2021-04-26

Multi-tissue integrative analysis of personal epigenomes [preprint]

Joel Rozowsky
Yale University

Et al.

Multi-tissue integrative analysis of personal epigenomes

Joel Rozowsky[1,2], Jorg Drenkow[3], Yucheng T Yang[1,2], Gamze Gursoy[1,2], Timur Galeev[1,2], Beatrice Borsari[4], Charles B Epstein[5], Kun Xiong[1,2], Jinrui Xu[1,2], Jiahao Gao[1,2], Keyang Yu[6], Ana Berthel[1,2], Zhanlin Chen[1,2], Fabio Navarro[1,2], Jason Liu[1,2], Maxwell S Sun[1,2], James Wright[7], Justin Chang[1,2], Christopher JF Cameron[1,2], Noam Shoresh[5], Elizabeth Gaskell[5], Jessika Adrian[8], Sergey Aganezov[9], Gabriela Balderrama-Gutierrez[10], Samridhi Banskota[5], Guillermo Barreto Corona[5], Sora Chee[11], Surya B Chhetri[12], Gabriel Conte Cortez Martins[1,2], Cassidy Danyko[3], Carrie A Davis[3], Daniel Farid[1,2], Nina P Farrell[5], Idan Gabdank[8], Yoel Gofin[6], David U Gorkin[11], Mengting Gu[1,2], Vivian Hecht[5], Benjamin C Hitz[8], Robbyn Issner[5], Melanie Kirsche[9], Xiangmeng Kong[1,2], Bonita R Lam[8], Shantao Li[1,2], Bian Li[1,2], Tianxiao Li[1,2], Xiqi Li[6], Khine Zin Lin[8], Ruibang Luo[13], Mark Mackiewicz[14], Jill E Moore[15], Jonathan Mudge[16], Nicholas Nelson[5], Chad Nusbaum[5], Ioann Popov[1,2], Henry E Pratt[15], Yunjiang Qiu[11], Srividya Ramakrishnan[9], Joe Raymond[5], Leonidas Salichos[1,2,17], Alexandra Scavelli[3], Jacob M Schreiber[18], Fritz J Sedlazeck[9,19,20], Lei Hoon See[3], Rachel M Sherman[9], Xu Shi[1,2], Minyi Shi[8], Cricket Alicia Sloan[8], J Seth Strattan[8], Zhen Tan[1,2], Forrest Y Tanaka[8], Anna Vlasova[4,21,22], Jun Wang[1,2], Jonathan Werner$_3$, Brian Williams[23], Min Xu[1,2], Chengfei Yan[1,2], Lu Yu[7], Christopher Zaleski[3], Jing Zhang[1,2,24], J Michael Cherry[8,] Eric M Mendenhall[12], William S Noble[18], Zhiping Weng[15], Morgan E Levine[1,25], Alexander Dobin[3], Barbara Wold[23], Ali Mortazavi[10], Bing Ren[11], Jesse Gillis[3], Richard M Myers[14], Michael P Snyder[8], Jyoti Choudhary[7], Aleksandar Milosavljevic[6], Michael C Schatz[9,19], Roderic Guigó[4,26], Bradley E Bernstein[5,27], Thomas R Gingeras[3], Mark Gerstein[1,2]

1 - Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
2 - Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
3 - Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
4 - Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
5 - Broad Institute of MIT and Harvard, Cambridge, MA, USA
6 - Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA
7 - Institute of Cancer Research, London, UK
8 - Department of Genetics, School of Medicine, Stanford University, Palo Alto, CA, USA
9 - Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD, USA
10 - Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, USA
11 - Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA, USA
12 - Biological Sciences, University of Alabama in Huntsville, Huntsville, AL, USA
13 - Department of Computer Science, The University of Hong Kong, Hong Kong, CHN
14 - HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
15 - Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA
16 - European Bioinformatics Institute, Cambridge, Cambridgeshire, GB
17 - Department of Biological and Chemical Sciences, New York Institute of Technology, Old Westbury, NY, USA
18 - Department of Genome Sciences, University of Washington, Seattle, WA, USA
19 - Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
20 - Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
21 - Comparative Genomics Group, Life Science Programme, Barcelona Supercomputing Centre, Barcelona, Spain
22 - Institute of Research in Biomedicine, Barcelona, Spain
23 - Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA
24 - Department of Computer Science, University of California, Irvine, CA, USA
25 - Department of Pathology, Yale University School of Medicine, New Haven, CT, USA
26 - Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
27 - Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

Abstract

Evaluating the impact of genetic variants on transcriptional regulation is a central goal in biological science that has been constrained by reliance on a single reference genome. To address this, we constructed phased, diploid genomes for four cadaveric donors (using long-read sequencing) and systematically charted noncoding regulatory elements and transcriptional activity across more than 25 tissues from these donors. Integrative analysis revealed over a million variants with allele-specific activity, coordinated, locus-scale allelic imbalances, and structural variants impacting proximal chromatin structure. We relate the personal genome analysis to the ENCODE encyclopedia, annotating allele- and tissue-specific elements that are strongly enriched for variants impacting expression and disease phenotypes. These experimental and statistical approaches, and the corresponding EN-TEx resource, provide a framework for personalized functional genomics.

The Human Genome Project assembled one representative haploid genome sequence 20 years ago (*1*). Since then, millions of individual genomes have been sequenced (*2*). Compared to the reference, a personal genome typically contains ~4.5 million variants (*3*). Understanding their functional impact is a fundamental question in biology and medicine. To this end, researchers have conducted many genome-wide association studies (GWASes) and expression quantitative trait loci (eQTL) analyses, associating genetic variants with changes in gene expression and phenotypic traits. In part(*4*)icular, the Genotype-Tissue Expression (GTEx) project performed RNA-seq on >40 human tissues from nearly 1000 individuals, allowing for the identification of >175K eQTLs (*5, 6*). In complementary fashion, the Encyclopedia of DNA Elements (ENCODE) project was initiated in 2003 to identify and annotate genomic regions (*7*). During the ensuing decades, the project utilized functional genomic techniques to chart the transcriptional and epigenomic landscapes of numerous human tissues and cell lines, producing a catalog of candidate cis-regulatory elements (cCREs) on the reference genome (*8-10*). These are widely used for predicting the impact of genetic variants (*10-13*). However, there is a lack of one-to-one correspondence between this epigenetic annotation, based on the generic reference genome, and genetic variants, which fundamentally relate to an individual's personal genome.

To overcome this limitation, we initiated the EN-TEx study (ENCODE assays applied to GTEx samples) to connect personal genomes and functional genomics. First, we built the diploid genomes for each of four individuals with long-read sequencing. Second, for each individual, we uniformly carried out a full range of functional genomic assays for 25 tissues, resulting in >1,500 datasets for histone modifications, gene expression, protein abundance, and three-dimensional genome structure. These raw data were processed in relation to each individual's personal genome, making the interpretation of genetic variants more direct.

In particular, by using an individual's diploid genome, heterozygous loci can distinguish reads that arise from each haplotype, assigning distinct molecular signals (e.g., RNA expression or TF binding) to each. The imbalance between the haplotypes can be accurately measured by taking the wild-type allele as a baseline, avoiding biological and technical biases, and if the imbalance is statistically significant, the heterozygous variant is termed allele-specific (AS). AS variants have been determined in numerous previous studies(*14-20*). (Note that only some AS variants are causal for the observed changes, such as those directly affecting TF-binding sites on one haplotype.)

Personal genomes & matched data matrix

Phasing & SVs. We sequenced the genomes of four individuals from the GTEx cohort (identified as 1 through 4), with a variety of sequencing technologies (10x Genomics linked-read, Illumina, and PacBio). After calling single-nucleotide variants (SNVs) and small insertions and deletions, we integrated the haplotype information from linked-reads and proximal ligation sequencing (Hi-C) to phase the variants (Fig. S1.1) (*21*). This step generated large blocks of phased variants across the genome, which were stitched together, forming phased personal genomes for the four individuals (Fig. 1A). We further determined the paternal/maternal origin of the phased segments by checking the AS expression levels of known imprinted loci (Fig. 1A and

Supplement). For individuals 2 and 3, we also identified 17,649 and 18,542 structural variants (SVs, greater than 50 bp; Supplement, Fig. 1B & S1.2), incorporating them into their personal genomes. We found that the SVs tended to be short (<1 Kb) and depleted in most functional regions (e.g., exons and cCREs), to be insertions, and to have typical allele-frequency spectra, all of which agree with previous findings (Fig. S1.2) (*22, 23*).

Diploid Mappings from >1500 Experiments. Next, we carried out a comprehensive set of 1635 experiments on the four individuals (i.e., ChIP-seq, ATAC-seq, Hi-C, DNase-seq, whole-genome bisulfite sequencing [WGBS], short and long-read RNA-seq, eCLIP, and labeled proteomic mass-spectrometry; Fig. 1D & S1.3a). All our datasets were processed according to both the personal diploid and reference genomes, giving rise to three mappings and signal tracks for each assay (maternal and paternal haplotypes and the reference; Fig. S1.4). When we applied strict mapping criteria (in terms of allowed mismatches) we found ~2.5% more reads mapped to the personal genomes than to the reference (Supplement). The increase was smaller in annotated regions (genes and cCREs) than in the genome overall. Still, mapping to the personal versus reference genome has an effect on gene expression quantification (e.g. resolving better the expression levels of immune-related genes; Fig. S1.4).

Measuring AS activity in diverse assays

(RNA/ChIP/ATAC/DNase)-seq. For the assays making up the bulk of the dataset, AS measurement involves the direct comparison of the number of mapped reads at a locus containing heterozygous SNVs (hetSNVs), and we report the number of significantly imbalanced hetSNVs relative to accessible hetSNVs (i.e., hetSNVs with enough sequencing depth to be able to detect statistically significant imbalances; Fig. 2A and Supplement). We performed these calculations uniformly on a large scale with a standard pipeline, making possible consistent call-set comparison, and with reads mapped to personal genomes, avoiding reference and ambiguous-mapping biases (Fig. 2 & Supplement) (*7, 8, 17, 24-27*). We also developed alternate call sets, including "high-power" ones based on joint calling across tissues (Fig. S2.1e). As shown in Fig. 2D, we consistently detected ~800 AS hetSNVs per sample, about 3% of the potential 27.5K accessible hetSNVs.

WGBS, Hi-C & Proteomics. For three assays we have had to assess AS activity in a specialized fashion. In particular, for WGBS, we accounted for base changes at potentially methylated CpG sites (Supplement and Fig. S2.2). We identified ~130K AS methylation events per sample. For Hi-C, we mapped the reads onto the personal genomes and generated haplotype-resolved contact matrices, partitioning the Hi-C contacts, where at least one of the contacting regions were AS, into AS interactions (Supplement & Fig. S2.3a). Of the average ~6.5M interactions per sample, ~500K showed significant AS behavior (Fig. S2.3). Finally, for proteomics we mapped peptides directly to the personal genomes, calling AS peptides in consistent fashion to the processing for AS RNA-seq (contrasting to other approaches (*28-30*)); in total, we found 2,028 potential AS peptides (Fig. S4.4c and Supplement).

Aggregating AS events, forming a catalog

AS Elements. In addition to determining AS activity at the SNV level, it is possible to pool the reads from multiple phased SNVs into a single genomic element, allowing the determination of AS elements (cCREs and genes, Fig. 2A and Fig. S3.1). In particular, for each individual and tissue, 182 cCREs and 351 genes showed a significant AS imbalance per assay; further aggregating across individuals resulted in ~400 AS elements per tissue (Fig. 2D). When comparing the resulting list with genes associated with specific diseases, we found sensible correlations; for example, TSHR, TG, and PAX8, which are associated with hyperthyroidism, showed AS behavior in thyroid (more examples in Supplement).

Tissue & Assay Merging. Next, we merged the 25 tissues, using a simple union of the tissue-specific AS call sets, detecting ~5.5K unique AS hetSNVs (for either binding or expression) and ~1K AS genomic elements for each individual per assay (Fig. 2C). Pooling the reads from each assay across all tissues dramatically increased (by >5X) the detection power, making it possible to identify ~27K AS hetSNVs per assay for each individual (Fig. 2D). Finally, merging across all assays provided a catalog of all loci where AS activity could be assessed in any of the tissues of the four individuals (Fig. 2D). For (RNA/ChIP/ATAC)-seq, the catalog contains 232K unique AS hetSNVs and 37K AS elements (28K cCREs and 9K genes, occurring in at least one donor and assay). The number of AS hetSNVs increases by ~2-fold (to 0.5M hetSNVs) when aggregating across tissues by pooling all the available reads. When AS sites from DNase-seq and methylation are added, the total number of hetSNVs increases to 1.3M (many in relation to previous efforts, Supplement).

Mining the catalog

Rare Variants. After constructing the catalog, we mined it for features associated with AS activity. First, consistent with previous studies (*17, 18, 26, 31*), we found that AS elements, particularly distal ones, were under less purifying selection (depleted in rare variants) than non-AS ones (Fig. 3A & S4.2). That said, a substantial number of AS variants are rare (8294 and 2961 for binding and expression, respectively; Supplement). Moreover, 14 of these were deleterious and pathogenic, based on inter-relating with ClinGen/ClinVar (Supplement) (*32*).

Model. We built a deep-learning model to predict whether a hetSNV position in an individual is AS in a particular assay based solely on the surrounding nucleotide sequence (*33*). In particular, the model was trained as a binary classifier in one individual and was then used to predict on non-shared hetSNVs in another (Supplement). As shown in Fig. 3B, the CTCF model has stronger performance than the ones for other assays (e.g. RNA-seq) and attaches higher importance to the central region surrounding the hetSNV, perhaps because of the well-defined CTCF binding motif.

Cross-assay Compatibility. We investigated the compatibility of AS activity across different assays, on a genomic scale, by assessing whether highly expressed alleles generally have more active promoters, indicated by a stronger active chromatin signal and weaker repressed one (Fig. 3C). As expected, there was substantial correlation in AS activity between H3K27ac and expression, with similar results for the other assays (e.g. expression with ATAC-seq or methylation), taking into account the opposing trend for repressive marks and methylation (Fig. 3D & S4.4a). A similar compatibility relation can also be found between AS activity and GTEx eQTLs, giving rise to annotations on the eQTLs (Fig. 3D & S4.4a for expression and binding for a given hetSNV, respectively). Fig. 3E summarizes the overall great degree of compatibility, with an associated list of strongly compatible gene-promoter pairs (Supplement). Counter to the overall trend is the low compatibility between AS mRNA expression and peptide expression, possibly reflecting a role of post-transcriptional regulation (*34, 35*).

Examples of coordinated AS activity across assays

Next, we describe specific examples of the coordination of AS activity, using diploid signal tracks.

Imprinted Locus. In most somatic cells, IGF2 and neighboring H19 are imprinted (*36*). In several EN-TEx tissues, we also observed that H19 is expressed only maternally, and IGF2, only paternally, due to AS CTCF binding at an imprinting control region (Fig. 4A) (*37*). Going beyond this, haplotype-resolved Hi-C showed that, on the maternal haplotype, a cCRE upstream of H19 interacts with this gene but not with IGF2. In contrast, on the paternal haplotype, the same cCRE interacts with IGF2 only.

Disease-associated Locus. A novel example of coordinated AS activity is found for *DNAH11, a* gene associated with Ciliary dyskinesia (OMIM #611884). We observed AS methylation in the promoter regions to be in the opposite direction of AS expression and activity of H3K4me3 and H3K27ac, consistent with transcriptional down-regulation. In fact, some of the AS hetSNVs, lying in the promoter, have been identified as eQTLs by GTEx (Fig. 4B).

Coordination over X-chromosome. On chromosome X, we observed gene expression, active histone marks, POL2R and CTCF binding skewed toward one haplotype, with repressive marks skewed to the other one (Fig. 4C & S5.1a-b). There are notable exceptions: genes on pseudoautosomal regions (e.g DHRSX) and documented “escaper” genes (e.g. KDM6A (*38*)). In addition, haplotype-specific Hi-C also manifest strong differences in AS interactions on the X-chromosome at some loci (illustration for XACT, Supplement & Fig. S5.1c).

Relating SVs to chromatin & expression

While most AS activity is assessed using hetSNVs, as in examples above, some is associated with indels and SVs (Fig. S6.1a). SVs are distributed over the diploid genome unevenly, and on a large scale, their association with chromatin can be different for different haplotypes (Fig. 5A).

At specific loci, SVs can potentially impact chromatin and gene expression in a (potentially) causal fashion.

Between Haplotypes. Fig. 5B shows an example of a heterozygous deletion removing an activating region (H3K27ac peak on one haplotype) that potentially reduces the expression of the nearby gene (ZFAND2A). As this SV overlaps a known SV eQTL (*39*), the peak removal might represent a mechanism for the QTL. Fig. 5C shows a similar example: a heterozygous deletion removing an activating region near PSCA. Here the deletion is not a known eQTL but has a similar allele frequency to that of nearby eQTL SNVs and thus might represent the causal variant associated with them. (Fig. S6.1c lists additional examples of SV-eQTL connections.)

Between Individuals. Figures 5D & 5E show the analogous situation for homozygous deletions between two individuals. The first one shows an SV removing an active region and the corresponding downregulation of a nearby lncRNA. The second shows an SV removing a likely repressive region (H3K9me3 peak) in an intron of PCCB. This SV is adjacent to a number of GTEx sQTL (splicing QTL) sites and potentially affects splicing (Fig. S6.1e). Moreover, long-read RNA-seq indicates that both individuals have novel splice isoforms near the SV location. It is notable that the EN-TEx dataset makes it possible to directly compare SVs, determined by long-read DNA sequencing, with their impact on transcript structure, determined by long-read RNA-seq.

TEs & Chromatin. Having demonstrated the potential impact of specific SVs, we next evaluated the relationship between SVs and the neighboring chromatin, genome-wide. We assessed whether chromatin significantly changes around SVs (Fig. S6.2a). We grouped SVs based on their length, genotype, type (insertion or deletion), or TE involvement (whether they involve transposable elements). The first three factors showed little or no relationship to chromatin changes (Fig. S6.2b). However, in regions of generally open chromatin, the neighborhood of TE SVs showed reduced openness compared to that of non-TE SVs (Fig. 5F & S6.2b). This agrees with findings that cells repress active chromatin to suppress the activity of TEs (*40-42*). Alternatively, it could also indicate that TEs tend to insert in regions of closed chromatin.

Decorating the ENCODE Encyclopedia

Overview. The previous sections detail the functional genomic activity of diploid genomes. In this section, we relate this activity to the haploid genome, in the form of "decorations" on the cCRE Registry, a major component of the ENCODE Encyclopedia.

In particular, we can decorate cCREs (on the reference genome) in a tissue- and allele-specific fashion (*15*). Moreover, the existing Registry provides only active annotations (*10*). Given the uniform nature of EN-TEx data across tissues and the comprehensive assays, which include repressive marks, we can add a fuller description of regulation to the encyclopedia (i.e., active, repressed, or bivalent); we can also consistently assess the activity variation of an element across tissues and individuals. Finally, our decorated elements can be used to interpret the functional variants, such as eQTL and GWAS SNVs.

Approach. Starting with a tissue-independent list of regulatory elements (0.9M cCREs in the Registry) (*10*), we used active histone marks, such as H3K27ac, to determine elements active in particular EN-TEx tissues (Fig. 6A & S7.2). Moreover, we defined a stringent subset of these based on a strong double-peak H3K27ac pattern (Supplement) (*43*). In addition, using repressive histone marks and methylation we could define tissue-specific subsets of Registry elements as repressed or bivalent. Note that in order to be included in the encyclopedia Registry an element must have been identified as active in at least one ENCODE biosample. Thus, even if an element is repressed in all EN-TEx tissues, it is still capable of being active. Alternatively, we can identify genomic regions with no active marks in all of ENCODE and only repressive ones in EN-TEx (i.e. not in the Registry at all; Fig. S7.3).

Finally, we validated our active and repressive decorations using tissue-matched Hi-C (Fig. 6B). Then, most importantly, we used our AS catalog to further decorate elements as AS.

Variation over Individuals, Tissues & Assays. For the decorated cCREs, we quantified their differential activity in different tissues of the same individuals, enabling calibrated inter-tissue, inter-individual, and inter-assay cross-comparisons. Overall, the decorated cCREs exhibit much larger differences in functional genomic activity across tissues than between the same tissues of different individuals, using joint analysis of variance (*44*) (Fig. 6C).

We also developed a regression-based approach to consistently quantify the difference (expressed as the fraction of unexplained variance) between any two types of functional activities (across cCREs or genes) between individuals, tissues and assays (Supplement & Fig. S7.5; all pairwise comparisons available in a matrix). For instance, our approach demonstrated that for spleen the H3K27ac variation across cCREs of one individual explained 86% of the variation in another. This similarity in explained variation is larger than the degree to which the variation across cCREs in spleen could explain the analogous variation in transverse colon in the same individual (65%). On average, histone marks between individuals differed by 17% in terms of unexplained variation, and between tissues, by 26% (Fig. S7.5b); these numbers were larger than the amount of RNA-seq variation unexplained (across genes). We can also compare cross-assay, finding, for instance, that H3K27ac in the spleen explains much more of the variation in H3K4me3 than H3K4me1. Finally, we can consistently expand these comparisons to correlating RNA with protein abundance, finding that it varies considerably between tissues, in line with previous studies (Fig. S7.5c) (*45*).

Measuring tissue specificity

The uniform nature of EN-TEx data is ideal for measuring tissue-specificity across the body. We used a simple approach, applicable across diverse annotations, coding and non-coding genes, cCREs, TSSs, and epigenomic profiles (Fig. 7A & S8.1, Supplement). As expected (*46, 47*), only a small percentage of protein-coding genes had activity in a single tissue (~8%; by RNA-seq or mass spectrometry); in contrast, pseudogenes and lncRNAs were more tissue-specific. Active regulatory elements generally exhibited even higher tissue specificity, while repressive

elements tended to be more ubiquitous. As expected, active distal elements were more tissue-specific than the proximal ones.

AS Elements. AS genes and cCREs were more tissue-specific than corresponding non-AS ones, across a wide variety of assays (Fig. 7A & S8.1d). That said, we identified 43 elements that are AS across all available tissues (23 cCREs and 20 genes, 14 of which are associated with housekeeping genes; Fig. S8.2c,d (*48*)). For these 43, the direction of AS imbalance was consistent across tissues (Fig. 7C-D & S8.2). This fact, and that we did not observe many loci where the imbalance direction flipped across tissues, supports our aggregation strategy for calling AS events (Fig. 2 & Supplement).

Conservation. Finally, we explored the relationship between tissue specificity and purifying selection (Fig. 7B & Supplement). As expected, ubiquitously active elements were under the strongest negative selection, and ubiquitously repressed elements were the least evolutionarily conserved. Elements that were active or repressed but tissue-specific demonstrated intermediate conservation (Fig. S8.3).

Relating Encyclopedia decorations to QTLs & GWAS loci

eQTLs/sQTLs. We analyzed the relationship of our decorated regulatory elements with eQTL and GWAS SNVs. First, we systematically estimated the enrichment of eQTL and sQTL variants in active cCREs from the matched tissue type (Fig. 8A & S9.1a). There was stronger proximal than distal enrichment, especially, as expected, for sQTLs (*49*). Also, the EN-TEx decorated cCREs exhibited somewhat stronger enrichments than matched ones from the Roadmap project, probably due to their compact size (Fig. S9.1c). Next, we compared eQTL/sQTL enrichment in AS elements to non-AS ones (Fig. 8A), finding higher enrichment in AS subsets. In particular, for distal active cCREs, the AS ones showed significantly higher enrichment across all tissue types. On average, the improvement was >2X for the subset with CTCF bound, with some tissues considerably larger. (Note, this subset has more TF-binding sites than other cCREs, suggesting greater regulatory importance [Fig. S9.3].)

In addition, we found that the differential histone modifications across EN-TEx tissues can potentially be used to predict the tissue specificity of eQTLs; in particular, the presence of an H3K27ac signal at a GTEx eQTL SNV is strongly correlated with the tissues the eQTL is active in (Fig. 8B & S9.1d).

GWAS. Similar to eQTLs, we demonstrated that AS decorations can produce significantly better GWAS enrichment for disease traits (Fig. 8C & S9.2). As a baseline, we estimated the GWAS enrichment in tissue-specific regulatory elements across diverse traits (using tag-SNP and LDSC approaches (*50*)). Several of our results recapitulated known biology (Supplement & Fig. S9.2b). We then compared the GWAS enrichment patterns in cCREs that were and were not AS, finding the AS subset had stronger enrichments. For example, in the coronary artery AS elements had higher enrichment for cardiovascular disease compared to non-AS ones (*51-54*). (We also compared this to Roadmap annotations.) Finally, for specific immune-associated

traits, we show AS cCREs achieved the better enrichment tissue specificity for spleen compared to non-AS ones (Fig. 8C & S9.2f).

Discussion

Many recent efforts have demonstrated that functional annotation of the reference genome improves the fine-mapping of disease variants (*55-58*). The unique advantage of EN-TEx is that the genetic variants and their functional annotations are determined for the same individual. This matching provides potentially direct insights into the mechanism by which the variants influence gene activity. Moreover, EN-TEx demonstrates that diploid functional genomics can be brought back to traditional haploid reference genome in a useful form, improving the detection and tissue characterization of functional genetic variants.

Resource. A key contribution of EN-TEx is the creation of a broadly useful resource, enabling additional analyses outside of the scope of this paper. For example, the methylation data can be combined with biological clocks and related to the ages of the EN-TEx individuals (Fig. S10.5). In another example, one can see the cross-tissue epigenetics of the genes associated with COVID-19 (Fig. S10.6).

All of the EN-TEx data are fully open-consented (Supplement); this consent is invaluable and is different from that in many other resources (e.g. GTEx, Roadmap, TCGA, and PsychENCODE) and has been shown to greatly improve data set utilization. In particular, all the raw sequencing data, annotations and decorations from downstream analyses, and associated tools can be freely downloaded (Supplement). These include the diploid and reference signal tracks, TAD annotations, a complete catalog of AS events, lists of tissue-specific active and repressive regulatory elements (Supplement & Fig. S1.3g). Finally, we developed two interactive applications for visualizing the resource (Supplement, Fig. S10.3 & S10.4): The first performs a variety of dimensionality reductions on the data (e.g. VAE and UMAP); the second is a chromosome “painting” tool for visualizing large-scale diploid maps of functional genomics signals.

Selection & Buffering. A central question addressed by the resource is the biological impact of genetic variants. As expected, broadly and consistently active elements (e.g. housekeeping genes and ubiquitously active promotors) tend to be under strong constraint (i.e. purifying selection) and are less likely to be AS (Fig. S9.3a). The buffering hypothesis posits a mechanism for this (*16*): redundancy in important elements may dampen the impact of variants. For instance, redundant TF binding sites in cCREs can buffer the loss of a site from a genetic variant. However, this buffering is not always perfect: the subset of cCREs that are AS are likely not fully buffered, allowing the AS variants to have stronger functional impact. This is consistent with our observed enrichment of eQTL SNPs in this subset and provides a rationalization for how our AS decorations can refine the regulatory regions responsible for driving expression changes.

Rare Variants. Taken together, the four EN-TEx individuals have millions of genetic variants. Although considered healthy, they are expected to contain many rare, deleterious, or even disease-associated mutations, especially recessive ones. These rare variants are not normally accessible to traditional QTL studies, which are best powered for common variants. In contrast, the matched functional genomics data and AS analysis in EN-TEx can provide information on rare variants and in this regard is particularly informative to precision medicine. Moreover, in the future, the approach piloted by EN-TEx could be readily scaled up to more individuals, providing information on more rare variants -- in contrast to the situation for common variants, where a scale-up would provide diminishing amounts of information on new variants.

Overall, ENTEX provides comprehensive tissue-specific maps of molecular phenotypes with matched genotypes. These data are valuable for developing and testing models of genotype-phenotype relationships, which is fundamental to population genetics and precision medicine. Moreover, the EN-TEx approach can be readily extended to model organisms (*59-68*).

Figure Captions

Fig. 1. Personal genome & data matrix.
(A) Illustration of the personal genome in contrast to the reference genome. The personal diploid genome of individual 3 is shown at the top of the panel. Its construction explicitly considers SNVs, indels, and SVs. The chromosomes are phased with long-read sequencing to form phased blocks. With known imprinting events (yellow) in the human genome, the maternal (blue) or paternal (red) origin of many of the phased haplotype blocks can be identified (see Supplement). Taking chromosome 13 (chr13) as an example, it contains two phased blocks, one of which has its maternal and paternal haplotypes identified based on an imprinted gene's AS expression patterns. A schematic diagram of a genomic region in chr13 shows the detailed differences between the personal diploid genome and the reference haploid genome. Because of the genetic variants in the personal genome, the diploid representation and the reference have different coordinate systems.
(B) Mapping reads to personal and reference genomes. Due to the genetic variants in the diploid personal genome, some reads can be only mapped to this but not to the reference genome. Such reads (yellow) are referred to as “personal genome unique”. Compared to the reference genome, using the personal genome results in ~2.5% more mapped reads for various functional genomic experiments. This fraction ranges from ~1% for RNA-seq to ~4% for Hi-C data (more details in the Supplement and Fig. S1.4).
(C) Classification of SVs detected in individual 3. The SVs are classified according to their type (e.g., insertion vs. deletion) and mechanism (e.g. transposable elements or not).
(D) Data matrix for the EN-TEx resource. Each voxel in the cube stack corresponds to a functional genomic assay for a particular tissue from one of the four individuals. A dark grey voxel indicates that the functional genomic data have been generated. The total experimental counts per assay across the tissues and individuals are shown above the cube stack. The total experimental counts per tissue type are shown on the left. The color scheme of the tissues is adopted from GTEx (*6*).

Fig. 2. Measuring AS Activity, Aggregating AS events & Forming a Catalog

(A) Schematic illustrating the measurement of AS events at heterozygous SNV (hetSNV) loci and in genomic elements (e.g., cCREs). The two haplotypes for a genomic region are labeled as Hap1 and Hap2. In this region, there are three hetSNVs; thus, the reads overlapping these SNVs can be unambiguously mapped to either Hap1 or Hap2. For each hetSNV locus, we can test the resulting imbalance between the reads for Hap1 or Hap2. As shown, the hetSNV to the right lying outside of the cCRE does not have a significant imbalance (equal number of reads from Hap1 and Hap2). In contrast, the two hetSNV loci (to the left) in the cCRE have significant imbalances, i.e., 13 reads from Hap1 vs. 1 from Hap2 and 14 vs. 2, and thus are AS. Moreover, for the cCRE element as a whole, we can aggregate the reads from these two loci (i.e., 27:3 in this example). The imbalance test using this new ratio indicates whether the element as a whole is AS (see the Supplement for technical details). A standard pipeline is used for the primary assays such as RNA/ChIP/DNase/ATAC-seq, maximizing the consistency of the results; moreover, the pipeline is configured to reduce ambiguous read mapping bias (see Supplement

and Fig. S1.1) (*17, 69*). Specialized pipelines are used for methylation, proteomics, and Hi-C (see Supplement and Fig. S2.2 & S2.3).

(B) Frequencies of the AS loci in the personal genome. With panel A procedures, the AS hetSNV loci and AS cCREs are identified for the different functional genomic data from each tissue of the four individuals (using the GTEx coloring for the tissues and the shapes to indicate the different individuals, from Fig. 1). Taking H3K27ac data from the spleen from individual 1 as an example, there are 2.6k AS hetSNV loci and 0.7k AS cCREs. Most of the assay types have similar numbers of AS loci across EN-TEx tissues and individuals, as indicated by the box plots.

(C) Aggregating different tissues for AS detection. Here hetSNVs are indicated in red if they are significantly imbalanced (i.e. AS) and gray, if not. One approach to aggregation is to take the union of all AS hetSNVs (i.e., significantly imbalanced hetSNVs) from different tissues to identify unique AS loci in the genome. The hetSNVs that are AS in any tissue are considered as AS, and this rule can also be applied for detecting AS elements. This approach is referred to as the "union approach." Another approach is "pooling" the haplotype-specific reads at an hetSNV locus (regardless of its being AS or not in particular tissue) across the tissues to conduct the imbalance test. Such reads may not be significantly imbalanced (and thus not AS [gray in the figure]) in any given tissue sample, but pooling across many samples may increase the statistical power to observe an AS event at the locus (indicated in red).

(D) Using the two aggregation approaches to construct the AS catalog. (1) As a reference, H3K27ac from the spleen in individual 1 is re-shown (from panel B) using the direct ascertainment shown in panel A. (2) For H2K27ac, aggregating the spleens from the four individuals (union approach in panel C) results in 9.3K AS hetSNVs and 2.5K AS elements. (3) For H3K27ac, aggregating the different tissues of the same individual with the union approach results in 11K AS hetSNVs. In contrast, using the pooling approach to aggregate the different tissues of the same individual leads to 27K AS hetSNVs. (4) For H3K27ac, aggregating different tissues within the same individual and the tissues across all four individuals simultaneously with the union approach results in 30K AS hetSNVs. Using the pooling approach for tissues and the union approach for individuals together increases these numbers to 79K. These four different strategies focus on either spleen and/or H3K27ac. The average numbers of AS loci detected for other tissues and assays are reported in the panel's right column. For each number in this column, the red-green-blue operation icon indicates whether we are doing a union, pooling, or averaging over individuals, tissues and assays. Finally, at the bottom of the right column, we aggregate all the available assays and tissues from all the individuals to determine a final set of AS hetSNVs, either by union or pooling (more detail in the Supplement). Note that the AS loci reported in this figure are identified using RNA/ChIP/ATAC-seq.

Fig. 3. Mining the catalog.

(A) Analysis of purifying selection on AS loci. The horizontal axis measures purifying selection in terms of the fraction of rare variants, relative to the total number of variants, in a genomic region (see Supplement for details).

(B) The performance of a model predicting AS events for three assays: CTCF, RNA-seq, and POLR2A ChIP-seq. The model is trained on AS SNVs from individual 3 and then tested on non-overlapping SNVs in other individuals. The left panel shows the accuracy in terms of AUROC score. The right panel shows the average attention score of the model for the three assays. The score is averaged over 2,000 randomly selected test samples. The attention score reflects the

weights that the model attaches to different nucleotide positions in the input sequences. As expected, the score peaks around the center at the hetSNV for CTCF. Note how more attention at the center is given to CTCF (strongest motif), then ATAC-seq and then RNA-seq (which has no focusing motif).
(C) Determining compatibility between AS expression and AS promoter state. Note how the compatible expression direction changes for repressive marks.
(D) Relating AS expression to the AS H3K27ac modification and eQTL effect. On left, for AS genes with promoters accessible to H3K27ac, the fraction of H3K27ac ChIP-seq reads in the promoter mapped to Hap1 are plotted against the fraction of Hap1 RNA-seq reads. The exact read counts giving rise to these fractions for two genes (LAMTOR1 and RPL4P4) are indicated in panel C. For the eQTL effect, the slope (beta coefficient) of the leading eQTL associated with an AS gene (*6*) is correlated with the fraction of RNA-seq reads mapped to the alternative allele on that gene (overall Pearson's correlation coefficient = 0.6, p = 0.01). Note that genetic variants with slopes around 0 are unlikely to have statistical significance to be identified as eQTLs. This also holds for hetSNVs with read fractions of approximately 0.5.
(E) Compatibility between AS expression and promoter AS chromatin activity. The compatibility is measured by the fraction of genes for which AS expression is compatible with promoter AS chromatin state or AS peptide expression (see Supplement and Fig. S4.4a). We randomly paired genes with promoters (and peptides) to calculate a z-score (more details in the Supplement). Compatibility between AS expression and AS methylation (meCpG), H3K9me3, and H3K27me3 is weak, potentially because these marks of repressed chromatin can also be associated with genes poised for transcription or genes that are actively transcribed (*70, 71*).

Fig. 4. Examples of coordinated AS activity across assays.
(A) AS events detected at a known imprinted locus: H19/IGF2. Functional genomic signals are measured in the gastrocnemius medialis of individual 2. In agreement with the known imprinting mechanism (top), we detected AS expression of H19 on haplotype 2 (maternally expressed) and expression of IGF2 on haplotype 1 (paternally expressed) along with CTCF in the imprinting control region of haplotype 2. Consistent with the expression data, we found an AS Hi-C interaction between an upstream cCRE and H19 on haplotype 2 (the bold blue arc), and an AS interaction between the same cCRE and IGF2 on haplotype 1 (the bold red arc). Other neighboring AS Hi-C interactions (which have fewer read counts or connect near but not onto the relevant genes) are also depicted with shaded arcs.

(B) AS events detected at an uncharacterized, disease-associated locus: the DNAH11 gene associated with ciliary dyskinesia. The polarity of the AS DNA methylation in the promoter region is in the opposite direction to that of the AS expression and chromatin state in the gene body, consistent with transcriptional downregulation in cis and the repressive nature of regulatory DNA methylation. Also, the active epigenetic marks H3K4me3 and H3K27ac demonstrate consistent AS imbalances. Moreover, most of the AS hetSNVs associated with DNAH11 are known eQTLs from GTEx. One such hetSNV (rs11760336) lies within the DNAH11 promoter, consistent with its effect on gene expression. Also indicated is that some of the AS hetSNVs overlap with known GWAS variants.
(C) Coordinated AS activity across a chromosome. Haplotype 1 of ChrX is inactivated in individual 3. The signal tracks show that, in the tibial nerve, Hap1 ChrX generally has lower expression levels, lower H3K27ac levels, and higher H3K27me3 levels than Hap2 ChrX. In

contrast, the histograms on the left show that the autosomes display balanced gene expression and histone modification levels between the two haplotypes. The top inset bar graphs show the expression of five example genes. DHRSX, located in the pseudoautosomal region (pink bars at the ChrX ends of the signal track), and KDM6A, known to escape ChrX-inactivation, show balanced expression between haplotypes. In contrast, TBL1X and SLC25AC fall in the inactivated region of ChrX, showing lower Hap1 expression. Expression of XIST is known to induce ChrX inactivation and is, therefore, higher in haplotype 1.

Fig. 5. Relating SVs to chromatin & expression.
(A) The chromosomal distribution of SVs in individual 3 is shown as SV density. The zoom-in views add (to the SV density) the average H3K27ac level and expression level of RNA. Two examples are shown in panels B and C.
(B) A 2.6-kb deletion in Hap2 reduces ZFAND2A expression in the thyroid, consistent with previous work (*39*). This deletion removes several H3K27ac peaks from Hap2 (more detail in Fig. S6.1b).
(C) A 98-bp deletion in Hap2 removes an H3K27ac peak downstream of PSCA, potentially contributing to the lower gene expression in Hap2. This is in the transverse colon of individual 3. The heights of the green bars indicate the allele frequencies of the deletion and the surrounding GTEx eQTLs in Hap2. The frequencies are similar, suggesting the SV is potentially in linkage disequilibrium with the eQTL variants. (Note, the height of a green bar plus its corresponding magenta bar equals 1.) Similar results are observed in two other tissues (see Fig. S6.1c).
(D) A 2.3-kb homozygous deletion in individual 3. This SV removes an H3K27ac peak downstream of lncRNA RP11-362F19.1, which has lower expression in this individual. The nearby CTCF peaks suggest a potential interaction between the H3K27ac peaks downstream and upstream of the lncRNA and may indicate a mechanism by which the loss of the H3K27ac peak reduces the gene expression. See Fig. S6.1d for more detail.
(E) A 5.2-kb homozygous deletion in individual 2. This SV removes an H3K9me3 peak, potentially increasing PCCB expression in individual 2. The sashimi plots show examples of novel splicing isoforms identified by long-read RNA-seq, which may be associated with the SV. Fig. S6.1e shows more detail on the splicing isoforms. The differences in splicing between individuals 2 and 3 could reflect the SV disrupting regions important to splicing, as suggested by the known GTEx sQTL sites nearby (*72*). Alternatively, the differences in splicing between the two samples may be caused by tissue specificity: the individual 2 sample is from the adrenal gland, and the individual 3 sample is from the heart left ventricle. See Fig. S6.1f for another example of novel splicing variants that are potentially associated with an SV.
(F) The genomic regions neighboring TE SVs show reduced chromatin accessibility. The change in accessibility is determined by comparing the accessibility (from ATAC-seq) of individuals 2 and 3, taking as a reference the one without the SV (see Supplement & Fig. S6.2a). The x-axis is the chromatin openness in the reference individual. P-values are based on the Chi-squared test.

Fig. 6. Decorating the ENCODE Encyclopedia.
(A) Workflow for decorating cCREs by personal functional genomic data. The workflow starts with the master list of cCREs from the ENCODE Registry. These 0.9 million cCREs are not necessarily tissue-specific. The spleen is used as an example to illustrate the tissue-specific

decoration of these cCREs. In this tissue, 290k cCREs have functional genomic signals and can be categorized as active (117k), repressed (154k), or bivalent (19k). Then for each category, the cCREs can be consecutively classified according to their genomic locations (proximal or distal), CTCF binding, and allele specificity. (This classification can occur in any order.) As a result, the spleen has 2,866 AS cCREs with various genomic activities and locations. More details on the workflow are described in the Supplement. The numbers of active and repressed cCREs are comparable in each tissue (i.e., on average ~202k active and ~166k repressed cCREs). Only a small subset of the active cCREs exhibit AS activity (~2.5% or 1,750 cCREs, averaging across tissues; see Fig. S7.2d for all available tissues).

(B) The relationship between decorated cCREs and chromatin compartments from Hi-C. Active and tissue-specific cCREs are more likely to be located in the active compartment (A), while repressed tissue-specific cCREs are relatively skewed to compartment B (see Fig. S7.4 for all available tissues).

(C) Functional genomic data across individuals, tissues and assays. The joint analysis of variation is used to aggregate the functional data (JIVE) (*44*) into two-dimensional projections (showing each individual-tissue pair as a dot) for each assay and then into a single overall combined projection for all assays. (The tissues are colored according the GTEx convention from Fig. 1 and for each tissue the number of colored dots indicates the number of individuals, up to four.) In addition, a linear-regression-based approach is used to measure the difference between functional genomic signals in different tissues, assays, or individuals (details in Supplement). The difference is measured by one minus the explained variance of the regression. For example, on average, for H2K27ac 14% of the variation across the cCREs between the spleens of two individuals was unexplained (resulting in a dissimilarity between individuals). The corresponding dissimilarity was also small (15%) between the transverse colons of two individuals. In contrast, the average dissimilarity between these two types of tissue (in the same individual) increases to 35%. For H3K27me3, the dissimilarity between different tissues is also high (42%). The dissimilarities between different types of functional activity are even higher: in particular, the dissimilarity between H3K27ac and H3K4me3 is 51%. A similar analysis can also be done comparing RNA and protein abundances. A large-scale analysis for all available assays and tissues indicates that the dissimilarity is markedly assay and tissue-specific (see Supplement and Fig. S7.5).

Fig. 7. Measuring tissue specificity.

(A) Tissue specificity of various genomic annotations. For a given element, tissue specificity is measured by the number of tissues for which the element is active (see Supplement and Fig. S8.1 for more detail). Then by determining the fraction of elements in an annotation category (e.g. distal active cCREs) that are active in just one tissue, we determine fractional uniqueness for that category. A smaller uniqueness indicates a category that is more ubiquitous. The plot shows uniqueness values for many diverse annotation categories.

(B) Purifying selection on active and repressed cCREs in relation to tissue specificity. Purifying selection is measured as the fraction of rare variants to the total number of variants in an element (details in the Supplement). Rare variants are determined from 1,000 Genomes. A subset of active cCREs with a strong "peak-valley-peak" pattern of H3K27ac signal is identified ("stringent subset", see Supplement), and this set of cCREs has increased purifying selection.

(C) The direction of AS imbalance across accessible cCREs. The imbalance between the two haplotypes is measured by the fraction of AS reads that map to haplotype 1. A zoomed-in view of the most ubiquitous (non tissue-specific) AS cCREs shows that the imbalance direction is consistent across tissues. However, a few tissue-specific cCREs show directional flips between tissues.
(D) Distribution of cCREs with AS H3K27ac in tissues of individual 3 is shown in an UpSet plot. Most of these AS cCREs are detected only in a single tissue but are not AS in other tissues. However, a few AS cCREs are observed across many tissues.

Fig. 8. Relating Encyclopedia decorations to QTLs & GWAS loci.
(A) Comparing the effect of AS decoration for eQTL (top left) and sQTL (bottom left) enrichment. Colored dots show the enrichment for each tissue (using the GTEx colors from Fig. 1). Each bar shows the median enrichment over all tissues for a given decorated annotation subset. Overall, AS elements show stronger enrichments compared to non-AS ones. Median enrichment of Roadmap “Enh” and “TssA” annotations are shown as dashed and dotted lines, respectively, as a reference. The enrichments for the liver are highlighted. To estimate the robustness of this particular calculation, we resampled the genetic variants as described in Fig S9.1, estimating a range of enrichments, shown with whiskers. (Also, see Fig. S9.1a for enrichment results on all tissues and the Supplement for more detail.)

(B) The relationship between the tissue-specificity of GTEx eQTLs and H3K27ac signal. For each eQTL SNV we observe H3K27ac signal present in most tissues where the eQTL is active (right-hand-side) and absent in the tissues where the eQTL is not active (left-hand-side). Thus H3K27ac signal in different tissues can be used as a predictor of tissue activity for each eQTL SNV (see also Fig. S9.1d).
(C) GWAS tag SNP and LDSC enrichment of EN-TEx AS decorations. The heatmap (center) shows the GWAS tag SNP enrichment of distal active CTCF+ cCREs. The tissues (colored according to the GTEx convention from Fig. 1) run along the bottom, and phenotypes (usually diseases, not explicitly labeled) run along the vertical axis. Around the heatmap, we show zoomed-in views highlighting various comparisons. At the bottom, we show that the tissue specificity of the enrichment for the trait Granulocyte % of Myeloid WBC is much stronger for AS versus non-AS cCREs. On the left, we show higher LDSC enrichment for AS elements compared to the corresponding non-AS ones for one tissue (coronary artery) across many associated traits. Left-bottom extends this comparison (for a single tissue-trait pair) to include roadmap annotations.

1. F. S. Collins, E. D. Green, A. E. Guttmacher, M. S. Guyer, U. S. N. H. G. R. Institute, A vision for the future of genomics research. *Nature* **422**, 835-847 (2003).
2. Z. D. Stephens *et al.*, Big Data: Astronomical or Genomical? *PLoS Biol* **13**, e1002195 (2015).
3. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
4. A. Almas, J. Moller, R. Iqbal, Y. Forsell, Effect of neuroticism on risk of cardiovascular disease in depressed persons - a Swedish population-based cohort study. *BMC Cardiovasc Disord* **17**, 185 (2017).
5. G. TEx Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
6. G. TEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
7. Encode Project Consortium *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
8. Encode Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
9. C. A. Davis *et al.*, The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).
10. Encode Project Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
11. D. M. Roden, R. F. Tyndale, Genomic medicine, precision medicine, personalized medicine: what's in a name? *Clin Pharmacol Ther* **94**, 169-172 (2013).
12. C. A. Steward *et al.*, Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med* **9**, 49 (2017).
13. A. Z. Dayem Ullah *et al.*, SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res* **46**, W109-W113 (2018).
14. M. Pirinen *et al.*, Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31**, 2497-2504 (2015).
15. Y. Baran *et al.*, The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25**, 927-936 (2015).
16. M. T. Maurano *et al.*, Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* **47**, 1393-1401 (2015).
17. J. Chen *et al.*, A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7**, 11101 (2016).
18. V. Onuchic *et al.*, Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* **361**, (2018).
19. S. E. Castel *et al.*, A vast resource of allelic expression data spanning human tissues. *Genome Biol* **21**, 234 (2020).
20. R. Chen *et al.*, Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307 (2012).
21. P. Edge, V. Bafna, V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-812 (2017).
22. P. A. Audano *et al.*, Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675 e619 (2019).
23. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).

24. J. F. Degner *et al.*, Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212 (2009).
25. J. Rozowsky *et al.*, AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
26. M. B. Gerstein *et al.*, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
27. E. Khurana *et al.*, Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
28. J. Shi *et al.*, Determining Allele-Specific Protein Expression (ASPE) Using a Novel Quantitative Concatamer Based Proteomics Method. *J Proteome Res* **17**, 3606-3612 (2018).
29. L. Wu, M. Snyder, Impact of allele-specific peptides in proteome quantification. *Proteomics Clin Appl* **9**, 432-436 (2015).
30. J. Barber, M. R. Russell, A. Rostami-Hodjegan, B. Achour, Characterization of CYP2B6 K262R allelic variants by quantitative allele-specific proteomics using a QconCAT standard. *J Pharm Biomed Anal* **178**, 112901 (2020).
31. Y. Fu *et al.*, FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).
32. P. Pawliczek *et al.*, ClinGen Allele Registry links information about genetic variants. *Hum Mutat* **39**, 1690-1701 (2018).
33. Y. Ji , Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **btab083**, (2021).
34. S. Ghaemmaghami *et al.*, Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
35. D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117 (2003).
36. H. Sasaki, K. Ishihara, R. Kato, Mechanisms of Igf2/H19 imprinting: DNA methylation, chromatin and long-distance gene regulation. *J Biochem* **127**, 711-715 (2000).
37. J. M. Autuoro, S. P. Pirnie, G. G. Carmichael, Long noncoding RNAs in imprinting and X chromosome inactivation. *Biomolecules* **4**, 76-100 (2014).
38. Y. Itoh *et al.*, The X-linked histone demethylase Kdm6a in CD4+ T lymphocytes modulates autoimmunity. *J Clin Invest* **129**, 3852-3863 (2019).
39. C. Chiang *et al.*, The impact of structural variation on human gene expression. *Nat Genet* **49**, 692-699 (2017).
40. N. Zamudio, D. Bourc'his, Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity (Edinb)* **105**, 92-104 (2010).
41. H. L. Levin, J. V. Moran, Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**, 615-627 (2011).
42. J. L. Goodier, H. H. Kazazian, Jr., Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23-35 (2008).
43. A. Sethi *et al.*, Supervised enhancere prediction with epigenetic pattern recognition and targeted validation. *Nat Methods* **17**, 807-814 (2020).
44. K. H. Hellton, M. Thoresen, Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics* **17**, 537-548 (2016).
45. I. Kosti, N. Jain, D. Aran, A. J. Butte, M. Sirota, Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci Rep* **6**, 24799 (2016).
46. M. Mele *et al.*, Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).
47. D. Wang *et al.*, A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **15**, e8503 (2019).

48. B. W. Hounkpe, F. Chenou, F. de Lima, E. V. De Paula, HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* **49**, D947-D955 (2021).
49. S. Whalen, K. S. Pollard, Most chromatin interactions are not in linkage disequilibrium. *Genome Res* **29**, 334-343 (2019).
50. B. K. Bulik-Sullivan *et al.*, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
51. S. S. Khan *et al.*, Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity. *JAMA Cardiol* **3**, 280-287 (2018).
52. L. Emilsson, B. Lebwohl, J. Sundstrom, J. F. Ludvigsson, Cardiovascular disease in patients with coeliac disease: A systematic review and meta-analysis. *Dig Liver Dis* **47**, 847-852 (2015).
53. A. Terracciano, C. E. Lockenhoff, A. B. Zonderman, L. Ferrucci, P. T. Costa, Jr., Personality predictors of longevity: activity, emotional stability, and conscientiousness. *Psychosom Med* **70**, 621-627 (2008).
54. T. R. Einarson, A. Acs, C. Ludwig, U. H. Panton, Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007-2017. *Cardiovasc Diabetol* **17**, 83 (2018).
55. T. Amariuta *et al.*, Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat Genet* **52**, 1346-1354 (2020).
56. K. Watanabe, E. Taskesen, A. van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
57. E. Cano-Gamez, G. Trynka, From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet* **11**, 424 (2020).
58. T. Lappalainen, Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res* **25**, 1427-1431 (2015).
59. modENCODE Consortium *et al.*, Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797 (2010).
60. W. Huang *et al.*, Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* **24**, 1193-1208 (2014).
61. F. Yue *et al.*, A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).
62. J. J. Crowley *et al.*, Analysis of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**, 353-360 (2015).
63. STAR Consortium *et al.*, SNP and haplotype mapping for genetic analysis in the rat. *Nat Genet* **40**, 560-566 (2008).
64. K. Lindblad-Toh *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
65. J. Plassais *et al.*, Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun* **10**, 1489 (2019).
66. G. Zhang *et al.*, Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320 (2014).
67. M. Li *et al.*, Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res* **27**, 865-874 (2017).
68. M. Schmid *et al.*, Third report on chicken genes and chromosomes. *Cytogenet Genome Res* **145**, 78-179 (2015).
69. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-1063 (2015).

70. S. L. Berger, The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412 (2007).
71. M. M. Suzuki, A. Bird, DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476 (2008).
72. D. Garrido-Martín, B. Borsari, M. Calvo, F. Reverter, R. Guigó, Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun* **12**, 727 (2021).

Fig. 1

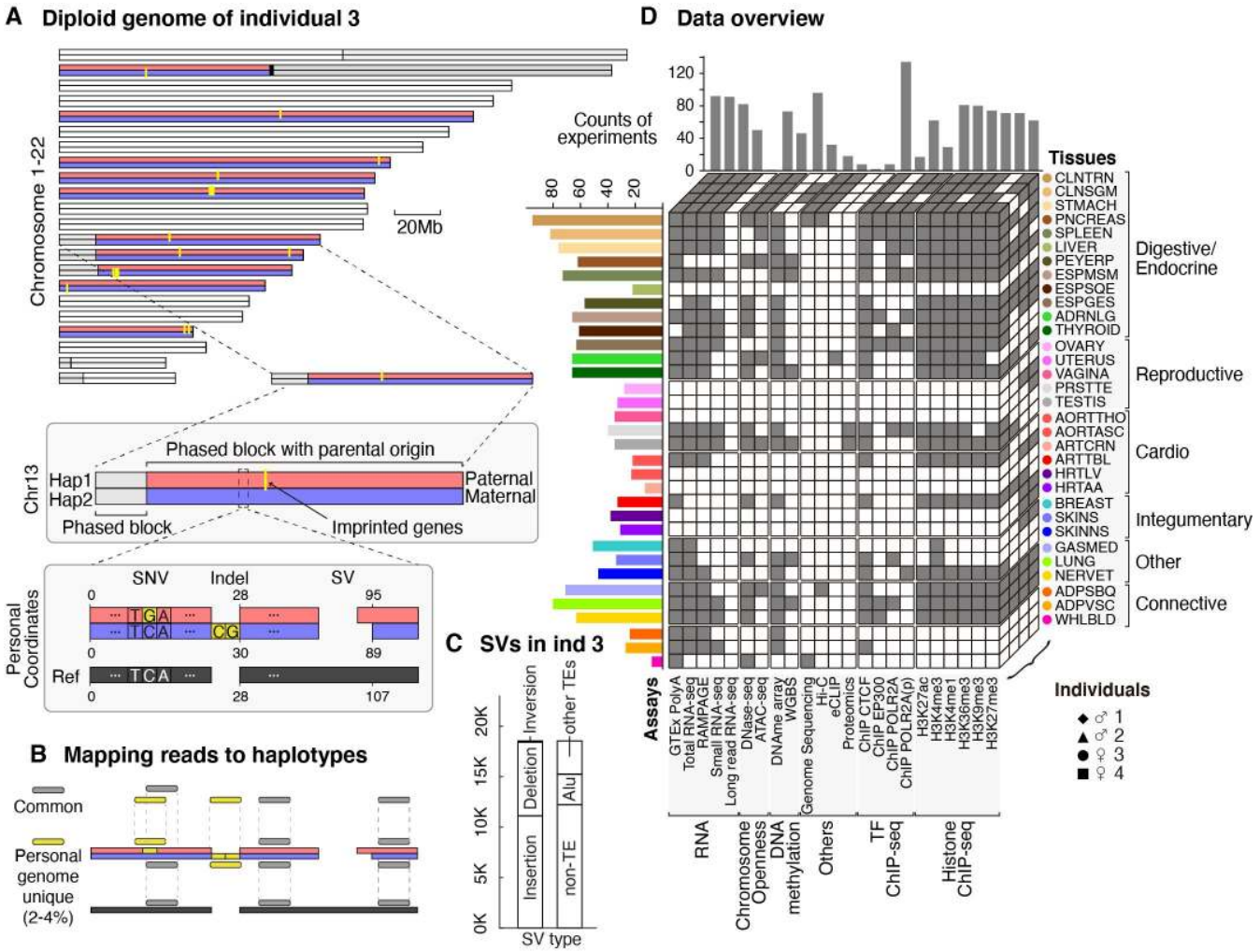

A Diploid genome of individual 3
Chromosome 1-22
20Mb
Chr13
Phased block with parental origin
Hap1
Hap2
Paternal
Maternal
Phased block
Imprinted genes
Personal Coordinates
SNV
Indel
SV
Ref
B Mapping reads to haplotypes
Common
Personal genome unique (2-4%)
C SVs in ind 3
Inversion
Deletion
Insertion
other TEs
Alu
non-TE
SV type
D Data overview
Counts of experiments
Tissues
CLNTRN
CLNSGM
STMACH
PNCREAS
SPLEEN
LIVER
PEYERP
ESPMSM
ESPSQE
ESPGES
ADRNLG
THYROID
OVARY
UTERUS
VAGINA
PRSTTE
TESTIS
AORTTHO
AORTASC
ARTCRN
ARTTBL
HRTLV
HRTAA
BREAST
SKINS
SKINNS
GASMED
LUNG
NERVET
ADPSBQ
ADPVSC
WHLBLD
Digestive/ Endocrine
Reproductive
Cardio
Integumentary
Other
Connective
Individuals
Assays
GTEx PolyA
Total RNA-seq
RAMPAGE
Small RNA-seq
Long read RNA-seq
DNase-seq
ATAC-seq
DNAme array
WGBS
Genome Sequencing
Hi-C
eCLIP
Proteomics
ChIP CTCF
ChIP EP300
ChIP POLR2A
ChIP POLR2A(p)
H3K27ac
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
RNA
Chromosome Openness
DNA methylation
Others
TF ChIP-seq
Histone ChIP-seq

Fig. 2

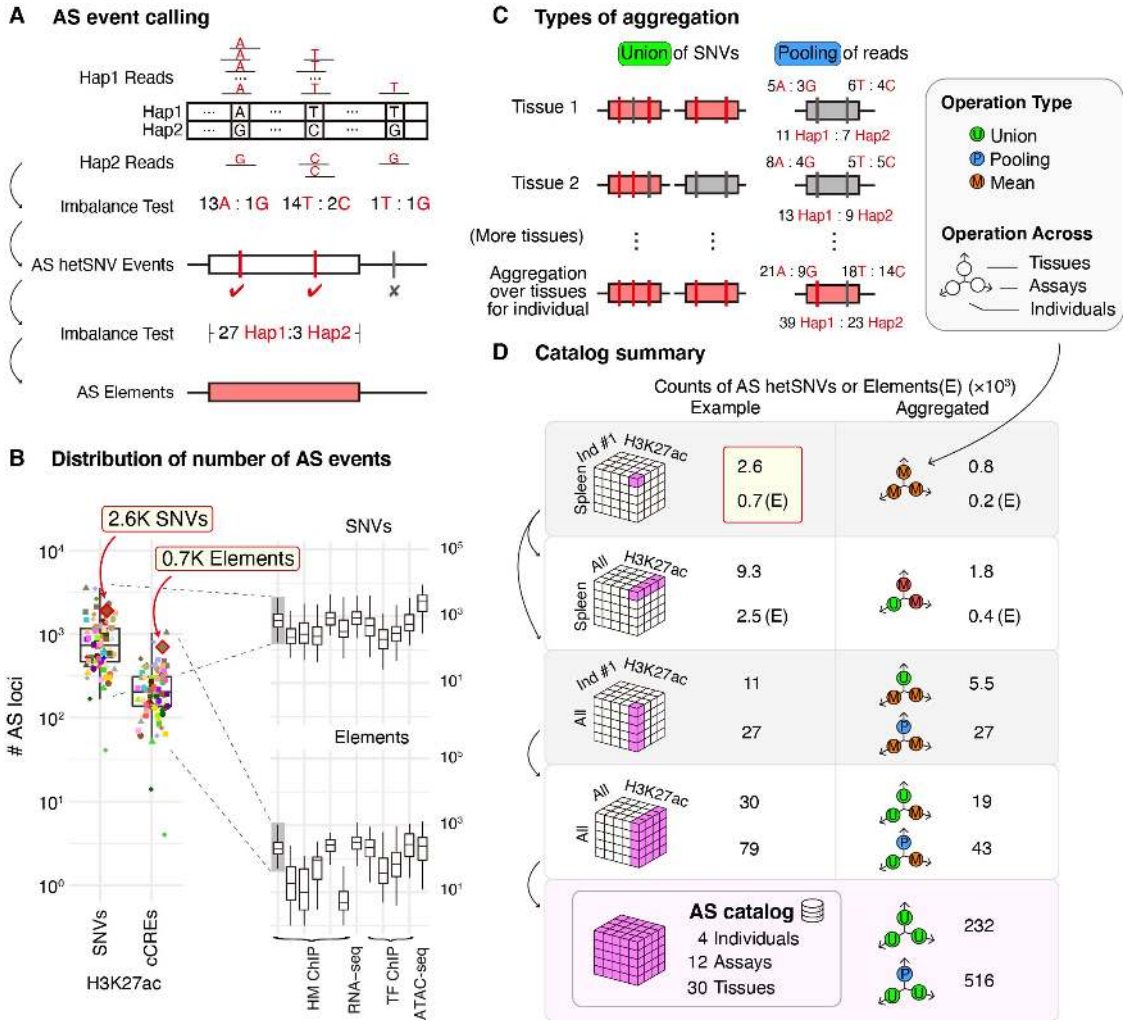

A AS event calling
Hap1 Reads
Hap1
Hap2
Hap2 Reads
Imbalance Test
13A : 1G
14T : 2C
1T : 1G
AS hetSNV Events
Imbalance Test
27 Hap1:3 Hap2
AS Elements
B Distribution of number of AS events
2.6K SNVs
0.7K Elements
SNVs
Elements
AS loci
SNVs
cCREs
H3K27ac
HM ChIP
RNA-seq
TF ChIP
ATAC-seq
C Types of aggregation
Union of SNVs
Pooling of reads
Tissue 1
5A : 3G
6T : 4C
11 Hap1 : 7 Hap2
Tissue 2
8A : 4G
5T : 5C
13 Hap1 : 9 Hap2
(More tissues)
Aggregation over tissues for individual
21A : 9G
18T : 14C
39 Hap1 : 23 Hap2
Operation Type
Union
Pooling
Mean
Operation Across
Tissues
Assays
Individuals
D Catalog summary
Counts of AS hetSNVs or Elements(E) (×10³)
Example
Aggregated
Spleen
Ind #1
H3K27ac
2.6
0.7 (E)
0.8
0.2 (E)
All
9.3
2.5 (E)
1.8
0.4 (E)
11
27
5.5
27
30
79
19
43
AS catalog
4 Individuals
12 Assays
30 Tissues
232
516

Fig. 3

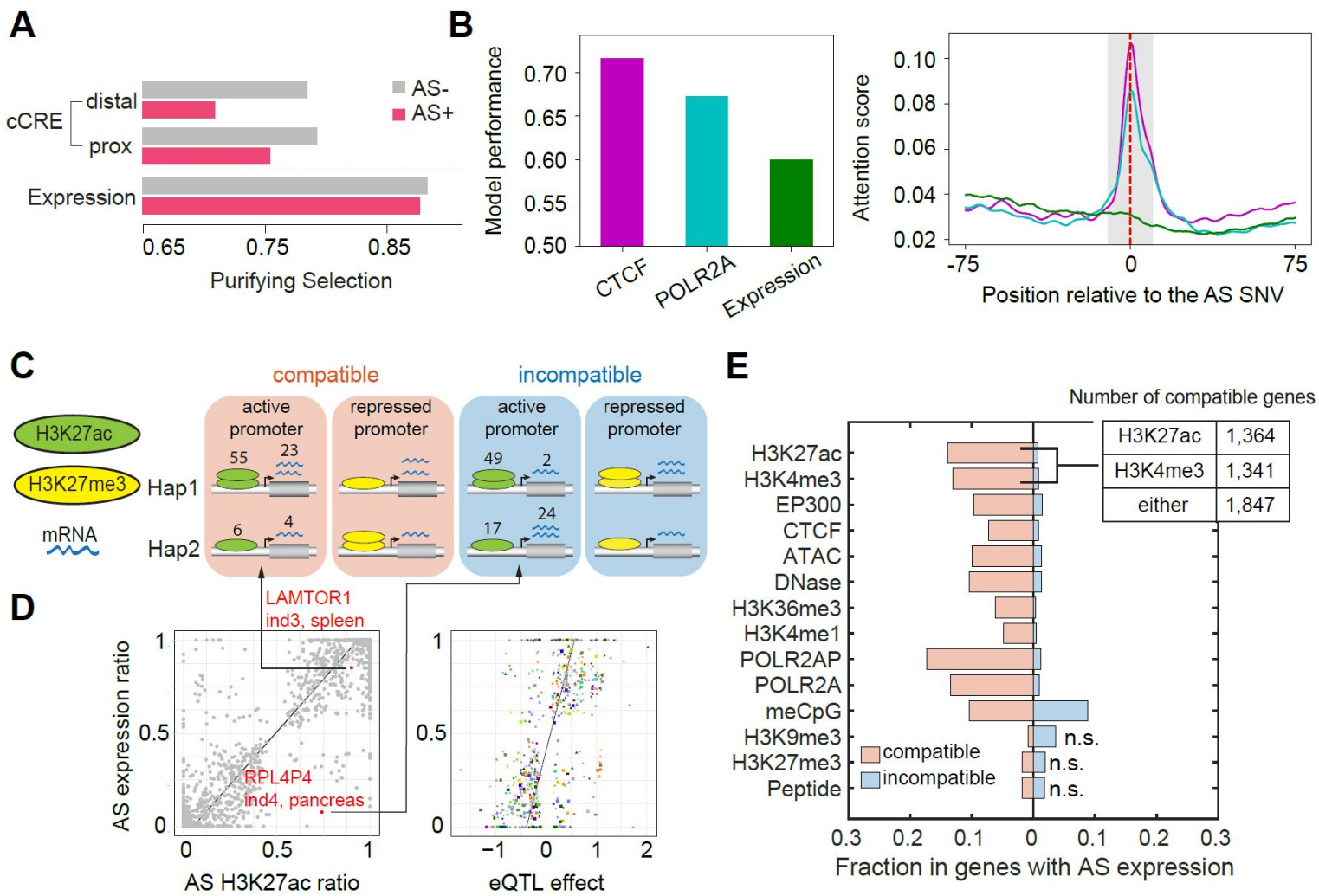

A
cCRE
distal
prox
Expression
AS-
AS+
0.65
0.75
0.85
Purifying Selection
B
Model performance
0.70
0.65
0.60
0.55
0.50
CTCF
POLR2A
Expression
Attention score
0.10
0.08
0.06
0.04
0.02
-75
0
75
Position relative to the AS SNV
C
compatible
incompatible
active promoter
repressed promoter
active promoter
repressed promoter
H3K27ac
H3K27me3
mRNA
Hap1
Hap2
55
23
6
4
49
2
17
24
D
LAMTOR1
ind3, spleen
RPL4P4
ind4, pancreas
AS expression ratio
AS H3K27ac ratio
eQTL effect
E
Number of compatible genes
H3K27ac 1,364
H3K4me3 1,341
either 1,847
H3K27ac
H3K4me3
EP300
CTCF
ATAC
DNase
H3K36me3
H3K4me1
POLR2AP
POLR2A
meCpG
H3K9me3
H3K27me3
Peptide
compatible
incompatible
n.s.
n.s.
n.s.
0.3
0.2
0.1
0
0.1
0.2
0.3
Fraction in genes with AS expression

Fig. 4

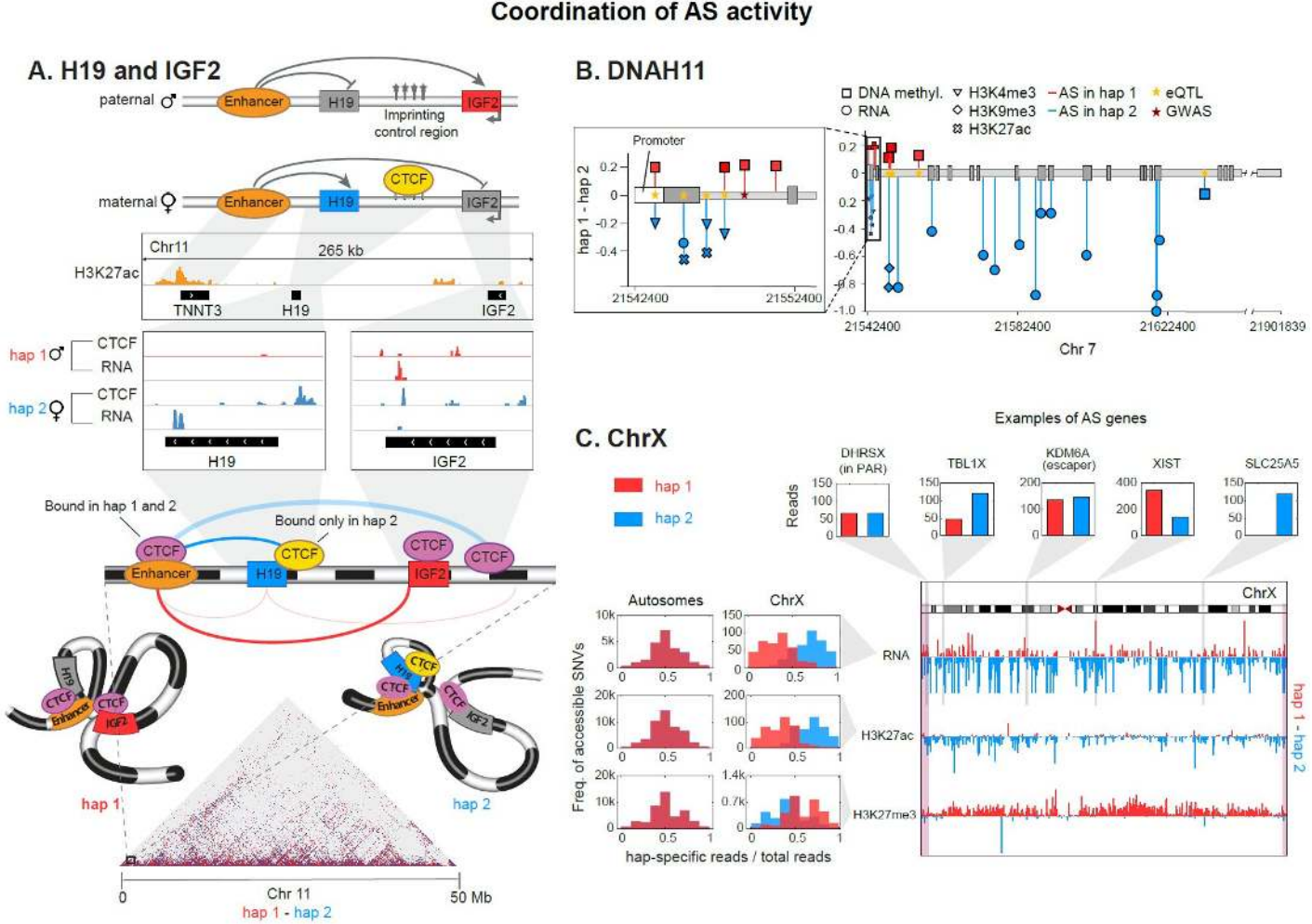

Coordination of AS activity
A. H19 and IGF2
paternal ♂
Enhancer
H19
Imprinting control region
IGF2
maternal ♀
CTCF
Chr11
265 kb
H3K27ac
TNNT3
hap 1 ♂
hap 2 ♀
RNA
Bound in hap 1 and 2
Bound only in hap 2
hap 1
hap 2
Chr 11
0
50 Mb
hap 1 - hap 2
B. DNAH11
DNA methyl.
RNA
H3K4me3
H3K9me3
H3K27ac
AS in hap 1
AS in hap 2
eQTL
GWAS
Promoter
hap 1 - hap 2
21542400
21552400
21582400
21622400
21901839
Chr 7
C. ChrX
Examples of AS genes
DHRSX (in PAR)
TBL1X
KDM6A (escaper)
XIST
SLC25A5
Reads
Autosomes
ChrX
Freq. of accessible SNVs
hap-specific reads / total reads
RNA
H3K27ac
H3K27me3
hap 1 - hap 2

Fig. 5

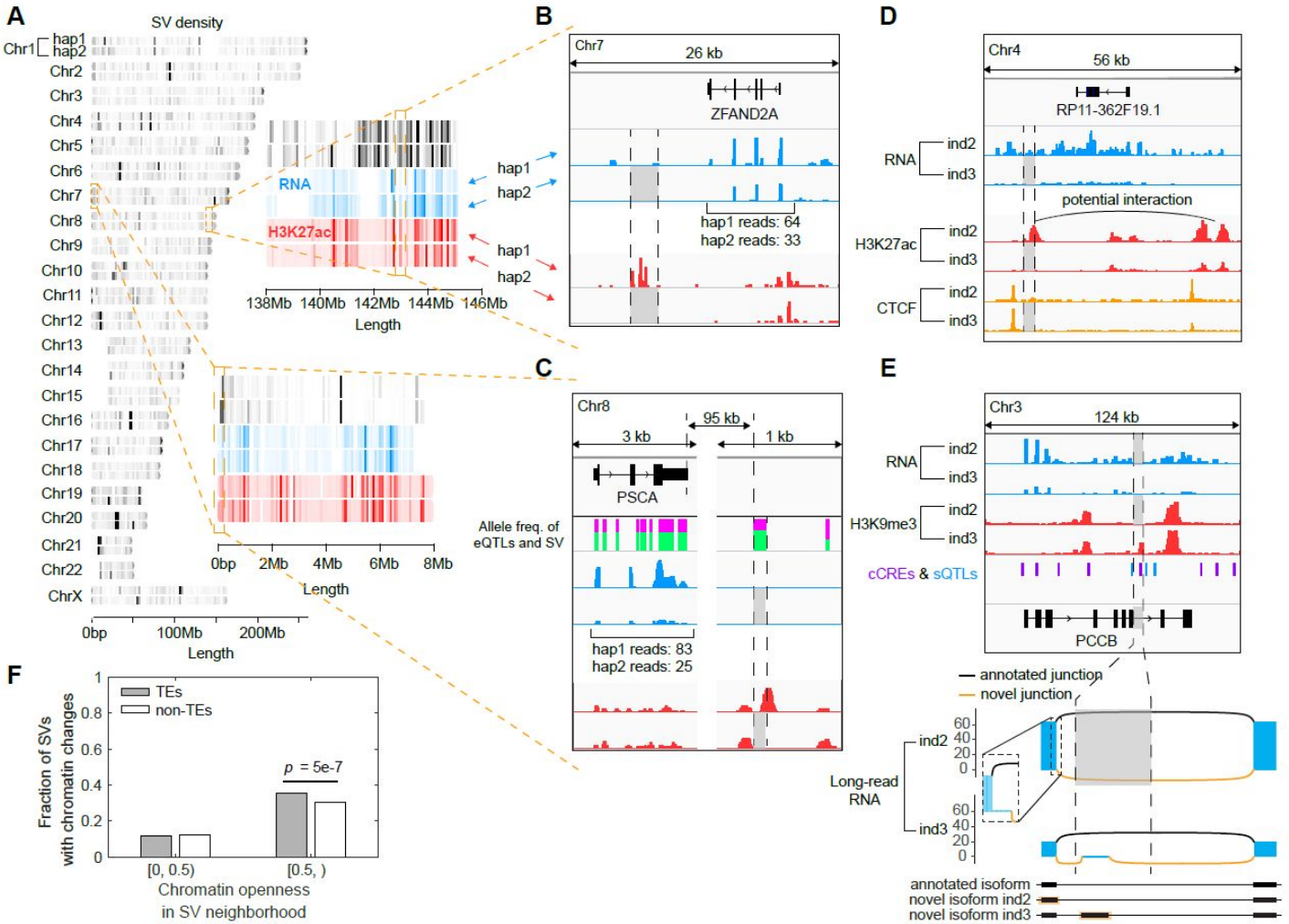

A
SV density
Chr1
hap1
hap2
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
0bp
100Mb
200Mb
Length
RNA
H3K27ac
hap1
hap2
138Mb 140Mb 142Mb 144Mb 146Mb
Length
0bp 2Mb 4Mb 6Mb 8Mb
Length
B
Chr7
26 kb
ZFAND2A
hap1 reads: 64
hap2 reads: 33
C
Chr8
3 kb
95 kb
1 kb
PSCA
Allele freq. of eQTLs and SV
hap1 reads: 83
hap2 reads: 25
D
Chr4
56 kb
RP11-362F19.1
RNA
ind2
ind3
potential interaction
H3K27ac
CTCF
E
Chr3
124 kb
RNA
H3K9me3
cCREs & sQTLs
PCCB
annotated junction
novel junction
Long-read RNA
annotated isoform
novel isoform ind2
novel isoform ind3
F
TEs
non-TEs
p = 5e-7
Fraction of SVs with chromatin changes
[0, 0.5)
[0.5,)
Chromatin openness in SV neighborhood

Fig. 6

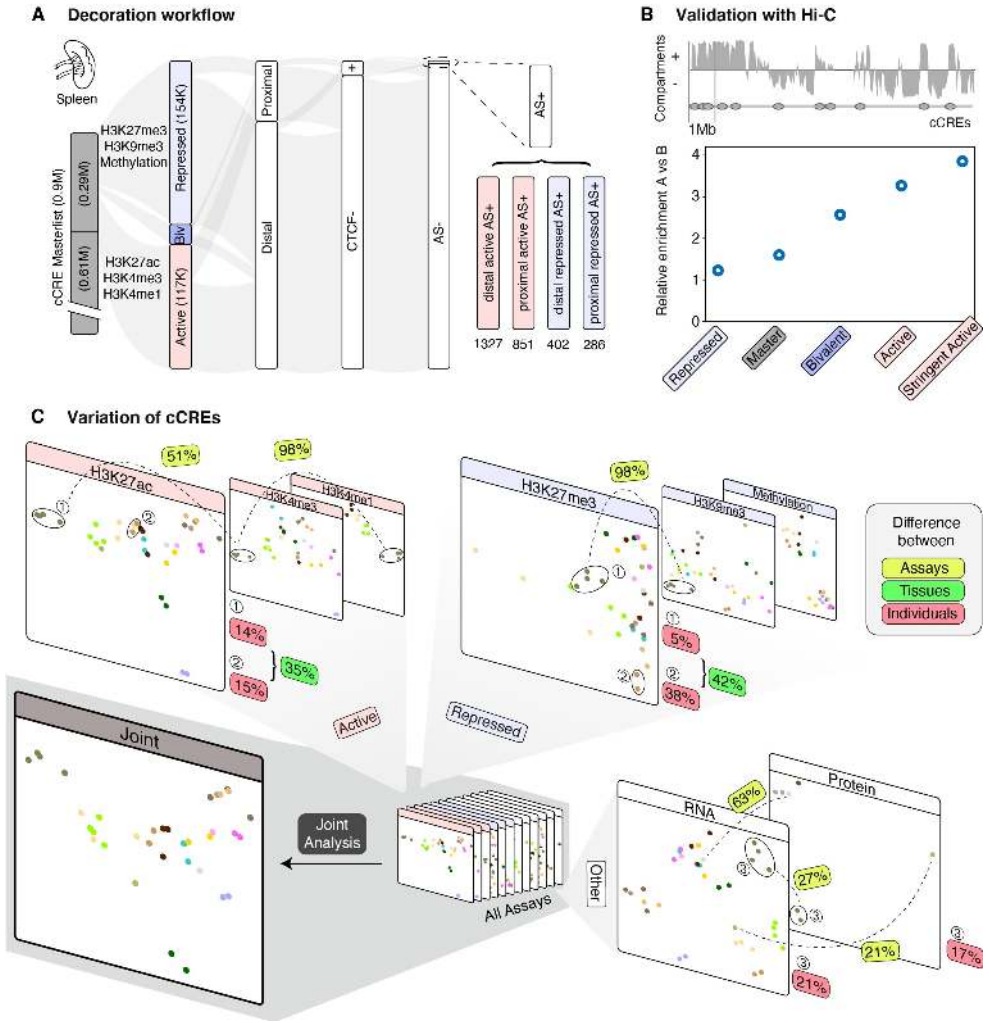

A Decoration workflow
Spleen
cCRE Masterlist (0.9M)
(0.29M)
(0.61M)
H3K27me3
H3K9me3
Methylation
H3K27ac
H3K4me3
H3K4me1
Repressed (154K)
Biv
Active (117K)
Proximal
Distal
CTCF+
CTCF-
AS+
AS-
distal active AS+
proximal active AS+
distal repressed AS+
proximal repressed AS+
1327
851
402
286
B Validation with Hi-C
Compartments
1Mb
cCREs
Relative enrichment A vs B
Repressed
Master
Bivalent
Active
Stringent Active
C Variation of cCREs
H3K27ac
H3K4me3
H3K4me1
51%
98%
14%
35%
15%
H3K27me3
H3K9me3
Methylation
98%
5%
42%
38%
Difference between
Assays
Tissues
Individuals
Active
Repressed
Joint
Joint Analysis
All Assays
Other
RNA
Protein
68%
27%
21%
17%
21%

Fig. 7

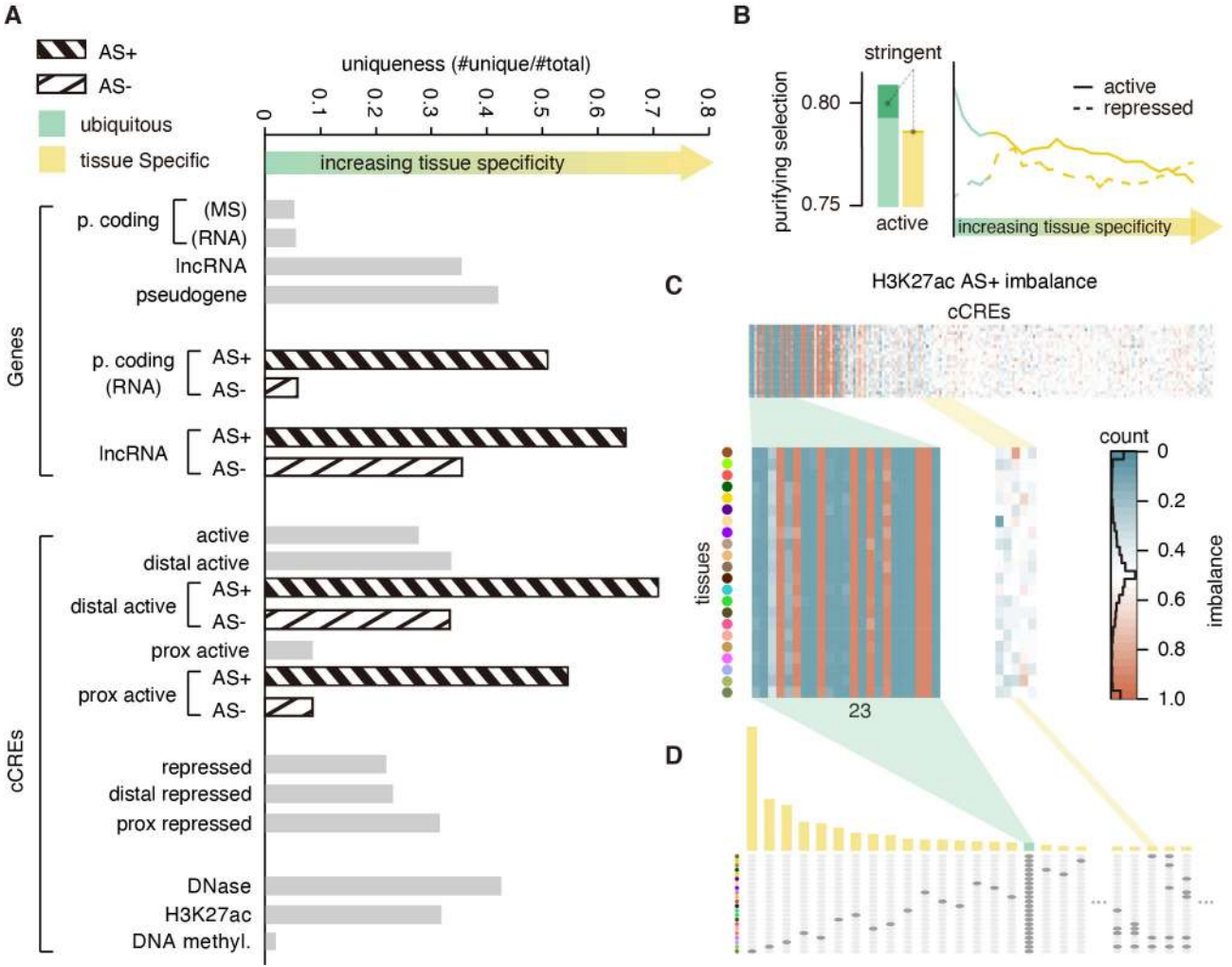
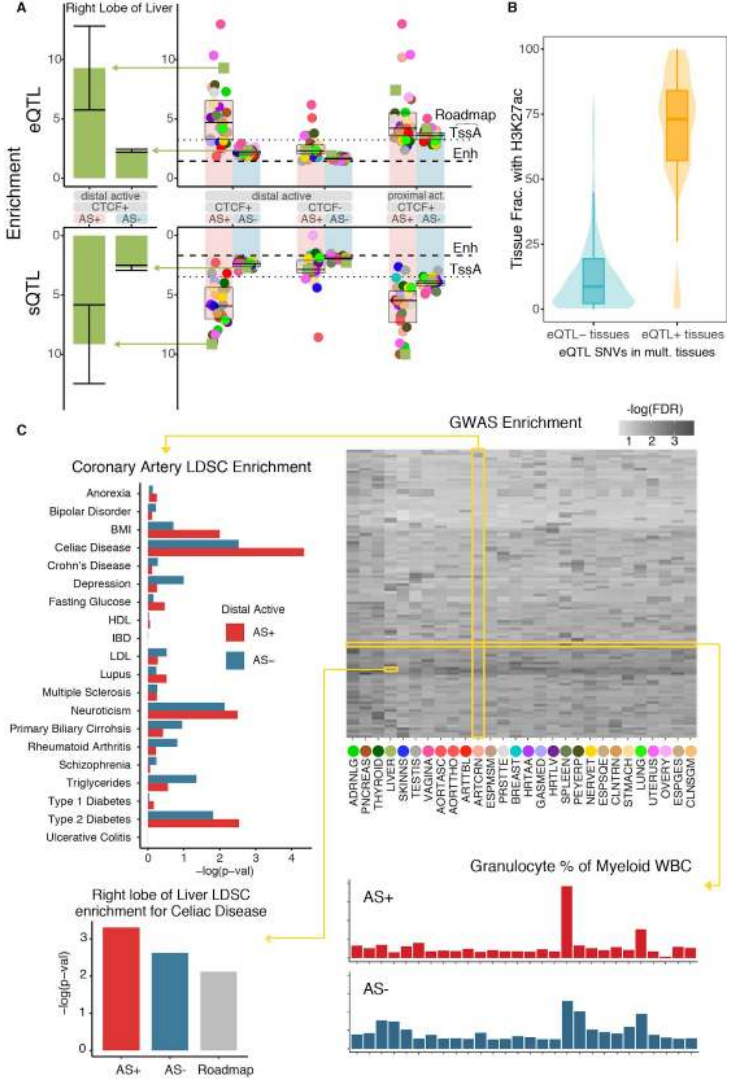

A
AS+
AS-
ubiquitous
tissue Specific
uniqueness (#unique/#total)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
increasing tissue specificity
Genes
p. coding
(MS)
(RNA)
lncRNA
pseudogene
p. coding (RNA)
AS+
AS-
lncRNA
AS+
AS-
cCREs
active
distal active
distal active
AS+
AS-
prox active
prox active
AS+
AS-
repressed
distal repressed
prox repressed
DNase
H3K27ac
DNA methyl.
B
stringent
purifying selection
0.80
0.75
active
active
repressed
increasing tissue specificity
C
H3K27ac AS+ imbalance
cCREs
tissues
23
count
0
0.2
0.4
0.6
0.8
1.0
imbalance
D

Fig. 8


A
Right Lobe of Liver
eQTL
sQTL
Enrichment
Roadmap
TssA
Enh
distal active
CTCF+
AS+
AS-
distal active
CTCF+
CTCF-
proximal act.
CTCF+
B
Tissue Frac. with H3K27ac
eQTL– tissues
eQTL+ tissues
eQTL SNVs in mult. tissues
C
GWAS Enrichment
-log(FDR)
Coronary Artery LDSC Enrichment
Anorexia
Bipolar Disorder
BMI
Celiac Disease
Crohn's Disease
Depression
Fasting Glucose
HDL
IBD
LDL
Lupus
Multiple Sclerosis
Neuroticism
Primary Biliary Cirrohsis
Rheumatoid Arthritis
Schizophrenia
Triglycerides
Type 1 Diabetes
Type 2 Diabetes
Ulcerative Colitis
Distal Active
AS+
AS-
-log(p-val)
Granulocyte % of Myeloid WBC
AS+
AS-
Right lobe of Liver LDSC
enrichment for Celiac Disease
-log(p-val)
AS+
AS-
Roadmap