RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

# A human cell atlas of fetal chromatin accessibility

Silvia Domcke*, Andrew J. Hill*, Riza M. Daza*, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell†, Darren A. Cusanovich†, Jay Shendure†

**INTRODUCTION:** In recent years, the single-cell genomics field has made incredible progress toward disentangling the cellular heterogeneity of human tissues. However, the overwhelming majority of effort has been focused on single-cell gene expression rather than the chromatin landscape that shapes and is shaped by gene expression. Toward advancing our understanding of the regulatory programs that underlie human cell types, we set out to generate single-cell atlases of both chromatin accessibility (this study) and gene expression (Cao *et al.*, this issue) from a broad range of human fetal tissues.

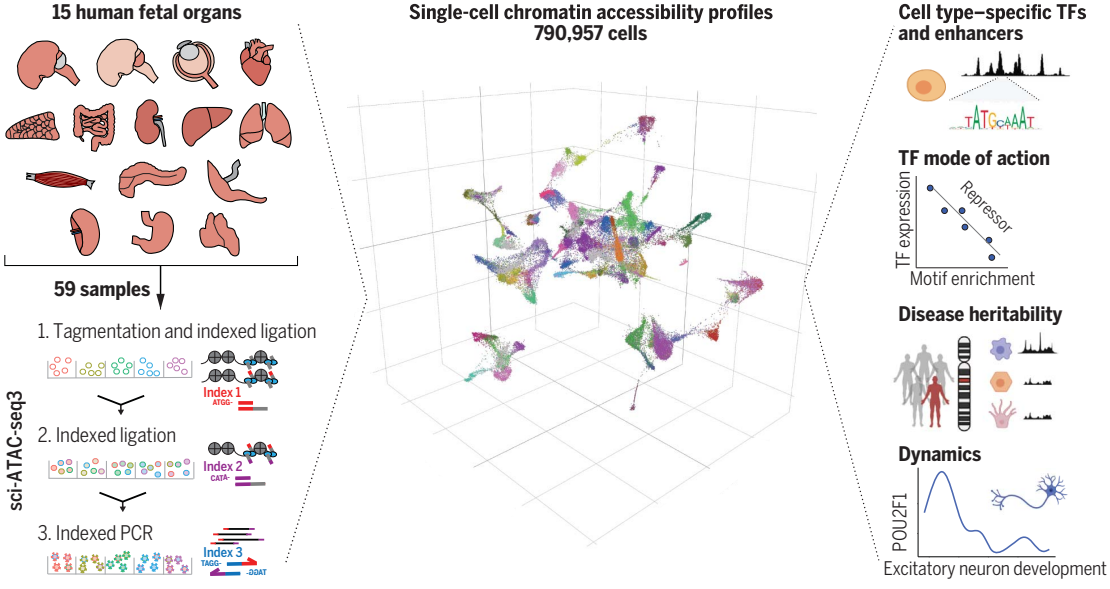
**RATIONALE:** Regions of accessible chromatin in our genome, such as enhancers, play key roles in the determination and maintenance of cell fates. Accessible regions are also markedly enriched for genetic variation that contributes to common disease heritability. The vast majority of chromatin accessibility data collected to date lacks single-cell resolution, limiting our ability to infer patterns such as which cell types are most relevant to each common disease. We previously demonstrated single-cell profiling of chromatin accessibility using combinatorial indexing, based on two rounds of in situ molecular barcoding. Here, we describe an improved assay that uses three levels of combinatorial indexing and does not rely on custom reagents. The method, sci-ATAC-seq3, reduces costs and opens the door to the scales necessary for generating a human cell atlas of chromatin accessibility.

**RESULTS:** We applied sci-ATAC-seq3 to 59 human fetal samples ranging from 89 to 125 days in estimated postconceptual age and representing 15 organs, altogether obtaining high-quality chromatin accessibility profiles from ~800,000 single cells. Gene expression data collected on an overlapping set of tissues were leveraged to annotate cell types. We asked which transcription factor (TF) motifs found in the accessible sites of each cell best explain its cell type affiliation, revealing both known and potentially previously unknown regulators of cell fate specification and/or maintenance. Many TFs could be putatively assigned as activators or repressors depending on whether their expression and the accessibility of their cognate motif were positively or negatively correlated across cell types. Comparing chromatin accessibility from cell types that appear in multiple tissues revealed that whereas blood cell types are highly similar across organs, endothelial cells exhibit organ-specific chromatin accessibility, which appears to be controlled combinatorially by several TFs with overlapping expression patterns. We leveraged our master set of 1.05 million accessible sites, spanning 532 Mb or 17% of the reference human genome, to score cell type–specific links between candidate enhancers and genes based on coaccessibility, to detect cell type–specific enrichment of heritability for specific common human diseases, and to identify genetic variants affecting chromatin accessibility in cis. Comparisons with chromatin accessibility in corresponding adult tissues allowed us to identify fetal-specific cell subtypes and nominate POU2F1 as a potential regulator of excitatory neuron development.

**CONCLUSION:** Sci-ATAC-seq3 adds to a growing repertoire of single-cell methods that use combinatorial indexing, a technical paradigm whose advantages include exponential scaling and greater range to profile diverse aspects of single-cell biology. We anticipate that the intersection of single-cell chromatin accessibility and gene expression will critically accelerate the field's long-term goal of establishing a deep, predictive understanding of gene regulation. An interactive website facilitates the exploration of these freely available data by tissue, cell type, locus, or motif (descartes.brotmanbaty.org). ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: darrenc@email.arizona.edu (D.A.C.); coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)
Cite this article as S. Domcke *et al.*, *Science* **370**, eaba7612 (2020). DOI: 10.1126/science.aba7612

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aba7612




**A human cell atlas of fetal chromatin accessibility enables the exploration of in vivo gene regulation across diverse cell types.** We devised a three-level combinatorial indexing assay (sci-ATAC-seq3) and profiled chromatin accessibility in ~800,000 single cells from 15 fetal organs. This rich resource enables, for example, identification of cell type–specific regulatory elements and TFs, classification of TFs into activators and repressors, and quantification of cell type–specific enrichments of complex trait heritability, as well as chromatin accessibility dynamics.

RESEARCH ARTICLE

HUMAN GENOMICS

# A human cell atlas of fetal chromatin accessibility

Silvia Domcke[1]*, Andrew J. Hill[1]*, Riza M. Daza[1]*, Junyue Cao[1], Diana R. O'Day[2], Hannah A. Pliner[3], Kimberly A. Aldinger[2,4], Dmitry Pokholok[5], Fan Zhang[5], Jennifer H. Milbank[1], Michael A. Zager[3,6], Ian A. Glass[2,3,4], Frank J. Steemers[5], Dan Doherty[2,3,4], Cole Trapnell[1,3,7]†, Darren A. Cusanovich[1,8,9]†, Jay Shendure[1,3,7,10]†

The chromatin landscape underlying the specification of human cell types is of fundamental interest. We generated human cell atlases of chromatin accessibility and gene expression in fetal tissues. For chromatin accessibility, we devised a three-level combinatorial indexing assay and applied it to 53 samples representing 15 organs, profiling ~800,000 single cells. We leveraged cell types defined by gene expression to annotate these data and cataloged hundreds of thousands of candidate regulatory elements that exhibit cell type–specific chromatin accessibility. We investigated the properties of lineage-specific transcription factors (such as POU2F1 in neurons), organ-specific specializations of broadly distributed cell types (such as blood and endothelial), and cell type–specific enrichments of complex trait heritability. These data represent a rich resource for the exploration of in vivo human gene regulation in diverse tissues and cell types.

In recent years, the single-cell genomics field has made incredible progress toward disentangling the cellular heterogeneity of human tissues. However, the overwhelming majority of effort has been focused on single-cell gene expression, with far fewer investigations of the chromatin landscape that shapes and is shaped by gene expression. This is in part because of a relative paucity of scalable methods for profiling chromatin accessibility, transcription factor (TF) binding, and/or histones at single-cell resolution.

The single-cell combinatorial indexing ("sci-") (*1*) framework involves the splitting and pooling of cells or nuclei to wells in which molecular barcodes are introduced in situ to a species of interest at each round. Through successive rounds of in situ molecular barcoding, species within the same cell are concordantly labeled with a distinct combination of barcodes. Sci- assays have been developed for profiling chromatin accessibility [sci-ATAC-seq (ATAC-seq, assay for transposase-accessible chromatin with high-throughput sequencing)], gene expression [sci-RNA-seq (RNA-seq, RNA-sequencing)], nuclear architecture, genome sequence, methylation, histone marks and other phenomena, as well as sci- co-assays—for example, for profiling chromatin accessibility and gene expression jointly (*1*–*12*) ["CoBatch," "Split-seq," "Paired-seq," and "dscATAC-seq" also effectively rely on single-cell combinatorial indexing (*8*–*10*, *12*)]. Although we and others have profiled chromatin accessibility in >100,000 mammalian cells (*9*, *12*, *13*), the methods used require custom-loading of the Tn5 enzyme with barcoded adapters and/or are limited to $10^4$ to $10^5$ cells per experiment by collisions—cells receiving the same combination of barcodes.

We developed an improved assay for single-cell profiling of chromatin accessibility that both uses three levels of combinatorial indexing and, in contrast with previous iterations of sci-ATAC-seq and related methods (*1*, *6*, *9*, *12*), does not rely on molecularly barcoded Tn5 complexes (sci-ATAC-seq3) (Fig. 1A and fig. S1A). Rather, the first two rounds of indexing are achieved through ligation to either end of the conventional, uniformly loaded Tn5 transposase complex (standard Nextera), whereas the final round of indexing remains through polymerase chain reaction (PCR). Relative to two-level sci-ATAC-seq but similar to sci-RNA-seq3, sci-ATAC-seq3 reduces the per-cell cost of library preparation (fig. S1B) as well as the rate of collisions (fig. S1, C and D), opening the door to experiments on the scale of $10^6$ cells. This protocol no longer requires cell sorting, and we also optimized ligase and polymerase choice, kinase concentration, and oligo designs and concentrations to maximize the number of fragments recovered from each cell. While maintaining an enrichment in accessible regions, we made the explicit choice to maximize complexity at the expense of specificity for accessible sites (Fig. 1B and fig. S1, E to G). In particular, we found that the fixation conditions could be tuned to adjust the sensitivity (complexity) versus specificity (enrichment in accessible sites) of the assay (fig. S1H).

As one step toward a comprehensive cell atlas of human development (*14*), we set out to generate single-cell atlases of both gene expression and chromatin accessibility using diverse human tissues obtained during mid-gestation [DESCARTES, Developmental Single Cell Atlas of gene Regulation and Expression; descartes.brotmanbaty.org (*15*)]. For chromatin accessibility, we applied sci-ATAC-seq3 to 59 fetal samples representing 15 organs, altogether profiling 1.6 million cells (Fig. 1C). We also describe profiling of gene expression in 5 million cells from the same organs, using an overlapping set of samples (*16*). The profiled organs span diverse systems. However, some systems were not accessible; bone marrow, bone, gonads, and skin are notably absent.

The rapid and uniform processing of heterogeneous fetal tissues presents a challenge. We developed a method for extracting nuclei directly from cryopreserved tissues that works across a variety of tissue types and produces homogenates suitable for both sci-ATAC-seq3 and sci-RNA-seq3. For sci-ATAC-seq3, we used tissue samples obtained from 23 fetuses ranging from 89 to 125 days in estimated post-conceptual age (Fig. 1, D and E, and table S1). All samples were karyotypically normal. Samples were processed in three batches; a mix of the same sentinel human fetal brain tissue and a mouse suspension cell line was included in each experiment to control for batch effects and estimate collision rates.

We sequenced sci-ATAC-seq3 libraries from the three experimental batches across five Illumina NovaSeq 6000 sequencing runs, generating just over 110 billion reads (55 billion read pairs). We compared these data at the tissue level, before splitting into single cells, against single-ended ENCODE deoxyribonuclease-sequencing (DNase-seq) data (fig. S2A) (*17*). Although sci-ATAC-seq3 data were somewhat less enriched in peaks (median reads in peaks: 29% for sci-ATAC-seq3; 35% for ENCODE DNase-seq) (fig. S2B), samples from the same tissue were comparably correlated for the two assays (median Spearman correlation: 0.93 for two samples from the same tissue for sci-ATAC-seq3; 0.91 for DNase-seq), with greater technical reproducibility for sci-ATAC-seq3 (median Spearman correlation: 0.95) (fig. S2C). Furthermore, samples clustered into their respective tissues from these aggregate profiles, whether analyzing the sci-ATAC-seq3 samples alone (Fig. 1F) or the sci-ATAC-seq3 and DNase-seq samples together (fig. S2D).

Splitting reads by sci- barcodes, we identified 1,568,018 cells (table S1), and from the barnyard control, we estimated collision rates of 1 to 4% for the three experiments (fig. S2E) (*18*). We observed no obvious batch effects (fig. S2F)

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA. [3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [4]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA. [5]Illumina, San Diego, CA, USA. [6]Center for Data Visualization, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [7]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. [8]Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, USA. [9]Asthma and Airway Disease Research Center, University of Arizona, Tucson, AZ, USA. [10]Howard Hughes Medical Institute, Seattle, WA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: darrenc@email.arizona.edu (D.A.C.); coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)

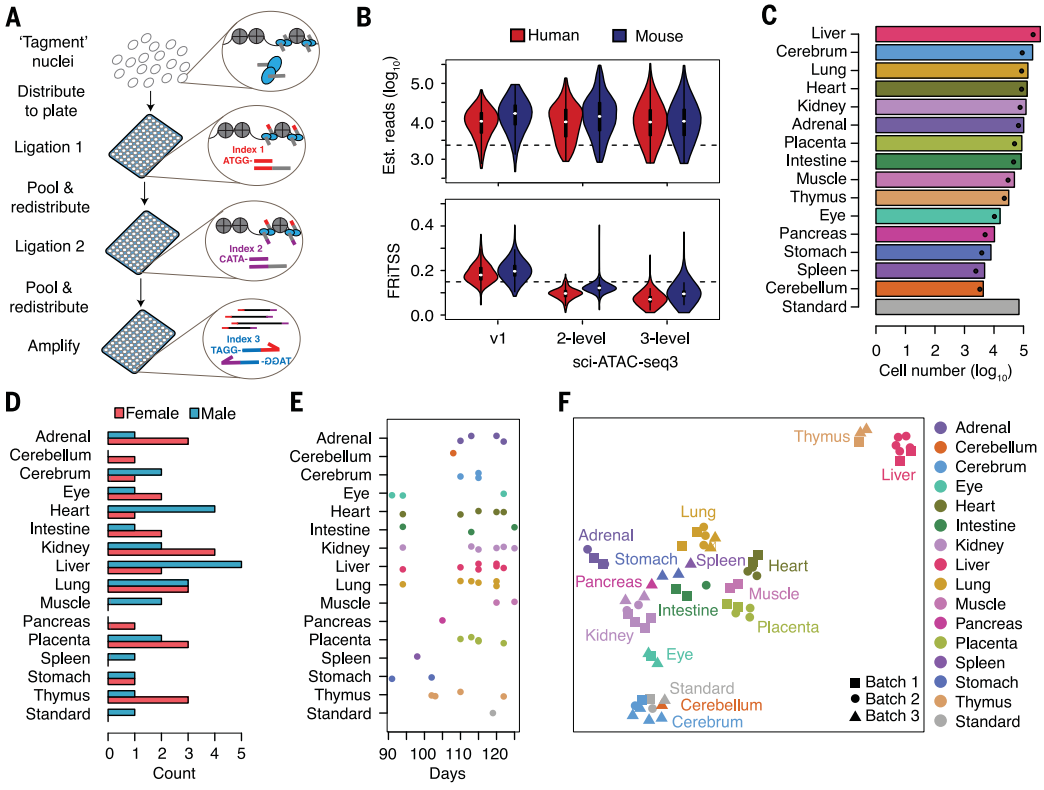



**Fig. 1. Design of three-level sci-ATAC-seq and application to chromatin accessibility profiling of 1.6 million cells from 59 fetal samples.** (**A**) Schematic of sci-ATAC-seq3. Nuclei are tagmented with Tn5 transposase in bulk. The first two rounds of indexing are achieved with successive ligations to each end of the Tn5 transposase complex, and the third round is achieved with PCR. (**B**) Comparison of complexity and specificity achieved with different versions of the sci-ATAC-seq protocol in mixing experiments of mouse and human suspension cell lines. The estimated total nonduplicate reads ("complexity") for each cell were calculated with Picard and are displayed as violin plots on a log10 scale (*115*). The fraction of reads in TSSs (FRiTSS) was calculated for each cell in the same experiments (bottom). Reads within 500 bp of a Gencode TSS were considered within the TSS. v1: species mixing experiment by using our previously published two-level sci-ATAC-seq protocol (*13*); 2-level: two-level version of the new protocol with simultaneous ligations; and 3-level: three-level version of the new protocol. (**C**) Barplot showing number of cells profiled per organ (log10 scale). Dots indicate the number of cells remaining after QC filtering procedures. Standard: sentinel tissue (trisomy 18 cerebrum) was included in all three experiments. (**D**) Barplot showing the distribution of sexes for samples corresponding to each organ. (**E**) Stripchart showing the estimated post-conceptual age of each sample. Samples are arranged by organ and slightly jittered to avoid overplotting. (**F**) UMAP visualization of aggregated chromatin accessibility profiles of single cells from each of the samples, colored by organ. Normalized accessibility at a master set of peaks was quantified for each "pseudobulk" sample and used as an input to UMAP. Shapes indicate the processing batch of each sample.

and dropped three samples on account of poor nucleosomal banding of their fragment size distribution (fig. S2G) and a further two samples that captured few cells. For the remaining samples, we observed a median of 5742 nonduplicate reads per cell (fig. S2H) and estimate that we sequenced a median of 88% of all nonduplicate reads per cell in these sci-ATAC-seq3 libraries (fig. S2I).

We identified peaks of accessibility on a tissue-by-tissue basis and then merged these to generate a master set of 1.05 million sites (data file S1). We filtered out lower-quality cells, which left 790,957 single-cell chromatin accessibility profiles from 53 fetal samples (data file S2). The total number of high-quality cells per tissue ranged from 2421 for spleen to 211,450 for liver (Fig. 1C). The median number of nonduplicate fragments per cell for this set is 6042, with a median of 49% overlapping the master set of accessible sites and 19% falling near a transcription start site (TSS) (±1 kb). We subjected high-quality cells to latent semantic indexing (*19*, *20*), linear correction (*21*), and Louvain clustering, initially obtaining 172 clusters across all tissues. We further reduced the dimensionality of each tissue dataset using Uniform Manifold Approximation and Projection (UMAP) (*22*).

## Annotating cell types

The annotation of cell types in scATAC-seq (sc, single cell) datasets can be simplified by leveraging scRNA-seq datasets (*13*, *23*–*25*). In order to partially automate cell type annotations for our sci-ATAC-seq data, we first annotated cell types within our sci-RNA-seq data for the same tissues (*16*). Second, we computed gene-level accessibility scores for our sci-ATAC-seq data, aggregating the number of transposition events falling within gene bodies extended by 2 kb upstream of their TSS. Third, we used the gene-by-cell matrices for each data type as input to an approach for finding likely correspondences between clusters on the basis of non-negative least squares (NNLS) regression (*26*), effectively resulting in a "lift-over" set of automated annotations for our sci-ATAC-seq clusters. Last, we manually reviewed these automated annotations by examining pileups around marker genes for each cell type within each tissue, making modifications to assigned labels as deemed necessary (Fig. 2A and fig. S3A). Although other approaches have shown considerable promise for multimodal integration of single-cell data (*23*), we found this cluster-to-cluster NNLS method (*26*) sufficient for our purposes here and much less computationally intensive.

Altogether, we were able to annotate 150 of the 172 clusters (87%), or 163 of 172 (95%) if we include lower-confidence labels. Some clusters received the same annotation within the same tissue and were merged, resulting in 124 annotations across all tissues. Of these, some annotations were present across multiple tissues (Fig. 2B). Collapsing across tissues resulted in 54 distinct cell type annotations that map 1:1 to "main cell type" annotations made in our sci-RNA-seq dataset (or 59 if we include lower-confidence labels and 1:2 mappings) (Fig. 2B). Many of the sci-RNA-seq cell types that were not found in the sci-ATAC-seq data at this level of resolution are small clusters that may not have been sufficiently sampled to be detectable, owing to the lower number of cells profiled here [~4 million RNA (*16*) versus ~800,000 ATAC high-quality cells] (fig. S3B). However, most of the nine sci-ATAC-seq clusters that remained fully unannotated appear to be due to unfiltered doublets because they are characterized by accessibility in marker genes for multiple

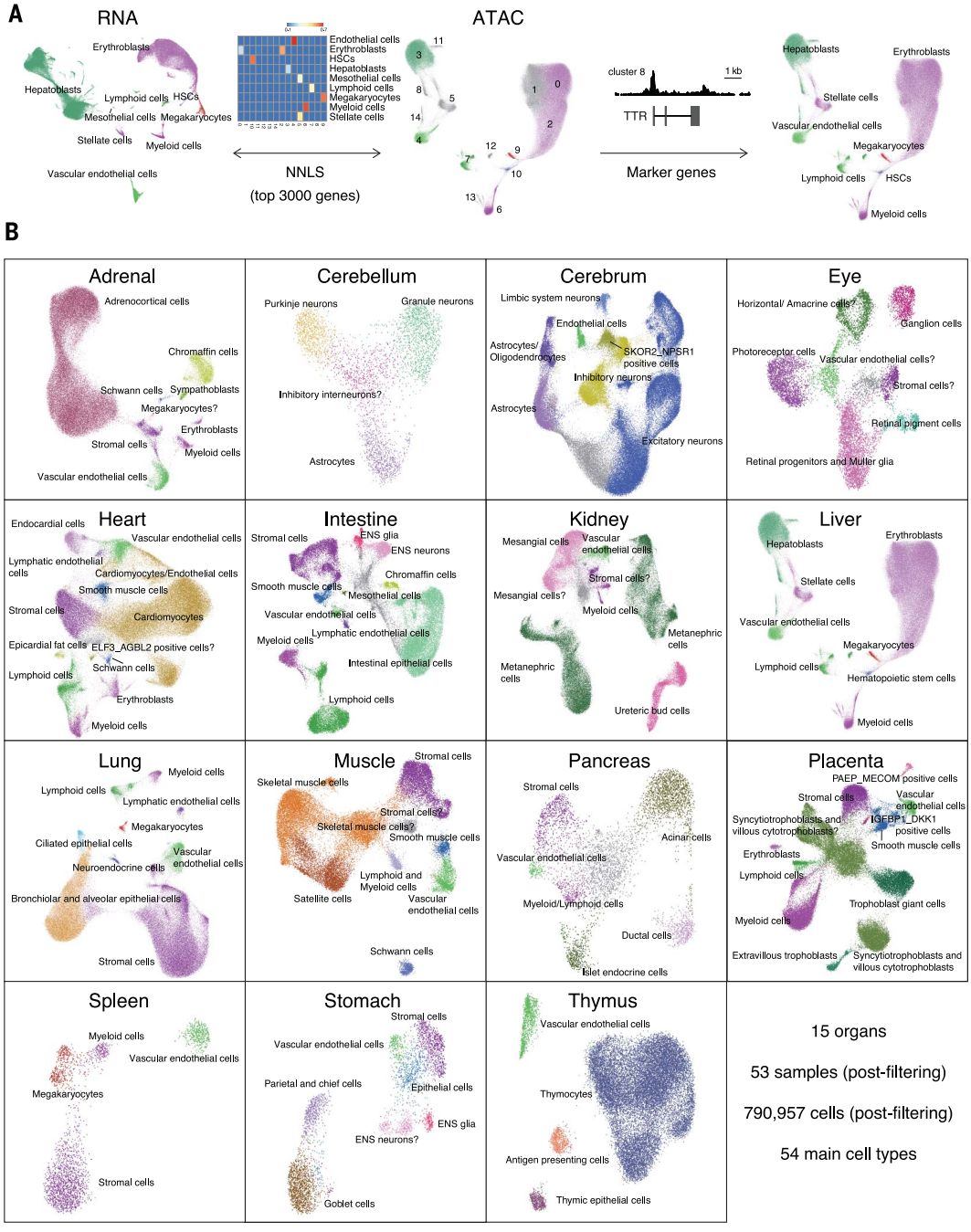



**Fig. 2. Identifying cell types across 15 human organs.** (**A**) Summary of annotation strategy. (Left) Cell types were first annotated in sci-RNA-seq data (*16*) gathered from matching tissues according to marker gene expression. (Middle) Louvain clusters were identified in sci-ATAC-seq data for each tissue. Next, gene-level accessibility scores were calculated for each of these clusters and matched to RNA clusters on the basis of NNLS regression, in some cases leading to merging of Louvain clusters. (Right) These first-pass automated annotations were refined by manually reviewing the cluster-specific accessibility landscape around marker genes—for example, initially unannotated cluster 8 exhibited specific accessibility at the *TTR* locus—and was therefore merged with cluster 3 (hepatoblasts). (**B**) UMAP visualization and annotation of 790,957 cells profiled across 15 organs. The colors correspond to the 54 main cell types that were identified across the different organs.

adjacent cell types in the UMAP representation (fig. S3A).

The nature of ATAC-seq data allows sexing of cells on the basis of Y chromosome reads. In the placenta in particular, we found that three cell types—*PAEP*$^+$, *MECOM*$^+$ and *IGFBP*$^+$, *DKK*$^+$ cells (both initially unannotated in the RNA data, although the labels readily lifted over to clusters in the ATAC data), as well as placental lymphoid cells—exhibited a significantly lower ratio of Y chromosome–derived reads in tissues derived from male fetuses (fig. S3C). Consistent with what is known about *PAEP* (glycodelin) and *IGFBP1*, these cell types likely correspond to maternally derived endometrial epithelial and decidualized stromal cells, respectively (*27*). This was confirmed with genotype inference with souporcell (*28*), which additionally identifies a subgroup of placental myeloid cells as likely to be of maternal origin (fig. S3D).

### Identifying cell type–specific TFs

We next sought to integrate and compare chromatin accessibility in cell types across all 15 organs. To mitigate the effects of gross differences in the numbers of cells per organ and/or cell type, we randomly sampled 800 cells per cell type per organ (including unannotated clusters; in cases in which fewer than 800 cells of a given cell type were represented in a given organ, all cells were taken), and we performed UMAP visualization (Fig. 3A). Reassuringly, cell types represented in multiple organs clustered together—for example, stromal cells (nine organs), endothelial cells (13 organs), lymphoid cells (seven organs), and myeloid cells (10 organs)—rather than by batch or individual (fig. S4). Developmentally and functionally related cell types also colocalized, such as diverse blood cells, secretory cells, peripheral nervous system neurons, and central nervous system neurons.

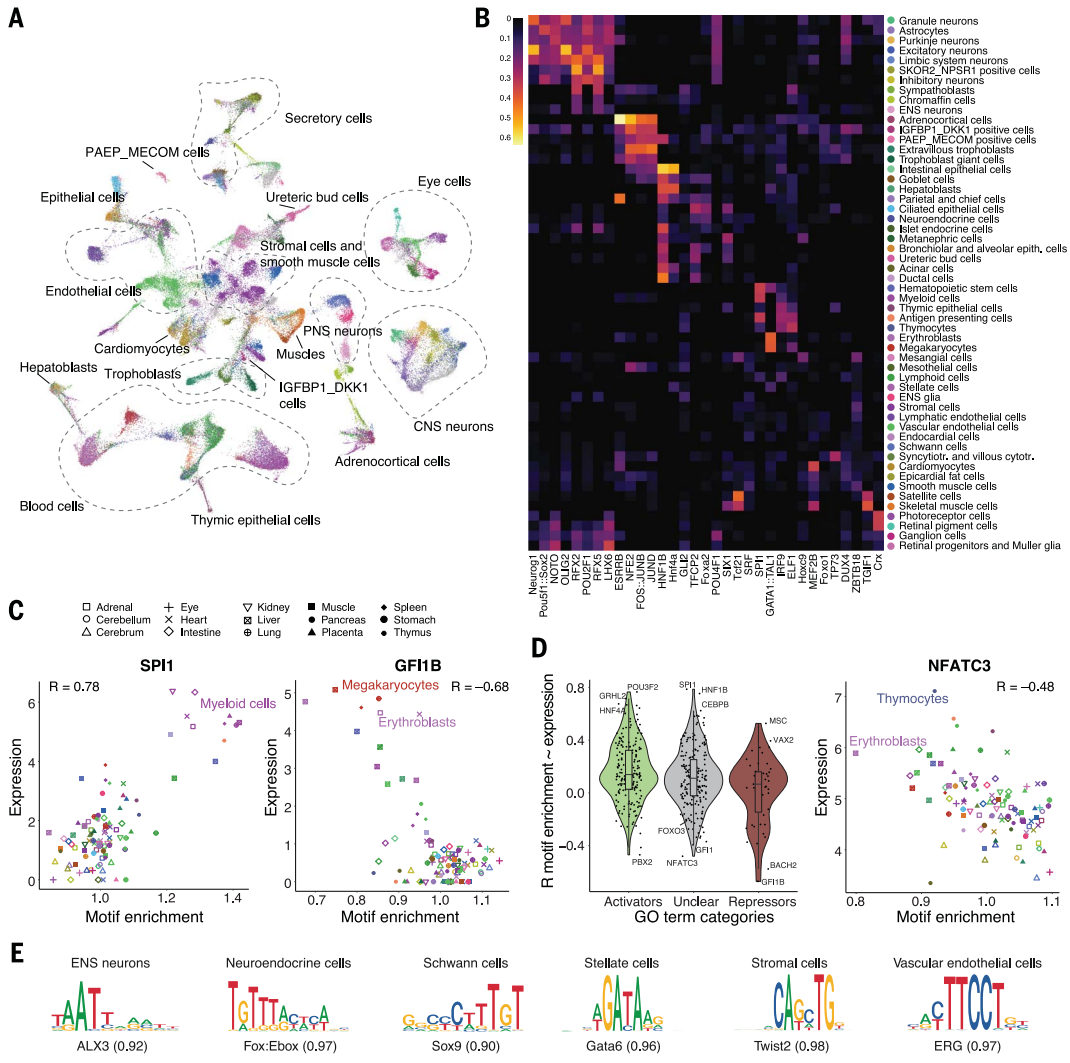



**Fig. 3. Identifying key TF regulators of cell type–specific chromatin accessibility and their modes of action.** (**A**) Combined UMAP of the entire dataset subsampled to a maximum of 800 cells per initial cluster ID. Cells are colored by 54 main cell types as in (B). Groups of related cell types are circled. (**B**) Fold-change of the top enriched TF motif in cell type–specific peaks for all 54 main cell types. Cell types (rows) are ordered by hierarchical clustering of the motif enrichment matrix (log10-scaled fold-change of the mean motif occurrence in peaks of this cell type relative to the rest of the dataset, $q < 0.01$). Additional enriched TF motifs in cell type–specific peaks are provided in data file S3. (**C**) Examples of an (Left) activating versus (Right) repressive TF whose expression levels are positively or negatively correlated with motif accessibility across cell types and tissues. Each point indicates a cell type from a specific tissue [color code as in (B); shape code above]. Motif enrichment corresponds to fold-change of the mean motif occurrence in peaks of this cell type relative to the rest of the dataset. Expression values for the TFs are from sci-RNA-seq data collected in matching cell types as described in (*16*) (natural log of CPM+1). Correlation coefficient ($R$) values are Pearson correlations. (**D**) Correlation of motif enrichment and expression can be used to predict the mode of action of unclassified TFs. (Left) TFs were automatically assigned to the category of activator, repressor, or unclear on the basis of their associated GO terms. Pearson correlation values of motif enrichment and TF expression were calculated across all cell types in all tissues and are shown by category for all 455 TFs for which we have both values. Most TFs show positive correlation values. Annotated repressors have lower median $R$ values than those of activators, with many of the outliers being due to missing or misleading GO term annotations. (Right) A high absolute $R$ value can serve to classify TFs with unknown mode of action. An example is NFATc3, a likely repressor based on this analysis. (**E**) Position weight matrices (PWMs) identified by de novo motif search for exemplary cell types with no strong enrichment in (B). De novo motif enrichment was performed with homer (*48*) in the 2000 most specific peaks for each cell type by using CpG-matched genomic sequences as background. The closest known motif and the score for the motif matching process are indicated below. Further details as well as PWMs for all cell types are provided in fig. S7.

A central question in developmental biology is which TFs are involved in generating and maintaining a diversity of cell types from an invariant genome. We sought to leverage these data to systematically assess which TF motifs are differentially accessible and thus nominate key regulators of cell fate specification and/or maintenance in the context of in vivo human development. Differential motif accessibility is not proof of TF binding, so further experimental validation will be needed to confirm the below observations.

As a first approach, we used a linear regression model to ask which TF motifs found in the accessible sites of each cell best explain its cell type affiliation. Initially treating each tissue independently, we identified the most highly enriched motifs and TFs from the JASPAR database for each of 124 cell type clusters across all tissues, which revealed both known and potentially previously unknown regulators (fig. S5). For example, in the placenta, the motif of SPI1/PU.1, an established regulator of myeloid lineage development (*29*), is highly enriched in peaks of myeloid cells; the motif of TWIST-1, which is required for the formation of stromal progenitors (*30*), is enriched in peaks of stromal cells; and the FOS::JUN motif is associated with chromatin accessibility in extravillous trophoblasts, a cell type in which the corresponding AP1 complex has been described to be specifically active (*31*, *32*).

An unannotated cluster within the placenta is enriched for GATA1::TAL1 motifs, which are established regulators of erythropoiesis (*33*). These cells cluster with erythroblasts from other tissues in the global UMAP (Fig. 3A and fig. S6A), and upon further inspection, key erythroid marker genes exhibited specific promoter accessibility (fig. S6B). In the NNLS-guided workflow, this cluster was not annotated because an erythroblast cluster was not detected in the

placenta in the sci-RNA-seq study [possibly because the placenta is one of the few tissues for which we have more cells with ATAC than RNA data (*16*)]. Thus, motif enrichment can assist in cell type annotation, if the key regulators of a cell type are known.

We repeated this regression analysis on the 54 main cell types observed across all tissues, after collapsing cell types appearing in multiple tissues (Fig. 3B and data file S3; descartes.brotmanbaty.org) (*15*). As expected, the top motifs remained consistent with the tissue-specific analyses as well as the literature—for example, SPI1/PU.1 in myeloid cells (*29*), CRX in retinal pigment and photoreceptor cells (*34*), MEF2B in cardiomyocytes and skeletal muscle cells (*35*), and SRF in endocardial and smooth muscle cells (*36*). Whereas most motifs are enriched in only one or two cell types, neuronal TF motifs (*37*–*39*) are enriched in multiple neuronal cell types (Fig. 3B, top left cluster). Another exception to the cell-type specificity of motifs is HNF1B, which is conventionally associated with kidney and pancreas development (*40*, *41*) and whose motif is enriched in 13 cell types that span a range of specialized epithelial and secretory roles (*34*).

POU2F1 (POU class 2 homeobox 1) is an example of a TF that has not previously been associated with a particular developmental branch but rather has been suggested to be an exception within the POU family—broadly expressed and controlling no specific trajectory (*42*). By contrast, we found that in developing human tissues, its motif is enriched in several neuronal cell types. Lending further support, *POU2F1* is more highly expressed in those same cell types (fig. S6C).

Extending on this observation, we sought to leverage an atlas of gene expression (*16*) to more generally ask whether TFs are differentially expressed in a pattern consistent with the differential accessibility of their motifs. For example, looking across all cell types annotated in the same tissue in both datasets, the expression of the myeloid pioneer factor SPI1/PU.1 is strongly positively correlated with the enrichment of its motif at accessible sites (Fig. 3C, left). This analysis also revealed TFs with a negative correlation between their expression and motif enrichment (table S2). Upon closer inspection, these TFs tended to be repressors. For example, GFI1B has been described to act as a repressor crucial to erythroblast and megakaryocyte development by recruiting histone deacetylase upon binding its motif and inducing closing of the chromatin, such as at the embryonic hemoglobin locus (*43*). Consistent with this, we observed its expression to be negatively correlated with its motif enrichment at accessible sites (Fig. 3C, right).

Categorizing TFs as "activators" or "repressors" from GO terms, we found that TF expression and motif accessibility tend to be positively correlated for annotated activators and negatively correlated for annotated repressors (Fig. 3D, left). Exceptions can largely be explained by missing or conflicting GO terms, whereas literature searches are consistent with the observed correlation. Accordingly, this kind of analysis provides a systematic approach for classifying TFs as activators or repressors. For example, NFATc3 is generally described as an activator (*44*), but our analysis points toward a repressive mode of action, especially in developing T cells, where it is highly expressed yet its motif is depleted in accessible sites (Fig. 3D, right). Apart from a general classification, we also gained insight into the cell-type contexts in which a TF might variably act as an activator or repressor. For example, TFs including FOXO3 have been proposed to act as activators in their unmodified state but as repressors when phosphorylated (*45*), which might explain its more ambiguous relationship between expression and accessibility (fig. S6D). We only classified TFs as repressors if their presence is linked to a reduction in accessible chromatin, yet there are also TFs that have been reported to repress transcription while maintaining an accessible state at their binding sites, such as REST (*46*, *47*). This group of repressors is not distinguished from activators by our analysis (fig. S6E) because this would require further linking to the transcriptional effect on target genes.

A limitation of the above-described linear regression strategy for associating cell types with TF motifs is that it relies on databases of known TF motifs. As a different approach, we calculated specificity scores for each accessible site (*13*), selected the 2000 most specific peaks for each cell type, and searched de novo for enriched motifs within this set compared with CpG-matched background genomic sequences (fig. S7 and data file S4) (*48*). In general, the top de novo motifs for individual cell types agree with the top known motifs identified with linear regression. Some cell types that did not have strong matches to known motifs by means of the regression strategy were nonetheless associated with de novo motifs (such as endothelial, stromal, and Schwann cells) (Fig. 3E, and fig. S7). For endothelial cells in particular, this result is discussed further below.

### Cross-tissue analyses of blood cells and endothelial cells

The nature of this dataset creates an opportunity to investigate organ-specific differences in chromatin accessibility within broadly appearing cell types such as blood and endothelial cells. In our first pass of cell type annotations for blood cells, we were able to differentiate between myeloid cells, lymphoid cells, erythroblasts, megakaryocytes, and hematopoietic stem cells (Fig. 2B). Extracting and reclustering these blood lineages from all organs allowed us to additionally identify macrophages, B cells, natural killer (NK)/type 3 innate lymphoid (ILC3) cells, T cells, and dendritic cells, once again adopting an RNA-assisted annotation approach (analyzing similar cell types from multiple tissues necessitated an additional doublet cleaning step) (Fig. 4A). Macrophages could be further separated into groups associated with tissue of origin, as previously observed (*49*), as well as phagocytic macrophages. This latter group was identified mainly in the spleen, followed by the liver and the adrenal gland (fig. S8A). In contrast to the RNA, we did not detect a separate microglia cluster in cerebrum, likely because this is a very rare cell type (~0.25%) (*16*).

Of particular interest within the blood lineages are erythroblasts, owing to the spatiotemporal dynamics of erythropoiesis during fetal development. We initially detected this lineage in the liver, adrenal gland, heart, and placenta (Fig. 2B); our cross-tissue analysis additionally identified erythroblasts in the shallowly profiled spleen (where only megakaryocytes and myeloid cells were originally annotated). The ratio of erythroblasts within the blood lineages of a tissue is highest in the liver, which is in line with this organ being the primary site of erythropoiesis at this developmental stage, followed by the spleen and adrenal gland (fig. S8A) (*50*, *51*), phenocopying the trend in the RNA data described in (*16*).

Further investigating erythroblasts, we observed that regions proximal to both the adult β- and fetal γ-globin genes are accessible at this stage of development, whereas the embryonic ε-globin gene's promoter is inaccessible (fig. S8B). The erythroblast cluster could be further subdivided into five major Louvain clusters with differential chromatin accessibility, including a distinct erythroblast progenitor cluster (Fig. 4A and fig. S8A). Accessible sites in the erythroblast progenitor cluster as well as in the adjacent early erythroblast cluster (erythroblast_3) are enriched for GATA1::TAL1 as well as other GATA motifs (Fig. 4B). Comparison of expression levels of various GATA factors in erythroblast progenitors allows us to nominate GATA1/2 as the TFs likely responsible for this motif enrichment (fig. S8C). The other erythroblast clusters, corresponding to later stages of erythropoiesis, show motif enrichment for NFE2/NFE2L2 (erythroblast_1) and NFYB/KLF1 factors (erythroblast_2/4) but a marked absence of enrichment for GATA motif accessibility. A scRNA-seq study on the mouse hematopoietic system reported induction of *GATA2* early in erythropoiesis, with a subsequent decrease in *GATA2* yet stable *GATA1* expression (*52*). By contrast, a study of sorted bulk human in vitro cultured erythroid populations revealed a decrease in *GATA1* expression from progenitors to differentiated erythroblasts, as well as increased *KLF1* and *NFE-2* levels in

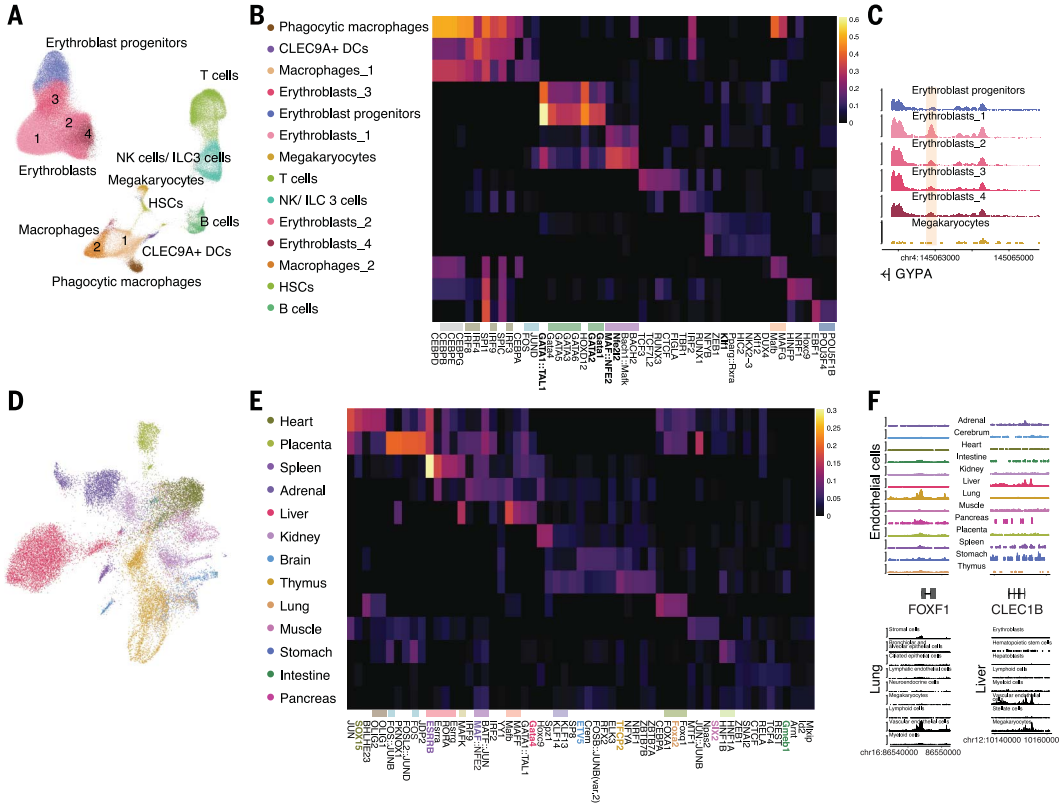



**Fig. 4. Identifying major subgroups and associated TFs in broadly distributed lineages.** (**A**) UMAP visualization of 152,649 blood cells extracted from all organs, colored and annotated by Louvain clusters. (**B**) Five TF motifs most strongly enriched in peaks of each Louvain cluster in (A) (log10-scaled fold-change of the mean motif occurrence in peaks of this cluster relative to the rest of the dataset, $q < 10^{-6}$). Highly similar motifs, as determined from RSAT matrix-clustering of the JASPAR vertebrate motif collection (*116*), are indicated with horizontal bars. (**C**) Example locus upstream of *GYPA* with differential accessibility across erythroblast populations. Accessibility is summed for all cells in each Louvain cluster, and the scale is normalized to account for differences in total reads per cell as well as cell numbers across clusters. Other blood cell types, including megakaryocytes (shown), have negligible accessibility at this region. (**D**) UMAP visualization of 27,576 vascular endothelial cells extracted from all organs and colored by tissue of origin. Colors are as in (E). Only the top 20,000 endothelial-specific peaks as determined in each tissue were used for clustering, merged to 94,023 distinct peaks across all tissues. (**E**) Five TF motifs most strongly enriched in peaks of each tissue group in (D) (log10-scaled fold-change of the mean motif occurrence in peaks of this tissue group relative to the rest of the dataset, $q < 10^{-4}$). Highly similar motifs, as determined from RSAT matrix-clustering of the JASPAR vertebrate motif collection (*116*), are indicated with horizontal bars. Motifs whose TFs (or TFs with highly similar motifs) are most highly expressed in endothelial cells from the same tissue in sci-RNA-seq data are highlighted (colors correspond to tissues). (**F**) Example loci showing specific accessibility in (left) lung or (right) liver endothelial cells. These sites also exhibit tissue-specific accessibility in their tissue of origin (bottom) and thus are unlikely to be consequent to residual doublets or free DNA contamination from other cell types. The *CLEC1B* locus is also accessible in the small cluster of megakaryocytes in liver and is known to be expressed in platelets (*117*). Accessibility is summed for all cells in each Louvain cluster, and the scale is normalized to account for differences in total reads per cell as well as cell numbers across clusters.

later-stage erythroblasts (*53*). Our observations align with the bulk in vitro human data on this point and indicate that there might be epigenetically distinct subpopulations of differentiated erythroblasts (subclusters 1, 2, and 4) in which the accessibility landscape is shaped by non-GATA factors (Fig. 4B). For example, a distal regulatory element upstream of *GYPA*, which is used as an erythrocyte invasion receptor by the malaria parasite (*54*), is most accessible in the erythroblast_1 population and contains a motif that resembles the NFE-2 motif (Fig. 4C).

Pseudotime analysis of hematopoietic stem cells (HSCs) and erythroblast subpopulations confirmed the order of progenitors and early erythroblasts in the HSC-to-erythroblast transition; late erythroblast clusters exhibited similar median pseudotimes, suggesting that they might represent subpopulations of differentiated erythroblasts rather than a succession of states (fig. S9, A and B) (*55*). This analysis also nominated candidate regulatory elements that open or close over the course of erythropoiesis (fig. S9C). Some of the top HSC- or erythroid progenitor–specific peaks (fig. S9C) are also accessible in bulk DNase profiles of fetal—but not adult—adrenal tissue (*56*–*58*), supporting the adrenal gland as a site of fetal hematopoiesis during normal mammalian development (fig. S9D) (*16*).

Another pervasive cell type is the vascular endothelium, which needs to perform both constitutive and highly specialized functions across organs, such as gas exchange in the lung or fluid filtration in the kidney. No TF has been described to be exclusively expressed in vascular endothelial cells, suggesting that the endothelial-specific transcriptome is controlled combinatorially by several TFs with overlapping expression in the endothelium (*59*). Consistent with this, we failed to observe any single, strong enrichment in endothelial cells in our analysis of JASPAR motifs (Fig. 3B). However, de novo motif discovery on the 2000 most endothelial-specific peaks revealed enrichment over background genomic sequences for motifs resembling ERG [E-26 transformation–specific (ETS)–related gene] and SOX15 [SRY (sex determining region Y)–box 15] (fig. S7). These motifs were likely not weighted as strongly in our linear regression approach because they are not restricted to endothelial cells (the ERG motif is enriched in megakaryocytes, and SOX15 is enriched in several cell types), nor is expression of these TFs limited to this cell type (fig. S10A). In line with this, ERG has been described as a major regulator of endothelial function (*60*) but also drives transdifferentiation into megakaryocytes (*60*, *61*).

We detected endothelial cells in 13 out of 15 organs, the exceptions being the more shallowly profiled cerebellum and eye (Fig. 2B). In contrast with erythroblasts (fig. S8A), extracting endothelial cells and reclustering revealed a marked separation according to tissue of origin (fig. S10B), in spite of stringent iterative filtering steps to remove residual contaminating doublets. Consistent with this, we also observed tissue-specific aspects of endothelial gene expression in fetal tissues (*16*) and previously

found regions exhibiting tissue-specific chromatin accessibility in adult mouse endothelial cells (*13*). To exclude technical sources for the tissue-specific signal, we selected the 20,000 most endothelial-specific peaks determined within each of the 13 tissues, merged these to 94,023 distinct peaks, and then clustered extracted endothelial cells on the basis of these peaks (Fig. 4D). The cells continued to cluster by tissue, similar to when we used all peaks (fig. S10B).

Further supporting tissue-specific differences in the endothelial regulatory landscape, endothelial cells derived from nearly all organs exhibited specific TF motif enrichments within these peaks (Fig. 4E). The TFs for many of the enriched motifs are also most highly expressed in endothelial cells of the matching tissue in the RNA data (Fig. 4E) (*16*). These analyses are limited to TF motifs present in the JASPAR vertebrate database, and additional TFs appear differentially expressed (*16*). Last, peaks of accessibility closest to differentially expressed genes have higher endothelial specificity scores in the matching tissue for about half of the profiled organs in the ATAC data (fig. S10C). Examples include *FOXF1*, which is specifically expressed and accessible in lung endothelium and whose promoter proximal region contains a FOXA2 motif; and *CLEC1B*, which is both specifically expressed in liver endothelium and harbors a GATA motif–containing candidate regulatory element exhibiting liver endothelium–specific accessibility (Fig. 4F). Some, but not all, of the enriched motifs are also enriched in other cell types of the same tissue. Although we cannot exclude residual contamination contributing to this signal, this might also reflect the underlying biology, for example, consequent to heterogeneous origins (*62*).

Overall, these findings indicate that the general program of chromatin accessibility and gene expression in endothelial cells, a widely distributed cell type that needs to fill both general and organ-specific functions, is mediated by a combination of constitutive TFs as well as tissue-specific TFs that may drive additional specialization. These analyses also highlight the merit of combining both de novo motif and linear regression approaches across tissues to nominate the key regulators that shape the chromatin landscape in individual cell types.

### A catalog of accessible elements in the human genome during development

Altogether, our master set of 1.05 million sites spans 532 Mb, or 17.1% of the reference human genome (data file S1). This extensive catalog of accessible sites enabled several additional analyses. First, we used Cicero to generate coaccessibility and gene activity scores (*63*), analyzing each of 54 cell types separately. Because some of these were represented in several tissues, 101 Cicero maps were generated altogether. In total, we tested 159 million distinct pairs of accessible sites within 500 kb. At a coaccessibility score threshold (*63*) of 0.1, we obtained a catalog of 6.3 million distinct coaccessible pairs of sites across the 101 maps, with an average of ~139,000 pairs per cell type. This catalog includes 1.4 million (22%) promoter-distal, 4.8 million (76%) distal-distal, and ~94,000 (1.5%) promoter-promoter candidate interactions (data files S5 and S6; descartes.brotmanbaty.org) (*15*). For example, as expected at this stage of development, erythroblasts, but not other cell types, exhibited coaccessibility between the locus control region (LCR) and the fetal and adult, but not the embryonic, β-globin gene (fig. S11A) (*64*). A second example is the *FOXF1* promoter (Fig. 4F), at which endothelial cells from the lung, but not other tissues, exhibited coaccessibility with nearby distal elements (fig. S11B).

Second, a substantial proportion of heritability for common human diseases and traits partitions to accessible chromatin, particularly to regions that are specifically accessible in tissues or cell types related to the trait or disease in question (*65*–*67*). We previously intersected genome-wide association study (GWAS) signals for diverse human phenotypes with an adult mouse single-cell atlas of chromatin accessibility and found many anticipated relationships to be discoverable despite the considerable species difference (*13*). We repeated such an analysis on these data, applying partitioned linkage disequilibrium score regression (LDSC) (*67*) to detect enrichment of human heritability for 34 phenotypes from the UK Biobank (UKBB) within accessible chromatin for each of our 54 fetal cell types (Fig. 5A and table S3). Of the 54 cell types, 45 had a significant enrichment for at least one phenotype, whereas 32 of 34 phenotypes were enriched for at least one cell type (the exceptions being basal metabolic rate and sunburn, the latter in line with absence of skin tissue). As expected, for example, blood cell traits are maximally enriched in blood cell types, neurological phenotypes in neuronal cell types, and high cholesterol in hepatoblasts and intestinal epithelial cells. Further, type 2 diabetes is not only enriched in islet endocrine cells but also in pancreatic acinar and ductal cells, hepatoblasts, and stomach goblet cells; menopause age is maximally enriched in adrenocortical cells (fig. S11C). As similar single-cell atlases of chromatin accessibility are generated across the human life span, it will be interesting to explore at what time points these enrichments are maximal for each phenotype.

Third, we sought to evaluate the suitability of these data for identifying genetic variants that affect chromatin accessibility in cis. Although we generated data on many cells and tissues, they were collected from a relatively limited number of individuals, precluding the possibility of using an association framework. Instead, we sought to identify allelic imbalance within individuals at heterozygous positions (*68*). Specifically, we tested the liver and brain sample from two individuals, aggregating the reads for all cells from each cell type and testing for allelic imbalance across these aggregate measures. Overall, we found 586 single-nucleotide polymorphisms (SNPs) that exhibited a significant allelic imbalance [20% false discovery rate (FDR)] (tables S4 and S5). In general, the number of significant sites identified correlated with the number of reads from that cell type (fig. S12, A and B), and consequently, there were large differences in the power to detect allelic imbalance across cell types (fig. S12, C to F). Of the SNPs that were heterozygous in both individuals, sites that had significant imbalance in one individual were strongly enriched for significant imbalance in the same tissue in the other individual (49-fold over random for brain, hypergeometric test $P = 1.5 \times 10^{-36}$; 59-fold over random for liver, hypergeometric test $P = 2.3 \times 10^{-60}$), although there was a greater degree of sharing between the liver and brain of the same individual (69-fold over random for one individual, hypergeometric test $P = 1.2 \times 10^{-77}$; 78-fold enrichment, hypergeometric test $P = 5.7 \times 10^{-44}$ for the other individual). Although not significantly enriched ($P = 0.059$), 25 SNPs with allelic imbalance in at least one cell type were previously associated with complex traits in the National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS catalog (table S6) (*69*). For example, rs61851769 shows allelic imbalance in erythroblasts and hepatoblasts in one liver sample and was previously associated with mean corpuscular hemoglobin (Fig. 5B) (*69*, *70*). The SNP disrupts a TAL1 binding site and is upstream of *SLC30A1*, a gene implicated in erythropoiesis (*71*). Consistent with the erythroblast-specific nature of these annotations, we believe that the hepatoblast signal may come from contaminating erythroblasts because hepatoblast accessibility is lost after peak module–based doublet filtering. Another example is rs362649, which is significant in excitatory neurons of one individual, was previously associated with the volume of cerebellar vermal lobules VIII to X (*72*) and lies within an intron of *RELN*, which plays a role in neuronal migration (Fig. 5B) (*73*). There are many caveats to these analyses, including the large differences in power across cell types. Nonetheless, these results illustrate how single-cell chromatin accessibility data might be leveraged for the identification of functional noncoding genetic variation with cell-type resolution.

Fourth, analogous to grouping cells on the basis of their shared patterns of accessibility across sites (Fig. 3A), we can instead group sites by their shared accessibility across cells (*74*, *75*). To reduce the computational complexity of this task, we removed sites of <400 base pair (bp) width and then computed a "UMAP of

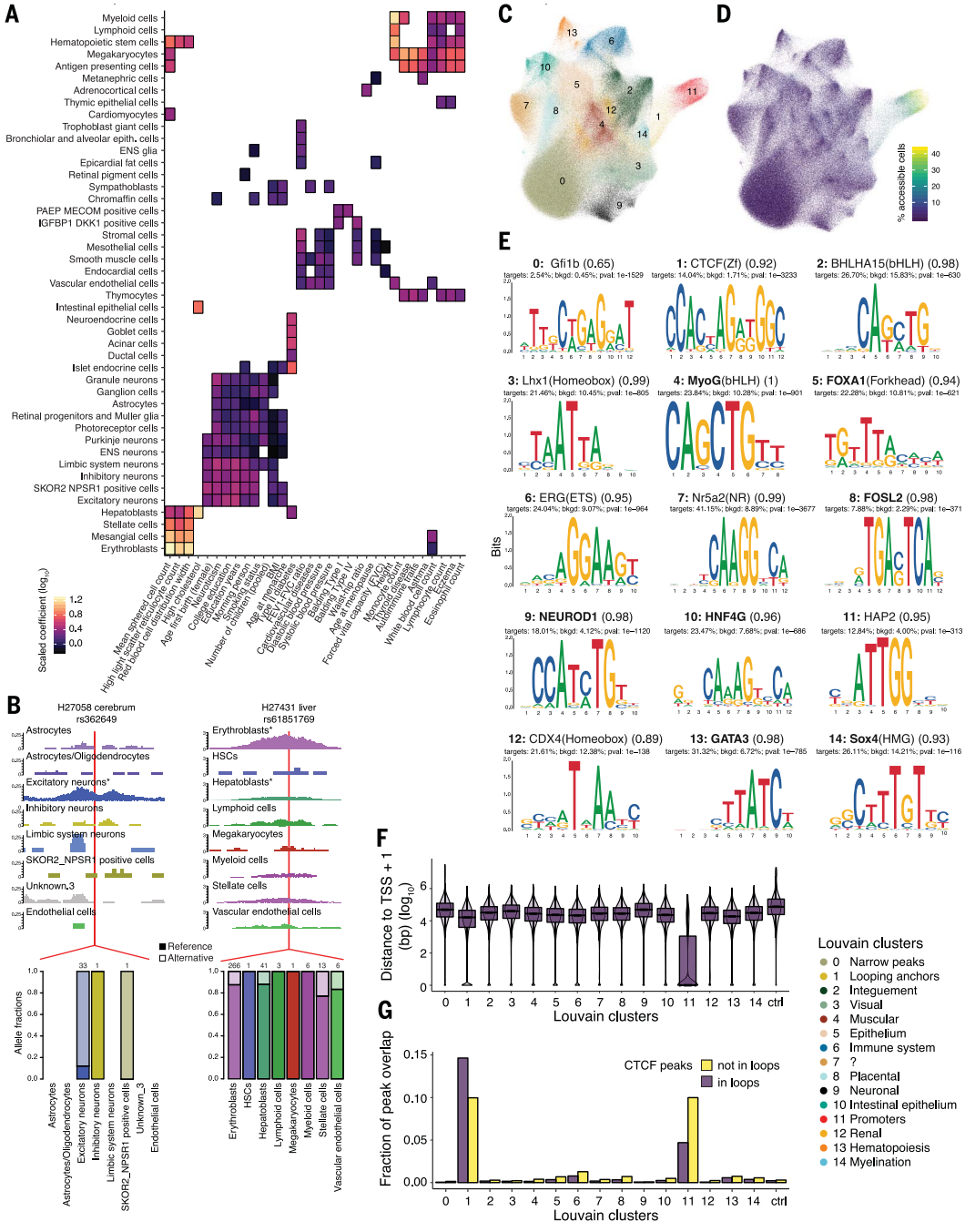



**Fig. 5. Heritability enrichment and coaccessibility of candidate regulatory regions.** (**A**) Enrichment of heritability for UK Biobank traits within top 10,000 specific sites for each cell type. Trait–cell type pairs with no significant positive enrichment ($q > 0.2$) are white. A full table of scaled coefficients and $q$ values for each trait–cell type pair is provided in table S3. (**B**) Example sites with allelic imbalance. Browser tracks of accessibility for the cell types in a (left) cerebrum and (right) liver sample are normalized to counts per million reads. Results are presented as unsmoothed base coverage. Asterisks indicate cell types with significant allelic imbalance. The red vertical line indicates the position of the SNP exhibiting allelic imbalance. The bar plots below show the relative portions of reads mapping to the reference and alternative allele at that position. Above each bar is the number of reads overlapping the SNP for each cell type. (**C**) UMAP visualization of a subset of accessible regions from the master set that are >400 bp (447,879 sites), by using accessibility profiles from the subsampled cell dataset in Fig. 3B (88,983 cells). Sites are colored by Louvain clusters, which are numbered according to decreasing size and annotated into broad categories on the basis of motif enrichment and lineage affiliation of enriched cells. Legend is at bottom right of the overall figure. Cluster 0 consists of narrower sites with the lowest accessibility across cells, is not enriched for a clear motif, and possibly reflects rare or transient cell states or biological or technical noise. (**D**) Same as (C), but colored by the percentage of cells in which sites are accessible. A version in which the accessible percentage is binned by content is shown in fig. S13C. (**E**) PWMs identified by means of de novo motif search in each of the clusters in (C). De novo motif search was performed with homer (*48*), using CpG-matched genomic sequences as background. The top PWM per cluster 0 to 14 is labeled by the closest known motif as determined by homer, with the score for the motif matching process indicated in brackets. Listed below are the percentage of sites within the cluster and CpG-matched background sequences that contain a match to the de novo PWM, and a $P$ value for the enrichment. Motifs associated with pioneer TFs are in boldface. The top motif for cluster 0 is only found in 2.5% of sites and has a poor matching score. (**F**) Violin plots of the distances of each group of sites in (C) to the nearest TSS is shown. 20,000 random regions located on autosomes with a width corresponding to the median width of all sites in (C) were used as control (ctrl). (**G**) Fraction of sites within each cluster overlapping with ENCODE-defined CTCF-bound peaks within versus outside of looping regions. All CTCF ChIP-seq peaks overlapping CTCF motifs in looping regions in GM12878 ($n = 8253$ peaks) and the same number of ChIP-seq peaks not overlapping looping regions but with the same ChIP-seq score were selected. For each cluster in (C), the fraction of sites overlapping these two CTCF-bound sets was calculated. The same control (ctrl) regions as in (F) were used.

sites," grouping 447,879 regions into 15 clusters (Fig. 5, C and D). Applying the aforedescribed linear regression and de novo motif search strategies, most of these 15 clusters were enriched for key TF regulators identified by our earlier analyses (Fig. 5E, fig. S13A, and data file S7). Correspondingly, when we determined "differential cells" (analogous to determining differential genes or peaks in conventional clustering of single-cell data), we found that cells from lineages that match the motif enrichments define most of these clusters (fig. S13B). Thus, most of these clusters represent sites specifically accessible in certain cell types or cell-type groups and therefore link to cell type–defining TFs.

The top cluster-defining TFs identified through de novo motif search include several pioneer factors, implying that sites bound by these TFs are more likely to be concurrently accessible.

However, a few of the clusters of sites were not enriched in a pattern that reflected a specific lineage. For example, cluster 11, comprising 10,983 or 2.5% of sites, clearly corresponds to commonly accessible promoters: Its sites are accessible in many cells (Fig. 5D and fig. S13C); 75% are within 1 kb of a TSS (Fig. 5F); and they are broader, CpG rich, and conserved (fig. S13D). In addition, this cluster is strongly enriched for motifs commonly found in promoters—such as various SP factors, KLF factors, NRF1, and ZFX (fig. S13A)—and the top identified de novo motif corresponds to the CCAAT promoter element (Fig. 5E). In particular, this cluster is enriched for housekeeping gene promoters [1.9-fold enriched, hypergeometric test $P = 6.5 \times 10^{-244}$; 80% of 3006 housekeeping TSSs defined by (*76*) are in this cluster].

Another case is cluster 1, whose 41,128 sites are not as commonly accessible as those of promoters (Fig. 5D) but are nonetheless less cell type–specific than other clusters (fig. S13B). These sites also have higher CpG content and are modestly broader and slightly nearer to TSSs than other nonpromoter clusters (Fig. 5F and fig. S13D). Although this might reflect a cluster of sites containing some promoters, motifs of promoter TFs are depleted in cluster 1 (data file S7). Its only significantly enriched motif is CTCF (Fig. 5E and fig. S13A). This suggests that these coaccessible sites correspond to TAD (topologically associating domain) boundaries and looping anchors, which are known to bind CTCF and to be largely but not entirely invariant across cell types (*77*).

To evaluate this hypothesis, we obtained CTCF-bound peak locations from ENCODE, as determined with chromatin immunoprecipitation–sequencing (ChIP-seq), as well as loop anchor locations from Hi-C data in GM12878 (*78*), and calculated the overlap of each cluster of sites with CTCF-bound peaks within versus outside of looping anchors (Fig. 5G). Most clusters showed limited overlap. A first exception was cluster 11 (promoters; 10% overlap with nonlooping peaks), which is in line with 20% of CTCF sites falling in promoters (*79*). A second exception was the CTCF-enriched cluster 1 (15% overlap with looping peaks, a number that would likely increase if Hi-C and ChIP-seq data from all profiled cell types were available). This was also the only cluster with greater overlap with looping than nonlooping CTCF-bound peaks. Taken together, profiling chromatin accessibility across many tissues reveals not only cell types but also sets of coaccessible regulatory elements—mostly lineage-specific sets, but also promoters and looping regions.

Fifth, we compared our master list of sites to orthogonally annotated functional regulatory regions in the human genome and accessible regions in other species. Of human accelerated regions (*80*), 66% overlap one of our peaks, as do 75% of human VISTA enhancers (*81*). Non-overlapping VISTA enhancers are slightly enriched for an absence of expression in transgenic mouse assays (1.2-fold; hypergeometric test $P = 6.9 \times 10^{-8}$). Peaks that we assigned to the visual, neuronal, and looping categories (Fig. 5C) are enriched for overlap with both human VISTA enhancers and accelerated regions, whereas narrow, rarely accessible peaks are depleted (fig. S14, A and B). We also compared our master list of sites to the peak set generated by profiling chromatin accessibility in 13 tissues from 8-week-old mice (*13*). Of the 23% of these mouse peaks that lift over to the human genome, 60% (61,396) overlap a human peak. The overlapping human peaks are significantly enriched for peaks associated with neuronal or myelination cell types, looping anchors, and promoters but not other cell types (such as immune or hematopoiesis); narrow rare peaks are depleted, as are placental peaks (placenta was not profiled in the mouse atlas) (fig. S14C). The result is consistent with the possibility that regulatory sites of some broad categories of cell types (such as neuronal cells) may have experienced less evolutionary turnover between mouse and human than others (such as immune cells) (*17*, *82*).

### Comparisons of accessibility across developmental stages

We next asked whether cell type–specific motif enrichments are shared across developmental stages. Many similar cell types show similar top motifs enriched in the mouse ATAC atlas, which was generated by profiling 13 tissues in 8-week-old mice (*13*), implying that these TFs have a role in cell fate maintenance that may be conserved across species (mouse versus human) as well as developmental stage (adult versus fetal) (fig. S15, A and B). POU2F1—the motif we suggest to be important for neuronal cells—is enriched in accessible sites of mouse excitatory neurons, in addition to B cells (fig. S15A). Motif enrichment patterns cluster largely by cell type rather than species in a shared heatmap (fig. S15C), with some exceptions. For example, whereas myeloid cells cluster together, mouse lymphoid cells cluster separately from human lymphoid cells, in part driven by a more pronounced enrichment of NFKB1/2 motifs in the mouse. This could be due to the difference in developmental stage because NFKB1 has been shown to be dispensable for the emergence of prenatal B-1 transitional cells yet essential later in development (*83*).

To investigate human developmental stage–specific chromatin accessibility, we compared our dataset with existing single-cell ATAC datasets in adult human tissues, namely blood and cortex (*84*, *85*). To remove strong batch effects observed, we selected overlapping peaks in the adult dataset, rescored our data on the basis of this peak set, identified anchors, and integrated the two datasets (fig. S16A) (*23*). After applying this integration strategy, blood cells clustered by cell type rather than stage, with fetal cells falling closer to naive subtypes in the UMAP visualization (fig. S16B). As with the comparison with the mouse atlas, and as expected given the relatively late stage of development that we were interrogating, we observed similar motifs enriched in many blood cell types, with some differences (fig. S16C). Again, adult B and T cells are more strongly enriched for NFKB1/2 (1.5- to 1.6-fold for adult B and T cells and 1.1-fold for fetal B cells; fetal T cells showed no enrichment). However, such comparisons are hampered by strong batch effects owing to different sample collection and processing as well as the removal of potentially meaningful dataset-specific differences in the integration workflow.

In our comparison of the developing versus adult cortex data (*85*), again several cell types overlap in the integrated UMAP representation (fig. S17A). However, some fetal subgroups, including the two largest excitatory neuron subgroups, do not overlap with the adult data (subgroups 1 and 2) (fig. S17B). The fetal cerebral UMAP contains more substructure than we annotated (as do other tissues and cell types), evidenced by cluster-specific accessibility at known neuronal subtype marker genes (Fig. 2B and fig. S17C). For a more in-depth analysis of one of the cell types, we sought to further annotate subgroups of the most prevalent fetal cerebral cell type: excitatory neurons. To this end, we first applied our NNLS-based cell-type annotation strategy using single-cell expression data from the Allen Brain Atlas, which was collected on post mortem adult brain samples (*86*). Whereas many clusters found a match, the largest excitatory neuron subgroup did not (subgroup 1) (fig. S17D). By contrast, when using single-cell expression data collected from developing cortex (gestational week 17 to 18) (*85*, *87*), we found that the two largest excitatory neuron subclusters match to newly formed migrating and maturing excitatory neurons, respectively (subgroups 1 and 2) (fig. S17E). Of the top 10 peaks specific to the migrating population (subgroup 1), four lie in the introns of neuronal genes, four lie in the introns of noncoding RNAs, and two are distal to transcriptional units but highly conserved in vertebrates (fig. S18, A and B). One example of the latter is a distal peak (>20 kb from nearest TSS) overlapping a conserved element listed as negative in the VISTA enhancer browser (fig. S18A) (*80*).

This finer annotation also enabled us to ask whether heritability for certain traits is differentially enriched across neuronal subtypes. We calculated enrichments of trait heritability for each Louvain cluster in the cerebrum,

instead of each cell type, compared with the entire dataset. As expected, we observed enrichment for various neurological traits in the neuronal cell types but not in non-neuronal cell types, such as brain endothelial cells (fig. S19A). Within our broadly annotated cell types, we observed variable enrichment for different Louvain clusters; for example, inhibitory neuron subtype 2 is strongly enriched for heritability of both bipolar disorder and number of children born to males (fig. S19B). As for the excitatory neurons, we found that heritability for educational attainment is more strongly enriched in accessible sites of differentiated deep-layer excitatory neurons than migrating or maturing excitatory neurons (fig. S19C). Conversely, anorexia heritability is more strongly enriched in accessible sites of maturing excitatory neurons (fig. S19C).

An inspection of TF motifs differentially enriched across excitatory neuron subgroups revealed that POU2F1, which we showed may be restricted to neurons, is most strongly enriched in the fetal-specific migrating group, suggesting that it might not only be involved in maintenance but also specification of neuronal fates (Fig. 6A). In line with this, an enhancer adjacent to *POU2F1* has been shown to be specifically active in mouse cortical progenitor cells (*88*). To further investigate the regulatory landscape during excitatory neuron development, we next generated a pseudotime trajectory, from migrating over maturing to differentiated deep-layer neurons (Fig. 6, B and C). Differences in the median pseudotime of all excitatory neuron cells from individual donors corresponded loosely to differences in gestational age (Fig. 6C), although the number of individuals ($n = 3$) was too small for detailed investigation of this interindividual heterogeneity. Thousands of excitatory neuron peaks open or close in a pseudotime-dependent manner (Fig. 6D). Dynamically accessible elements that open over pseudotime were enriched for motifs of Rfx- and Tal-related factors important for neuronal maturation, whereas elements that close over pseudotime are enriched, among others, for motifs belonging to paired-related homeodomain factors and POU factors, including POU2F1 (2.2-fold change, $q = 2.3 \times 10^{-4}$) (Fig. 6E). This dynamic is supported by the matched RNA data, in which *POU2F1* expression peaks early in the pseudotime trajectory of excitatory neuron development (Fig. 6F) (*16*). Similar analyses of developmental-specific cell populations, their associated candidate regulatory regions, and TF motifs could be applied to further tissues once the progenitor population has been identified.

## Discussion

Sci-ATAC-seq3 adds to a growing repertoire of single-cell methods that use combinatorial indexing, a technical paradigm whose advantages over other platforms include exponential scaling and greater range with which to profile diverse aspects of single-cell biology (*1*–*12*). Although libraries have limited complexity and sci- protocols suffer from loss of material during the pooling and washing steps, the results presented here and in (*16*) illustrate the power of sci- methods. All experiments were conducted by a handful of individuals in a nonproduction environment but nonetheless resulted in very large single-cell chromatin accessibility and gene expression datasets.

An overarching goal of the field is to develop an "atlas" of human gene regulation as it unfolds across development and across the human life span. Aside from scale, our studies differ from other recent single-cell atlasing reports in at least three respects. First, we sought to profile as many tissues as possible within the context of a single study rather than focus on a single organ. This was both to create a broadly useful reference atlas as well as to enable cross-tissue comparisons of widely distributed cell types. For example, we observed tissue-specific chromatin accessibility and gene expression for endothelial cells but not erythroblasts.

Second, we focused on tissues obtained during human development. The rationale for this choice is discussed in greater length in (*16*) but includes our goal of laying a foundation for the systematic investigation of genetic disorders of development, which account for a disproportionate proportion of pediatric disease (*89*, *90*). The further accumulation of similar data from additional developmental time

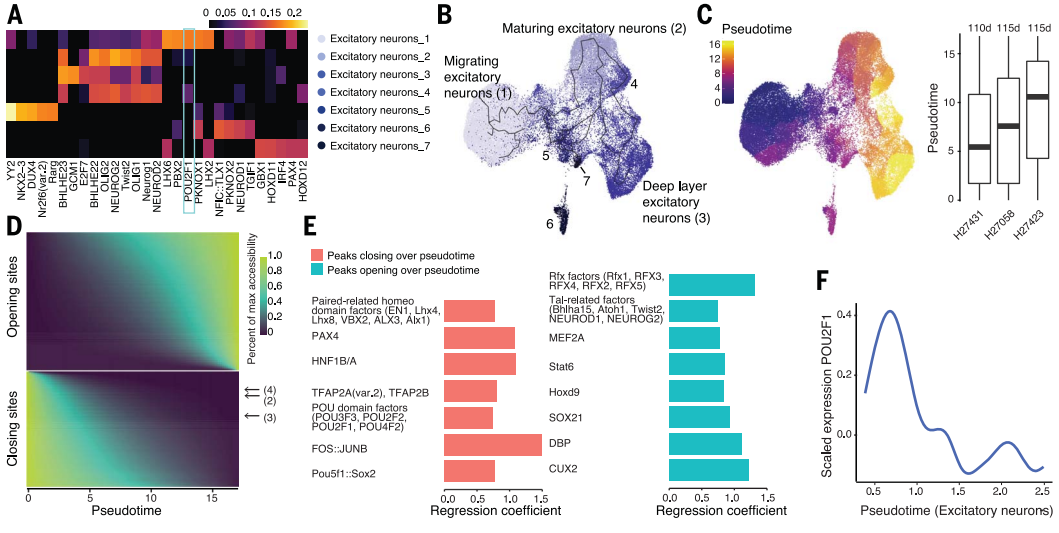



**Fig. 6. Chromatin accessibility dynamics in developing excitatory neurons.** (**A**) TF motifs enriched in excitatory neuron clusters. Fold-change of the top five enriched TF motifs in cluster-specific peaks for each of seven Louvain clusters that were annotated as excitatory neurons (log10-scaled fold-change of the mean motif occurrence in peaks of this cell subtype relative to the rest of the excitatory neurons, $q < 0.01$). POU2F1 enrichment is highlighted with a vertical box. (**B**) UMAP visualization and pseudotime trajectory path of 48,733 excitatory neurons colored by Louvain cluster. Color legend is in (A). (**C**) Pseudotime of excitatory neurons. (Left) UMAP visualization colored by pseudotime and (right) boxplots of median pseudotime per individual donor. Estimated gestational age is indicated above the boxplots. (**D**) Smoothed pseudotime-dependent accessibility curves of excitatory neurons, generated by a negative binomial regression and scaled as a percent of the maximum accessibility of each site. Sites (rows) are sorted by the pseudotime at which they first reach half their maximum accessibility. A random 10% of accessible sites was selected, and 3387 sites with pseudotime-dependent accessibility ($P < 0.05$, Wald test) are shown. Peaks from fig. S18A are indicated with arrows. (**E**) Motif enrichments in dynamically accessible sites from (D). Coefficients from logistic regression model by using the presence or absence of a given motif in each site to predict whether the site has a given accessibility trend. Plots show the top motifs ordered by Benjamini-Hochberg corrected $q$ value for each category ($q < 0.05$); similar motifs are grouped together. (**F**) Expression dynamics of *POU2F1* over pseudotime in excitatory neurons. Smoothed *POU2F1* expression in matching excitatory neurons from (*16*) was normalized by size factor in each single cell, then log-transformed and scaled.

points in both mouse and human will enable a systematic understanding of in vivo emergence and differentiation of mammalian cell types.

Third, we chose to study not only single-cell gene expression but also chromatin accessibility, in the same tissues and where possible from identical samples (*16*). Genomic regions exhibiting cell type–specific chromatin accessibility generally correspond to DNA regulatory elements such as enhancers and thus afford the opportunity to understand not only the "output" of the genome in particular cell types but also the regulatory program that underpins that output. The aggregate of all accessible regions that we identified spans 17% of the human genome, which is in line with recent bulk DNase-seq profiles from fetal tissues (*91*). Most of these ~1 million elements are cell type–specific or cell type–restricted in accessibility, although a large group of shared elements likely corresponds to looping anchors. Further studies (for example, those based on evolutionary conservation, massively parallel reporter assays, and/or CRISPR perturbation) are necessary to validate these candidate regulatory elements as well as their Cicero-based candidate linkages to target genes.

An interactive website facilitates the exploration of these data by tissue, cell type, locus, or motif (descartes.brotmanbaty.org) (*15*). Beyond constituting a rich and easily accessible resource for the field (for example, providing individual researchers with information on their gene, enhancer, or cell type of interest), this dataset also enables us to learn about more general aspects of gene regulation. For example, leveraging that we have matching chromatin accessibility and gene expression data spanning so many tissues and cell types allows us to study the mode of action of TFs as well as organ-specific differences in the regulatory landscape of cell types or cell type–specific disease heritability. Because the underlying methods are relatively new, there is currently a paucity of single-cell chromatin accessibility datasets in the public domain. We anticipate further comparisons to adult humans (*92*) or other species (*13*) as more such data become available.

The breadth and resolution of this dataset also provides insights into specific developmental processes. POU2F1 is one of the earliest described mammalian TFs (*93*). It is thought to be the only known POU family member not expressed in a specific temporal or spatial pattern, and in spite of being the subject of many studies, to date POU2F1's role has remained elusive (*42*, *94*). Although it has been suggested to be involved in housekeeping gene regulation or tumorigenesis, knockdowns in cancer cell lines showed no growth defect (*42*). The single-cell resolution provided by this study reveals that *POU2F1* is more highly expressed in neuronal cell types, and its motif is specifically enriched in neuronal regulatory regions. Because we captured developing neurons in our profiling window, we could observe that this motif is most highly enriched in the developing population of excitatory neurons, which is mirrored by *POU2F1* expression dynamics. *POU2F1* and its binding sites are highly conserved (*42*), and we also observed motif enrichment in mouse excitatory neurons, implying that this TF is a conserved inducer and maintainer of excitatory neuron cell fate. In line with this, POU2F1 deficiency is embryonic lethal (*95*). This example illustrates the power of combined chromatin accessibility and gene expression data at single-cell resolution. We anticipate that further such examples will emerge with more in-depth analyses of other tissue systems, stages, and cell types.

These and other downstream analyses used stratifications of accessibility that were based on our cell-type annotations. Although our assignments appear appropriate given that they generally recapitulate known biology in downstream analyses, they should be regarded as preliminary and will likely necessitate adjustments as more atlases and improved data become available. We intentionally kept our cell-type annotations rather broad, but there is more substructure in the data that could be explored further by subclustering—for example, as we show for blood cells and excitatory neurons. Although we are undoubtedly missing many cell types because of shallow profiling of several tissues or insufficiently aggressive clustering, we were nonetheless able to derive chromatin accessibility profiles and key regulators for some rare and potentially previously unknown cell types.

The analyses that we present here are only a starting point. Many other facets can be explored directly from these data—for example, nominating sets of TFs that must be coexpressed in the same cell type in order to bind regulatory regions cooperatively. In addition, these data can directly be used as input to machine learning models—for example, to predict the effect of all disease-associated variants identified in the human genome on chromatin accessibility across all cell types (*96*). We foresee that the true power of single-cell methods will lie in combining descriptive resources as we present here with both machine learning and high-throughput perturbation, with the long-term goal of establishing a deep, predictive understanding of gene regulation in human development and disease.

## Materials and methods

A more detailed version of materials and methods is provided as supplementary materials.

### *sci-ATAC-seq3*

A more detailed version of the full sci-ATAC-seq3 workflow is available on protocols.io (*97*) and in the supplementary materials.

### *Preparation of nuclei*

Human fetal tissues (89 to 125 days estimated post-conceptual age) were obtained by the University of Washington Birth Defects Research Laboratory (BDRL) under a protocol approved by the University of Washington Institutional Review Board. Tissues of interest were isolated and rinsed in 1X Hanks' balanced salt solution. Dried tissue was snap frozen in liquid nitrogen, manually pulverized on dry ice with a chilled hammer, aliquoted, and stored at −80°C until further processing. A subset of these aliquots were used for sci-ATAC-seq3, and others were used for sci-RNA-seq3, as described in the companion paper (*16*). For ATAC, nuclei were lysed with Omni-ATAC lysis buffer (*98*), cross-linked with 1% formaldehyde, and snap frozen in freezing buffer (*99*).

### *sci-ATAC-seq3 library construction and sequencing*

Frozen fixed nuclei were thawed, resuspended in Omni lysis buffer (*98*), and diluted in ATAC–resuspension buffer (RSB) buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MaCl2) supplemented with 0.1% Tween-20. For three-level indexing experiments at $384^3$, the nuclei input number was 4.8 million at 50,000 nuclei per well spread across 96 reactions. We profiled 24 individual tissue samples per batch (table S1), in which the 24th sample was a mixture of sentinel tissue (trisomy 18 cerebrum) and a mouse cell line (CH12-LX). For each sample, 200,000 nuclei were pelleted and resuspended in tagmentation reaction master mix (Nextera TD buffer, 1X Dulbecco's phosphate-buffered saline, 0.01% Digitonin, 0.1% Tween-20). Nuclei in tagmentation reaction master mix were aliquoted into four wells per tissue sample across a LoBind 96-well plate, 2.5 μl of Nextera v2 enzyme were added per well, and the plate was incubated at 55°C for 30 min. Tagmentation reactions were stopped by adding stop reaction mixture (40 mM EDTA with 1 mM Spermidine) and incubating at 37°C for 15 min. Tagmented nuclei from each sample were pooled (24 sample tubes in a batch), pelleted, washed, and resuspended in ATAC-RSB with 0.1% Tween-20. After adding phosphorylation master mix [1X polynucleotide kinase (PNK) buffer, 1 mM rATP, T4 PNK], the phosphorylation and nuclei reaction mix was aliquoted across a total of 16 wells in four LoBind 96-well plates and incubated at 37°C for 30 min. Ligation master mix (1X T7 ligase buffer, N5_splint, T7 DNA ligase enzyme) was added to the nuclei in the phosphorylation reaction followed by N5_oligos (384 distinct N5 barcodes). Sequences of all splint and barcode oligos used for sci-ATAC-seq3 are provided in table S7. Plates were incubated at 25°C for 1 hour. After this first round of ligation, stop reaction mixture was added, and the plates were incubated at 37°C for 15 min. Wells

were pooled, and nuclei were transferred into a 50-ml falcon tube, pelleted, and washed with ATAC-RSB with 0.1% Tween-20. The nuclei were then resuspended in N7 ligation master mix (1X T7 ligase buffer, N7_splint, T7 DNA ligase). This ligation and nuclei master mix was aliquoted into four 96-well LoBind plates, and N7_oligos (384 distinct N7 barcodes) were added to each well across four 96-well plates. Plates were incubated at 25°C for 1 hour before adding stop reaction mixture and incubating the plates at 37°C for 15 min. Wells were pooled and nuclei transferred into a 50-ml falcon tube, pelleted, and resuspended in Qiagen EB buffer. Then, 1000 to 3000 nuclei were aliquoted per well across four 96-well LoBind plates. To reverse cross-link the nuclei, we added a reverse cross-link master mix of EB buffer, PNK, and 1% SDS to each well. Plates were incubated at 65°C for 16 hours. A test PCR amplification was performed, and the reaction was monitored with SYBR green on several wells of a plate to determine the optimal cycle number (*1*). On the basis of this test PCR result, the rest of the reversed cross-linked plates were amplified with Nextera PCR Mastermix (NPM), bovine serum albumin, indexed P5 oligo, and indexed P7 oligo. Amplification products were pooled and purified first by using Zymo Clean & Concentrate-5 and then 1X AMPure beads. Final libraries were quantified on an Agilent 4200 Tapestation System. A 2 nM pool was created from equimolar pooling and sequenced with custom recipe and primers (sequences are provided in table S7) on an Illumina NovaSeq 6000 sequencer with custom sequencing recipe (read 1: 51 cycles, read 2: 51 cycles, index 1: 10 cycles+15 dark cycles+10 cycles, index 2: 10 cycles+15 dark cycles+10 cycles).

### Data processing for sci-ATAC-seq3

A more detailed version of all data processing and analysis steps is available in the supplementary materials. A demultiplexing script and tutorial are provided on Zenodo at (*100*).

Data processing for the barnyard experiments conducted to develop sci-ATAC-seq3 was done as previously described (*13*). Methods for processing sequencing data from the tissue samples closely follow the methods used in (*13*) as well, albeit with numerous optimizations to scale to larger datasets.

Cell barcodes were separated from the distribution of background barcodes by fitting a mixture of two negative binomials (noise versus signal). Nonduplicate fragment endpoints for each cell were used for peak calling in each sample by use of MACS2 (*101*). Peak calls from all samples included in downstream analysis were merged to form a master set of peaks. For each sample, we created sparse matrices counting (i) reads falling within the master set of peaks and (ii) reads falling within gene bodies extended by 2 kb upstream for each cell. We additionally tabulated the total number of reads from each cell coming from annotated TSSs (±1 kb around each TSS), ENCODE blacklist regions, and our set of merged peaks for quality control (QC) purposes. To filter out low-quality cells, we chose tissue-specific cut-offs for the fraction of nonduplicate reads in peaks (minimums ranging from 20 to 40%) and the fraction of nonduplicate reads falling in TSSs (minimums ranging from 5 to 15%) by means of visual inspection of their distributions for each sample (for example, for some tissues we observed a bimodal distribution for the fraction of nonduplicate reads in peaks and removed the lower mode), and a global cutoff of 0.5% of nonduplicate reads coming from ENCODE blacklist regions. All downstream steps were performed one tissue at a time by pooling cells passing QC from all samples of a given tissue. We used a modified version of the scrublet (*102*) algorithm to remove the cells most likely to be doublets.

### QC of sci-ATAC-seq3 data in bulk

After initial processing of the data, we assessed its quality relative to bulk DNase-seq profiles generated on fetal tissues procured from BDRL by the Roadmap Epigenomics consortium (*58*). After reprocessing the DNase-seq data in a comparable manner, we generated a master list of peaks across all DNase-seq and sci-ATAC-seq3 samples by merging all peaks called on each individual sample and generated a matrix of reads by peaks for each sample. This matrix of read counts was then used to calculate pairwise Spearman correlations to evaluate how similar samples were in their distributions of accessibility.

### Dimensionality reduction and clustering

For dimensionality reduction, we found that the implementation of latent semantic indexing [LSI; or equivalently, latent semantic analysis (LSA)] that we have previously applied (*13*) did not perform well on data collected in this study, likely owing to sparsity. Log-scaling the term-frequency term in LSI resulted in very similar performance to those of the other tools we tested (*103*, *104*). We suspect that this is due to the exponential distribution of total counts per cell and the impact of strong outliers on the principal components analysis (PCA) step of LSI in the absence of log scaling.

We performed LSI on the binarized peak-by-cell matrix for all cells passing QC from each tissue, one tissue at a time. We first weighted all the sites for individual cells by log(total number of peaks accessible in cell) (log-scaled "term frequency"). We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells), the "inverse document frequency." We used singular value decomposition on the term frequency–inverse document frequency matrix to generate a lower-dimensional representation of the data (PCA) by only retaining the 2nd through 50th dimensions (the first dimension tends to be highly correlated with read depth). L2 normalization was performed on the PCA matrix to further account for differences in the number of nonduplicate fragments per cell. This L2-normalized PCA matrix was used for all downstream steps.

Although we did not observe evidence for substantial batch effects between samples, we nonetheless applied the Harmony batch correction algorithm on the PCA space to correct batch effects between different samples (*20*, *21*). This corrected L2-normalized PCA space was used as input to Louvain clustering and UMAP as implemented in Seurat V3 (*105*).

### Cell type annotation

To transfer cell type labels for our Louvain clusters from the companion sci-RNA-seq data, we used an NNLS-based cluster-by-cluster annotation approach, which we have implemented previously to transfer labels between single-cell RNA-seq datasets (*26*). Briefly, we predicted the gene expression of target cell type in dataset A with the gene expression of all cell types in dataset B, and vice versa, and then multiplied the resulting βs to determine the matching of cell types between the two data sets with high specificity. To calculate gene level accessibility scores from ATAC data, we summed the accessibility over gene bodies extended by 2 kb upstream of their TSS. In addition to determining the NNLS score for each cell type or cluster, accessibility close to known cell type marker genes [described in (*16*)] was inspected for each cluster in each tissue (summed over all cells in that cluster). Clusters that had a high score in the NNLS and/or clear specific accessibility at matching marker genes were annotated accordingly. Clusters without strong NNLS signal and weaker or less specific marker gene accessibility received a less confident annotation. Clusters with no NNLS signal and no or only uninformative marker gene expression were left unannotated. In some cases, several Louvain clusters received the same cell type annotation within a tissue and were merged accordingly for downstream analyses.

The same strategy was applied to transfer labels from adult or fetal expression data of human cortex. Processed gene-by-cell matrices were downloaded from the Allen Brain Map website for the adult data and from (*87*) for the fetal data.

### Determining maternal contribution to cell types

To identify cell types with a lower fraction of Y chromosomal reads within a tissue, we selected all individual tissue samples with at least 800 cells and subsampled each cell type to 150 cells. For each tissue sample, we then calculated the ratio of Y chromosome over

autosome reads for these cells. As an additional line of evidence, we also used souporcell (*28*), a tool recently developed for clustering cells based on genotypes without a priori knowledge of individual genotypes.

### Specificity scores

Cell type–specificity scores for each site–cell type pair were calculated by using Jensen-Shannon divergence as previously described (*13*). A list of the top 10,000 most specific peaks per cell type is provided in data file S4. Similarly, we calculated tissue specificity scores for cross-tissue analysis of shared cell types (fig. S9C).

### Motif enrichments

We generated a binarized peak-by-motif matrix by identifying occurrences of motifs from the JASPAR vertebrate motif database (*106*) in each peak at a *P* value threshold of $10^{-7}$ using GC matched background nucleotide composition. A matrix of motif-by-cell counts was obtained by multiplying the peak-by-cell matrix with the peak-by-motif matrix. We downsampled the dataset so that a maximum of 800 cells per cell type, including unannotated clusters, were included to reduce computational cost and to reduce overrepresentation of very abundant cell types and tissues in computing enrichments in downstream steps. For each annotation, we then performed a negative binomial regression, predicting total motif counts using two input variables: an indicator column for the annotation as the main variable of interest and log(total number of nonzero entries in input peak matrix) for each cell as a covariate. We used the coefficient for the annotation indicator column and the intercept to estimate the fold change of the motif count of the annotation of interest relative to cells from all other annotations – exp(intercept + annotation_coefficient)/exp(intercept). This test was performed for all motifs in all groups, and *P* values were corrected by using the Benjamini Hochberg procedure.

For de novo motif finding, the top 2000 sites were selected by specificity score for each cell type (data file S4). We then ran homer (v4.11) using findMotifsGenome.pl with the specification -noknown -cpg (*48*). Matches to known motifs and scores were obtained from the standard homer output.

### Trajectory analysis

Cells were subjected to trajectory analysis with Monocle3, similar to as previously described (*26*, *55*, *63*). When determining sites that changed over pseudotime, 10% of accessible sites (~50,000) were sampled to reduce computational complexity. Motifs enriched in sites with significant changes over pseudotime were determined by use of a logistic regression model predicting accessibility trends (opening or closing) with the presence or absence of JASPAR motifs called in the peak-by-motif matrix. For comparison with excitatory neuron dynamics in sci-RNA-seq data, excitatory neurons were extracted and compared with single-cell RNA-seq data from (*107*) with a previously described NNLS-based matching technique to identify progenitor and mature neuron populations (*26*). Pseudotime was determined with Monocle3, and expression of the gene of interest was normalized by size factor in each single cell, then log-transformed, scaled, and plotted over pseudotime.

### Cicero models

Cicero coaccessibility analysis was performed for each of the 54 main cell types in each tissue (101 cell type–tissue combinations). Because coaccessibility scores are sensitive to false positives because of residual inter–cell type doublets driven by imperfect tissue dissociation, tissues were subjected to stringent doublet filtering on a per-tissue basis. To this end, the top 5000 specific peaks were selected per cell type (peak modules); each cell type was subclustered by using the union of these peak modules and a higher resolution, and subclusters with high accessibility in a peak module not matching the cell type in question were excluded. After applying this strategy, 91% of all cells were retained and subjected to coaccessibility score analysis. This stringently filtered dataset of 720,613 cells is provided at (*15*) and on the Gene Expression Omnibus (GEO) (GSE149683). The Cicero R package for Monocle3 (version 1.3.4.5) (*63*) was used to generate coaccessibility scores for each pair of sites within 500 kb in the linear genome and accessibility in at least 10 cells. Cicero gene activity scores were also generated by using default parameters.

### Heritability enrichments

Heritability enrichments were calculated similar to our previous work (*13*). Our input set of peaks for each annotated cell type (without any distinction between the same annotation occurring across different tissues) were the top 10,000 peaks as ranked by specificity score within this set of annotations. UKBB traits were downloaded from (*108*). Only traits with an estimated heritability of 0.01 or higher were carried forward for analysis. *P* values were calculated from *z* scores assigned to coefficients (assuming two-sided test), and coefficients were divided by the average per-SNP heritability for the trait associated with a given test, producing scaled coefficients. Tests were corrected for multiple hypothesis testing by using the Benjamini-Hochberg method, and only tests with a *q* value of 0.2 or lower were deemed to be significant.

### Testing for allelic imbalance

In order to identify genetic variation influencing chromatin accessibility levels, we used the "Quantitative Allele-Specific Analysis of Reads" (QuASAR) package in R (*68*). All reads from the selected samples were first remapped with HISAT2 (*109*), a SNP-aware aligner, to remove any positions showing an allelic mapping bias. Properly mapping reads were then sorted into individual bam files per cell type per individual. For each cell type, we generated a pileup file using Samtools (*110*) and considering all common SNPs from dbSNP144 (*111*). After filtering, we ran the analysis on all cell types from a sample at the same time using default parameters. The resulting *P* values were then FDR-corrected by using the Benjamini and Hochberg method (*112*) for each sample independently. Any *q* values less than 0.2 were considered significant. The method we used does not require prior genotyping, but because we had genotyped the individuals, we were able to confirm that sites inferred to be heterozygous in our analysis were also identified as heterozygous on the genotyping array: 2253 of 2254 (99.96%) overlapping SNPs for one individual and 6059 of 6065 (99.90%) for the other individual.

### UMAP of sites

To reduce computational complexity, we used the peak-by-cell matrix based on 88,983 subsampled cells (subsampling up to 800 cells per cell type in each tissue, including unannotated clusters) and filtered out peaks smaller than 400 bp because these were less conserved and found in fewer cells. We transposed the resulting peak-by-cell matrix (447,879 by 88,983) and proceeded with Louvain clustering and UMAP visualization as above. To determine overlap of sites with regions bound by CTCF in GM12878, we downloaded the ChIP-seq peak locations from ENCODE (*56*, *57*). To determine overlap with looping anchors from Hi-C data in GM12878, we compared with loop annotations returned by the loop-calling algorithm HiCCUPS (*78*).

### Comparison with mouse ATAC atlas

To compare peak coordinates, mouse peaks from (*13*) were lifted over from mm9 to hg19 by using the University of California, Santa Cruz liftOver tool (*113*). To compare cell type–specific motif enrichments, a peak-by-motif matrix was generated for all mouse peaks and cell types from (*13*) by using the same motif database and *P* value cutoff, and motif enrichment analysis was conducted as described above.

### Comparison with adult single-cell ATAC-seq data

Comparison with adult single-cell ATAC-seq data was performed on peak-by-cell matrices. For blood (63,883 nuclei from adult bone marrow and peripheral blood profiled on the 10X platform), this matrix was directly downloaded (*84*); for cortex (12,557 nuclei from post-mortem human brain profiled with sn-ATAC-seq), it was generated from the snap file and peak calls

provided by the authors (*85*, *104*). Fetal and adult peaks from both master lists were intersected, and 100,000 overlapping peaks were randomly sampled from the adult peak set to reduce computational complexity. Then the fetal data, downsampled to 1500 cells per Louvain cluster, was rescored on the basis of these peaks. Subsequently, the datasets were integrated by using integration anchors and the Signac package (*23*, *114*). For motif analysis, a peak-by-motif matrix was generated for all peaks in the adult blood dataset by using the same motif database and *P* value cutoff as described above, and motif enrichment calculation was carried out accordingly.

## REFERENCES AND NOTES


1. D. A. Cusanovich *et al*., Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015). doi: 10.1126/science.aab1601; pmid: 25953818
2. J. Cao *et al*., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017). doi: 10.1126/science.aam8940; pmid: 28818938
3. V. Ramani *et al*., Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017). doi: 10.1038/nmeth.4155; pmid: 28135255
4. R. M. Mulqueen *et al*., Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018). doi: 10.1038/nbt.4112; pmid: 29644997
5. Y. Yin *et al*., High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **76**, 676–690.e10 (2019). doi: 10.1016/j.molcel.2019.08.002; pmid: 31495564
6. J. Cao *et al*., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018). doi: 10.1126/science.aau0730; pmid: 30166440
7. S. A. Vitak *et al*., Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017). doi: 10.1038/nmeth.4154; pmid: 28135258
8. A. B. Rosenberg *et al*., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018). doi: 10.1126/science.aam8999; pmid: 29545511
9. C. A. Lareau *et al*., Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019). doi: 10.1038/s41587-019-0147-6; pmid: 31235917
10. Q. Wang *et al*., CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol. Cell* **76**, 206–216.e7 (2019). doi: 10.1016/j.molcel.2019.07.015; pmid: 31471188
11. J. Cao, W. Zhou, F. Steemers, C. Trapnell, J. Shendure, Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020). doi: 10.1038/s41587-020-0480-9; pmid: 32284584
12. C. Zhu *et al*., An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019). doi: 10.1038/s41594-019-0323-x; pmid: 31695190
13. D. A. Cusanovich *et al*., A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018). doi: 10.1016/j.cell.2018.06.052; pmid: 30078704
14. S. Behjati, S. Lindsay, S. A. Teichmann, M. Haniffa, Mapping human development at single-cell resolution. *Development* **145**, dev152561 (2018). doi: 10.1242/dev.152561; pmid: 29439135
15. descartes.brotmanbaty.org
16. J. Cao *et al*., A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020). doi: 10.1126/science.aba7721
17. F. Yue *et al*., A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014). doi: 10.1038/nature13992; pmid: 25409824
18. J. D. Bloom, Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ* **6**, e5578 (2018). doi: 10.7717/peerj.5578; pmid: 30202659
19. H. Chen *et al*., Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019). doi: 10.1186/s13059-019-1854-5; pmid: 31739806
20. A. J. Hill, Andrew John Hill Blog; http://andrewjohnhill.com/blog/2019/05/06/dimensionality-reduction-for-scatac-data.
21. I. Korsunsky *et al*., Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019). doi: 10.1038/s41592-019-0619-0; pmid: 31740819
22. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018). doi: 10.21105/joss.00861
23. T. Stuart *et al*., Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019). doi: 10.1016/j.cell.2019.05.031; pmid: 31178118
24. B. B. Lake *et al*., Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018). doi: 10.1038/nbt.4038; pmid: 29227469
25. L. T. Graybuck *et al*., Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven AAVs. *bioRxiv* (2019). doi: 10.1101/525014
26. J. Cao *et al*., The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019). doi: 10.1038/s41586-019-0969-x; pmid: 30787437
27. H. Suryawanshi *et al*., A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018). doi: 10.1126/sciadv.aau4788; pmid: 30402542
28. H. Heaton *et al*., Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020). doi: 10.1038/s41592-020-0820-1; pmid: 32366989
29. H. Iwasaki, K. Akashi, Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**, 726–740 (2007). doi: 10.1016/j.immuni.2007.06.004; pmid: 17582345
30. A. Arthur *et al*., Twist-1 Enhances Bone Marrow Mesenchymal Stromal Cell Support of Hematopoiesis by Modulating CXCL12 Expression. *Stem Cells* **34**, 504–509 (2016). doi: 10.1002/stem.2265; pmid: 26718114
31. S. J. Renaud, K. Kubota, M. A. K. Rumi, M. J. Soares, The FOS transcription factor family differentially controls trophoblast migration and invasion. *J. Biol. Chem.* **289**, 5025–5039 (2014). doi: 10.1074/jbc.M113.523746; pmid: 24379408
32. A.-M. Bamberger *et al*., Expression pattern of the activating protein-1 family of transcription factors in the human placenta. *Mol. Hum. Reprod.* **10**, 223–228 (2004). doi: 10.1093/molehr/gah011; pmid: 14985474
33. J. Palis, Primitive and definitive erythropoiesis in mammals. *Front. Physiol.* **5**, 3 (2014). doi: 10.3389/fphys.2014.00003; pmid: 24478716
34. A. K. Hennig, G.-H. Peng, S. Chen, Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Res.* **1192**, 114–133 (2008). doi: 10.1016/j.brainres.2007.06.036; pmid: 17662965
35. B. L. Black, E. N. Olson, Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.* **14**, 167–196 (1998). doi: 10.1146/annurev.cellbio.14.1.167; pmid: 9891782
36. Z. Niu *et al*., Serum response factor orchestrates nascent sarcomerogenesis and silences the biomineralization gene program in the heart. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17824–17829 (2008). doi: 10.1073/pnas.0805491105; pmid: 19004760
37. R. Kageyama, H. Shimojo, T. Ohtsuka, Dynamic control of neural stem cells by bHLH factors. *Neurosci. Res.* **138**, 12–18 (2019). doi: 10.1016/j.neures.2018.09.005; pmid: 30227160
38. G. Wilkinson, D. Dennis, C. Schuurmans, Proneural genes in neocortical development. *Neuroscience* **253**, 256–273 (2013). doi: 10.1016/j.neuroscience.2013.08.029; pmid: 23999125
39. M. Zou, S. Li, W. H. Klein, M. Xiang, Brn3a/Pou4f1 regulates dorsal root ganglion sensory neuron specification and axonal projection into the spinal cord. *Dev. Biol.* **364**, 114–127 (2012). doi: 10.1016/j.ydbio.2012.01.021; pmid: 22326227
40. M. G. De Vas *et al*., Hnf1b controls pancreas morphogenesis and the generation of Ngn3$^{+}$ endocrine progenitors. *Development* **142**, 871–882 (2015). doi: 10.1242/dev.110759; pmid: 25715395
41. A. Desgrange *et al*., HNF1B controls epithelial organization and cell polarity during ureteric bud branching and collecting duct morphogenesis. *Development* **144**, 4704–4719 (2017). doi: 10.1242/dev.154336; pmid: 29158444
42. K. Vázquez-Arreguín, D. Tantin, The Oct1 transcription factor and epithelial malignancies: Old protein learns new tricks. *Biochim. Biophys. Acta* **1859**, 792–804 (2016). doi: 10.1016/j.bbagrm.2016.02.007; pmid: 26877236
43. D. McClellan *et al*., Growth factor independence 1B-mediated transcriptional repression and lineage allocation require lysine-specific demethylase 1-dependent recruitment of the BHC complex. *Mol. Cell. Biol.* **39**, e00020–e19 (2019). doi: 10.1128/MCB.00020-19; pmid: 30988160
44. M. Oukka *et al*., The transcription factor NFAT4 is involved in the generation and survival of T cells. *Immunity* **9**, 295–304 (1998). doi: 10.1016/S1074-7613(00)80612-3; pmid: 9768749
45. Z. Li *et al*., Serine 574 phosphorylation alters transcriptional programming of FOXO3 by selectively enhancing apoptotic gene expression. *Cell Death Differ.* **23**, 583–595 (2016). doi: 10.1038/cdd.2015.125; pmid: 26470730
46. A. Jankowski, J. Tiuryn, S. Prabhakar, Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* **32**, 2419–2426 (2016). doi: 10.1093/bioinformatics/btw209; pmid: 27153645
47. Z. F. Chen, A. J. Paquette, D. J. Anderson, NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat. Genet.* **20**, 136–142 (1998). doi: 10.1038/2431; pmid: 9771705
48. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: 10.1016/j.molcel.2010.05.004; pmid: 20513432
49. Y. Lavin *et al*., Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326 (2014). doi: 10.1016/j.cell.2014.11.018; pmid: 25480296
50. D. B. Thomas, J. M. Yoffey, Human foetal haematopoiesis. II. Hepatic haematopoiesis in the human foetus. *Br. J. Haematol.* **10**, 193–197 (1964). doi: 10.1111/j.1365-2141.1964.tb00694.x; pmid: 14141618
51. L. O. Jacobson, E. L. Simmons, E. K. Marks, J. H. Eldredge, Recovery from radiation injury. *Science* **113**, 510–511 (1951). doi: 10.1126/science.113.2940.510; pmid: 14828383
52. B. K. Tusi *et al*., Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018). doi: 10.1038/nature25741; pmid: 29466336
53. V. P. Schulz *et al*., A unique epigenomic landscape defines human erythropoiesis. *Cell Rep.* **28**, 2996–3009.e7 (2019). doi: 10.1016/j.celrep.2019.08.020; pmid: 31509757
54. T. J. Satchwell, Erythrocyte invasion receptors for *Plasmodium falciparum*: New and old. *Transfus. Med.* **26**, 77–88 (2016). doi: 10.1111/tme.12280; pmid: 26862042
55. C. Trapnell *et al*., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014). doi: 10.1038/nbt.2859; pmid: 24658644
56. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
57. C. A. Davis *et al*., The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46** (D1), D794–D801 (2018). doi: 10.1093/nar/gkx1081; pmid: 29126249
58. Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
59. S. De Val, B. L. Black, Transcriptional control of endothelial cell development. *Dev. Cell* **16**, 180–195 (2009). doi: 10.1016/j.devcel.2009.01.014; pmid: 19217421
60. A. V. Shah, G. M. Birdsey, A. M. Randi, Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor ERG. *Vascul. Pharmacol.* **86**, 3–13 (2016). doi: 10.1016/j.vph.2016.05.003; pmid: 27208692
61. D. Siripin *et al*., Transdifferentiation of erythroblasts to megakaryocytes using FLI1 and ERG transcription factors. *Thromb. Haemost.* **114**, 593–602 (2015). doi: 10.1160/TH14-12-1090; pmid: 26063314
62. O. Goldman *et al*., Endoderm generates endothelial cells during liver development. *Stem Cell Reports* **3**, 556–565 (2014). doi: 10.1016/j.stemcr.2014.08.009; pmid: 25358784
63. H. A. Pliner *et al*., Cicero Predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018). doi: 10.1016/j.molcel.2018.06.044; pmid: 30078726
64. D. Noordermeer, W. de Laat, Joining the loops: Beta-globin gene regulation. *IUBMB Life* **60**, 824–833 (2008). doi: 10.1002/iub.129; pmid: 18767169
65. J. K. Pickrell, Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014). doi: 10.1016/j.ajhg.2014.03.004; pmid: 24702953

66. M. T. Maurano *et al*., Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012). doi: 10.1126/science.1222794; pmid: 22955828
67. H. K. Finucane *et al*., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015). doi: 10.1038/ng.3404; pmid: 26414678
68. C. T. Harvey *et al*., QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics* **31**, 1235–1242 (2015). doi: 10.1093/bioinformatics/btu802; pmid: 25480375
69. A. Buniello *et al*., The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** (D1), D1005–D1012 (2019). doi: 10.1093/nar/gky1120; pmid: 30445434
70. G. Kichaev *et al*., Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019). doi: 10.1016/j.ajhg.2018.11.008; pmid: 30595370
71. N. Tanimura *et al*., GATA/heme multi-omics reveals a trace metal-dependent cellular differentiation mechanism. *Dev. Cell* **46**, 581–594.e4 (2018). doi: 10.1016/j.devcel.2018.07.022; pmid: 30122630
72. B. Zhao *et al*., Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* **51**, 1637–1644 (2019). doi: 10.1038/s41588-019-0516-6; pmid: 31676860
73. F. Tissir, A. M. Goffinet, Reelin and brain development. *Nat. Rev. Neurosci.* **4**, 496–505 (2003). doi: 10.1038/nrn1113; pmid: 12778121
74. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017). doi: 10.1038/nmeth.4401; pmid: 28825706
75. M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, C. Trapnell, Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.* **11**, 1537 (2020). doi: 10.1038/s41467-020-15351-4; pmid: 32210240
76. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013). doi: 10.1016/j.tig.2013.05.010; pmid: 23810203
77. J. R. Dixon *et al*., Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). doi: 10.1038/nature11082; pmid: 22495300
78. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). doi: 10.1016/j.cell.2014.11.021; pmid: 25497547
79. T. H. Kim *et al*., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007). doi: 10.1016/j.cell.2006.12.048; pmid: 17382889
80. R. M. Gittelman *et al*., Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* **25**, 1245–1255 (2015). doi: 10.1101/gr.192591.115; pmid: 26104583
81. A. Visel, S. Minovitsky, I. Dubchak, L. A. Pennacchio, VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35** (Database), D88–D92 (2007). doi: 10.1093/nar/gkl822; pmid: 17130149
82. D. C. King *et al*., Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res.* **17**, 775–786 (2007). doi: 10.1101/gr.5592107; pmid: 17567996
83. E. Montecino-Rodriguez, K. Dorshkind, Formation of B-1 B cells from neonatal B-1 transitional cells exhibits NF-κB redundancy. *J. Immunol.* **187**, 5712–5719 (2011). doi: 10.4049/jimmunol.1102416; pmid: 22031760
84. A. T. Satpathy *et al*., Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019). doi: 10.1038/s41587-019-0206-z; pmid: 31375813
85. C. Luo *et al*., Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* 873398 [Preprint] 12 December 2019; .doi: 10.1101/2019.12.11.873398
86. Cell Types Database, RNA-Seq Data—brain-map.org; https://portal.brain-map.org/atlases-and-data/rnaseq#Human_Cortex.
87. D. Polioudakis *et al*., A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801.e8 (2019). doi: 10.1016/j.neuron.2019.06.011; pmid: 31303374
88. A. Bery, B. Martynoga, F. Guillemot, J.-S. Joly, S. Rétaux, Characterization of enhancers active in the mouse embryonic cerebral cortex suggests Sox/Pou cis-regulatory logics and heterogeneity of cortical progenitors. *Cereb. Cortex* **24**, 2822–2834 (2014). doi: 10.1093/cercor/bht126; pmid: 23720416
89. S. E. McCandless, J. W. Brunger, S. B. Cassidy, The burden of genetic disease on inpatient care in a children's hospital. *Am. J. Hum. Genet.* **74**, 121–127 (2004). doi: 10.1086/381053; pmid: 14681831
90. M. H. Wojcik *et al*., Genetic disorders and mortality in infancy and early childhood: Delayed diagnoses and missed opportunities. *Genet. Med.* **20**, 1396–1404 (2018). doi: 10.1038/gim.2018.17; pmid: 29790870
91. W. Meuleman *et al*., Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020). doi: 10.1038/s41586-020-2559-3; pmid: 32728217
92. M. R. Corces *et al*., Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016). doi: 10.1038/ng.3646; pmid: 27526324
93. G. J. Pruijn, W. van Driel, P. C. van der Vliet, Nuclear factor III, a novel sequence-specific DNA-binding protein from HeLa cells stimulating adenovirus DNA replication. *Nature* **322**, 656–659 (1986). doi: 10.1038/322656a0; pmid: 3748145
94. D. Tantin, C. Schild-Poulter, V. Wang, R. J. G. Haché, P. A. Sharp, The octamer binding transcription factor Oct-1 is a stress sensor. *Cancer Res.* **65**, 10750–10758 (2005). doi: 10.1158/0008-5472.CAN-05-2399; pmid: 16322220
95. V. E. H. Wang, T. Schmidt, J. Chen, P. A. Sharp, D. Tantin, Embryonic lethality, decreased erythropoiesis, and defective octamer-dependent promoter activation in Oct-1-deficient mice. *Mol. Cell. Biol.* **24**, 1022–1032 (2004). doi: 10.1128/MCB.24.3.1022-1032.2004; pmid: 14729950
96. D. Lee *et al*., A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015). doi: 10.1038/ng.3331; pmid: 26075791
97. S. Domcke, A. J. Hill, R. M. Daza, C. Trapnell, D. A. Cusanovich, J. Shendure, sci-ATAC-seq3. protocols.io (2020). doi: 10.17504/protocols.io.be8mjhu6
98. M. R. Corces *et al*., An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017). doi: 10.1038/nmeth.4396; pmid: 28846090
99. A. Saunders, L. J. Core, C. Sutcliffe, J. T. Lis, H. L. Ashe, Extensive polymerase pausing during *Drosophila* axis patterning enables high-level and pliable transcription. *Genes Dev.* **27**, 1146–1158 (2013). doi: 10.1101/gad.215459.113; pmid: 23699410
100. S. Domcke, A. Hill, shendurelab/human-atac: First release of sci-ATAC-seq3 demultiplexing. Zenodo (2020). doi: 10.5281/zenodo.4012524
101. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
102. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019). doi: 10.1016/j.cels.2018.11.005; pmid: 30954476
103. C. Bravo González-Blas *et al*., cisTopic: Cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019). doi: 10.1038/s41592-019-0367-1; pmid: 30962623
104. R. Fang *et al*., Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. *bioRxiv* 615179 [Preprint] 13 May 2019).
105. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018). doi: 10.1038/nbt.4096; pmid: 29608179
106. O. Fornes *et al*., JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020). doi: 10.1093/nar/gkz1001; pmid: 31701148
107. A. Zeisel *et al*., Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018). doi: 10.1016/j.cell.2018.06.021; pmid: 30096314
108. https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K
109. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019). doi: 10.1038/s41587-019-0201-4; pmid: 31375807
110. H. Li *et al*., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: 10.1093/bioinformatics/btp352; pmid: 19505943
111. S. T. Sherry *et al*., dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001). doi: 10.1093/nar/29.1.308; pmid: 11125122
112. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995). doi: 10.1111/j.2517-6161.1995.tb02031.x
113. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
114. T. Stuart, timoast/signac. GitHub (2020); https://github.com/timoast/signac.
115. Broad Institute, *Picard Tools*; http://broadinstitute.github.io/picard.
116. J. A. Castro-Mondragon, S. Jaeger, D. Thieffry, M. Thomas-Chollier, J. van Helden, RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* **45**, e119 (2017). doi: 10.1093/nar/gkx314; pmid: 28591841
117. L. Navarro-Núñez, S. A. Langan, G. B. Nash, S. P. Watson, The physiological and pathophysiological roles of platelet CLEC-2. *Thromb. Haemost.* **109**, 991–998 (2013). doi: 10.1160/TH13-01-0060; pmid: 23572154


## ACKNOWLEDGMENTS

We thank past and present members of the Shendure, Trapnell, and Cusanovich laboratories; A. Adey; the Brotman Baty Institute for Precision Medicine (BBI) Advanced Technology Laboratory; and the University of Washington Birth Defects Research Laboratory. We thank C. Spurrell, D. Ahrendsen, and R. Blecher for reviewing the sci-ATAC-seq3 protocol. We thank C. Cenik and Z. Ma from M. Snyder's laboratory for providing the CH12.LX cell line. We thank R. Pique-Regi for advice on implementing QuASAR. We acknowledge the ENCODE Consortium for access to their data. **Funding:** S.D. was supported by an EMBO Long-Term Fellowship and a HFSP Long-Term Fellowship. D.A.C. was supported in part by an NHLBI fellowship (T32HL007828). Aspects of this work were supported by funding from the BBI, the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), The Chan Zuckerberg Initiative (to C.T.) and the NIH (HD000836 to I.A.G.). J.S. is an investigator of the Howard Hughes Medical Institute. **Authors contributions:** R.M.D., S.D., and D.A.C. developed techniques and performed sci-ATAC-seq3 experiments with assistance from F.Z., D.P., F.J.S., and J.H.M.; D.R.O. performed tissue collection and nuclei extraction under supervision of D.D. and I.A.G.; S.D. and A.J.H. analyzed the data with assistance from D.A.C. and H.A.P.; J.C. provided sci-RNA-seq data; K.A.A. performed genotyping; M.A.Z. developed the website, with assistance from S.D.; S.D., D.A.C., and J.S. wrote the manuscript with input from all co-authors; J.S., D.A.C., and C.T. supervised the project. **Competing interests:** D.P., F.Z., and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina. J.S. has competing financial interests (paid consulting and/or equity) with Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnologies, and Stratos Genomics. One or more embodiments of one or more patents and patent applications filed by Illumina and the University of Washington may encompass the methods, reagents, and data disclosed in this manuscript. **Data and materials availability:** In contrast with previous versions of sci-ATAC-seq, custom Tn5 is not required for sci-ATAC-seq3. A detailed version of the sci-ATAC-seq3 protocol is available on protocols.io (*97*). A demultiplexing script and tutorial are provided on Zenodo at (100). Supplementary data files S1 to S7 are provided at GEO (accession no. GSE149683). Raw data are provided at dbGaP (accession no. phs002003.v1.p1). Processed data are available at GEO (accession no. GSE149683 (both unfiltered and peak module-based filtered dataset)). An interactive version of the peak module–based filtered dataset facilitates the exploration of these data by tissue, cell type, locus, or motif at descartes.brotmanbaty.org (*15*).

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/eaba7612/suppl/DC1
Materials and Methods
Figs. S1 to S19
References (*118*–*124*)
Tables S1 to S7

View/request a protocol for this paper from *Bio-protocol*.

3 January 2020; accepted 10 September 2020
10.1126/science.aba7612

# A human cell atlas of fetal chromatin accessibility

Silvia Domcke, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. ODay, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell, Darren A. Cusanovich, and Jay Shendure

*Science*, **370** (6518), eaba7612.
DOI: 10.1126/science.aba7612

**The genomics of human development**

Understanding the trajectory of a developing human requires an understanding of how genes are regulated and expressed. Two papers now present a pooled approach using three levels of combinatorial indexing to examine the single-cell gene expression and chromatin landscapes from 15 organs in fetal samples. Cao *et al.* focus on measurements of RNA in broadly distributed cell types and provide insights into organ specificity. Domcke *et al.* examined the chromatin accessibility of cells from these organs and identify the regulatory elements that regulate gene expression. Together, these analyses generate comprehensive atlases of early human development.

*Science*, this issue p. eaba7721, p. eaba7612