Title

A single-cell atlas of chromatin accessibility in the human genome

Permalink

https://escholarship.org/uc/item/7z7914tt

Journal

Cell, 184(24)

ISSN

0092-8674

Authors

Zhang, Kai
Hocker, James D
Miller, Michael
et al.

Publication Date

2021-11-01

DOI

10.1016/j.cell.2021.10.024

Peer reviewed

Published in final edited form as:

Cell. 2021 November 24; 184(24): 5985–6001.e19. doi:10.1016/j.cell.2021.10.024.

A single-cell atlas of chromatin accessibility in the human genome

Kai Zhang[1,6,8], **James D. Hocker**[1,2,3,8], **Michael Miller**[4], **Xiaomeng Hou**[4], **Joshua Chiou**[3,5], **Olivier B. Poirion**[4], **Yunjiang Qiu**[1], **Yang E. Li**[1,6], **Kyle J. Gaulton**[5,7], **Allen Wang**[4], **Sebastian Preissl**[4], **Bing Ren**[1,4,6,7,9,*]

[1] Ludwig Institute for Cancer Research, La Jolla, CA, USA

[2] Medical Scientist Training Program, University of California San Diego, La Jolla, CA, USA

[3] Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA

[4] Center for Epigenomics, University of California San Diego, La Jolla, CA, USA

[5] Department of Pediatrics, Pediatric Diabetes Research Center, University of California San Diego, La Jolla, CA, USA

[6] Department of Cellular and Molecular Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA

[7] Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA

[8] These authors contributed equally

[9] Lead contact

SUMMARY

Current catalogs of regulatory sequences in the human genome are still incomplete and lack cell type resolution. To profile the activity of gene regulatory elements in diverse cell types and tissues in the human body, we applied single-cell chromatin accessibility assays to 30 adult human tissue types from multiple donors. We integrated these datasets with single-cell chromatin accessibility data from 15 fetal tissue types to reveal the status of open chromatin for approximately 1.2 million candidate *cis*-regulatory elements (cCREs) in 222 distinct cell types comprised of >1.3 million nuclei. We used these chromatin accessibility maps to delineate cell type-specificity of

[*] Correspondence: Bing Ren, PhD (biren@health.ucsd.edu).

AUTHOR CONTRIBUTIONS

Study was conceived by: J.D.H., S.P., A.W., and B.R. Study supervision: B.R. Supervision of data generation: S.P., A.W. and B.R. Contribution to data generation: J.D.H., X.H., M.M. Contribution to data analysis: K.Z., J.D.H., J.C., O.P. Y.E.L., Y.Q. Contribution to web portal: Y.E.L., K.Z. Contribution to data interpretation: K.Z., J.D.H., S.P., A.W., K.J.G. Contribution to writing the manuscript: K.Z., J.D.H., B.R. All authors edited and approved the manuscript.

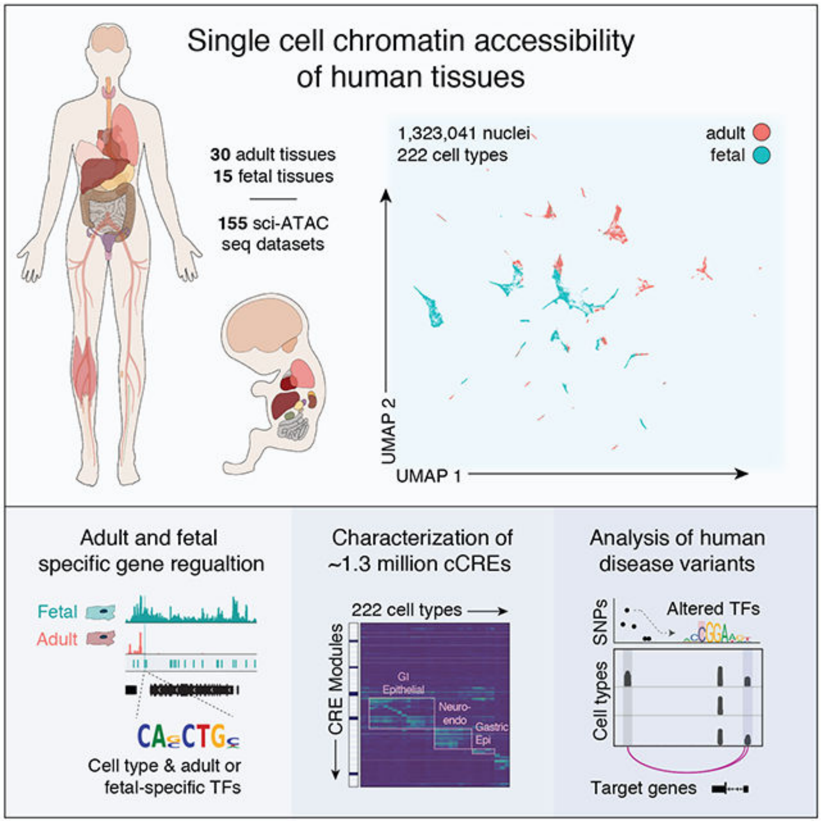
DECLARATION OF INTERESTS

B.R. is a shareholder and consultant of Arima Genomics, Inc., and a co-founder of Epigenome Technologies, Inc. K.J.G is a consultant of Genentech, and shareholder in Vertex Pharmaceuticals. These relationships have been disclosed to and approved by the UCSD Independent Review Committee.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

fetal and adult human cCREs and to systematically interpret the noncoding variants associated with complex human traits and diseases. This rich resource provides a foundation for the analysis of gene regulatory programs in human cell types across tissues, life stages, and organ systems.

Graphical Abstract


Single cell chromatin accessibility of human tissues
30 adult tissues
15 fetal tissues
155 sci-ATAC seq datasets
1,323,041 nuclei
222 cell types
adult
fetal
UMAP 2
UMAP 1
Adult and fetal specific gene regualtion
Fetal
Adult
CAsCTGs
Cell type & adult or fetal-specific TFs
Characterization of ~1.3 million cCREs
222 cell types
CRE Modules
GI Epithelial
Neuro-endo
Gastric Epi
Analysis of human disease variants
SNPs
Altered TFs
Cell types
Target genes

In Brief

A cell-type-resolved map of human cis-regulatory elements, derived from single cell analysis of diverse tissue types, facilitates functional interpretation of the noncoding variants associated with complex human traits and diseases.

Keywords

Single cell ATAC-seq; GWAS; cis regulatory elements; enhancers; chromatin accessibility; epigenome; noncoding variants

INTRODUCTION

The human body is comprised of various organs, tissues and cell types, each with highly specialized functions. The genes expressed in each tissue and cell type – and in turn their physiologic roles in the body – are regulated by *cis*-regulatory elements such as enhancers and promoters (Carter and Zhao, 2020). These sequences dictate the expression patterns of target genes by recruiting sequence specific transcription factors (TFs) in a cell-type specific manner (Shlyueva et al., 2014). Upon binding of TFs, *cis*-regulatory elements frequently adopt conformational changes such that they are more accessible to endonucleases or transposases, enabling genome-wide discovery of candidate *cis*-regulatory elements (cCREs) by combining assays incorporating these enzymes with high throughput sequencing (Buenrostro et al., 2013; John et al., 2013; Klemm et al., 2019). However, conventional assays have, in large part, used heterogeneous tissues as input materials to produce population average measurements, and consequently, the current catalogs of candidate regulatory sequences in the human genome (Andersson et al., 2014; Meuleman et al., 2020; Moore et al., 2020; Roadmap Epigenomics et al., 2015; Shen et al., 2012) still lack information about the cell type-specific activities of most elements. This limitation has hampered our ability to study gene regulatory programs in distinct human cell types and to interpret the noncoding DNA in the human genome.

Genome wide association studies (GWAS) have identified hundreds of thousands of genetic variants associated with a broad spectrum of human traits and diseases. The large majority of these variants are noncoding (Claussnitzer et al., 2020). Observations that annotated *cis*-regulatory elements in disease-relevant tissues and cell types are enriched for noncoding disease risk variants (Ernst et al., 2011; Maurano et al., 2012; Roadmap Epigenomics et al., 2015) led to the hypothesis that a major mechanism by which noncoding variants influence disease risk is by affecting transcriptional regulatory elements in specific cell types. However, annotation of these noncoding risk variants has been hindered by a lack of cell type-resolved maps of regulatory elements in the human genome. While innovative approaches to distinguish causal variants from local variants in linkage disequilibrium (LD) using fine mapping (Wakefield, 2009), and to link variants to target genes using co-accessibility of open chromatin regions in single-cells (Pliner et al., 2018) or 3-dimensional chromosomal contact-based linkage scores (Nasser et al., 2020), have made important strides toward the prioritization of causal variants and the prediction of their target genes, functional interpretation of the noncoding variants continues to be challenging.

Single-cell omics technologies, enabled by droplet-based, combinatorial barcoding or other approaches, have now enabled the profiling of transcriptome, epigenome and chromatin organization from complex tissues at single-cell resolution (Grosselin et al., 2019; Klein et al., 2015; Lake et al., 2018; Luo et al., 2017a; Macosko et al., 2015; Preissl et al., 2018). In particular, combinatorial cellular barcoding-based assays such as sci-ATAC-seq (Cusanovich et al., 2015) have permitted the identification of cCREs in single nuclei without the need for physical purification of individual cell types. The resulting data can be used to deconvolute cell types from mixed cell populations and to dissect cell type-specific transcriptomic and epigenomic states in primary tissues. While these tools have been applied to mammalian tissues including murine biosamples (Cusanovich et al., 2018; Lareau et al., 2019; Li et al., 2021; Preissl et al., 2018; Sinnamon et al., 2019), human fetal tissues (Domcke et al., 2020; Trevino et al., 2021), and a few individual adult human organ systems (Chiou et al., 2019; Corces et al., 2020; Hocker et al., 2020; Wang et al., 2020), we still lack comprehensive cell-type-resolved maps of cCREs from most primary tissues of the adult human body.

In the present study we used a modified sci-ATAC-seq protocol optimized for flash frozen primary tissues (Hocker et al., 2020; Preissl et al., 2018) to profile chromatin accessibility in 30 adult human tissue types from multiple donors. We profiled 615,998 nuclei from these tissues, grouped them into 111 distinct cell types based on similarity in chromatin landscapes, and identified a union of 890,130 open chromatin regions corresponding to cCREs from the resulting maps. We next integrated these data with a recent fetal cell atlas of chromatin accessibility (Domcke et al., 2020) to reveal open chromatin profiles for >1.3 million cells across the human lifespan, and chromatin accessibility maps at 1,154,611 cCREs covering 14.8% of the genome for 222 cell types. Finally, we used this cCRE atlas to interpret cell types and target genes for noncoding variants associated with 240 complex human traits and diseases, reveal cell type-disease associations and suggest relevant therapeutic targets in human cell types. We created an interactive web atlas to disseminate this resource [CATLAS, Cis-element ATLAS] http://catlas.org/humanenhancer.

RESULTS

Single-cell chromatin accessibility analysis of adult human primary tissues

To generate a cell atlas of cCREs in the adult human body, we performed sci-ATAC-seq (Cusanovich et al., 2015; Preissl et al., 2018) with primary tissue samples collected from 30 distinct anatomic sites in postmortem adult human donors (Figure 1A, Table S1). Tissue samples were chosen to survey a breadth of human organ systems which differed in their nuclear compositions and sensitivities to mechanical dissociation, posing a technical challenge. We thus optimized nuclear isolation methods and buffer conditions for different tissue types (Table S1, see STAR Methods). Subsequently, we generated sci-ATAC-seq datasets using a semi-automated workflow (Hocker et al., 2020; Preissl et al., 2018) and sequenced resulting libraries to 6,464 raw sequence reads per nucleus on average, with a median read duplication rate of 44.88% (Table S2). After filtering out lower quality nuclei and potential doublets, we finally obtained high quality open chromatin profiles for 615,998 nuclei, with a median of 2,822 unique open chromatin fragments per nucleus and an average transcription start site (TSS) enrichment score of 12.8 (±3.2) per nucleus (Figure 1B, Table S2, see STAR Methods and Data S1).

Analyzing large single-cell chromatin accessibility datasets has been challenging. In the latest development of SnapATAC (Fang et al., 2021), we further improved its scalability to handle millions of cells. Using this algorithm, we first identified 30 major cell groups (Figure 1B), 22 (73%) of which were found to consist of multiple subclusters during a second round of clustering analysis (see STAR Methods and Methods S1). Altogether we uncovered a total of 111 distinct cell clusters (Figure 1B-E).

Annotation of major and sub-classes of human cell types

To annotate the resulting cell clusters, we first curated a set of marker genes from the PanglaoDB marker gene database (Franzén et al., 2019) corresponding to expected human cell types. We utilized chromatin accessibility at the promoter as a proxy for gene activity and computed cell-type enrichment scores for each of the 111 clusters to create initial cell cluster annotations (see STAR Methods and Methods S1). We next manually reviewed these assignments based on focused consideration of marker gene accessibility (see Methods S1). Altogether, we annotated each of the 30 major cell groups and all 111 distinct clusters with a cell type label (Figure 1E, Table S3). For example, within the major cell group of gastrointestinal epithelial cells, higher resolution subclustering and annotation revealed three clusters of colon epithelial cells, one cluster of enterocytes from the small intestine, two clusters of goblet cells from the colon and small intestine respectively, and three rare populations with distinct chromatin accessibility profiles including enterochromaffin cells (0.060% of total nuclei), tuft cells (0.050% of total nuclei), and Paneth cells (0.045% of total nuclei) (Figure 1B-C).

Encouragingly, several prevalent cell types detected in most tissue samples such as endothelial cells and myeloid cells clustered based on cell type rather than tissue of origin or individual (Figure 1E). On the other hand, tissue-resident fibroblasts clustered into seven subtypes with diverse tissues of origin for each (Figure 1E). Notably, the majority of the 111 cell types exhibited high tissue specificity. For example, highly specialized cell types such as follicular cells, pneumocytes, and hepatocytes were restricted to only one tissue type, reflecting their tissue-specific functions (Figure 1E). Finally, we observed that the cell types we identified by sci-ATAC-seq are highly concordant with those identified by single-cell RNA-seq experiments on corresponding tissues (Data S1).

An atlas of cCREs in adult human cell types

To identify accessible chromatin regions in each of the 111 cell types, we aggregated chromatin accessibility profiles from all nuclei comprising each cell cluster and applied a peak calling procedure optimized for single-cell data (see STAR Methods and Methods S1). We then merged these accessible chromatin regions to obtain a list of 890,130 non-overlapping cCREs (Figure 2A). These cCREs covered 58.9% of the elements in the registry of cCREs published by the ENCODE consortium (Moore et al., 2020), and also included 420,152 previously unannotated elements (Figure S1A). To benchmark these cCREs, we next compared chromatin accessibility profiles between biosamples profiled by bulk DNase-seq and cell types identified by sci-ATAC-seq in the current study. In aggregate, sci-ATAC-seq cell types resembled primary cell type biosamples more closely than bulk tissue or immortalized cell line biosamples (Figure S1B), and prevalent cell types with higher tissue abundance defined by sci-ATAC-seq showed closer similarities to bulk DNase-seq biosamples than rare cell types did (Figure S1C). Out of the 111 cell types profiled in the current study, 44 (40%) did not show statistically significant correlation with any bulk biosample profiled by the ENCODE consortium (Figure S1D). Many of these cell types were rare: their median maximal tissue abundance was only 3.2%, and 36 (81.8%) of them constituted fewer than 10% of all cells in any tissue. Taken together, these findings suggest that our dataset contributes previously underrepresented cCREs from in vivo human cell types to existing catalogues, particularly from cell types with low abundance in bulk tissues.

To assess the potential function of these cCREs, we next compared them with catalogs of transgenic reporter-validated mammalian enhancers (Visel et al., 2007) and found that validated tissue-specific enhancers exhibited much higher chromatin accessibility in cell types comprising a large proportion of nuclei identified in the corresponding tissue (Figure 2B). For example, validated enhancers in heart showed higher average chromatin accessibility in atrial cardiomyocytes (z-score: 1.41) and ventricular cardiomyocytes (z-score: 1.43) compared with other cell types (Figure 2B), suggesting a good correlation between cell type-specific chromatin accessibility and tissue-specific enhancer activity. We further found that eQTLs from 49 adult tissue types (Consortium, 2020) were most commonly accessible in prevalent cell types, such as endothelial and smooth muscle cells. In addition, eQTLs from homogenous tissues, such as liver and thyroid, displayed strongest accessibility in the corresponding cell type which comprised a large proportion of nuclei identified in the tissue (Figure S2A-B). These results suggest that bulk tissue eQTLs best represent sequence variants associated with gene expression in abundant cell types and homogenous tissues, and may be less representative for rarer cell types within homogenous tissues or for unique cell types from heterogenous tissues.

We next categorized each cCRE based on distance to the nearest TSS as shown in Figure 2A. The majority (80.94%) of cCREs in the current catalogue resided more than 2,000 bp away from annotated TSSs. cCREs located directly over TSSs or near promoter regions displayed higher levels of sequence conservation and elevated chromatin accessibility (Figure 2C-D). By contrast, gene-distal cCREs were less accessible and showed larger variance relative to their accessibility (Figure 2D), suggesting the presence of shared programs of highly accessible promoter-proximal cCREs alongside variable programs of gene-distal cCREs across cell types and species. To further dissect cell-type specific chromatin signatures and regulatory programs, we applied an entropy-based strategy (Schug et al., 2005) to reveal 435,142 cCREs that demonstrated restricted accessibility in one or a few cell types (Figure 2E, see STAR Methods). We next applied GREAT ontology enrichment analysis and motif enrichment analysis on cell-type restricted cCREs to reveal putative biological processes and TFs of each cell type, which largely correlated with expected cell type-specific functions (FDR < 0.01). For instance, cCREs restricted to hepatocytes yielded biological process ontology terms such as steroid metabolic process (Figure 2F), and were enriched for the binding sites of hepatocyte nuclear factor TF family members HNF1A/B, HNF4A/G, and ONECUT1/2 (Figure 2G) (Costa et al., 2003).

Integrative analysis of adult and fetal chromatin accessibility

To examine transcriptional regulators and cCRE remodeling between fetal and adult stages, we re-processed data from a recent cell atlas of chromatin accessibility in 15 human fetal tissue types (Domcke et al., 2020) using the same quality control, clustering, and annotation strategies described above, which lead to the discovery of 111 fetal cell types and 802,025 cCREs (Figure S3, Table S3). Combining these cCREs with those identified from the adult cell types, we mapped a total of 1,154,611 distinct cCREs spanning 14.8% of the human genome in 222 fetal and adult cell types (Mendeley Data: 10.17632/yv4fzv6cnm.1). These cCREs covered 58.5% and 69.7% of the elements in the EpiMap (Boix et al., 2021) and the ENCODE cCRE registry (Moore et al., 2020), respectively. In addition, 34.8% and 51.0% of our cCREs were not annotated by the EpiMap and the ENCODE cCRE registry, respectively.

To compare the 222 fetal and adult cell types across the two atlases of chromatin accessibility, we utilized SnapATAC followed by batch-correction to obtain a low dimensional representation of the 1,323,041 nuclei from both fetal and adult tissues (Figure 3AB, see STAR Methods). We next performed phylogenetic analysis to place the fetal and adult cell types into different groups based on the distance defined in the low dimensional space (see STAR Methods, Figure S4A). In general, cell types belonging to different lineages separated into independent groups and harbored specific cCREs that were enriched for previously characterized lineage-specific TF motifs (Figure S4B). However, while many fetal cell types such as lymphoid, myeloid, and endothelial cells clustered near their adult counterparts in the tree, some cell types such as neurons and skeletal myocytes differed drastically between adult and fetal stages (Figure S4A), suggesting distinct cCRE usage by these cell types during development. To more systematically quantify differences in chromatin accessibility between adult and fetal cell types, we compared normalized accessibility across the list of 1,154,611 cCREs for each pair of fetal and adult cell types (Figure 3C-D, Figure S5). We found that fetal cell types such as immune and endothelial cells showed a relatively higher correlation with their adult counterparts than did other cell types such as neurons, glial cells, and skeletal myocytes (Figure 3D), consistent with the findings from our phylogenetic analysis. Together, these analyses suggest that the extent to which cCREs remodel to achieve developmental-stage specific functions varies greatly between human cell types.

To reveal the specific elements that may underlie fetal or adult-specific regulatory programs, we calculated life stage-specific cCREs for major cell groups which contained corresponding adult and fetal cell types (Figure 4A). Characterization of these elements revealed striking life stage-specific regulatory programs (Figure 4B-C). For example, skeletal myocytes differentiate substantially during pre and post-natal development (Chal and Pourquié, 2017) and showed lower global similarity between life stages than most other major cell types (Figure 3C-D). In total, we identified 72,648 differentially accessible (DA) cCREs between fetal and adult skeletal myocytes (Figure 4D). DA cCREs in fetal myocytes were associated with biological processes such as embryo development and response to wounding, and were enriched for motifs of myogenic regulatory TFs (MRFs) which orchestrate normal myogenesis (Mary Elizabeth Pownall et al., 2002) (Figure 4E-F), highlighting the role of these elements in regulating myogenic properties of fetal myocytes. On the other hand, adult skeletal myocyte DA cCREs were associated with biological processes related to muscle adaptation to contractile activity as well as insulin and steroid hormone response, and were enriched for MEF family members (P = 1e-424) and AP-1 complex members including FOSL1 (P = 1e-274) (Figure 4D-E), suggesting a role for these elements in regulating transcriptional responses to hormonal exposures and load bearing in adult skeletal muscle. In line with these ontology results and with established patterns of myosin isoform expression across the human lifespan (Schiaffino and Reggiani, 2011; Schiaffino et al., 2015; Stuart et al., 2016), we discovered DA cCREs at loci encoding marker genes of pre-natal myocytes including *MYH3* and *MYH8*, the heavy chains of embryonic and neonatal myosin respectively, as well as markers of type I (slow) and type II (fast) twitch adult myocytes including *MYH6/MYH7* and *MYH1/MYH2* respectively (Figure 4F). Taken together, these findings reveal the regulatory elements that may underlie the proliferative capacity and mature functionality of fetal and adult skeletal myocytes, respectively, and emphasize the value of this dataset alongside emerging human cell atlases collected at different timepoints along the lifespan for determining life stage-specific gene regulatory programs at cell type resolution.

Delineation of cell-type specificity of human cCREs

To characterize the cell-type specificity of cCREs across fetal and adult cell types, we organized the 1,154,611 cCREs into 150 clusters, referred to as *cis*-regulatory modules (CRMs), based on their normalized accessibility across the 222 cell types. While several CRMs displayed shared accessibility patterns across all cell types, most CRMs were limited either to single fetal or adult cell types or to groups of cell types that reflected shared cellular lineages (Figure 5A). To annotate putative functions of CRMs, we applied GREAT ontology enrichment analysis (McLean et al., 2010). Broadly, CRMs showing preferential accessibility in specific fetal and adult cell types were enriched for biological process ontology terms related to both cell type and life stage-specific cellular processes (FDR < 0.01) (Figure 5B-C).

To identify sequence features underlying these CRMs, we next measured the enrichment of 1,565 human TF motifs across the 150 CRMs to reveal putative master regulators of fetal and adult human cell types. This analysis revealed a comprehensive catalogue of fetal and adult cell and lineage-specific TF motifs. For example, a module with strong accessibility in adult CD8+ T cells and natural killer T cells was distinguished by enrichment for TBR, EOMES, and TBX TF family motifs (Module 8, P < 1e-84; Figure 5B-D), modules with strong accessibility in B cells were distinguished by enrichment for EBF family TF motifs (Module 13, P = 1e-27; Module 17, P = 1e-197) and a module with strong accessibility in adult mast cells was distinguished by GATA family member motif enrichment (Module 25, P = 1e-84) (Figure 5B-D). Further, the module with the strongest accessibility across all identified cell types was characterized by enrichment of the SP1 motif (Module 1, P = 1e-9180), consistent with the original description of SP1 as a regulator of ubiquitously-expressed housekeeping genes (Black et al., 2001). In addition to these well-characterized associations, we also report previously undefined TF associations with human cell types that are challenging to study in their in vivo tissue contexts: for example, motifs of the ESRR (Module 92, P = 1e-357; Module 93, P = 0.1) and FOX (Module 92, P = 1e-36; Module 93, P = 1e-255) TF family were preferably enriched in modules accessible in fetal (Module 92) and adult (Module 93) gastric epithelial cells respectively (Figure 5A), and motifs of the FOS and JUN families were enriched in modules accessible in fetal and adult adrenal cortical cells (Modules 135-138, P < 1e-10; Figure 5A).

Association of human cell types with complex traits and diseases

We next sought to use our 1.2 million cell type-resolved cCREs to interpret genetic variants associated with complex traits and multigenic disease phenotypes. We downloaded the NHGRI-EBI GWAS catalogue (Buniello et al., 2019) and retained 1,123 well-powered GWAS with 10 or more significant SNPs and over 20,000 cases (14% of 8,219 GWAS publications). We then used a hypergeometric test to measure the enrichment of trait-associated variants within cCREs identified from the 222 fetal and adult cell types. GWAS variants of 450 traits/diseases were found to be enriched in cCREs from at least one cell type (FDR < 0.1%) (Figure S6). As a comparison, EpiMap, a comprehensive enhancer catalogue comprising 833 epigenomic maps from bulk human tissue samples, primary cells and *ex vivo* cell lines (Boix et al., 2021), captured 457 GWAS studies (FDR < 0.1%) (Figure S6). For the 290 traits shared by both this study and EpiMap, our data captured the strongest GWAS enrichment in 74.8% of cases (217 of 290) and provided improved resolution by linking complex traits to specific cell type(s) (Figure S6). Further, for 160 additional traits, we were able to identify enrichments that were not detected in previous analyses (Figure S6), highlighting the added value of cell type-resolved cCREs maps.

The GWAS enrichment analysis above considered only index variants, *i.e.*, SNPs in genome-wide significant loci. However, the index variants may not represent the specific causal variants due to linkage disequilibrium (Schaid et al., 2018) and much of the heritability lies in SNPs with associations that do not reach genome-wide significance (Yang et al., 2010). We thus curated 240 GWAS studies with publicly available summary statistics and examined the enrichment of their associated SNPs within cCREs annotated in fetal and adult cell types using stratified linkage disequilibrium score regression (LDSC), a method for identifying functional enrichment from GWAS summary statistics using genome-wide information from all SNPs and explicitly modeling linkage disequilibrium (Finucane et al., 2015). This analysis revealed a total of 3,220 significant (FDR < 0.1) associations between fetal and adult cell types and human traits and disease phenotypes (Figure 6, Table S4). These enrichments revealed many expected cell type-disease phenotype relationships - for example, eczema risk variants were strongly enriched in adult T lymphocyte cCREs, atrial fibrillation risk variants were strongly enriched in both adult and fetal atrial and ventricular cardiomyocyte cCREs (FDR < 0.001), and thyroid stimulating hormone variants were enriched in follicular cell cCREs (Figure 6, Table S4). In addition to expected relationships, our analysis also revealed GWAS enrichment for human cell types not presently annotated by bulk DNase-seq or ATAC-seq data. These included a strong enrichment of coronary artery disease variants in adult vascular smooth muscle cCREs (FDR < 0.001) in addition to fetal and adult fibroblast, pericyte, and endothelial cell cCREs (FDR < 0.01), COPD variants in several adult stromal smooth muscle cell types (FDR < 0.01), triglyceride and HDL cholesterol level-associated variants in adult adipocyte cCREs (FDR < 0.01), and a nominal enrichment of ulcerative colitis variants in colon epithelial cell cCREs (P < 0.02). Interestingly, we detected substantial differences in the enrichment of disease and trait associated noncoding variants in subtypes of adult and fetal fibroblasts. These included a significant enrichment of variants associated with birth weight in fetal fibroblasts (FDR < 0.01) but not in adult fibroblasts (Table S4). Further, we detected differences in the enrichment of disease and trait variants in subtypes of adult fibroblasts, each of which displayed unique regulatory elements in addition to comparable chromatin accessibility at a set of core fibroblast cCREs (Figure S7). While all adult fibroblast populations were enriched for variants associated with standing height to an equivalent degree (FDR < 0.001), adult epithelial fibroblasts displayed a striking enrichment for variants associated with balding (FDR < 0.001) and only adult cardiac fibroblasts showed any enrichment for variants associated with myocardial fractal dimensions (FDR < 0.1; Table S4).

Systematic interpretation of molecular functions for noncoding risk variants

Many noncoding genetic variants enriched within cCREs from the analysis above are hypothesized to alter the expression of disease-associated genes by disrupting TF binding to *cis*-regulatory elements (Claussnitzer et al., 2020). To systematically interpret molecular mechanisms for the specific genetic variants associated with complex traits, we first applied the Activity-by-Contact (ABC) model (Fulco et al., 2019) to link the cCREs identified in 111 adult cell types to their target genes using our previously published promoter capture Hi-C data from 15 adult human tissues (Jung et al., 2019) (See STAR Methods). This analysis revealed 5,723,307 unique distal cCRE-to-gene linkages across the 111 adult cell types, with a median of 726,514 total linkages and 6,804 cell type-specific linkages per cell type (Figure S8). Second, we determined the probability that variants from 48 GWAS were causal for disease or trait association (Posterior probability of association, PPA) using Bayesian fine-mapping (Wakefield, 2009). We defined likely causal variants as those with a PPA > 0.1, and found that they were more likely to reside within cCREs than variants with low PPA (Figure S8A). Overall, we detected 3,096 likely causal variants residing within cCREs mapped in 111 adult human cell types (Figure 7A-B, Table S5), 2,096 of which were linked to putative target genes via the ABC model (Figure 7A, Table S5). Third, we applied our recently developed deltaSVM models for 94 TFs (Yan et al., 2021) to identify variants potentially disrupting binding by these regulators. This analysis revealed 527 TF binding sites predicted to be significantly altered by the likely causal variants (Figure 7A, Table S5). The intersection of these lists prioritized 361 likely causal variants that 1) resided within a human cell type cCRE, 2) significantly altered TF binding 3) and were linked to one or more target genes (Figure 7A-B, Table S5).

For example, one likely causal variant for ulcerative colitis (rs16940186) resided within an intergenic cCRE restricted to epithelial cells of the gastrointestinal tract, particularly colon epithelial cells, enterocytes, and goblet cells (Figure 7C). The cCRE containing rs16940186 was predicted to contact the TSS of *IRF8* (ABC score > 0.015), which encodes a TF involved in the regulation of immune cell maturation (Salem et al., 2020) and regulation of innate immunity in gastric epithelial cells (Yan et al., 2016). The rs16940186 risk allele is an eQTL associated with increased *IRF8* expression in human colon tissue and, consistent with these findings, deltaSVM models predicted this risk allele to create a binding site for the ETS family of activating TFs (Figure 7C), which are expressed in intestinal epithelia and have been suggested to regulate intestinal epithelial maturation (Jedlicka et al., 2009). One other prioritized likely causal risk variant for osteoarthritis (rs75621460) resided within a cCRE that was primarily accessible in immune cell types, was predicted to target the immunosuppressive cytokine gene *TGFB1,* and disrupted a binding site for the zinc-finger TF ERG1 (Figure 7D).

DISCUSSION

Detailed knowledge of the regulatory programs that govern gene expression in the human body has key implications for understanding human development and disease pathogenesis. Here, we used single-cell ATAC-seq to profile chromatin accessibility in 615,998 cells across 30 adult human tissues representing a wide range of human organ systems and integrated this dataset with single-cell chromatin accessibility data from human fetal tissues (Domcke et al., 2020). We mapped the state of activity for approximately 1.2 million cCREs across 222 fetal and adult cell types, bridging the key gap of cell type resolution in the annotation of candidate regulatory elements in the human genome. This work highlights the value of integrating human sci-ATAC-seq datasets from multiple sources and timepoints (Chiou et al., 2019; Domcke et al., 2020; Hocker et al., 2020; Wang et al., 2020) and, in the future, integration of these data along with new human single-cell datasets of increasing scale, breadth, and depth will enable a comprehensive understanding of gene regulatory features of human cell types throughout the lifespan.

While genome-wide association studies (GWAS) have been broadly used to enhance our understanding of polygenic human traits and reveal clinically-relevant therapeutic targets for complex diseases, to date the discovery of new variants has far outpaced our ability to interpret their molecular functions (Claussnitzer et al., 2020). Two central goals of the current study were thus to link individual human cell types to complex traits and to leverage cCRE maps to interpret the molecular functions of specific noncoding risk variants. By applying our datasets alongside cutting-edge methods to prioritize likely causal variants in LD, link distal cCREs to putative target genes, and predict motifs altered by risk variants, we revealed thousands of cell type-trait associations and created a framework to systematically interpret noncoding risk variants. For example, we highlight the likely causal ulcerative colitis-associated variant rs16940186. This risk variant may function to increase *IRF8* expression in gastrointestinal epithelial cells by creating a binding site for ETS family TFs in a GI epithelial-specific enhancer, and thereby alter the transcriptional responses of intestinal epithelial cells to inflammatory cytokines. Pending functional validation experiments, our results suggest that targeting *IRF8* in GI epithelial cells could be a potential therapeutic target for ulcerative colitis. As future GWAS in large cohorts with detailed phenotyping, whole genome sequencing efforts, and additional association studies employing long read technologies to capture structural variants become available, we anticipate that this combined resource and framework will be of continued utility for the interpretation of molecular functions for noncoding genetic variants. This resource thus lays the foundation for the analysis of gene regulatory programs across human organ systems at cell type resolution, and accelerates the interpretation of noncoding sequence variants associated with complex human diseases and phenotypes. The datasets can be accessed and explored at http://catlas.org/humanenhancer.

Limitations of the Study

The current study is still limited in several ways: firstly, we solely integrated data from two discrete life stages and in an incomplete sampling of organ systems. While we utilized tissue from anatomic sites corresponding directly to existing biosamples in large-scale databases (Carithers et al., 2015; Stranger et al., 2017), the size and diversity of adult human organ systems make it difficult to representatively sample them in their entirety. Additionally, our assay solely profiles chromatin accessibility in dissociated nuclei, and thus misses key orthogonal molecular and spatial information. Future assays that incorporate gene expression, chromatin accessibility, histone modifications, DNA methylation, chromosomal conformation, TF binding, and spatial information in the same single-cell will greatly enhance our understanding of gene regulation in human cell types (Zhu et al., 2020).

STAR METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and request for resources and reagents should be directed to and will be fulfilled by the lead contact, Bing Ren (biren@health.ucsd.edu).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- Single-nucleus ATAC-seq datasets generated in this study have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table and Table S2. This paper analyzes existing, publicly available data. These accession numbers for these datasets are listed in the key resources table. Raw data from Figures 1, 2, 4, 6, S4 and S6 were deposited on Mendeley at 10.17632/yv4fzv6cnm.1.
- All original code has been deposited at Github and is publicly available as of the date of publication. Links are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects—Adult human tissue samples were acquired by the ENTEx collaborative project (Stranger et al., 2017) via the GTEx collection pipeline (Carithers et al., 2015). Donor characteristics including age and sex are provided in Table S1. All human donors were deceased, and informed consent was obtained via next-of-kin consent for the collection and banking of deidentified tissue samples for scientific research. Donor eligibility requirements were as described previously (Carithers et al., 2015), and excluded individuals with metastatic cancer and individuals who had received chemotherapy for cancer within the prior two years.

METHOD DETAILS

Tissue feasibility testing for sci-ATAC-seq—Frozen tissue samples were sectioned on dry ice into two aliquots of equivalent mass. For nuclear isolation, one aliquot was subjected to manual pulverization via mortar and pestle while submerged in liquid nitrogen, and the other aliquot was homogenized in a gentleMACS M-tube (Miltenyi) on a gentleMACS Octo Dissociator (Miltenyi) using the “Protein_01_01” protocol in MACS buffer (5 mM CaCl2, 2 mM EDTA, 1X protease inhibitor (Roche, 05-892-970-001), 300 mM MgAc, 10 mM Tris-HCL pH 8, 0.6 mM DTT) and pelleted with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). Pulverized frozen tissue and pelleted nuclei from gentleMACS M-tubes were each split into two further aliquots. One aliquot from each of the two nuclear isolation conditions was then resuspended in 1 mL Nuclear Permeabilization

Buffer (1X PBS, 5% Bovine Serum Albumin, 0.2% IGEPAL CA-630 (Sigma), 1 mM DTT, 1X Protease inhibitor), and the other aliquot from the same nuclear isolation condition was resuspended in 1 mL OMNI Buffer (10mM Tris-HCL (pH 7.5), 10mM NaCl, 3mM MgCl2, 0.1% Tween-20 (Sigma), 0.1% IGEPAL-CA630 (Sigma) and 0.01% Digitonin (Promega) in water), yielding a total of four nuclear isolation/nuclear permeabilization buffer conditions tested for each tissue type. Nuclei were rotated at 4 °C for 5 minutes before being pelleted again with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). After centrifugation, permeabilized nuclei were resuspended in 500 μL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer. Concentration was adjusted to 2,000 nuclei/9 μl, and 2,000 nuclei were dispensed 12 wells of a 96-well plate per nuclear isolation/permeabilization condition (samples were processed in batches of 4 nuclear isolation/permeabilization conditions per 2 different tissue samples). For tagmentation, 1 μL barcoded Tn5 transposomes (Table S6) were added using a BenchSmart™ 96 (Mettler Toledo), mixed five times, and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 μL of 40 mM EDTA (final 20mM) were added to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 μL of 2x sort buffer (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All 12 wells from each nuclear isolation/permeabilization condition were combined into a separate FACS tube, and stained with Draq7 at 1:150 dilution (Cell Signaling). For each nuclear isolation/permeabilization condition, we used a SH800 (Sony) to sort four wells containing 0 nuclei per well and four wells containing 80 nuclei per well into one 96-well plate (total of 768 wells) containing 10.5 μL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma)). After addition of 1 μL 0.2% SDS using a BenchSmart™ 96 (Mettler Toledo), the 96 well plate was incubated at 55 °C for 7 min with shaking (500 rpm). 1 μL 12.5% Triton-X was added to each well to quench the SDS. Next, 12.5 μL NEBNext High-Fidelity 2× PCR Master Mix (NEB) were added to each well and samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72°C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were assayed for DNA library concentration using the PerfeCTa NGS Quantification RT-qPCR Kit (Quanta Biosciecnces) according to manufacturer's protocols, and subsequently returned to the thermal cycler for a second round of PCR amplification (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72°C 60 s) × 4 cycles, held at 12 °C). After the second PCR amplification, for each nuclear isolation/permeabilization condition, wells containing 0 nuclei were combined and wells containing 80 nuclei were combined. The resulting DNA libraries were purified according to the MinElute PCR Purification Kit manual (Qiagen) and size selection was performed with SPRISelect reagent (Beckmann Coulter, 0.55x and 1.5x). Final libraries were quantified using a Qubit fluorimeter (Life technologies) and a nucleosomal pattern of fragment size distribution was verified using a Tapestation (High Sensitivity D1000, Agilent). We calculated a signal to noise ratio for final feasibility test libraries using LightCycler® 480 SYBR Green I Master Mix (Roche) along with custom primers for the promoter of human *GAPDH* (5'-CATCTCAGTCGTTCCCAAAGT-3', 5'-TTCCCAGGACTGGACTGT-3') and a heterochromatic gene desert region (5'-CCCAAACTCTGAGAGGCTTATT-3', 5'-GAGCCATCATCTAGACACCTTC-3'). For each tissue type, the nuclear isolation/permeabilization condition that resulted in optimized nuclear yield (nuclei/mg tissue), library concentrations > 50 pM per 80 sorted nuclei, nucleosomal distribution pattern of fragments, and a $\log_2$(signal to noise ratio) > 3.3 was selected for combinatorial indexing-assisted single nucleus ATAC-seq (Table S1).

Combinatorial indexing-assisted single nucleus ATAC-seq—Combinatorial indexing-assisted single nucleus ATAC-seq was performed as described previously (Preissl et al., 2018) with slight modifications (Hocker et al., 2020) and using new sets of oligos for tagmentation and PCR (Table S6). Nuclei were isolated and permeabilized according to the optimized conditions from feasibility testing (Table S1). After resuspension in permeabilization buffer, nuclei were rotated at 4 °C for 5 minutes before being pelleted again with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). After centrifugation, permeabilized nuclei were resuspended in 500 µL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer. Concentration was adjusted to 2,000 nuclei/9 µl, and 2,000 nuclei were dispensed into each well of a 96-well plate per sample (96 tagmentation wells/sample, samples were processed in batches of 2-4 samples). For tagmentation, 1 µL barcoded Tn5 transposomes (Table S6) were added using a BenchSmart™ 96 (Mettler Toledo), mixed five times, and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA (final 20mM) were added to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 µL of 2x sort buffer (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells were combined into a separate FACS tube for each sample, and stained with Draq7 at 1:150 dilution (Cell Signaling). Using a SH800 (Sony), 20 nuclei per sample were sorted per well into eight 96-well plates (total of 768 wells) containing 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma)). Preparation of sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter). After addition of 1 µL 0.2% SDS, samples were incubated at 55 °C for 7 min with shaking (500 rpm). 1 µL 12.5% Triton-X was added to each well to quench the SDS. Next, 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix (NEB) were added and samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72°C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries were purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRISelect reagent (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRISelect reagent (Beckman Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life technologies) and a nucleosomal pattern of fragment size distribution was verified using a Tapestation (High Sensitivity D1000, Agilent). Libraries were sequenced on a NextSeq500 or HiSeq4000 sequencer (Illumina) using custom sequencing primers with following read lengths: 50 + 10 + 12 + 50 (Read1 + Index1 + Index2 + Read2). Primer and index sequences are listed in Table S6.

Demultiplexing of single nucleus ATAC-seq sequencing reads—For each sequenced single nucleus ATAC-Seq library, we obtained four FASTQ files, two for paired end DNA reads and two for the combinatorial indexes for i5 and T7 (768 and 364 indices, respectively). We selected all reads with up to 2 mismatches per i5 and T7 index (Hamming distance between each pair of indices is 4) and integrated the concatenated barcode at the beginning of the read name in the demultiplexed FASTQ files. The customized scripts can be found at: https://gitlab.com/Grouumf/ATACdemultiplex/.

Quality control metrics: TSS enrichment and unique fragments—TSS positions were obtained from the GENCODE database v31 (Frankish et al., 2019). Tn5 corrected insertions were aggregated ±2000 bp relative (TSS strand-corrected) to each unique TSS genome wide. Then this profile was normalized to the mean accessibility ± (1900 to 2000) bp from the TSS and smoothed every 11 bp. The max of the smoothed profile was taken as the TSS enrichment. We then filtered out all single-cells that had fewer than 1,000 unique fragments and/or a TSS enrichment of less than 7 for all data sets.

Overall clustering strategy—We utilized multiple rounds of clustering analysis to identify cell clusters. The first round of clustering analysis was performed on individual samples. We divided the genome into 5kb consecutive windows and then scored each cell for any insertions in these windows, generating a window by cell binary matrix for each sample. We filtered out those windows that are generally accessible in all cells for each sample using z-score threshold 1.65. Based on the filtered matrix, we then carried out dimension reduction followed by graph-based clustering to identify cell clusters. We called peaks for each cluster using the aggregated profile of accessibility and then merged the peaks from all clusters to generate a union peak list. Based on the peak list, we generated a cell-by-peak count matrix and used Scrublet (Wolock et al., 2019) to remove potential doublets. Next, to carry out the second round of clustering analysis, we merged peaks called from all samples to form a reference peak list. We then generated a single binary cell-by-peak matrix using cells from all samples and again performed the dimension reduction followed by graph-based clustering to obtain the major cell groups across the entire dataset. To further dissect cell-type heterogeneity within the major cell groups, we then performed another round of clustering analysis for each of the identified major cell group to identify subclusters.

Doublet removal—We applied Scrublet to the cell-by-peak count matrix with default parameters. Doublet scores returned by Scrublet were then used to fit a two-component Gaussian mixture model using the “BayesianGaussianMixture” function from the python package “scikit-learn”. The component with larger mean doublet score is presumably formed by doublets and cells belonging to it were removed from downstream analysis.

Dimension reduction—To find the low-dimensional manifold of the single-cell data, we adapted our previously published method, SnapATAC (Fang et al., 2020), to reduce the dimensionality of the peak by cell count matrix. The previous iteration of SnapATAC utilized spectral embedding for dimension reduction. To increase scalability of spectral embedding, we applied the Nyström method (Bouneffouf and Birol, 2016) for handling large datasets. Specifically, we first randomly sampled 35,000 cells as the training data. We then computed the Jaccard index between each pair of cells in the training set and constructed the similarity matrix S. We computed the matrix $P = D^{-1}S$, where D is the diagonal matrix such that $D_{ii} = \Sigma_j S_{ij}$. The eigendecomposition was performed on P and the eigenvector with eigenvalue 1 was discarded. From the rest of the eigenvectors, we took the first 30 of them corresponding to the largest eigenvalues as the spectral embedding of the training data. We utilized the Nyström method to extend the embedding to the data outside the training set. Given a set of unseen samples, we computed the similarity matrix S' between the new samples and the training set. The embedding of the new samples is given by $U' = S'U\Lambda^{-1}$, where U and Λ are the eigenvectors and eigenvalues of P obtained in the previous step.

Correction of Batch Effects—We performed batch correction for each tissue separately. Inspired by the mutual nearest neighbor batch-effect-correction method (Haghverdi et al., 2018), we developed a variant using mutual nearest centroids to iteratively correct for batch effects in multiple donor samples. Specifically, after dimension reduction we performed k-means clustering on individual replicate or donor sample with k equal to 20. We choose this number because the number of major clusters in a given tissue sample is typically less than 20. We then computed the centroid for each cluster and identified pairs of mutual nearest centroids across different batches. These mutual nearest centroids were used as the anchors to match the cells between different batches and correct for batch effects as described previously (Haghverdi et al., 2018). We found that the result can be further improved by performing above steps iteratively. However, too many iterations may lead to over-correction. We therefore used two iterations in this study.

Graph-based clustering algorithm—We constructed the k-nearest neighbor graph (k-NNG) using low-dimensional embedding of the cells with k equal to 50. We then applied the Leiden algorithm (Traag et al., 2019) to find communities in the k-NNG corresponding to cell clusters. The Leiden algorithm can be configured to use different quality functions. The modularity model is a popular choice but it suffers from the issue of resolution-limit, particularly when the network is large (Traag et al., 2011). Therefore, we used the modularity model only in the first round of clustering analysis to identify initial clusters. In the final round of clustering, we chose the constant Potts model as the quality function since it is resolution-limit-free and is better suited for identifying rare populations in a large dataset (Traag et al., 2011). To determine the optimal number of clusters, we varied the resolution parameter in the Leiden algorithm and computed the clustering stability and average silhouette score under each resolution. Cluster stability was defined as the consistency, measured by the average adjusted rand index, of results from five independent clustering analyses on perturbed inputs. The perturbation was introduced in a way that 2% of the edges were randomly selected and subjected to removal. We selected the resolution that leads to both high average silhouette score and high clustering stability as well as biological considerations, e.g., number of known cell types in the tissue, marker gene accessibility.

Peak calling and peak filtering—For each cell cluster, initial peak calling was performed on Tn5-corrected single-base insertions (each end of the Tn5-corrected fragments) using the MACS2 (Zhang et al., 2008) callpeak command with parameters “–shift −100 –extsize 200 –nomodel –call-summits –nolambda – keep-dup all”, filtered by the ENCODE hg38 blacklist (accession: ENCFF356LFX). Due to the varying abundance of cell types in each tissue, single-cell assays typically profile different cell types at different sequencing depths. To account for these differences, we adapted peak calling cutoffs to different sequencing depths. Specifically, we choose a cutoff of FDR less than 0.1, 0.05, 0.025, 0.01, and 0.001, corresponding to the situations when the number of reads is with then range of 0-5 million, 5-25 million, 25-50 million, 50-100 million, and 100 million and above. Using simulated datasets, we found that this procedure achieved good balance between the sensitivity and specificity for detecting peaks under different sequencing depths. Next, based on the chromatin accessibility at the single cell level, we developed a peak filtering procedure to further reduce the false positive rate by retaining only those peaks that were accessible in a significant fraction of the cells compared to background regions. To do so, we first randomly selected 1 million regions from the genome and for each of these regions we calculated the fraction of cells that are accessible. These were used to fit a beta distribution as the null model. We then computed the fraction of accessible cells and its significance level for each candidate peak identified by MACS2. Candidate peaks with FDR < 0.01 were included in the final peak list.

Generating the union peak set—To compile a union peak set, we combined peaks from all clusters and extended the peak summits by 200 bp on either side. Overlapping peaks were then handled using an iterative removal procedure. First, the most significant peak, *i.e.*, the peak with the smallest p-value, was kept and any peak that directly overlapped with it was removed. Then, this process was iterated to the next most significant peak and so on until all peaks were either kept or removed due to direct overlap with a more significant peak.

Computing relative accessibility scores—We define an accessible locus as the minimal genomic region that can be bound and cut by the Tn5 enzyme. We use $L \subset N$ to represent the set of all accessible loci. We further define a pseudo-locus as the set of accessible loci that relates to each other in certain meaningful way (for example, nearby loci, loci from different alleles). In this example, pseudo-loci correspond to peaks. We use $\{d_i \setminus d_i \subset L\}$ to represent the set of all pseudo-loci. Let a_l be the accessibility of accessible locus l, where $l \in L$. We define the accessibility of pseudo-locus d_i as $A_i = \Sigma_{k \in di}\, a_k$, *i.e.*, the sum of accessibility of accessible loci associated with di. Let C_j be the library complexity (the number of distinct molecules in the library) of cell j. Assuming unbiased PCR amplification, then the probability of being sequenced for any fragment in the library is: $s_j = 1 - (1 - \frac{1}{C_j})k_j$, where k_j is the total number of reads for cell j. If we assume that the probability of a fragment present in the library is proportional to its accessibility and the complexity of the library, then we can deduce that the probability of a given locus l in cell j being sequenced is: $p_{lj} \propto a_l C_j S_j$. For any pseudo-locus d_i, the number of reads in d_i for cell j follows a Poisson binomial distribution, and its mean is $m_{ij} = \Sigma_{k \in d_i} P_{kj} \propto C_j S_j$ $\Sigma k \in d_i\, a_k = C_j S_j A_i$. Given a pseudo-locus (or peak) by cell count matrix 0, we have: $\Sigma_j O_{ij}$ $= \Sigma_j m_{ij}$. Therefore, $A_i = Z \frac{\sum_j O_{ij}}{\sum_j C_j s_j}$, where Z is a normalization constant. When comparing across different samples the relative accessibility may be desirable as they sum up to a constant, *i.e.*, $\Sigma_i A_i = 1 \times 10^6$. In this case, we can derive $A_i = \frac{\sum_j O_{ij}}{\sum_{ij} O_{ij}} * 10^6$.

Assigning cell types to cell clusters—To annotate the cell clusters, we first curated a set of marker genes from the PanglaoDB (Franzén et al., 2019) corresponding to expected cell types. We aggregated open chromatin fragments from each cluster and utilized the promoter accessibility, defined as RPM of +/− 1kb around TSS, as the proxy for gene activity. We then computed the raw cell type enrichment score as the logarithm of the geometric mean of marker genes' activity. The final enrichment scores were obtained by applying two rounds of z-score transformation, first across cell types and then across cell clusters, on raw enrichment scores. For each cluster, we picked the cell type that showed strongest enrichment to make initial assignments. Finally, we manually reviewed these assignments and made adjustments based on focused consideration of marker gene accessibility in conjunction with information about tissue(s) of origin.

Identification of cell type-restricted peaks—We used a Shannon entropy-based method (Schug et al., 2005) to identify cell type-specific peaks. Given the relative accessibility scores of a peak across clusters, we first converted the scores to probabilities: $p_i = q_i/\Sigma_i q_i$. The entropy was then calculated by: $H_p = -\Sigma_t p_t \log_2(p_t)$. The specificity score is $Q_{p|t} = H_p - \log_2(p_t)$. To estimate the statistical significance of specificity scores, we assumed that under the null hypothesis each peak has an average accessibility level across all cell types and that the log base 2 of the cell-type-dependent fold changes from the average level follow a normal distribution with mean equal to zero and standard deviation **s**. The value of **s** was estimated using the top 50% least variable peaks, and 500,000 samples were then drawn to form the empirical distribution of Q_p that are used to determine the p-values of specificity scores. The cell-type-restricted peaks were then identified using a p-value cutoff of 0.025.

Cell-type enrichment analysis of fine-mapped GTEx eQTLs—The fine-mapped eQTLs (GTEx Analysis V8) in each of the 49 tissues or cell lines were downloaded from the GTEx portal (https://gtexportal.org). For each tissue, we first identified the overlapping cCREs with its eQTLs. We then calculated the average of log-transformed accessibility scores of these peaks in each of the 111 cell types. This yielded a tissue by cell-type table containing raw cell-type enrichment scores of eQTLs from each tissue. The raw enrichment scores were then normalized row-wise using z-score transformation. For each tissue, we define the maximum cell-type enrichment as the largest value of z-scores across 111 cell types. In general, we found that homogenous tissues tend to have higher maximum cell-type enrichment than tissues that are more heterogenous.

Differential peak analysis—To carry out differential peak analysis between foreground set and background set, we first removed all peaks with fold changes of relative accessibility less than 2. For each peak, we then built a full model and a reduced model.

$$\log\frac{P_{full}}{1 - P_{full}} = \beta_0 + \beta_1 r + \beta_2 c$$

$$\log\frac{P_{reduced}}{1 - P_{reduced}} = \beta_0 + \beta_1 r$$

$P_{reduced}$ and P_{full} represent the likelihood of the reduced model and full model respectively. r contains the logarithm of number of fragments. c is categorical variable indicating if the cell comes from foreground or background. We then used a likelihood ratio test framework to determine whether the full model provided a significantly better fit of the data than the reduced model. We selected the sites using a 5% FDR threshold (Benjamini-Hochberg method).

Identification of fibroblast core signature and subtype-specific signatures— We first performed pairwise differential peak analysis for the seven fibroblast subtypes. We then defined fibroblast core signature as peaks that are shared by all subtypes and were not called as differentially accessible in any of the pairwise comparison. Likewise, we defined the specific signature for a subtype as peaks that are differentially more accessible in the given subtype for every pairwise comparison.

Measuring the similarity of chromatin accessibility profiles between cell types identified by sci-ATAC-seq and bulk biosamples—We downloaded bulk DNase-seq data from the ENCODE portal. We excluded samples collected at embryonic stage or originated from kidney, bladder or brain tissues, as we did not perform experiments on those tissues. As a result, 638 datasets were kept for downstream analysis. For each of the DNase-seq datasets, we calculated its Pearson correlation coefficient with 111 identified cell types based on RPKM values at identified cCREs. These correlation scores were then scaled using z-score transformation across 111 cell types. We used the maximum of scaled correlation scores to represent each biosample’s overall similarity with sci-ATAC-seq cell types.

Identification of cCRE modules—A cCRE module is defined as co-accessible regions that share similar accessibility pattern across cell types. To identify cCRE modules, we first performed quantile normalization on the log2 transformed matrix containing accessibilities of 1,154,611 cCREs in 222 fetal and adult cell types. For each cCRE, we then divided its accessibility vector by the L2 norm, which allowed us to better extract the accessibility pattern from the data. Next we applied the k-mean algorithm to this matrix to identify clusters of cCREs. Using the “elbow” method, we determined the number of clusters to be 150.

Motif enrichment analysis—We measured the enrichment of 1565 human TF motifs consisting of the JASPAR (2018) core non-redundant vertebrate motifs, the HOCOMOCO v1156 human motif set and the SELEX motifs by Jolma et al.. We computed the enrichments for each of the 1565 motifs relative to a joint cCRE background and filtered the list using FDR cutoff 0.01. For each motif. We reported the motif with the highest enrichment for each of the 286 previously identified motif archetypes (Vierstra et al., 2020).

Identification of candidate driver TFs—We used the Taiji pipeline (Zhang et al., 2019) to identify candidate driver TFs in each cell cluster. Briefly, for each cell type cluster, we constructed the TF regulatory network by scanning TF motifs at the accessible chromatin regions and linking them to the nearest genes. The network is directed with edges from TFs to target genes. The genes’ weights in the network were determined based on the relative accessibility of their promoters. The weights of the edges were calculated by the relative accessibility of the promoters of the source TFs. We then used the personalized PageRank algorithm to rank the TFs in the network.

Integration of adult and fetal datasets—To integrate our dataset with the recent cell atlas of fetal chromatin accessibility (Domcke et al., 2020), we downloaded the fragment files for 63 fetal samples spanning 15 tissues and converted the genomic coordinates from GRC37 (hg19) to GRCh38 using the UCSC liftOver tool. We then performed the quality control, cell filtering and cell clustering using the same pipeline described above and identified 111 fetal cell types. Next, we combined the QC passed cells from adult and fetal datasets and performed the joint embedding using the SnapATAC algorithm. We considered fetal or adult cells as belonging to different batches, and used a linear model to remove technical batch effects for each dimension in the reduced dimensional space. Using these batch-corrected lower-dimensional representations, we applied the UMAP algorithm to visualize the cells in a 2D space and used the FASTME algorithm (Guindon and Gascuel, 2003) to construct the phylogenetic tree for adult and fetal cell types.

Differential peak analysis between fetal and adult cells—To perform differential peak analysis between fetal and adult samples, we modified the likelihood-ratio test framework described above to account for technical batch effects between two datasets. We started with three set of cells. The first two sets of cells corresponded to foreground and background sets that are subject to the differential test. The third set was the auxiliary set corresponding to remaining cells that were not from the first two sets. The auxiliary set served as a proxy to estimate the batch effects. For instance, when performing differential test between two sub-trees of the phylogenetic tree of fetal and adult cell types, for each sub-tree we randomly sampled an equal number of cells for each cell type in the sub-tree. The cells sampled from one branch were considered as foreground and those from the other were considered as background. The remaining cells did not belong to the two sub-trees form the auxiliary set. For each peak, we then built a full model and a reduced model.

$$\log\frac{P_{full}}{1-P_{full}} = \beta_0 + \beta_1 r + \beta_2 s + \beta_3 t + \beta_4 c$$

$$\log\frac{P_{reduced}}{1-P_{reduced}} = \beta_0 + \beta_1 r + \beta_2 s + \beta_3 t$$

$P_{reduced}$ and P_{full} represent the likelihood of the reduced model and full model respectively. r contains the logarithm of number of fragments. s is a categorical variable indicating whether the cell comes from the fetal tissue or the adult tissue. t indicates whether the cell comes from the auxiliary set. c indicates whether the cell comes from foreground set. We then used a likelihood ratio test framework to determine whether the full model provided a significantly better fit of the data than the reduced model. We selected the sites using a 1% FDR threshold (Benjamini-Hochberg method).

Generation of bigwig tracks—Each Tn5-corrected insertion was extended in both directions by 100 bp to form a 200-bp fragment. We then counted the number of fragments overlapping with each base on the genome and generated a bedgraph file. The bedgraph file was converted to bigwig file using the “bedGraphToBigWig” tool.

Linking cCREs to target genes—We downloaded the chromosome interactions called from published promoter capture Hi-C data in 14 human tissues (Jung et al., 2019). In each tissue, we first filtered the chromosome interactions using a lenient p-value cutoff of 0.1. We then created the chromosome interaction matrix using the normalized interaction frequency. The interaction matrices from 14 tissues were then averaged to get the final interaction matrix. We applied the Activity-by-Contact (ABC) Model (Fulco et al., 2019) to compute the ABC Score for each cCRE-gene pair as the product of Activity (chromatin accessibility) and Contact (interaction frequency), normalized by the product of Activity and Contact for all other cCREs. We retained all distal cCRE-gene connections with an ABC score greater than 0.015.

Stratified linkage disequilibrium (LD) score regression—We used LD score regression (Bulik-Sullivan et al., 2015) v1.0.1 to estimate genome-wide GWAS enrichment for disease and non-disease phenotypes within cell type resolved cCREs (peaks called on each cell cluster via MACS2 (Zhang et al., 2008) using the above parameters). We compiled published GWAS summary statistics for complex diseases (Bentham et al., 2015; Bronson et al., 2016; Consortium, 2019; Cordell et al., 2015; Jansen et al., 2019; Ji et al., 2017; Jin et al., 2016; Luo et al., 2017b; Mahajan et al., 2018; Malik et al., 2018; Michailidou et al., 2017; Nielsen et al., 2018; Nikpay et al., 2015; Okada et al., 2014; Paternoster et al., 2015; Pividori et al., 2019; Sakornsakolpat et al., 2019; Schafmayer et al., 2019; Shadrina et al., 2019; Tachmazidou et al., 2019; Tin et al., 2019; Watanabe et al., 2019; Wiberg et al., 2019; Wuttke et al., 2019) and endophenotypes (Astle et al., 2016; Hoffmann et al., 2018; Kemp et al., 2017; Kilpeläinen et al., 2016; Manning et al., 2012; Saxena et al., 2010; Shrine et al., 2019; Strawbridge et al., 2011; Teumer et al., 2018; Warrington et al., 2019) within European populations. Using cell type resolved cCREs as a binary annotation, we created custom partitioned LD score files by following the steps outlined in the LD score estimation tutorial. As background annotations, we included all baseline annotations in the baseline-LD model v1.2 as well as partitioned LD scores created from all merged cCREs. For each trait, we used LD score regression to then estimate coefficient p-value for each cell type relative to the background annotations and used the Benjamini-Hochberg procedure to correct for multiple tests.

GWAS enrichment analysis—We downloaded the NHGRI-EBI GWAS catalogue (Buniello et al., 2019) (downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads on July 7, 2021) and pruned the catalogue using an approach described previously (Boix et al., 2021). Specifically, for each trait and PMID combination, we ranked associations by their significance (P value) and added SNPs iteratively if they were not within 5 kb of previously added SNPs. We then compiled a compendium of 1,123 well-powered GWAS with 10 or more significant SNPs and over 20,000 cases (14% of 8,219 GWAS publications) that capture over 81,057 GWAS loci.

For each cell type and trait combination, we computed the number of intersections between trait associated SNPs and cell-type associated cCREs. We compared this number with the number of intersections between SNPs and the entire set of cCREs from all cell types, using a hypergeometric test to evaluate the statistical significance of enrichments. To estimate the false discovery rate, we generated 1,000 null GWAS with the same lead SNP set size by randomly shuffling the trait associations across GWAS locations. We then computed the null association P values for each permuted GWAS and used the 0.1% top quantiles as the cut-off.

Fine mapping—We performed genetic fine mapping for GWAS of diseases and endophenotypes that had sufficient coverage (i.e., were at least imputed into 1000 Genomes). For GWAS with available fine mapping data, we took 99% credible sets directly from the supplemental tables. For GWAS without available fine mapping data, we calculated approximate Bayes factors (Wakefield, 2009) (ABF) for each variant assuming prior variance $\omega = 0.04$. For every trait, we obtained index variants for each locus from the supplemental tables of the respective study. We extracted all variants in at least low linkage disequilibrium ($r2 > 0.1$ using the European subset of 1000 Genomes Phase 3 (Auton et al., 2015)) in a large window (±2.5 Mb) around each index variant. We calculated posterior probabilities of association (PPA) for each variant by dividing its ABF by the cumulative ABF for all variants within the locus. We then defined 99% credible sets for each locus by sorting variants by descending PPA and keeping variants adding up to a cumulative PPA of 0.99.

Predicting the effects of noncoding variants on TF binding—To identify SNPs that affect TF binding, we employed deltaSVM models as described previously (Yan et al., 2021). Briefly, 40 bp sequences centered on each SNP were used as input to 94 previously trained and validated TF models. For each SNP, we predicted the binding scores for both alleles by running "gkmpredict". A SNP was considered to be bound if the binding score passed the pre-defined threshold for either allele. Among those SNPs, deltaSVM scores were calculated using the "deltasvm.pl" script and SNPs with deltaSVM scores passing the threshold for the corresponding model are predicted to affect TF binding.

External genome browser track data—Genome browser tracks displaying ChIP-seq and DNase-seq signal from bulk transverse colon datasets and human primary T cell datasets were downloaded from ENCODE with the following identifiers: ENCSR340MRJ, ENCSR557OWY, ENCSR500QVK, ENCSR792VLP, ENCSR627UDJ, ENCSR902BOX, ENCSR218OEZ, ENCSR222QLW.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters were reported either in individual figures or corresponding figure legends. Statistical details of experiments can be found in “METHOD DETAILS”. All statistical analyses were performed in either R or Python.

ADDITIONAL RESOURCES

The raw data and analyzed results are available at our interactive web portal: http://catlas.org/humanenhancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the ENCODE consortium, in particular Mike Pazin (NHGRI) and Idan Gadbank (Stanford), Kristin Ardlie (Broad Institute) and Ellen Gelfand (Broad Institute), for providing the human tissue samples for the present study. We thank B. Li for bioinformatics support. We thank S. Kuan for sequencing support. We thank B. Chen for valuable discussion and feedbacks in earlier versions of the manuscript. We thank the QB3 Macrolab at UC Berkeley for purification of the Tn5 transposase. This work was supported by the Ludwig Institute for Cancer Research (to B.R.), National Human Genome Research Institute (3U54HG006997-04S2 to B.R.) and Foundation for the National Institutes of Health (AMP T2D RFP 14 to K.J.G). J.D.H. was supported in part by a Ruth L. Kirschstein Institutional National Research Service Award T32 GM008666 from the National Institute of General Medical Sciences. Work at the UCSD Center for Epigenomics was supported by the UC San Diego School of Medicine.

REFERENCES

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461. [PubMed: 24670763]

Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415–1429.e1419. [PubMed: 27863252]

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR (2015). A global reference for human genetic variation. Nature 526, 68–74. [PubMed: 26432245]

Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen L, et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet 47, 1457–1464. [PubMed: 26502338]

Black AR, Black JD, and Azizkhan-Clifford J (2001). Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer. Journal of Cellular Physiology 188, 143–160. [PubMed: 11424081]

Boix CA, James BT, Park YP, Meuleman W, and Kellis M (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature 590, 300–307. [PubMed: 33536621]

Bouneffouf D, and Birol I (2016). Theoretical analysis of the Minimum Sum of Squared Similarities sampling for Nyström-based spectral clustering. In 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3856–3862.

Bronson PG, Chang D, Bhangale T, Seldin MF, Ortmann W, Ferreira RC, Urcelay E, Pereira LF, Martin J, Plebani A, et al. (2016). Common variants at PVT1, ATG13-AMBRA1, AHI1 and CLEC16A are associated with selective IgA deficiency. Nat Genet 48, 1425–1429. [PubMed: 27723758]

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218. [PubMed: 24097267]

Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM, and Schizophrenia Working Group of the Psychiatric Genomics, C. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics 47, 291–295. [PubMed: 25642630]

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012. [PubMed: 30445434]

Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, et al. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank 13, 311–319. [PubMed: 26484571]

Carter B, and Zhao K (2020). The epigenetic basis of cellular heterogeneity. Nature Reviews Genetics.

Chal J, and Pourquié O (2017). Making muscle: skeletal myogenesis *in vivo* and *in vitro*. Development 144, 2104–2122. [PubMed: 28634270]

Chiou J, Zeng C, Cheng Z, Han JY, Schlichting M, Huang S, Wang J, Sui Y, Deogaygay A, Okino M-L, et al. (2019). Single cell chromatin accessibility reveals pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. bioRxiv, 693671.

Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, et al. (2020). A brief history of human disease genetics. Nature 577, 179–189. [PubMed: 31915397]

Consortium, G. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318. [PubMed: 32913098]

Consortium, I.M.S.G. (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science 365.

Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Frésard L, Granja JM, Louie BH, Eulalio T, Shams S, Bagdatli ST, et al. (2020). Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet 52, 1158–1168. [PubMed: 33106633]

Cordell HJ, Han Y, Mells GF, Li Y, Hirschfield GM, Greene CS, Xie G, Juran BD, Zhu D, Qian DC, et al. (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun 6, 8019. [PubMed: 26394269]

Costa RH, Kalinichenko VV, Holterman AX, and Wang X (2003). Transcription factors in liver development, differentiation, and regeneration. Hepatology 38, 1331–1347. [PubMed: 14647040]

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, and Shendure J (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910. [PubMed: 25953818]

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309–1324.e1318. [PubMed: 30078704]

Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al. (2020). A human cell atlas of fetal chromatin accessibility. Science 370, eaba7612. [PubMed: 33184180]

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49. [PubMed: 21441907]

Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F, et al. (2020). SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq. bioRxiv, 615179.

Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F, et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun 12, 1337. [PubMed: 33637727]

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228–1235. [PubMed: 26414678]

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47, D766–D773. [PubMed: 30357393]

Franzén O, Gan L-M, and Björkegren J (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database The Journal of Biological Databases and Curation 2019, 46.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Patwardhan TA, Nguyen TH, et al. (2019). Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. bioRxiv, 529990.

Furtado MB, Costa MW, Pranoto EA, Salimova E, Pinto AR, Lam NT, Park A, Snider P, Chandran A, Harvey RP, et al. (2014). Cardiogenic genes expressed in cardiac fibroblasts contribute to heart development and repair. Circulation research 114, 1422–1434. [PubMed: 24650916]

Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, Dahmani A, Lameiras S, Reyal F, Frenoy O, et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet 51, 1060–1066. [PubMed: 31152164]

Guindon S, and Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696–704. [PubMed: 14530136]

Haghverdi L, Lun ATL, Morgan MD, and Marioni JC (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427. [PubMed: 29608177]

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, and Glass CK (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589. [PubMed: 20513432]

Hocker JD, Poirion OB, Zhu F, Buchanan J, Zhang K, Chiou J, Wang T-M, Hou X, Li YE, Zhang Y, et al. (2020). Cardiac Cell Type-Specific Gene Regulatory Programs and Disease Risk Association. bioRxiv, 2020.2009.2011.291724.

Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, Kvale MN, Kwok PY, Schaefer C, Krauss RM, et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. Nat Genet 50, 401–413. [PubMed: 29507422]

Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hägg S, Athanasiu L, et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 51, 404–413. [PubMed: 30617256]

Jedlicka P, Sui X, Sussel L, and Gutierrez-Hartmann A (2009). Ets transcription factors control epithelial maturation and transit and crypt-villus morphogenesis in the mammalian intestine. Am J Pathol 174, 1280–1290. [PubMed: 19264912]

Ji SG, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, Kumasaka N, Atkinson EJ, Schlicht EM, Liu JZ, et al. (2017). Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nat Genet 49, 269–273. [PubMed: 27992413]

Jin Y, Andersen G, Yorgov D, Ferrara TM, Ben S, Brownson KM, Holland PJ, Birlea SA, Siebert J, Hartmann A, et al. (2016). Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. Nat Genet 48, 1418–1424. [PubMed: 27723757]

John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP, Thurman RE, et al. (2013). Genome-Scale Mapping of DNase I Hypersensitivity. Current Protocols in Molecular Biology 103, 21.27.21–21.27.20.

Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet 51, 1442–1449. [PubMed: 31501517]

Kemp JP, Morris JA, Medina-Gomez C, Forgetta V, Warrington NM, Youlten SE, Zheng J, Gregson CL, Grundberg E, Trajanoska K, et al. (2017). Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. Nat Genet 49, 1468–1475. [PubMed: 28869591]

Kilpeläinen TO, Carli JF, Skowronski AA, Sun Q, Kriebel J, Feitosa MF, Hedman Å K, Drong AW, Hayes JE, Zhao J, et al. (2016). Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. Nat Commun 7, 10494. [PubMed: 26833098]

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, and Kirschner MW (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201. [PubMed: 26000487]

Klemm SL, Shipony Z, and Greenleaf WJ (2019). Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics 20, 207–220.

Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, Duong TE, Gao D, Chun J, Kharchenko PV, et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 36, 70–80. [PubMed: 29227469]

Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, Pokholok D, Aryee MJ, Steemers FJ, Lebofsky R, et al. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology 37, 916–924.

Li YE, Preissl S, Hou X, Zhang Z, Zhang K, Qiu Y, Poirion OB, Li B, Chiou J, Liu H, et al. (2021). An atlas of gene regulatory elements in adult mouse cerebrum. Nature 598, 129–136. [PubMed: 34616068]

Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, Castanon R, Lucero J, Nery JR, Sandoval JP, et al. (2017a). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 357, 600–604. [PubMed: 28798132]

Luo Y, de Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, Lamb CA, McCarthy S, Ahmad T, Edwards C, et al. (2017b). Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat Genet 49, 186–192. [PubMed: 28067910]

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214. [PubMed: 26000488]

Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 50, 1505–1513. [PubMed: 30297969]

Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van der Laan SW, Gretarsdottir S, et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 50, 524–537. [PubMed: 29531354]

Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, Rybin D, Liu CT, Bielak LF, Prokopenko I, et al. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659–669. [PubMed: 22581228]

Pownall Mary Elizabeth, Gustafsson Marcus K., and Emerson Charles P., J (2002). Myogenic Regulatory Factors and the Specification of Muscle Progenitors in Vertebrate Embryos. Annual Review of Cell and Developmental Biology 18, 747–783.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, NY) 337, 1190–1195.

McInnes L, Healy J, and Melville J (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, and Bejerano G (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28, 495–501. [PubMed: 20436461]

Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Teodosiadis A, et al. (2020). Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251. [PubMed: 32728217]

Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, Lemaçon A, Soucy P, Glubb D, Rostamianfar A, et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92–94. [PubMed: 29059683]

Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710. [PubMed: 32728249]

Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen TH, Ulirsch JC, Natri HM, et al. (2020). Genome-wide maps of enhancer regulation connect risk variants to disease genes. bioRxiv, 2020.2009.2001.278093.

Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet 50, 1234–1239. [PubMed: 30061737]

Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 47, 1121–1130. [PubMed: 26343387]

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381. [PubMed: 24390342]

Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, Curtin JA, Bønnelykke K, Tian C, Takahashi A, et al. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet 47, 1449–1456. [PubMed: 26482879]

Perrino C, and Rockman HA (2006). GATA4 and the Two Sides of Gene Expression Reprogramming. Circulation Research 98, 715–716. [PubMed: 16574910]

Pividori M, Schoettler N, Nicolae DL, Ober C, and Im HK (2019). Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. Lancet Respir Med 7, 509–522. [PubMed: 31036433]

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell 71, 858–871.e858. [PubMed: 30078726]

Pollard KS, Hubisz MJ, Rosenbloom KR, and Siepel A (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20, 110–121. [PubMed: 19858363]

Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, et al. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nature neuroscience 21, 432–439. [PubMed: 29434377]

Roadmap Epigenomics, C., Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330. [PubMed: 25693563]

Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, Shrine N, Qiao D, Bartz TM, Kim DK, et al. (2019). Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat Genet 51, 494–505. [PubMed: 30804561]

Salem S, Salem D, and Gros P (2020). Role of IRF8 in immune cells functions, protection against infections, and susceptibility to inflammatory diseases. Human Genetics 139, 707–721. [PubMed: 32232558]

Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU, et al. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42, 142–148. [PubMed: 20081857]

Schafmayer C, Harrison JW, Buch S, Lange C, Reichert MC, Hofer P, Cossais F, Kupcinskas J, von Schönfels W, Schniewind B, et al. (2019). Genome-wide association analysis of diverticular disease points towards neuromuscular, connective tissue and epithelial pathomechanisms. Gut 68, 854–865. [PubMed: 30661054]

Schaid DJ, Chen W, and Larson NB (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19, 491–504. [PubMed: 29844615]

Schiaffino S, and Reggiani C (2011). Fiber types in mammalian skeletal muscles. Physiol Rev 91, 1447–1531. [PubMed: 22013216]

Schiaffino S, Rossi AC, Smerdu V, Leinwand LA, and Reggiani C (2015). Developmental myosins: expression patterns and functional significance. Skeletal muscle 5, 22–22. [PubMed: 26180627]

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, and Stoeckert CJ Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol 6, R33. [PubMed: 15833120]

Shadrina AS, Sharapov SZ, Shashkova TI, and Tsepilov YA (2019). Varicose veins of lower extremities: Insights from the first large-scale genetic study. PLoS Genet 15, e1008110. [PubMed: 30998689]

Shen T, Aneas I, Sakabe N, Dirschinger RJ, Wang G, Smemo S, Westlund JM, Cheng H, Dalton N, Gu Y, et al. (2011). Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. The Journal of Clinical Investigation 121, 4640–4654. [PubMed: 22080862]

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120. [PubMed: 22763441]

Shlyueva D, Stampfel G, and Stark A (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272–286. [PubMed: 24614317]

Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, et al. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet 51, 481–493. [PubMed: 30804560]

Singh MK, Christoffels VM, Dias JM, Trowe M-O, Petry M, Schuster-Gossler K, Bürger A, Ericson J, and Kispert A (2005). *Tbx20* is essential for cardiac chamber differentiation and repression of *Tbx2*. Development 132, 2697–2707. [PubMed: 15901664]

Sinnamon JR, Torkenczy KA, Linhoff MW, Vitak SA, Mulqueen RM, Pliner HA, Trapnell C, Steemers FJ, Mandel G, and Adey AC (2019). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. Genome Research 29, 857–869. [PubMed: 30936163]

Song M, Pebworth M-P, Yang X, Abnousi A, Fan C, Wen J, Rosen JD, Choudhary MNK, Cui X, Jones IR, et al. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature.

Song M, Yang X, Ren X, Maliskova L, Li B, Jones IR, Wang C, Jacob F, Wu K, Traglia M, et al. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nature Genetics 51, 1252–1262. [PubMed: 31367015]

Stranger BE, Brigham LE, Hasz R, Hunter M, Johns C, Johnson M, Kopen G, Leinweber WF, Lonsdale JT, McDonald A, et al. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. Nature Genetics 49, 1664–1670. [PubMed: 29019975]

Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, et al. (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes 60, 2624–2634. [PubMed: 21873549]

Stuart CA, Stone WL, Howell ME, Brannon MF, Hall HK, Gibson AL, and Stone MH (2016). Myosin content of individual human muscle fibers isolated by laser capture microdissection. Am J Physiol Cell Physiol 310, C381–389. [PubMed: 26676053]

Tachmazidou I, Hatzikotoulas K, Southam L, Esparza-Gordillo J, Haberland V, Zheng J, Johnson T, Koprulu M, Zengini E, Steinberg J, et al. (2019). Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. Nat Genet 51, 230–236. [PubMed: 30664745]

Teumer A, Chaker L, Groeneweg S, Li Y, Di Munno C, Barbieri C, Schultheiss UT, Traglia M, Ahluwalia TS, Akiyama M, et al. (2018). Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. Nat Commun 9, 4455. [PubMed: 30367059]

Tin A, Marten J, Halperin Kuhns VL, Li Y, Wuttke M, Kirsten H, Sieber KB, Qiu C, Gorski M, Yu Z, et al. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat Genet 51, 1459–1474. [PubMed: 31578528]

Traag VA, Van Dooren P, and Nesterov Y (2011). Narrow scope for resolution-limit-free community detection. Phys Rev E Stat Nonlin Soft Matter Phys 84, 016114. [PubMed: 21867264]

Traag VA, Waltman L, and van Eck NJ (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233. [PubMed: 30914743]

Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Paşca AM, Kundaje A, et al. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell.

Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al. (2020). Global reference mapping of human transcription factor footprints. Nature 583, 729–736. [PubMed: 32728250]

Visel A, Minovitsky S, Dubchak I, and Pennacchio LA (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35, D88–92. [PubMed: 17130149]

Wakefield J (2009). Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 33, 79–86. [PubMed: 18642345]

Wang A, Chiou J, Poirion OB, Buchanan J, Valdez MJ, Verheyden JM, Hou X, Kudtarkar P, Narendra S, Newsome JM, et al. (2020). Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. Elife 9.

Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland Ø, Laurin C, Bacelis J, Peng S, Hao K, Feenstra B, et al. (2019). Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet 51, 804–814. [PubMed: 31043758]

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, and Posthuma D (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet 51, 1339–1348. [PubMed: 31427789]

Wiberg A, Ng M, Schmid AB, Smillie RW, Baskozos G, Holmes MV, Künnapuu K, Mägi R, Bennett DL, and Furniss D (2019). A genome-wide association analysis identifies 16 novel susceptibility loci for carpal tunnel syndrome. Nat Commun 10, 1030. [PubMed: 30833571]

Wolock SL, Lopez R, and Klein AM (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems 8, 281–291.e289. [PubMed: 30954476]

Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, Tin A, Wang L, Chu AY, Hoppmann A, et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet 51, 957–972. [PubMed: 31152163]

Yan J, Qiu Y, Santos A.M.R.d., Yin Y, Li YE, Vinckier N, Nariai N, Benaglio P, Raman A, Li X, et al. (2021). Systematic Analysis of Transcription Factor Binding to Noncoding Variants in the Human Genome. Nature *in press*.

Yan M, Wang H, Sun J, Liao W, Li P, Zhu Y, Xu C, Joo J, Sun Y, Abbasi S, et al. (2016). Cutting Edge: Expression of IRF8 in Gastric Epithelial Cells Confers Protective Innate Immunity against Helicobacter pylori Infection. J Immunol 196, 1999–2003. [PubMed: 26843324]

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42, 565–569. [PubMed: 20562875]

Zhang K, Wang M, Zhao Y, and Wang W (2019). Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. Science Advances 5, eaav3262. [PubMed: 30944857]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9, R137. [PubMed: 18798982]

Zhu C, Preissl S, and Ren B (2020). Single-cell multimodal omics: the power of many. Nature Methods 17, 11–14. [PubMed: 31907462]

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D (2002). The human genome browser at UCSC. Genome Res 12, 996–1006. [PubMed: 12045153]

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

Litvinukova M, Talavera-Lopez C, Maatz H, Reichart D, Worth CL, Lindberg EL, Kanda M, Polanski K, Heinig M, Lee M, et al. (2020). Cells of the adult human heart. Nature 588, 466–472. [PubMed: 32971526]

Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, Andrade-Navarro MA, Buenrostro JD, and Pinello L (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biology 20, 241. [PubMed: 31739806]

Highlight

- >1.3 million chromatin accessibility profiles from 30 adult and 15 fetal human tissues
- An atlas of ~1.2 million candidate *cis*-regulatory elements across 222 cell types
- Cell type specificity of fetal and adult candidate *cis*-regulatory elements
- Interpretation of noncoding variants associated with complex traits and diseases

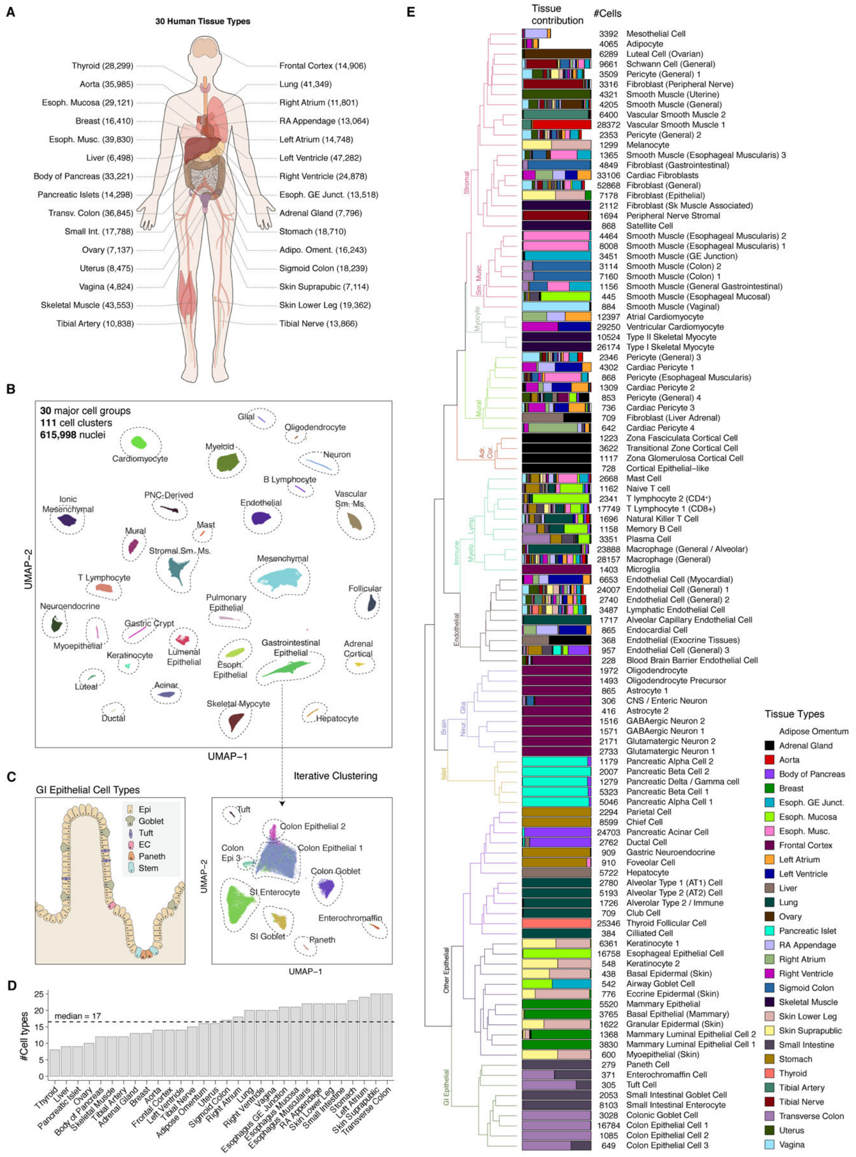


Figure 1 |. Single-cell chromatin accessibility analysis of 30 adult human primary tissues.

A) A total of 92 biosamples from 30 tissue types, were used for sci-ATAC-seq. The number of nuclei profiled per tissue is denoted in parentheses. **B)** Clustering of 615,998 nuclei revealed 30 major cell groups. Each dot represents a nucleus colored by cluster ID. Embedding was created by Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). **C)** An example illustrating subclusters within the major cell group of gastrointestinal (GI) epithelial cells revealed by iterative clustering. **D)** Bar plot showing the number of cell types identified in each of the 30 human tissues, counting only cell types constituting >0.2% of all cells in the given tissue. **E)** Distribution of cell types across human tissues. The dendrogram on the left was created by hierarchical clustering of cell clusters based on chromatin accessibility. The bar chart represents relative contributions of tissues to cell clusters. Raw data are available on Mendeley Data: 10.17632/yv4fzv6cnm.1.

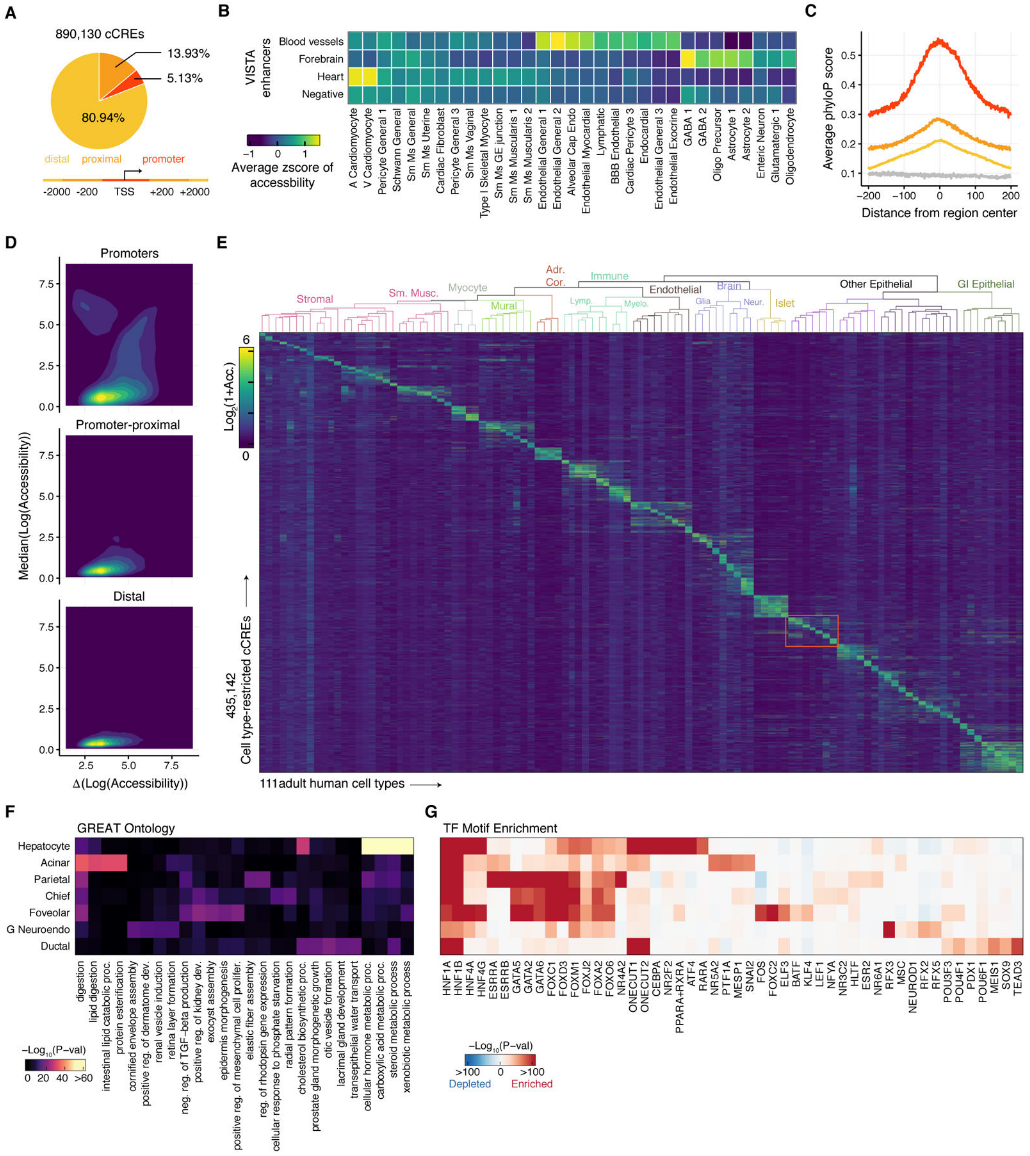


Figure 2 |. An atlas of cCREs in adult human cell types.

A) Classification of 890,130 cCREs across the human genome based on their distances to annotated TSSs. **B)** Heatmap showing the average chromatin accessibility for each of four groups (blood vessel, forebrain, heart, negative control) of validated tissue-specific enhancers from the VISTA database (Visel et al., 2007) across indicated cell types. Z-scores were calculated using all 111 cell types. The top 10 cell types in each validated enhancer group are shown. **C)** Average phyloP (Pollard et al., 2010) conservation scores of cCREs stratified by groups defined in **A.** Genomic background is indicated in gray. **D)** Two-dimensional density plot showing the median chromatin accessibility compared with the range (difference between maximum and minimum) of chromatin accessibility across 111 cell clusters for 890,130 cCREs, stratified by groups defined in **A. E)** Heatmap representation of 435,142 cCREs showing cell-type-restricted patterns in 111 cell types. Color represents $\log_2$-transformed chromatin accessibility. **F,G)** Heatmaps showing GO terms **(F)** and TF motifs **(G)** with maximal enrichment in cell-type-restricted cCREs of selected cell types. Only the most enriched TF motif in each of the previously identified motif archetypes (Vierstra et al., 2020) was selected as the representative and the top 10 motifs were selected for each cell type. Color represents $-\mathrm{Log}_{10}P$. Full GO and motif enrichments are available on Mendeley Data: 10.17632/yv4fzv6cnm.1.

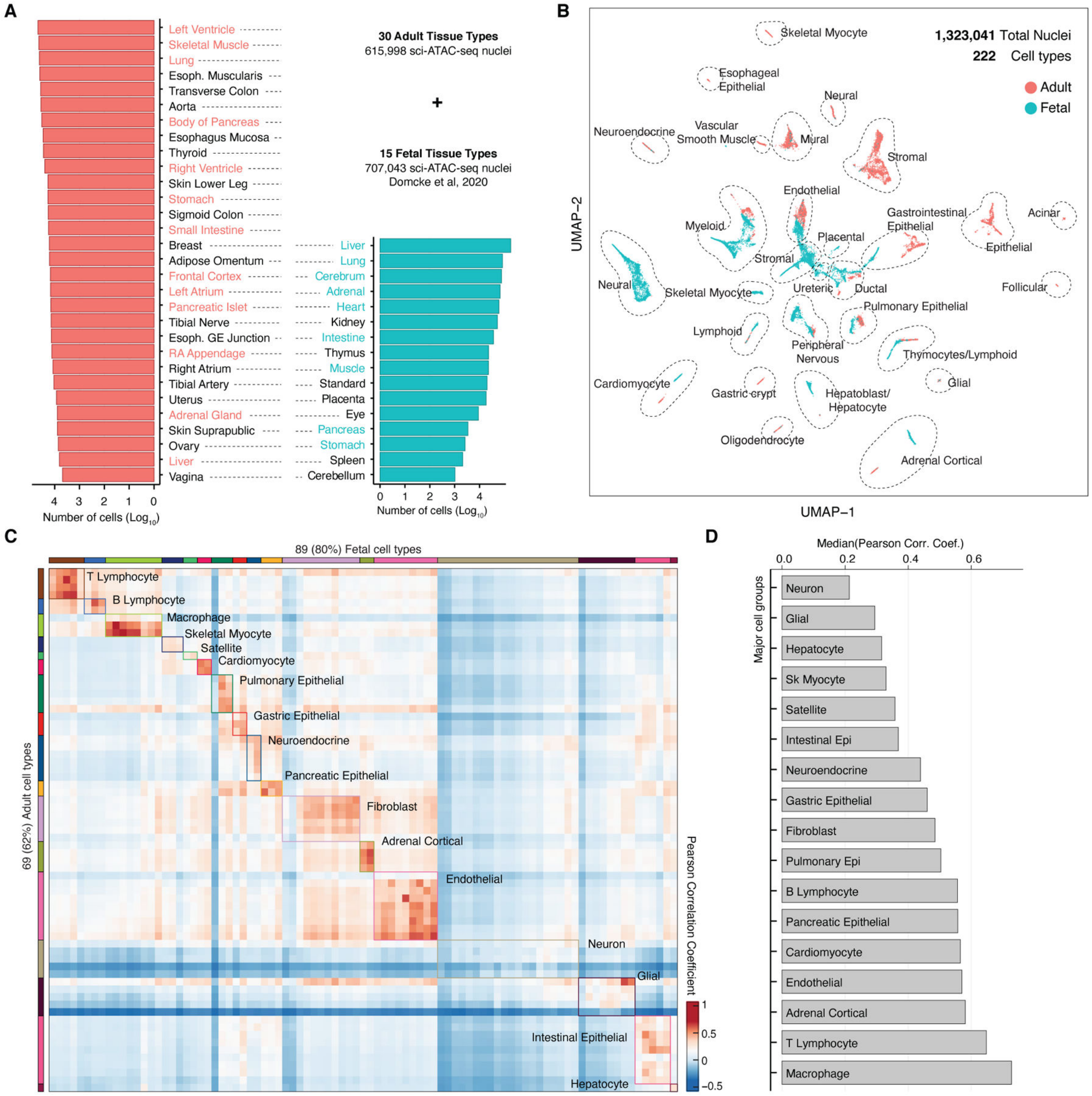


Figure 3 |. Integrative analysis of adult and fetal single-cell chromatin accessibility atlases.

A) Number of sci-ATAC-seq cells per tissue type for 30 adult and 15 human fetal tissue types that were integrated. Matching tissue types between adult and fetal datasets are highlighted in red or blue respectively. Standard: sentinel tissue (trisomy 18 cerebrum). B) UMAP embedding of 1,323,041 nuclei from fetal and adult tissues. Each dot in the scatter plot represents a nucleus, colored by life stage. C) Heatmap showing Pearson correlation coefficient (PCC) between 69 adult cell types and 89 fetal cell types from 17 manually defined cell groups that are present in both adult and fetal tissues. A comprehensive heatmap is provided in Figure S5. D) Bar plot showing the median PCC for each major cell group indicated in C.

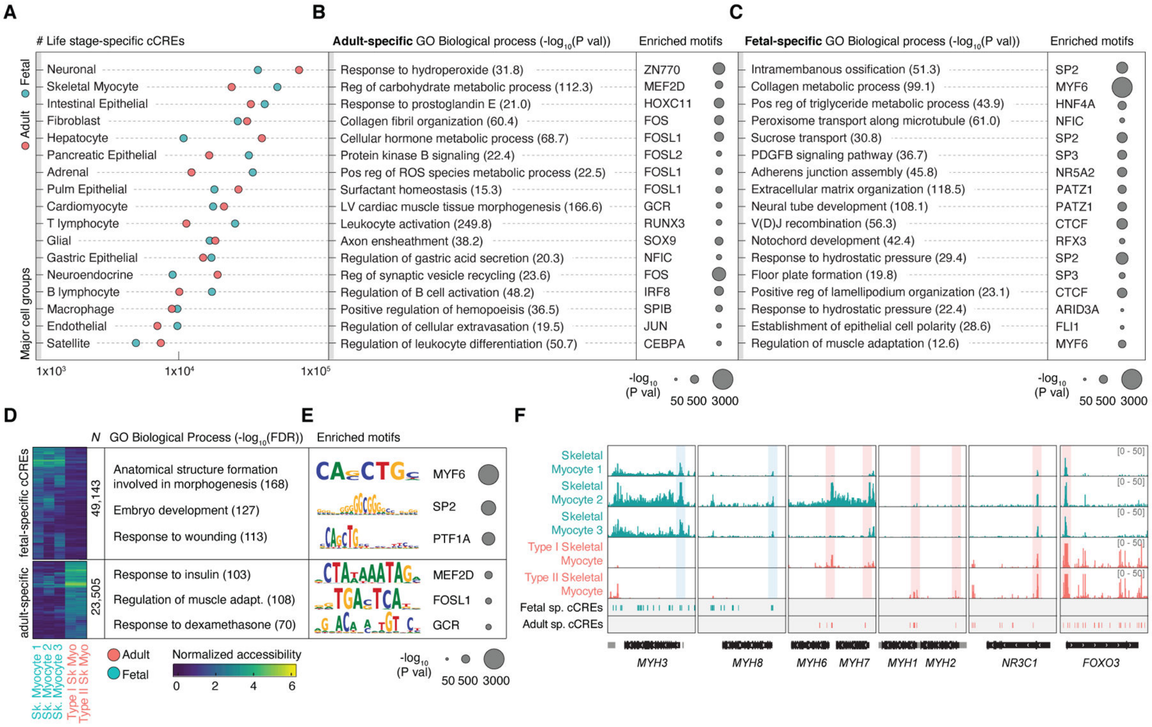


Figure 4 |. Differential chromatin accessibility landscapes in adult and fetal human cell types.

A) Dot plot showing the number of adult and fetal specific cCREs detected for each major cell group indicated in **C**. **B-C)** Significant GO biological process ontology terms and transcription factor motif enrichments for adult-specific (**B**) and fetal-specific (**C**) cCREs. **D)** Heatmap representation of 72,648 differentially accessible (DA) cCREs between fetal and adult skeletal myocytes along with significant GREAT biological process ontology enrichments (McLean et al., 2010). Color represents log-transformed normalized signal. **E)** Significantly enriched TF motifs within fetal and adult skeletal myocyte DA cCREs. The most enriched motif within each motif archetype (Vierstra et al., 2020) was selected and the top three were displayed. **F)** Genome browser tracks showing chromatin accessibility for fetal and adult skeletal myocytes along with DA cCREs between the adult and fetal skeletal myocytes. Indicated genes are shown in black, other genes are shown in gray. TSSs of the indicated genes are shaded in red and blue.

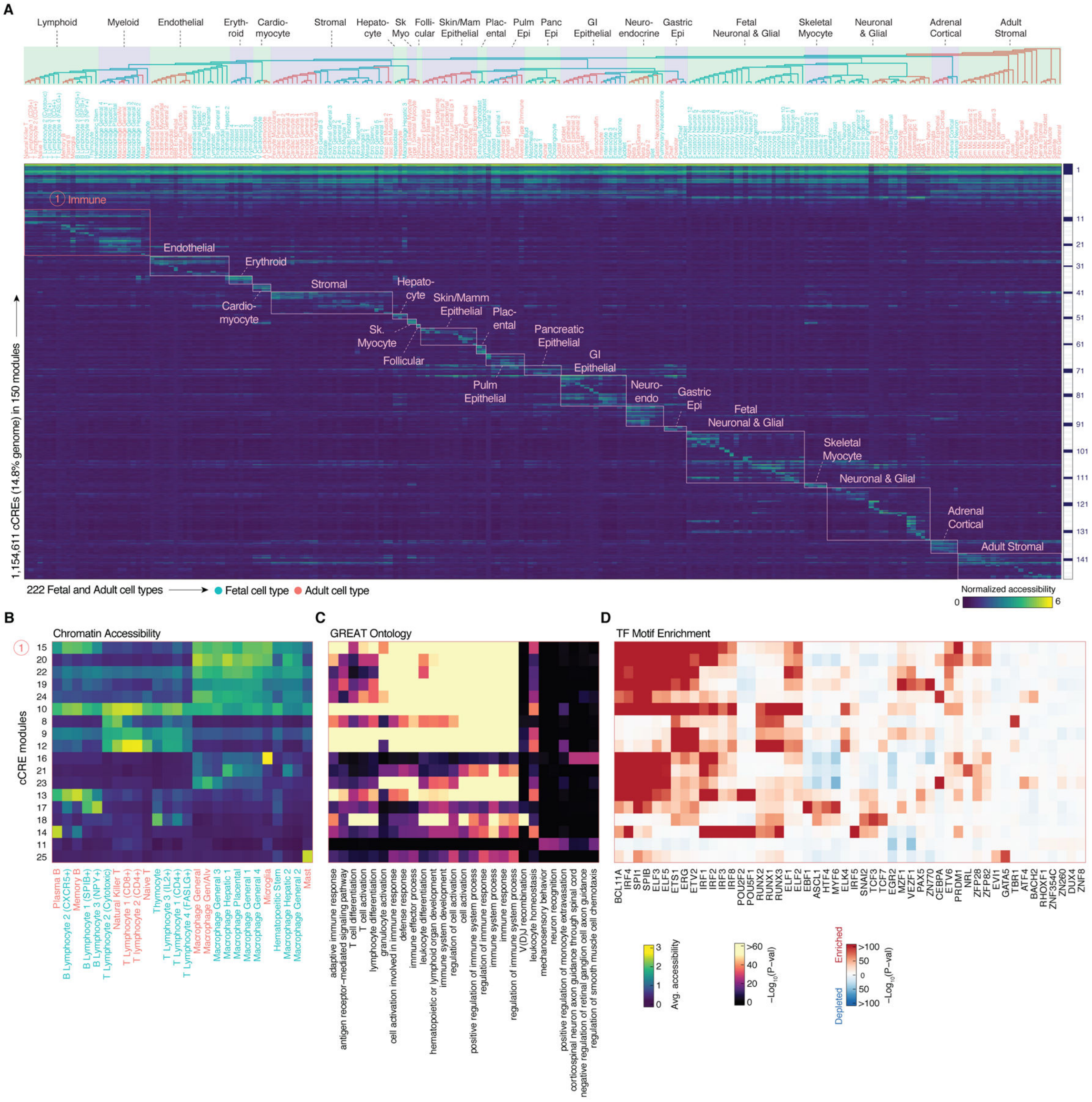


Figure 5 |. Delineation of CRE modules across 222 fetal and adult human cell types.

A) Heatmap representation of chromatin accessibility for 1,154,611 cCREs across 222 fetal and adult cell types. Color represents normalized chromatin accessibility. cCREs were organized into 150 modules by K-means clustering, indicated by the color bars on the right. 20 groups of lineage-specific modules (colored boxes) are highlighted. B-D) Heatmaps showing chromatin accessibility (B), GO terms (C) and motifs (D) with maximal enrichment in a subset of CRE modules (rows) for immune cell types. The GO and motif heatmaps are colored by enrichment $-\log_{10}$P. Only the most enriched TF motif in each of the previously identified motif archetypes (Vierstra et al., 2020) was selected as the representative and the top 5 motifs were selected for each module. Full GO and motif enrichments are available on Mendeley Data: 10.17632/yv4fzv6cnm.1.

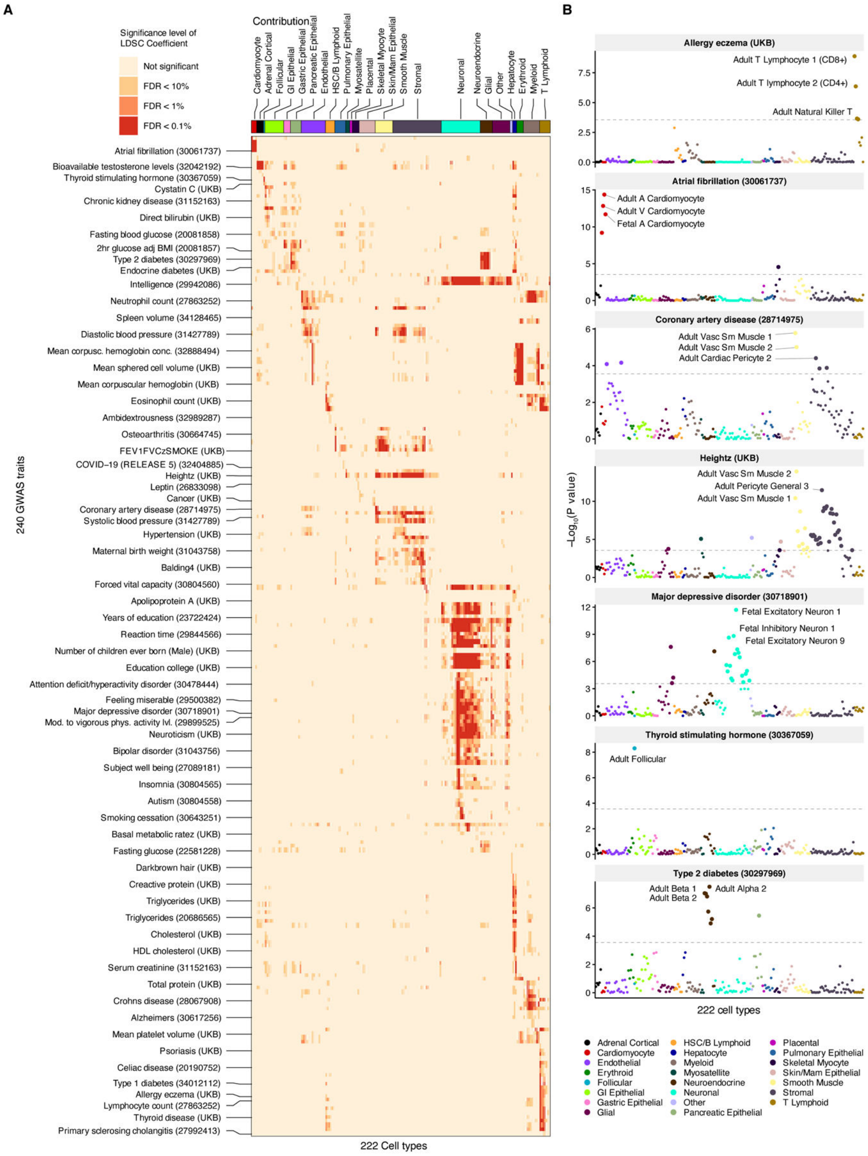
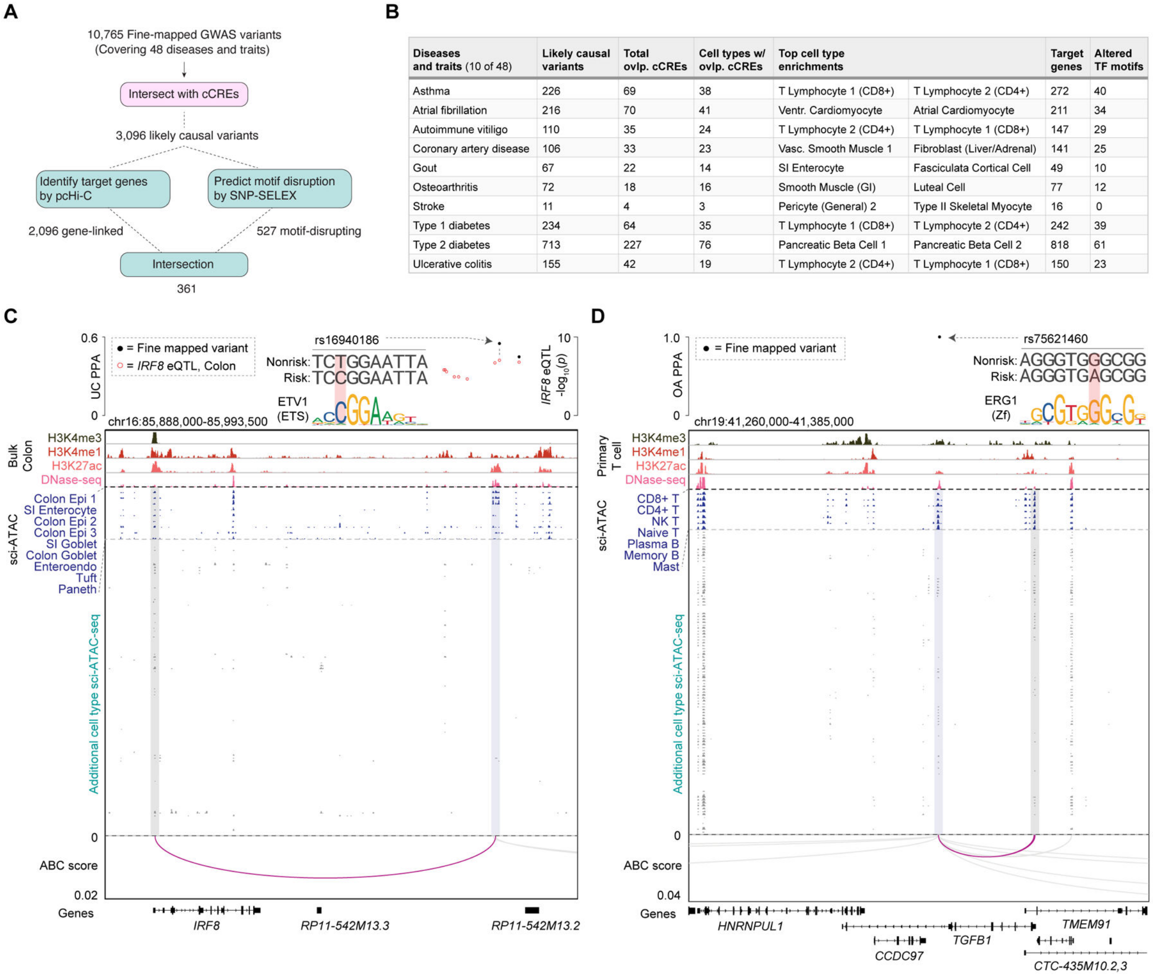


Figure 6 |. Association of fetal and adult human cell types with complex traits and diseases.

A) Heatmap showing enrichment of risk variants associated with disease and non-disease traits from genome wide association studies in human cell type-resolved cCREs. Cell type-stratified linkage disequilibrium score regression (LDSC) analysis was performed using GWAS summary statistics for 240 phenotypes. Total cCREs identified independently from each fetal and adult cell type were used as input for analysis. P-values were corrected using the Benjamini Hochberg procedure for multiple tests. FDRs of LDSC coefficient are displayed. 66 selected traits were highlighted on the left, with PubMed identifiers (PMIDs) or “UKB”, indicating summary statistics downloaded from the UK Biobank, enclosed in parentheses. Numerical results are reported in Table S4. B) Dot plots showing significance of enrichment for selected traits from panel A within cCREs from 222 fetal and adult cell types. Each circle represents a cell type. Large circles pass the cutoff of FDR < 1% at

$-\log_{10}(P) = 3.55$. The top 3 most highly associated cell types are labeled for each trait. Comprehensive data are provided in Table S4.



Diseases and traits (10 of 48)	Likely causal variants	Total ovlp. cCREs	Cell types w/ ovlp. cCREs	Top cell type enrichments		Target genes	Altered TF motifs
Asthma	226	69	38	T Lymphocyte 1 (CD8+)	T Lymphocyte 2 (CD4+)	272	40
Atrial fibrillation	216	70	41	Ventr. Cardiomyocyte	Atrial Cardiomyocyte	211	34
Autoimmune vitiligo	110	35	24	T Lymphocyte 2 (CD4+)	T Lymphocyte 1 (CD8+)	147	29
Coronary artery disease	106	33	23	Vasc. Smooth Muscle 1	Fibroblast (Liver/Adrenal)	141	25
Gout	67	22	14	SI Enterocyte	Fasciculata Cortical Cell	49	10
Osteoarthritis	72	18	16	Smooth Muscle (GI)	Luteal Cell	77	12
Stroke	11	4	3	Pericyte (General) 2	Type II Skeletal Myocyte	16	0
Type 1 diabetes	234	64	35	T Lymphocyte 1 (CD8+)	T Lymphocyte 2 (CD4+)	242	39
Type 2 diabetes	713	227	76	Pancreatic Beta Cell 1	Pancreatic Beta Cell 2	818	61
Ulcerative colitis	155	42	19	T Lymphocyte 2 (CD4+)	T Lymphocyte 1 (CD8+)	150	23

Figure 7 |. Systematic interpretation of molecular functions of noncoding risk variants.

A) Schematic illustrating the workflow for annotating fine-mapped noncoding risk variants. **B)** Table showing the number of likely causal variants (PPA > 0.1), number of cCREs overlapping likely causal variants, number of cell types in which overlapping cCREs are accessible, top cell types variants are enriched in based on LD score regression (Bulik-Sullivan et al., 2015), number of predicted target genes for likely causal variants, and significantly altered motifs predicted by deltaSVM model trained using SNP-SELEX data for 10 examples out of 48 total fine-mapped diseases and traits. Comprehensive data are provided in Table S5. **C,D)** Fine mapping and molecular characterization of an ulcerative colitis (UC) risk variant (**C**) in a gastrointestinal (GI) epithelial cell cCRE and an osteoarthritis variant (**D**) in an immune cell cCRE. Genome browser tracks (GRCh38) display ChIP-seq and DNase-seq from ENCODE human colon datasets (**C**) and primary T cell datasets (**D**) as well as chromatin accessibility profiles for cell types from sci-ATAC-seq. Chromatin interaction tracks show linkages between the variant-containing cCREs and genes from promoter capture Hi-C data via Activity-by-Contact (ABC) (Fulco et al., 2019) analysis. All linkages shown have an ABC score > 0.015. PPA: Posterior probability of association.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Adult Human Tissue Samples	ENTEx Collaboration	https://www.encodeproject.org/entex-matrix/?type=Experiment&status=released&internal_tags=ENTEx
Chemicals, peptides, and recombinant proteins		
Tn5 transposase	QB3 Macrolab at UC Berkeley	N/A
SPRISelect reagent	Beckman Coulter	Cat# B23319
DRAQ7	Cell Signaling	Cat# 7406
NEBNext High-Fidelity PCR Master Mix	topNEB	Cat# M0541L
LightCycler 480 SYBR Green I Master Mix	Roche	Cat# 04707516001
Critical commercial assays		
gentleMACS Octo Dissociator	Miltenyi	Cat# 130-095-937
Deposited data		
sci-ATAC-seq data of human adult tissues	This paper	GEO: GSE184462
sci-ATAC-seq data of human lung samples	(Wang et al., 2020)	dbGaP: phs001961
sci-ATAC-seq data of human heart samples	(Hocker et al., 2021)	dbGaP: phs002204
sci-ATAC-seq data of human islet samples	(Chiou et al., 2021)	GEO: GSE160472
sci-ATAC-seq data of human fetal tissues	(Domcke et al., 2020)	https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/ dbGaP: phs002003
raw data used to produce the figures	This paper	Mendeley Data: 10.17632/yv4fzv6cnm.1
Oligonucleotides		
Custom Tagmentation Oligos	This paper	Table S18
Custom PCR Primers	This paper	Table S18
Custom Sequencing Primers	This paper	Table S18
Software and algorithms		
BWA version 0.7.17	(Li and Durbin, 2009)	https://github.com/lh3/bwa
Samtools version 1.9	(Danecek et al., 2021)	https://github.com/samtools/samtools
Taiji version 1.3.0	(Zhang et al., 2019)	https://github.com/Taiji-pipeline/Taiji
Python package: taiji-utils version 0.2.3	This paper	https://pypi.org/project/taiji-utils/
liftOver	(Kent et al., 2002)	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver
ATACdemultiplex version 0.46.12	This paper	https://gitlab.com/Grouumf/ATACdemultiplex/
bedGraphToBigWig	(Kent et al., 2002)	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/bedGraphToBigWig
R version 4.0.5	R Foundation for Statistical Computing	https://www.r-project.org/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python version 3.8	Python Software Foundation	https://www.python.org/
Python package: umap-learn version 0.5.0		https://pypi.org/project/umap-learn/
Python package: scrublet	(Wolock et al., 2019)	https://github.com/swolock/scrublet
GREAT version 4.0.4	(McLean et al., 2010)	http://great.stanford.edu/public/html/