An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis

Guanjue Xiang[1]*, Cheryl A. Keller[1]*, Elisabeth Heuston[2], Belinda M. Giardine[1], Lin An[1], Alexander Q. Wixom[1], Amber Miller[1], April Cockburn[1], Jens Lichtenberg[2], Berthold Göttgens[3], Qunhua Li[4], David Bodine[2], Shaun Mahony[1], James Taylor[5], Gerd A. Blobel[6], Mitchell J. Weiss[7], Yong Cheng[7], Feng Yue[8], Jim Hughes[9], Douglas R. Higgs[9], Yu Zhang[4], Ross C. Hardison[1]**

[1]Department of Biochemistry and Molecular Biology, [4]Department of Statistics, Program in Bioinformatics and Genomics, Center for Computational Biology and Bioinformatics, The Pennsylvania State University, University Park, PA; [2]NHGRI Hematopoiesis Section, Genetics and Molecular Biology Branch, National Institutes of Health, Bethesda, MD; [3]Welcome and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK; [5]Departments of Biology and Computer Science, Johns Hopkins University, Baltimore, MD; [6]Department of Pediatrics, Children's Hospital of Philadelphia and University of Pennsylvania School of Medicine, Philadelphia, PA; [7]Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN; [8]Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA; [9]MRC Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK

* = equal contributions

**Corresponding author: Ross Hardison, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 304 Wartik Lab, University Park, PA 16802, Phone: 814-863-0113; E-mail: rch8@psu.edu

Running Title: Concise representation of regulatory landscapes

Key words: epigenetics, gene regulatory elements, dimensional reduction, regulatory potential

Abstract

Thousands of epigenomic datasets have been generated in the past decade, but it is difficult for researchers to effectively utilize all the data relevant to their projects. Systematic integrative analysis can help meet this need, and the VISION project was established for **V**al**I**dated **S**ystematic **I**ntegrati**ON** of epigenomic data in hematopoiesis. Here, we systematically integrated extensive data recording epigenetic features and transcriptomes from many sources, including individual laboratories and consortia, to produce a comprehensive view of the regulatory landscape of differentiating hematopoietic cell types in mouse. By employing IDEAS as our **I**ntegrative and **D**iscriminative **E**pigenome **A**nnotation **S**ystem, we identified and assigned epigenetic states simultaneously along chromosomes and across cell types, precisely and comprehensively. Combining nuclease accessibility and epigenetic states produced a set of over 200,000 candidate *cis*-regulatory elements (cCREs) that efficiently capture enhancers and promoters. The transitions in epigenetic states of these cCREs across cell types provided insights into mechanisms of regulation, including decreases in numbers of active cCREs during differentiation of most lineages, transitions from poised to active or inactive states, and shifts in nuclease accessibility of CTCF-bound elements. Regression modeling of epigenetic states at cCREs and gene expression produced a versatile resource to improve selection of cCREs potentially regulating target genes. These resources are available from our VISION website (usevision.org) to aid research in genomics and hematopoiesis.

Introduction

Recent work from both individual laboratories and major consortia (e.g., The_ENCODE_Project_Consortium 2012; Cheng *et al.* 2014; Yue *et al.* 2014; Roadmap Epigenomics *et al.* 2015; Stunnenberg *et al.* 2016; The_ENCODE_Project_Consortium *et al.* 2019) have produced, from diverse cell types, thousands of genome-wide datasets on transcriptomes and many epigenetic features, including nuclease accessibility, histone modifications, transcription factor occupancy. However, it is challenging for individual investigators to find all the data relevant to their projects or, once found, to incorporate the data effectively into analyses and hypothesis generation. One approach to address this challenge of overwhelming data is to integrate these deep and diverse datasets (Ernst and Kellis 2010; Ernst and Kellis 2012; Hoffman *et al.* 2012; Hoffman *et al.* 2013; Zhou and Troyanskaya 2015; Greenside *et al.* 2018; Lee *et al.* 2018; Ludwig *et al.* 2019). An effective integration will produce simplified representations of the data that facilitate discoveries and lead to testable hypotheses about functions of genomic elements and mechanisms of regulatory processes. Our multi-lab project called VISION (for **V**al**I**dated **S**ystematic **I**ntegrati**ON** of hematopoietic epigenomes) is endeavoring to meet this challenge by focusing on an important biological system, hematopoietic differentiation. Not only is hematopoietic differentiation an important biological and medical system with abundant epigenetic data available (e.g., Cheng *et al.* 2009; Fujiwara *et al.* 2009; Yu *et al.* 2009; Wilson *et al.* 2010; Pilon *et al.* 2011; Tijssen *et al.* 2011; Wong *et al.* 2011; Wu *et al.* 2011; Kowalczyk *et al.* 2012; Su *et al.* 2013; Lara-Astiaso *et al.* 2014; Pimkin *et al.* 2014; Wu *et al.* 2014; Corces *et al.* 2016; Huang *et al.* 2016; Heuston *et al.* 2018; Ludwig *et al.* 2019), but it also provides a powerful framework for validation of the integrative modeling. Specifically, work over prior decades has established key concepts that a successful modeling effort should recapitulate, and predictions of the modeling can be tested genetically in animals and cell lines. Here, we report on our initial systematic integrative modeling of mouse hematopoiesis.

The production of many distinct blood cell types from a common stem cell (hematopoiesis) is critically important for human health (Orkin and Zon 2008). This process has been studied intensively in humans and mouse. Despite some differences between these species (An *et al.* 2014; Cheng *et al.* 2014; Pishesha *et al.* 2014), the mouse system has served as a good model for many aspects of hematopoiesis in humans and mammals in general (Sykes and Scadden 2013). In adult mammals, all blood cells are produced from mesodermally-derived, self-renewing hematopoietic stem cells (HSCs) located in the bone marrow (Till and McCulloch 1961; Kondo *et al.* 2003). Several distinct populations of multilineage progenitor cells with different capacities for differentiation to specific lineages have been purified using cell surface markers (Weissman and Shizuru 2008). In a common model derived from these cell populations, hematopoietic differentiation proceeds from HSC through progenitor cells with progressively more restricted lineage potential, eventually committing to a single cell lineage (Reya *et al.* 2001). More recent analyses of single cell transcriptomes have revealed heterogeneity in each of these populations, and in some cases, they uncovered a bias in multilineage progenitors toward a single cell type (Sanjuan-Pla *et al.* 2013; Psaila *et al.* 2016). Overall, analysis of single cell transcriptomes support an ensemble of pathways for differentiation (Nestorowa *et al.* 2016; Laurenti and Gottgens 2018). After lineage commitment, cells progress through multiple stages to form mature, circulating blood cells. Regardless of the complexity in cell-fate pathways, it is clear that changes in patterns of gene expression drive the differentiation program (Cantor and Orkin 2002; Graf and Enver 2009). Mis-regulation of those gene expression patterns can cause diseases such as leukemias and anemias (Higgs 2013; Lee and Young 2013), and thus efforts to better understand the molecular mechanisms regulating gene expression can help uncover the processes underlying cancers and blood disorders.

Comprehensive epigenomic and transcriptomic data can be used to describe how both the patterns of gene expression and the regulatory landscapes change during hematopoietic differentiation. Previous publications provided many insights and datasets on epigenomic changes during hematopoiesis in mouse (e.g., Lara-Astiaso *et al.* 2014) and in human (e.g., Adams *et al.* 2012; Corces *et al.* 2016). Additional informative datasets have come from detailed studies in cell line models of hematopoietic differentiation. In the intensively studied process of hematopoiesis, such comprehensive datasets could encompass virtually all the recognized regulatory and transcriptional changes that occur during differentiation. However, elucidating from these comprehensive datasets the regulatory events most critical to producing the transcriptional patterns needed for distinctive cell types is still a major challenge.

Here, our major aim is to determine the value of systematic integration of the extensive epigenomic data to improve accessibility and understanding of the data and to facilitate the generation of novel hypotheses about changes in the regulatory landscape during hematopoietic differentiation. We used the **I**ntegrative and **D**iscriminative **E**pigenomic **A**nnotation **S**ystem (Zhang *et al.* 2016; Zhang and Hardison 2017) to learn and assign epigenetic states, which are common combinations of features such as nuclease accessibility, histone modifications, and CTCF occupancy, jointly along chromosomes and across 20 hematopoietic cell types. This methodological choice was guided by previous results showing that, compared to other segmentation approaches (Ernst and Kellis 2010; Ernst and Kellis 2012; Hoffman *et al.* 2013), IDEAS provides significant improvement in the precision and consistency of state assignments (Zhang *et al.* 2016; Zhang and Hardison 2017). It also is able to assign states in a principled, effective way despite missing data (Zhang and Mahony 2019).

The resulting segmentations provide a readily interpretable “painting” of the epigenomic landscape across selected hematopoietic cell types.

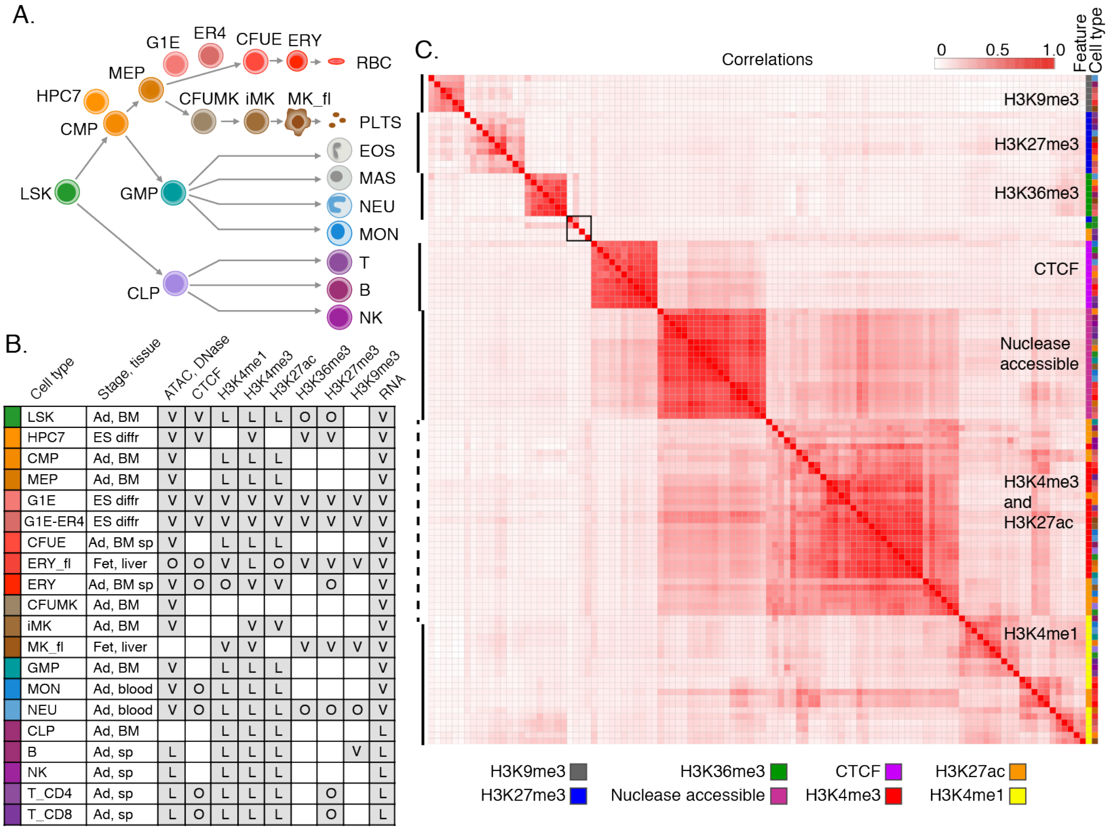
Furthermore, we combined the integrated features in the form of epigenetic states with peaks of nuclease accessibility to produce an initial compendium of over 200,000 candidate *Cis*-Regulatory Elements (cCREs) active in one or more hematopoietic lineages in mouse. Comparison with other datasets indicate these cCREs cover many of the known and likely regulatory elements, suggesting that this compendium is valuable for further study of individual loci and genome-wide assessments. Investigation of state transitions in the cCREs across differentiation revealed insights into epigenetic dynamics, including progressions from poised to active or inactive enhancers and loss of nuclease accessibility at some CTCF-bound sites. Furthermore, exploration of the correlations of cCRE states and gene expression produced a flexible, user-tunable resource for assigning cCREs to candidate target genes in the investigated cell types, which in turn can help explain the impacts of genetic variation in noncoding regions, including eQTLs and trait associated variants from genome-wide association studies.

Results

Epigenomic and transcriptomic datasets of mouse hematopoietic cells

A large number of genome-wide determinations of RNA levels and epigenetic features related to gene expression across hematopoietic cell types have been published. We reasoned that integrative analysis of these data should provide a more accessible view of the information that would help investigators utilize the multiple diverse datasets, and it may lead to novel insights into gene regulation. In order to build a set of data for integrative and discriminative analysis, we collated the raw sequence data for 150 determinations of relevant epigenetic features (104 experiments after merging replicates), including histone modifications and CTCF by ChIP-seq, nuclease accessibility of DNA in chromatin by ATAC-seq and DNase-seq, and 20 experiments on transcriptomes by RNA-seq. The data were gathered from a set of purified cell populations including LSK (Lin$^-$Sca1$^+$Kit$^+$, which includes hematopoietic stem cells or HSC and multipotent progenitor cells or MPP), several multilineage progenitor cells (common myeloid progenitor cells or CMP, granulocyte monocyte progenitor cells or GMP, megakaryocyte erythrocyte progenitor cells or MEP, and common lymphoid progenitor cells or CLP), and committed cells of the major blood cell lineages at different stages of maturity (for erythroblasts or ERY and megakaryocytes or MK) (**Figure 1A**). Most of the cell populations were from adult bone marrow or spleen, but some cell populations were from the hematopoietic organ fetal liver. We included data from three immortalized cell lines used extensively in mechanistic studies of gene regulation at distinct stages of differentiation and maturation. These were HPC7 cells, which are models for multilineage myeloid progenitor cells (Pinto do *et al.* 1998; Wilson *et al.* 2010), G1E cells, which are a model for early erythroid committed cells blocked in maturation by a knockout of the *Gata1* gene, and G1E-ER4 cells, a rescued subline of G1E that partially matures to erythroblast-like cells in a GATA1-dependent manner upon estradiol treatment to activate a GATA1-ER hybrid protein (Weiss *et al.* 1997; Gregory *et al.* 1999). This collection of cell populations was heavily weighted toward the erythroid and myeloid lineages, but representatives of some of the major lymphoid lineages were included to provide a broad context for the resources built from our integrative modeling.

To establish key signatures of epigenomic and transcriptomic data in various hematopoietic cell types, data were gathered from many different sources, including individual laboratories and consortia (**Figure 1B** and **Supplementary Tables**). The initially gathered data had quality metrics within the ENCODE recommendations (see **Materials and Methods** and **Supplementary Tables**). However, the diversity of sources presented a challenge for data analysis, since each experiment differed widely in sequencing depth, fraction of reads on target, signal-to-noise ratio, presence of replicates, and other properties (Xiang et al. 2019), all of which can impact downstream analyses. Two strategies were employed to improve the comparability of these heterogeneous datasets. First, the sequencing reads from each type of assay were uniformly processed, using pipelines similar to or adapted from current ENCODE pipelines (see **Materials and Methods**). One notable difference is that our VISION pipelines allow reads to map to genes and genomic intervals that are present in multiple copies, thereby allowing interrogation of duplicated chromosomal segments, including multigene families with highly similar genes such as those encoding globins, as well as regions subject to deletions and amplifications. Second, for the ChIP-seq and nuclease accessibility data, we applied a new normalization method, S3norm, that simultaneously adjusts for differences in sequencing depths and signal-to-noise ratios in the collected data (Materials and Methods and Xiang et al. 2019). As with other normalization procedures, the S3norm method was designed to give similar signals in common peaks for an epigenetic feature, but it does so without inflating the background signal. Preventing an increased background was necessary to avoid introducing spurious signals during the genome-wide modeling of the epigenetic landscape.



	Cell type	Stage, tissue	ATAC, DNase	CTCF	H3K4me1	H3K4me3	H3K27ac	H3K36me3	H3K27me3	H3K9me3	RNA
🟩	LSK	Ad, BM	V	V	L	L	L	O	O		V
🟧	HPC7	ES diffr	V	V		V		V	V		V
🟧	CMP	Ad, BM	V		L	L	L				V
🟧	MEP	Ad, BM	V		L	L	L				V
🟥	G1E	ES diffr	V	V	V	V	V	V	V	V	V
🟥	G1E-ER4	ES diffr	V	V	V	V	V	V	V	V	V
🟥	CFUE	Ad, BM sp	V		L	L	L				V
🟥	ERY_fl	Fet, liver	O	O	V	L	O	V	V	V	V
🟥	ERY	Ad, BM sp	V	O	O	V	V		O		V
🟫	CFUMK	Ad, BM	V								V
🟫	iMK	Ad, BM	V			V	V				V
🟫	MK_fl	Fet, liver			V	V		V	V	V	V
🟦	GMP	Ad, BM	V		L	L	L				V
🟦	MON	Ad, blood	V	O	L	L	L				V
🟦	NEU	Ad, blood	V	O	L	L	L	O	O	O	V
🟪	CLP	Ad, BM			L	L	L				L
🟪	B	Ad, sp	L		L	L	L			V	L
🟪	NK	Ad, sp	L		L	L	L				L
🟪	T_CD4	Ad, sp	L	O	L	L	L		O		L
🟪	T_CD8	Ad, sp	L	O	L	L	L		O		L

Figure 1. Hematopoietic cell types and datasets, and correlations among normalized data. **A.** Schematic representation of the main lineage commitment steps in hematopoiesis, along with three immortalized cell lines (HPC7, G1E, G1E-ER4) and their approximate position relative to the primary cell types shown. **B.** Available hematopoietic cell types and datasets. Shown in each row: Cell type along with its representative color, tissue stage (Ad = adult, ES diff = Embryonic stem cell derived, differentiated) and source (BM = bone marrow, sp = spleen, liver, blood). Shaded boxes indicate the presence of the dataset, and letters denote the source (V = VISION, L = Lara-Astiaso et. al 2014, O = other). See Supplementary Tables for more information. **C.** Correlations across all features (S3norm normalized) and cell types. The genome-wide Pearson correlation coefficients *r* were computed for each cell type-feature pair and displayed as a heatmap after hierarchical clustering (using 1-*r* as the distance measure). The features are indicated by a characteristic color (first column on right), and the cell types are indicated in the second column to the right using the same colors as panel **B**. The more intensely red colors within the matrix represent higher correlations.

An overview of the similarities across all the datasets showed that most clustered by epigenetic features across cell types (**Figure 1C**). These groupings within epigenetic features were more apparent after S3norm normalization (**Supplementary Figure 1**), which supports the effectiveness of the normalization. Determinations across cell types for nuclease accessibility, CTCF, the H3K9me3 heterochromatin mark, H3K27me3 Polycomb repressive mark, or the H3K36me3 transcriptional elongation mark were highly correlated. In contrast, the signature marks for promoters and enhancers, H3K4me3 and H3K4me1, respectively, formed groups interspersed with the H3K27ac modification, which is characteristic of active enhancers and promoters. These intermingled groups (e.g. H3K4me1 and H3K27ac) tended to form within related cell types, such as maturing erythroid cells or lymphoid cells, as expected for the cell type-specificity of enhancer-associated marks (Heintzman *et al.* 2009; Yue *et al.* 2014). The similarity of patterns for a particular feature across cell types suggests that examination of a single epigenetic mark may have limited power to find patterns distinctive to a cell type, whereas combinations of features appear to be more effective.

Despite our quality checks on the initially compiled experiments, four datasets were problematic. They failed to cluster with other datasets for that feature (H3K27ac for CD4 and CD8 T cells) or formed an unexpected group such as H3K27me3 with H3K36me3 for LSK (enclosed in a gray box in **Figure 1C**), even after normalization. Inclusion of these datasets in the integrative modeling (described below) generated chromatin states that were highly enriched only in those cell types, unlike the other states, suggesting that they contained artifactual signals. Thus, these four datasets were therefore excluded from the integrative and discriminative modeling.

In summary, our compilation of signal tracks, peak calls, estimates of transcript levels, and other material established a unified, consistently processed data resource for mouse hematopoiesis, which can be accessed at our VISION website (http://usevision.org).

Simultaneous integration in two dimensions of non-binary epigenomic data

The frequent co-occurrence of some histone modifications have led to discrete models for epigenetic structures of candidate *cis*-regulatory elements, or cCREs (reviewed in Noonan and McCallion 2010; Hardison and Taylor 2012; Long *et al.* 2016). Moreover, the co-occurrences can be modeled formally using genome segmentation to learn the most frequently occurring, unique combinations of epigenetic features, called epigenetic states, and assigning each segment of DNA in each cell type to an epigenetic state. Computational tools such as chromHMM (Ernst and Kellis 2012), Segway (Hoffman *et al.* 2012), and Spectacle (Song and Chen 2015) provide informative segmentations primarily in one dimension, usually along chromosomes. The **I**ntegrative and **D**iscriminative **E**pigenome **A**nnotation **S**ystem, or IDEAS (Zhang *et al.* 2016; Zhang and Hardison 2017) expands the capability of segmentation tools in several ways. It integrates the data simultaneously in two dimensions, along chromosomes and across cell types, thus improving the precision of state assignments. It uses continuous (not binarized) data as the input, and the number of epigenetic states is determined automatically. Also, when confronted with missing data, it can make state assignments with good accuracy (Zhang and Mahony 2019).

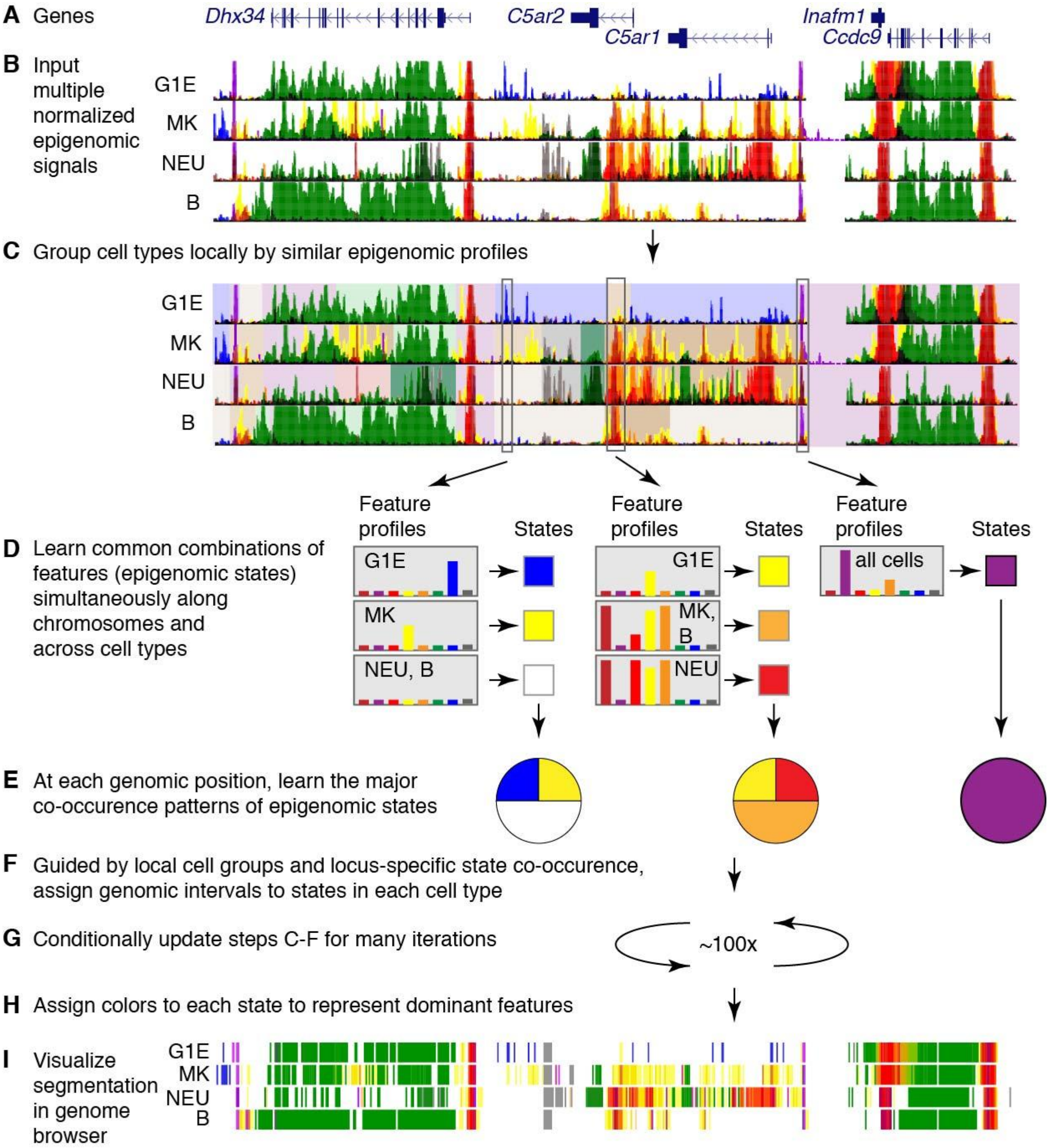

A Genes
Dhx34
C5ar2
C5ar1
Inafm1
Ccdc9
B Input multiple normalized epigenomic signals
G1E
MK
NEU
B
C Group cell types locally by similar epigenomic profiles
G1E
MK
NEU
B
D Learn common combinations of features (epigenomic states) simultaneously along chromosomes and across cell types
Feature profiles
States
G1E
MK
NEU, B
Feature profiles
States
G1E
MK, B
NEU
Feature profiles
States
all cells
E At each genomic position, learn the major co-occurence patterns of epigenomic states
F Guided by local cell groups and locus-specific state co-occurence, assign genomic intervals to states in each cell type
G Conditionally update steps C-F for many iterations
~100x
H Assign colors to each state to represent dominant features
I Visualize segmentation in genome browser
G1E
MK
NEU
B

Figure 2. Major steps in integrative and discriminative modeling of epigenomic signals using IDEAS. **A.** Gene models in a 100kb region centered on two complement receptor genes (position chr7:16,190,001-16,290,000 in GRCm38/mm10). **B.** In four cell types (G1E, MK, NEU, and B cells), the normalized signal for each of the eight epigenomic features was given a distinctive color (burgundy for ATAC-seq, purple for CTCF, red for H3K4me3, yellow for H3K4me1, orange for H3K27ac, green for H3K36me3, blue for

H3K27me3, and gray for H3K9me3), and the eight tracks were overlaid for each cell type using the Track Collections tool of the UCSC Genome Browser (Haeussler *et al.* 2019). **C.** The grouping of cell types locally based on their epigenetic profiles is illustrated by distinctive background colors, mauve background for chromosomal segments that have similar profiles across all cell types and different colors in backgrounds for segments with differing profiles. **D.** The epigenetic feature profiles that occur most commonly are illustrated for three genomic positions as bar graphs representing the intensity of signal for each of the eight features (each with the distinctive color listed in **B**). Those combinations of quantitative signals define an epigenetic state, illustrated as a colored square. The epigenetic state at a given position can be constant or different across cell types. **E.** The frequencies of occurrence of the states at the three genomic positions are illustrated as pie diagrams; the colors in the pie diagrams represent particular states. Panels **F**, **G**, and **H** indicate steps for assigning genomic intervals to epigenetic states in each cell type and giving them informative colors. **I.** The resulting segmentation for the four cell types at this locus is shown as a track in dense mode for a genome browser.

We employed IDEAS to analyze the normalized signals for nuclease accessibility, CTCF occupancy, and histone modifications (**Figure 2**). One key step in the IDEAS modeling is to group cell types *locally* based on the epigenomic profiles, finding regions that are similar across subgroups of cell types or that are similar across all cell types **(Figure 2 A-C**). Importantly, cell types that were more similar in one locus can differ in another locus. IDEAS then learned the epigenetic states, which were defined by the signal strength, not just the presence or absence of each feature (**Figure 2D**), while retaining position-specific information (**Figure 2E**). Leveraging the information about local cell type groups and the position-specific distributions of states, genomic intervals in each cell type were assigned to an epigenetic state in an iterative process (**Figure 2F-I**).

The coloring of each state was determined automatically, generating an informative representation for each state by mixing colors from a palette of distinctive colors for each feature. The colors provide a visual representation of the contribution of each epigenetic feature to the state. The output of the IDEAS segmentation effectively paints the epigenome of each cell type in a distinctive pattern, providing a compact and concise display of function-associated states along the chromosomes of each cell type. Importantly, the segmentation provides a simplified, integrated representation of over 100 tracks of epigenomic data, enabling investigators examine the entire dataset in a concise form.

The resulting 27 epigenetic states included many expected ones, as well as others that have been less frequently studied. The IDEAS model summary shows the prevalence of each of the eight epigenetic features as a heatmap, organized by similarity among the states (**Figure 3A**). These states described an informative landscape, distinguishing multiple states signatures representing distinct classes of regulatory elements (including enhancers, promoters and boundary elements). For example, six states showed a promoter-like signature, with high frequency of H3K4me3 (states 18, 21, 10, 15, 24, and 11); these are displayed in different shades of red, and P is the initial character in the explicit label. These six states distinguished promoter-like signatures by the presence or absence of other features with functional implications. For instance, four promoter-like states were also nuclease accessible (states 21, 10, 15, and 24), four also had the H3K27ac mark associated with active promoters (states 18, 21, 10, and 24), one (state 24) also had CTCF, and three had the H3K4me1 modification that flanks active promoters as well as marks enhancers (states 21, 24, and 19). Two states (19 and 23) had equivalent frequencies of the tri- and monomethylated H3K4, and these were categorized as a mix of promoter-like and enhancer-like signatures, colored orange. States 22 and 20 had a high frequency of the Polycomb repressive mark H3K27me3 along with methylated H3K4, and they were categorized as bivalent states (Bernstein *et al.* 2006). Similarly, multiple states related to CTCF occupancy (shades of purple), enhancer-like signatures (shades of yellow and orange), transcriptional elongation (shades of green), polycomb repression (shades of blue), and H3K9me3-associated heterochromatin (shades of gray) were learned in the IDEAS modeling. Several states did not fall exclusively into one of these common categories. While H3K9me3 is frequently associated with heterochromatin, state 17 had the H3K9me3 modification together with the transcription elongation mark H3K36me3. This state is unlikely to be in repressed heterochromatin, but it is reminiscent of a previously reported association of H3K9 methylation with transcriptional elongation (Vakoc *et al.* 2005), a combination that was also described for KRAB-zinc finger genes (Hahn *et al.* 2011) and found more generally by Segway (Hoffman *et al.* 2012). Other states identified by IDEAS have not been previously considered in detail. One state had the expected co-occurrence of CTCF and nuclease accessibility (state 13), but an even more common state had CTCF without nuclease accessibility (state 7). While states predominated by H3K27me3 alone (state 3) or H3K9me3 alone (state 2) were common, state 16 had both repressive marks. Thus, the IDEAS segmentation learned and assigned a diverse set of states that not only included previously described epigenetic signatures but also identified some new states.

The fraction of the genome in each state reveals the proportion of a genome associated with a particular activity. The most common state in all the epigenomes is quiescence, i.e. state 0 with low signals for all the features (**Figure 3B**). The mean percentage of the genome in this state was 86%, with values ranging from 85% to 92% in individual cell types. Interestingly, 60% of the genome was in this state in all cell types examined, indicating that in hematopoietic cells, about 40% of the mouse genome is incorporated within chromatin with the dynamic histone modifications identified in this study. The most common non-quiescent states were transcribed, heterochromatic, and Polycomb repressed (**Figure 3B**). The remaining portion of the genome was populated with a large number of active states, comprising ~4% of the genome. Thus, only a small proportion of the genome in each cell type was found in chromatin associated with the dynamic histone modifications assayed here. Importantly, this small fraction of the genome is probably responsible for much of the regulated gene expression characteristic of each cell type.

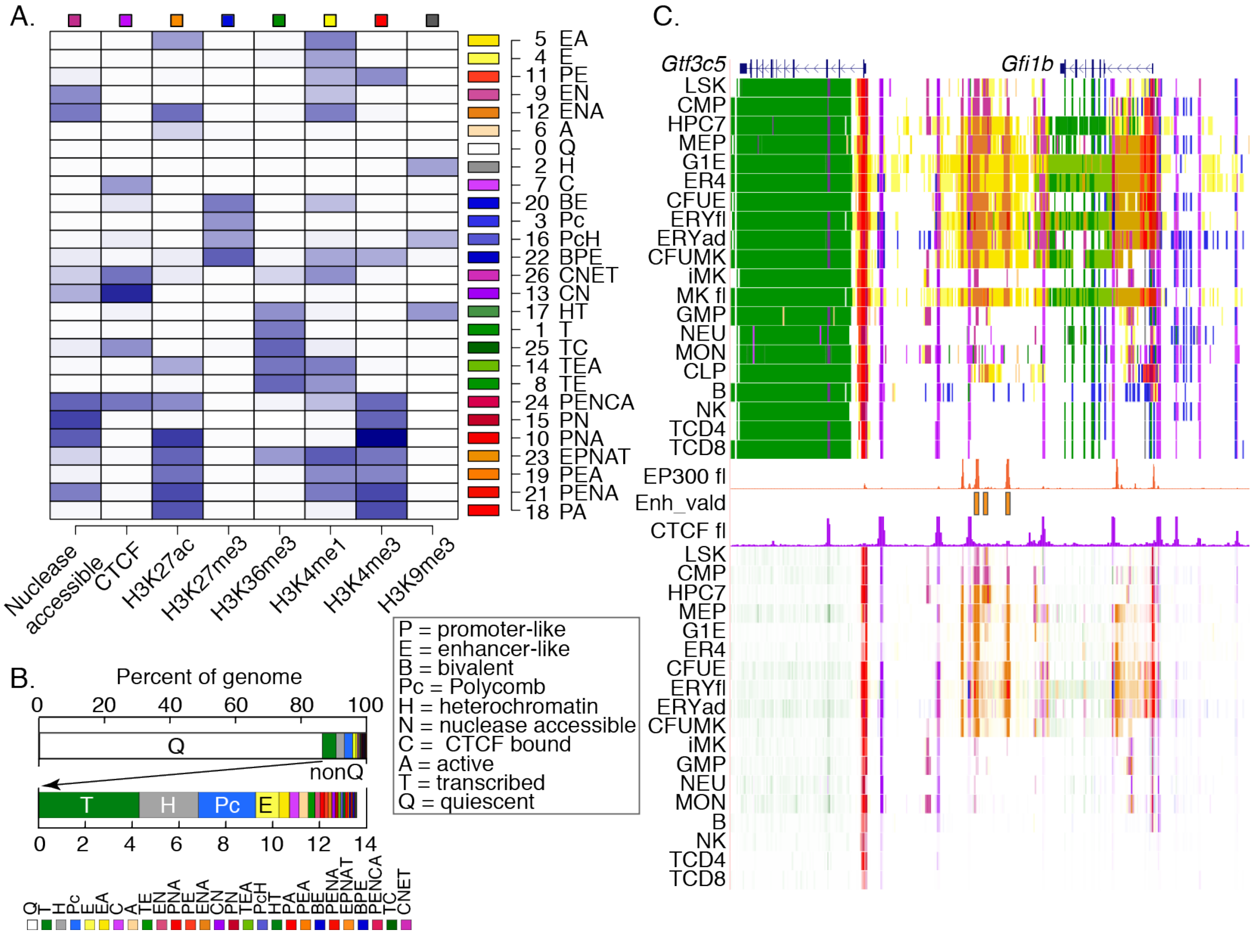


Figure 3. Segmentation of the epigenomes of hematopoietic cells after integrative modeling with IDEAS. **A.** Heatmap of the emission frequencies of each of the 27 states, with state number and function-associated labels. Each letter in the label indicates a function associated with the combination of features in each state, defined in the *box*. The indicator for transcribed is H3K36me3, active is H3K27ac, enhancer-like is H3K4me1>H3K4me3, promoter-like is H3K4me3>H3K4me1, heterochromatin is H3K9me3, and polycomb is H3K27me3. **B.** Bar graphs of the average coverage of genomes by each state. The top graph emphasizes the high abundance of state Q, and the second graph shows the abundances of the 26 non-quiescent states. The key for annotated colors is the same order as the states in the bar graph. **C.** Segmentation pattern at an exemplar locus, *Gfi1b*, covering 70kb from

chr2:28,565,001-28,635,000 in GRCm38/mm10. Signal tracks for EP300 (ENCSR982LJQ, ENCODE consortium) and CTCF from mouse fetal liver were included for validation and confirmation, along with the locations of enhancers shown to be active (Enh_vald; Moignard *et al.* 2013). The lower set of tracks shows nuclease accessibility, with more intense signal representing greater accessibility, colored by the epigenetic state assignments.

Visualizing the regulatory landscape across hematopoietic cell types as defined by the IDEAS segmentation

The IDEAS segmentation gives an informative view of the chromatin activity landscape along chromosomes and across cell types. The usual display assigns the distinctive color for each state to a DNA segment (**Figure 3C**). For example, genes transcribed in all cell types, such as *Gtf3c5* and *Tsc1*, were painted red at the active promoter and green for regions of transcriptional elongation. Within and between the transcription units are short purple segments Indicating CTCF binding, aligning with the CTCF occupancy data available for tissues like fetal liver and providing a prediction for CTCF binding in other cell types. The gene *Gfi1b*, encoding a transcription factor required in specific hematopoietic lineages, shows different state assignments across the cell types, with active promoters (red), intronic enhancers (orange), and transcribed regions (green) in CMP, erythroid, and megakaryocytic cells but fewer active states in other cell types. Downstream (left) of *Gfi1b* was a large region with many DNA segments assigned to enhancer-associated states; these were model-generated candidates for regulating expression of *Gfi1b*. The potential role of the intronic and downstream candidate enhancers was supported by the binding by the coactivator EP300 observed both in mouse fetal liver and MEL cells (Yue *et al.* 2014; The_ENCODE_Project_Consortium *et al.* 2019), information that was not included in training the model. Furthermore, previous studies of cross-regulation between GATA2 and GFI1B revealed three enhancers downstream of the *Gfi1b* gene by reporter gene assays in transgenic mouse and transfected cells (Moignard *et al.* 2013). These enhancers overlap with the model-predicted enhancers and provide strong experimental validation of the predictions from the IDEAS segmentation.

Given the frequent association between nuclease sensitivity and *cis*-regulatory elements, we combined the IDEAS segmentation results with nuclease accessibility (ATAC-seq or DNase-seq) signal intensity to generate a view of the epigenetic landscape with an emphasis on DNA segments likely to be involved in regulation. The intensity of the state-associated color was adjusted by the signal of the nuclease-sensitivity data, thereby showing more highly accessible regions as more intense colors (lower portion of **Figure 3C**). Thus, in a single track one can view both the level of accessibility and the state assignment determined by the integration of epigenetic features. This view emphasizes the candidate regulatory elements.

cCREs across mouse hematopoiesis

While genomic regions potentially involved in gene regulation can be discerned from the segmentation views of regulatory landscapes, it is important to assign discrete genomic intervals as candidate *cis*-regulatory elements (cCREs) to clarify assessments and validations of regulatory elements and to empower systematic modeling of regulatory systems. Therefore, we combined our nuclease sensitivity data with IDEAS segmentation to infer a set of 205,019 cCREs in the 20 cell types.

A cCRE was defined as a DNA segment assigned as a validated peak by ATAC-seq or DNase-seq that was not in a quiescent epigenetic state in all cell types. We considered ATAC-seq or DNase-seq data to be validated when peaks were called in each replicate. Some peaks were assigned to the quiescent state in all cell types, and these were removed from the set of cCREs. No cCREs could be called in mature MK or CLP cells because no ATAC-seq or DNase-seq data were available for these cell types; however, we inferred the epigenetic states in these two cell types for the DNA segments predicted to be cCREs in other cell types. This information about the locations and epigenetic states of cCREs in hematopoietic cell types provides a valuable resource for detailed studies of regulation both at individual loci and globally across the genome.

Because a wide range of hematopoietic cells were interrogated for epigenetic features, we expected that the set of cCREs from the VISION project would expand and enhance other collections of cCREs. Thus, we compared the VISION cCRE set with the Blood Cell Enhancer Catalog, which contains 48,396 candidate enhancers based on iChIP data in sixteen mouse hematopoietic cell types (Lara-Astiaso *et al.* 2014), and a set of 56,467 cCREs from mouse fetal liver released by the ENCODE project (The_ENCODE_Project_Consortium *et al.* 2019). Furthermore, we examined the set of 431,202 cCREs across all assayed mouse tissues and cell types in the SCREEN cCRE catalog from ENCODE (The_ENCODE_Project_Consortium *et al.* 2019). The overlapping DNA intervals in the four datasets were merged to generate a common set of DNA intervals for comparison; this merger reduced the number of cCREs in each set. The overlapping DNA intervals among several combinations of datasets revealed substantial consistency among these sets of inferred cCREs (**Figure 4A**). A large portion of the VISION cCREs (70,445 or 41.5%) were also in the iChIP Blood Enhancer Catalog and/or the SCREEN fetal liver cCREs. Conversely, a majority of the cCREs in the iChIP catalog (78.7%) were also in VISION cCREs, as expected given the large contribution of iChIP data to the VISION compilation. An even larger proportion (84%) of the SCREEN fetal liver catalog was also in VISION cCREs. The cCREs that are common to each collection, despite differences in data input and data processing, are strongly supported as candidate regulatory elements.

The VISION cCRE set is substantially larger than either the iChIP Blood Enhancer Catalog or the SCREEN fetal liver cCREs, and we hypothesized that the larger size reflected the inclusion of greater numbers of cell types and features in the VISION catalog. This hypothesis predicts that VISION cCREs that were not in the other blood cell cCRE sets may be found in larger collections of cCREs, and we tested this prediction by comparing VISION cCREs to the entire set of ENCODE SCREEN cCREs. Indeed, we found another 58,504 (34.5%) VISION cCREs matching this catalog across mouse tissues, supporting the interpretation that the VISION cCRE set is more comprehensive than other current blood cell cCRE collections. Overall, the comparisons with other collections supported the specificity and accuracy of the VISION cCRE set.

To further assess the quality of the VISION cCRE set, we evaluated its ability to capture known *cis*-regulatory elements (CREs) and independently determined DNA elements associated with gene regulation. Using a collection of 212 experimentally determined erythroid CREs curated from the literature (Dogan *et al.* 2015) as known erythroid CREs, we found that while the iChIP Blood Enhancer catalog captured only a small portion, the VISION and SCREEN fetal liver cCREs overlapped with almost all the erythroid cREs (**Figure 4B**). The latter two collections were built from datasets that included highly erythroid tissues, such as fetal liver, which may explain their more complete coverage than the Blood Enhancer Catalog, which was built from datasets from fewer erythroid cell types. Increasing the number of cCREs to over 400,000 in the SCREEN mouse cCREs did not substantially increase the number of known CREs that overlap. Thus, the VISION cCREs efficiently captured known erythroid CREs.

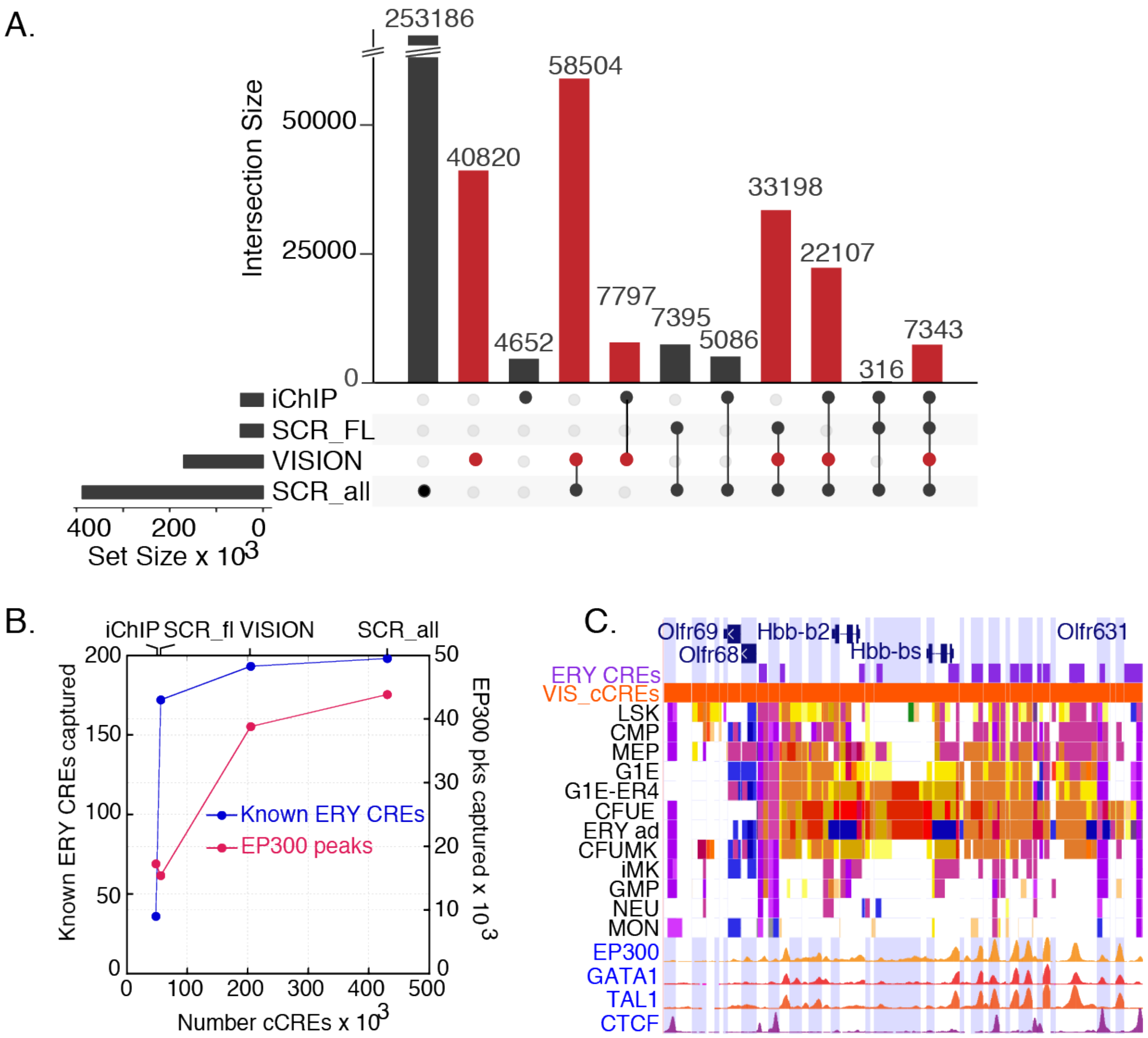


Figure 4. Comparative analysis of VISION cCREs. **A.** Overlaps of the VISION cCREs with three other cCRE catalogs. The overlapping cCREs in all four datasets were merged. The numbers of merged cCREs in each set were labeled on each row, and the numbers in each level of overlap were shown in columns, visualized using an UpSet plot (Lex *et al*. 2014). The sets compared with the VISION cCREs were the Blood Enhancer Catalog derived from iChIP data (iChIP; Lara-Astiaso *et al.* 2014), the SCREEN cCREs specific to mouse fetal liver at E14.5 (SCR_FL), and those for all tissues and cell types in mouse (SCR_all). **B.** The VISION cCREs capture known regulatory elements and orthogonal predicted cCREs. The number of known CREs that are also present in each cCRE collection was plotted against the number of regulatory elements (known or inferred) in each dataset. The EP300 peaks were deduced

from EP300 ChIP-seq data from ENCODE, reprocessed by VISION pipelines, from FL E14.5, MEL, and CH12 cells. Replicated peaks were combined into one dataset and merged, to get over 60,000 peaks. The number of known EP300 peaks that were also present in each cCRE collection was plotted against the number of cCREs in each dataset. **C.** A cCRE-centric view of epigenomic data revealing support for the cCRE calls and state assignments. The multi-region view feature of the UCSC Genome Browser (Haeussler *et al.* 2019) was used to display only the DNA intervals in the VISION cCRE dataset; alternating cCREs have white or light blue backgrounds to distinguish them. Only the portions of gene models that overlap cCRE intervals are shown. The IDEAS state assignment in selected cell types are shown followed by independent data on binding of the coactivator EP300, GATA1, and TAL1. CTCF binding tracks (used in the IDEAS modeling) are also shown to illustrate binding in these CTCF-associated states. The region covered in this view was 250kb from chr7:103,700,001-103,950,000 in mm10, encompassing the *Hbb* gene complex and flanks. This region was reduced to 17.6kb when showing only the cCREs.

The co-activator EP300 catalyzes the acetylation of histone H3K27, and it is associated with many active enhancers. We used ChIP-seq data on EP300 in three blood-related cell types from mouse as a comparison set of blood cell candidate enhancers that were determined independently of the data analyzed in VISION. The ENCODE consortium has released replicated datasets of EP300 ChIP-seq data determined in two cell lines, MEL cells representing maturing proerythroblasts and CH12 cells representing B cells (Yue *et al.* 2014), and one tissue, mouse fetal liver from day E14.5 (The_ENCODE_Project_Consortium *et al.* 2019). After re-processing the ChIP-seq data using the VISION project pipelines, replicated peaks were merged across the cell types to generate a set of over 60,000 EP300 peaks in blood related cells. The VISION cCRE set efficiently captured the EP300 peaks, hitting almost two-thirds of these proxies for regulatory elements, a much larger fraction than captured by the Blood Enhancer catalog or ENCODE fetal liver cCREs (**Figure 4B**). Expanding the number of cCREs to over 400,000 gave only a small increase in the number of EP300 peaks captured. The EP300 peaks not captured by the VISION cCREs tended to have lower signal strength and were less associated with ontology terms such as those for mouse phenotype (Supplementary Material), suggesting that the more likely functional EP300 peaks were captured by VISION cCREs.

These analyses show that the VISION cCREs included almost all known erythroid CREs and they captured a large fraction of potential enhancers identified in relevant cell types by a different feature (EP300). Both these observations supported the quality of the VISION cCRE set. More generally, the cCREs were expected to overlap with transcriptional regulatory proteins, such as transcription factors, co-activators, and CTCF. This expected overlap was apparent when the binding profiles for such regulatory proteins were observed in a cCRE-centric view of the genome. By restricting the genome browser to show only the cCREs, all the dispersed candidate regulatory elements can be viewed together, e.g. focusing on the *Hbb* gene complex (**Figure 4C**). The cCREs in enhancer-like states tend to be co-bound by GATA1, TAL1, and EP300, while those in the CTCF-associated states were indeed bound by CTCF. Also, the epigenetic states of cCREs change across cell types, consistent with changes in expression of target genes.

Global comparisons of regulatory landscapes and transcriptomes

The collection of cCREs and transcriptomes in VISION provide an opportunity to examine the relationships between cell types, including both purified populations of primary cells and cell lines. The cCREs are a prominent feature of the regulatory landscape, and therefore we used the correlations between the nuclease accessibility signals in cCREs across cell types to group the cell types by hierarchical clustering (**Figure 5A**). All erythroid cell types, including the G1E and G1E-ER4 cell lines, and the MEPs clustered to the exclusion of other cell types. The remaining cell types form two groups. One consists of hematopoietic stem and multilineage progenitor cells (LSK, CMP and GMP) along with early progenitor (CFUMK) and immature (iMK) megakaryocytic cells. The other contains both innate (NEU, MON) and acquired (B, NK, T-CD4, T-CD8) immune cells. Comparisons using a dimensional reduction approach (principal component analysis or PCA) confirmed these groupings and revealed additional insights (**Figure 5B**). The first principal component (PC1) captured a substantial fraction (82%) of the variation, placing the cell types along an axis with many multilineage progenitor cells on one end and many mature cells on the other. As explored in more detail below, this axis relates to the numbers of active cCREs, with more cCREs in the multilineage progenitor and megakaryocytic cells and fewer in other maturing lineages. The second component separated erythroid cells (to the left) from other cells, and the third component tended to separate multilineage progenitor cells (toward the top) from more mature cells (toward the bottom). Thus, both the PCA and hierarchical clustering of nuclease sensitivity data in cCREs largely supported the groupings of megakaryocytic cells with progenitor cells along with separate clusters of erythroid and immune cells.

The gene expression landscape was then compared across cell types, using estimates of gene transcript levels from RNA-seq data in a subset of 12 cell types interrogated by the same method within our VISION laboratories. RNA-seq data on acquired immunity cells were not included because the assay was done by a substantially different procedure (Lara-Astiaso *et al.* 2014), and this difference in RNA-seq methodology dominated the combined comparison. The hierarchical clustering results (**Figure 5C**) and PCA (**Figure 5D**) revealed three clusters that were consistent with the analysis of the regulatory landscape, grouping megakaryocytic cells with multilineage progenitors while keeping primary erythroid cells (CFUE and ERY) and innate immune cells (NEU and MON) in distinct groups. In contrast, MEP cells grouped with progenitor cells in the transcriptome profiles whereas they grouped with erythroid cells by nuclease sensitivity data. MEP cells have a pronounced erythroid bias in differentiation (Psaila *et al.* 2016), and this difference in the grouping of MEPs suggests that the regulatory landscape of

MEP has shifted toward the erythroid lineage prior to reflecting that bias in the transcriptome data. G1E and G1E-ER4 cell lines, which are models for GATA1-dependent erythroid differentiation, also were placed differently based on cCRE and transcriptome data, forming a separate cluster in the transcriptome data. While that result reveals a difference in the overall RNA profiles between G1E and G1E-ER4 cells versus primary cells, their grouping with primary erythroid cells by cCRE landscape supports the use of these cell lines in specific studies of erythroid differentiation.

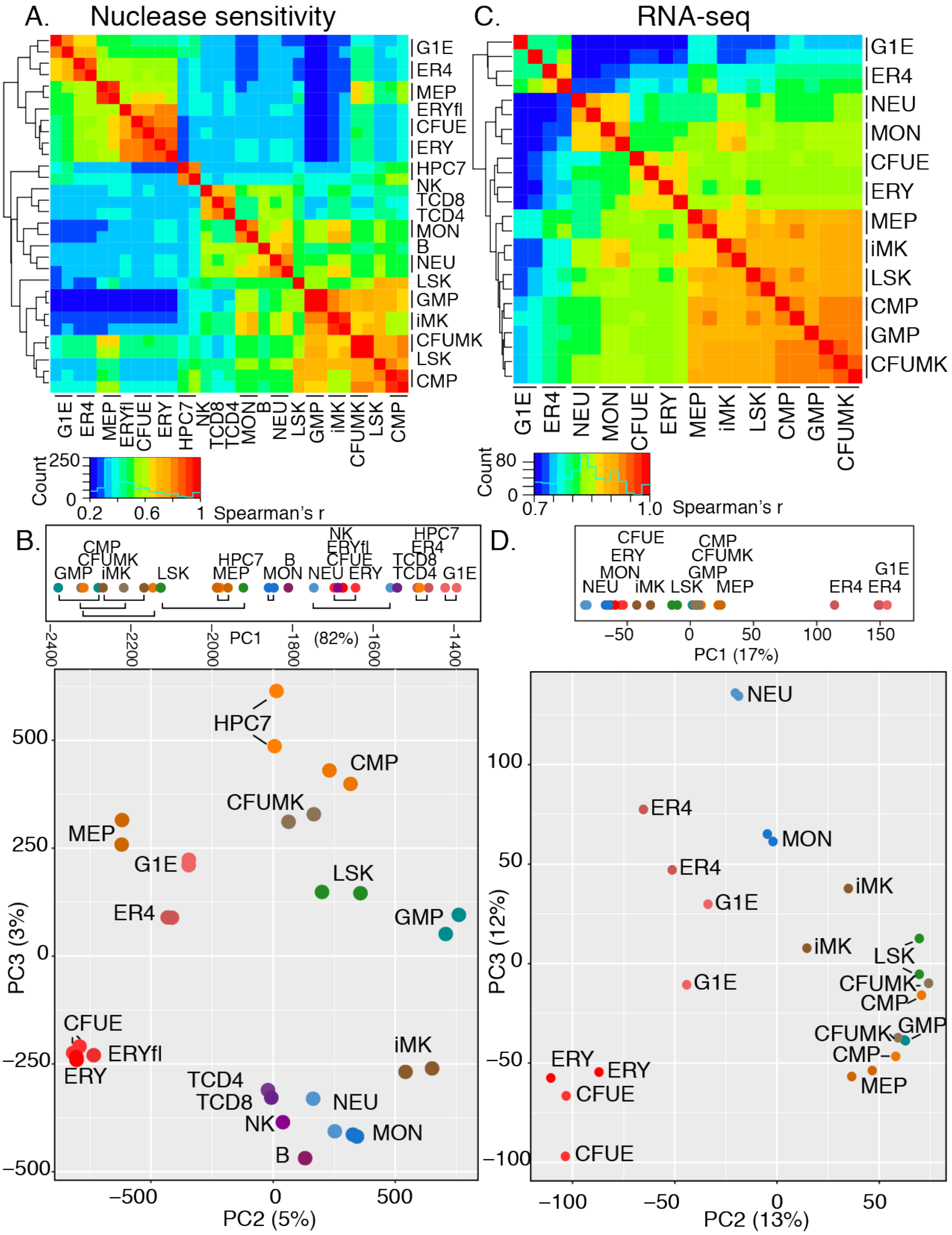

A.
Nuclease sensitivity
G1E
ER4
MEP
ERYfl
CFUE
ERY
HPC7
NK
TCD8
TCD4
MON
B
NEU
LSK
GMP
iMK
CFUMK
LSK
CMP
Count
250
0
0.2
0.6
1
Spearman's r
C.
RNA-seq
G1E
ER4
NEU
MON
CFUE
ERY
MEP
iMK
LSK
CMP
GMP
CFUMK
Count
80
0
0.7
1.0
Spearman's r
B.
CMP
CFUMK
GMP
iMK
LSK
HPC7
MEP
B
MON
NK
ERYfl
CFUE
NEU
ERY
HPC7
ER4
TCD8
TCD4
G1E
-2400
-2200
-2000
PC1
-1800
(82%)
-1600
-1400
D.
CFUE
ERY
MON
NEU
iMK
CMP
CFUMK
GMP
LSK
MEP
G1E
ER4
ER4
-50
0
50
100
150
PC1 (17%)
HPC7
CMP
CFUMK
MEP
G1E
LSK
ER4
GMP
CFUE
ERYfl
ERY
iMK
TCD4
TCD8
NEU
NK
MON
B
500
250
0
-250
-500
PC3 (3%)
-500
0
500
PC2 (5%)
NEU
ER4
MON
ER4
iMK
G1E
iMK
G1E
LSK
CFUMK
CMP
CFUMK
GMP
CMP
ERY
ERY
MEP
CFUE
CFUE
100
50
0
-50
-100
PC3 (12%)
-100
-50
0
50
PC2 (13%)

Figure 5. Global comparisons of nuclease accessibility profiles and transcriptomes across mouse hematopoietic cell types. **A.** Heatmap of the hierarchical clustering of nuclease-sensitive elements (ATAC-seq and DNase-seq, using S3norm for normalization), with Spearman's rank correlation *r* as the similarity measure, and 1-*r* as the distance measure for hierarchical clustering across 18 cell types. Results include replicates for cell types with replicated data (indicated by bars next to the cell type name). **B.** PCA to show groups of cell types, using ATAC-seq and DNase-seq profiles. **C.** Heatmap of the hierarchical clustering of RNA-seq (TPM values for all genes, quantile normalized, showing replicates), with Spearman's *r* as the similarity measure. **D.** PCA to show groups of cell types, using RNA-seq.

Numbers of cCREs and their states vary across cell types in an informative manner.

The VISION catalog of cCREs, annotated by their presence and epigenetic state in each cell type, can be used to track during differentiation when regulatory elements become active or lose activity and to follow the transitions in epigenetic states. This information can provide insights into mechanisms of regulation, e.g. which CREs are likely to be inducing or repressing a target gene at a specific stage of differentiation. The full scope of state transitions in cCREs across cell types is quite complex, and ongoing work aims to provide resources and tools to enable interpretations of the transitions. In this section, we focus on major transitions contributing to changes in the numbers of cCREs.

One major change in global regulatory landscapes was a striking reduction in the numbers of active cCREs (nuclease-accessible regions in a non-quiescent epigenetic state) after commitment to a single cell lineage. The number of active cCREs was consistently higher in multilineage progenitor cell populations and cells on the megakaryocytic lineage (CFUMK and iMK) than in the other lineage-committed cell types (**Figure 6A**). This trend was observed after normalizing for differences in sequencing depth, and was robust to changes in thresholds for peak calling (**Supplementary Figures 2 and 3**). Furthermore, the nuclease sensitive peaks showed enrichment for histone modifications that support the accuracy of the peak calls (**Supplementary Figure 4**). The values for PC1 in the nuclease accessibility analysis (**Figure 4B**) were strongly associated with the decrease in the number of nuclease sensitivity peaks during differentiation (**Figure 6B**, Pearson's correlation *r*=0.92), indicating that the numbers of nuclease sensitive elements were a strong contributor to this principal component that explained a large proportion (82%) of the variation.

The decrease in numbers of cCREs during differentiation and maturation could be associated with a decrease in numbers of genes expressed. We tested this and found that the numbers of expressed protein-coding genes were high (8800 to 10,000) in the progenitor (LSK, CMP, GMP, MEP) and megakaryocytic (CFUMK and iMK) cells, medium (~ 8000) in MON and NEU, and low (~6500) in erythroid cells (CFUE and ERY) (**Figure 6C**). A larger number of genes (8000 to 8500) were expressed in the ES-derived cell lines, G1E and G1E-ER4, than in the primary erythroid cells. A similar decline was observed over a ten-fold range of thresholds for declaring a gene as expressed (TPM exceeding 1, 5 or 10). The parallel decreases in numbers of active cCREs and expressed coding genes led to a strong positive association between these two features (**Figure 6C**; Pearson correlation *r*= 0.90 or 0.78 when values for G1E and G1E-ER4 cells were excluded and included, respectively, in a linear fit). The numbers of noncoding genes expressed were also positively associated with the numbers of nuclease accessible peaks (**Figure 6D**, Pearson's correlation *r*= 0.64 or 0.55 when values for G1E and G1E-ER4 cells were excluded and included, respectively, in a linear fit). These reductions in expressed genes and active cCREs indicated a progressive decrease in the breadth of transcription during differentiation, and furthermore the loss of activity of cCREs may contribute to the decrease in numbers of genes expressed. Similar results were reported for transitions during megakaryopoiesis and erythropoiesis in Heuston et al (2018) based on peak calls for histone modification and nuclease accessibility. Our results based on integrative modeling confirm these conclusions and show that they apply more broadly across hematopoiesis.

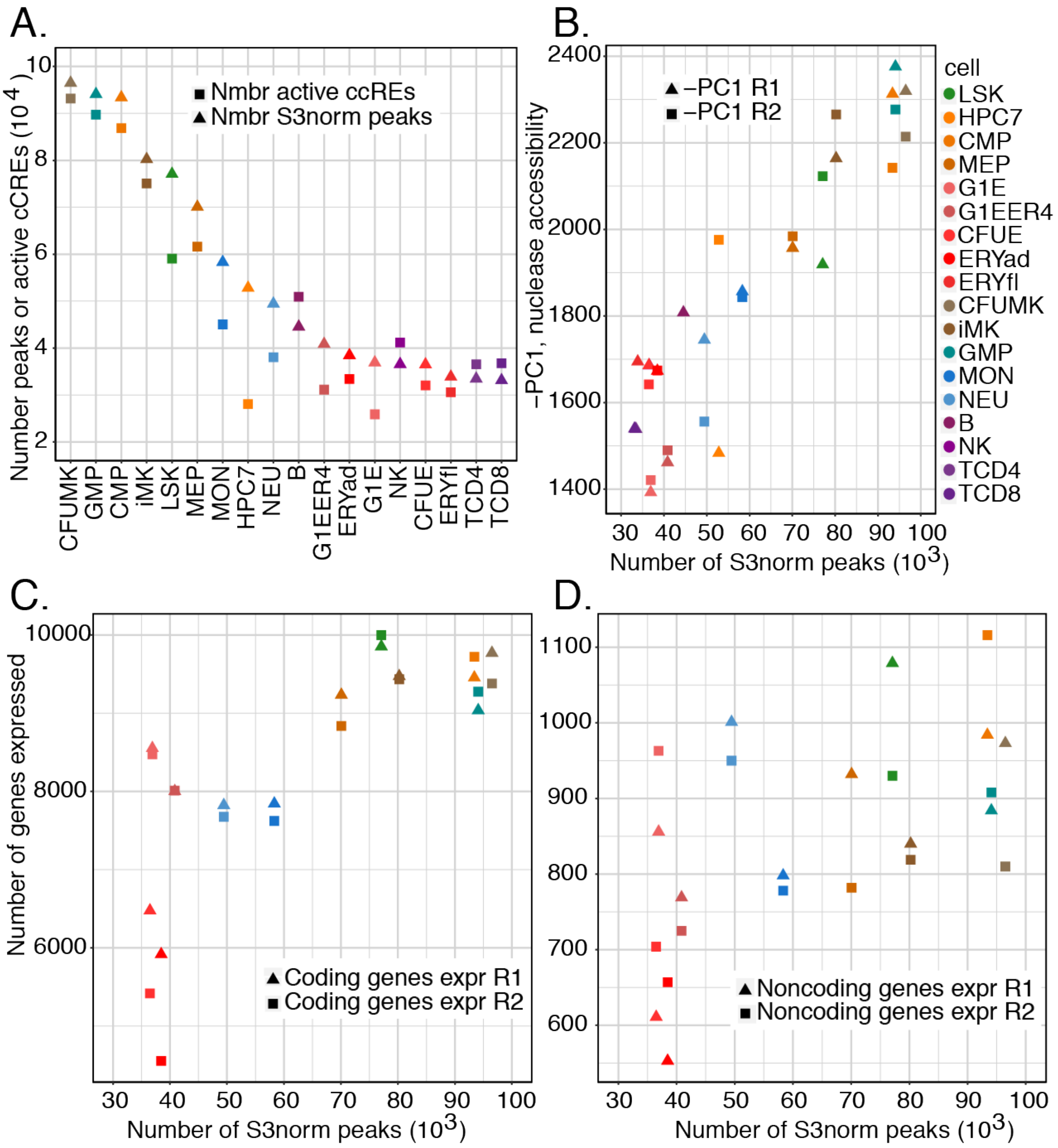


Figure 6. Concordant decreases during hematopoietic differentiation in nuclease accessibility and gene expression. **A.** Numbers of active cCREs and nuclease accessible regions (after S3norm normalization) in each cell type. **B.-D.** Positive association between numbers of nuclease accessible peaks in each cell

type and (**B**) negative of PC1 values in PCA of nuclease accessibility, (**C**) numbers of expressed protein-coding genes (TPM>=1), and (**D**) numbers of expressed non-protein-coding genes (TPM>=1) . Values for determinations in replicates are shown in panels **B-D**. The color code for cell types is displayed in panel **B**. R1 and R2 refer to replicates.

Next, we investigated transitions in epigenetic states throughout hematopoietic differentiation that help explain some of the reduction in active cCREs. Within the dominant pattern of decreasing numbers of active cCREs during commitment and maturation of lineages (except MK), the reduction was particularly pronounced for cCREs in state 9 (EN) and state 13 (CN) (**Figure 7A**) while the numbers of cCREs in other states did not show consistent trends (**Figure 7B)**. These state-specific reductions suggested that many cCREs in progenitor and MK cells were in a poised enhancer mode (state 9 EN) or in a CTCF-bound nuclease accessible state (state 13 CN). We then determined the states into which these cCREs tended to transition by examining all state transitions in cCREs between all pairs of cells. In the case of CMP cells differentiating to ERY, we found that cCREs in the poised enhancer state 9 in CMP did not stay in state 9, but rather they most frequently transitioned to states 12 (active enhancer), 3 (polycomb), and 0 (quiescent) in ERY (**Supplementary Figure 8A**). These classes of state transitions were strongly supported by examination of the underlying signals for the nuclease sensitivity and histone modifications (**Figure 7C**). This systematic analysis of transitions in epigenetic states across cell types helps uncover the differentiation history of cCREs and provides mechanistic insights into regulation. For example, using SeqUnwinder (Kakumanu *et al.* 2017) to discover discriminative motifs, we found that the CMP cCREs that transition from poised to active enhancer in the erythroid lineage were enriched for the GATA transcription factor binding site motif, whereas those that transition to a polycomb state were enriched in motifs for binding ETS transcription factors such as PU.1. Furthermore, these results illustrate specific mechanisms for the recent report of substantial changes in epigenomic landscape during differentiation of CMP to ERY (Heuston *et al.* 2018).

Another major state of cCREs in progenitor and megakaryocytic cells was CTCF-bound and nuclease accessible (state 13). Surprisingly, much of the decrease in numbers of cCREs in this state occurred through a loss of accessibility while retaining occupancy by CTCF (state 7, **Supplementary Figure 8**). The frequent observation of transitions from a nuclease-accessible, CTCF-bound state to a non-accessible state still bound by CTCF raised the question of whether the nuclease accessibility data were sufficiently sensitive, i.e. could the apparent lack of accessibility reflect false negatives in the input data? We leveraged the diversity of data in VISION to address this concern by examining two independent types of nuclease accessibility data. Specifically, we examined DNase-seq data on ERY from fetal liver (fl) and ATAC-seq data on ERY from adult bone marrow (ad) in the context of the chromatin state transitions. As a positive control, we first examined the set of 8081 cCREs that stayed in state 13 in both LSK and ERY. As expected, they show consistently strong signals for both nuclease accessibility and CTCF ChIP-seq both in multilineage progenitor cells (LSK, CMP, and MEP) as well as differentiating CFUE and ERY from both developmental stages (**Figure 7D**). In contrast, the set of 6354 cCREs that transition from state 13 in LSK to state 7 in ERY (adult bone marrow) showed a consistent loss of nuclease accessibility both for ATAC-seq signal in adult bone marrow ERY and for DNase-seq signal in fetal liver ERY while retaining a strong CTCF signal (**Figure 7D**). Furthermore, the loss of nuclease accessibility was observed in the CFUE precursor to ERY. Because the loss of nuclease accessibility was robust to different methods of determination and was observed in multiple related cells at different developmental stages, we concluded that the state 13 to state 7 transition was not an artifact of poor sensitivity of the accessibility assays. The loss of nuclease accessibility at this subset of CTCF-bound sites occurred between MEP and CFUE stages, suggesting that it could be connected to the process of erythroid commitment.

In summary, the number of active cCREs declines dramatically as cells differentiate from stem and progenitor cells to committed, maturing blood cells. This decrease in cCREs is strongly associated with a reduction in the numbers of expressed genes in committed cells. Our analysis of epigenetic states in cCREs across this process revealed major declines in two states. First, the poised enhancer state is prevalent in cCREs in stem and progenitor cells, and it has two major fates. One is a transition to an active enhancer state, and in the erythroid lineage this transition is associated with GATA transcription factor binding site motifs, as expected for activation of erythroid genes. The other fate is to lose nuclease sensitivity and switch to a repressed state. The occurrence of those transitions is not a novel observation, but our extensive annotation of the cCREs allows investigators to identify which cCREs around genes of interest are making those transitions. Second, another state prevalent in stem and progenitor cells is a CTCF-bound and nuclease accessible state, and the number of cCREs in that state declines during differentiation. Surprisingly, we found that a transition to a state with CTCF still bound but no longer nuclease accessible was a strong contributor to the decline in numbers of active cCREs. Further studies are needed to better understand the roles of these different classes of CTCF-bound sites. Many other state transitions were observed, and we have developed a tool to help organize and visualize these results (**Supplementary Figures 9 and 10**).

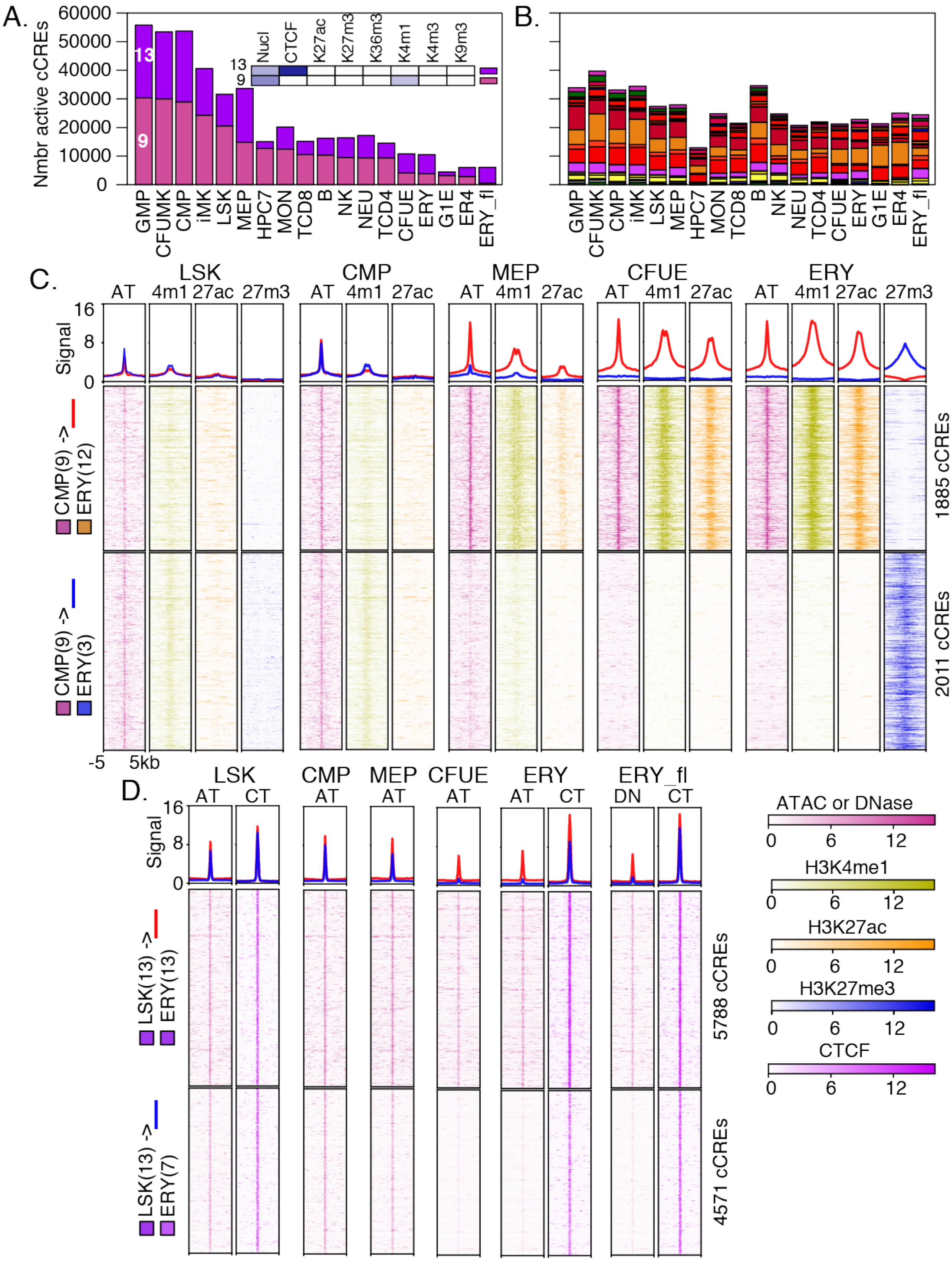

A.
Nmbr active cCREs
60000
50000
40000
30000
20000
10000
0
13
9
Nucl
CTCF
K27ac
K27m3
K36m3
K4m1
K4m3
K9m3
GMP
CFUMK
CMP
iMK
LSK
MEP
HPC7
MON
TCD8
B
NK
NEU
TCD4
CFUE
ERY
G1E
ER4
ERY_fl
B.
C.
LSK
CMP
MEP
CFUE
ERY
AT
4m1
27ac
27m3
Signal
16
8
0
CMP(9) ->
ERY(12)
CMP(9) ->
ERY(3)
1885 cCREs
2011 cCREs
-5
5kb
D.
LSK
CMP
MEP
CFUE
ERY
ERY_fl
AT
CT
DN
LSK(13) ->
ERY(13)
LSK(13) ->
ERY(7)
5788 cCREs
4571 cCREs
ATAC or DNase
H3K4me1
H3K27ac
H3K27me3
CTCF
0
6
12

Figure 7. Transitions in epigenetic states at cCREs across hematopoietic differentiation. **A and B.** The numbers of cCREs in each cell type are colored by their IDEAS epigenetic state, emphasizing decreases in numbers of cCREs in states 9 and 13 (**A),** while numbers in other states less variable (**B**). **C**. Aggregated and individual signal profiles for cCREs in the poised enhancer state 9 in CMPs as they transition from LSK through CMP and MEP to CFUE and ERY. Profiles for up to four relevant epigenetic features are presented. Data for H3K27me3 are not available for CMP, MEP, or CFUE. The first graph in each panel shows the aggregated signal for all cCREs in a class, and graphs beneath it are heatmaps representing signal intensity in individual cCREs. In the aggregated signal, red lines show signals for cCREs that transition from poised state 9 to active enhancer-like state 12, and blue lines show signals for cCREs that transition from poised state 9 to polycomb repressed state 3. **D**. Aggregated and individual signal profiles for CTCF-bound cCREs that either retain or lose nuclease accessibility during differentiation from LSK to ERY. In the aggregated signal, red lines show signals for cCREs that stay in the CTCF-bound, nuclease sensitive state 13, and blue lines show signals for CTCF-bound cCREs that lose nuclease sensitivity as they transition from state 13 to state 7. Signals were normalized by S3norm. Abbreviations are AT=ATAC, 4m1=H3K4me1, 27ac=H3K27ac, 27m3=H3K27me3, CT=CTCF.

Estimating regulatory output and assigning target genes to cCREs

Having established that the VISION collection of mouse hematopoietic cCREs overlaps well with other indicators of regulatory elements, we investigated the effectiveness of the cCREs in explaining levels of gene expression. We developed a modeling approach to evaluate how well the cCREs could account for levels of expression in the twelve cell types for which the RNA-seq measurements were determined in the same manner. This modeling approach had the additional benefit of making predictions of target genes for each cCRE.

We reasoned that the epigenetic state assignments for each cCRE DNA interval in each cell type could serve as a versatile proxy for cCRE regulatory activity, since the states were based on a systematic integration of multiple epigenetic signals. As explained in detail in the **Materials and Methods** (*Mapping cCREs to Genes*), we estimated cCRE effects on expression by treating the states as categorical variables and training a multivariate linear model of gene expression on the states. Each cCRE could be composed of multiple epigenetic states (**Figure 8A**), and we used the proportion of pooled cCREs covered by a state as the predictor variable for that state (**Figure 8B**). All cCREs within 1Mb of the TSS of a gene were initially included in the modeling, while allowing for separate effects of proximal (within 1 kb on either side of the transcription start site or TSS) and distal (within 1 Mb but beyond 1 kb of the TSS) cCREs for each gene. Not all cCREs within the 2 Mb region surrounding a gene's TSS were expected to influence expression. Thus, CREs predicted to have limited contribution to explaining expression were removed via a sub-selection strategy, maximizing leave-one-out prediction accuracy, during iterations of model fitting (**Figure 8B, Supplementary Figure 11B**).

The regression coefficients, beta, determined for the epigenetic states showed some expected trends (a full set of values is presented in **Supplementary Figure 11C**). For example, the coefficients for the set of differentially expressed genes were high for most promoter-like and enhancer-like states and low for most polycomb-repressed and heterochromatin states (**Figure 8B)**. The coefficients for some states differed between proximal and distal cCREs. The weighted sum of the state-specific regression coefficients can be considered an initial approximation of an epigenetic regulatory potential (eRP) score (**Figure 8B, Supplementary Figure 11D**), which measures the impact of each cCRE on expression relative to a quiescent state (whose impact is assumed to be 0). In this formulation, each cCRE has an eRP score for each potential target gene (distal or proximal, depending on distance from the potential target) in each cell type.

We evaluated the accuracy of predicting gene expression from the eRP scores using a leave-one-out strategy. Specifically, we trained a linear model on data from eleven of the twelve cell types, used that model to predict expression levels in the twelfth cell type, and then computed the adjusted r^2 for the accuracy of the predicted expression levels compared to the actual expression levels in the left-out cell type. This procedure was repeated leaving out each of the cell types in turn. Models were trained using only proximal cCREs, only distal cCREs, or a combination of proximal and distal cCREs. For expression of all genes, the prediction accuracy was around 50% for proximal cCREs only or distal cCREs only, and it improved to about 60% when proximal and distal cCREs were combined (**Figure 8C**, graph All genes).

Some portion of the explanatory power was expected to derive from the strong differences in epigenetic signals for expressed *versus* silent genes. In an effort to remove this effect from the predictions of accuracy, we repeated the linear regression modeling and evaluations on four categories of genes separately, specifically those with (1) consistently low, (2) differentially low, (3) differentially high, and (4) consistently high expression across cell types. The values of beta varied across the four categories and for distal versus proximal cCREs (**Supplementary Figure 11C**). Using the eRP scores for the appropriate gene category, the accuracy of predicting expression levels in the leave-one-out strategy showed a much smaller impact of the proximal cCREs (**Figure 8C**, graphs 1-4), suggesting that a major effect of the epigenetic states around the TSSs was to establish expression or silencing. In contrast, the distal cCREs did contribute to expression variation within the gene categories, especially for differentially expressed genes (**Figure 8C**, graphs 2 and 3). Overall, these evaluations indicate that proximal cCREs contributed strongly to the broad expression category (expressed or not, differential or constitutive), and distal cCREs contributed to the expression level of each gene within a category.

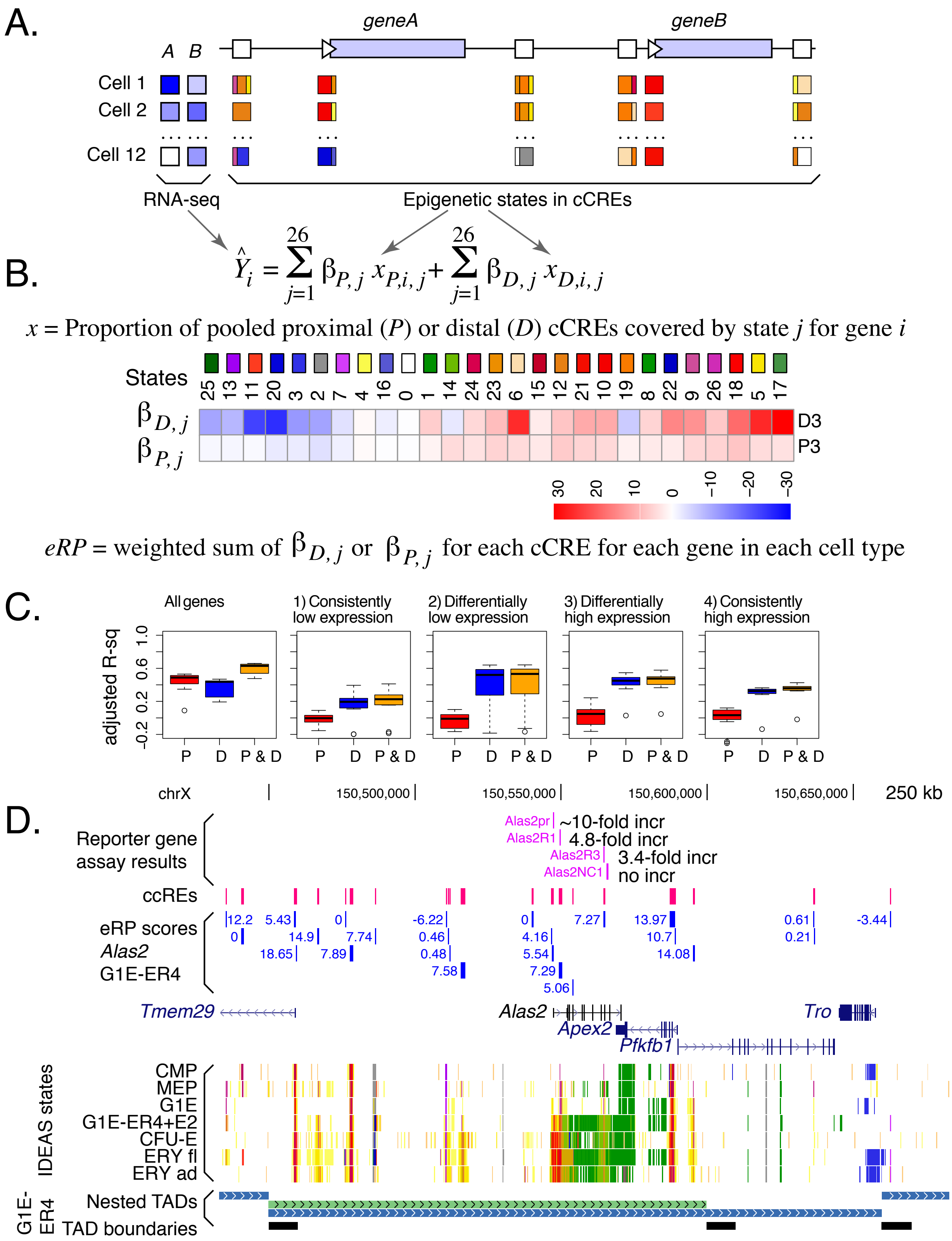

A.
geneA
geneB
A B
Cell 1
Cell 2
Cell 12
RNA-seq
Epigenetic states in cCREs
B.
$\hat{Y}_i = \sum_{j=1}^{26} \beta_{P,j}\, x_{P,i,j} + \sum_{j=1}^{26} \beta_{D,j}\, x_{D,i,j}$
x = Proportion of pooled proximal (P) or distal (D) cCREs covered by state j for gene i
States
25 13 11 20 3 2 7 4 16 0 1 14 24 23 6 15 12 21 10 19 8 22 9 26 18 5 17
$\beta_{D,j}$
$\beta_{P,j}$
D3
P3
30 20 10 0 −10 −20 −30
eRP = weighted sum of $\beta_{D,j}$ or $\beta_{P,j}$ for each cCRE for each gene in each cell type
C.
All genes
1) Consistently low expression
2) Differentially low expression
3) Differentially high expression
4) Consistently high expression
adjusted R-sq
−0.2 0.2 0.6 1.0
P D P & D
D.
chrX
150,500,000
150,550,000
150,600,000
150,650,000
250 kb
Reporter gene assay results
Alas2pr ~10-fold incr
Alas2R1 4.8-fold incr
Alas2R3 3.4-fold incr
Alas2NC1 no incr
ccREs
eRP scores Alas2 G1E-ER4
12.2 5.43 0 -6.22 0 7.27 13.97 0.61 -3.44
0 14.9 7.74 0.46 4.16 10.7 0.21
18.65 7.89 0.48 5.54 14.08
7.58 7.29
5.06
Tmem29
Alas2
Apex2
Pfkfb1
Tro
IDEAS states
CMP
MEP
G1E
G1E-ER4+E2
CFU-E
ERY fl
ERY ad
G1E-ER4
Nested TADs
TAD boundaries

Figure 8. Initial estimates of regulatory output and target gene prediction using regression models of IDEASs states in cCREs versus gene expression. **A.** Illustration of cCREs around two potential target genes, showing expression profiles of the genes across cell types (shades of blue, *left*) and cCREs with one or more epigenetic states assigned in each cell type. Note that cCREs that are proximal to one gene can be distal to another gene. **B.** Multivariate linear regression of proportion of pooled cCREs in each state against expression levels of potential target genes, keeping proximal and distal cCREs separate and learning the regression coefficients iteratively in a sub-selection strategy. Values of the regression coefficients beta for each epigenetic state for proximal and distal cCREs for differentially expressed genes. The values of the regression coefficients for each epigenetic state are presented as a blue to red heatmap, with the coefficients expressed relative to that for state 0 (quiescent). **C.** Ability of eRP scores of cCREs to explain levels of expression on chr1-chr19 and chrX in the twelve cell types for all genes and (1-4) in the four categories of genes. showing the effects. A leave-one-out strategy was employed to calculate the accuracy predicting expression. The distribution of adjusted r^2 values are shown as box-plots for proximal, distal, and combined cCREs. **D.** Illustration of eRP scores for cCREs in and around the *Alas2* gene, including a comparison with previously measured enhancer and promoter activities. Nested TADs called by OnTAD (An *et al.* 2019) are shown in the bottom tracks.

The positive predictive power of these initial estimates of eRP scores suggest that they have utility in assigning candidates for target genes for cCREs. The estimated eRP scores can serve as one indicator of the potential contribution of each cCRE to the regulation of a gene in its broad vicinity. Thus, a set of likely cCRE-target gene pairs can be obtained at any desired eRP threshold. We provide a large table of potential cCRE-target gene pairs at the VISION project website, along with a versatile filtering tool for finding cCREs potentially regulating a specified gene in a particular cell type. The filtering tool also allows further restriction of cCREs to those within the same topological associated domain or compartment as the candidate target gene.

The example from the *Alas2* locus (**Figure 8D**) illustrates how these eRP scores are consistent with results from previous experimental assays for CREs within the gene (Wang *et al.* 2006), and they raise the possibility of additional, distal cCREs regulating the gene. These data-driven, integrated resources should allow users to make informed decisions about important but challenging issues such as finding the set of cCREs likely to regulate a particular gene.

Discussion

One goal of the VISION project is to gather information from our laboratories, other laboratories, and consortia to conduct systematic integrative analysis to produce resources of high utility to investigators of genome biology, blood cell differentiation, and other processes. In this study, we have compiled and generated epigenomic and transcriptomic data on cell types across hematopoietic differentiation in mouse. Most of the data were from purified populations of primary blood cells, and we further included data from some cell lines that have been widely used in mechanistic studies. After uniform data processing and normalization, the data were systematically analyzed by the IDEAS method to assign genomic intervals to epigenetic states in the 20 cell types examined. Each state was defined by a quantitative spectrum of nuclease sensitivity, histone modifications, and CTCF occupancy. Most of these combinations of epigenetic features have been associated with specific regulatory elements or events, such as active promoters, poised enhancers, transcribed regions, or quiescent zones. As such, the assignments to epigenetic states provide a guide to potential functions of each genomic interval in each cell type. In effect, the IDEAS segmentation pipeline described here has reduced 150 dimensions (or tracks) of epigenomic data to 20 dimensions, i.e. the number of cell types. Thus, investigators now have a simplified way to view the large amount of data, e.g. in a genome browser, and they can operate computationally on the state assignments. We further simplified the epigenomic data by constructing an initial registry of 205,000 cCREs; these are discrete genomic intervals with features predictive of a potential regulatory role in at least one hematopoietic cell type. Our evaluation of these cCREs against other groups of candidate regulatory regions indicates that this initial set is valuable for identifying likely regulatory regions, especially in multilineage progenitors and the erythroid and myeloid cell lineages.

These resources should be valuable to many investigators, and we provide multiple ways to access and interact with the data via our VISION website (usevision.org). The raw and normalized data tracks can be downloaded for custom analysis. Examination of particular genes is enabled by the custom genome browser, which is built on the familiar framework of the UCSC Genome Browser (Haeussler *et al.* 2019). Tables of annotated cCREs and their associations with specific genes by regression are available for download, and cCREs for specific genes and genomic intervals can be obtained by queries at the website. Links are provided to additional resources such as CODEX for more extensive transcription factor occupancy and histone modification data (Sanchez-Castillo *et al.* 2015), the 3D Genome Browser for visualizing matrices of chromatin interaction frequencies (Wang *et al.* 2018), and the ENCODE registry of cCREs (The_ENCODE_Project_Consortium *et al.* 2019).

We chose IDEAS as the systematic integration method because its joint segmentation along chromosomes and across cell types retains position-specific information, thereby providing more precision to the state assignments (Zhang *et al.* 2016; Zhang and Hardison 2017). Furthermore, the IDEAS method does not require determination of all features in all cell types, and thus cell types with missing data were included. Even an extreme case of a cell type for which the only epigenomic dataset was ATAC-seq, CFUMK, was assigned a meaningful segmentation pattern. The local clustering of cell types by their epigenomic profiles are key to the approach to handling missing data. When data for a feature are missing from one cell type, IDEAS learns the signal distribution for that feature in locally related cell types, and it uses those signal distributions when assigning likely states to genomic intervals in cell types with missing data. A recent systematic study shows that IDEAS is able to produce reliable segmentation despite missing data and without resorting to imputation (Zhang and Mahony 2019). While full determination of all biochemical features in each cell type is preferred, attaining complete coverage across a wide range of cell types, especially for rare cell types, is difficult, and it may be impossible in some cases. Indeed, many integrative analysis projects are contending with the challenges of missing data (Ernst and Kellis 2015; Schreiber *et al.* 2019; The_ENCODE_Project_Consortium *et al.* 2019), and we expect that missing data will continue to be a challenge in the future. We suggest that the IDEAS method provides a principled approach with good utility for integrative analyses in the face of missing data. Future work will test experimentally the IDEAS predictions about maps of epigenetic features, such as CTCF occupancy or specific histone modifications, in cell types for which those data were missing in this study.

Our initial collection of cCREs in mouse blood cells appears to be robust, in that it captures virtually all known erythroid regulatory elements and it includes a majority of potential enhancers predicted by EP300 occupancy in fetal liver and cell line models for erythroid and B cells. The scope of the IDEAS analysis and cCRE predictions was greatly expanded by including the iChIP data (Lara-Astiaso *et al.* 2014) on histone modification profiles in the rare multilineage progenitor cells. Thus, despite the larger number of cell types on the erythroid and megakaryocytic lineages, the cCRE collection is not limited to those lineages. However, the initial cCRE registry is unlikely to be complete. Parallel efforts, such as the Immunological Genome Project (Yoshida *et al.* 2019), are generating complementary resources that can be incorporated to expand the cCRE collection. Only DNA intervals in nuclease accessible chromatin were assigned as cCREs, and thus, any regulatory elements that function in nuclease inaccessible regions will be missed. Such elements may be discovered by further studies on inaccessible regions that are bound by transcription factors. Given the absence of comprehensive reference sets of known regulatory elements, neither the completeness nor the specificity of the cCRE collections can be evaluated rigorously. Future work will evaluate experimentally the impact of cCREs on gene expression via genome editing, such as direct targeting of mutagenesis to cCREs or saturation mutagenesis of loci.

Each cCRE has been annotated with its presence or absence in each cell type, as well as its epigenetic state. Furthermore, an initial estimate of the epigenomic regulatory potential (eRP) from each cCRE for regulating candidate target genes was derived from a multivariate regression and sub-selection procedure. These results can be used to identify potential cCREs for any gene in the investigated cell types. These initial eRP scores were derived from an effort to leverage the correlations between epigenetic states and gene expression to find potential target genes. The eRP scores for combined distal and proximal cCREs can explain a substantial portion of variance in gene expression, but a considerable amount of expression variance remains unexplained. Future work should improve eRP estimates by bringing in additional information such as enrichment for transcription factor binding site motifs (Weirauch *et al.* 2014), transcription factor occupancy (Dogan *et al.* 2015), and patterns in multi-species genome sequence alignments (Taylor *et al.* 2006). Nevertheless, these initial results have provided a much smaller number of potential cCREs for genes compared to all cCREs within the broad vicinity of a TSS, and the eRP scores track well with previous experimental results on regulatory elements around hematopoietic genes. The target gene assignments can be refined by inclusion of data on chromatin interaction frequencies, e.g. simply by restricting cCRE-gene pairs to those within a topologically associated domain, or TAD (Oudelaar *et al.* 2017). The VISION project has analyzed Hi-C data in G1E-ER4 cells (Hsu *et al.* 2017) and HPC7 cells (Wilson *et al.* 2016) to provide coordinates of TADs (An *et al.* 2019) and compartments (Zheng and Zheng 2018); and our query interface allows users to use this information to refine choices of cCREs for specific genes. Further improvements can leverage the higher resolution interactions obtained with capture approaches (Hughes *et al.* 2014).

The IDEAS segmentation results across cell types revealed many expected states and transitions between states, such as poised enhancers in multilineage progenitor cells either shifting to active enhancers or losing their pre-activation signatures to become repressed or quiescent in more differentiated cells. However, one of the most common transitions has not been described previously (to our knowledge). Of the CTCF-bound sites in LSK that were also accessible to nuclease, a substantial proportion became much less nuclease accessible while retaining CTCF occupancy in differentiated cells. This loss of accessibility was observed for both ATAC-seq and DNase-seq data. The reduction in accessibility reflects a change in the chromatin structure to a more closed state, but unexpectedly, the CTCF protein remains bound. Further studies can examine whether the CTCF-bound-but-inaccessible sites are enriched at structurally or functionally important regions, such as boundaries of topologically associated domains, the base of chromatin loops, mitotically stable or unstable sites, or promoters of active or inactive genes.

We found a substantially larger number of cCREs in hematopoietic progenitor cells than in mature cells, with the notable exception of megakaryocytic cells. The reduction in numbers of cCREs coincides with the decrease in the size of the nucleus during differentiation and maturation after commitment to a single lineage (Baron and Barminko 2016), which is indicative of a decrease in transcriptional activity. Indeed, the decrease in number of cCREs correlates with a decrease in the number of genes being expressed. This reduction in active genes and numbers of active regulatory elements appears to be a common feature of lineage-committed, maturing cells in most lineages. Megakaryocytic cells present a strikingly different pattern, as they retain many aspects of the regulatory landscape and transcriptomes of multilineage progenitor cells. This similarity of MK to multilineage progenitor cells is a robust result, having been discerned previously based on phenotypic similarities (Huang and Cantor 2009), transcriptome data (Sanjuan-Pla *et al.* 2013; Psaila *et al.* 2016), global epigenetic profiles (Heuston *et al.* 2018), as well as from our integrative segmentations, cCRE predictions, and transcriptomes. These recent studies have indicated that MK cells can be derived from multiple stages of progenitor cells, including HSC, CMP, and MEP (Sanjuan-Pla *et al.* 2013; Psaila *et al.* 2016). It is intriguing to speculate that the similarity of MK to multilineage progenitor cells may indicate that multiple stages of progenitor cells could differentiate into MK without substantial changes to the regulatory landscape. Such a flexible process differs remarkably from other lineage commitment and maturation processes that involved substantial changes to the epigenome and reduction in numbers of genes expressed.

A large majority of the genome in each cell type was assigned to the quiescent, low signal state. A low-signal state covering most of the genome was observed in previous studies (Wu *et al.* 2011; Ernst and Kellis 2012; Hoffman *et al.* 2013; Yue *et al.* 2014; Roadmap Epigenomics *et al.* 2015; Zhang and Hardison 2017), but the interpretation has ranged from this representing artifacts due to high repeats and low mappability (Ernst and Kellis 2012) to a true, dramatic under-representation of dynamic histone modifications, CTCF, and open chromatin in most of the genome. We favor the latter interpretation, and suggest that much of the quiescent chromatin is repressed, but in a state not subject to histone modifications that are revealed by conventional ChIP-seq. Nevertheless, the fraction of the genome in a quiescent state may be overestimated if current assays are not fully recording some modifications in chromatin (Becker *et al.* 2017). For example, the H3K9me3 modification in highly compacted heterochromatin may be less accessible to the antibodies during chromatin immunoprecipitation, or heterochromatin may not be sheared adequately to solubilize the compacted chromatin to produce DNA fragments that are sequenced efficiently. Even if current methods preclude the identification of some modified chromatin because of such issues, it is still the case that the DNA in the quiescent is distinctly different from that in other states.

The systematic integration of 150 tracks of epigenetic data on mouse hematopoietic cells has produced an easily interpretable representation of the regulatory landscapes across these cell types. Further refinement led to calls of discrete candidate regulatory elements, which in turn were annotated with both their epigenetic states in each cell type and an estimate of their regulatory output toward each potential target gene. While we expect to improve these resources in future work, our initial applications of the resources suggest they will have good utility for a broad user community. Similar systematic integration of epigenetic data in human blood cells is on-going, which will generate equivalent resources. Catalogs of cCREs, annotated by cell type activity and indicators of potential target genes, should provide guidance on many important problems, such as suggesting specific hypotheses for mechanisms by which genetic variants in non-coding regions can be associated with complex traits and diseases (Ulirsch *et al.* 2016; Bao *et al.* 2019).

Materials and Methods

Cell Isolation

All primary hematopoietic cell populations were enriched from 5-8 week old C57BL6 male mice. LSK, CMP, MEP, GMP, CFUE, ERY, CFU-MK, and iMK populations were harvested and isolated from bone marrow (BM) as described (Heuston *et al.* 2018). Neutrophils (NEU) and monocytes (MON) were isolated from peripheral blood by as described (Heuston *et al.* 2018). Isolation of other cell populations was described in Lara-Astiaso *et al.* (2014).

Sources of epigenomic and transcriptomic data

The datasets were collected from many different sources, including individual laboratories and consortia (**Figure 1B**). Much of the data were published or released previously. We started with work from our own laboratories in both the cell lines and primary erythroid and megakaryocytic cells at various stages (Cheng *et al.* 2009; Wilson *et al.* 2010; Wu *et al.* 2011; Cheng *et al.* 2014; Pimkin *et al.* 2014; Wu *et al.* 2014; Yue *et al.* 2014; Hsiung *et al.* 2015; Jain *et al.* 2015; Stonestrom *et al.* 2015; Wilson *et al.* 2016; Heuston *et al.* 2018). The numbers of hematopoietic stem and progenitor cells that could be isolated by FACS on selected surface markers are small relative to those for maturing, lineage-committed cells, which presents a limitation for ChIP-seq analyses. Thus, for histone modifications in these multilineage stem and progenitors cell populations, we used the ChIP-seq data obtained using the iChIP method for interrogating small numbers of cells (Lara-Astiaso *et al.* 2014). The iChIP data also were the primary source for epigenomic information on mouse lymphoid cells. Additional datasets obtained through the CODEX compendium (Sanchez-Castillo *et al.* 2015), the GEO database (Barrett *et al.* 2009) and the ENCODE data portal (Sloan *et al.* 2016; Davis *et al.* 2018) filled in more features for several cell types. A ChIP-seq determination on CTCF in LSK cells is a new dataset for this paper. Almost all the experiments from the VISION project and ENCODE projects were done in replicates, but experiments without replicates from any source were included if they passed quality checks.

Several types of RNA-seq data were compiled across these cell types, including those using various strategies for RNA-seq on polyA+ RNA (Lara-Astiaso *et al.* 2014; Paralkar *et al.* 2014) and on total RNA (Heuston *et al.* 2018). Comparisons between the RNA-seq collections determined by different methods or using different sources of RNA (total vs polyA+) were problematic; the differences attributable to method of preparation or RNA source exceeded differences between cell populations. Thus, the set of transcriptomes determined in replicate on total RNA from twelve cell types and populations using the same procedure in the same laboratories (Heuston *et al.* 2018) were the primary source of expression level data used in evaluating our integration products and target gene assignments.

ChIP-seq

For CTCF in LSK cells, approximately 5M cells were fixed in 0.4% formaldehyde (16% methanol-free, Thermo Scientific) for 15 minutes before quenching in 125 mM glycine for 5 minutes. Cells were washed in 2X PIC (Roche mini-tabs, 1 tab in 5 ml = 2X) and stored at -80°C. Cells were then lysed (10 mM Tris-HCl, pH 8.0, 10 nM NaCl, 0.2% NP40) for 10 min on ice, washed once in 1x PBS, followed by nuclear lysis (50 mM Tris-HCl 8.0, 10 mM EDTA, 1% SDS) for 10 min on ice. Chromatin was then diluted further with Immunoprecipitation Buffer (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, 0.01% SDS) and a 1x Protease Inhibitor Cocktail set V, EDTA-free (Calibiochem, La Jolla, CA). Samples were sheared for 10-13 cycles of 30 sec on, 30 sec off sonication at medium output power at 4^0C. Sonicated chromatin was pre-cleared overnight at 4°C with 10 µg of normal rabbit IgG (Santa Cruz Biotechnology, Santa Cruz, CA; sc-2027). 5 ul of CTCF serum (Millipore Sigma, Cat# 07-729) was also pre-bound to protein G agarose beads overnight at 4°C. For binding, pre-cleared chromatin was added to the antibody:bead complex and incubated with rotation at 4°C for 4 hours. After binding, the beads were washed with Wash Buffer I (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 50 mM NaCl, 1% Triton X-100, 0.1% SDS), High Salt Wash Buffer (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, 0.1% SDS), Wash Buffer II (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 250 mM LiCl, 1% NP40, 1% deoxycholate), and 1x TE. DNA:protein complexes were then eluted with Elution Buffer (1% SDS, 100 mM NaHCO3). Reverse crosslinking of immunoprecipitated chromatin was accomplished by the addition of NaCl to ChIP and input samples, followed by incubation overnight at 65 °C with 1µg RNase A. To remove proteins, each sample was treated with 6 µg Proteinase K for 2 hours at 45 °C. Immunoprecipitated DNA was purified using the Qiagen PCR Purification Kit. All samples were processed for library construction using Illumina's TruSeq ChIP Sample Preparation Kit according to manufacturer's instructions. The DNA libraries were sequenced on NextSeq 500 using Illumina's kits and reagents as appropriate.

ChIP-seq data processing

The sequencing reads from both the new ChIP-seq experiments and many previously published ChIP-seq were processed through the VISION pipeline. Reads were mapped to mouse genome assembly mm10, using a pipeline that contains essential elements of the ENCODE mapping pipeline but adjusted to allow for multiple mapping reads, which allows interrogation of duplicated genes and repetitive elements. Specifically, the mapping pipeline consisted of Bowtie 0.12.8 (Langmead *et al.* 2009) for mapping, then filtering to remove chrM, unplaced chromosomes, and unmapped reads. The alignment is converted to bam format using Samtools 0.1.8 (Li *et al.* 2009). MACS 1.3.7.1 (Zhang *et al.* 2008) is used to generate the wiggle tracks and call peaks. The wigToBigWig program from the UCSC Genome Browser (Haeussler *et al.* 2019) is used to convert the wiggle file to a bigWig.

ATAC-seq from VISION project

ATAC-seq (Buenrostro *et al.* 2013) from LSK, CMP, MEP, GMP, CFUE, ERY, CFU-MK, iMK, NEU, and MON populations were generated as described (Heuston *et al.* 2018). The sequence reads were processed using a pipeline consisting of trimming the reads to 30 base pairs and then mapped using bowtie 0.12.8. The alignment is filtered to remove chrM, unplaced chromosomes, and unmapped reads and is converted to bam format using Samtools 0.1.8. Bedtools 2.16.2 (Quinlan and Hall 2010) is used to convert the alignments to bed format for F-Seq 1.85 (Boyle *et al.* 2008), which generates the wiggles. Local peaks of high frequency cleavage were determined using HOMER 1.0 (Heinz *et al.* 2010), as used previously (Ramirez *et al.* 2017).

RNA-seq from VISION project

RNA-seq from LSK, CMP, MEP, GMP, CFUE, ERY, CFU-MK, iMK, NEU, and MON populations were generated as described (Heuston *et al.* 2018). The sequence reads were processed using the ENCODE3 long RNA-seq pipeline (https://www.encodeproject.org/pipelines/ENCPL002LPE/). In brief, reads were mapped to the mouse genome (mm10 assembly) using STAR 2.5.1b_modified (Dobin *et al.* 2013), followed by RSEM-1.2.28 (Li and Dewey 2011) for gene quantifications. UCSC's bedGraphToBigWig is used to convert the bedgraph files to bigwigs for display in the browser.

Replication and quality evaluation

Experiments arising from laboratories in the VISION consortium and ENCODE were conducted on biological replicates (e.g. either cells isolated from different groups of mice on different days, or from the same group of mice, but collected in different aliquots from the sorter and processed separately). Experiments from Lara-Astiaso *et al.* (2014) were determined once. The read coverage for all experiments exceeded the recommended level (Landt *et al.* 2012), providing over 4.2 billion mapped reads (1,952,990,660 for RNA-seq, 1,897,113,048 for ATAC-seq, and 353,030,332 for ChIP-seq) supporting the results of the experiments. The data were high quality, as evaluated by metrics currently recommended by the ENCODE Project Consortium.

All ATAC-seq datasets had a FRiP score of 0.27 (27%) or greater, and all ChIP-seq datasets had a FRiP score of 0.03 (3%) or greater, consistent with the currently accepted ENCODE standards for ATAC-seq, as described at https://www.encodeproject.org/data-standards/. All DNase-seq datasets had a FRiP score of 0.21 (21%) or greater.

The replicates within RNA-seq experiments were highly correlated, with Spearman correlation coefficients equal to or greater than 0.93 for almost all experiments. The exceptions were RNA-seq for CFUE and ERY, for which the replicate correlation was 0.89. This slightly lower correlation values may result from the fact that RNA isolated from these cell types exhibited consistently lower RIN scores, which may reflect the presence of older, degraded transcripts due to the nuclear condensation process during erythroid maturation.

Comparisons of epigenetic and transcriptional profiles across hematopoietic cell types

The signal strengths of ATAC-seq and DNase-seq peaks among eighteen cell types were compared pairwise by computing the Spearman correlation coefficient (*r*) across the comprehensive set of ATAC-seq peaks for each pair of cell types, and using 1- *r* as the distance measure. Hierarchical clustering of the pairwise comparisons was performed using heatmap.2 in R.

Using the RNA-seq data on total RNA from twelve cell types (Heuston *et al.* 2018), we estimated transcript levels for each gene annotated by Gencode (M4) (Harrow *et al.* 2012), including both protein-coding and non-coding genes, using the program RSEM (Li and Dewey 2011). We compared the global transcriptomes across the cell types, again using 1 – *r* as the distance measure, and performed hierarchical clustering.

The pairwise analyses reduced each comparison between cell types to a single value (r) summarizing the relationships among the ATAC-seq or RNA-seq signals. To capture the genome-wide information more completely, we also analyzed the ATAC-seq signal matrix and RNA-seq transcript levels across replicates and cell types by principal component analysis (PCA), using the tool prcomp in R.

Normalization of ChIP-seq and nuclease sensitivity data

The datasets for ChIP-seq, ATAC-seq, and DNase-seq came from heterogeneous sources with considerable differences in sequencing depth and signal-to-noise ratio. We developed a new method, called S3norm, to simultaneously adjust for both sequencing depth and signal-to-noise ratio (Xiang et al. 2019). In contrast to other normalization methods, which are designed to rescale signals mainly by focusing on either background regions or peak regions (Liang and Keles 2012; Meyer *et al.* 2012; Shao *et al.* 2012; Dillies *et al.* 2013), S3norm simultaneously matches the mean signals of both background and peak regions across multiple datasets). Maintaining a low background signal is particularly important genome segmentations [ref], since an inflation of the background could lead to assigning the background regions with increased noise to low signal-containing states. We also explored other normalization methods, such as the widely used quantile normalization (Bolstad *et al.* 2003), which forces the datasets to have an identical distribution across datasets. Unlike quantile normalization, the S3norm method retains the variation in the overall signal distribution across datasets. Thus, when normalizing datasets with a different number of peaks, for example, S3norm is less likely than quantile normalization to create false positive peaks in the dataset with fewer peaks.

In brief, the S3norm method converts raw signal (number of mapped reads per genomic interval) to p-values, selects reference datasets to use as standards, and normalizes via a nonlinear transform to adjust background and foreground simultaneously. This latter adjustment was effectively accomplished by rotation of a regression line through a scatter plot of signal strengths for each window in the dataset being normalized and the proxy reference, such that the mean signals for common peaks were the same between normalized and reference datasets, and the mean signals for common backgrounds were also the same for the two datasets (Xiang et al. 2019). To maintain a balance between replicated data for some cell types and non-replicated data for others, replicate data were merged after conversion to p-values so that only one dataset was used for each feature in each cell type. The S3norm tool is available from Github at the link https://github.com/guanjue/S3norm .

IDEAS segmentation

IDEAS utilizes a Bayesian nonparametric hierarchical latent-class mixed-effect model to achieve segmentations simultaneously along chromosomes and across cell types (Zhang *et al.* 2016; Zhang and Hardison 2017). The computational approach has a linear time solution with respect to the number of cell types, which allows it to scale to hundreds of cell types simultaneously. For the segmentation runs described in this paper, signals in terms of numbers of mapped reads per 200 bp bin for 8 epigenetic features (histone modification and CTCF ChIP-seq, ATAC-seq or DNase-seq) were compiled from 20 cell types to produce a set of 150 tracks of data, including replicates. The replicates of the same epigenetic feature were merged (via Fishers' method that emphasizes the replicate with the better signal-to-noise level), and signals were normalized using the S3norm procedure (Xiang et al. 2019) to generate 104 datasets. The normalized datasets were used as input for IDEAS to generate chromatin segmentation. The current version of the IDEAS method includes a preliminary, simple assessment of the most common combinations of epigenetic features to initialize the model building. We first binarized the signals of each feature by peak calling at FDR 0.05 using a negative-Binomial distribution as the null, where the parameters of the negative-Binomial distribution was estimated from the bottom 99% of the signals for the feature. From the combinations of 0s and 1s of multiple features at each position, we identified the distinct combinations that correspond to a preliminary set of epigenetic states. For *k* features, there are 2^*k* distinct combinations of 0s and 1s. We removed the rare combinations with <0.1% occurrence, and we used the remaining set of preliminary epigenetic states as the initial states for IDEAS. The removed rare states were replaced by a random sample of the common states. We also applied a relatively high threshold for inclusion of signals into the IDEAS modeling to produce a simpler, more interpretable model. Lowering the threshold generated many more states that were small variations on the states described here. Using the higher threshold did reduce the coverage of the genome by non-quiescent state assignments (from 19% to 14% on average). While higher coverage could be desired for some applications, for the current study we felt that the decreased coverage was off-set by the improved ability to interpret the model. The current software is available from Github, at the link https://github.com/guanjue/IDEAS_2018.

When dealing with datasets that range in quality, the IDEAS segmentation will sometimes “discover” states containing almost all features, including ones associated with opposite functions. Unlike the situation for most states, the DNA intervals assigned to such states also have high variation in the signal for each feature, indicating that the state may be further split to substates. We refer those states as ‘heterogeneous states’. When such heterogeneous states were returned, we revised our IDEAS pipeline as follows. (1) We identified the most common patterns of epigenetic peaks as introduced above from the cell types with all marks available, and we calculated the mean signal in each pattern as the initial parameters to train the IDEAS model. (2) After the first round of IDEAS run, we identified and removed the potential heterogeneous state and used the remaining states as priors to retrain IDEAS for a second round. The final segmentation is given by the second round of IDEAS run.

cCREs

Peaks called by Homer (Heinz *et al.* 2010) in the DNase-seq and ATAC-seq datasets were filtered to remove mitochondrial reads (which map to chrM) and blacklist regions. For datasets that have replicates, only peaks called in both replicates were retained. The remaining peaks from all cell types were combined, and peaks overlapping by at least one nucleotide were merged. This merger caused only a modest increase the size of the peak intervals; the median sizes were 150 bp before and 263 bp after merging. The ATAC-seq signal in the DNA interval corresponding to each ATAC-seq peak in this comprehensive set was determined by aggregating the ATAC-seq reads mapping to that interval in each cell type. Of the 215,120 merged ATAC-seq peaks, 207,690 were not in a quiescent state (state 0) from the IDEAS segmentation. Specifically, if more than 50% of a peak interval was in the quiescent state, it was not included as a candidate regulatory element. These non-quiescent, reproducible (if replicates were available), merged ATAC-seq peaks constitute the set of candidate Cis-Regulatory Elements (cCREs).

Other cCRE datasets were obtained from the Blood Enhancer Catalog (Lara-Astiaso *et al.* 2014), and the SCREEN website for ENCODE cCREs (The_ENCODE_Project_Consortium *et al.* 2019). The SCREEN cCREs were downloaded in July of 2018, obtaining one set for all mouse cCREs and another set restricted to those with DNase-seq data for "C57BL/6 liver male embryo (14.5 days)", which are referred to as the SCREEN fetal liver cCREs.

Mapping cCREs to Genes

We developed a novel method to use gene expression in 12 cell types to score the cCREs for their regulatory potentials based on epigenetic states, map the cCREs to candidate genes, and further select the most likely subset of cCREs for predicting gene expression. All genes, both expressed and silent, were included so that all of the 27 IDEAS states were covered.

1. Initial calculation of eRP scores to identify an inclusion threshold for cCREs

Let $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{26})$ denote the proportions (between [0,1]) of each IDEAS state within the 200bp regions that cover the TSSs of genes. Each $\boldsymbol{x}_i$ corresponds to the ith IDEAS state, excluding state 0 (the quiescent state). Let $\boldsymbol{Y}$ denote the log(y+0.001) transformed tag-per-million (TMP) value of RNA-seq data. We first used the regression model $\boldsymbol{Y}=\alpha+\beta\boldsymbol{X}+\varepsilon$ to obtain the coefficient β for each state, which represents the relative impact of the state in TSSs on expression. Given the β coefficient for each epigenetic state (where state 0 has coefficient 0), we assign an initial eRP$_0$ score to each cCRE in the genome in each cell as the weighted sum of β coefficients, and the weight of each state is the proportion of that state occurred within the cCRE. For every genome location that has at least one cCRE in some cell types, we further assigned the eRP$_0$ scores to the DNA segments at the same location in all other cell types. This yields a 12-dimensional vector of eRP$_0$ scores for each location with at least one cCRE.

2. Pre-selection of cCRE-gene pairs

We next used the 12-dimensional vector of eRP$_0$ scores at each cCRE location to calculate its correlations with gene expression of all genes within 1Mb. Our hypothesis is that the TSS-derived eRP$_0$ scores of a cCRE must be correlated with the expression of its target gene as if the cCRE is brought to the promoter region through chromatin looping. As such, our initial assignment of target genes for each cCRE is the genes whose correlation with the cCRE eRP$_0$ is >0.2. This threshold was suggested by a power curve for predicting expression, which showed increased adjusted r^2 values above this threshold.

3. Refinement of cCRE-gene pairs

Because we have only 12 cell types for correlation analysis, our initial application of a filter based on marginal correlation has limited power to accurately predict target genes, i.e., we do not expect all the cCREs passing that filter to be equally predictive of gene expression. We therefore developed a novel selection procedure to identify a subset of cCRE-gene pairs that are most likely capturing the regulatory relationships between cCREs and genes. Our principles are that the true cCRE-gene pairs should better predict gene expression, and that each cCRE impacts expression through their epigenetic states, thus cCREs with the same state composition should have the same impact on target gene expression. As such, the number of parameters in our model will only depend on the number of epigenetic states but not the number of cCREs assigned to each gene.

We assume that the impacts of distal cCREs (1kb away from TSS) may be different from the impacts of TSS regions (1kb on each side of TSS) on expression. Let $\boldsymbol{X^P}=(\boldsymbol{x^P}_1,\ldots,\boldsymbol{x^P}_{26})$ denote the state proportions observed at the TSS regions, and $\boldsymbol{X^D}=(\boldsymbol{x^D}_1,\ldots,\boldsymbol{x^D}_{26})$ denote the state proportions observed at the distal cCRE regions assigned to each gene. Initially, $\boldsymbol{X^D}$ is calculated by pooling all cCREs assigned to each gene together (based on the marginal correlation threshold), and our task is to remove some of the assigned cCREs and recalculate $\boldsymbol{X^D}$ so to maximize predictive power. Our model is in a regression form

$$\boldsymbol{Y}=\alpha+\beta^P\boldsymbol{X^P}+\beta^D\boldsymbol{X^D}+\varepsilon$$

where $\boldsymbol{Y}$ is the observed gene expression, β^{P} and β^{D} are unknown coefficients to be estimated from the model for the effects of proximal and distal elements, respectively, and ε is a Gaussian error term with mean 0. The intercept term α is set to be 0, so the predicted value of $\boldsymbol{Y}$ depends only on the proportions of states of the cCREs. Note, however, that this is not a standard regression model: though $\boldsymbol{X^P}$ is fixed, we will be updating $\boldsymbol{X^D}$ by adding or removing distal cCREs to each gene in order to maximize predictive power. Nevertheless, given $\boldsymbol{X^D}$, the coefficients β^{P} and β^{D} will be estimated by least squares.

4. Selecting distal cCREs for each gene

After initial assignment of cCRE-gene pairs based on marginal correlation, there were on average 136 cCREs (passing the cor>0.2 filter) assigned to each gene, whereas there were 216 cCREs per gene without the cor>0.2 filter. We will inevitably over fit our model by adding or removing cCREs from $\boldsymbol{X^D}$ if our objective is to maximize the model fitting. Instead, we used cross-validation accuracy as our criterion to select cCREs. Specifically, we only used data in (K-

1) cell types to train the model but we evaluated the model performance, and thus whether a cCRE should be included or excluded from a gene, by the held-out cell type. At each iteration, for each gene, we calculated the change of prediction r^2 of expression in the held-out cell type by adding or removing each cCRE to or from the current list for the gene. We then add or remove each cCRE to and from the list in the direction of increasing prediction r^2. This is repeated for all genes and the iteration is repeated for 100 times. To speed up the calculation, we updated the assignment of multiple cCREs simultaneously based on the previous assignment configuration. Each change in cCRE assignment may thus be suboptimal, as the assignment of other cCREs may have also changed. We however gradually reduced the number of simultaneous cCRE changes to 1 per gene as the iteration increases.

After fitting the model, the regression coefficients β^P and β^D can be used to recalculate eRP scores for cCREs by weighted sums. These final eRP scores would better reflect their regulatory potentials, as the coefficients are estimated from the selected cCRE-gene pairs that better predicted gene expression. In addition, we call the eRPs calculated from β^P as proximal eRPs, and the eRPs calculated from β^D as distal eRPs.

5. Out-sample evaluation of prediction accuracy

To evaluate if the above method can indeed improve prediction of gene expression by sub-selecting cCRE-gene pairs, we ran the method in 11 out of 12 cell types, and then used the model to predict the gene expression in the 12th cell type (not to be confused with the held-out cell type, which is one of the 11 cell types used in cross-validation). We repeated this by leaving each cell type out once, predicting its expression from the model trained in the other cell types, and calculating an overall $r^2$of the predicted expression for all 12 cell types against the observed expression.

6. Classification of genes

To evaluate if the eRP scores differ by genes, we classified the genes into four categories: 1) consistently lowly expressed genes (mean <= -4, standard deviation or sd <= 2); 2) differentially lowly expressed (mean <= -4, sd > 2); 3) differentially highly expressed (mean > -4, sd > 2); and 4) consistently highly expressed (mean > -4, sd <= 2). We used in-sample expression data (from 11 cell types) to classify the genes into four groups first, and then we ran our method within each gene group separately.

Data Access

The full lists of experiments and datasets are presented in the **Supplementary Tables**, along with information about replication structure of each experiment, read counts, quality metrics, literature citations, GEO and ENCODE dataset identification numbers. These lists are currently online at this link:

https://docs.google.com/spreadsheets/d/1q7wwrTfHQlEWCq301yaF-YZk3cQMb0cRksUmgluB9kQ/edit?usp=sharing

The **Supplementary Tables** also provide ids for datasets available from the ENCODE data portal and/or the Gene Expression Omnibus, to facilitate access to sequencing reads.

Results of data processing such as the signal tracks (before and after normalization) and peaks for ATAC-seq and ChIP-seq data and the estimates of transcript levels from RNA-seq (Li and Dewey 2011) are available at the VISION Project website (http://usevision.org). Signal tracks, peaks, and ranges of transcript levels can be visualized at the customized genome browser at the VISION Project website.

Acknowledgments

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (grant number R24DK106766-01A1), the National Human Genome Research Institute (NHGRI) intramural funds, NHGRI U54HG006998.

Disclosure Declarations

The authors have no conflicts of interest to declare.

References

Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A et al. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* **30**: 224-226.

An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, Li Q, Zhang Y. 2019. Hierarchical Domain Structure Reveals the Divergence of Activity among TADs and Boundaries. *Genome Biology; bioRxiv* doi:https://doi.org/10.1101/361147: under review.

An X, Schulz VP, Li J, Wu K, Liu J, Xue F, Hu J, Mohandas N, Gallagher PG. 2014. Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood* **123**: 3466-3477.

Bao EL, Cheng AN, Sankaran VG. 2019. The genetics of human hematopoiesis and its disruption in disease. *EMBO molecular medicine* **11**: e10316.

Baron MH, Barminko J. 2016. Chromatin Condensation and Enucleation in Red Blood Cells: An Open Question. *Dev Cell* **36**: 481-482.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**: D885-890.

Becker JS, McCarthy RL, Sidoli S, Donahue G, Kaeding KE, He Z, Lin S, Garcia BA, Zaret KS. 2017. Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol Cell* **68**: 1023-1037 e1015.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315-326.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185-193.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537-2538.

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

Cantor A, Orkin S. 2002. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**: 3368-3376.

Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515**: 371-375.

Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D et al. 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19**: 2172-2184.

Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ et al. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**: 1193-1203.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**: D794-D801.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14**: 671-683.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, Keller CA, Cheng Y, Jain D, Visel A et al. 2015. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**: 16.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215-216.

Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**: 364-376.

Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ, Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**: 667-681.

Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**: 587-594.

Greenside P, Shimko T, Fordyce P, Kundaje A. 2018. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* **34**: i629-i637.

Gregory T, Yu C, Ma A, Orkin SH, Blobel GA, Weiss MJ. 1999. GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-xL expression. *Blood* **94**: 87-96.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**: D853-D858.

Hahn MA, Wu X, Li AX, Hahn T, Pfeifer GP. 2011. Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. *PLoS One* **6**: e18844.

Hardison RC, Taylor J. 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**: 469-483.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

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

Heuston EF, Keller CA, Lichtenberg J, Giardine B, Anderson SM, Center NIHIS, Hardison RC, Bodine DM. 2018. Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. *Epigenetics Chromatin* **11**: 22.

Higgs DR. 2013. The molecular basis of alpha-thalassemia. *Cold Spring Harbor perspectives in medicine* **3**: a011718.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473-476.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827-841.

Hsiung CC, Morrissey CS, Udugama M, Frank CL, Keller CA, Baek S, Giardine B, Crawford GE, Sung MH, Hardison RC et al. 2015. Genome accessibility is widely preserved and locally modulated during mitosis. *Genome Res* **25**: 213-225.

Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries. *Mol Cell* **66**: 102-116 e107.

Huang H, Cantor AB. 2009. Common features of megakaryocytes and hematopoietic stem cells: what's the connection? *J Cell Biochem* **107**: 857-864.

Huang J, Liu X, Li D, Shao Z, Cao H, Zhang Y, Trompouki E, Bowman TV, Zon LI, Yuan GC et al. 2016. Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev Cell* **36**: 9-23.

Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR. 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**: 205-212.

Jain D, Mishra T, Giardine BM, Keller CA, Morrissey CS, Magargee S, Dorman CM, Long M, Weiss MJ, Hardison RC. 2015. Dynamics of GATA1 binding and expression response in a GATA1-induced erythroid differentiation system. *Genom Data* **4**: 1-7.

Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that discriminate between overlapping regulatory annotations. *PLoS Comput Biol* **13**: e1005795.

Kondo M, Wagers AJ, Manz MG, Prohaska SS, Scherer DC, Beilhack GF, Shizuru JA, Weissman IL. 2003. Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annu Rev Immunol* **21**: 759-806.

Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D et al. 2012. Intragenic enhancers act as alternative promoters. *Mol Cell* **45**: 447-458.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813-1831.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S et al. 2014. Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**: 943-949.

Laurenti E, Gottgens B. 2018. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**: 418-426.

Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* **152**: 1237-1251.

Lee YS, Wong AK, Tadych A, Hartmann BM, Park CY, DeJesus VA, Ramos I, Zaslavsky E, Sealfon SC, Troyanskaya OG. 2018. Interpretation of an individual functional genomics experiment guided by massive public data. *Nat Methods* **15**: 1049-1052.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**: 1983-1992.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liang K, Keles S. 2012. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**: 199.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170-1187.

Ludwig LS, Lareau CA, Bao EL, Nandakumar SK, Muus C, Ulirsch JC, Chowdhary K, Buenrostro JD, Mohandas N, An X et al. 2019. Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. *Cell reports* **27**: 3228-3240 e3227.

Meyer SU, Kaiser S, Wagner C, Thirion C, Pfaffl MW. 2012. Profound effect of profiling platform and normalization strategy on detection of differentially expressed microRNAs--a comparative study. *PLoS One* **7**: e38946.

Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ et al. 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol* **15**: 363-372.

Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG, Gottgens B. 2016. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**: e20-31.

Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* **11**: 1-23.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**: 631-644.

Oudelaar AM, Hanssen LLP, Hardison RC, Kassouf MT, Hughes JR, Higgs DR. 2017. Between form and function: the complexity of genome folding. *Hum Mol Genet* **26**: R208-R215.

Paralkar VR, Mishra T, Luan J, Yao Y, Kossenkov AV, Anderson SM, Dunagin M, Pimkin M, Gore M, Sun D et al. 2014. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood* **123**: 1927-1937.

Pilon AM, Ajay SS, Kumar SA, Steiner LA, Cherukuri PF, Wincovitch S, Anderson SM, Center NCS, Mullikin JC, Gallagher PG et al. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood* **118**: e139-148.

Pimkin M, Kossenkov AV, Mishra T, Morrissey CS, Wu W, Keller CA, Blobel GA, Lee D, Beer MA, Hardison RC et al. 2014. Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res* **24**: 1932-1944.

Pinto do OP, Kolterud A, Carlsson L. 1998. Expression of the LIM-homeobox gene LH2 generates immortalized steel factor-dependent multipotent hematopoietic precursors. *Embo J* **17**: 5744-5756.

Pishesha N, Thiru P, Shi J, Eng JC, Sankaran VG, Lodish HF. 2014. Transcriptional divergence and conservation of human and mouse erythropoiesis. *Proc Natl Acad Sci U S A* **111**: 4103-4108.

Psaila B, Barkas N, Iskander D, Roy A, Anderson S, Ashley N, Caputo VS, Lichtenberg J, Loaiza S, Bodine DM et al. 2016. Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol* **17**: 83.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. 2017. Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Syst* **4**: 416-429 e413.

Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* **414**: 105-111.

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

Sanchez-Castillo M, Ruau D, Wilkinson AC, Ng FS, Hannah R, Diamanti E, Lombard P, Wilson NK, Gottgens B. 2015. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* **43**: D1117-D1123.

Sanjuan-Pla A, Macaulay IC, Jensen CT, Woll PS, Luis TC, Mead A, Moore S, Carella C, Matsuoka S, Jones TB et al. 2013. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature* **502**: 232-236.

Schreiber J, Bilmes J, Noble WS. 2019. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *submitted to Nature Biotechnology* doi:https://doi.org/10.1101/533273.

Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. 2012. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol* **13**: R16.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726-732.

Song J, Chen KC. 2015. Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol* **16**: 33.

Stonestrom AJ, Hsu SC, Jahn KS, Huang P, Keller CA, Giardine BM, Kadauke S, Campbell AE, Evans P, Hardison RC et al. 2015. Functions of BET proteins in erythroid gene expression. *Blood* **125**: 2825-2834.

Stunnenberg HG, International Human Epigenome C, Hirst M. 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**: 1145-1149.

Su MY, Steiner LA, Bogardus H, Mishra T, Schulz VP, Hardison RC, Gallagher PG. 2013. Identification of biologically relevant enhancers in human erythroid cells. *J Biol Chem* **288**: 8433-8444.

Sykes SM, Scadden DT. 2013. Modeling human hematopoietic stem cell biology in the mouse. *Seminars in hematology* **50**: 92-100.

Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, Chiaromonte F. 2006. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* **16**: 1596-1604.

The_ENCODE_Project_Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

The_ENCODE_Project_Consortium, Moore J, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A et al. 2019. Expanded Encyclopedias of DNA Elements in the Human and Mouse Genomes. *Nature*: under review.

Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK et al. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20**: 597-609.

Till JE, McCulloch EA. 1961. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res* **14**: 213-222.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS et al. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**: 1530-1545.

Vakoc CR, Mandat SA, Olenchock BA, Blobel GA. 2005. Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell* **19**: 381-391.

Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B et al. 2006. Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res* **16**: 1480-1492.

Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, Li D, Choudhary MNK, Li Y, Hu M et al. 2018. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* **19**: 151.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431-1443.

Weiss MJ, Yu C, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol* **17**: 1642-1651.

Weissman IL, Shizuru JA. 2008. The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases. *Blood* **112**: 3543-3553.

Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E et al. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**: 532-544.

Wilson NK, Schoenfelder S, Hannah R, Castillo MS, Schutte J, Ladopoulos V, Mitchelmore J, Goode DK, Calero-Nieto FJ, Moignard V et al. 2016. Integrated genome-scale analysis of the transcriptional regulatory landscape in a blood stem/progenitor cell model. *Blood*: in press.

Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF. 2011. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood* **118**: e128-138.

Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, Morrissey C, Dorman CM, Chen KB, Drautz D et al. 2011. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* **21**: 1659-1671.

Wu W, Morrissey CS, Keller CA, Mishra T, Pimkin M, Blobel GA, Weiss MJ, Hardison RC. 2014. Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res* **24**: 1945-1962.

Xiang G, Keller CA, Giardine B, An L, Hardison RC, Zhang Y. 2019. S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *submitted to Nature Biotechnology* doi:https://doi.org/10.1101/506634.

Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, Desland F, Chudnovskiy A, Mortha A, Dominguez C et al. 2019. The cis-Regulatory Atlas of the Mouse Immune System. *Cell* **176**: 897-912 e820.

Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, Yu D, Hardison R, Weiss MJ, Orkin SH et al. 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**: 682-695.

Yue F Cheng Y Breschi A Vierstra J Wu W Ryba T Sandstrom R Ma Z Davis C Pope BD et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355-364.

Zhang Y, An L, Yue F, Hardison RC. 2016. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* **44**: 6721-6731.

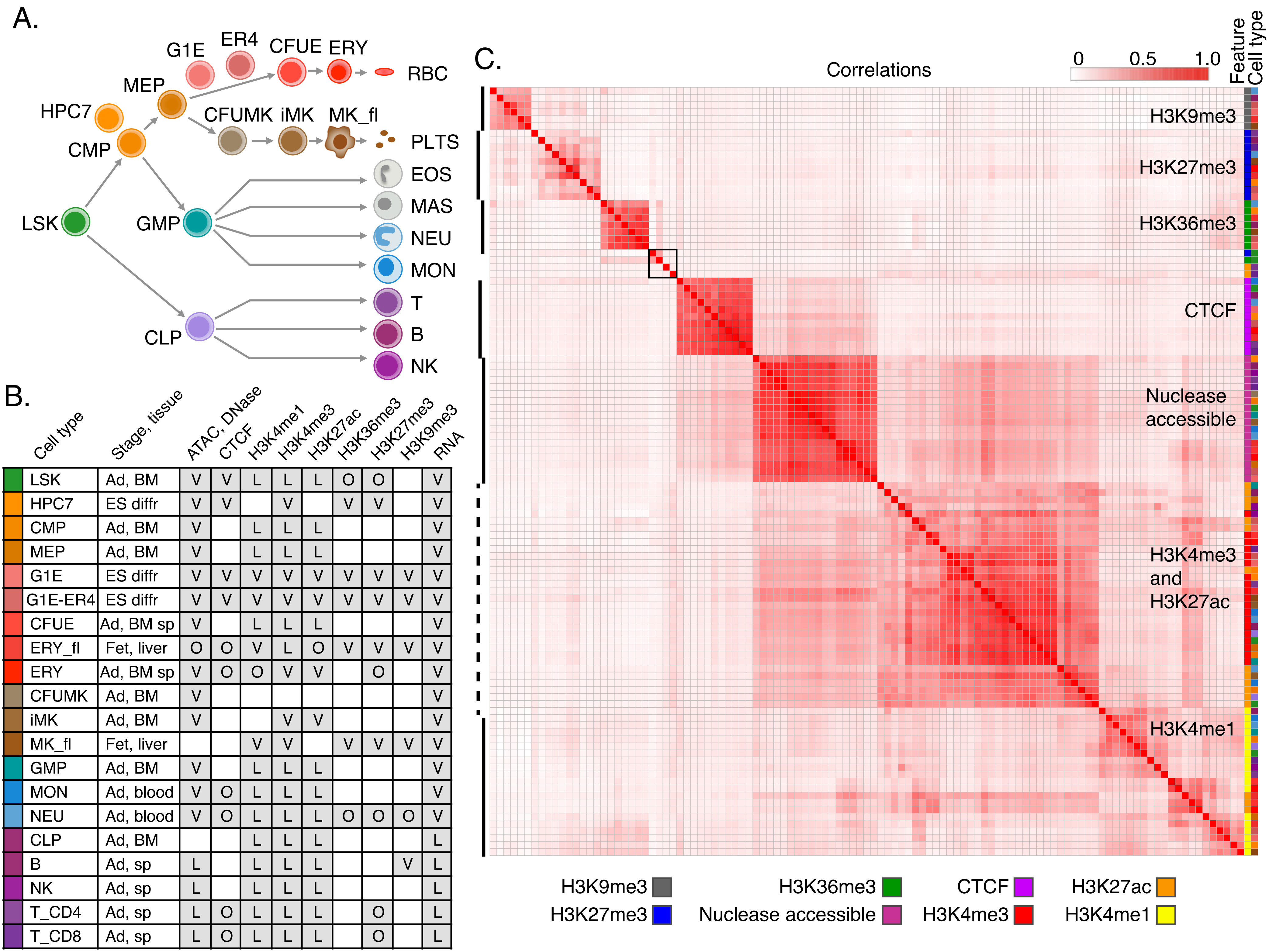
Zhang Y, Hardison RC. 2017. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res* **45**: 9823-9836.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhang Y, Mahony S. 2019. Direct prediction of regulatory elements from partial data without imputation. *PLoS Comput Biol; bioRxiv*: under review.

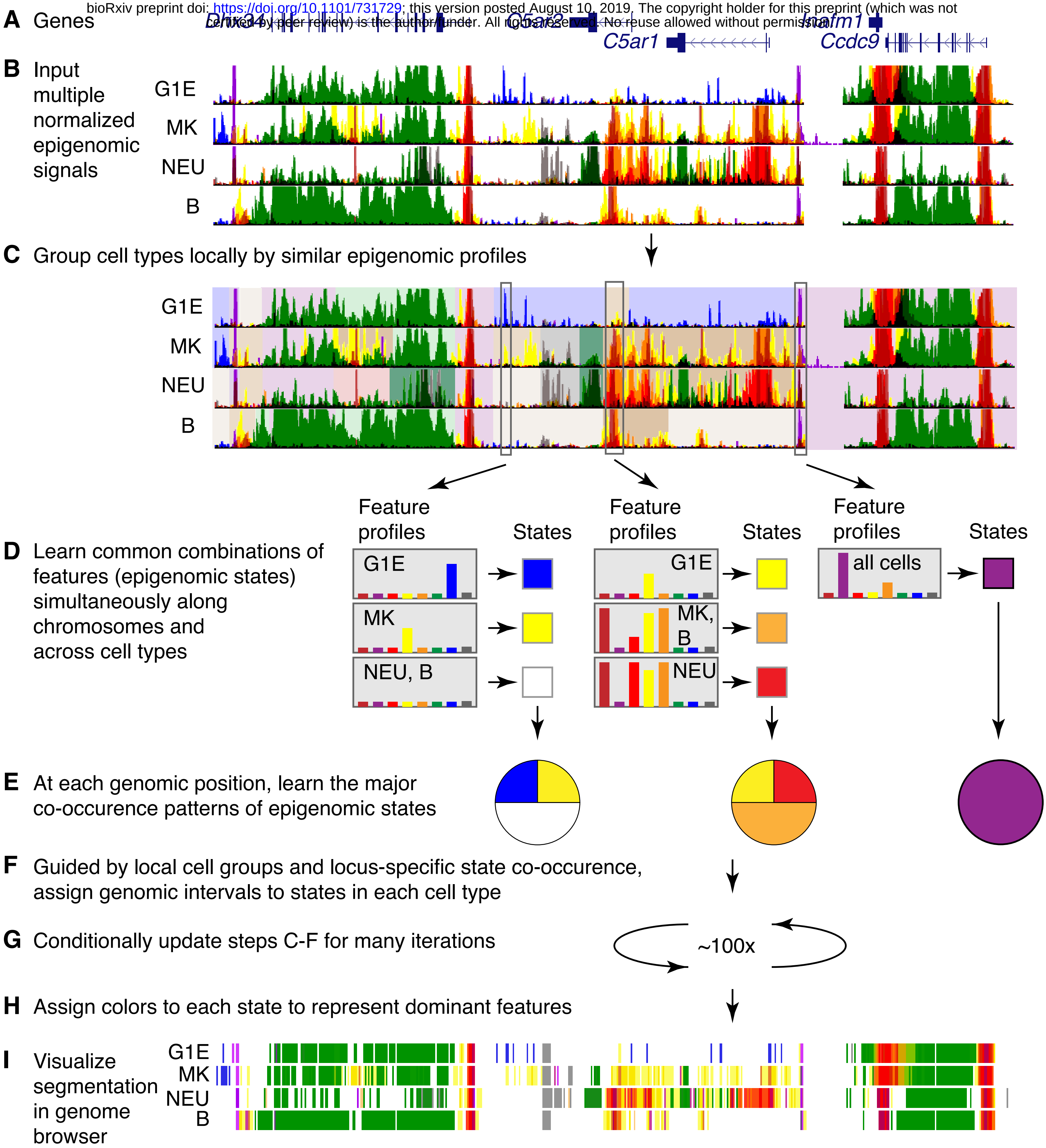
Zheng X, Zheng Y. 2018. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics* **34**: 1568-1570.

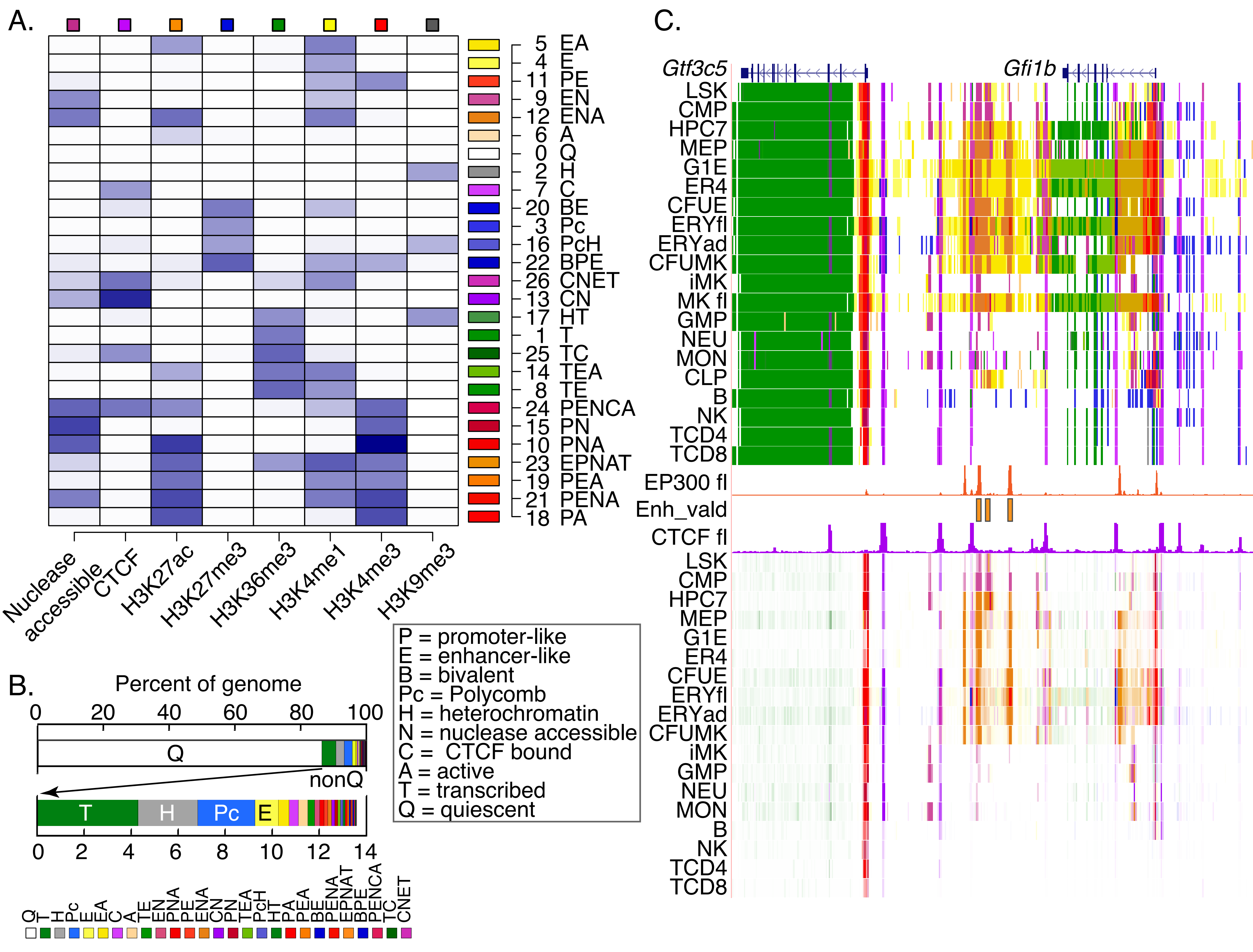
Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931-934.

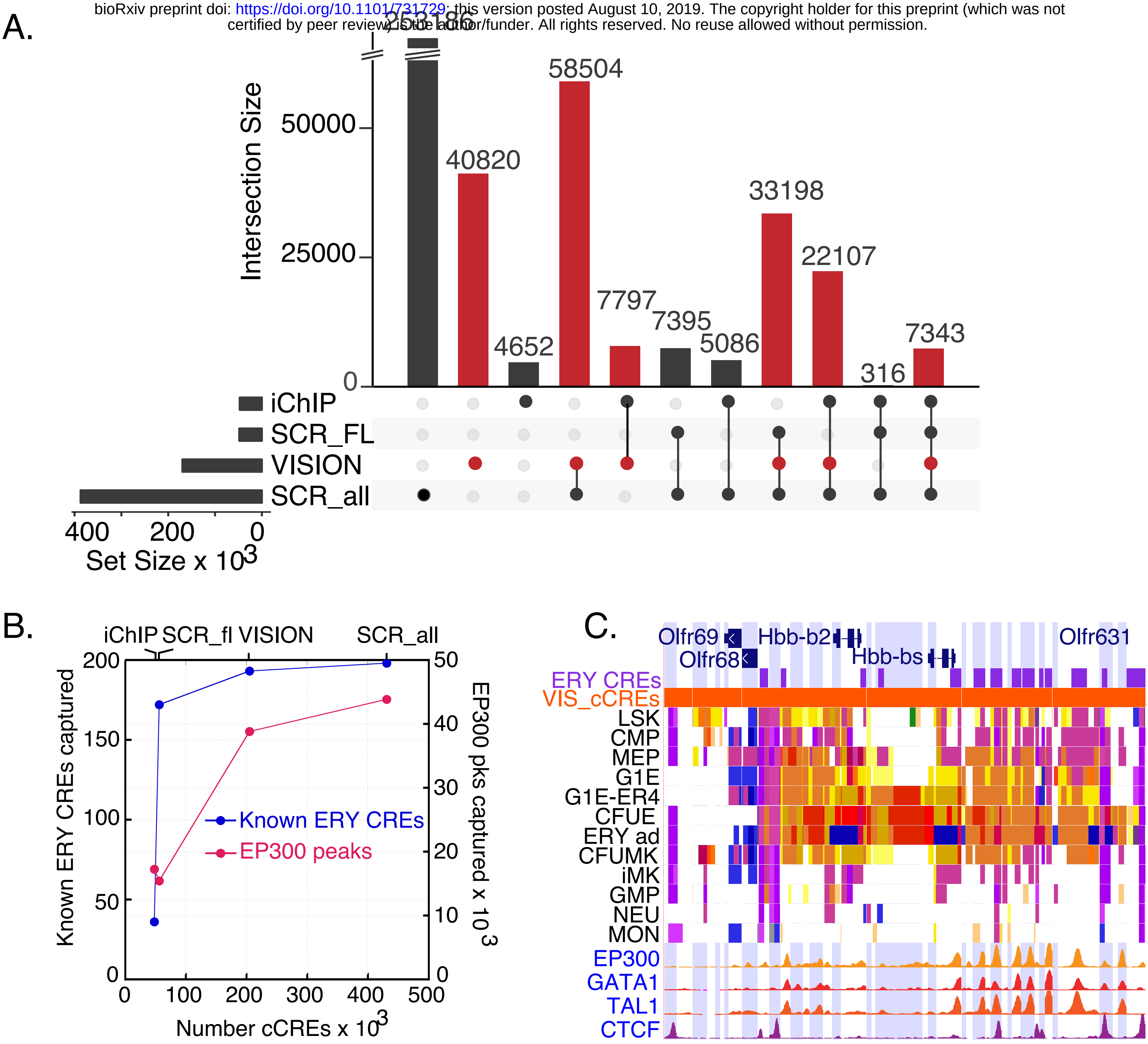


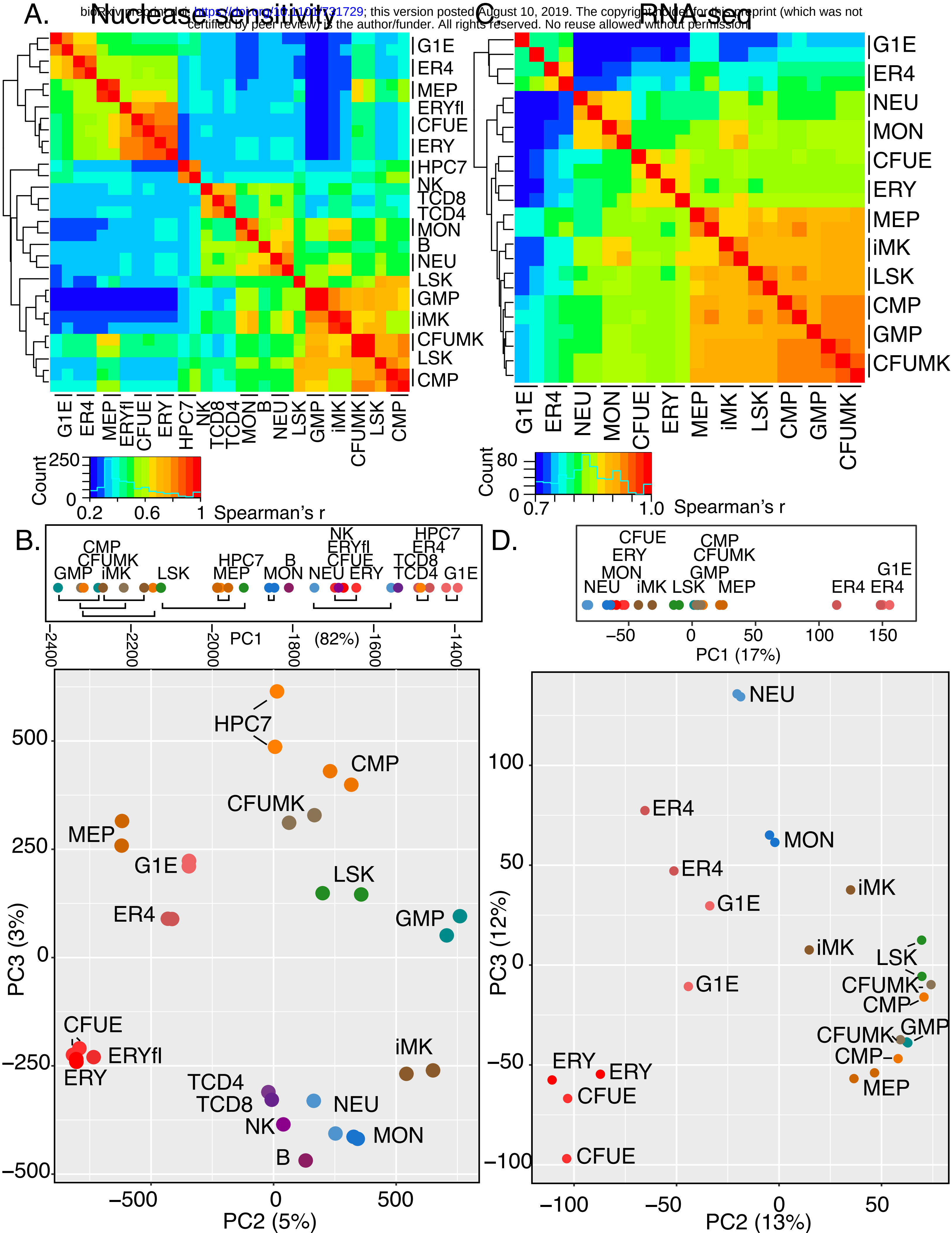
B.

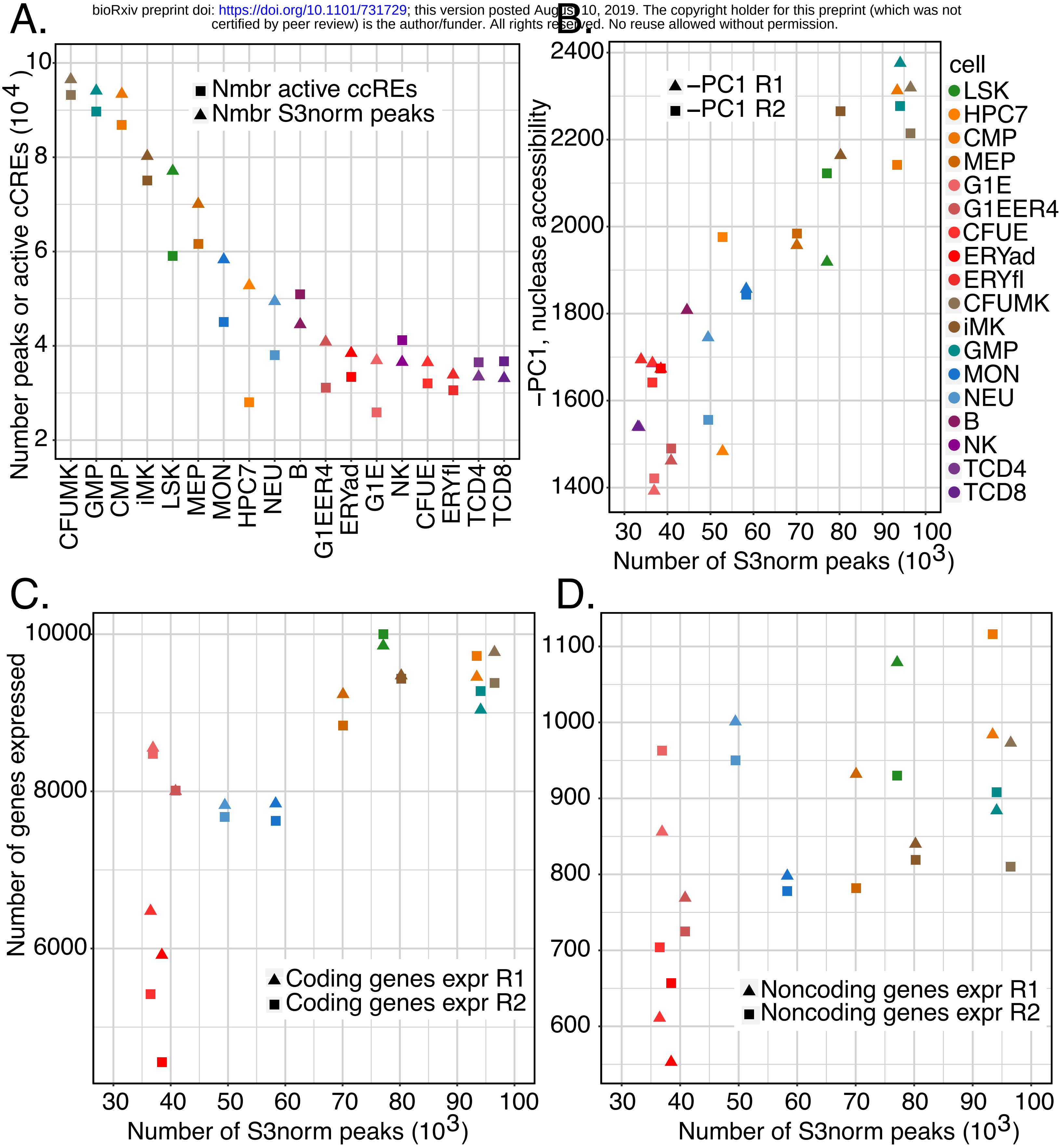
Cell type		Stage, tissue	ATAC, DNase	CTCF	H3K4me1	H3K4me3	H3K27ac	H3K36me3	H3K27me3	H3K9me3	RNA
🟩	LSK	Ad, BM	V	V	L	L	L	O	O		V
🟧	HPC7	ES diffr	V	V		V		V	V		V
🟧	CMP	Ad, BM	V		L	L	L				V
🟧	MEP	Ad, BM	V		L	L	L				V
🟥	G1E	ES diffr	V	V	V	V	V	V	V	V	V
🟥	G1E-ER4	ES diffr	V	V	V	V	V	V	V	V	V
🟥	CFUE	Ad, BM sp	V		L	L	L				V
🟥	ERY_fl	Fet, liver	O	O	V	L	O	V	V	V	V
🟥	ERY	Ad, BM sp	V	O	O	V	V		O		V
🟫	CFUMK	Ad, BM	V								V
🟫	iMK	Ad, BM	V			V	V				V
🟫	MK_fl	Fet, liver			V	V		V	V	V	V
🟦	GMP	Ad, BM	V		L	L	L				V
🟦	MON	Ad, blood	V	O	L	L	L				V
🟦	NEU	Ad, blood	V	O	L	L	L	O	O	O	V
🟪	CLP	Ad, BM			L	L	L				L
🟪	B	Ad, sp	L		L	L	L			V	L
🟪	NK	Ad, sp	L		L	L	L				L
🟪	T_CD4	Ad, sp	L	O	L	L	L		O		L
🟪	T_CD8	Ad, sp	L	O	L	L	L		O		L

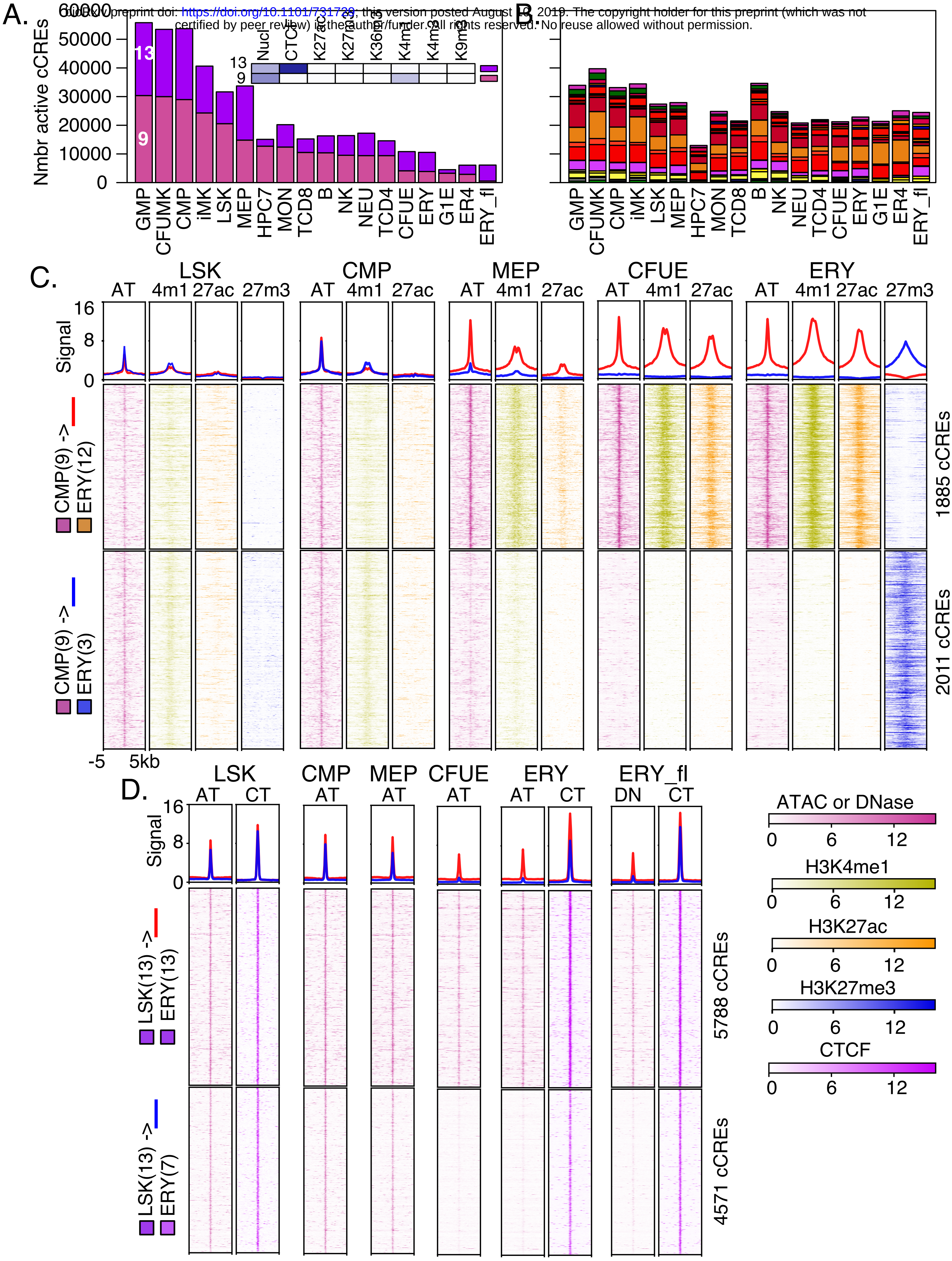

A Genes
Dhx34
C5ar2
C5ar1
Inafm1
Ccdc9
B Input multiple normalized epigenomic signals
G1E
MK
NEU
B
C Group cell types locally by similar epigenomic profiles
G1E
MK
NEU
B
Feature profiles
States
G1E
MK
NEU, B
Feature profiles
States
G1E
MK, B
NEU
Feature profiles
States
all cells
D Learn common combinations of features (epigenomic states) simultaneously along chromosomes and across cell types
E At each genomic position, learn the major co-occurence patterns of epigenomic states
F Guided by local cell groups and locus-specific state co-occurence, assign genomic intervals to states in each cell type
G Conditionally update steps C-F for many iterations
~100x
H Assign colors to each state to represent dominant features
I Visualize segmentation in genome browser
G1E
MK
NEU
B

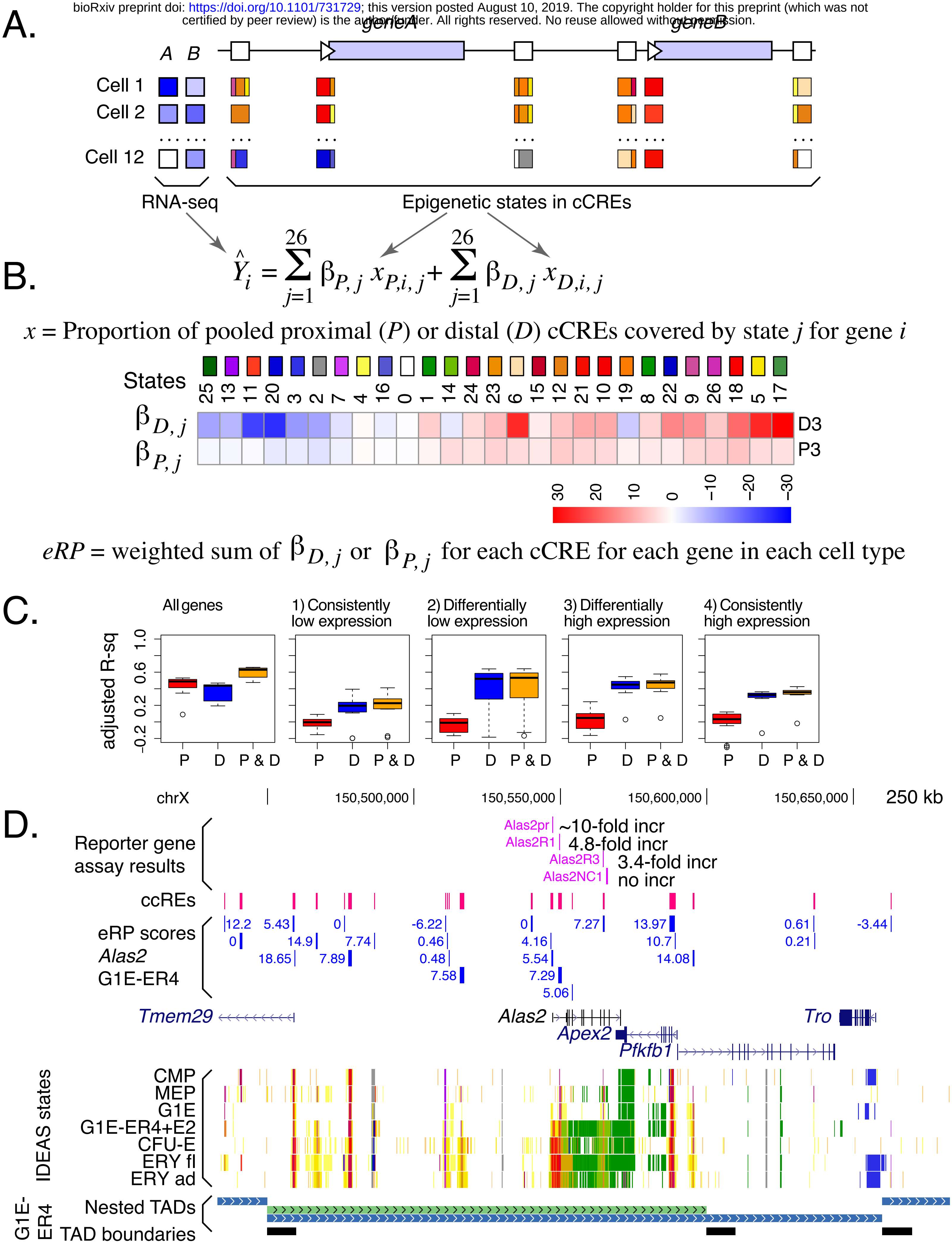

A.
5 EA
4 E
11 PE
9 EN
12 ENA
6 A
0 Q
2 H
7 C
20 BE
3 Pc
16 PcH
22 BPE
26 CNET
13 CN
17 HT
1 T
25 TC
14 TEA
8 TE
24 PENCA
15 PN
10 PNA
23 EPNAT
19 PEA
21 PENA
18 PA
Nuclease accessible
CTCF
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3
B.
Percent of genome
0
20
40
60
80
100
Q
nonQ
T
H
Pc
E
0
2
4
6
8
10
12
14
Q T H Pc E EA C A TE EN PNA PE ENA CN PN TEA PcH HT PA PEA BE PENA EPNAT BPE PENCA TC CNET
P = promoter-like
E = enhancer-like
B = bivalent
Pc = Polycomb
H = heterochromatin
N = nuclease accessible
C = CTCF bound
A = active
T = transcribed
Q = quiescent
C.
Gtf3c5
Gfi1b
LSK
CMP
HPC7
MEP
G1E
ER4
CFUE
ERYfl
ERYad
CFUMK
iMK
MK fl
GMP
NEU
MON
CLP
B
NK
TCD4
TCD8
EP300 fl
Enh_vald
CTCF fl
LSK
CMP
HPC7
MEP
G1E
ER4
CFUE
ERYfl
ERYad
CFUMK
iMK
GMP
NEU
MON
B
NK
TCD4
TCD8


A.
253186
40820
4652
58504
7797
7395
5086
33198
22107
316
7343
Intersection Size
50000
25000
0
iChIP
SCR_FL
VISION
SCR_all
400
200
0
Set Size x 10^3
B.
iChIP
SCR_fl
VISION
SCR_all
200
150
100
50
0
Known ERY CREs captured
50
40
30
20
10
0
EP300 pks captured x 10^3
Known ERY CREs
EP300 peaks
0
100
200
300
400
500
Number cCREs x 10^3
C.
Olfr69
Olfr68
Hbb-b2
Hbb-bs
Olfr631
ERY CREs
VIS_cCREs
LSK
CMP
MEP
G1E
G1E-ER4
CFUE
ERY ad
CFUMK
iMK
GMP
NEU
MON
EP300
GATA1
TAL1
CTCF


A.
Nuclease Sensitivity
G1E
ER4
MEP
ERYfl
CFUE
ERY
HPC7
NK
TCD8
TCD4
MON
B
NEU
LSK
GMP
iMK
CFUMK
LSK
CMP
Count
250
0
0.2
0.6
1
Spearman's r
C.
RNA-seq
G1E
ER4
NEU
MON
CFUE
ERY
MEP
iMK
LSK
CMP
GMP
CFUMK
Count
80
0
0.7
1.0
Spearman's r
B.
CMP
CFUMK
GMP
iMK
LSK
HPC7
MEP
B
MON
NK
ERYfl
CFUE
NEU
ERY
HPC7
ER4
TCD8
TCD4
G1E
−2400
−2200
−2000
PC1
−1800
(82%)
−1600
−1400
500
250
0
−250
−500
PC3 (3%)
−500
0
500
PC2 (5%)
HPC7
CMP
CFUMK
MEP
G1E
LSK
ER4
GMP
CFUE
ERYfl
ERY
iMK
TCD4
TCD8
NEU
NK
MON
B
D.
CFUE
ERY
MON
NEU
CMP
CFUMK
GMP
iMK
LSK
MEP
G1E
ER4
ER4
−50
0
50
100
150
PC1 (17%)
100
50
0
−50
−100
PC3 (12%)
−100
−50
0
50
PC2 (13%)
NEU
ER4
MON
ER4
iMK
G1E
iMK
LSK
G1E
CFUMK
CMP
GMP
CFUMK
CMP
MEP
ERY
ERY
CFUE
CFUE


A.
Number peaks or active cCREs (10^4)
10
8
6
4
2
■ Nmbr active ccREs
▲ Nmbr S3norm peaks
CFUMK
GMP
CMP
iMK
LSK
MEP
MON
HPC7
NEU
B
G1EER4
ERYad
G1E
NK
CFUE
ERYfl
TCD4
TCD8
B.
–PC1, nuclease accessibility
2400
2200
2000
1800
1600
1400
▲ –PC1 R1
■ –PC1 R2
30
40
50
60
70
80
90
100
Number of S3norm peaks (10^3)
cell
● LSK
● HPC7
● CMP
● MEP
● G1E
● G1EER4
● CFUE
● ERYad
● ERYfl
● CFUMK
● iMK
● GMP
● MON
● NEU
● B
● NK
● TCD4
● TCD8
C.
Number of genes expressed
10000
8000
6000
▲ Coding genes expr R1
■ Coding genes expr R2
30
40
50
60
70
80
90
100
Number of S3norm peaks (10^3)
D.
1100
1000
900
800
700
600
▲ Noncoding genes expr R1
■ Noncoding genes expr R2
30
40
50
60
70
80
90
100
Number of S3norm peaks (10^3)


A.
Nmbr active cCREs
60000
50000
40000
30000
20000
10000
0
13
9
Nucl
CTCF
K27ac
K27me3
K36me3
K4me1
K4me3
K9me3
GMP
CFUMK
CMP
iMK
LSK
MEP
HPC7
MON
TCD8
B
NK
NEU
TCD4
CFUE
ERY
G1E
ER4
ERY_fl
B.
C.
LSK
CMP
MEP
CFUE
ERY
AT
4m1
27ac
27m3
Signal
16
8
0
CMP(9) ->
ERY(12)
CMP(9) ->
ERY(3)
1885 cCREs
2011 cCREs
-5
5kb
D.
LSK
CMP
MEP
CFUE
ERY
ERY_fl
AT
CT
DN
LSK(13) ->
ERY(13)
LSK(13) ->
ERY(7)
5788 cCREs
4571 cCREs
ATAC or DNase
H3K4me1
H3K27ac
H3K27me3
CTCF
0
6
12


A.
geneA
geneB
A
B
Cell 1
Cell 2
Cell 12
RNA-seq
Epigenetic states in cCREs
B.
$\hat{Y}_i = \sum_{j=1}^{26} \beta_{P,j}\, x_{P,i,j} + \sum_{j=1}^{26} \beta_{D,j}\, x_{D,i,j}$
x = Proportion of pooled proximal (P) or distal (D) cCREs covered by state j for gene i
States
25 13 11 20 3 2 7 4 16 0 1 14 24 23 6 15 12 21 10 19 8 22 9 26 18 5 17
$\beta_{D,j}$
$\beta_{P,j}$
D3
P3
30 20 10 0 −10 −20 −30
eRP = weighted sum of $\beta_{D,j}$ or $\beta_{P,j}$ for each cCRE for each gene in each cell type
C.
All genes
1) Consistently low expression
2) Differentially low expression
3) Differentially high expression
4) Consistently high expression
adjusted R-sq
-0.2 0.2 0.6 1.0
P D P & D
D.
chrX
150,500,000
150,550,000
150,600,000
150,650,000
250 kb
Reporter gene assay results
Alas2pr ~10-fold incr
Alas2R1 4.8-fold incr
Alas2R3 3.4-fold incr
Alas2NC1 no incr
ccREs
eRP scores
Alas2
G1E-ER4
12.2 5.43 0 -6.22 0 7.27 13.97 0.61 -3.44
0 14.9 7.74 0.46 4.16 10.7 0.21
18.65 7.89 0.48 5.54 14.08
7.58 7.29
5.06
Tmem29
Alas2
Apex2
Pfkfb1
Tro
IDEAS states
CMP
MEP
G1E
G1E-ER4+E2
CFU-E
ERY fl
ERY ad
G1E-ER4
Nested TADs
TAD boundaries