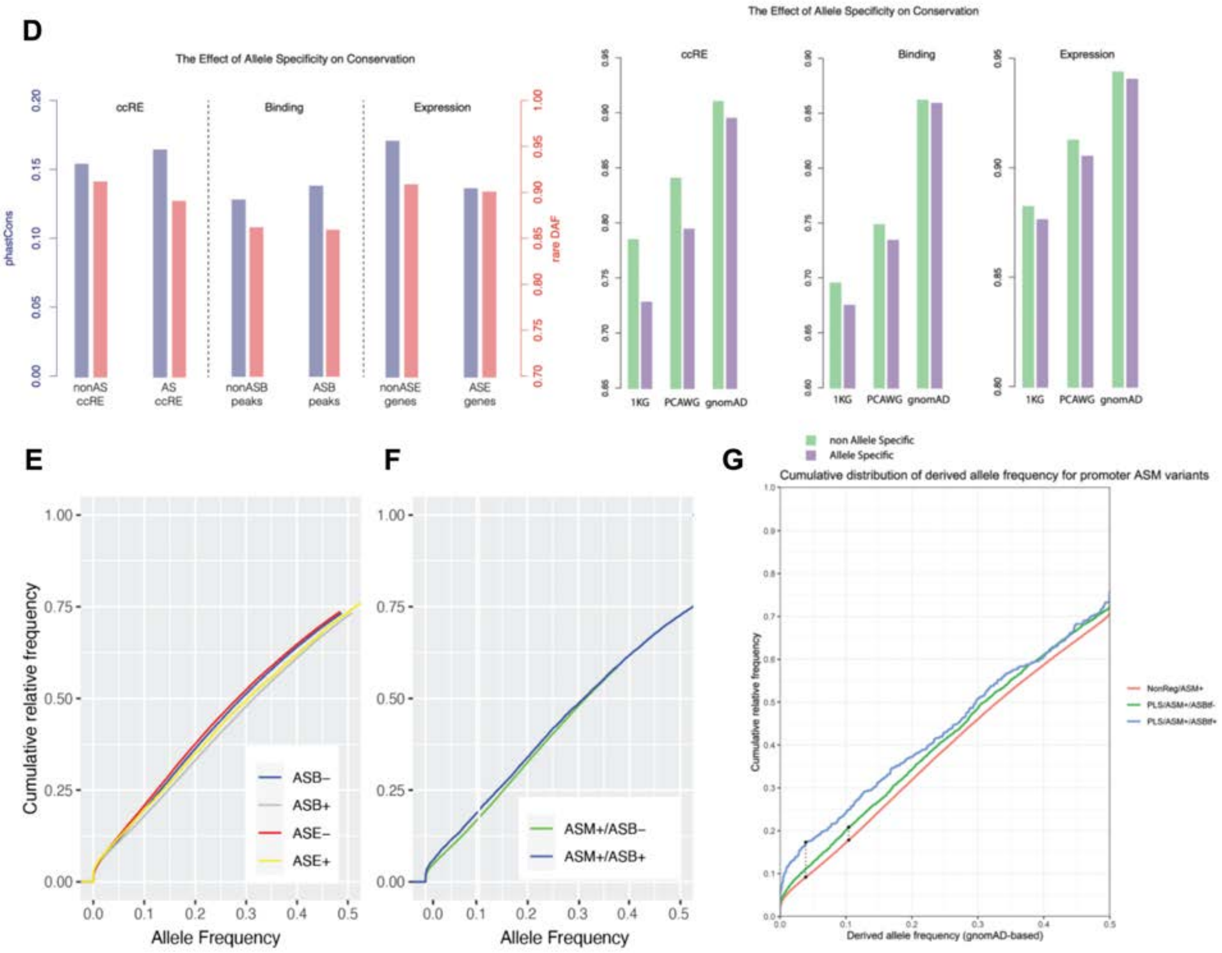

Conservation of Active, Bivalent, and Repressive CREs
Fraction of Rare Variants
0.94
0.92
0.90
0.88
genome avg
active
bivalent
repressive
cCRE count log10
4
2
0
SNP count log10
6
4
2
0
1
29
Tissue Count (shared)

B


Conservation Analysis of Enhancer Decorations
phastCons
0.25
0.20
0.15
0.10
0.05
0.00
Fraction of Rare Variants
0.75
0.80
0.85
0.90
0.95
Active Distal CTCF
Active Distal nonCTCF
Active Proximal CTCF
Active Proximal nonCTCF
Bivalent Distal CTCF
Bivalent Distal nonCTCF
Bivalent Proximal CTCF
Bivalent Proximal nonCTCF
Repressed Distal CTCF
Repressed Distal nonCTCF
Repressed Proximal CTCF
Repressed Proximal nonCTCF
Tissue Specific
Constitutive

C


Conservation of decorated ccREs
Active
Repressed
Fraction of Rare Variants
0.95
0.90
0.85
0.80
0.75
1KG
PCAWG
gnomAD
1KG
PCAWG
gnomAD
Ubiquitous
Tissue Specific



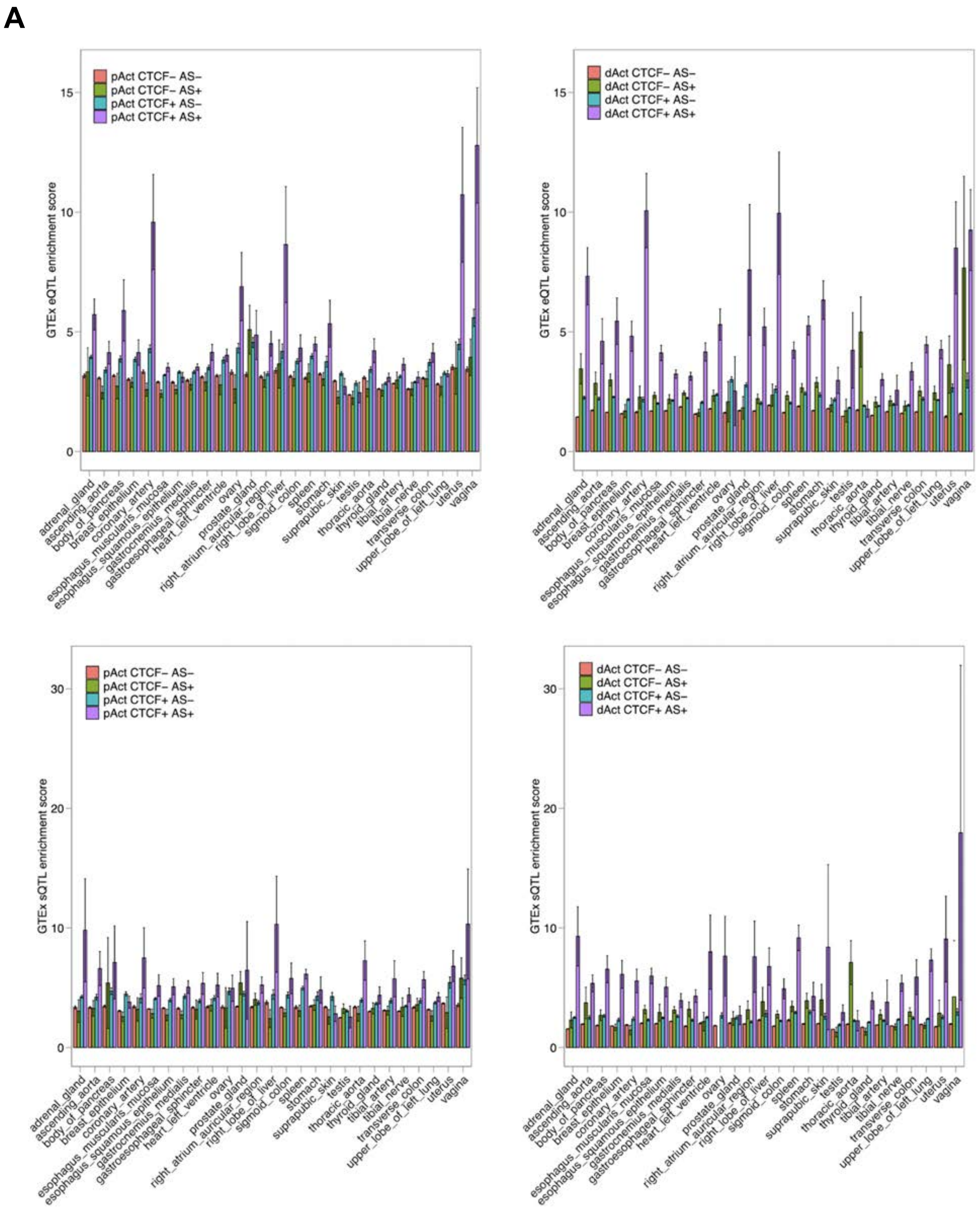
Data S23. Conservation of cCREs, related to Figure 3 and STAR Methods “Tissue Specificity” Section

(A) Rare derived allele frequency (DAF) for active, bivalent, and repressed cCREs in increasing tissue count. Fraction of rare variants was calculated as # rare variants / (# rare variants + # common variants). Total cCRE and SNP (taking into account all SNPs, common and rare) counts are shown for tissue count as well. We additionally performed a more in-depth analysis on the correlation as shown in Figure 3. We show the correlation for tissue specificity and conservation for active, bivalent, and repressed methylation groups of cCREs. The correlations are -0.90, 0.05, and 0.84 and the p-values are 1.9e-11, 0.60, and 9.76e-9, respectively.

(B) Conservation of enhancer decorations. The conservation was calculated in terms of the phastCons score and fraction of rare variants, based on the DAF in the gnomAD database. The annotations are from Figure S5A.

(C) Conservation of active and repressed cCREs for tissue-specific and ubiquitous categories. Dark red shows an increase in conservation for more stringently defined cCREs (selected via the top 1% of Matched Filter signals; see STAR Methods “Tissue Specificity” Section). The databases for this calculation include 1,000 Genomes (1KG), the Pan-Cancer Analysis of Whole Genomes (PCAWG), and gnomAD.

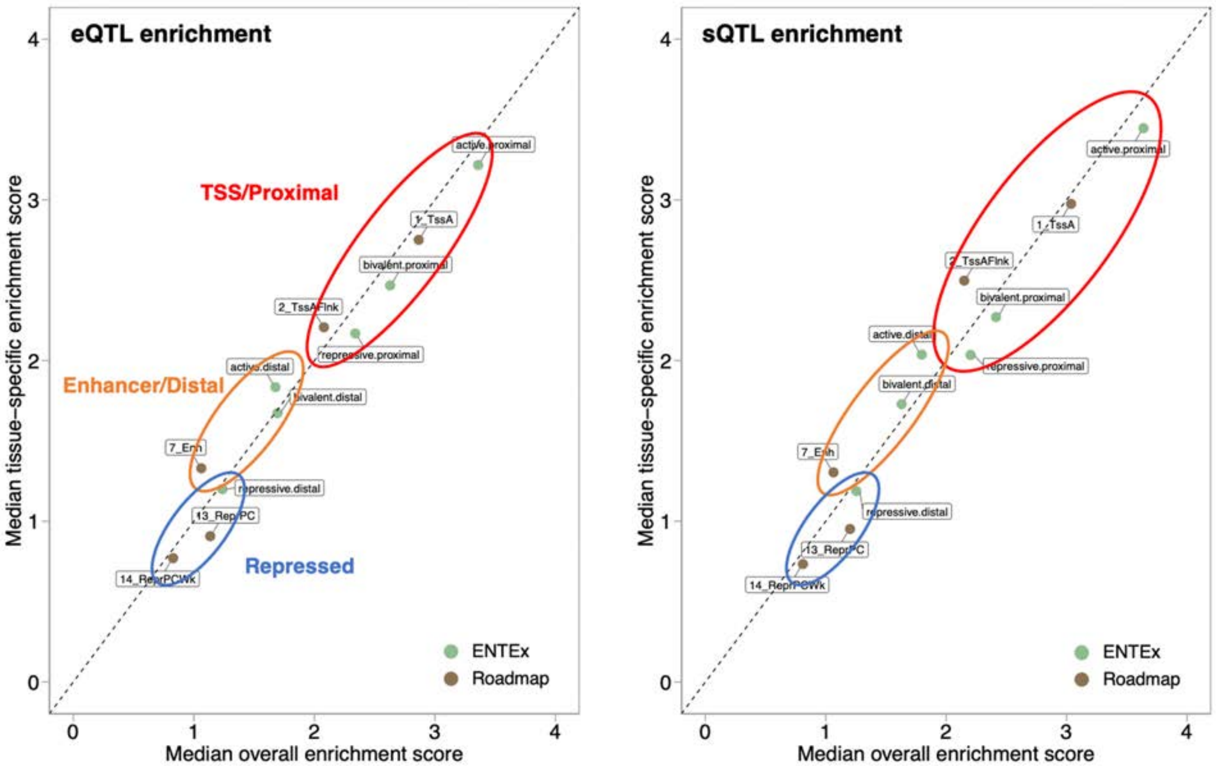
(D) - (G) Conservation of regions exhibiting AS activity. (A) The conservation of various AS annotations was calculated using phastCons and the fraction of rare variants based on different population variants. Specifically, we considered AS/non-AS cCREs, ASB/non-ASB peaks from H3K27ac, and AS/non-AS genes. An alternate way to observe the same phenomenon is to determine the cumulative relative frequency of variants, shown in (B). Here, we see that non-AS events demonstrate stronger purifying selection than AS events, shown by the higher cumulative frequency curve. However, ASM/ASB variants demonstrate more consistency and higher purifying selection as compared to ASM/non-ASB events, shown in (C)**.** Finally, in (D) we show that the effect in (C) is amplified when exclusively considering promoter regions.


A
pAct CTCF− AS−
pAct CTCF− AS+
pAct CTCF+ AS−
pAct CTCF+ AS+
GTEx eQTL enrichment score
0
5
10
15
dAct CTCF− AS−
dAct CTCF− AS+
dAct CTCF+ AS−
dAct CTCF+ AS+
GTEx eQTL enrichment score
0
5
10
15
pAct CTCF− AS−
pAct CTCF− AS+
pAct CTCF+ AS−
pAct CTCF+ AS+
GTEx sQTL enrichment score
0
10
20
30
dAct CTCF− AS−
dAct CTCF− AS+
dAct CTCF+ AS−
dAct CTCF+ AS+
GTEx sQTL enrichment score
0
10
20
30
adrenal_gland
ascending_aorta
body_of_pancreas
breast_epithelium
coronary_artery
esophagus_muscularis_mucosa
esophagus_squamous_epithelium
gastrocnemius_medialis
gastroesophageal_sphincter
heart_left_ventricle
ovary
prostate_gland
right_atrium_auricular_region
right_lobe_of_liver
sigmoid_colon
spleen
stomach
suprapubic_skin
testis
thoracic_aorta
thyroid_gland
tibial_artery
tibial_nerve
transverse_colon
upper_lobe_of_left_lung
uterus
vagina

B

Roadmap_ID	Roadmap_name	EN-Tex_name	GTEx_name
E065	Aorta	ascending_aorta	Artery_Aorta
E098	Pancreas	body_of_pancreas	Pancreas
E119	HMEC Mammary Epithelial Primary C	breast_epithelium	Breast_Mammary_Tissue
E079	Esophagus	esopahgus_muscularis_muco	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_lung	Lung

C

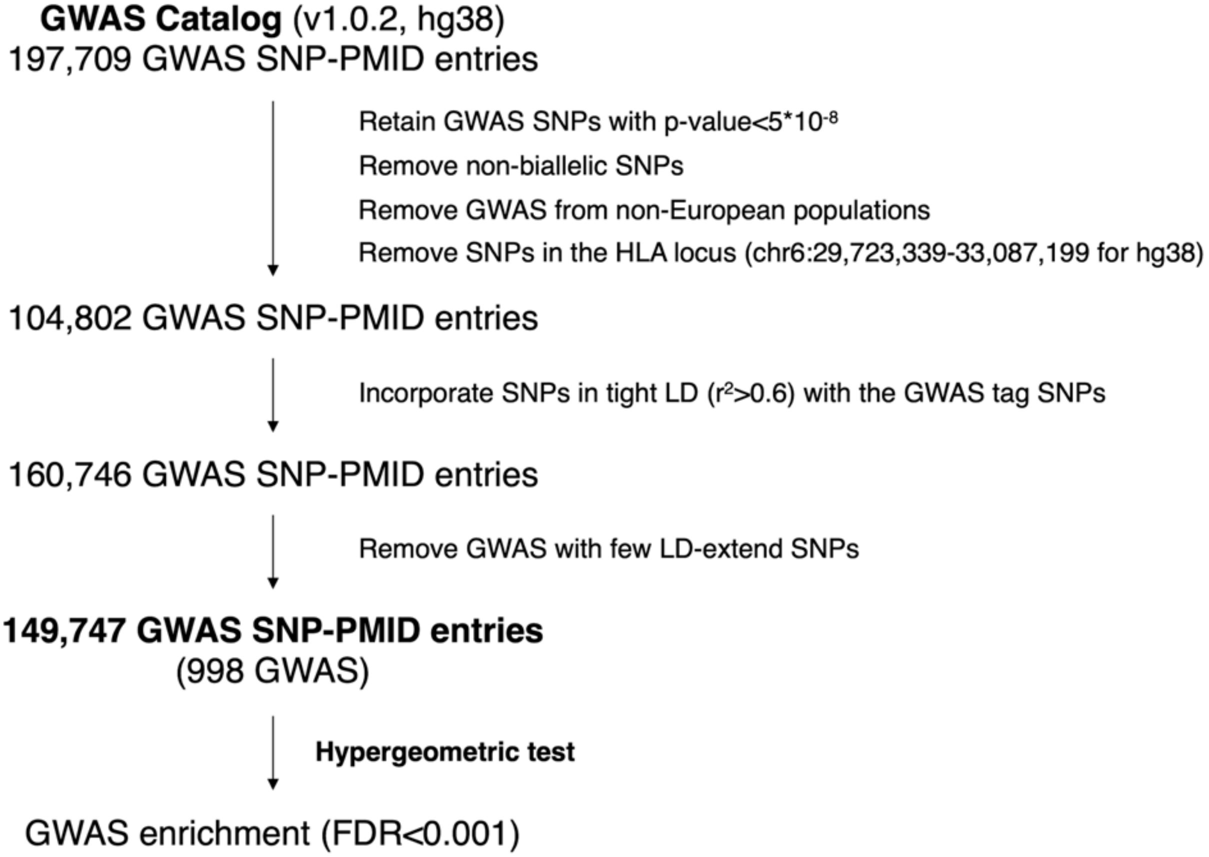


Data S24. eQTLs in cCREs, related to Figure 4 and STAR Methods "Decoration Enrichments" Section

(A) eQTL and sQTL enrichment in cCREs. We computed 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 were 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 was repeated 30 times to calculate the standard deviation, and the values are indicated by the whiskers. 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 AS patterns. Note that the decoration terms are defined in Figure 5A.
(B) Roadmap annotations. We selected 14 tissue types that are matched across the EN-TEx, GTEx, and Roadmap projects to compare the QTL enrichment in the EN-TEx cCREs and Roadmap regulatory annotations. We used the 15-state Roadmap annotations in the analysis.
(C) QTL enrichment in cCREs: EN-TEx vs. Roadmap. We compared 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 EN-TEx and Roadmap as shown in panel B.

A



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

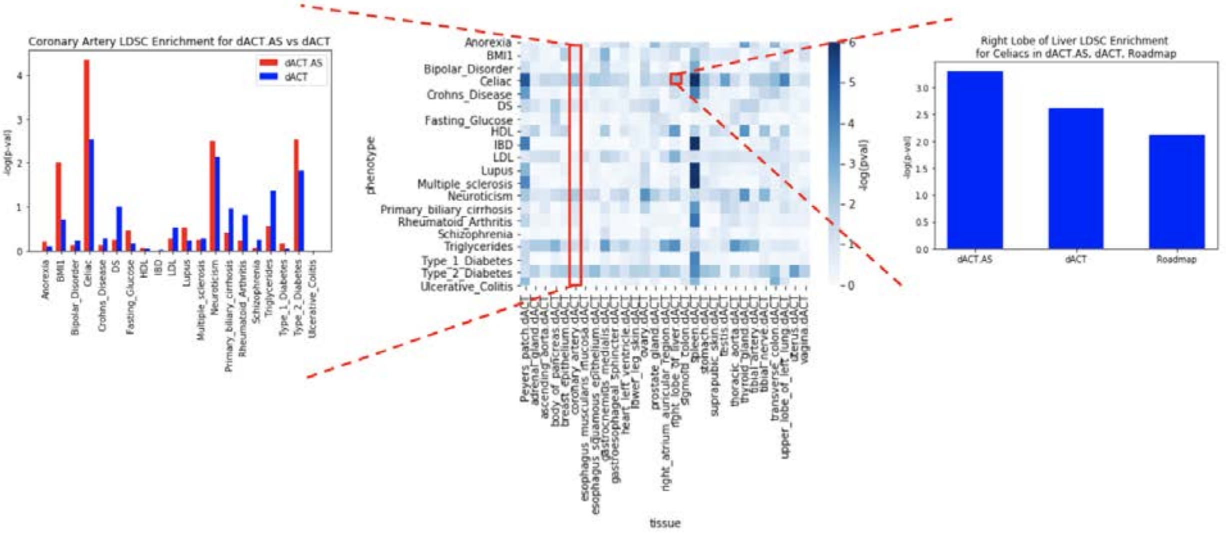
Remove GWAS with few LD-extend SNPs

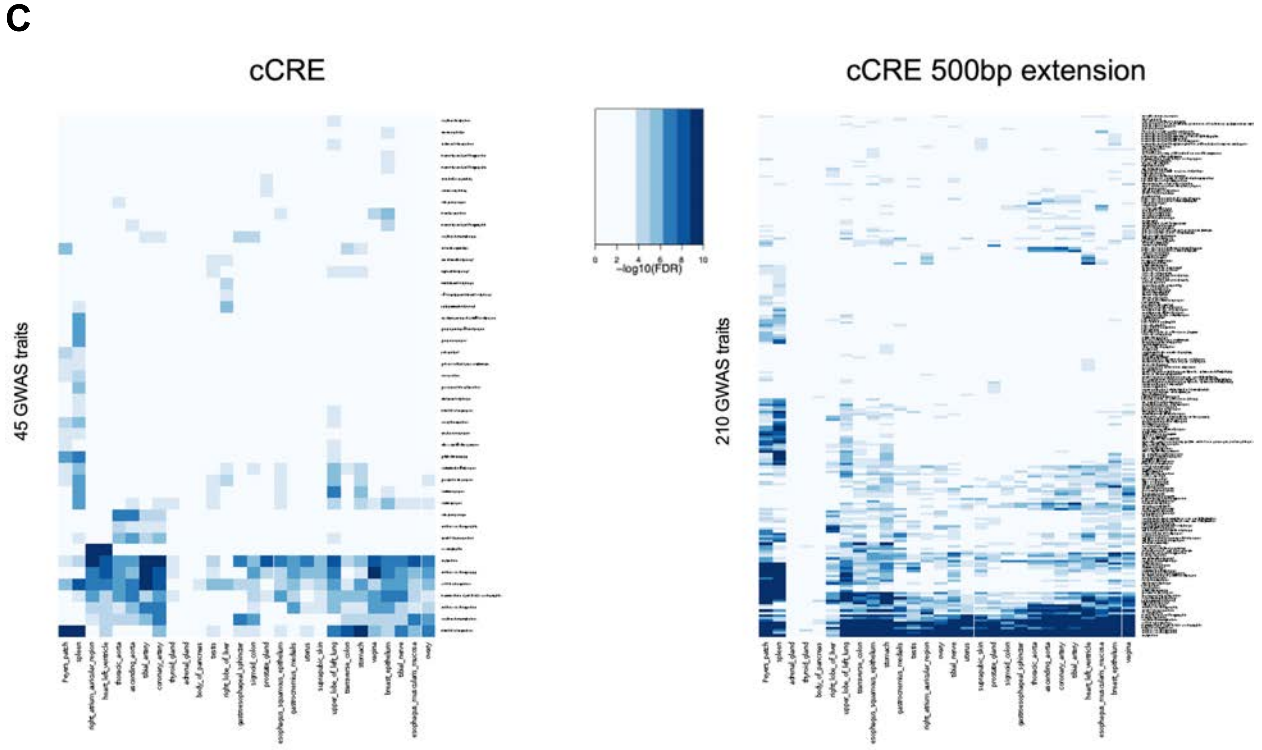
149,747 GWAS SNP-PMID entries
(998 GWAS)

Hypergeometric test

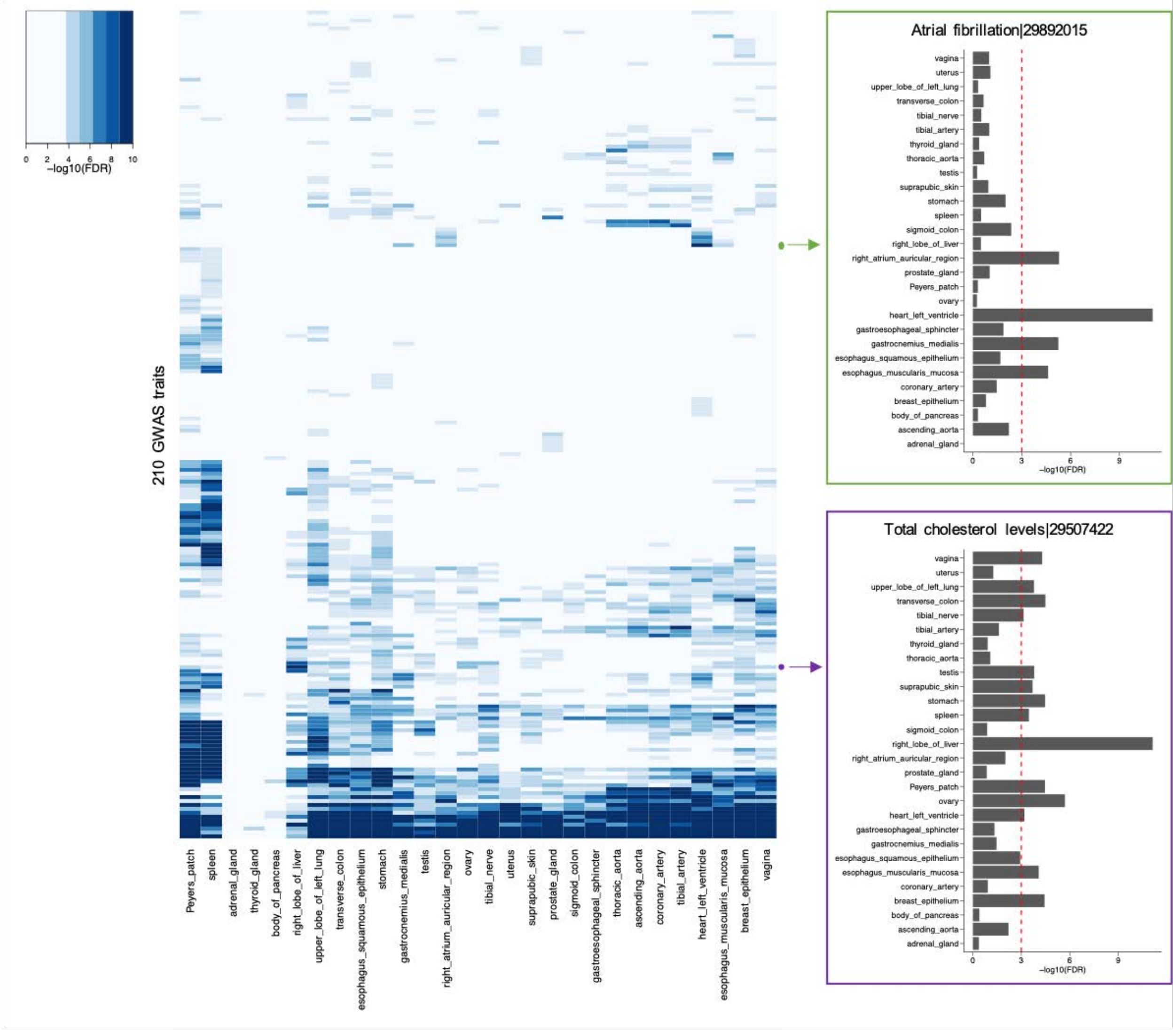
GWAS enrichment (FDR<0.001)

B



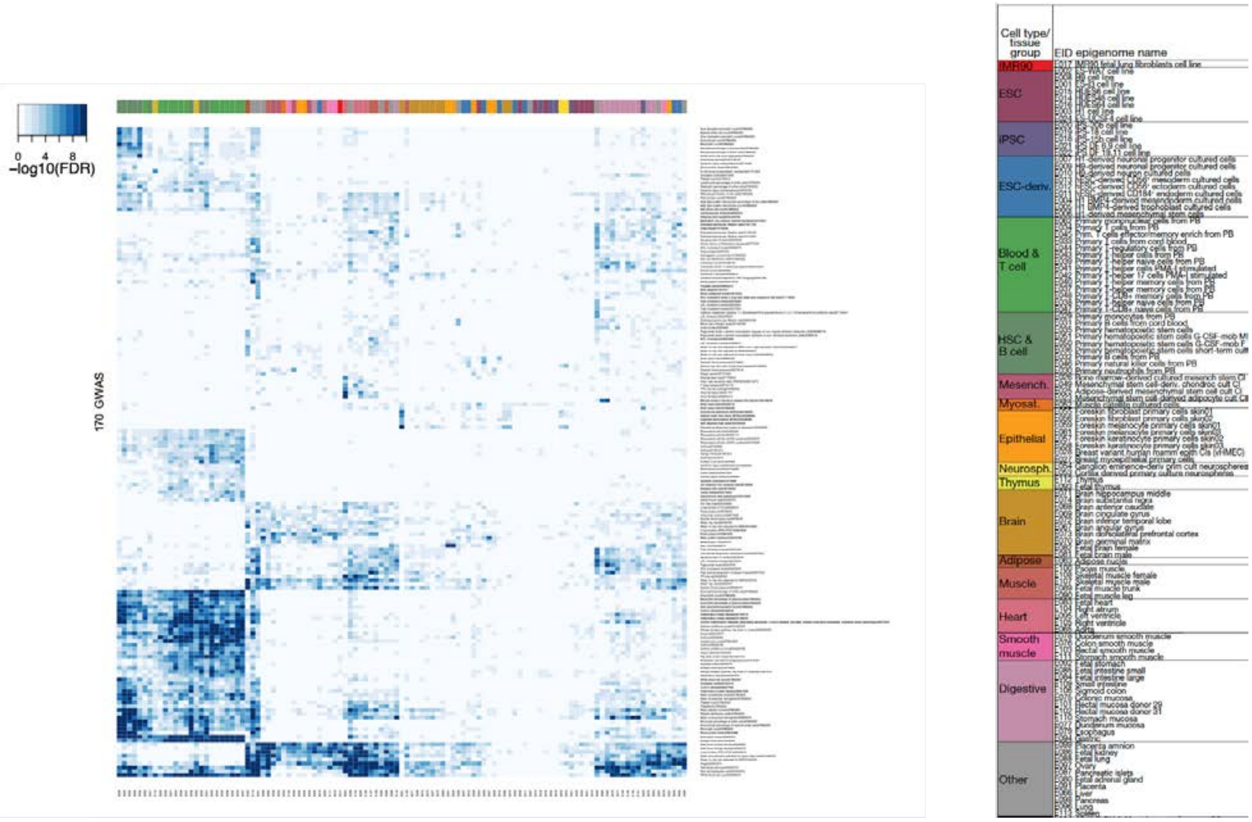

C
cCRE
cCRE 500bp extension
45 GWAS traits
210 GWAS traits
0
2
4
6
8
10
-log10(FDR)
Peyers_patch
spleen
right_atrium_auricular_region
heart_left_ventricle
thoracic_aorta
ascending_aorta
tibial_artery
coronary_artery
thyroid_gland
adrenal_gland
body_of_pancreas
testis
right_lobe_of_liver
gastroesophageal_sphincter
sigmoid_colon
prostate_gland
esophagus_squamous_epithelium
gastrocnemius_medialis
uterus
suprapubic_skin
upper_lobe_of_left_lung
transverse_colon
stomach
vagina
breast_epithelium
tibial_nerve
esophagus_muscularis_mucosa
ovary
Peyers_patch
spleen
adrenal_gland
thyroid_gland
body_of_pancreas
right_lobe_of_liver
upper_lobe_of_left_lung
transverse_colon
esophagus_squamous_epithelium
stomach
gastrocnemius_medialis
testis
right_atrium_auricular_region
ovary
tibial_nerve
uterus
suprapubic_skin
prostate_gland
sigmoid_colon
gastroesophageal_sphincter
thoracic_aorta
ascending_aorta
coronary_artery
tibial_artery
heart_left_ventricle
esophagus_muscularis_mucosa
breast_epithelium
vagina

D

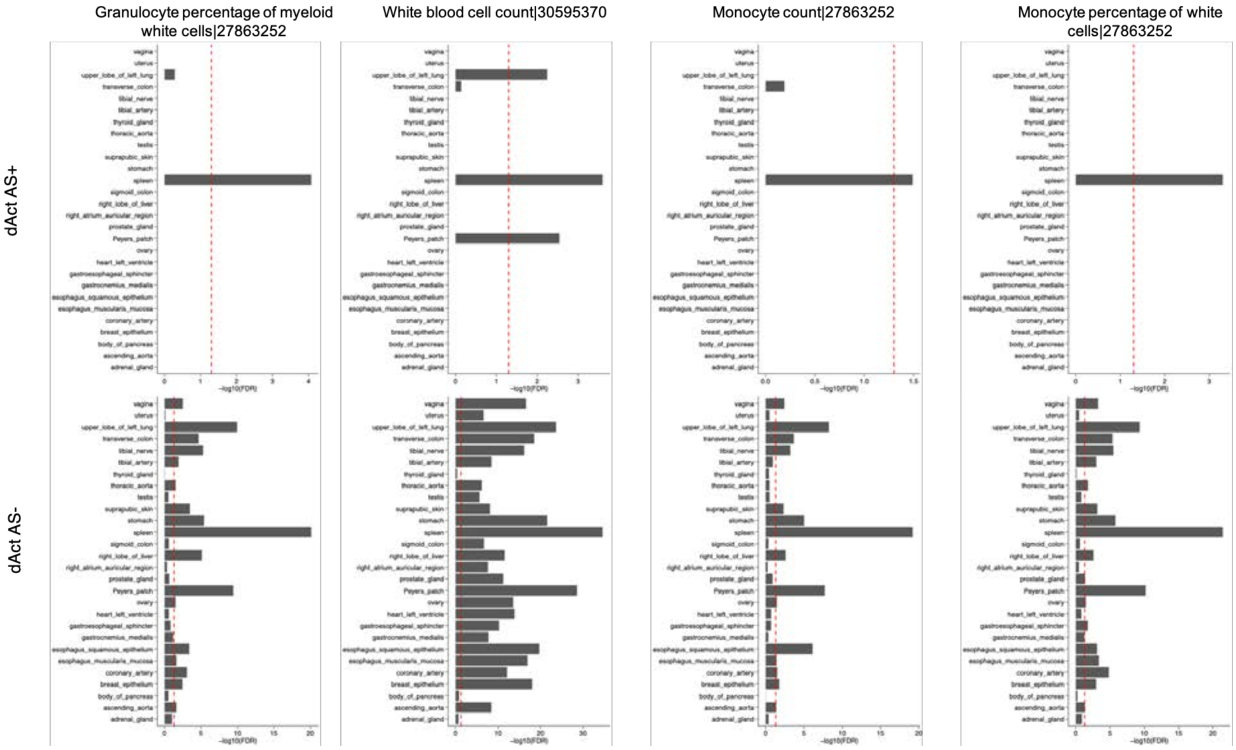


0
2
4
6
8
10
−log10(FDR)
210 GWAS traits
Peyers_patch
spleen
adrenal_gland
thyroid_gland
body_of_pancreas
right_lobe_of_liver
upper_lobe_of_left_lung
transverse_colon
esophagus_squamous_epithelium
stomach
gastrocnemius_medialis
testis
right_atrium_auricular_region
ovary
tibial_nerve
uterus
suprapubic_skin
prostate_gland
sigmoid_colon
gastroesophageal_sphincter
thoracic_aorta
ascending_aorta
coronary_artery
tibial_artery
heart_left_ventricle
esophagus_muscularis_mucosa
breast_epithelium
vagina
Atrial fibrillation|29892015
vagina
uterus
upper_lobe_of_left_lung
transverse_colon
tibial_nerve
tibial_artery
thyroid_gland
thoracic_aorta
testis
suprapubic_skin
stomach
spleen
sigmoid_colon
right_lobe_of_liver
right_atrium_auricular_region
prostate_gland
Peyers_patch
ovary
heart_left_ventricle
gastroesophageal_sphincter
gastrocnemius_medialis
esophagus_squamous_epithelium
esophagus_muscularis_mucosa
coronary_artery
breast_epithelium
body_of_pancreas
ascending_aorta
adrenal_gland
0
3
6
9
−log10(FDR)
Total cholesterol levels|29507422
vagina
uterus
upper_lobe_of_left_lung
transverse_colon
tibial_nerve
tibial_artery
thyroid_gland
thoracic_aorta
testis
suprapubic_skin
stomach
spleen
sigmoid_colon
right_lobe_of_liver
right_atrium_auricular_region
prostate_gland
Peyers_patch
ovary
heart_left_ventricle
gastroesophageal_sphincter
gastrocnemius_medialis
esophagus_squamous_epithelium
esophagus_muscularis_mucosa
coronary_artery
breast_epithelium
body_of_pancreas
ascending_aorta
adrenal_gland
0
3
6
9
−log10(FDR)

E


0 4 8
−log10(FDR)
170 GWAS
Cell type/ tissue group
EID epigenome name
IMR90
E017 IMR90 fetal lung fibroblasts cell line
ESC
ES-WA7 cell line
H9 cell line
H1 cell line
HUES48 cell line
HUES6 cell line
HUES64 cell line
H1 cell line
ES-UCSF4 cell line
iPSC
iPS-20b cell line
iPS-18 cell line
iPS-15b cell line
iPS DF 6.9 cell line
iPS DF 19.11 cell line
ESC-deriv.
H1-derived neuronal progenitor cultured cells
H9-derived neuronal progenitor cultured cells
H9-derived neuron cultured cells
hESC-derived CD56⁺ mesoderm cultured cells
hESC-derived CD56⁺ ectoderm cultured cells
hESC-derived CD184⁺ endoderm cultured cells
H1 BMP4-derived mesendoderm cultured cells
H1 BMP4-derived trophoblast cultured cells
H1-derived mesenchymal stem cells
Blood & T cell
Primary mononuclear cells from PB
Primary T cells from PB
Prim. T cells effector/memory enrich from PB
Primary T cells from cord blood
Primary T regulatory cells from PB
Primary T helper cells from PB
Primary T helper naive cells from PB
Primary T helper cells PMA-I stimulated
Primary T helper 17 cells PMA-I stimulated
Primary T helper memory cells from PB
Primary T helper memory cells from PB
Primary T CD8+ memory cells from PB
Primary T helper naive cells from PB
Primary T CD8+ naive cells from PB
HSC & B cell
Primary monocytes from PB
Primary B cells from cord blood
Primary hematopoietic stem cells
Primary hematopoietic stem cells G-CSF-mob M
Primary hematopoietic stem cells G-CSF-mob F
Primary hematopoietic stem cells short-term cult
Primary B cells from PB
Primary natural killer cells from PB
Primary neutrophils from PB
Mesench.
Bone marrow-derived cultured mesench stem Cl
Mesenchymal stem cell-deriv. chondroc cult Cl
Adipose-derived mesenchymal stem cell cult Cl
Mesenchymal stem cell derived adipocyte cult Cl
Myosat.
Muscle satellite cultured cells
Epithelial
Foreskin fibroblast primary cells skin01
Foreskin fibroblast primary cells skin02
Foreskin melanocyte primary cells skin01
Foreskin melanocyte primary cells skin03
Foreskin keratinocyte primary cells skin02
Foreskin keratinocyte primary cells skin03
Breast variant human mamm epith Cls (vHMEC)
Breast myoepithelial primary cells
Neurosph.
Ganglion eminence-deriv prim cult neurospheres
Cortex derived primary culture neurospheres
Thymus
E112 Thymus
Fetal thymus
Brain
Brain hippocampus middle
Brain substantia nigra
Brain anterior caudate
Brain cingulate gyrus
Brain inferior temporal lobe
Brain angular gyrus
Brain dorsolateral prefrontal cortex
Brain germinal matrix
Fetal brain female
Fetal brain male
Adipose
Adipose nuclei
Muscle
Psoas muscle
Skeletal muscle female
Skeletal muscle male
Fetal muscle trunk
Fetal muscle leg
Heart
Fetal heart
Right atrium
Left ventricle
Right ventricle
Aorta
Smooth muscle
Duodenum smooth muscle
Colon smooth muscle
Rectal smooth muscle
Stomach smooth muscle
Digestive
Fetal stomach
Fetal intestine small
Fetal intestine large
Small intestine
Sigmoid colon
Colonic mucosa
Rectal mucosa donor 29
Rectal mucosa donor 31
Stomach mucosa
Duodenum mucosa
Esophagus
Gastric
Other
Placenta amnion
Fetal kidney
Fetal lung
Ovary
Pancreatic islets
Fetal adrenal gland
Placenta
Liver
Pancreas
Lung
Spleen

F


Granulocyte percentage of myeloid white cells|27863252
White blood cell count|30595370
Monocyte count|27863252
Monocyte percentage of white cells|27863252
dAct AS+
dAct AS-
vagina
uterus
upper_lobe_of_left_lung
transverse_colon
tibial_nerve
tibial_artery
thyroid_gland
thoracic_aorta
testis
suprapubic_skin
stomach
spleen
sigmoid_colon
right_lobe_of_liver
right_atrium_auricular_region
prostate_gland
Peyers_patch
ovary
heart_left_ventricle
gastroesophageal_sphincter
gastrocnemius_medialis
esophagus_squamous_epithelium
esophagus_muscularis_mucosa
coronary_artery
breast_epithelium
body_of_pancreas
ascending_aorta
adrenal_gland
-log10(FDR)
0 1 2 3 4
0 1 2 3
0.0 0.5 1.0 1.5
0 1 2 3
0 5 10 15 20
0 10 20 30
0 5 10 15 20
0 5 10 15 20

Data S25. GWAS enrichment, related to Figure 4 and STAR Methods "Decoration Enrichments" Section

(A) Framework of GWAS enrichment analysis.

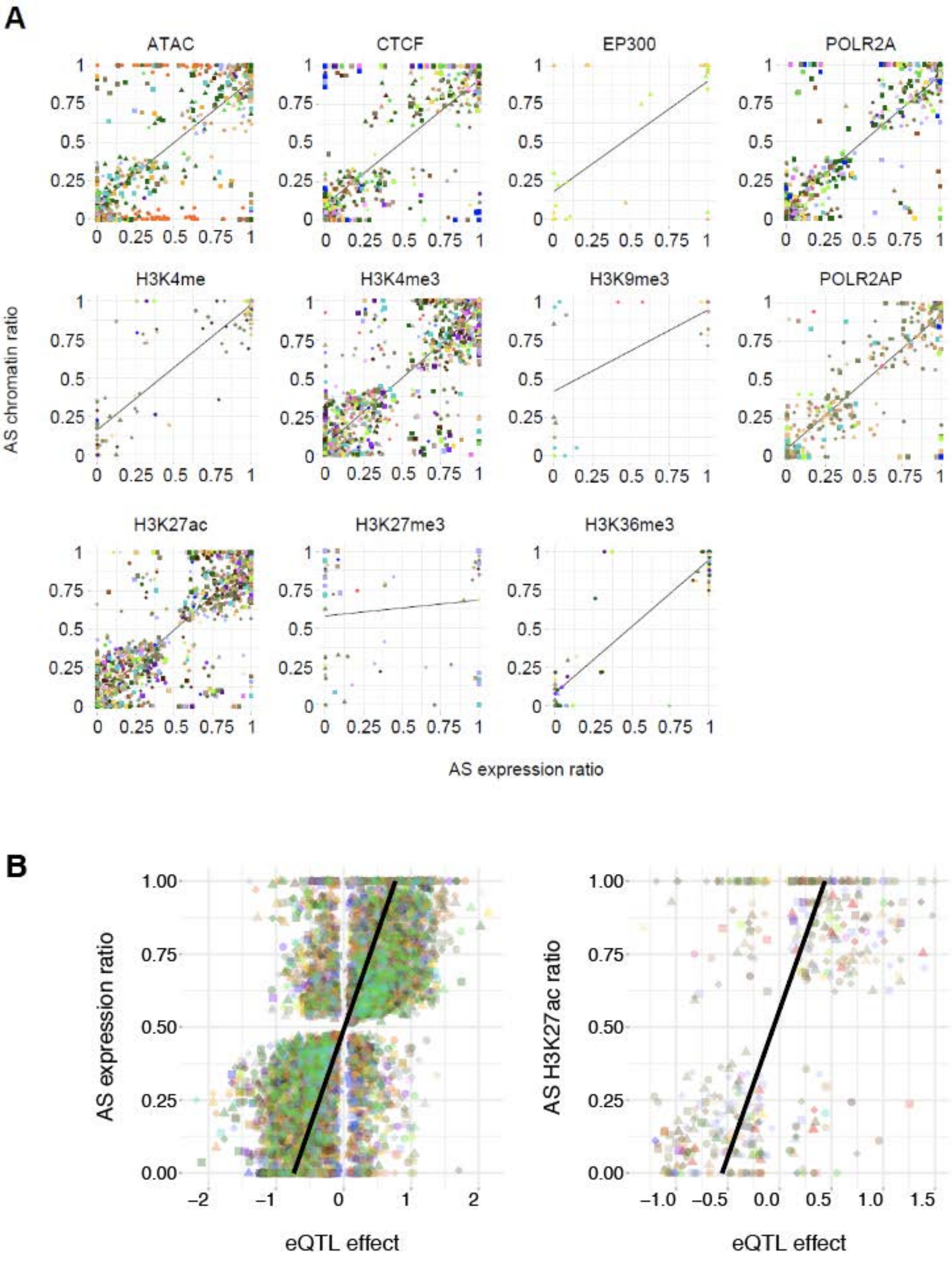
(B) Stratified linkage disequilibrium score regression (LDSC) enrichment: Comparing EN-TEx AS, non-AS, and Roadmap annotations. This is a shadow figure for Figure 10B in the main text. The central heatmap is the stratified LDSC enrichment of various GWAS traits over distal active elements of all EN-TEx tissues. In the left panel, we compare LDSC enrichment of distal active AS and non-AS over all traits for the coronary artery. In the right panel, we compare LDSC enrichment of distal active AS, non-AS, and Roadmap annotations in the right lobe of liver.

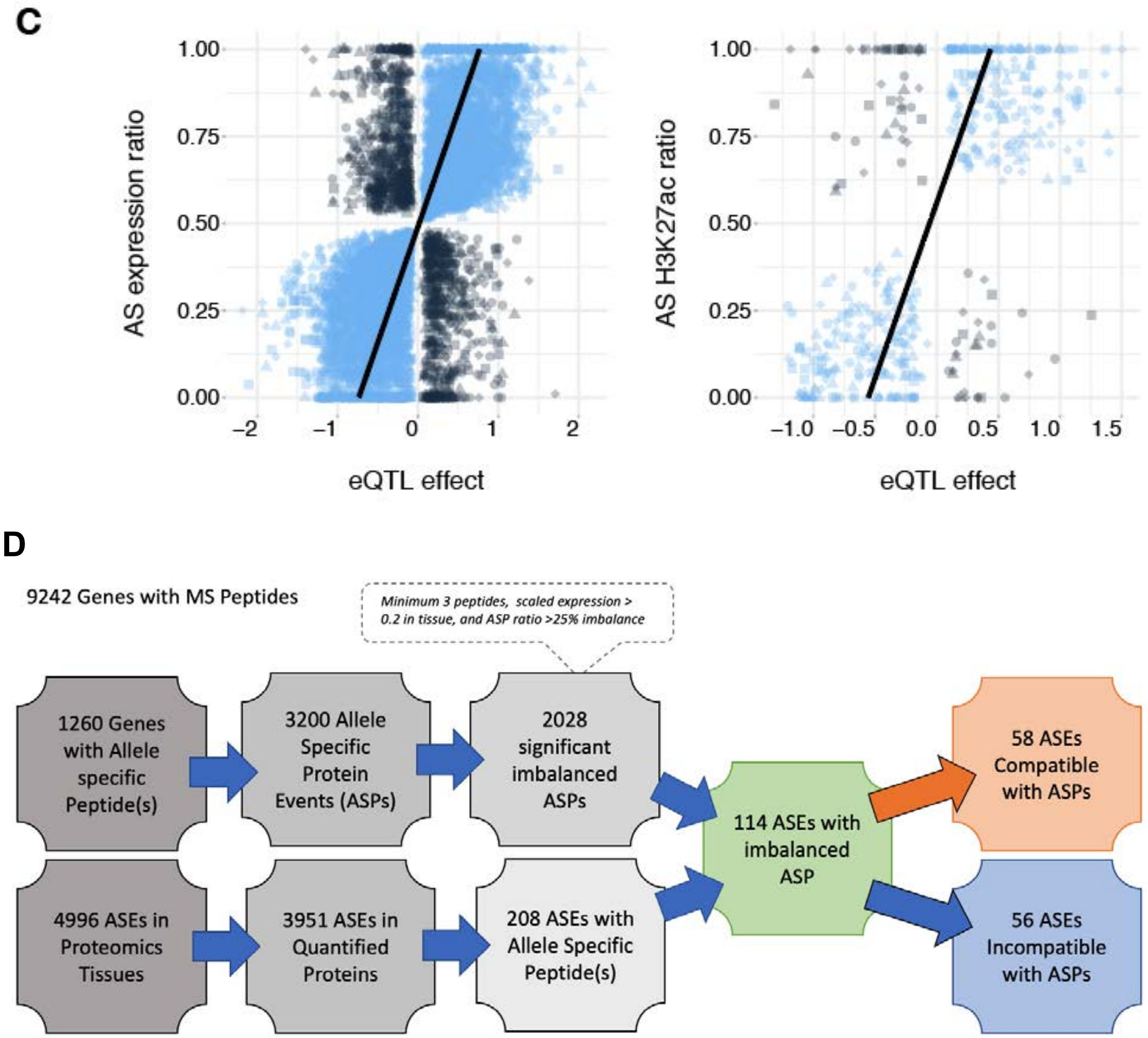
(C) GWAS enrichment: cCREs vs. cCREs with 500 bp extensions. We performed GWAS enrichment analysis on the original cCRE regions and the cCRE regions with a 500 bp extension on both sides. More significantly enriched GWAS traits can be identified on the cCRE regions with extensions, suggesting that it is necessary to include the flanking regions in the GWAS enrichment analysis.

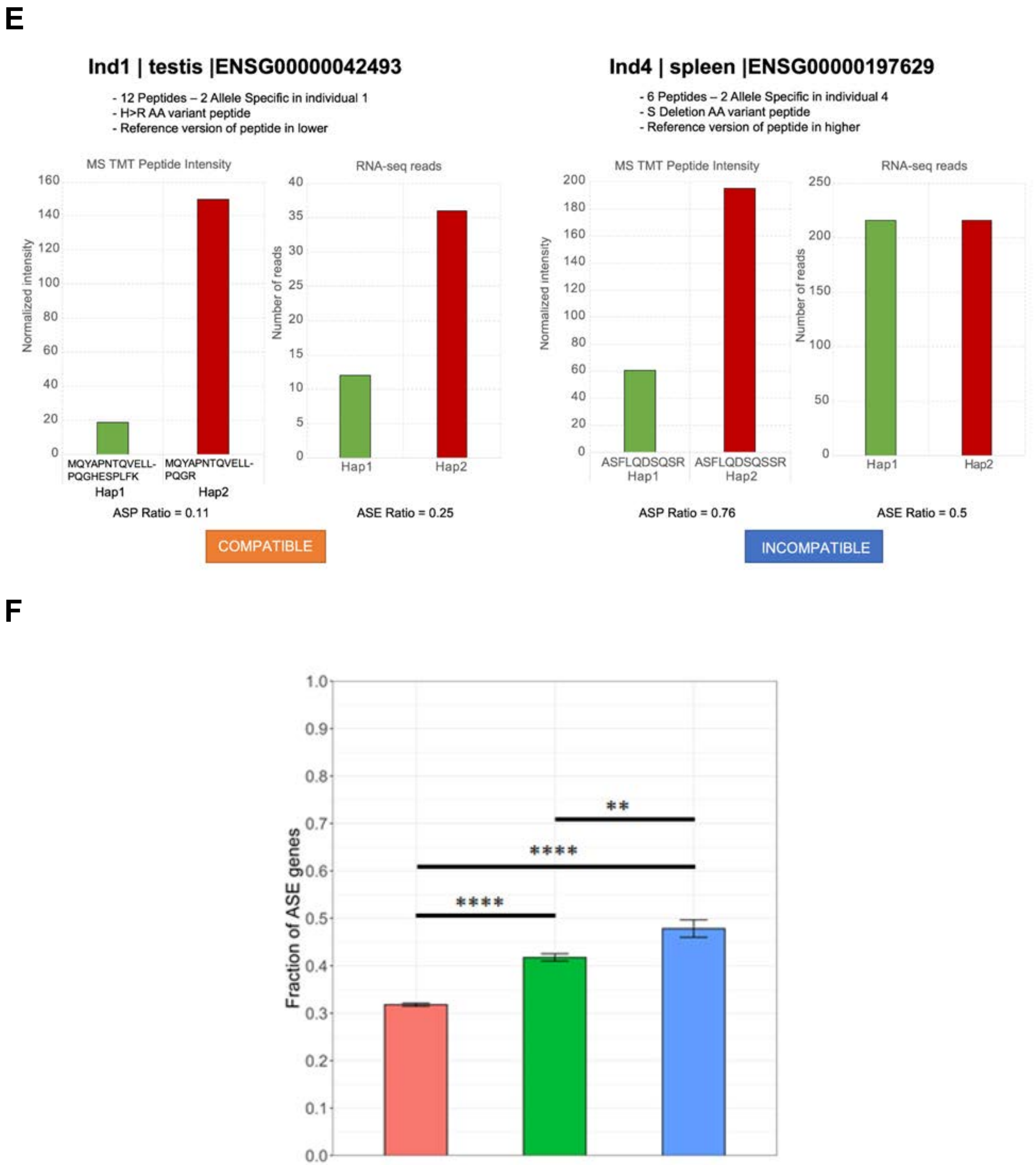
(D) GWAS enrichment across tissues. We selected two GWAS traits, atrial fibrillation and total cholesterol levels, to show their enrichment scores across all the tissue types.

(E) GWAS enrichment for Roadmap annotations. We performed GWAS enrichment analysis on the enhancer annotations from the 127 cell and tissue types from the Roadmap Epigenomics Project. Tissue names are on the horizontal axis and traits are on the vertical axis. Simple clustering of this matrix reveals a blocky structure with sets of traits associated with groups of tissues.

(F) GWAS enrichment for AS vs. non-AS cCREs. We compared the GWAS enrichment scores on the distal active cCREs with (upper) and without (lower) AS signatures using the GWAS tag SNPs from blood-associated traits. Note that the decoration terms are defined in Figure S5A.


A
ATAC
CTCF
EP300
POLR2A
H3K4me
H3K4me3
H3K9me3
POLR2AP
H3K27ac
H3K27me3
H3K36me3
AS chromatin ratio
AS expression ratio
0
0.25
0.5
0.75
1
B
AS expression ratio
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
eQTL effect
AS H3K27ac ratio
−1.0
−0.5
0.0
0.5
1.0
1.5
eQTL effect

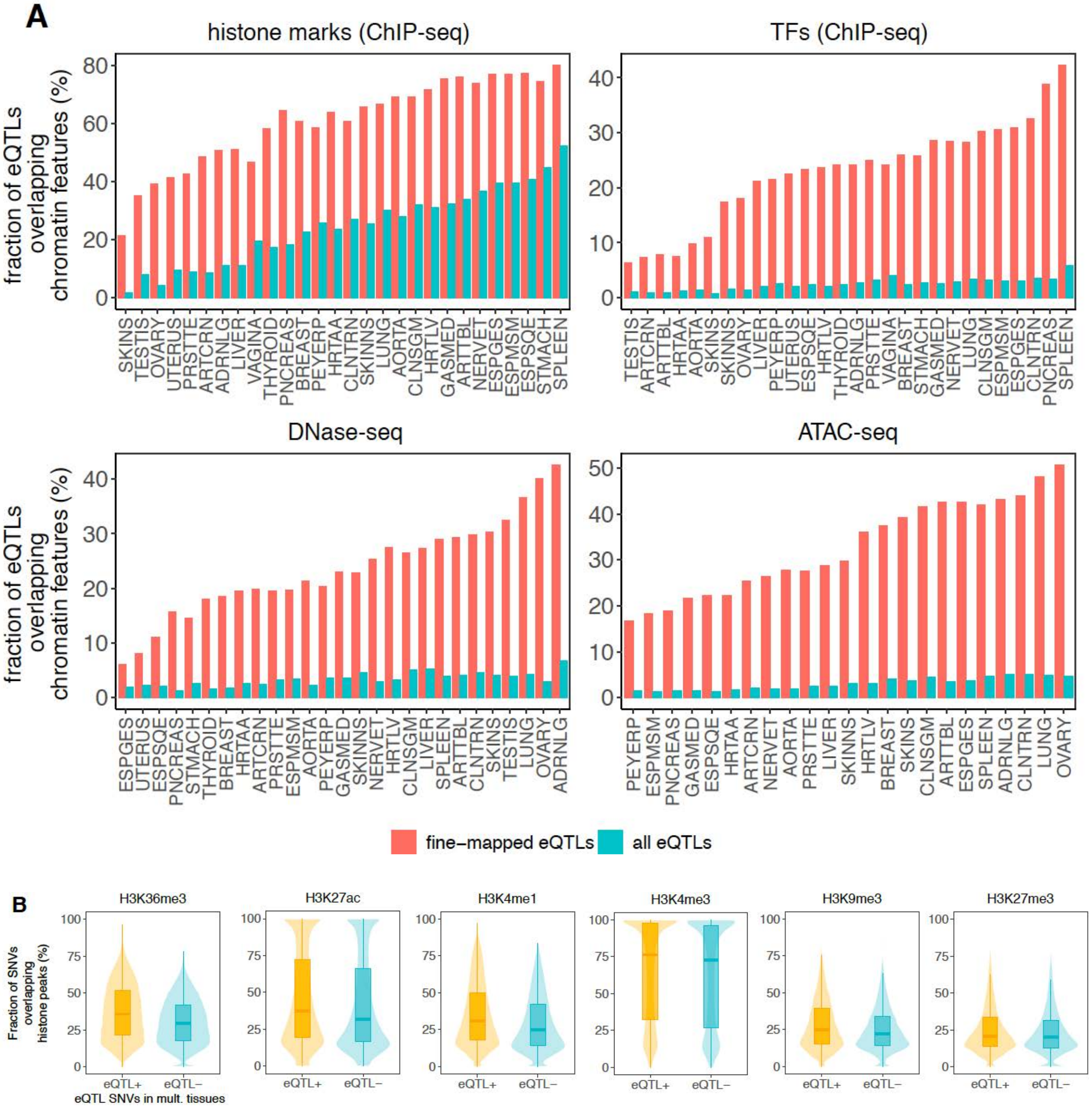

C
AS expression ratio
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
eQTL effect
AS H3K27ac ratio
1.00
0.75
0.50
0.25
0.00
−1.0
−0.5
0.0
0.5
1.0
1.5
eQTL effect
D
9242 Genes with MS Peptides
Minimum 3 peptides, scaled expression > 0.2 in tissue, and ASP ratio >25% imbalance
1260 Genes with Allele specific Peptide(s)
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



Data S26. Compatibility between allelic events, related to Figure 4, Figure S5, and STAR Methods "Compatibility" Section

(A) Compatibility between AS chromatin state of the promoters (+/- 2 kb from the TSS) and the ASE of the corresponding genes. The AS chromatin ratio is the fraction of hap1 ChIP-seq reads among the total number of reads. The ASE ratio is the fraction of hap1 RNA-seq reads among the total number of reads. Each dot is a gene-promoter pair in a given tissue (marked by colors) and individual (marked by shape). See Figure 1A for details regarding the colors and shapes.

(B) (Left) ASE ratio of known GTEx eGenes [20] that are AS in the EN-TEx individuals (fraction of reads of the haplotype with the alternative allele) and eQTL effect size. (Right) AS H3K27ac at hetSNVs that are known GTEx eQTLs. The y-axis shows the fraction of H3K27ac ChIP-seq reads mapped to the alternative allele among the total reads mapped to either allele of an eQTL.
(C) Same plots as Figure 4D (left) and Data S26B (right) but re-colored to show whether or not the eQTL effect (beta coefficient) and the ASE (left)/ASB (right) are compatible.
(D) Flow chart of filtering and ASE/AS proteomics comparison. Proteomics data were mapped at the gene level and filtered for proteins containing allele-specific peptides (ASPs). 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. The p-value was calculated as 0.7 using z-scores.
(E) Compatibility between AS mRNA and ASP calculations. (Left) Example of a compatible ASP and ASE ratio. Both the proteomics and transcriptomics results indicate that the second allele is expressed at a higher level. (Right) Example of an incompatible ASP/ASE pairing. The transcriptomics result does not show any bias in the gene expression; however, the second allele is more highly expressed at the protein level.
(F) Enrichment of ASE genes near ASM promoters. Enrichment of ASE genes near ASM promoters with (blue) (χ2-test, OR = 1.96) or without (green) (χ2-test, OR = 1.54) AS TF binding, relative to genes near ASM non-cCREs (red). ** p < 0.01, **** p < 0.0001.

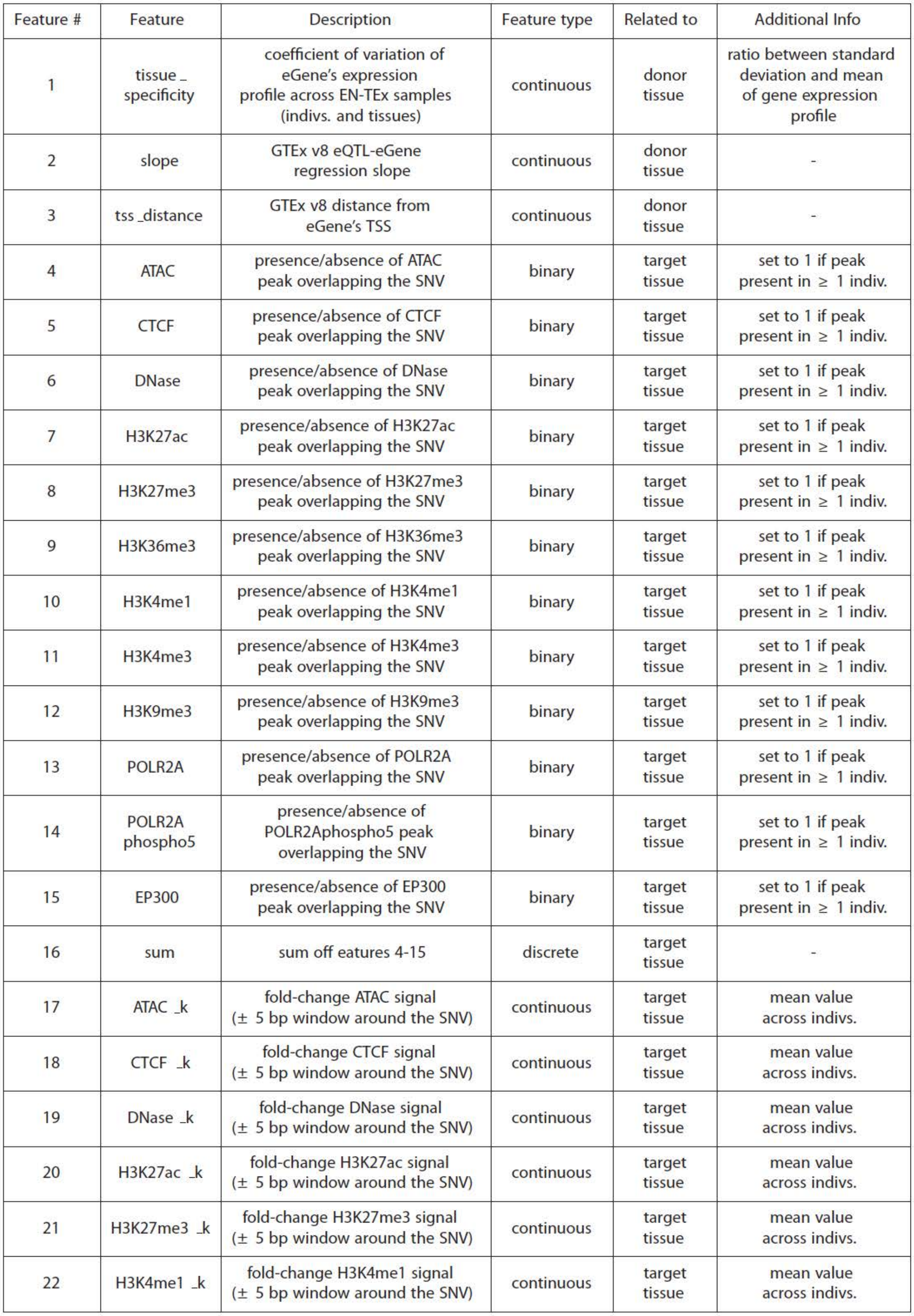


Data S27. Correlation between chromatin features and eQTL activity, related to Figure S6 and STAR Methods "transferQTL Model" Section

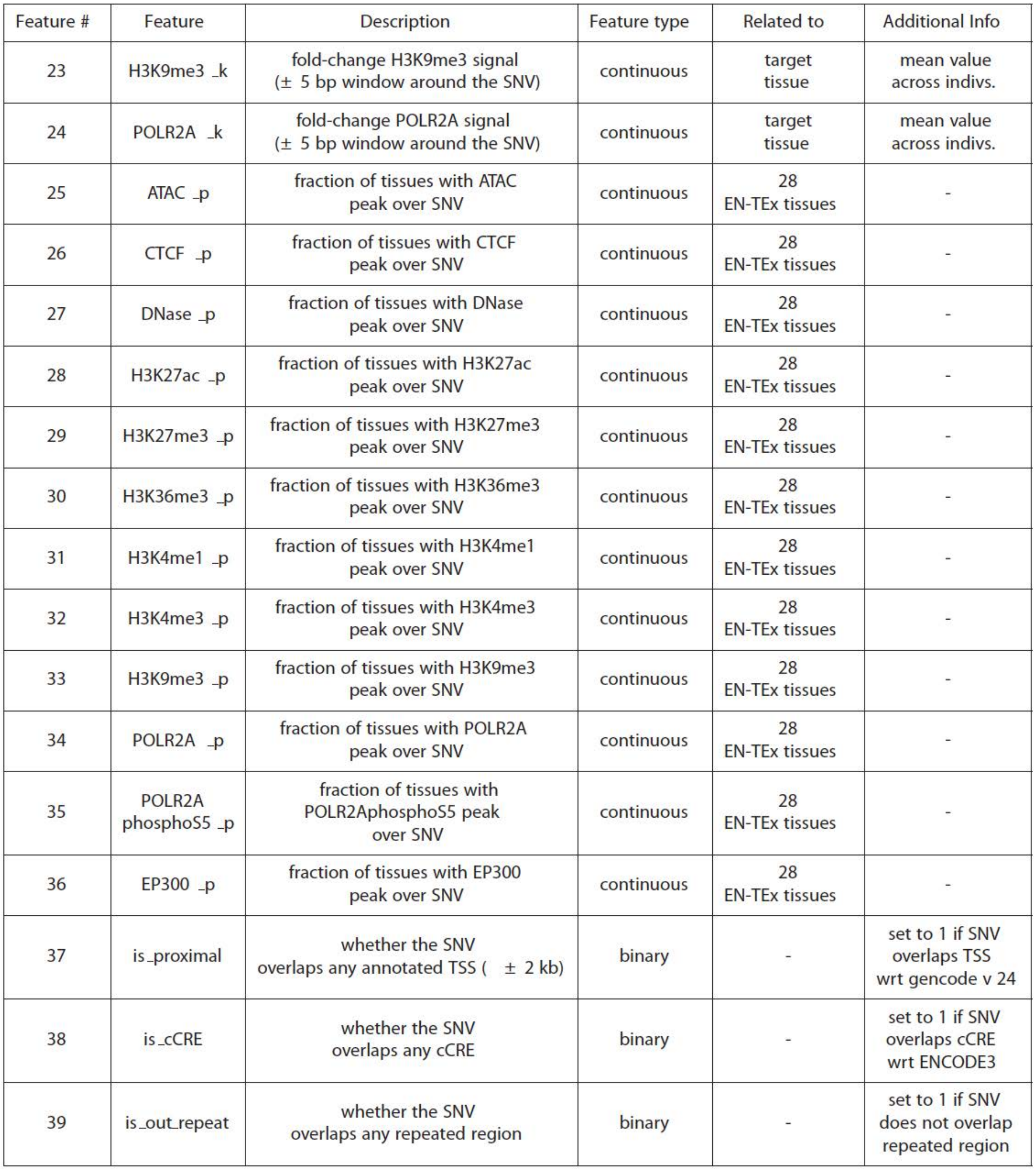
(A) Chromatin features can help prioritize causal eQTLs. Barplots showing the percentage of eQTLs overlapping a given feature. The fraction of fine-mapped (causal) eQTLs overlapping chromatin features is higher compared with the total set of GTEx eQTLs reported in a given tissue. In the case of histone marks and TFs, we report the proportion of eQTLs/fine-mapped eQTLs overlapping any of the six histone marks (H3K27ac, H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3) and any of the four TFs (CTCF, EP300, POLR2A, POLR2APhosphoS5) assayed by the EN-TEx project, respectively.

(B) Chromatin-marked loci associated with eQTL activity. We identified 1,353,101 SNVs that show tissue-specific eQTL activity. These SNVs are GTEx eQTLs in ≥ 5 EN-TEx tissues and are not GTEx eQTLs in ≥ 5 other EN-TEx tissues. Thus, for every SNV we defined two groups of tissues: (1) tissues in which the SNV is an eQTL (eQTL+, orange) and (2) tissues in which the SNV is not an eQTL (eQTL-, cyan). Next, for each histone mark we only considered SNVs that overlap with chromatin peaks in ≥ 10% of all EN-TEx ChIP-seq samples for that particular histone mark. We observed that SNVs are more likely to be marked by a given histone modification in the tissues in which they are eQTLs, compared with the tissues in which they are not eQTLs (p-value < 2.2e-16 for all histone marks, Wilcoxon paired test). We indicate n as the number of SNVs in the violin plots of each histone mark (H3K36me3: n = 232,610; H3K27ac: n = 176,260; H3K4me1: n = 191,689; H3K4me3: n = 64,650; H3K9me3: n = 50,236; H3K27me3: n = 50,973). All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a band inside the box showing the median value and whiskers representing 1.5x the interquartile range.

A



Feature #	Feature	Description	Feature type	Related to	Additional Info
1	tissue _ specificity	coefficient of variation of eGene's expression profile across EN-TEx samples (indivs. and tissues)	continuous	donor tissue	ratio between standard deviation and mean of gene expression profile
2	slope	GTEx v8 eQTL-eGene regression slope	continuous	donor tissue	-
3	tss _distance	GTEx v8 distance from eGene's TSS	continuous	donor tissue	-
4	ATAC	presence/absence of ATAC peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
5	CTCF	presence/absence of CTCF peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
6	DNase	presence/absence of DNase peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
7	H3K27ac	presence/absence of H3K27ac peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
8	H3K27me3	presence/absence of H3K27me3 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
9	H3K36me3	presence/absence of H3K36me3 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
10	H3K4me1	presence/absence of H3K4me1 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
11	H3K4me3	presence/absence of H3K4me3 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
12	H3K9me3	presence/absence of H3K9me3 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
13	POLR2A	presence/absence of POLR2A peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
14	POLR2A phospho5	presence/absence of POLR2Aphospho5 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
15	EP300	presence/absence of EP300 peak overlapping the SNV	binary	target tissue	set to 1 if peak present in ≥ 1 indiv.
16	sum	sum off eatures 4-15	discrete	target tissue	-
17	ATAC _k	fold-change ATAC signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
18	CTCF _k	fold-change CTCF signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
19	DNase _k	fold-change DNase signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
20	H3K27ac _k	fold-change H3K27ac signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
21	H3K27me3 _k	fold-change H3K27me3 signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
22	H3K4me1 _k	fold-change H3K4me1 signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.

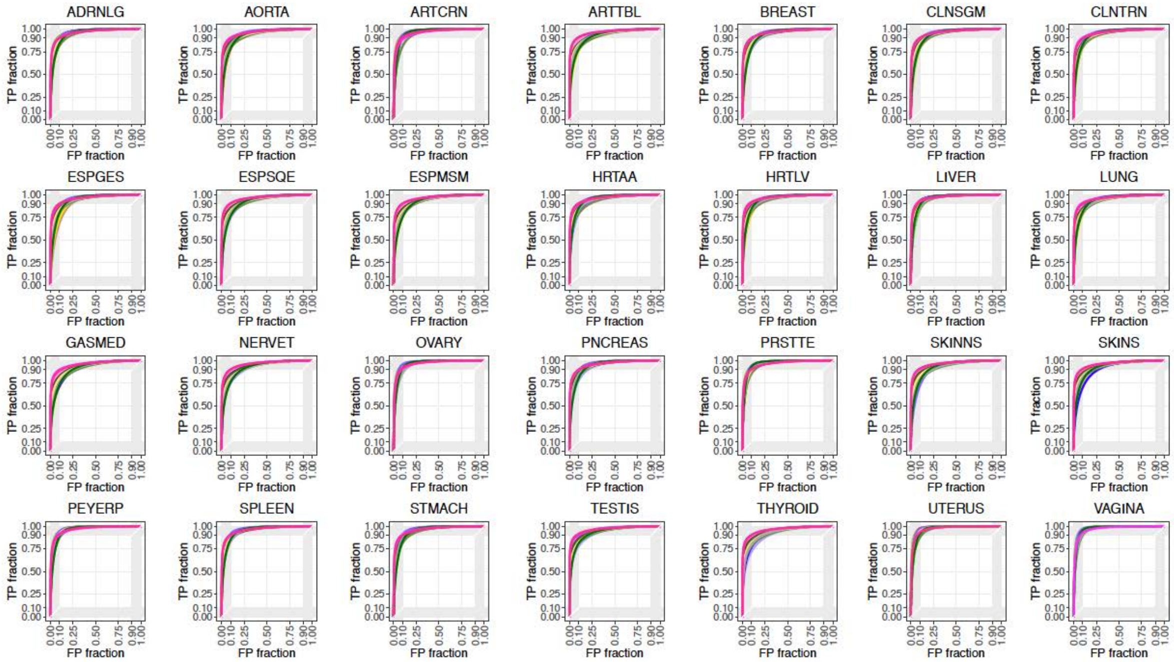


Feature #	Feature	Description	Feature type	Related to	Additional Info
23	H3K9me3 _k	fold-change H3K9me3 signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
24	POLR2A _k	fold-change POLR2A signal (± 5 bp window around the SNV)	continuous	target tissue	mean value across indivs.
25	ATAC _p	fraction of tissues with ATAC peak over SNV	continuous	28 EN-TEx tissues	-
26	CTCF _p	fraction of tissues with CTCF peak over SNV	continuous	28 EN-TEx tissues	-
27	DNase _p	fraction of tissues with DNase peak over SNV	continuous	28 EN-TEx tissues	-
28	H3K27ac _p	fraction of tissues with H3K27ac peak over SNV	continuous	28 EN-TEx tissues	-
29	H3K27me3 _p	fraction of tissues with H3K27me3 peak over SNV	continuous	28 EN-TEx tissues	-
30	H3K36me3 _p	fraction of tissues with H3K36me3 peak over SNV	continuous	28 EN-TEx tissues	-
31	H3K4me1 _p	fraction of tissues with H3K4me1 peak over SNV	continuous	28 EN-TEx tissues	-
32	H3K4me3 _p	fraction of tissues with H3K4me3 peak over SNV	continuous	28 EN-TEx tissues	-
33	H3K9me3 _p	fraction of tissues with H3K9me3 peak over SNV	continuous	28 EN-TEx tissues	-
34	POLR2A _p	fraction of tissues with POLR2A peak over SNV	continuous	28 EN-TEx tissues	-
35	POLR2A phosphoS5 _p	fraction of tissues with POLR2AphosphoS5 peak over SNV	continuous	28 EN-TEx tissues	-
36	EP300 _p	fraction of tissues with EP300 peak over SNV	continuous	28 EN-TEx tissues	-
37	is_proximal	whether the SNV overlaps any annotated TSS (± 2 kb)	binary	-	set to 1 if SNV overlaps TSS wrt gencode v 24
38	is_cCRE	whether the SNV overlaps any cCRE	binary	-	set to 1 if SNV overlaps cCRE wrt ENCODE3
39	is_out_repeat	whether the SNV overlaps any repeated region	binary	-	set to 1 if SNV does not overlap repeated region

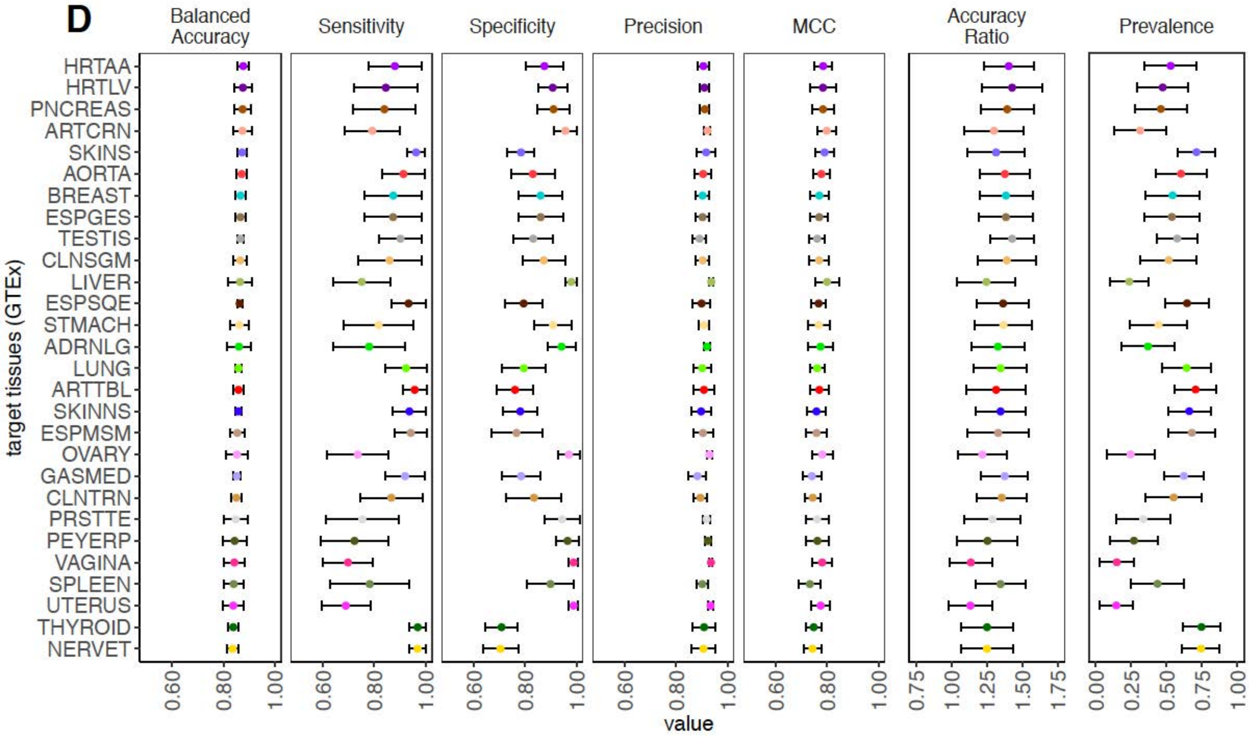
B

Metric	Description
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Precision	$\frac{TP}{TP + FP}$
Balanced Accuracy	$\frac{\text{sensitivity} + \text{specificity}}{2}$
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
No-information Rate	the largest proportion of the observed classes
Accuracy Ratio	$\frac{\text{Accuracy}}{\text{No-information Rate}}$
Prevalence	$\frac{TP + FN}{TP + TN + FP + FN}$

C

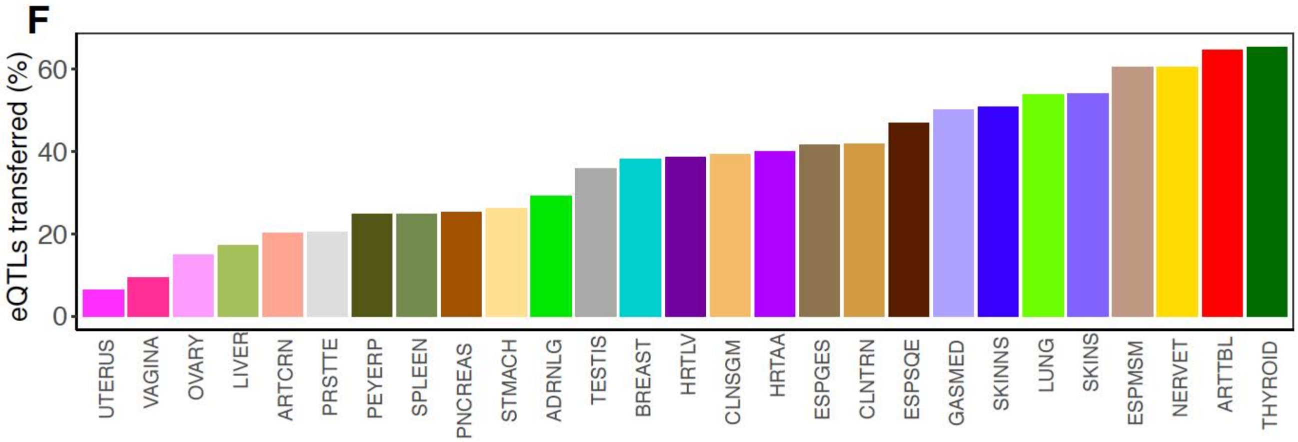


D



E

		Donor Tissue	Label	N. of samples	N. of GTEx eQTLs	N. of likely eQTLs
1	■	Adrenal Gland	ADRNLG	233	422,213	81,933
2	■	Artery Aorta	AORTA	387	724,353	202,090
3	■	Artery Coronary	ARTCRN	213	336,341	65,570
4	■	Artery Tibial	ARTTBL	584	911,849	261,926
5	■	Breast - Mammary Tissue	BREAST	396	586,535	174,403
6	■	Colon - Sigmoid	CLNSGM	318	573,063	160,106
7	■	Colon - Transverse	CLNTRN	368	607,436	190,233
8	■	Esophagus - Gastroesophageal Junction	ESPGES	330	579,297	173,447
9	■	Esophagus - Mucosa	ESPSQE	497	843,932	239,419
10	■	Esophagus - Muscularis	ESPMSM	465	829,886	252,055
11	■	Heart - Atrial Appendage	HRTAA	372	641,282	162,689
12	■	Heart - Left Ventricle	HRTLV	386	577,859	131,671
13	■	Liver	LIVER	208	293,783	37,803
14	■	Lung	LUNG	515	776,962	231,600
15	■	Muscle - Skeletal	GASMED	706	869,283	259,837
16	■	Nerve - Tibial	NERVET	532	1,007,246	296,071
17	■	Ovary	OVARY	167	272,146	43,010
18	■	Pancreas	PNCREAS	305	552,290	122,906
19	■	Prostate	PRSTTE	221	361,651	73,230
20	■	Skin - Not Sun Exposed (Suprapubic)	SKINNS	517	861,598	259,642
21	■	Skin - Sun Exposed (Lower leg)	SKINS	605	967,288	239,637
22	■	Small Intestine - Terminal Ileum	PEYERP	174	290,755	51,824
23	■	Spleen	SPLEEN	227	520,408	122,690
24	■	Stomach	STMACH	324	483,265	121,167
25	■	Testis	TESTIS	322	932,078	208,046
26	■	Thyroid	THYROID	574	998,513	287,261
27	■	Uterus	UTERUS	129	163,537	21,789
28	■	Vagina	VAGINA	141	171,342	21,898



G

Target Tissue	Total	Predicted	Novel
THYROID	1,467,572 (1,290,756)	958,740 (841,058)	816,420 (698,738)
ARTTBL	1,467,916 (1,286,326)	947,245 (828,900)	837,659 (719,314)
NERVET	1,467,940 (1,289,188)	887,180 (777,348)	758,986 (649,154)
ESPMSM	1,468,001 (1,289,266)	886,159 (776,500)	799,913 (690,254)
SKINS	1,468,004 (1,288,495)	793,089 (695,178)	683,870 (585,959)
LUNG	1,467,841 (1,289,600)	789,415 (692,658)	720,074 (623,317)
SKINNS	1,467,830 (1,288,516)	744,235 (652,032)	656,856 (564,653)
GASMED	1,467,849 (1,284,844)	734,548 (641,891)	646,988 (554,331)
ESPSQE	1,467,929 (1,285,876)	688,868 (603,916)	615,039 (530,087)
CLNTRN	1,467,703 (1,288,973)	612,148 (536,318)	575,364 (499,534)
ESPGES	1,467,849 (1,289,115)	609,024 (534,229)	575,021 (500,226)
HRTAA	1,467,844 (1,287,852)	586,424 (515,551)	549,573 (478,700)
CLNSGM	1,467,970 (1,289,237)	574,549 (504,154)	544,411 (474,016)
HRTLV	1,467,925 (1,286,240)	565,778 (496,803)	533,022 (464,047)
BREAST	1,467,981 (1,288,553)	560,010 (490,672)	526,559 (457,221)
TESTIS	1,467,715 (1,293,072)	526,338 (461,012)	460,638 (395,312)
ADRNLG	1,467,819 (1,289,468)	427,103 (377,276)	413,087 (363,260)
STMACH	1,467,763 (1,289,089)	382,833 (335,265)	366,193 (318,625)
PNCREAS	1,467,795 (1,287,656)	369,129 (323,967)	349,283 (304,121)
PEYERP	1,467,623 (1,290,110)	364,123 (322,879)	357,375 (316,131)
SPLEEN	1,467,822 (1,290,788)	364,926 (319,817)	347,251 (302,142)
PRSTTE	1,467,921 (1,289,186)	299,553 (263,773)	290,817 (255,037)
ARTCRN	1,467,752 (1,287,251)	297,262 (260,061)	291,965 (254,764)
LIVER	1,467,913 (1,285,031)	251,342 (220,727)	245,376 (214,761)
OVARY	1,467,879 (1,289,709)	219,686 (193,123)	215,232 (188,669)
VAGINA	1,467,893 (1,290,434)	136,579 (117,983)	134,710 (116,114)
UTERUS	1,467,947 (1,288,382)	94,299 (83,539)	93,199 (82,439)

Data S28. Building a predictive model that transfers eQTLs from a donor tissue to a target tissue: features, performance, and validation, related to Figure 5, Figure S6, and STAR Methods "transferQTL Model" Section

(A) List of predictive features employed by the random forest model to predict eQTL activity in a target tissue. Features employed to predict which donor-tissue eQTLs can be transferred to a target tissue. For features 25-36, the fraction is computed over those of the 28 tissues with available data for the relevant experiment (e.g., if no ATAC-seq experiments were performed for lung tissue, then lung is not included in the calculation of ATAC_p).

(B) Description of the metrics used to evaluate the random forest model. These metrics have been used in the submodel evaluations shown in Figure 5B and Data S28D.

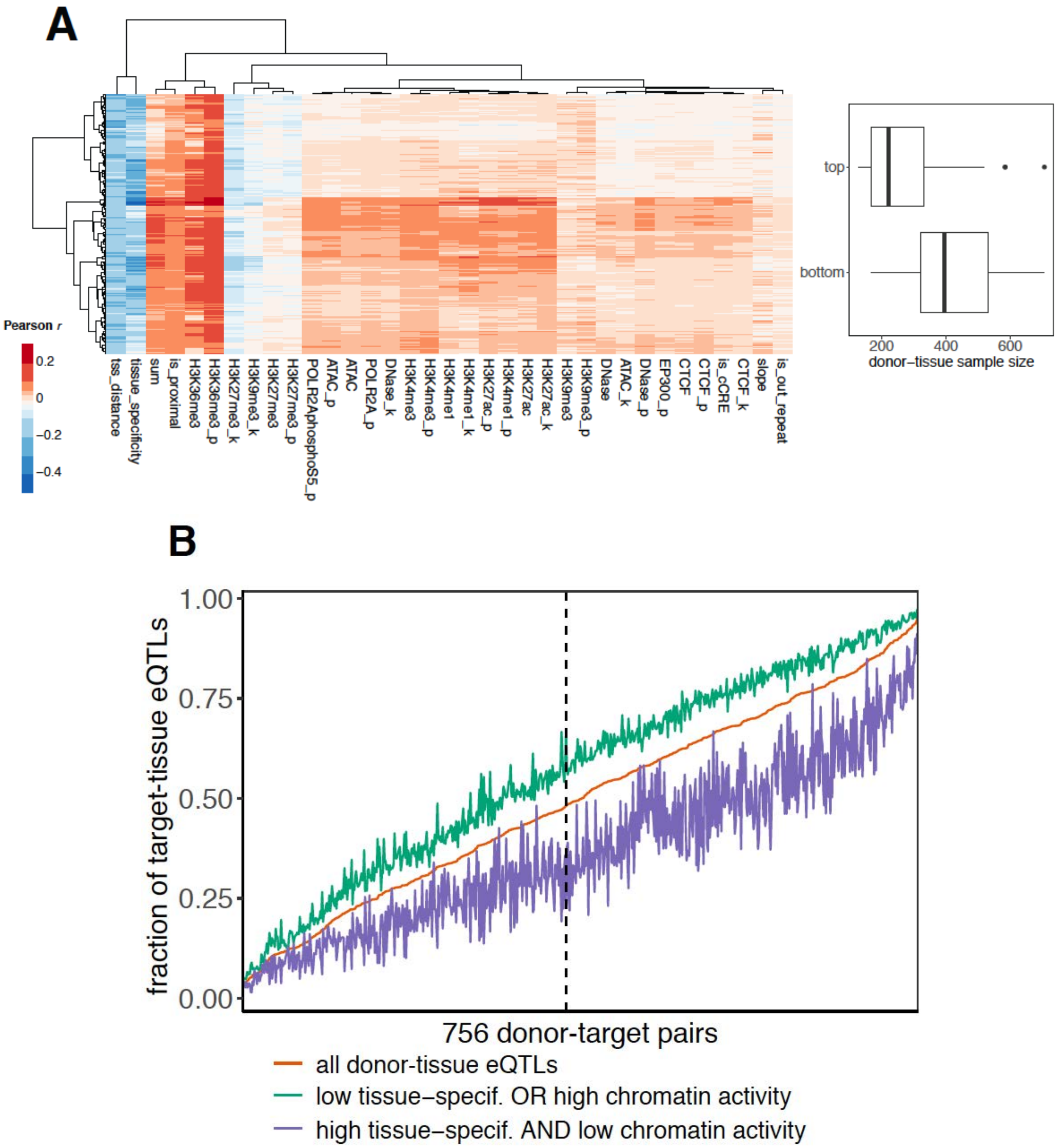
(C) Performance of the random forest submodels by donor tissue. Each plot shows receiver operating characteristic (ROC) curves from multiple target-tissue submodels obtained using the same donor tissue. For instance, the first plot shows ROC curves obtained from all submodels using adrenal gland (ADRNLG) as the donor tissue. ROC curves for each target tissue were computed on a five-fold cross-validation schema and are color-coded in the figure (see Data S28E for a correspondence between tissues and colors).

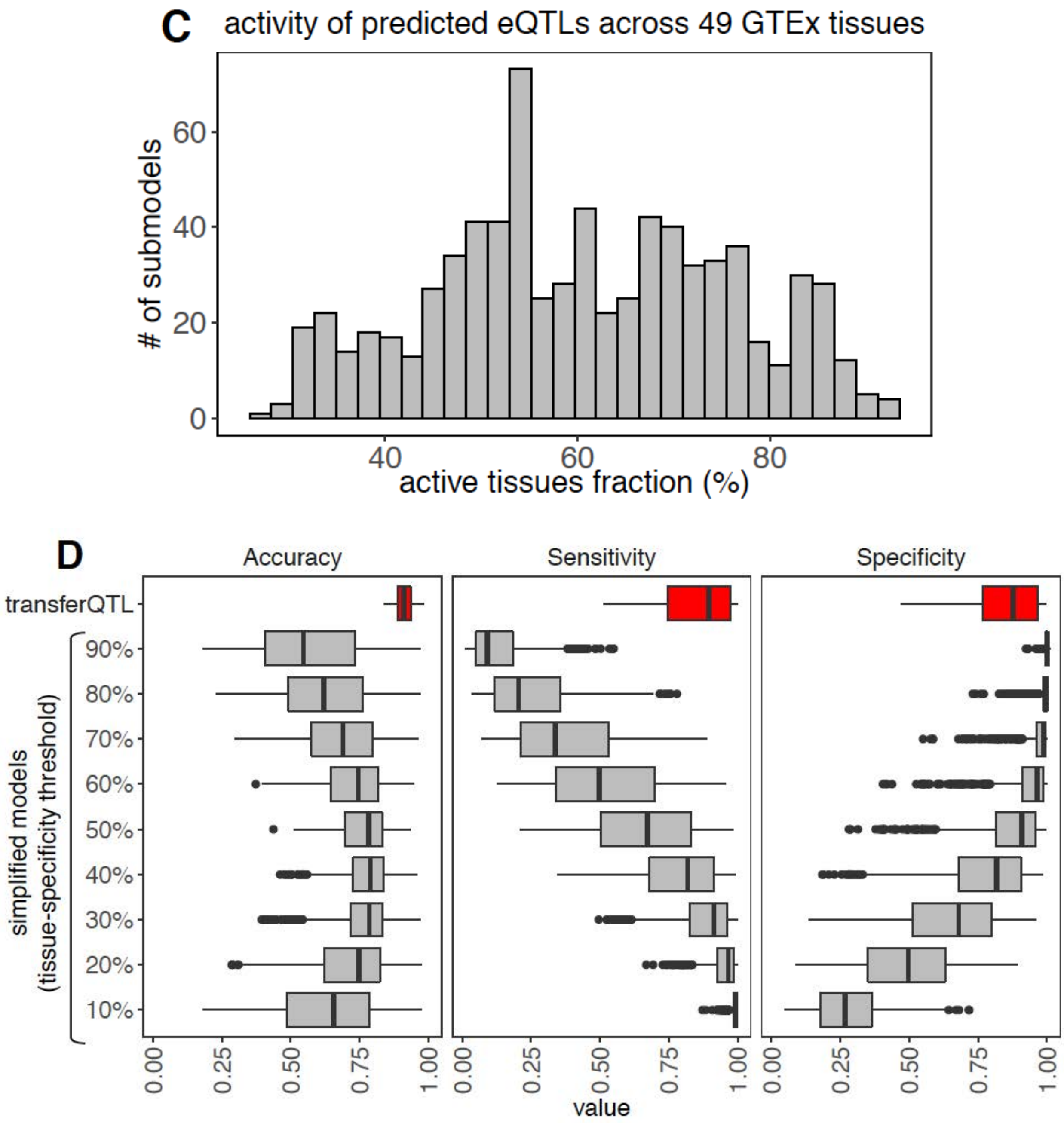
(D) Performance of the random forest submodels by target tissue. Dotplot reporting, for each target tissue (y-axis), performance metrics (x-axis) of the models obtained by using different donor tissues. For a particular target tissue, we report mean and standard deviation of the metric computed using different donor tissues. See Data S28B for a detailed description of each performance metric.

(E) Number of eQTLs per donor tissue. For each of the 28 deeply sampled EN-TEx tissues, we list (1) the corresponding number of samples with the individual's genotype used by GTEx to perform eQTL analyses (column "N. of samples") (2) the number of eQTLs used by the model when setting the relevant tissue as a "donor tissue" (column "N. of GTEx eQTLs"). More specifically, this number corresponds to the set of eQTLs associated with one single eGene in the donor tissue. We also require this single eGene to have non-missing (e.g., "not NA") coefficient of variation for gene expression across EN-TEx individuals and tissues. In the fifth column we list the number of novel "likely" eQTLs for each tissue. These are the union of all "likely" eQTLs across all submodels' test sets for a given tissue. These are eQTLs not present in the original GTEx catalog for a given tissue, but that were predicted by our model to be active in that tissue. These lists are available on the EN-TEx portal (perTissue.likely.eQTLs.tsv).

(F) Proportion (%) of blood eQTLs from Vosa et al. [21] that can be transferred to each EN-TEx tissue. In this analysis, we aimed to predict the activity of 1,547,430 blood eQTLs from Vosa et al. in every EN-TEx tissue. To do so, we applied, for every target tissue, the submodel previously trained on GTEx data that uses artery aorta as the donor tissue (since blood is not among the EN-TEx tissues, we do not currently have a model using blood as the donor tissue). For each EN-TEx tissue (x-axis) other than artery aorta, we report the proportion of blood eQTLs (y-axis) predicted to be active by the model specific to each target tissue. Because of some overlap between this catalog of blood eQTLs and the GTEx catalogs, we computed these results after excluding the blood eQTLs that were also contained in the original training set used for every target tissue (see also Data S28G).

(G) Number of potentially novel eQTLs predicted for each EN-TEx tissue that are not present in the GTEx catalog. This table reports, for every target tissue, (1) the total number of blood eQTLs (from Vosa et al. [21]) analyzed after removing those contained in the training set (column "total"), (2) the number of blood eQTLs transferred, i.e., predicted to be active (column "predicted"), and (3) within the predicted eQTLs, the number of novel eQTLs that are not contained in the GTEx eQTL catalog for the relevant tissue. We identified 496,477 novel eQTLs on average across tissues. Because not all blood eQTLs from Vosa et al. [21] might also have been tested by GTEx for gene associations in the relevant tissue, we also report (in parentheses) the numbers of total, predicted, and novel eQTLs out of those SNVs also tested by GTEx in a particular tissue.

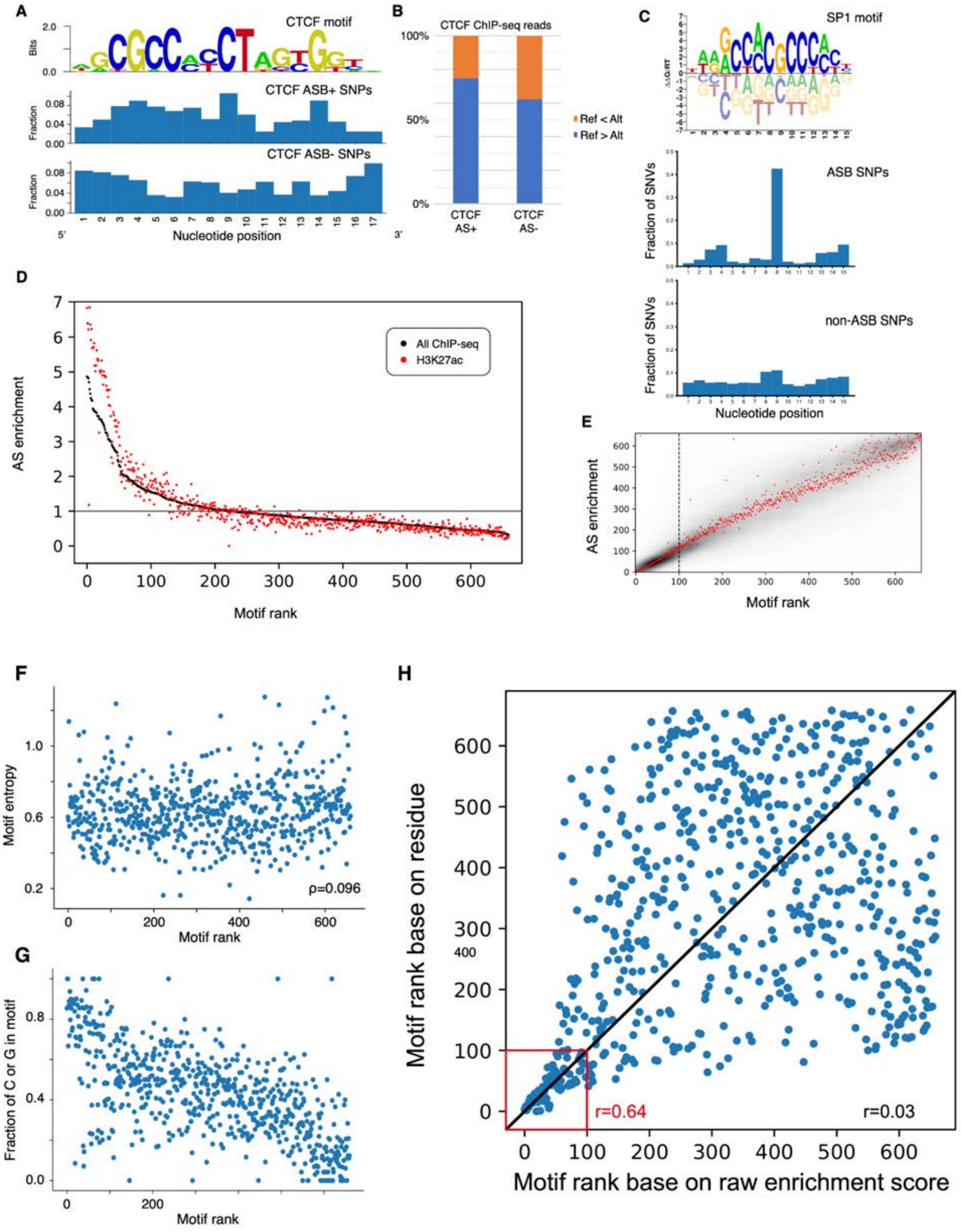

A
Pearson r
0.2
0
−0.2
−0.4
tss_distance
tissue_specificity
sum
is_proximal
H3K36me3
H3K36me3_p
H3K27me3_k
H3K9me3_k
H3K27me3
H3K27me3_p
POLR2AphosphoS5_p
ATAC_p
ATAC
POLR2A_p
DNase_k
H3K4me3
H3K4me3_p
H3K4me1
H3K4me1_k
H3K27ac_p
H3K4me1_p
H3K27ac
H3K27ac_k
H3K9me3
H3K9me3_p
DNase
ATAC_k
DNase_p
EP300_p
CTCF
CTCF_p
is_cRE
CTCF_k
slope
is_out_repeat
top
bottom
200
400
600
donor−tissue sample size
B
1.00
0.75
0.50
0.25
0.00
fraction of target-tissue eQTLs
756 donor-target pairs
all donor-tissue eQTLs
low tissue−specif. OR high chromatin activity
high tissue−specif. AND low chromatin activity



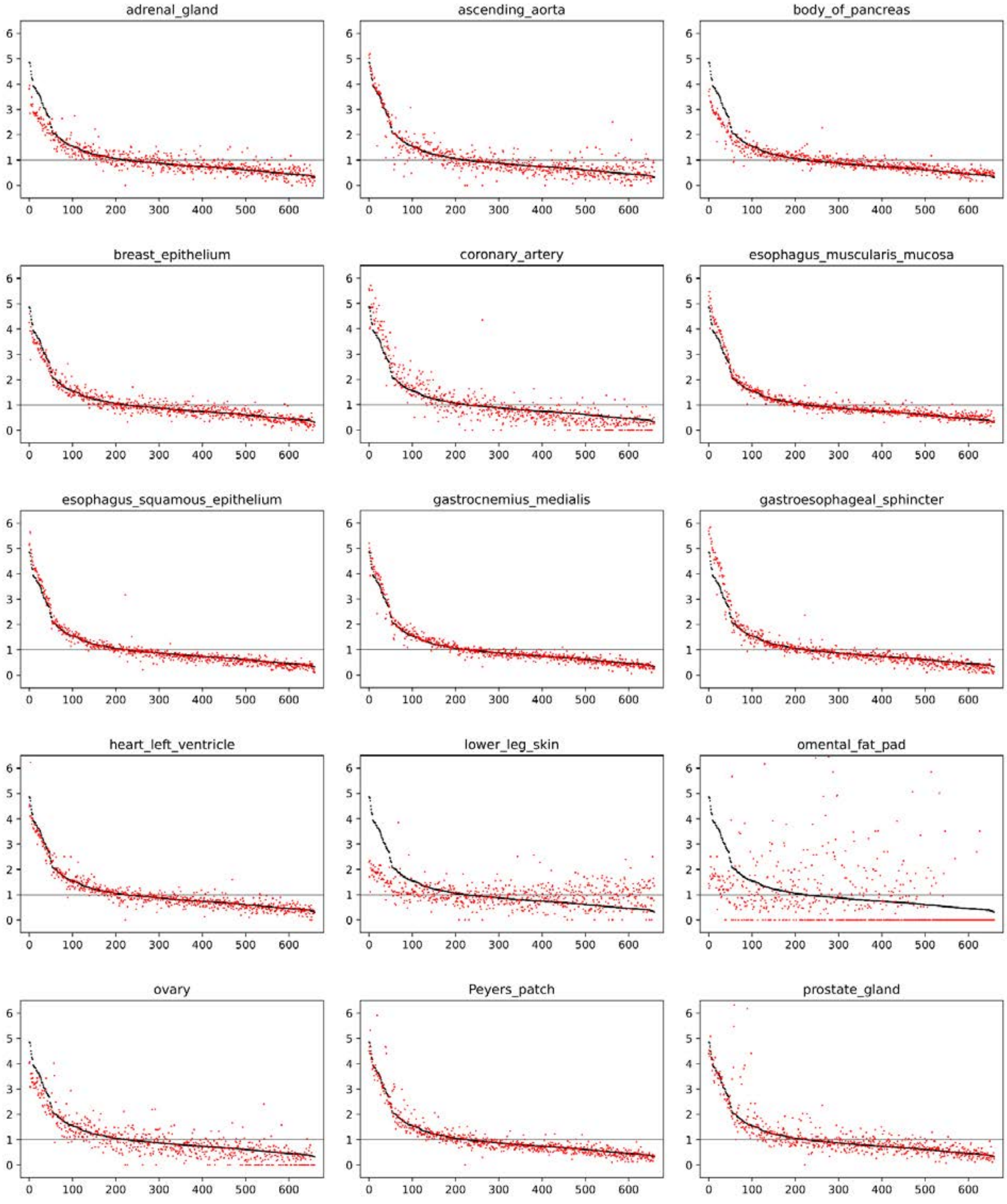
Data S29. Model interpretation and evaluating the impact of tissue specificity on predicted eQTLs, related to Figure 5, Figure S6, and STAR Methods “transferQTL Model” Section

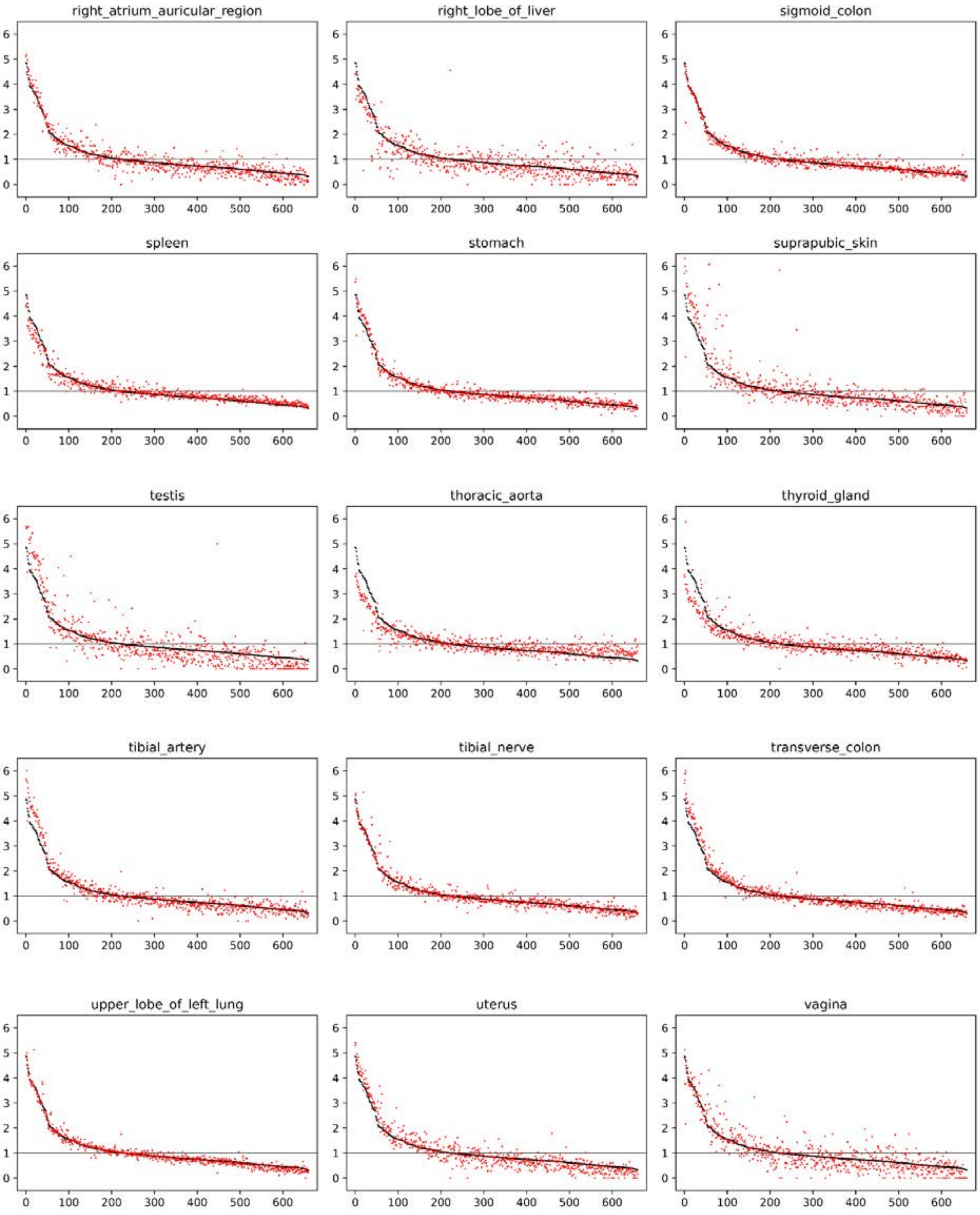
(A) Dissecting the contribution of features to predicting tissue-specific eQTL activity. (Left) Heatmap showing, for each submodel (rows), Pearson’s correlation coefficients between the level of predictive features (columns) at donor-tissue eQTLs and the probability of donor-tissue eQTLs being classified as eQTLs in the target tissue (clustering method: “Ward.D2”, clustering distance: “manhattan”). (Right) Boxplot showing the sample size of the donor tissues used for the submodels in the top and bottom (row) clusters of the heatmap. The sample size of each tissue is reported in Data S28E (column “N. of samples”).

(B) By analyzing the chromatin activity of donor-tissue eQTLs in the target tissue and the tissue specificity of their eGenes, we can identify eQTLs active in the target tissue. Donor-tissue eQTLs either associated with housekeeping eGenes, or those that have high chromatin activity in the target tissue (green line), are more frequently also eQTLs in the target tissue, compared with donor-tissue eQTLs associated with tissue-specific eGenes and that have low chromatin activity in the target tissue (purple line). These enrichments are compared with the proportion of target-tissue eQTLs out of all donor-tissue eQTLs used across the 756 donor-target tissue pairs (orange line). Green line: donor-tissue eQTLs with “tissue specificity” < 0.8 or “sum” (chromatin marking) ≥ 3. Purple line: “tissue specificity” > 5 and “sum” = 0. See also Data S28A for the definition of “sum” and “tissue specificity” features. The dashed vertical line corresponds to the results shown in Figure 5G, i.e., the case using testis as the donor tissue and thyroid as the target tissue.
(C) Activity of predicted eQTLs across 49 GTEx tissues. We computed the tissue specificity of eQTLs predicted by each submodel. The tissue specificity corresponds to the percentage of tissues (out of 49 GTEx tissues) in which an eQTL is found to be active based on the GTEx catalog. We report the distribution of median tissue specificity across all 756 submodels.
(D) Evaluating the impact of tissue specificity on predicted eQTLs. We built a simplified model that transfers eQTLs to a given tissue based on their degree of tissue specificity. We evaluated the performance of this simplified model by using different thresholds of tissue specificity. For instance, 10% corresponds to a model that transfers to a given tissue eQTLs that are active in at least 10% of the GTEx tissues; 90% corresponds to a model that transfers to a given tissue eQTLs that are active in at least 90% of the GTEx tissues. The performance (accuracy, sensitivity, and specificity) is compared to our random-forest model trained on multi-omics EN-TEx data (transferQTL). See also Figure S6D for a comparison based on balanced accuracy.

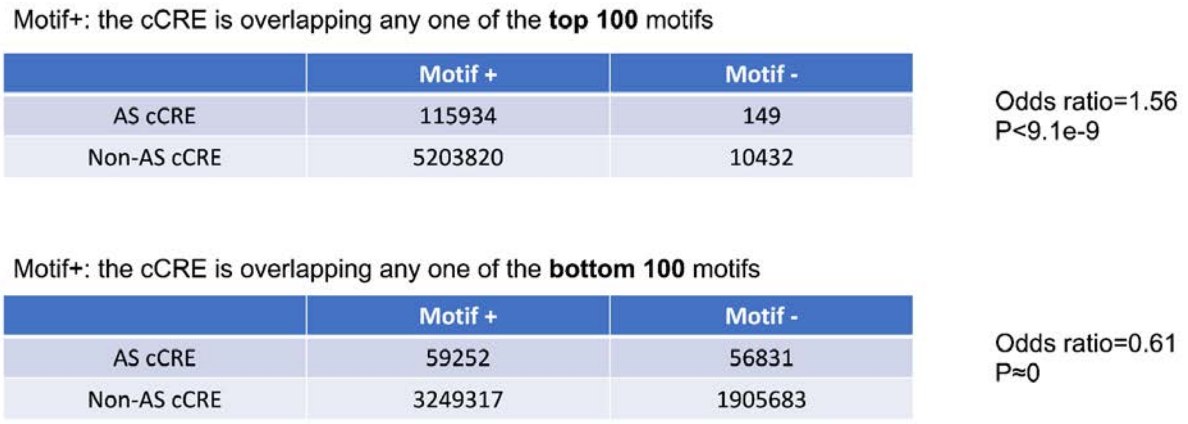

A
CTCF motif
2.0
Bits
0.0
CTCF ASB+ SNPs
Fraction
0.08
0.04
0.00
CTCF ASB- SNPs
Fraction
0.08
0.04
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
5'
Nucleotide position
3'
B
CTCF ChIP-seq reads
100%
50%
0%
Ref < Alt
Ref > Alt
CTCF AS+
CTCF AS-
C
SP1 motif
ΔΔG/RT
ASB SNPs
Fraction of SNVs
non-ASB SNPs
Fraction of SNVs
Nucleotide position
D
All ChIP-seq
H3K27ac
AS enrichment
7
6
5
4
3
2
1
0
0 100 200 300 400 500 600
Motif rank
E
AS enrichment
600
500
400
300
200
100
0
0 100 200 300 400 500 600
Motif rank
F
Motif entropy
1.0
0.6
0.2
ρ=0.096
0 200 400 600
Motif rank
G
Fraction of C or G in motif
0.8
0.4
0.0
0 200
Motif rank
H
Motif rank base on residue
600
500
400
300
400
200
100
0
r=0.64
r=0.03
0 100 200 300 400 500 600
Motif rank base on raw enrichment score

I


adrenal_gland
ascending_aorta
body_of_pancreas
breast_epithelium
coronary_artery
esophagus_muscularis_mucosa
esophagus_squamous_epithelium
gastrocnemius_medialis
gastroesophageal_sphincter
heart_left_ventricle
lower_leg_skin
omental_fat_pad
ovary
Peyers_patch
prostate_gland
0
1
2
3
4
5
6
0
100
200
300
400
500
600


right_atrium_auricular_region
right_lobe_of_liver
sigmoid_colon
spleen
stomach
suprapubic_skin
testis
thoracic_aorta
thyroid_gland
tibial_artery
tibial_nerve
transverse_colon
upper_lobe_of_left_lung
uterus
vagina
0
1
2
3
4
5
6
0
100
200
300
400
500
600

J



Motif+: the cCRE is overlapping any one of the **top 100** motifs

	Motif +	Motif -
AS cCRE	115934	149
Non-AS cCRE	5203820	10432

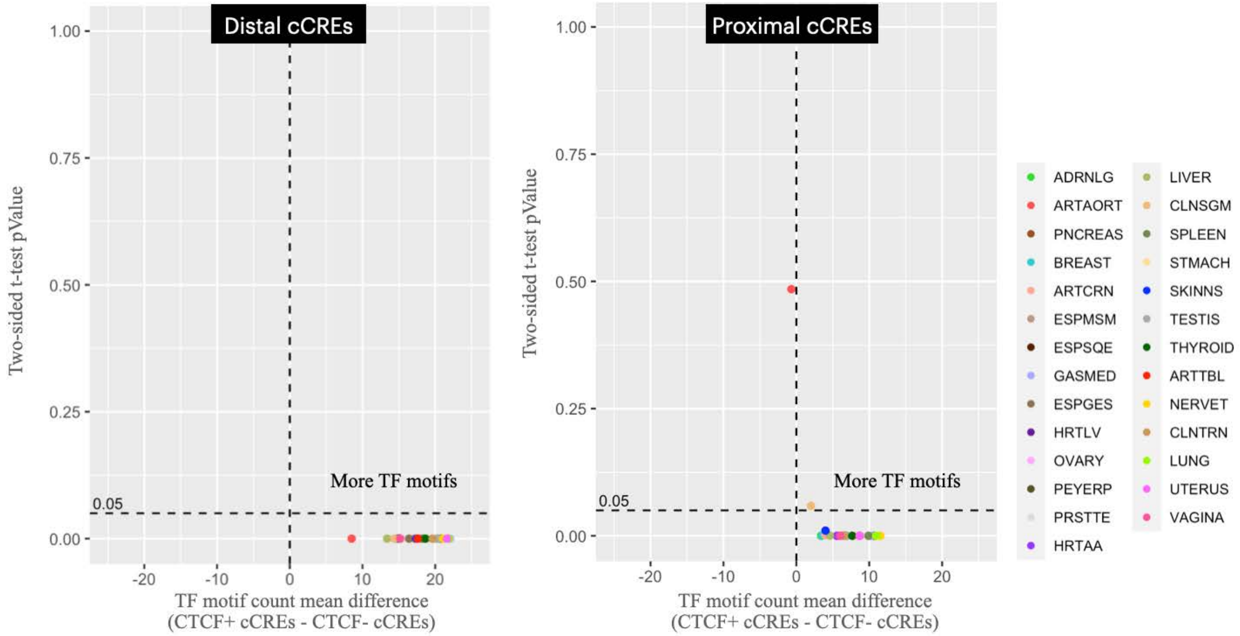
Odds ratio=1.56
P<9.1e-9

Motif+: the cCRE is overlapping any one of the **bottom 100** motifs

	Motif +	Motif -
AS cCRE	59252	56831
Non-AS cCRE	3249317	1905683

Odds ratio=0.61
P≈0

K



Data S30. Motif analysis, related to Figure 6, Figure S7, and STAR Methods "Sensitive Motifs" Section

(A) Similar to Figure 6A, among the accessible SNPs from CTCF ChIP-seq, AS SNPs occur more frequently in the key positions of the CTCF motif.

(B) For all CTCF accessible SNPs intersecting with the CTCF motif, >70% of the AS SNPs have more reads in the reference allele than the alternative allele. This number is ~60% for non-AS SNPs, indicating that the observation of AS is likely to be caused by the disruption of the motifs.

(C) The motif of transcription factor SP1. The logo is downloaded from the Cis-BP database. Among all AS SNVs that overlap with the SP1 motif, >40% occur at position 9 of the motif. For non-AS SNVs, they occur relatively randomly across all the positions of the motif.

(D) Similar to Figure 6A, for each TF motif, we made a 2-by-2 contingency table of the number of SNVs: the SNVs falling in motif or non-motif regions and SNVs being AS or non-AS. The odds ratio of the table indicates the enrichment of AS SNVs for that motif. Figure shows the result using all accessible SNVs of all ChIP-seq (black) or using accessible SNVs from H3K27ac ChIP-seq (red) only. Each dot represents one motif. The x-axis represents the 660 TF motifs in the order based on the y-axis, the AS enrichment score.

(E) Similar to Figure S7A, we examined the motif ranks in each tissue individually. Unlike the bar plot in Figure S7A, we show the approximate density of the dots in the figure, demonstrating that the top 100 TF motifs have more consistent ranking across tissues.

(F) Motif ranking based on the enrichment of any ChIP-seq AS SNPs is not correlated with the entropy of the motifs.

(G) The same ranking is negatively correlated with the GC content, determined by the number of positions where C or G is the most frequent base, divided by the motif length.

(H) Comparison of motif ranking between using raw enrichment score and model residual. By using the residual, the potential effects of the C/G content and motif entropy are removed. The top 100 TF motifs from the original ranking are more consistent with the new ranking (with a Pearson correlation 0.637, p<1.1e-12, n=100 motif ranks), than the rest (560) of the motifs (Pearson correlation 0.03, p<0.416, n=560 motif ranks). While the choice of 100 for the top motifs is somewhat arbitrary, there are three interlinked justifications for using it: (i) In this panel, the differential corrected performance between the first 100 and remaining motifs. (ii) In panel (I) and Figure 6, the enrichment score follows a linear trend at the tail but bends and rises rapidly around rank 100. (iii) In panel (E), greater cross-tissue consistency is evident for the first 100 motifs versus the remaining ones.

(I) Motif rank in each individual tissue. We examined the motif rank in each tissue (red dots) compared to the motif rank based on all aggregate tissue (black dots). In most cases, these ranks were similar to each other. The x-axis shows the 660 TF motifs ordered by the all-tissue-based rank. The y-axis shows the AS enrichment score by aggregating all experiments.

(J) Enrichment of AS sensitive or AS non-sensitive motifs in cCREs. We intersected all cCREs with the top 100 or bottom 100 motifs from the motif ranking. AS cCREs were significantly enriched with the top 100 motifs, while the non-AS cCREs contained more AS non-sensitive motifs (Fisher’s exact test).

(K) 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, n=27 tissues) more TF motifs than CTCF- cCREs.

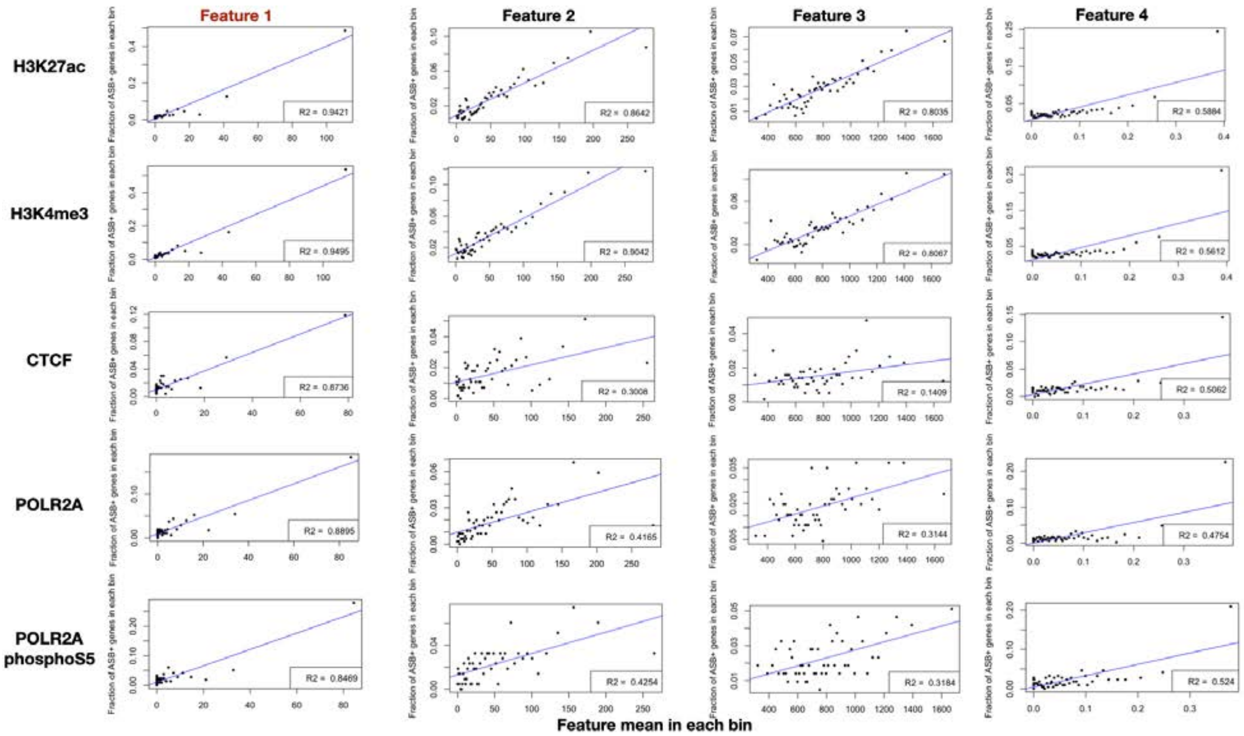
A

RF model on protein-coding genes			Positive Size	Negative Size	Accuracy	Sensitivity	Specificity	MCC	F1	Precision	ROC_AUC	PR_AUC
H3K27ac	No cross-individual	Feature 1+2+3	1224	41005	0.8080	0.7895	0.8265	0.6164	0.8043	0.8198	0.8797	0.8967
		Feature 1+2+3+4	1323	45889	0.8249	0.7926	0.8571	0.6512	0.8191	0.8474	0.8922	0.9088
	Cross-individual testing	Enc1	424	12489	0.8518	0.6527	0.8586	0.2498	0.2246	0.1357	0.8137	0.3335
		Enc2	277	10373	0.8756	0.5025	0.8855	0.1872	0.1737	0.1051	0.7574	0.1969
		Enc3	275	11629	0.8549	0.6041	0.8608	0.1958	0.1615	0.0932	0.8171	0.1888
		Enc4	351	11388	0.8404	0.7504	0.8432	0.2665	0.2197	0.1287	0.8479	0.3665
	Independent testing	STL002_3	44	1070	0.7239	0.8726	0.7177	0.2504	0.2006	0.1135	0.8829	0.2842
H3K4me3	No cross-individual	Feature 1+2+3	1607	42579	0.7731	0.7449	0.8014	0.5472	0.7665	0.7896	0.8466	0.8721
		Feature 1+2+3+4	1738	48343	0.7995	0.7594	0.8397	0.6011	0.7912	0.8257	0.8686	0.8899
	Cross-individual testing	Enc1	593	14424	0.8299	0.6428	0.8375	0.2429	0.2300	0.1401	0.7983	0.3510
		Enc2	330	10799	0.8563	0.5304	0.8662	0.1909	0.1797	0.1082	0.7703	0.1968
		Enc3	404	12503	0.8388	0.5878	0.8469	0.2031	0.1859	0.1104	0.8032	0.2126
		Enc4	416	10610	0.8216	0.7285	0.8252	0.2661	0.2357	0.1406	0.8285	0.3661
	Independent testing	STL002_3	23	51	0.6854	0.8872	0.6771	0.2317	0.1827	0.1019	0.8799	0.3082
CTCF	No cross-individual	Feature 1+2+3	384	25040	0.7058	0.7017	0.7098	0.4119	0.7044	0.7077	0.7759	0.7903
		Feature 1+2+3+4	430	27755	0.7236	0.6978	0.7494	0.4480	0.7162	0.7359	0.7897	0.8156
	Cross-individual testing	Enc1	64	6509	0.7491	0.6367	0.7502	0.0875	0.0472	0.0245	0.7661	0.0552
		Enc2	78	5907	0.8007	0.2885	0.8075	0.0276	0.0364	0.0194	0.5996	0.0207
		Enc3	152	7298	0.7364	0.5741	0.7398	0.1005	0.0817	0.0440	0.7120	0.0775
		Enc4	136	8033	0.7343	0.5804	0.7369	0.0918	0.0680	0.0361	0.6878	0.0674
POLR2A	No cross-individual	Feature 1+2+3	404	21092	0.7014	0.6884	0.7144	0.4032	0.6974	0.7070	0.7679	0.7948
		Feature 1+2+3+4	424	22436	0.7495	0.7192	0.7799	0.5002	0.7416	0.7659	0.8199	0.8445
	Cross-individual testing	Enc1	54	4630	0.7878	0.5570	0.7905	0.0906	0.0571	0.0301	0.7435	0.0760
		Enc2	55	2914	0.8208	0.3738	0.8292	0.0722	0.0719	0.0398	0.6792	0.0664
		Enc3	145	8195	0.7831	0.6674	0.7852	0.1424	0.0969	0.0523	0.7945	0.1095
		Enc4	170	6699	0.7724	0.6288	0.7760	0.1489	0.1207	0.0668	0.7554	0.1402
POLR2A phosphoS5	No cross-individual	Feature 1+2+3	235	10153	0.6611	0.6445	0.6776	0.3226	0.6552	0.6669	0.7204	0.7537
		Feature 1+2+3+4	241	10538	0.7172	0.6840	0.7505	0.4357	0.7074	0.7330	0.7856	0.8107
	Cross-individual testing	Enc1	76	3400	0.7412	0.6102	0.7441	0.1178	0.0937	0.0508	0.7208	0.1163
		Enc2	15	601	0.7818	0.5721	0.7871	0.1335	0.1136	0.0631	0.7483	0.1364
		Enc3	57	3334	0.7778	0.4835	0.7829	0.0827	0.0685	0.0369	0.6904	0.0819
		Enc4	93	3207	0.7552	0.6024	0.7597	0.1386	0.1222	0.0680	0.7307	0.1688

B

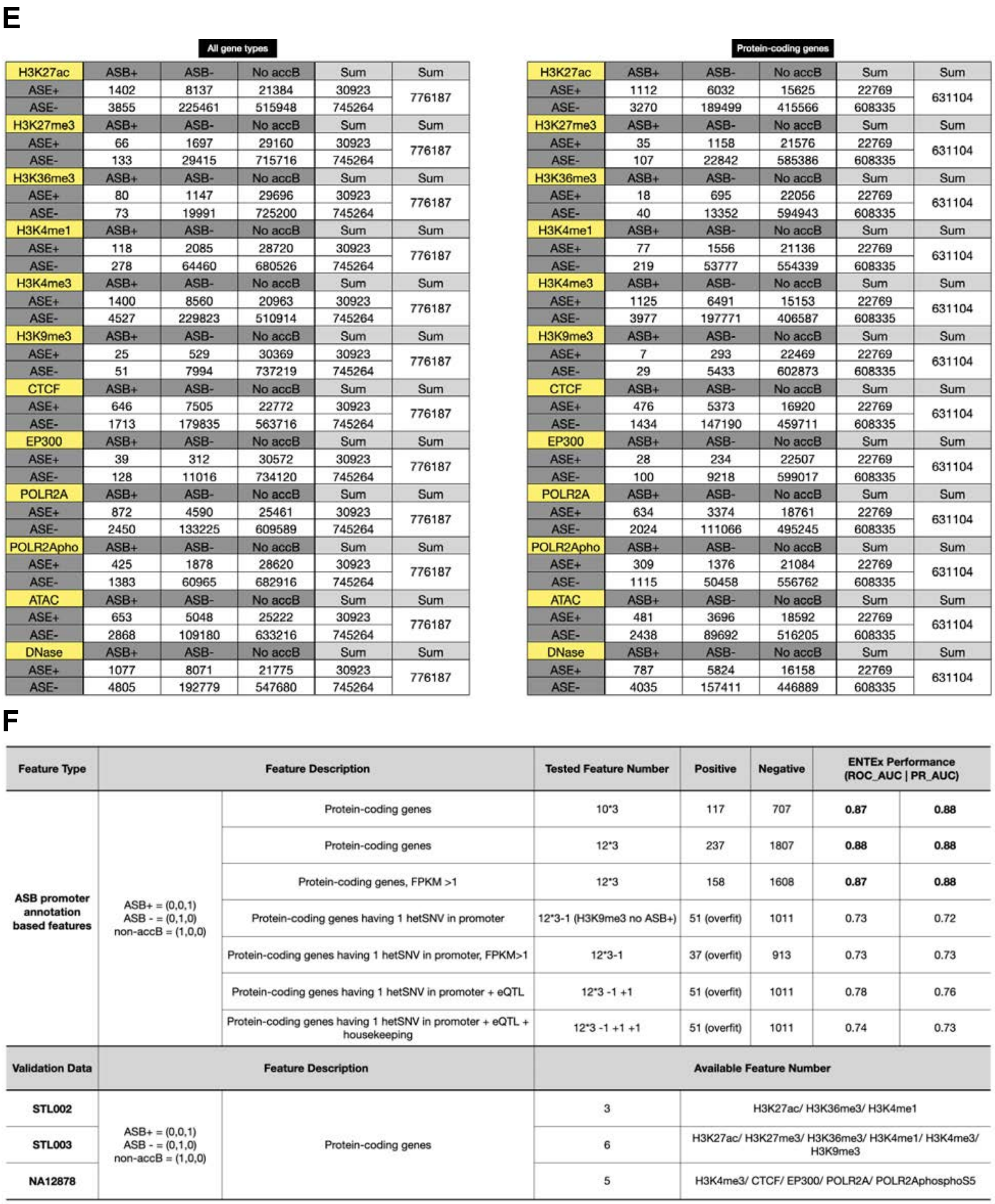
Association of features with ASB event		Positive_mean	Negative_mean	t.test p-value	R2 score	Assay	RF feature importance
Feature 1	Top100TF_motifSum_w_hetSNV	43	4	< 2.2E-16	0.9421	H3K27ac	0.3298
		42	4	< 2.2E-16	0.9495	H3K4me3	0.3127
		16	4	< 2.2E-16	0.8736	CTCF	0.1929
		24	4	< 2.2E-16	0.8895	POLR2A	0.2359
		26	4	1.156E-15	0.8469	POLR2AphosphoS5	0.2442
Feature 2	Top100TF_motifSum_nearby_hetSNV	91	49	< 2.2E-16	0.8642	H3K27ac	0.2263
		89	48	< 2.2E-16	0.9042	H3K4me3	0.2285
		63	44	1.092E-08	0.3008	CTCF	0.2500
		81	52	< 2.2E-16	0.4165	POLR2A	0.2371
		73	49	2.757E-09	0.4254	POLR2AphosphoS5	0.2347
Feature 3	motifSum_distal_to_hetSNV	910	775	< 2.2E-16	0.8035	H3K27ac	0.2129
		898	780	< 2.2E-16	0.8067	H3K4me3	0.2303
		817	769	0.0004748	0.1409	CTCF	0.2677
		833	769	5.254E-06	0.3144	POLR2A	0.2201
		854	778	0.0002361	0.3184	POLR2AphosphoS5	0.2302
Feature 4	abs(RNAseq_hap1_allele_ratio -0.5)	0.1403	0.0617	< 2.2E-16	0.5884	H3K27ac	0.2311
		0.1289	0.0633	< 2.2E-16	0.5612	H3K4me3	0.2285
		0.1461	0.0625	< 2.2E-16	0.5062	CTCF	0.2894
		0.1638	0.0631	< 2.2E-16	0.4754	POLR2A	0.3069
		0.1423	0.0610	7.64E-14	0.5420	POLR2AphosphoS5	0.2909

C



D

ASE features	P value to be ASE+ gene (Redundant to **Feature 4-** hap1 allelic ratio)
	Gene expression level in FPKM
TF motif features	All 660 TF motif sum in the promoter
	All 660 TF motif sum hit by the hetSNV in the promoter
	All 660 TF motif sum hit by accE hetSNV in the promoter
	Top 30 ranked TF motif sum hit by the hetSNV in the promoter
hetSNV features	If the hetSNV is GTEx eQTL
	eQTL effect (slope absolute value)
Individual features	The ratio of the promoter to be ASB+ in other EN-TEx individuals



E

All gene types

H3K27ac	ASB+	ASB-	No accB	Sum	Sum
ASE+	1402	8137	21384	30923	776187
ASE-	3855	225461	515948	745264	
H3K27me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	66	1697	29160	30923	776187
ASE-	133	29415	715716	745264	
H3K36me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	80	1147	29696	30923	776187
ASE-	73	19991	725200	745264	
H3K4me1	ASB+	ASB-	No accB	Sum	Sum
ASE+	118	2085	28720	30923	776187
ASE-	278	64460	680526	745264	
H3K4me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	1400	8560	20963	30923	776187
ASE-	4527	229823	510914	745264	
H3K9me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	25	529	30369	30923	776187
ASE-	51	7994	737219	745264	
CTCF	ASB+	ASB-	No accB	Sum	Sum
ASE+	646	7505	22772	30923	776187
ASE-	1713	179835	563716	745264	
EP300	ASB+	ASB-	No accB	Sum	Sum
ASE+	39	312	30572	30923	776187
ASE-	128	11016	734120	745264	
POLR2A	ASB+	ASB-	No accB	Sum	Sum
ASE+	872	4590	25461	30923	776187
ASE-	2450	133225	609589	745264	
POLR2Apho	ASB+	ASB-	No accB	Sum	Sum
ASE+	425	1878	28620	30923	776187
ASE-	1383	60965	682916	745264	
ATAC	ASB+	ASB-	No accB	Sum	Sum
ASE+	653	5048	25222	30923	776187
ASE-	2868	109180	633216	745264	
DNase	ASB+	ASB-	No accB	Sum	Sum
ASE+	1077	8071	21775	30923	776187
ASE-	4805	192779	547680	745264	

Protein-coding genes

H3K27ac	ASB+	ASB-	No accB	Sum	Sum
ASE+	1112	6032	15625	22769	631104
ASE-	3270	189499	415566	608335	
H3K27me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	35	1158	21576	22769	631104
ASE-	107	22842	585386	608335	
H3K36me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	18	695	22056	22769	631104
ASE-	40	13352	594943	608335	
H3K4me1	ASB+	ASB-	No accB	Sum	Sum
ASE+	77	1556	21136	22769	631104
ASE-	219	53777	554339	608335	
H3K4me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	1125	6491	15153	22769	631104
ASE-	3977	197771	406587	608335	
H3K9me3	ASB+	ASB-	No accB	Sum	Sum
ASE+	7	293	22469	22769	631104
ASE-	29	5433	602873	608335	
CTCF	ASB+	ASB-	No accB	Sum	Sum
ASE+	476	5373	16920	22769	631104
ASE-	1434	147190	459711	608335	
EP300	ASB+	ASB-	No accB	Sum	Sum
ASE+	28	234	22507	22769	631104
ASE-	100	9218	599017	608335	
POLR2A	ASB+	ASB-	No accB	Sum	Sum
ASE+	634	3374	18761	22769	631104
ASE-	2024	111066	495245	608335	
POLR2Apho	ASB+	ASB-	No accB	Sum	Sum
ASE+	309	1376	21084	22769	631104
ASE-	1115	50458	556762	608335	
ATAC	ASB+	ASB-	No accB	Sum	Sum
ASE+	481	3696	18592	22769	631104
ASE-	2438	89692	516205	608335	
DNase	ASB+	ASB-	No accB	Sum	Sum
ASE+	787	5824	16158	22769	631104
ASE-	4035	157411	446889	608335	

F

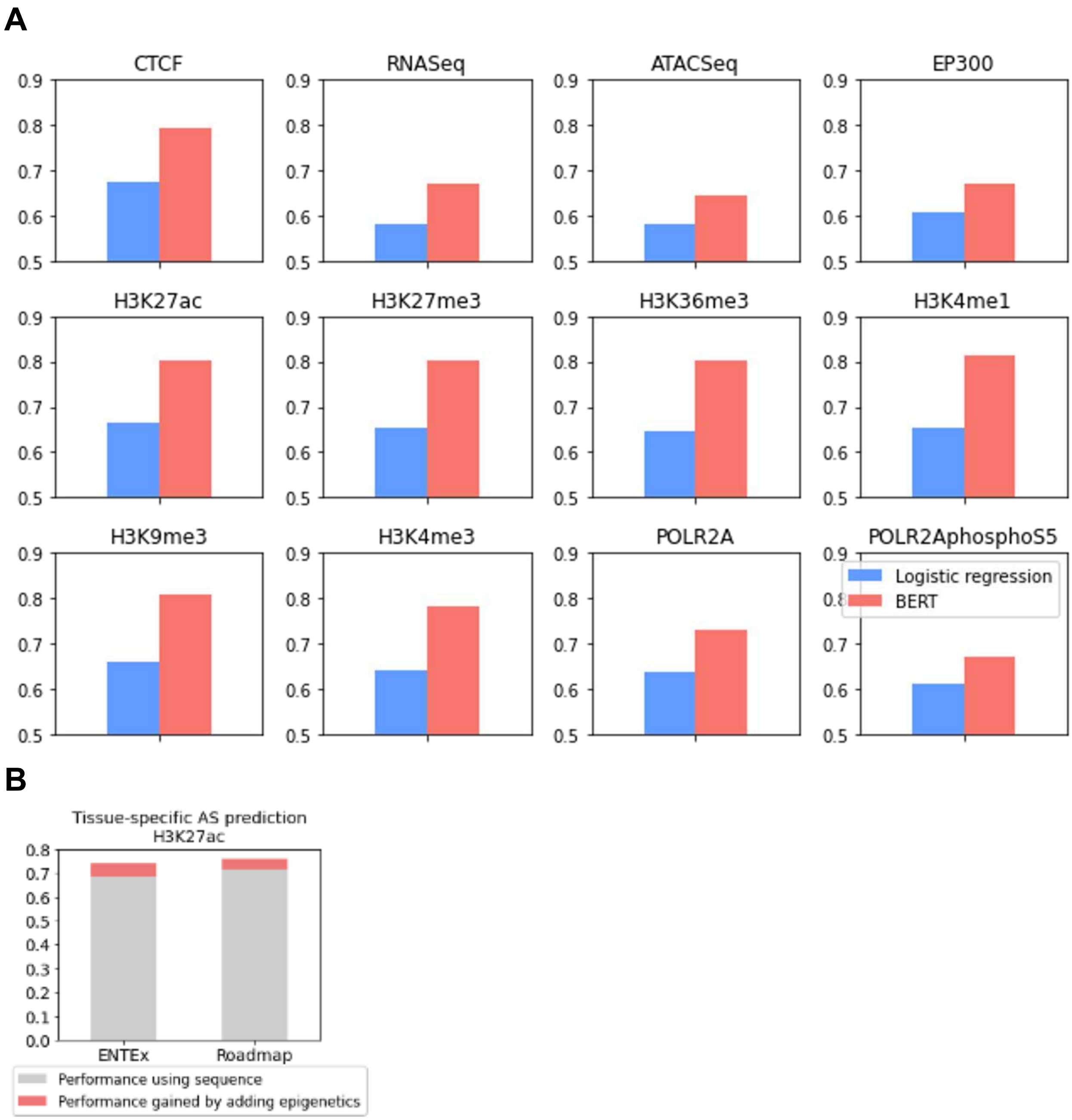
Feature Type	Feature Description		Tested Feature Number	Positive	Negative	ENTEx Performance (ROC_AUC | PR_AUC)	
ASB promoter annotation based features	ASB+ = (0,0,1) ASB - = (0,1,0) non-accB = (1,0,0)	Protein-coding genes	10*3	117	707	**0.87**	**0.88**
		Protein-coding genes	12*3	237	1807	**0.88**	**0.88**
		Protein-coding genes, FPKM >1	12*3	158	1608	**0.87**	**0.88**
		Protein-coding genes having 1 hetSNV in promoter	12*3-1 (H3K9me3 no ASB+)	51 (overfit)	1011	0.73	0.72
		Protein-coding genes having 1 hetSNV in promoter, FPKM>1	12*3-1	37 (overfit)	913	0.73	0.73
		Protein-coding genes having 1 hetSNV in promoter + eQTL	12*3 -1 +1	51 (overfit)	1011	0.78	0.76
		Protein-coding genes having 1 hetSNV in promoter + eQTL + housekeeping	12*3 -1 +1 +1	51 (overfit)	1011	0.74	0.73
Validation Data	**Feature Description**		**Available Feature Number**				
STL002	ASB+ = (0,0,1) ASB - = (0,1,0) non-accB = (1,0,0)	Protein-coding genes	3	H3K27ac/ H3K36me3/ H3K4me1			
STL003			6	H3K27ac/ H3K27me3/ H3K36me3/ H3K4me1/ H3K4me3/ H3K9me3			
NA12878			5	H3K4me3/ CTCF/ EP300/ POLR2A/ POLR2AphosphoS5			

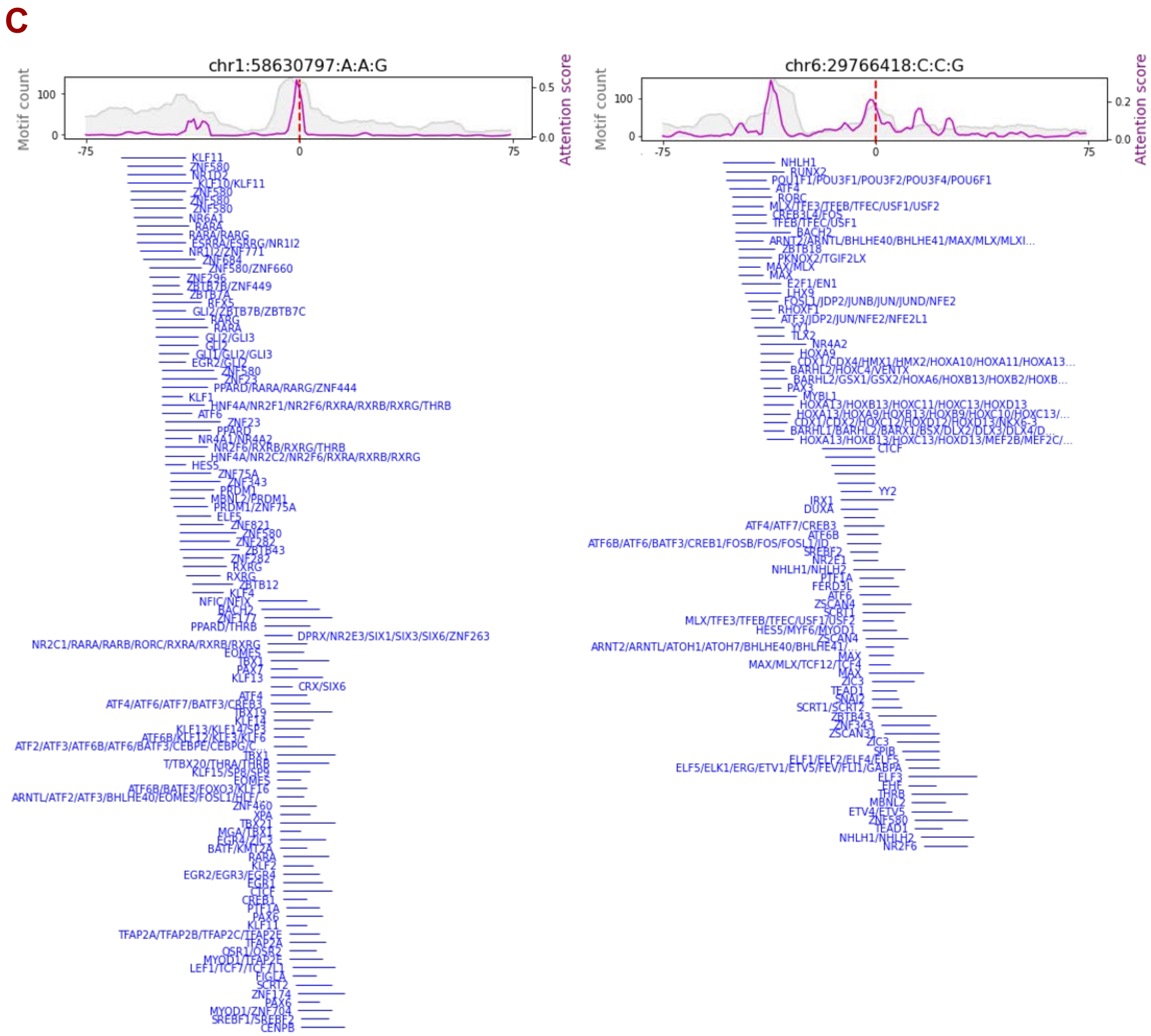
Data S31. Predicting ASB from ASE, related to Figure 6, Figure S7, and STAR Methods “AS Promoter” Section

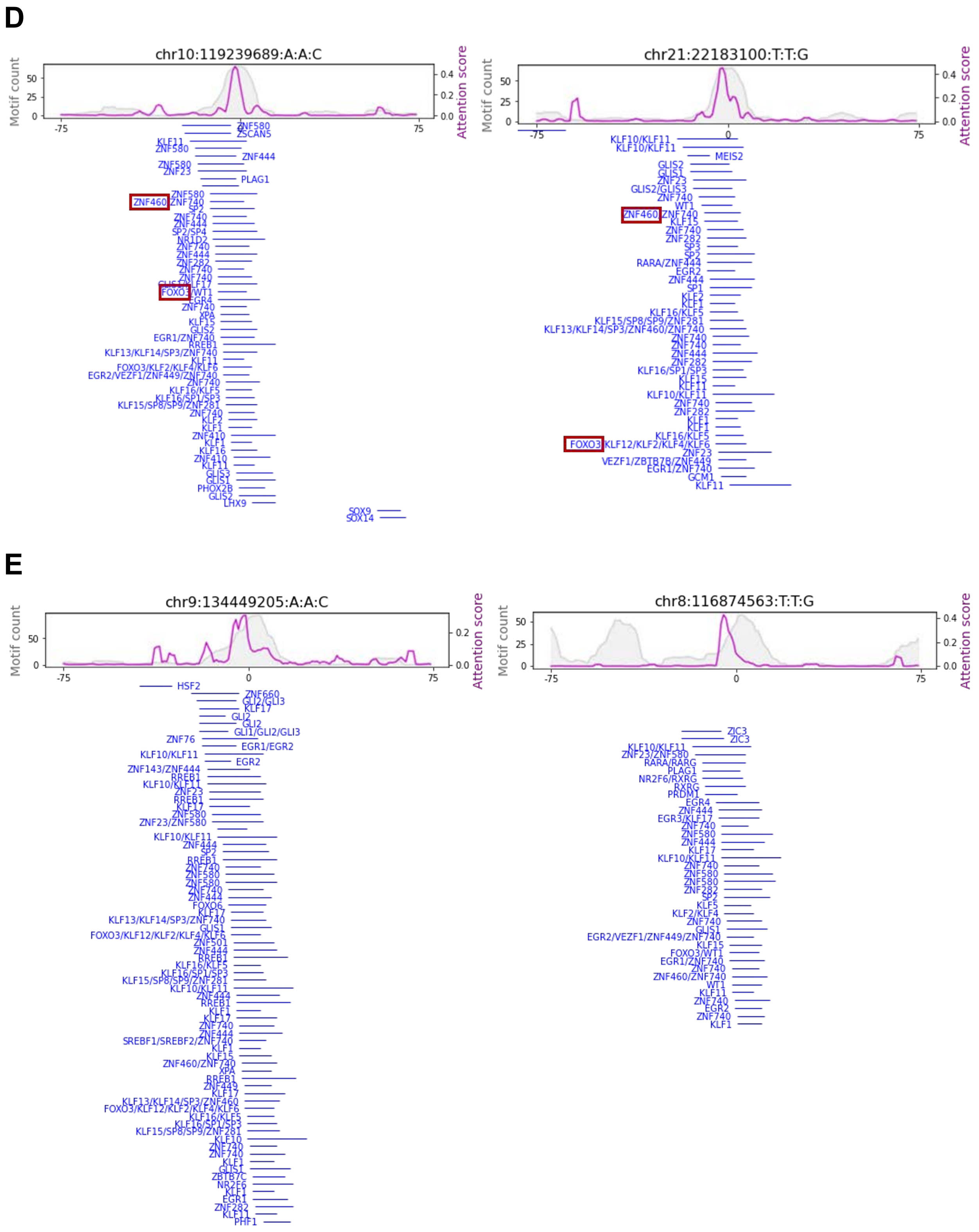
(A) Performance validation of models predicting AS bound promoters for each assay. Models were trained separately for each assay with enough training data, including H3K27ac, H3K4me3, CTCF, POLR2A, and POLR2AphosphoS5. We targeted protein-coding genes with

exactly one hetSNV in the promoter region (±1 Kb of the TSS), and the ASB states of the hetSNV and the promoter were examined for consistency. For "no cross-individual" training/testing, sub-models were trained and tested with balanced data composed of the same set of positives and different subsets of negatives. For each sub-model, a five-fold cross validation strategy was used. Different sets of features (see panel B for feature description) were tested for improved model performance. For cross-individual testing, for each of the four EN-TEx individuals, sub-models were trained with balanced data from the other three individuals and tested on the imbalanced data from the targeting individual. For independent testing, sub-models were trained with balanced EN-TEx data, including the four individuals, and tested on the imbalanced integrated data from Roadmap individuals STL002 and STL003. For all the metrics, the average performances of the sub-models are shown in the table.
(B) Association analysis between features and the promoter ASB events. To predict the ASB state of a gene promoter, a random forest model was built using four features: three TF motif-based features of the promoter region and one ASE feature of the gene. The three motif-based features are the total number of top-100-ranked TF motifs intersecting the hetSNV in the promoter, the total number of top-100-ranked TF motifs nearby (200 bp window centered on the hetSNV) but not intersecting the hetSNV in the promoter, and the total number of all 660 human TF motifs distal to the hetSNV in the promoter; the ASE feature is the imbalance ratio of gene expression between the haplotypes. Other features (including gene expression level, eQTL, all 660 non-ranked TF features in the promoter) were tested but proven not to be informative (panel D). For each assay, we investigated the association of each feature with ASB promoters. Welch's two-sample t-test (two-tailed) was performed for each feature between ASB and non-ASB promoters. To test if each feature is positively correlated with ASB promoters, an R2 score was calculated (panel C). A larger R2 value indicates a stronger association between the targeted feature and the ASB event of the promoters. Datasets were shuffled 100 times and an averaged R2 score is shown in the table. For each feature, a random forest-based feature importance score is shown, which is the average of sub-random forest models trained from the balanced EN-TEx data composed of the same set of positives and different subsets of negatives. Protein-coding genes with one hetSNV in their promoters were used to train the model.
(C) Scatter plots for the association analysis between features and the promoter ASB events. First, for each assay, the whole dataset including all ASB and non-ASB promoters was ranked by the value of the targeting feature in an ascending order. Second, the ranked dataset was split into 50 bins. For each bin, we calculated the mean value of the targeting feature (x-axis) and the ratio of ASB promoters in the bin (y-axis), followed by a scatter plot shown above. For all assays, feature 1 (total number of top-100-ranked TF motifs intersecting the hetSNV in the promoter) showed the strongest association with the ASB promoters.
(D) Features tested but not informative for the prediction of ASB promoters. To build a machine-learning model to predict ASB promoters, many other features were tested but not included in the final model since they did not improve the model performance.
(E) Contingency table for ASE genes and ASB promoters. For each assay, the number of ASE genes with ASB promoters is shown. "No accB" represents promoters that are not accessible for the assay.

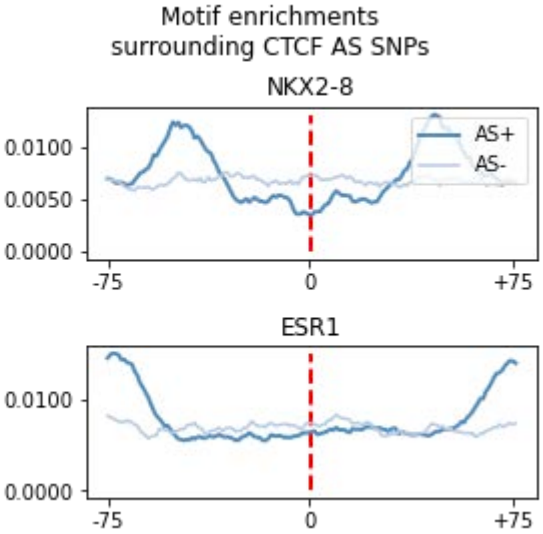
(F) Features and performance of a model to predict ASE for a gene from ASB on the associated promoter. In addition to the ‘ASB from ASE’ model, we constructed a model to predict ASE from ASB (going in the "forward" direction). The features of this model are summarized in this table; here, we used epigenetics as opposed to sequence features (as we did in the "reverse" model in panel B). As shown in the feature description, different gene sets were tested to train the model. Other features that were tested but not informative included whether the hetSNV in the promoter was eQTL or whether the gene was a housekeeping gene. Overall, the model performed well on the EN-TEx samples, but we do not have enough validation data on STL002, STL003, or NA12878 to properly evaluate the model.


A
CTCF
RNASeq
ATACSeq
EP300
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K9me3
H3K4me3
POLR2A
POLR2AphosphoS5
0.9
0.8
0.7
0.6
0.5
Logistic regression
BERT
B
Tissue-specific AS prediction
H3K27ac
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ENTEx
Roadmap
Performance using sequence
Performance gained by adding epigenetics


C
chr1:58630797:A:A:G
Motif count
100
0
-75
0
75
0.5
0.0
Attention score
KLF11
ZNF580
NR1D2
KLF10/KLF11
ZNF580
ZNF580
ZNF580
NR6A1
RARA
RARA/RARG
ESRRA/ESRRG/NR1I2
NR1I2/ZNF771
ZNF684
ZNF580/ZNF660
ZNF296
ZBTB7B/ZNF449
ZBTB7A
RFX5
GLI2/ZBTB7B/ZBTB7C
RARG
RARA
GLI2/GLI3
GLI2
GLI1/GLI2/GLI3
EGR2/GLI2
ZNF580
ZNF23
PPARD/RARA/RARG/ZNF444
KLF1
HNF4A/NR2F1/NR2F6/RXRA/RXRB/RXRG/THRB
ATF6
ZNF23
PPARD
NR4A1/NR4A2
NR2F6/RXRB/RXRG/THRB
HNF4A/NR2C2/NR2F6/RXRA/RXRB/RXRG
HES5
ZNF75A
ZNF343
PRDM1
MBNL2/PRDM1
PRDM1/ZNF75A
ELF5
ZNF821
ZNF580
ZNF282
ZBTB43
ZNF282
RXRG
RXRG
ZBTB12
KLF4
NFIC/NFIX
BACH2
ZNF177
PPARD/THRB
DPRX/NR2E3/SIX1/SIX3/SIX6/ZNF263
NR2C1/RARA/RARB/RORC/RXRA/RXRB/RXRG
EOMES
TBX1
PAX7
KLF13
CRX/SIX6
ATF4
ATF4/ATF6/ATF7/BATF3/CREB3
TBX19
KLF14
KLF13/KLF14/SP3
ATF6B/KLF12/KLF3/KLF6
ATF2/ATF3/ATF6B/ATF6/BATF3/CEBPE/CEBPG/C...
TBX1
T/TBX20/THRA/THRB
KLF15/SP8/SP9
EOMES
ATF6B/BATF3/FOXO3/KLF16
ARNTL/ATF2/ATF3/BHLHE40/EOMES/FOSL1/HLF/...
ZNF460
XPA
TBX21
MGA/TBX1
EGR4/ZIC3
BATF/KMT2A
RARA
KLF2
EGR2/EGR3/EGR4
EGR1
CTCF
CREB1
PTF1A
PAX6
KLF11
TFAP2A/TFAP2B/TFAP2C/TFAP2E
TFAP2A
OSR1/OSR2
MYOD1/TFAP2E
LEF1/TCF7/TCF7L1
FIGLA
SCRT2
ZNF174
PAX6
MYOD1/ZNF704
SREBF1/SREBF2
CENPB
chr6:29766418:C:C:G
Motif count
100
0
-75
0
75
0.2
0.0
Attention score
NHLH1
RUNX2
POU1F1/POU3F1/POU3F2/POU3F4/POU6F1
ATF4
RORC
MLX/TFE3/TFEB/TFEC/USF1/USF2
CREB3L4/FOS
TFEB/TFEC/USF1
BACH2
ARNT2/ARNTL/BHLHE40/BHLHE41/MAX/MLX/MLXI...
ZBTB18
PKNOX2/TGIF2LX
MAX/MLX
MAX
E2F1/EN1
LHX9
FOSL1/JDP2/JUNB/JUN/JUND/NFE2
RHOXF1
ATF3/JDP2/JUN/NFE2/NFE2L1
YY1
TLX2
NR4A2
HOXA9
CDX1/CDX4/HMX1/HMX2/HOXA10/HOXA11/HOXA13...
BARHL2/HOXC4/VENTX
BARHL2/GSX1/GSX2/HOXA6/HOXB13/HOXB2/HOXB...
PAX3
MYBL1
HOXA13/HOXB13/HOXC11/HOXC13/HOXD13
HOXA13/HOXA9/HOXB13/HOXB9/HOXC10/HOXC13/...
CDX1/CDX2/HOXC12/HOXD12/HOXD13/NKX6-3
BARHL1/BARHL2/BARX1/BSX/DLX2/DLX3/DLX4/D...
HOXA13/HOXB13/HOXC13/HOXD13/MEF2B/MEF2C/...
CTCF
YY2
IRX1
DUXA
ATF4/ATF7/CREB3
ATF6B
ATF6B/ATF6/BATF3/CREB1/FOSB/FOS/FOSL1/JD...
SREBF2
NR2E1
NHLH1/NHLH2
PTF1A
FERD3L
ATF6
ZSCAN4
SCRT1
MLX/TFE3/TFEB/TFEC/USF1/USF2
HES5/MYF6/MYOD1
ZSCAN4
ARNT2/ARNTL/ATOH1/ATOH7/BHLHE40/BHLHE41/...
MAX
MAX/MLX/TCF12/TCF4
MAX
ZIC3
TEAD1
SNAI2
SCRT1/SCRT2
ZBTB43
ZNF343
ZSCAN31
ZIC3
SPIB
ELF1/ELF2/ELF4/ELF5
ELF5/ELK1/ERG/ETV1/ETV5/FEV/FLI1/GABPA
ELF3
EHF
THRB
MBNL2
ETV4/ETV5
ZNF580
TEAD1
NHLH1/NHLH2
NR2F6


D
chr10:119239689:A:A:C
Motif count
Attention score
50
25
0
0.4
0.2
0.0
-75
0
75
ZNF580
ZSCAN5
KLF11
ZNF580
ZNF444
ZNF580
ZNF23
PLAG1
ZNF580
ZNF460
ZNF740
SP2
ZNF740
ZNF444
SP2/SP4
NR1D2
ZNF740
ZNF444
ZNF282
ZNF740
ZNF740
KLF17
FOXO3/WT1
EGR4
ZNF740
XPA
KLF15
GLIS2
EGR1/ZNF740
RREB1
KLF13/KLF14/SP3/ZNF740
KLF11
FOXO3/KLF2/KLF4/KLF6
EGR2/VEZF1/ZNF449/ZNF740
ZNF740
KLF16/KLF5
KLF16/SP1/SP3
KLF15/SP8/SP9/ZNF281
ZNF740
KLF2
KLF1
ZNF410
KLF1
KLF16
ZNF410
KLF11
GLIS3
GLIS1
PHOX2B
GLIS2
LHX9
SOX9
SOX14
chr21:22183100:T:T:G
Motif count
Attention score
50
25
0
0.4
0.2
0.0
-75
0
75
KLF10/KLF11
KLF10/KLF11
MEIS2
GLIS2
GLIS1
ZNF23
GLIS2/GLIS3
ZNF740
WT1
ZNF460
ZNF740
KLF15
ZNF740
ZNF282
SP3
SP2
RARA/ZNF444
EGR2
ZNF444
SP1
KLF2
KLF1
KLF16/KLF5
KLF15/SP8/SP9/ZNF281
KLF13/KLF14/SP3/ZNF460/ZNF740
ZNF740
ZNF740
ZNF444
ZNF282
KLF16/SP1/SP3
KLF15
KLF11
KLF10/KLF11
ZNF740
ZNF282
KLF1
KLF1
KLF16/KLF5
FOXO3
KLF12/KLF2/KLF4/KLF6
ZNF23
VEZF1/ZBTB7B/ZNF449
EGR1/ZNF740
GCM1
KLF11
E
chr9:134449205:A:A:C
Motif count
Attention score
50
0
0.2
0.0
-75
0
75
HSF2
ZNF660
GLI2/GLI3
KLF17
GLI2
GLI2
GLI1/GLI2/GLI3
ZNF76
EGR1/EGR2
KLF10/KLF11
EGR2
ZNF143/ZNF444
RREB1
KLF10/KLF11
ZNF23
RREB1
KLF17
ZNF580
ZNF23/ZNF580
KLF10/KLF11
ZNF444
SP2
RREB1
ZNF740
ZNF580
ZNF580
ZNF740
ZNF444
FOXO6
KLF17
KLF13/KLF14/SP3/ZNF740
GLIS1
FOXO3/KLF12/KLF2/KLF4/KLF6
ZNF501
ZNF444
RREB1
KLF16/KLF5
KLF16/SP1/SP3
KLF15/SP8/SP9/ZNF281
KLF10/KLF11
ZNF444
RREB1
KLF1
KLF17
ZNF740
ZNF444
SREBF1/SREBF2/ZNF740
KLF1
KLF15
ZNF460/ZNF740
XPA
RREB1
ZNF449
KLF17
KLF13/KLF14/SP3/ZNF460
FOXO3/KLF12/KLF2/KLF4/KLF6
KLF16/KLF5
KLF16/SP1/SP3
KLF15/SP8/SP9/ZNF281
KLF10
ZNF740
ZNF740
KLF1
GLIS1
ZBTB7C
NR2F6
KLF1
EGR1
ZNF282
KLF11
PHF1
chr8:116874563:T:T:G
Motif count
Attention score
50
25
0
0.4
0.2
0.0
-75
0
75
ZIC3
ZIC3
KLF10/KLF11
ZNF23/ZNF580
RARA/RARG
PLAG1
NR2F6/RXRG
RXRG
PRDM1
EGR4
ZNF444
EGR3/KLF17
ZNF740
ZNF580
ZNF444
KLF17
KLF10/KLF11
ZNF740
ZNF580
ZNF580
ZNF282
SP2
KLF5
KLF2/KLF4
ZNF740
GLIS1
EGR2/VEZF1/ZNF449/ZNF740
KLF15
FOXO3/WT1
EGR1/ZNF740
ZNF740
ZNF460/ZNF740
WT1
KLF11
ZNF740
EGR2
ZNF740
KLF1

F



Data S32. Deep-learning model predicting AS activity from nucleotide sequence, related to Figure 7, Figure S7, and STAR Methods “Transformer Model” Section

(A) Performance of trained allelic effect prediction models. “Logistic regression” results were derived from simple logistic regression on the dna2vec embedding of the input sequence; “BERT” results were derived from the fine-tuned DNABERT model. Both models were trained on SNPs from individual 3, and the results are reported for the validation sets from all four individuals.

(B) Tissue-specific performance (AUROC) in H3K27ac (with Roadmap external validation).

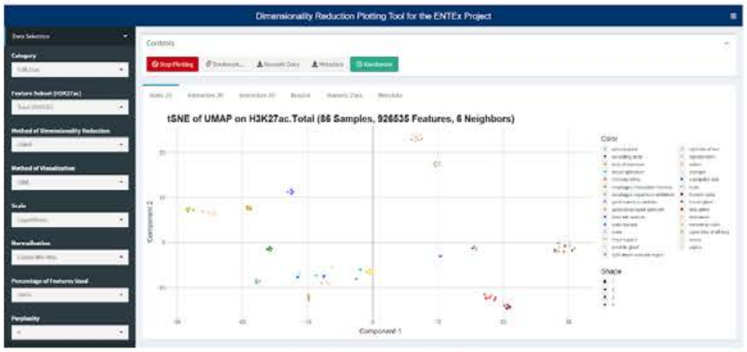
(C) Additional examples of attention patterns learned by the model (CTCF). The upper panel shows the attention pattern (red line) and motif count (gray) of the proximal region of the given SNV. The lower panel shows the motifs discovered in the region. Note the overlap between attention peaks and the enrichment of motif instances. Some motifs are shown to have generically higher enrichment at the SNV position as well (see Figure 7).

(D) Additional examples of attention patterns learned by the model (H3K4me3). FOXO3 and ZNF460 are highlighted.

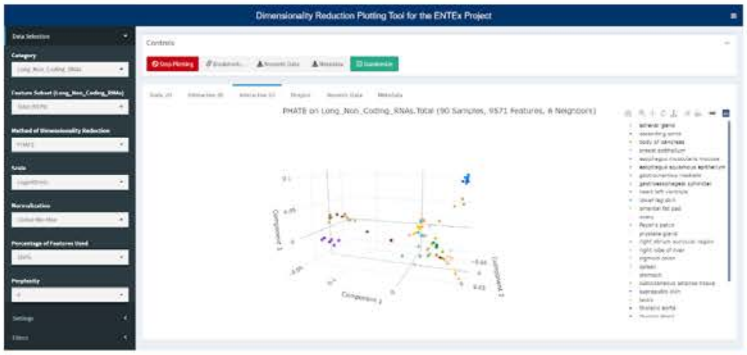
(E) Additional examples of attention patterns learned by the model (H3K9me3).

(F) Motifs that peak in the proximity of the AS CTCF SNPs.

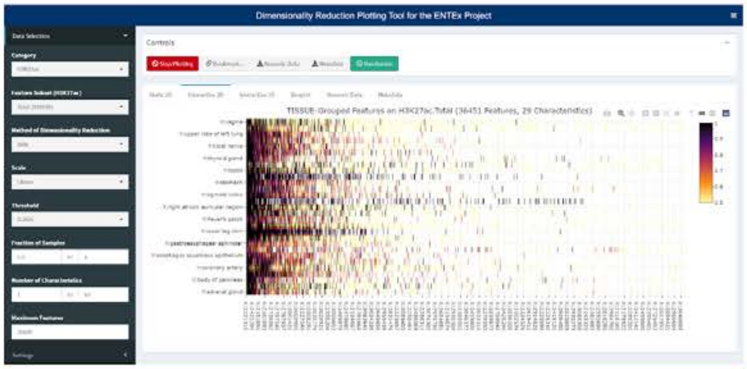
A


Dimensionality Reduction Plotting Tool for the ENTEx Project
Data Selection
Category
Feature Subset (H3K27ac)
Method of Dimensionality Reduction
Method of Visualization
Scale
Normalization
Percentage of Features Used
Perplexity
Controls
Stop Plotting
Bookmark...
Numeric Data
Metadata
Randomize
tSNE of UMAP on H3K27ac.Total (86 Samples, 926535 Features, 6 Neighbors)
Component 2
Component 1
Color
Shape

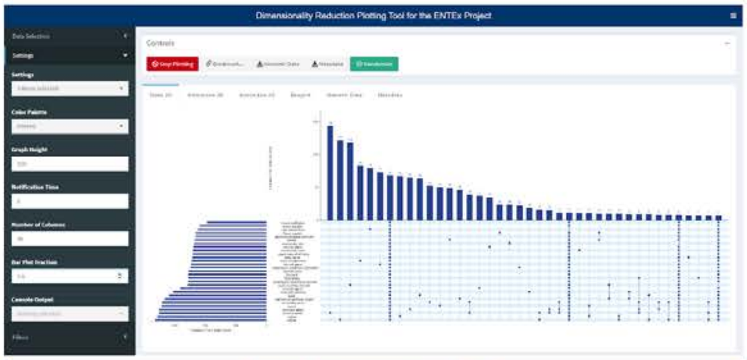
B


Dimensionality Reduction Plotting Tool for the ENTEx Project
Data Selection
Category
Feature Subset (Long_Non_Coding_RNAs)
Method of Dimensionality Reduction
Scale
Normalization
Percentage of Features Used
Perplexity
Settings
Filters
Controls
Stop Plotting
Bookmark...
Numeric Data
Metadata
Randomize
PHATE on Long_Non_Coding_RNAs.Total (90 Samples, 9571 Features, 6 Neighbors)
Component 1
Component 2
Component 3

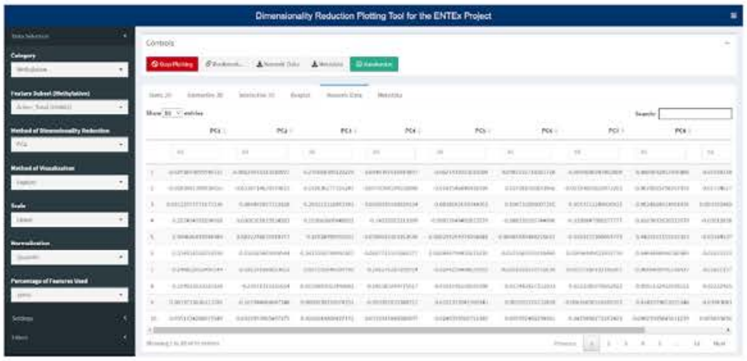
C

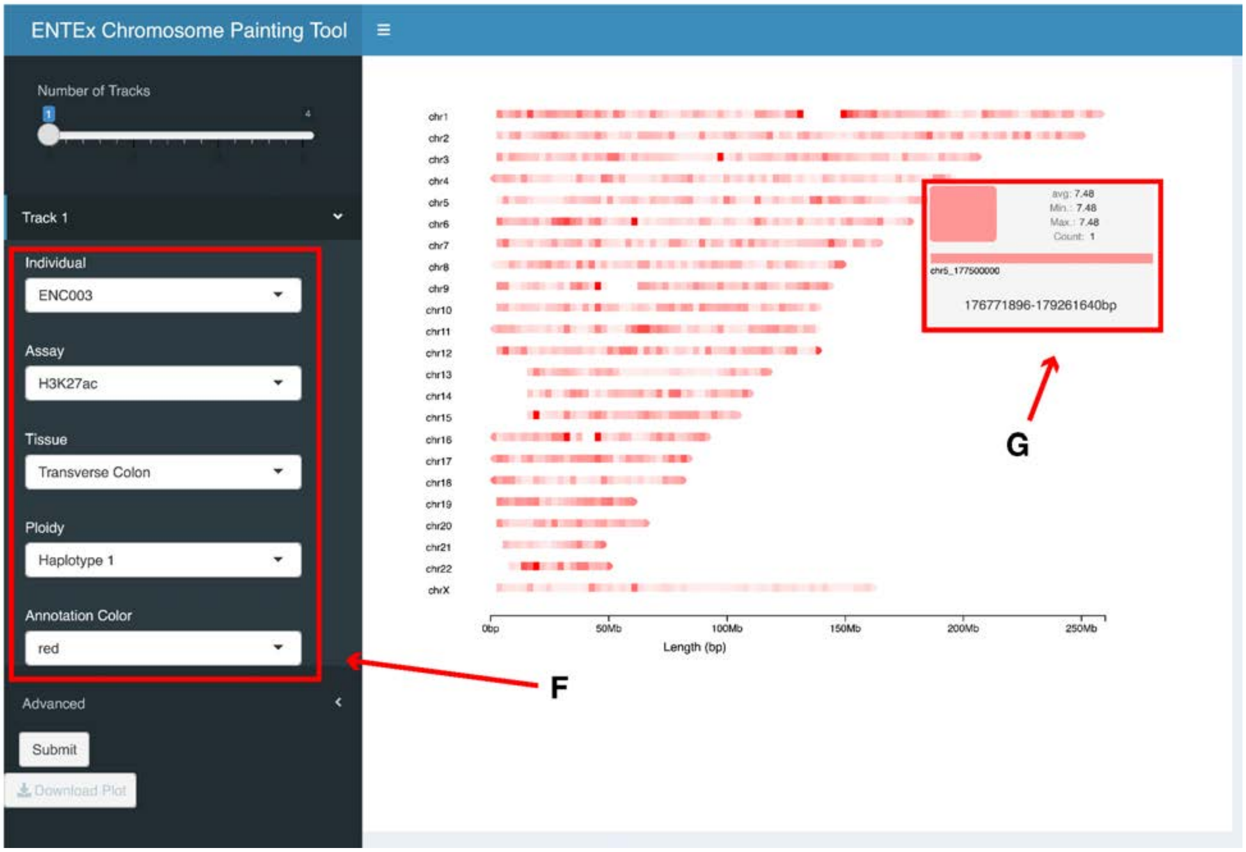

Dimensionality Reduction Plotting Tool for the ENTEx Project
Data Selection
Category
Feature Subset (H3K27ac)
Method of Dimensionality Reduction
Scale
Threshold
Fraction of Samples
Number of Characteristics
Maximum Features
Settings
Controls
Stop Plotting
Bookmark...
Numeric Data
Metadata
Randomize
TISSUE-Grouped Features on H3K27ac.Total (36451 Features, 29 Characteristics)
vagina
upper lobe of left lung
tibial nerve
thyroid gland
testis
stomach
sigmoid colon
right atrium auricular region
Peyer's patch
lower leg skin
gastroesophageal sphincter
esophagus squamous epithelium
coronary artery
body of pancreas
adrenal gland

D


Dimensionality Reduction Plotting Tool for the ENTEx Project
Data Selection
Settings
Settings
Color Palette
Graph Height
Notification Time
Number of Columns
Bar Plot Fraction
Console Output
Filters
Controls
Stop Plotting
Bookmark...
Numeric Data
Metadata
Randomize

E


Dimensionality Reduction Plotting Tool for the ENTEx Project
Data Selection
Category
Feature Subset (Methylation)
Method of Dimensionality Reduction
Method of Visualization
Scale
Normalization
Percentage of Features Used
Settings
Filters
Controls
Stop Plotting
Bookmark...
Numeric Data
Metadata
Randomize
Show 10 entries
Search:
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
Previous
1
2
3
4
5
...
Next



Individual: ENC003

Assay: H3K27ac

Tissue: Transverse Colon

Ploidy: Haplotype 1

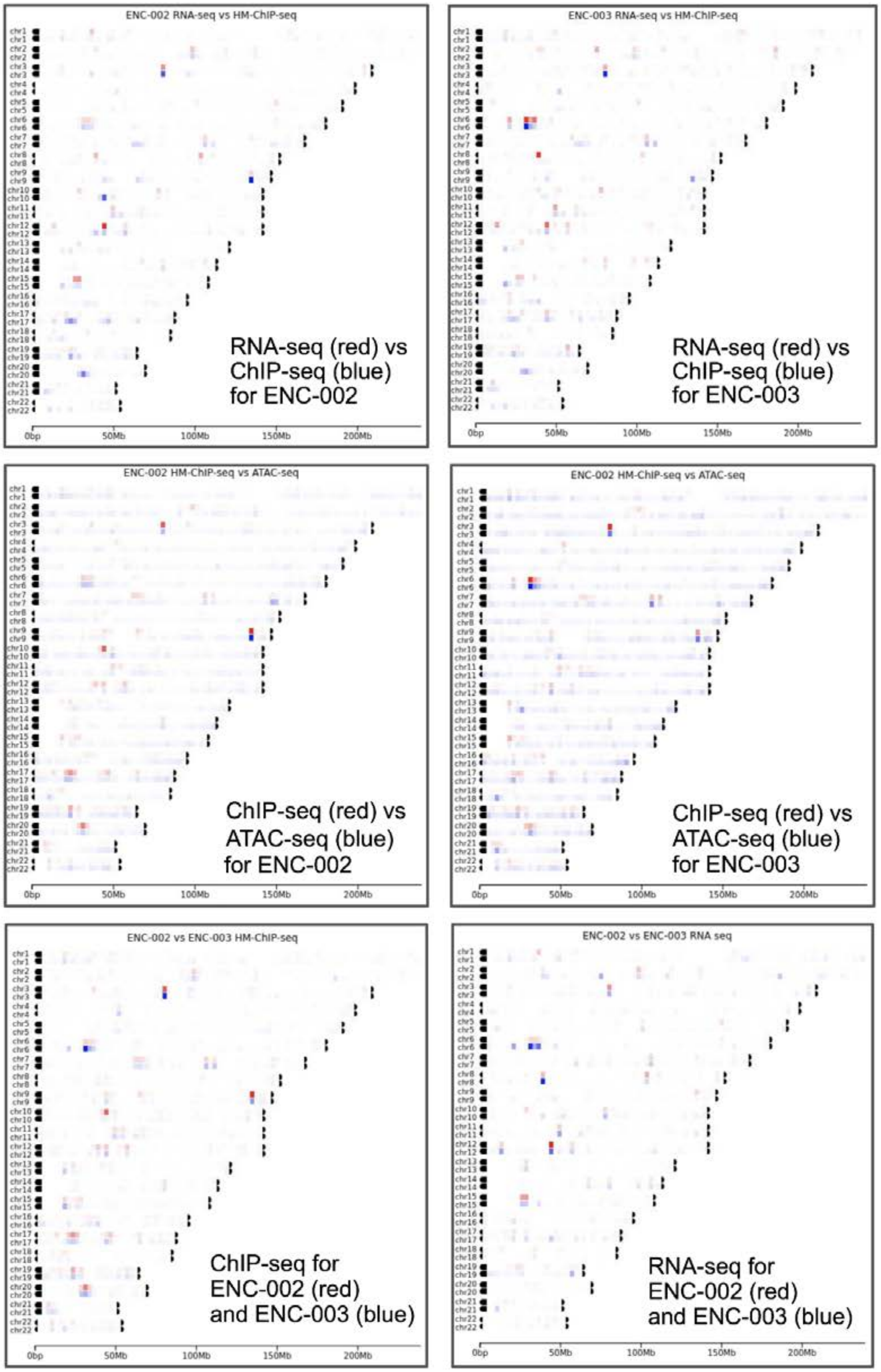
Annotation Color: red

avg: 7.48
Min.: 7.48
Max.: 7.48
Count: 1

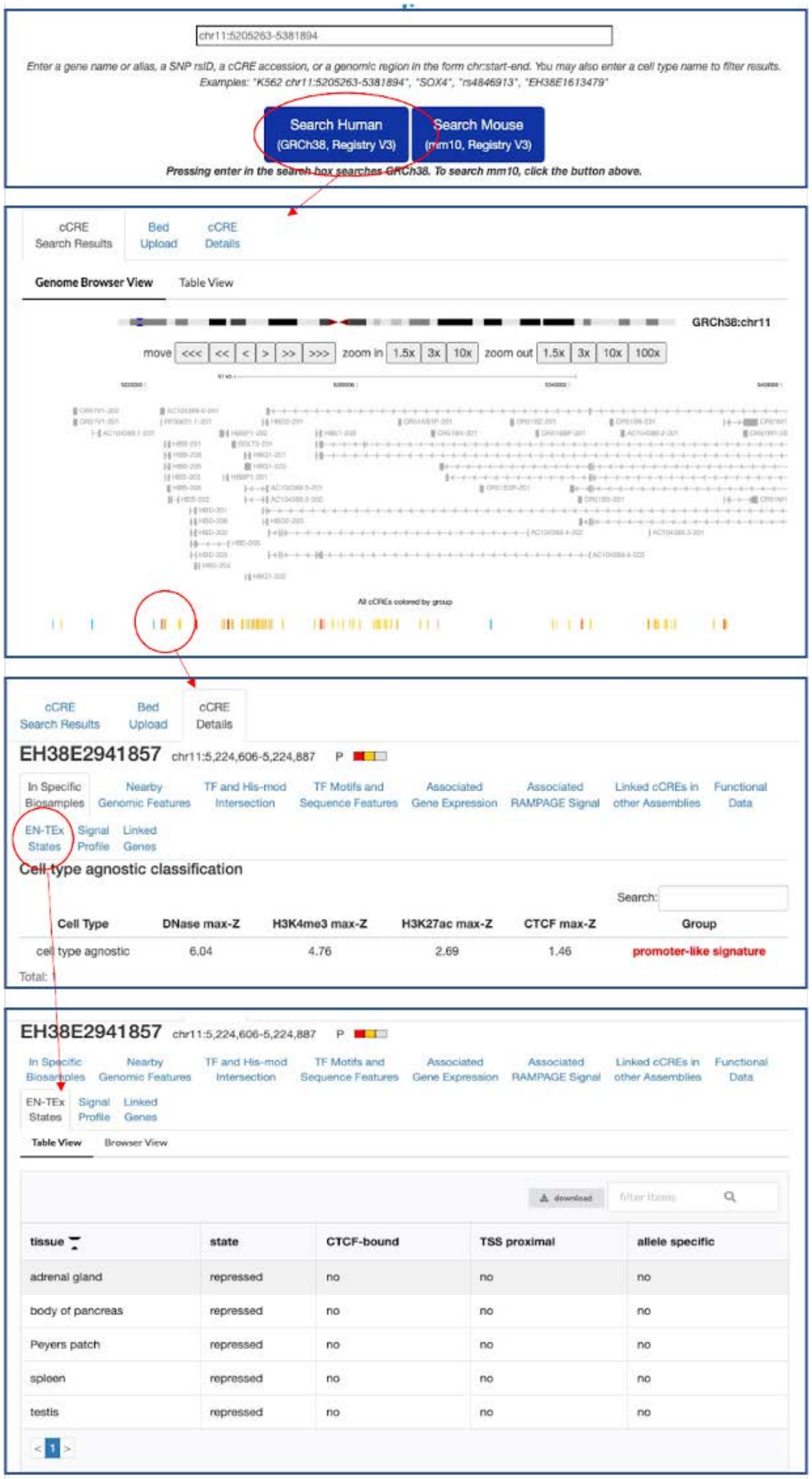
chr5_177500000

176771896-179261640bp

H


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

I



chr11:5205263-5381894

Enter a gene name or alias, a SNP rsID, a cCRE accession, or a genomic region in the form chr:start-end. You may also enter a cell type name to filter results. Examples: "K562 chr11:5205263-5381894", "SOX4", "rs4846913", "EH38E1613479"

Search Human (GRCh38, Registry V3)
Search Mouse (mm10, Registry V3)

Pressing enter in the search box searches GRCh38. To search mm10, click the button above.

EH38E2941857 chr11:5,224,606-5,224,887 P

In Specific Biosamples
Nearby Genomic Features
TF and His-mod Intersection
TF Motifs and Sequence Features
Associated Gene Expression
Associated RAMPAGE Signal
Linked cCREs in other Assemblies
Functional Data
EN-TEx States
Signal Profile
Linked Genes

Cell type agnostic classification

Search:

Cell Type	DNase max-Z	H3K4me3 max-Z	H3K27ac max-Z	CTCF max-Z	Group
cell type agnostic	6.04	4.76	2.69	1.46	promoter-like signature

Total: 1

EH38E2941857 chr11:5,224,606-5,224,887 P

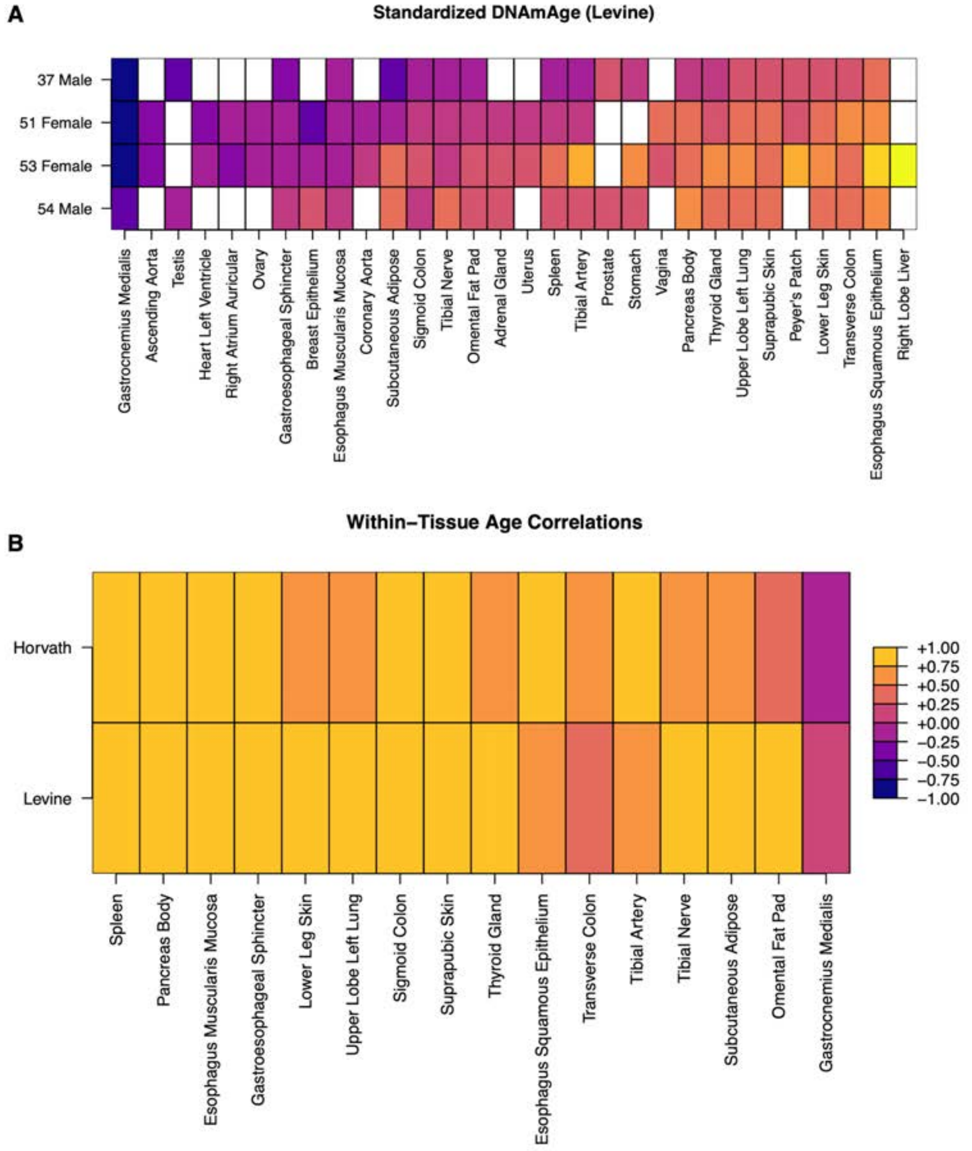
In Specific Biosamples
Nearby Genomic Features
TF and His-mod Intersection
TF Motifs and Sequence Features
Associated Gene Expression
Associated RAMPAGE Signal
Linked cCREs in other Assemblies
Functional Data
EN-TEx States
Signal Profile
Linked Genes
Table View
Browser View
download
filter items

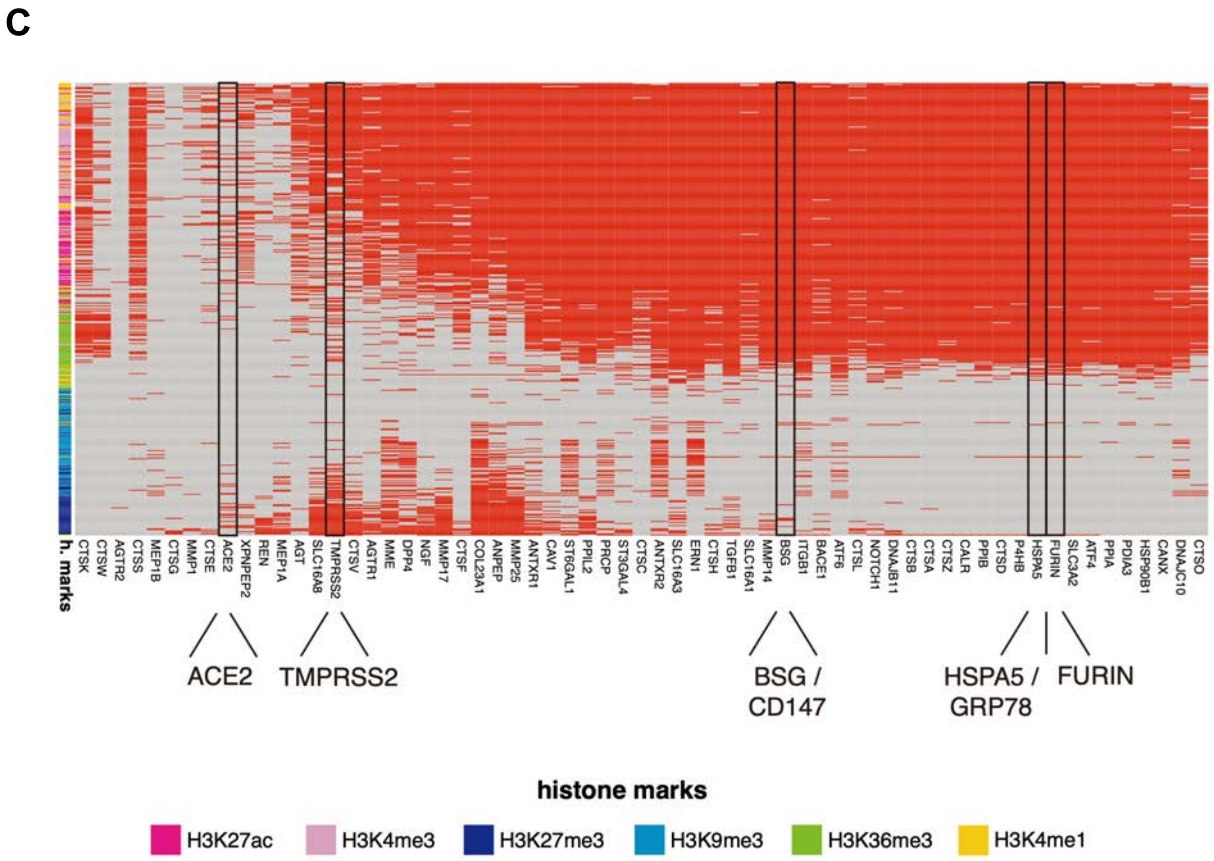
tissue	state	CTCF-bound	TSS proximal	allele specific
adrenal gland	repressed	no	no	no
body of pancreas	repressed	no	no	no
Peyers patch	repressed	no	no	no
spleen	repressed	no	no	no
testis	repressed	no	no	no

< 1 >

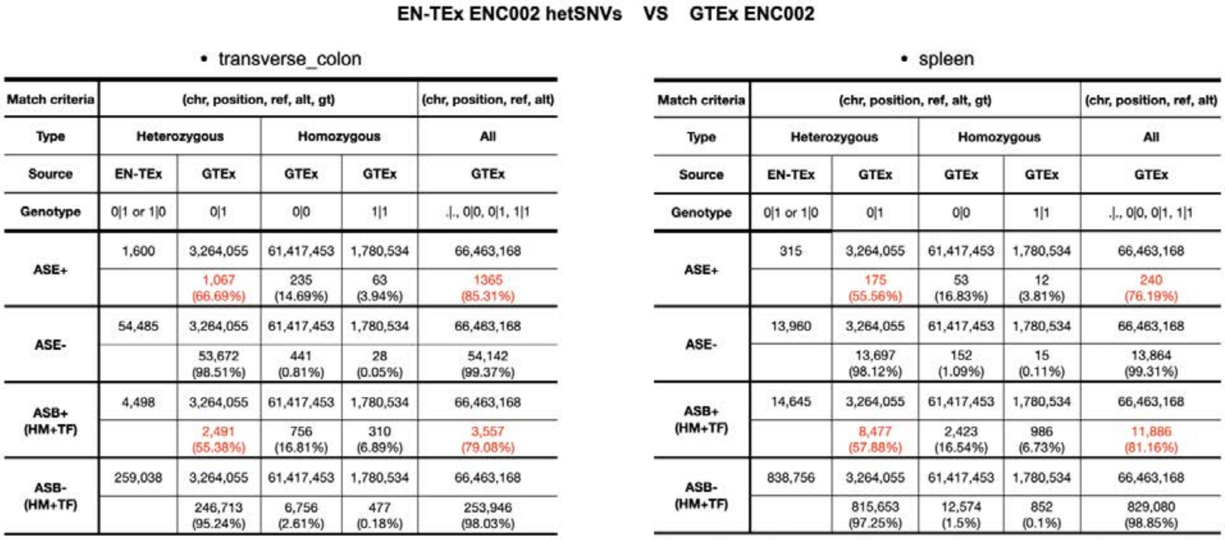
Data S33. Visualization of EN-TEx data, related to STAR Methods "Portal" Section
(A) - (E) Explorer tool. (A) Dimensionality reduction of the EN-TEx Explorer Tool allows for the generation of low-dimensional plots of several assays comprising cCREs, genomic expression,

and proteomic expression. Data are primarily reduced to ten dimensions through PCA, VAE, UMAP, or PHATE. Components of the result can be plotted against each other (e.g., principal component 1 vs. principal component 2 on a scatter plot), summarized based on the reduction method, or reduced further with t-SNE. It is also possible to rapidly view different configurations of preprocessing parameters (scaling, normalization, feature variance) or hyperparameters through extensive precomputation. (B) Interactive reduction 2D and 3D visualizations are also included for intuitively exploring the data. (C) UpSetR plots visualize the intersection of genes in various tissues, replacing the traditional Venn diagram for larger sets. In the context of EN-TEx, these tools apply user-defined thresholds for each gene, consider the fraction of samples for which a gene is present in a particular tissue, and then calculate the UpSetR plot. (D) Heatmaps, which can also have dendrograms applied, visualize the data that are aggregated in the UpSetR plot. (E) The numeric data and metadata for all results can be bookmarked or downloaded for rapid sharing or analysis. Note that the input files for the explorer tool are available from the EN-TEx portal.
(F) - (G) Screenshot of the EN-TEx chromosome painting tool. (F) Parameters for data visualization of the EN-TEx data. (G) Plots generated by the chromosome painting tool are interactive.
(H) Examples of the chromosome painting tool.
(I) Viewing EN-TEx decoration on the SCREEN website. This shows views of the SCREEN website (screen.encodeproject.org) that indicate how to access the EN-TEx decoration of the ENCODE cCREs.


A
Standardized DNAmAge (Levine)
37 Male
51 Female
53 Female
54 Male
Gastrocnemius Medialis
Ascending Aorta
Testis
Heart Left Ventricle
Right Atrium Auricular
Ovary
Gastroesophageal Sphincter
Breast Epithelium
Esophagus Muscularis Mucosa
Coronary Aorta
Subcutaneous Adipose
Sigmoid Colon
Tibial Nerve
Omental Fat Pad
Adrenal Gland
Uterus
Spleen
Tibial Artery
Prostate
Stomach
Vagina
Pancreas Body
Thyroid Gland
Upper Lobe Left Lung
Suprapubic Skin
Peyer's Patch
Lower Leg Skin
Transverse Colon
Esophagus Squamous Epithelium
Right Lobe Liver
B
Within−Tissue Age Correlations
Horvath
Levine
+1.00
+0.75
+0.50
+0.25
+0.00
−0.25
−0.50
−0.75
−1.00
Spleen
Pancreas Body
Esophagus Muscularis Mucosa
Gastroesophageal Sphincter
Lower Leg Skin
Upper Lobe Left Lung
Sigmoid Colon
Suprapubic Skin
Thyroid Gland
Esophagus Squamous Epithelium
Transverse Colon
Tibial Artery
Tibial Nerve
Subcutaneous Adipose
Omental Fat Pad
Gastrocnemius Medialis



Data S34. Additional analyses with EN-TEx data but outside of the scope of this study
(A) - (B) Predicting the ages of tissues from their DNA methylation status. The statistical model developed by Levine et al. [22] was used to predict the ages of the different tissues from the four individuals. The different tissues of the same individuals showed 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 actual aging. The high correlation is also observed using other predictive models [23]. Taken together, these results suggest that the different tissues age at different speeds.
(C) Histone ChIP-seq data for COVID19-related genes. Chromatin marking of COVID-19-related genes. The heatmap represents the presence/absence (red/gray) of 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) [24]. 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 in Sungnak et al. [25].



EN-TEx ENC002 hetSNVs VS GTEx ENC002

- transverse_colon

Match criteria	(chr, position, ref, alt, gt)				(chr, position, ref, alt)
Type	Heterozygous		Homozygous		All
Source	EN-TEx	GTEx	GTEx	GTEx	GTEx
Genotype	0|1 or 1|0	0|1	0|0	1|1	.|., 0|0, 0|1, 1|1
ASE+	1,600	3,264,055	61,417,453	1,780,534	66,463,168
		1,067 (66.69%)	235 (14.69%)	63 (3.94%)	1365 (85.31%)
ASE-	54,485	3,264,055	61,417,453	1,780,534	66,463,168
		53,672 (98.51%)	441 (0.81%)	28 (0.05%)	54,142 (99.37%)
ASB+ (HM+TF)	4,498	3,264,055	61,417,453	1,780,534	66,463,168
		2,491 (55.38%)	756 (16.81%)	310 (6.89%)	3,557 (79.08%)
ASB- (HM+TF)	259,038	3,264,055	61,417,453	1,780,534	66,463,168
		246,713 (95.24%)	6,756 (2.61%)	477 (0.18%)	253,946 (98.03%)

- spleen

Match criteria	(chr, position, ref, alt, gt)				(chr, position, ref, alt)
Type	Heterozygous		Homozygous		All
Source	EN-TEx	GTEx	GTEx	GTEx	GTEx
Genotype	0|1 or 1|0	0|1	0|0	1|1	.|., 0|0, 0|1, 1|1
ASE+	315	3,264,055	61,417,453	1,780,534	66,463,168
		175 (55.56%)	53 (16.83%)	12 (3.81%)	240 (76.19%)
ASE-	13,960	3,264,055	61,417,453	1,780,534	66,463,168
		13,697 (98.12%)	152 (1.09%)	15 (0.11%)	13,864 (99.31%)
ASB+ (HM+TF)	14,645	3,264,055	61,417,453	1,780,534	66,463,168
		8,477 (57.88%)	2,423 (16.54%)	986 (6.73%)	11,886 (81.16%)
ASB- (HM+TF)	838,756	3,264,055	61,417,453	1,780,534	66,463,168
		815,653 (97.25%)	12,574 (1.5%)	852 (0.1%)	829,080 (98.85%)

Data S35. Coverage of EN-TEx AS hetSNVs in GTEx-corresponding individual tissue, related to STAR Methods “AS Calling” Section

This comparison was performed on two tissues from individual 2. For each of the four categories of hetSNVs called in EN-TEx, the number and percentage of EN-TEx hetSNVs detected in GTEx were calculated. For ASB hetSNVs, the call sets from all available histone marks (HMs) and TFs were integrated without duplications. Note that the somatic mutations in the colon can be potentially measured by comparing with the genome sequencing done in the blood by GTEx. The slightly lower overlap for the colon active region as compared to the spleen might reflect the greater number of somatic mutations in the colon as compared to blood (from which the GTEx sequencing is derived).

References

1. Muller-Felber, W., Vill, K., Schwartz, O., Glaser, D., Nennstiel, U., Wirth, B., Burggraf, S., Roschinger, W., Becker, M., Durner, J., et al. (2020). Infants Diagnosed with Spinal Muscular Atrophy and 4 SMN2 Copies through Newborn Screening - Opportunity or Burden? J Neuromuscul Dis *7*, 109-117. 10.3233/JND-200475.
2. Son, Y.S., Choi, K., Lee, H., Kwon, O., Jung, K.B., Cho, S., Baek, J., Son, B., Kang, S.M., Kang, M., et al. (2019). A SMN2 Splicing Modifier Rescues the Disease Phenotypes in an In Vitro Human Spinal Muscular Atrophy Model. Stem Cells Dev *28*, 438-453. 10.1089/scd.2018.0181.
3. Mujahid, N., Liang, Y., Murakami, R., Choi, H.G., Dobry, A.S., Wang, J., Suita, Y., Weng, Q.Y., Allouche, J., Kemeny, L.V., et al. (2017). A UV-Independent Topical Small-Molecule Approach for Melanin Production in Human Skin. Cell Rep *19*, 2177-2184. 10.1016/j.celrep.2017.05.042.
4. Hansen, J., Snow, C., Tuttle, E., Ghoneim, D.H., Yang, C.S., Spencer, A., Gunter, S.A., Smyser, C.D., Gurnett, C.A., Shinawi, M., et al. (2015). De novo mutations in SIK1 cause a spectrum of developmental epilepsies. Am J Hum Genet *96*, 682-690. 10.1016/j.ajhg.2015.02.013.
5. Proschel, C., Hansen, J.N., Ali, A., Tuttle, E., Lacagnina, M., Buscaglia, G., Halterman, M.W., and Paciorkowski, A.R. (2017). Epilepsy-causing sequence variations in SIK1 disrupt synaptic activity response gene expression and affect neuronal morphology. Eur J Hum Genet *25*, 216-221. 10.1038/ejhg.2016.145.
6. Dianzani, U., Massaia, M., Pileri, A., Grossi, C.E., and Clement, L.T. (1986). Differential expression of ecto-5' nucleotidase activity by functionally and phenotypically distinct subpopulations of human Leu-2+/T8+ lymphocytes. J Immunol *137*, 484-489.
7. Fang, F., Hodges, E., Molaro, A., Dean, M., Hannon, G.J., and Smith, A.D. (2012). Genomic landscape of human allele-specific DNA methylation. Proc Natl Acad Sci U S A *109*, 7332-7337. 10.1073/pnas.1201310109.
8. Mugford, J.W., Starmer, J., Williams, R.L., Jr., Calabrese, J.M., Mieczkowski, P., Yee, D., and Magnuson, T. (2014). Evidence for local regulatory control of escape from imprinted X chromosome inactivation. Genetics *197*, 715-723. 10.1534/genetics.114.162800.
9. Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature *550*, 244-248. 10.1038/nature24265.
10. Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., Santoni, F., and Antonarakis, S.E. (2018). Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. Proc Natl Acad Sci U S A *115*, 13015-13020. 10.1073/pnas.1806811115.
11. Zhang, X., Hong, D., Ma, S., Ward, T., Ho, M., Pattni, R., Duren, Z., Stankov, A., Bade Shrestha, S., Hallmayer, J., et al. (2020). Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage. Proc Natl Acad Sci U S A *117*, 4864-4873. 10.1073/pnas.1910003117.
12. Zito, A., Roberts, A.L., Visconti, A., Rossi, N., Andres-Ejarque, R., Nardone, S., Moustafa, J.E.S., Falchi, M., and Small, K.S. (2021). Escape from X-inactivation in twins exhibits intra- and inter-individual variability across tissues and is heritable. bioRxiv, 2021.2010.2015.463586. 10.1101/2021.10.15.463586.
13. Werner, J.M., Ballouz, S., Hover, J., and Gillis, J. (2022). Variability of cross-tissue X-chromosome inactivation characterizes timing of human embryonic lineage specification events. Dev Cell *57*, 1995-2008 e1995. 10.1016/j.devcel.2022.07.007.
14. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019).

Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663-675 e619. 10.1016/j.cell.2018.12.019.

15. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Consortium, G.T., et al. (2017). The impact of structural variation on human gene expression. Nat Genet *49*, 692-699. 10.1038/ng.3834.
16. Reese, F., and Mortazavi, A. (2021). Swan: a library for the analysis and visualization of long-read transcriptomes. Bioinformatics *37*, 1322-1323. 10.1093/bioinformatics/btaa836.
17. Garrido-Martin, D., Palumbo, E., Guigo, R., and Breschi, A. (2018). ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS Comput Biol *14*, e1006360. 10.1371/journal.pcbi.1006360.
18. Beraldi, R., Meyerholz, D.K., Savinov, A., Kovacs, A.D., Weimer, J.M., Dykstra, J.A., Geraets, R.D., and Pearce, D.A. (2017). Genetic ataxia telangiectasia porcine model phenocopies the multisystemic features of the human disease. Biochim Biophys Acta Mol Basis Dis *1863*, 2862-2870. 10.1016/j.bbadis.2017.07.020.
19. Hounkpe, B.W., Chenou, F., de Lima, F., and De Paula, E.V. (2021). 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. 10.1093/nar/gkaa609.
20. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. 10.1126/science.aaz1776.
21. Vosa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet *53*, 1300-1310. 10.1038/s41588-021-00913-z.
22. Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) *10*, 573-591. 10.18632/aging.101414.
23. Horvath, S. (2015). Erratum to: DNA methylation age of human tissues and cell types. Genome Biol *16*, 96. 10.1186/s13059-015-0649-6.
24. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res *47*, D607-D613. 10.1093/nar/gky1131.
25. Sungnak, W., Huang, N., Becavin, C., Berg, M., Queen, R., Litvinukova, M., Talavera-Lopez, C., Maatz, H., Reichart, D., Sampaziotis, F., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med *26*, 681-687. 10.1038/s41591-020-0868-6.