

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

Guanjue Xiang,^{1,10} Cheryl A. Keller,^{1,10} Elisabeth Heuston,² Belinda M. Giardine,¹ Lin An,¹ Alexander Q. Wixom,¹ Amber Miller,¹ April Cockburn,¹ Michael E.G. Sauria,³ Kathryn Weaver,³ Jens Lichtenberg,² Berthold Göttgens,⁴ Qunhua Li,⁵ David Bodine,² Shaun Mahony,¹ James Taylor,³ Gerd A. Blobel,⁶ Mitchell J. Weiss,⁷ Yong Cheng,⁷ Feng Yue,⁸ Jim Hughes,⁹ Douglas R. Higgs,⁹ Yu Zhang,⁵ and Ross C. Hardison¹

¹Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; ²NHGRI Hematopoiesis Section, Genetics and Molecular Biology Branch, National Institutes of Health, Bethesda, Maryland 20892, USA; ³Departments of Biology and Computer Science, Johns Hopkins University, Baltimore, Maryland 20218, USA; ⁴Welcome and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 1TN, United Kingdom; ⁵Department of Statistics, Program in Bioinformatics and Genomics, Center for Computational Biology and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; ⁶Department of Pediatrics, Children's Hospital of Philadelphia and University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA; ⁷Department of Hematology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA; ⁸Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA; ⁹MRC Weatherall Institute of Molecular Medicine, Oxford University, Oxford OX3 9DS, United Kingdom

Thousands of epigenomic data sets have been generated in the past decade, but it is difficult for researchers to effectively use all the data relevant to their projects. Systematic integrative analysis can help meet this need, and the VISION project was established for validated systematic integration 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 using IDEAS as our integrative and discriminative epigenome annotation system, 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 more than 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 to aid research in genomics and hematopoiesis.

[Supplemental material is available for this article.]

Individual laboratories and major consortia (e.g., The ENCODE Project Consortium 2012; Cheng et al. 2014; Yue et al. 2014; Roadmap Epigenomics Consortium et al. 2015; Stunnenberg et al. 2016; The ENCODE Project Consortium et al. 2020) have produced thousands of genome-wide data sets on transcriptomes and many epigenetic features, including nuclease accessibility, histone modifications, and transcription factor occupancy, across diverse cell types. However, it is challenging for individual investigators to find all the data relevant to their projects or to incorporate the data effectively into analyses and hypothesis generation. One approach to address this challenge of overwhelming data is to integrate the deep and diverse data sets (Ernst and Kellis 2010,

2012; Hoffman et al. 2012, 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 multilaboratory project called VISION (validated systematic integration 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, 2014; Kowalczyk et al. 2012; Su et al. 2013; Lara-Astiaso et al. 2014; Pimkin et al. 2014; Corces et al. 2016; Huang et al. 2016;

¹⁰These authors contributed equally to this work.
Corresponding author: rch8@psu.edu

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.255760.119>. Freely available online through the *Genome Research* Open Access option.

© 2020 Xiang et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

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), and it has been studied intensively in humans and mouse. Despite some differences between these species (An et al. 2014; Cheng et al. 2014; Pishsha et al. 2014), the mouse system has served as a good model for many aspects of hematopoiesis in humans and mammals (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). Studies of populations of multilineage progenitor cells, purified using cell surface markers (Weissman and Shizuru 2008), show that hematopoietic differentiation proceeds from HSCs 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 cell populations (Sanjuan-Pla et al. 2013; Psaila et al. 2016). Overall, analyses of single cell transcriptomes support an ensemble of pathways for differentiation (Nestorowa et al. 2016; Laurenti and Göttgens 2018). 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). Misregulation of gene expression patterns can cause diseases such as leukemias and anemias (Higgs 2013; Lee and Young 2013; Ling and Crispino 2020), 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 reports provided many insights and data sets 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 data sets have come from detailed studies in cell line models of hematopoietic differentiation. In the intensively studied process of hematopoiesis, such comprehensive data sets could encompass virtually all the regulatory and transcriptional changes that occur during differentiation. However, distilling the regulatory events that are most critical to producing the transcriptional patterns needed for distinctive cell types is still a major challenge. Here, our major aim is to systematically integrate 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 determined epigenetic states, which are common combinations of epigenetic features, to generate a readily interpretable “painting” of the epigenomic landscape across selected mouse hematopoietic cell populations. The state assignments coupled with peaks of nuclease accessibility produced an initial compendium of more than 200,000 candidate *cis*-regulatory elements (cCREs) active in one or more hematopoietic lineages in mouse, which are valuable for further studies of hematopoietic gene regulation.

Results

Epigenomic and transcriptomic data sets of mouse hematopoietic cells

We reasoned that integrative analysis of the large number of genome-wide determinations of RNA levels and epigenetic features should provide an accessible view of the information that would help investigators use these diverse data sets, and it may lead to novel insights into gene regulation. To conduct the integrative and discriminative analysis, we collated the raw sequence data for 150 determinations of relevant epigenetic features (104 experiments after merging replicates) across 20 cell types or populations (Fig. 1A), including histone modifications and CTCF by ChIP-seq, nuclease accessibility of DNA in chromatin by ATAC-seq and DNase-seq, and transcriptomes by RNA-seq. The purified cell populations and cell lines are described in detail in the Supplemental Material, section 1 (Supplemental Fig. S1).

The epigenomic data were gathered from many different sources, including individual laboratories and consortia (Fig. 1B; Supplemental Tables; Supplemental Fig. S2). These data had quality metrics within the ENCODE recommendations (see Supplemental Material, section 2; Supplemental Tables S1–S5). However, this diversity of sources presented a challenge for data analysis because 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. 2020), all of which can impact downstream analyses. We used two strategies to improve the comparability of these heterogeneous data sets. First, the sequencing reads from each type of assay were uniformly processed, using pipelines similar to or adapted from current ENCODE pipelines (see Supplemental Material, section 2). 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 and 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 (Methods) (Xiang et al. 2020). As with other normalization procedures, the S3norm method gives similar signals in common peaks for an epigenetic feature, but it does so without inflating the background signal (Supplemental Material, section 3; Supplemental Fig. S3). Preventing an increased background was necessary to avoid introducing spurious signals during the genome-wide modeling of the epigenetic landscape.

An overview of the similarities across all the data sets showed that most clustered by epigenetic features across cell types (Supplemental Fig. S4). For example, nuclease accessibility was highly correlated among the cell types examined, showing the global similarity in this primary feature of the regulatory landscape in blood cells (Fig. 1C). Other features such as CTCF and the signature marks for active promoters (H3K4me3) and enhancers (H3K27ac) showed notable but substantially lower correlations with the nuclease accessibility signal. In contrast, the H3K9me3 heterochromatin mark, the H3K27me3 Polycomb repressive mark, and H3K36me3 had almost no correlation with nuclease sensitivity, and H3K4me1 showed modest correlation. The groupings within epigenetic features were more apparent after S3norm normalization (Supplemental Fig. S5), which supports the effectiveness of the normalization. The similarity of patterns for a particular

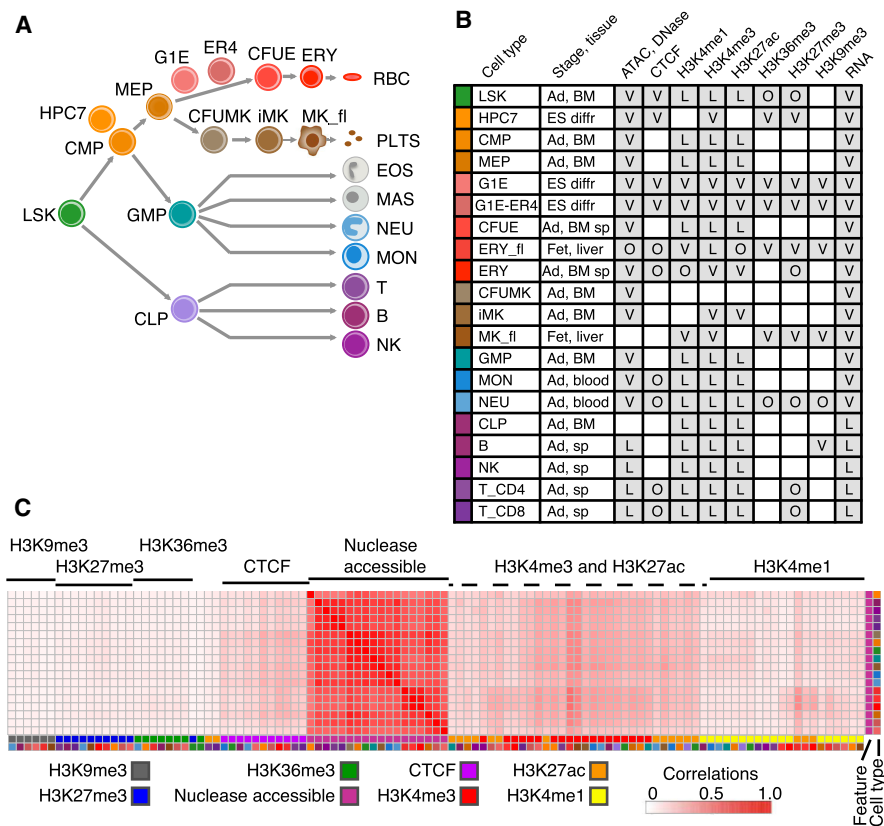


Figure 1. Hematopoietic cell types and data sets used for integrative analysis. (A) Schematic representation of the main lineage commitment steps in hematopoiesis, along with three immortalized cell lines (HPC7, G1E, G1E-ER4) and their approximate position relative to the primary cell populations shown. Abbreviations for cell populations are as follows: (LSK) Lin⁻Sca1⁺Kit⁺, which includes hematopoietic stem cells and multipotent progenitor cells; (CMP) common myeloid progenitor cells; (GMP) granulocyte monocyte progenitor cells; (MEP) megakaryocyte erythrocyte progenitor cells; (CLP) common lymphoid progenitor cells; (CFUE) colony forming unit erythroid; (ERY) erythroblasts; (RBC) red blood cells; (CFUMK) colony forming unit megakaryocyte; (iMK) immature megakaryocytes; (MK_fl) maturing megakaryocytes from fetal liver; (PLTS) platelets; (EOS) eosinophils; (MAS) mast cells; (NEU) neutrophils; (MON) monocytes; (T_CD8) CD8⁺ T cells; (T_CD4) CD4⁺ T cells; (B) B cells; (NK) natural killer cells. (B) Available hematopoietic data sets. 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 data set, and letters denote the source ([V] VISION; [L] Lara-Astiaso et al. (2014); [O] other). For more information, see Supplemental Table S1. (C) Correlations of nucleosome-accessible signals with all features (S3norm normalized) and across 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 full correlation matrix of all features across all cell types is in Supplemental Figure S4.

feature across cell types suggested that combinations of features may be more effective than a single epigenetic mark to find patterns distinctive to a cell type.

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 nonbinary epigenomic data

The frequent co-occurrence of some histone modifications has led to discrete models for epigenetic structures of cCREs (for review,

see 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 integrative and discriminative epigenome annotation system (Zhang and Hardison 2017; Zhang et al. 2016), or IDEAS, 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 (Supplemental Fig. S6). Also, when confronted with missing data, it can make state assignments with good accuracy (Zhang and Mahony 2019).

When applied to the normalized epigenomic data from the 20 hematopoietic cell types, IDEAS learned 27 epigenetic states, including many expected ones as well as others that have been less frequently studied. The IDEAS model summary shows the prevalence of the eight epigenetic features in each state as a heatmap, organized by similarity among the states (Fig. 2A). The epigenetic state assignments were well supported by the underlying epigenomic data (Fig. 2B; Supplemental Fig. S3C). The epigenetic states described an informative landscape, distinguishing multiple state 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, the four promoter-like states that were also nucleosome accessible (states 21, 10, 15, and 24) may encompass the nucleosome-depleted region found adjacent to the transcriptional start site (TSS). Supporting this interpretation, three of these states (states 21, 10, and 24) also had the H3K27ac mark that frequently flanks the nucleosome-depleted region of active promoters. For all the major categories of chromatin associated with gene expression and regulation, including bivalent promoters, CTCF occupancy,

and these were removed from the set of cCREs. No cell type-specific 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 across the genome globally.

Because a wide range of hematopoietic cells was 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 16 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. 2020). 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. 2020). The overlapping DNA intervals among combinations of data sets revealed substantial consistency in the inferred cCREs (Fig. 3A). A large proportion of the VISION cCREs (70,445 or 41.5%) were 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 in VISION cCREs. The cCREs that are common among these collections, despite differences in data input and analysis, 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. By using a collection of 212 experimentally determined, erythroid CREs curated from the literature (Dogan et al. 2015) as known erythroid CREs, we found that although the iChIP Blood Enhancer catalog captured only a small portion, the VISION and SCREEN fetal liver cCREs overlapped with almost all the erythroid CREs (Fig. 3B). The latter two collections were built from data sets 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 data sets from fewer erythroid cell types. Increasing the number of cCREs to more than 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.

The coactivator EP300 catalyzes the acetylation of histone H3K27, and it is associated with many active enhancers. We used ChIP-seq data on EP300 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 data sets of EP300 ChIP-seq data determined in three blood-related cell types from mouse, MEL cells representing maturing proerythroblasts, CH12 cells representing B cells, and mouse fetal liver from embryonic day 14.5 (Yue et al. 2014; The ENCODE Project Consortium et al. 2020). After reprocessing the ChIP-seq data using the VISION project pipelines, replicated peaks were merged across the cell types to generate a set of more than 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 (Fig. 3B). Expanding the number of SCREEN cCREs to more than 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 (Supplemental Fig. S9), suggesting that VISION cCREs captured the more likely functional EP300 peaks.

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).

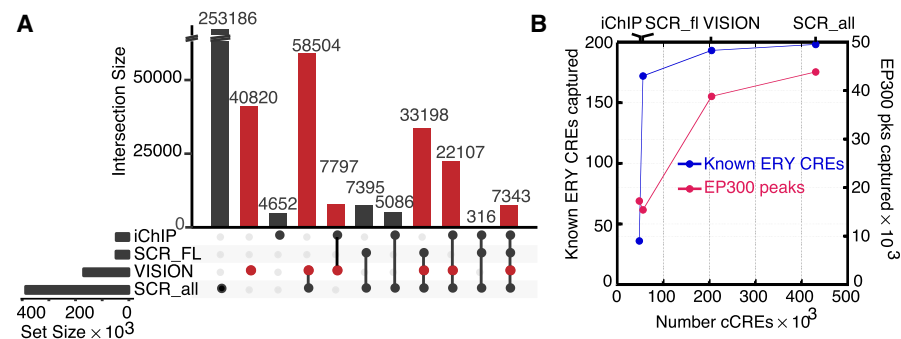


Figure 3. Comparative analysis of VISION cCREs. (A) Overlaps of the VISION cCREs with three other cCRE catalogs. The overlapping cCREs in all four data sets were merged. The numbers of merged cCREs in each set are labeled on each row, and the numbers in each level of overlap are shown in columns, visualized 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 data set. 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 data set and merged to get more than 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 data set.

Global comparisons of regulatory landscapes and transcriptomes

The collection of cCREs and transcriptomes in VISION provided an opportunity to examine the relationships between cell types, including both purified populations of primary cells and cell lines. In conducting this analysis, we distinguished a cCRE from an active cCRE. A cCRE, which is a DNA interval predicted to be a regulatory element in any cell type, is present in all cell types, just as a gene is present in all cell types. However, a cCRE can show evidence of activity (either positive or negative) differentially across cell types, just as genes may be active in only some cell types. Thus, we refer to cCREs in epigenetic states indicative of regulatory activity as active cCREs, including states with either positive or negative associations with gene expression.

The epigenetic modifications at cCREs are a prominent feature of the regulatory landscape. Thus, to compare the regulatory landscape across cell types, we used the correlations between the nuclease accessibility signals in cCREs across cell types to group the cell types by hierarchical clustering (Fig. 4A). All erythroid cell types, including the G1E and G1E-ER4 cell lines, clustered with MEP to the exclusion of other cell types. The remaining cell types formed two groups. One consisted of hematopoietic stem and multilineage progenitor cells (LSK, CMP, and GMP) along with early progenitor (CFUMK) and immature (iMK) megakaryocytic cells. The other contained both innate (NEU, MON) and acquired (B, NK, T-CD4, T-CD8) immunity cells. Comparisons using a dimensional reduction approach (principal component analysis or PCA) also supported these groupings (Supplemental Fig. S10A).

Furthermore, the PCA and subsequent analyses showed that a substantial reduction in the number of active cCREs was a major contributor to the differences in the landscape of nuclease accessibility during hematopoietic differentiation. The first principal component (PC1) captured a large 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 (Supplemental Fig. S10A). That PC1 axis was highly correlated with the numbers of active cCREs (Supplemental Fig. S10B), and a direct comparison showed a progressive decline in numbers of cCREs active in most maturing blood cells (Fig. 4B). We conclude that a reduction in numbers of active cCREs is a major trend during mouse hematopoietic differentiation.

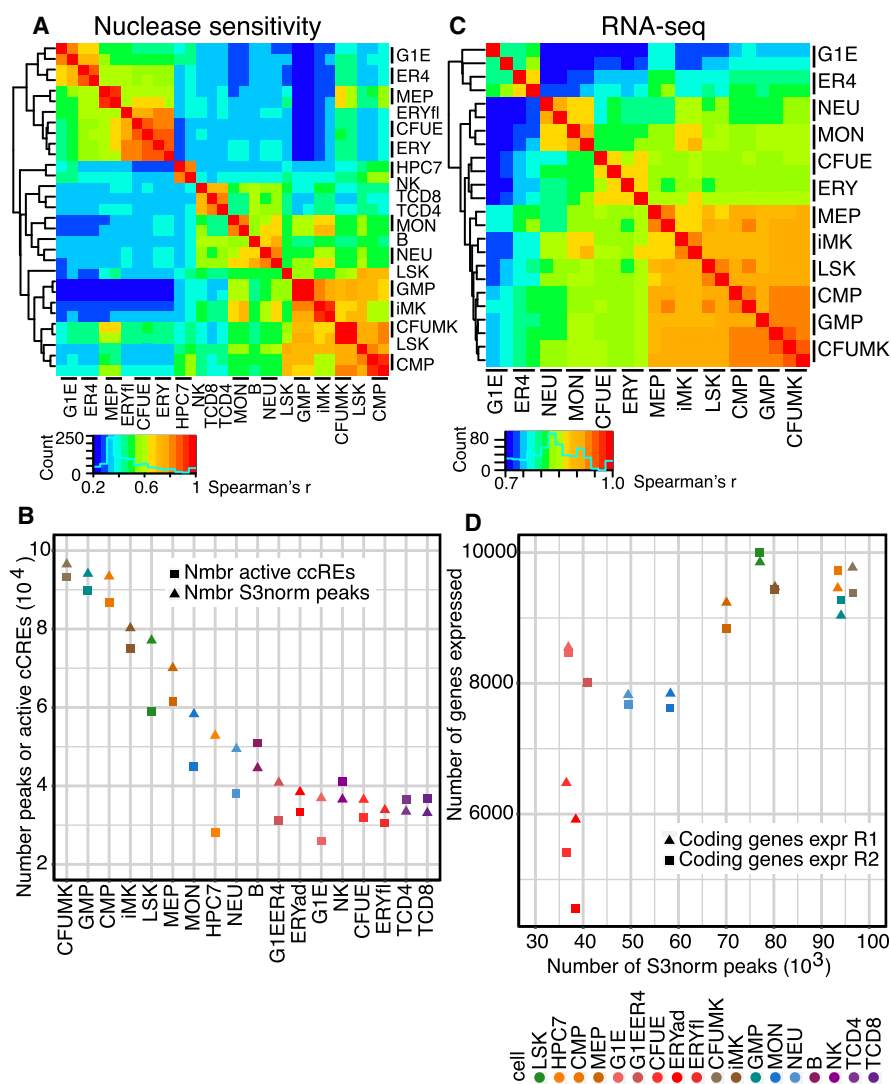


Figure 4. Global comparisons of nuclease accessibility profiles and transcriptomes across mouse hematopoietic cell types. (A) Heatmap of the hierarchical clustering of nuclease-sensitive elements (ATAC-seq and DNase-seq, using S3norm for normalization), with Spearman's rank correlation r as the similarity 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) Numbers of dynamic cCREs in each cell type, determined from ATAC-seq and DNase-seq profiles and analyzed as both peak calls from HOMER (Heinz et al. 2010) or from peaks after S3 normalization. (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) Concordant decreases during hematopoietic differentiation in nuclease accessibility and expressed genes, shown as the association between numbers of genes expressed and numbers of dynamic cCREs across cell populations and types.

The gene expression landscape was also 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 performed 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 (Fig. 4C) and PCA (Supplemental Fig. S10C) revealed three clusters that were largely 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 before 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. Although 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.

The decrease in numbers of cCREs during differentiation and maturation was associated with a decrease in numbers of genes expressed. The highest numbers of protein-coding genes were expressed in the progenitor (LSK, CMP, GMP, MEP) and megakaryocytic (CFUMK and iMK) cells, with fewer in MON and NEU, and the lowest number in erythroid cells (CFUE and ERY) (Fig. 4D). A larger number of genes 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 10-fold range of thresholds for declaring a gene as expressed (TPM exceeding one, five, or 10). The parallel decreases in numbers of active cCREs and expressed genes led to a strong positive association between these two features (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) (for coding genes, see Fig. 4D; for noncoding genes, see Supplemental Fig. S10D). 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 the reduction in numbers of expressed genes and active cCREs was observed broadly across hematopoiesis. By considering specifically genes encoding hematopoietic regulators, we found that this general decline in transcription led to a reduction in the number of hematopoietic regulators produced in differentiated, maturing erythroid cells but not in other hematopoietic cell types (Supplemental Fig. S11). We conclude that the breadth of transcription declines during differentiation, and furthermore, the loss of activity of cCREs may contribute to the decrease in numbers of genes expressed.

Epigenetic states of cCREs vary across cell types in an informative manner

The VISION catalog of cCREs, annotated by their epigenetic state in each cell type, can be used to track both the timing and types of transitions in epigenetic states during differentiation, which provide insights into regulatory mechanisms, for example, which cCREs are likely to be inducing or repressing a target gene. The full scope of state transitions in cCREs across cell types is complex, and in this section, we focus on major transitions contributing to changes in the numbers and state of 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) (Fig. 5A), whereas changes in the numbers of cCREs in other states were more modest (Fig. 5B; Supplemental Fig. S12A,B). These state-specific reductions suggested that many

active 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 tend to stay in state 9, but rather they more frequently transitioned to states 12 (active enhancer), 3 (polycomb), and 0 (quiescent) in ERY (Supplemental Fig. S12C). These classes of state transitions were strongly supported by examination of the underlying signals for the nuclease sensitivity and histone modifications (Fig. 5C). 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, by 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 SPI1 (also known as PU.1) (Supplemental Fig. S13). These results are consistent with the known antagonism between GATA1 and SPI1 in erythroid versus myeloid differentiation (Rekhtman et al. 1999; Zhang et al. 1999). Thus, they illustrate the value of machine-learning approaches, such as assigning epigenetic states systematically and finding discriminative motifs, to uncover relationships from genome-wide data that fit with models derived from decades of experimentation.

Another major state of cCREs in progenitor and megakaryocytic cells was CTCF bound and nuclease accessible (state 13). Much of the decrease in numbers of cCREs in this state occurred through a loss of accessibility while retaining occupancy by CTCF (state 7) (Supplemental Fig. S12C,D). To eliminate the possibility that the inferred loss of nuclease sensitivity was an artifact of low sensitivity in the ATAC-seq data, we examined these cCREs for DNase sensitivity in an independent experiment conducted on ERY from fetal liver (ERY_{fl}). We found that the cCREs undergoing the transition from state 13 to state 7 had low nuclease sensitivity in ERY by both assays, as well as in CFUE, while retaining a strong CTCF signal (Fig. 5D). Thus, 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 the MEP and CFUE stages, suggesting that it could be connected to the process of erythroid commitment. By examining genes in the vicinity of the CTCF-bound cCREs, we found that this loss of nuclease sensitivity at CTCF-bound sites occurred in more gene-poor regions, and it was associated to some extent with gene repression (Supplemental Fig. S14). The CTCF-bound cCREs that retained nuclease accessibility during differentiation were enriched at topologically associated domain (TAD) boundaries that were common across myelo-erythroid differentiation (Supplemental Fig. S15).

In summary, the number of active cCREs declined as cells differentiated from stem and progenitor cells to committed, maturing blood cells. This decrease in cCREs was 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 was prevalent in cCREs in stem and progenitor cells, and it had two major fates. One was a transition to an active enhancer state, and in the erythroid lineage this transition was associated with GATA transcription factor binding site motifs, as

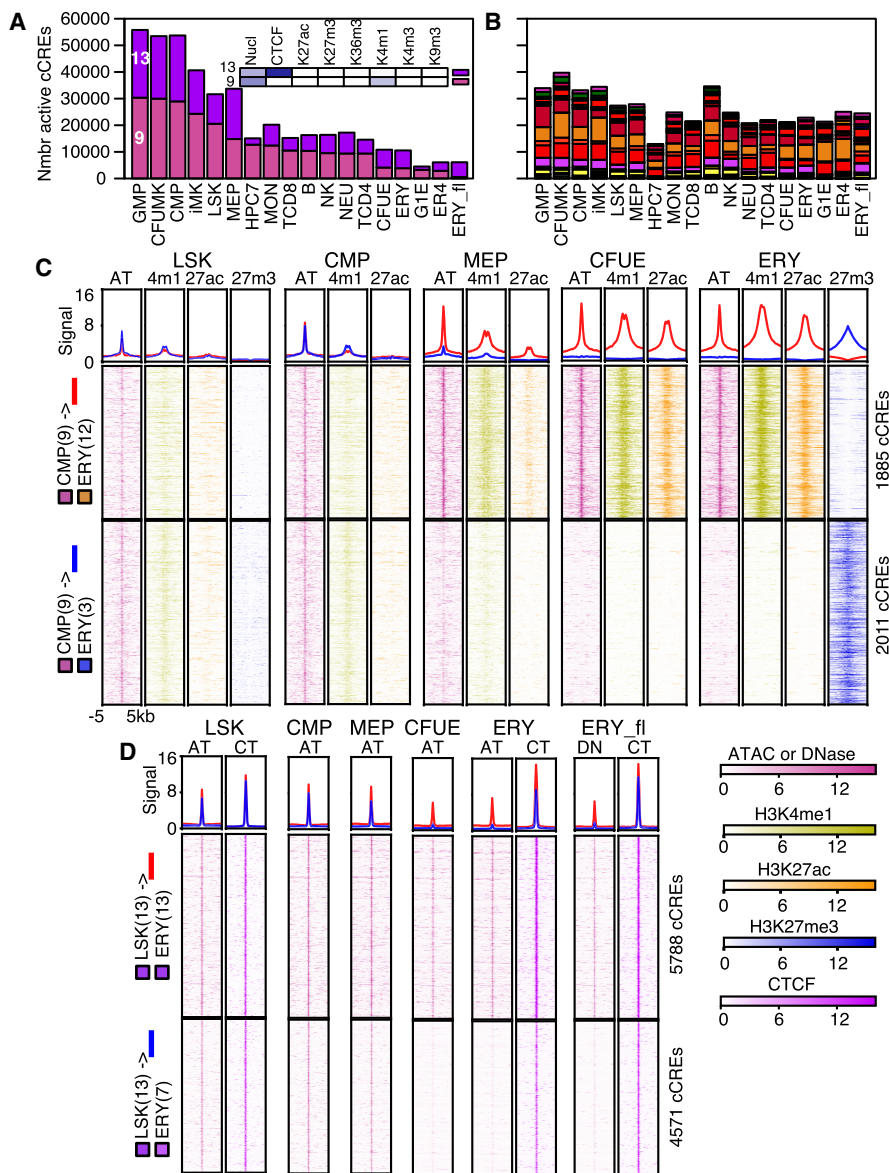


Figure 5. Transitions in epigenetic states at cCREs across hematopoietic differentiation. (A,B) The numbers of cCREs in each cell type are colored by their IDEAS epigenetic state, emphasizing decreases in numbers of cCREs in states 9 and 13 (A), whereas numbers in other states are less variable (B). (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 below 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 polycomb repressed state 3, and blue lines show signals for cCREs that transition from poised state 9 to nucleosome-sensitive state 12. Signals were normalized by S3norm. (AT) ATAC; (4m1) H3K4me1; (27ac) H3K27ac; (27m3) H3K27me3; (CT) CTCF.

expected for activation of erythroid genes. The other fate was to lose nuclease sensitivity and switch to a repressed state. Those state transitions were not novel observations, but our extensive annotation of the cCREs allows investigators to identify which cCREs around genes of interest make those transitions. Second, another state prevalent in stem and progenitor cells was a

CTCF-bound and nuclease-accessible state. The number of cCREs in that state declined during differentiation, with many cCREs transitioning to a state with CTCF still bound but no longer nuclease accessible. Further studies are needed to better understand the roles of these different classes of CTCF-bound sites.

Estimating regulatory output and assigning target genes to cCREs

We investigated the effectiveness of the collection of mouse hematopoietic cCREs from VISION in explaining levels of gene expression. We developed a modeling approach to evaluate how well the cCREs, in conjunction with promoters, could account for levels of expression in the 12 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 because the states were based on a systematic integration of multiple epigenetic signals. As explained in detail in the Supplemental Material, section 16, we estimated promoter and 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 and promoter could be composed of multiple epigenetic states (Fig. 6A), and we used the proportion of promoters and the proportion of pooled cCREs covered by a state as the predictor variable for that state (Fig. 6B). However, in our subselection training, a given cCRE is represented by a single state rather than a weighted sum of states (Supplemental Material). All cCREs within 1 Mb of the TSS of a gene were initially considered and then filtered by a minimum correlation to that gene's expression. Not all cCREs within the 2-Mb region surrounding a gene's TSS were expected to influence expression. Thus, cCREs predicted to have limited contribution to explaining expression were removed

via a subselection strategy during iterations of model fitting (Fig. 6B; Supplemental Fig. S16B).

The regression coefficients, β , determined for the epigenetic states showed some expected trends. 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-

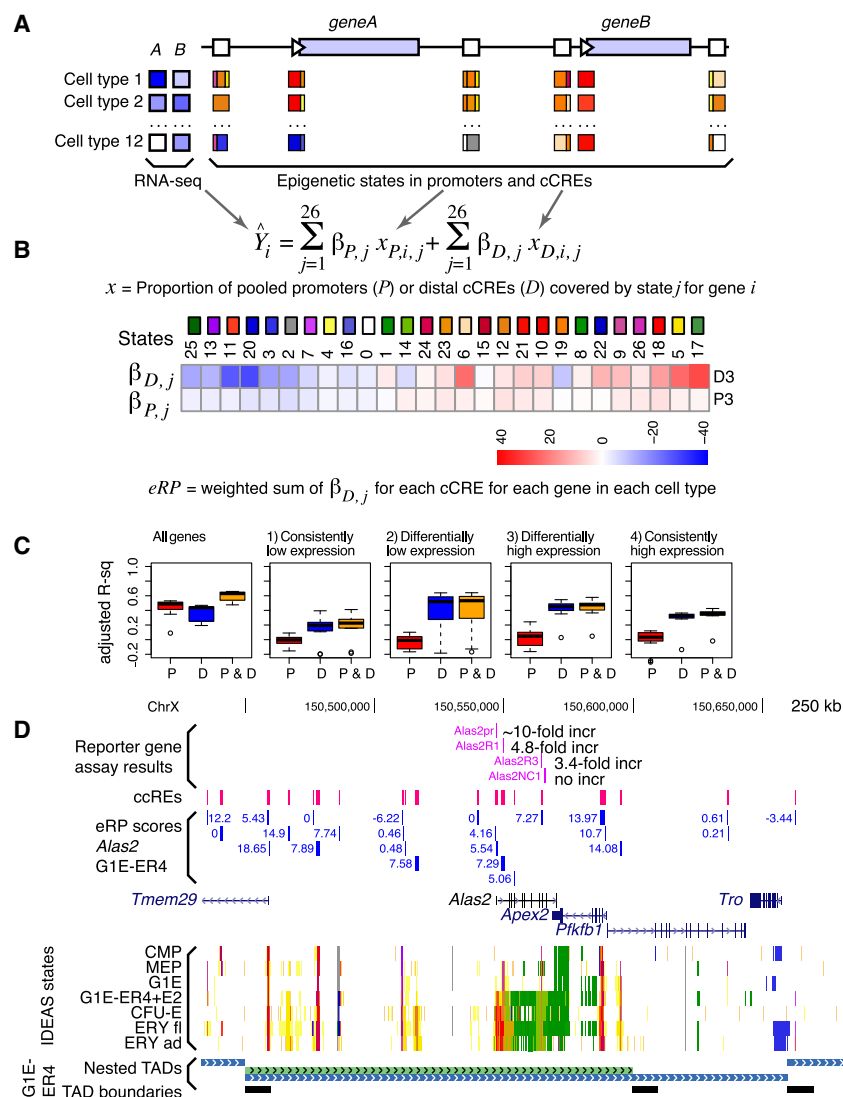


Figure 6. Initial estimates of regulatory output and target gene prediction using regression models of IDEAS states in promoters and cCREs versus gene expression. (A) Illustration of promoters and cCREs around two potential target genes, showing expression profiles of the genes across cell types (shades of blue; left) and promoters/cCREs with one or more epigenetic states assigned in each cell type. (B) Multivariate linear regression of proportion of promoters and pooled cCREs in each state against expression levels of potential target genes, keeping promoters and cCREs separate and learning the regression coefficients iteratively in a subselection strategy. Values of the regression coefficients' beta for each epigenetic state for promoters and cCREs for differentially expressed genes. The values of the regression coefficients for each epigenetic state are presented as a blue-to-red heatmap. (C) Ability of eRP scores of cCREs to explain levels of expression on Chr 1–Chr 19 and Chr X in the 12 cell types for all genes and in the four categories of genes (1–4). A leave-one-out strategy was used to calculate the accuracy predicting expression. The distributions of adjusted r^2 values are shown as box-plots for promoters, distal cCREs, and combined. (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.

repressed and heterochromatin states (Fig. 6B; for a full set of values, see Supplemental Fig. S16D).

We evaluated the accuracy of predicting gene expression from the weighted sum of the state-specific regression coefficients using a leave-one-out strategy. Specifically, we trained a linear model on data from 11 of the 12 cell types, minimizing mean squared error (MSE), and then computed the adjusted r^2 for the accuracy of the predicted expression levels compared with the actual expression

levels in the left-out cell type. This procedure was repeated, leaving out each of the cell types in turn. Coefficients were calculated using only promoters, only cCREs, or a combination of both. In the case of the cCRE trained model, we defined the sum of coefficients weighted by each cCRE state proportion as the epigenetic regulatory potential (eRP) score. The predicted expression for each gene was the mean of the eRP scores for all paired cCREs. For expression of all genes, the prediction accuracy was ~50% for promoters only or eRPs only, and it improved to ~60% when both were combined (Fig. 6C, 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 β varied across the four categories (Supplemental Fig. S16D). Using gene category partitioning, the accuracy of predicting expression levels in the leave-one-out strategy showed a much smaller impact of the promoters (Fig. 6C, 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 (Fig. 6C, graphs 2 and 3). Overall, these evaluations indicated that promoters 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.

By considering these linear regression coefficients as proxies for the regulatory output of cCREs in a particular epigenetic state, we used them to estimate the impact of histone modifications around cCREs close to differentially expressed genes. Many expected associations were found, but in addition, this analysis revealed that H3K27ac was the histone modification at cCREs most distinctly associated with gene activation, CTCF at a cCRE was associated with repression, and H3K4me1 and nuclease accessibility were about equally frequent in states with positive or negative impacts on expression (Supplemental Fig. S17).

The positive predictive power of these initial estimates of eRP scores supported their 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 (Fig. 6D) illustrates how these eRP scores were 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 and produce resources of high utility to investigators of genome biology, blood cell differentiation, and other processes. In this study, we compiled and generated epigenomic and transcriptomic data on cell types across hematopoietic differentiation in mouse. The data were systematically analyzed by the IDEAS method to assign genomic intervals to epigenetic states in 20 cell types, with each state defined by a quantitative spectrum of nuclease sensitivity, histone modifications, and CTCF occupancy. Most of these combinations of epigenetic features are associated with specific regulatory elements or events, such as active promoters, poised enhancers, transcribed regions, or quiescent zones, and thus, the epigenetic state assignments provide a guide to potential functions of each genomic interval in each cell type. In effect, the IDEAS segmentation pipeline reduced 150 dimensions (or tracks) of epigenomic data to 20 dimensions, that is, the number of cell types examined. Although the cell populations studied can be conceptualized as cell “types,” it is important to keep in mind that these populations, especially of stem and progenitor cells, are heterogeneous, and thus our integrative analyses do not delve into all the stages of hematopoietic differentiation and maturation. We further focused the epigenomic data by constructing an initial registry of 205,019 cCREs, which are discrete genomic intervals with features predictive of a potential regulatory role in one or more hematopoietic cell types, along with state assignments and initial estimates of regulatory output for candidate target genes in each cell type. Investigators now have simplified ways to view the large amount of data, for example, in a genome browser, and to operate computationally on the state assignments and cCREs.

We provide multiple ways for investigators to access and interact with the data via our VISION website (<http://usevision.org>). The raw and normalized data tracks can be downloaded for further analysis. The regulatory and transcriptomic landscapes around individual genes can be viewed in our 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 can be downloaded, 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 (Sánchez-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. 2020).

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 and Hardison 2017; Zhang et al. 2016). Furthermore, the IDEAS method does not require determination of all features in all cell types, and thus cell types with missing data were included (Zhang and Mahony 2019). Even an extreme case of the cell type CFUMK, for which the only epigenomic data set was ATAC-seq, was assigned a meaningful segmentation pattern. The local clustering of cell types by their epigenomic profiles in IDEAS allows the system to learn the signal distribution for a feature missing in one cell type from the available signal in locally related cell types and then use that signal distribution when assigning likely states in the cell type with missing data. Although full determination of all biochemical features in each cell type is preferred, attaining complete coverage is difficult, especially for rare cell types. Indeed, many integrative analysis projects are contending with the challenges of missing data (Ernst and Kellis 2015; The ENCODE Project Consortium et al. 2020; Schreiber et al. 2020). We suggest that the IDEAS method provides a principled approach with good utility for integrative analyses in the face of missing data.

Our collection of cCREs in mouse blood cells efficiently captures known erythroid regulatory elements and potential enhancers predicted by available EP300 occupancy data. However, this initial cCRE registry is unlikely to be complete, especially for cell lineages underrepresented in our collection. The VISION resources can be useful for analysis of new data from users, such as searching for overlaps of the cCREs with peaks from new data sets. Also, parallel efforts, such as the Immunological Genome Project (Yoshida et al. 2019), are generating complementary resources that can expand the cCRE registry. 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 evaluating experimentally the impact of cCREs on gene expression will provide a more complete assessment of the quality of the registry.

Each cCRE has been annotated with its epigenetic state in each cell type and an initial estimate of the eRP score for regulating candidate target genes. These initial eRP scores for cCREs, derived from a multivariate regression and subselection procedure, can explain a substantial portion of variance in gene expression, but a considerable amount of expression variance remains unexplained. Estimates for regulatory output could be improved by incorporating transcription factor binding site motifs (Weirauch et al. 2014), transcription factor occupancy (Dogán et al. 2015), and patterns in multispecies genome sequence alignments (Taylor et al. 2006). The target gene assignments can be refined by inclusion of data on chromatin interaction frequencies, for example, by restricting cCRE–gene pairs to those within a 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.

The IDEAS segmentation results across cell types revealed some known transitions between states, such as poised enhancers in multilineage progenitor cells either shifting to active enhancers or losing their preactivation 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 nucleases, a substantial proportion became much less nuclease accessible while retaining CTCF occupancy in differentiated cells. The reduction in accessibility reflects a change in the chromatin structure to a more closed state, but unexpectedly, the CTCF protein remains bound. Initial studies suggested that the CTCF-bound-but-inaccessible sites were associated with repressed, gene-poor regions, whereas the CTCF-bound-and-accessible sites were enriched at constitutive TAD boundaries. However, further studies are needed to more fully investigate the functions of different categories of CTCF-bound sites.

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) and a decrease in the number of genes being expressed (Fig. 4D). Although this reduction in numbers of active genes and regulatory elements appears to occur in most lineages of blood cells, it was not observed in megakaryocytic cells, which retain aspects of the regulatory landscape and transcriptomes of multilineage progenitor cells. Similarity of MK to multilineage progenitor cells has been discerned previously from phenotypic similarities (Huang and Cantor 2009), transcriptome data (Sanjuan-Pla et al. 2013; Psaila et al. 2016), and global epigenetic profiles (Heuston et al. 2018). Recent studies have shown 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 conservative process differs from other lineage commitment and maturation processes that involved substantial changes to the epigenome and reduction in numbers of genes expressed.

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 along with predictions of and annotations of candidate regulatory elements. Similar systematic integration of epigenetic data in human blood cells is ongoing, which will generate equivalent resources. Such resources should provide guidance on many important problems, such as suggesting specific hypotheses for mechanisms by which genetic variants in noncoding regions can be associated with complex traits and diseases (Ulirsch et al. 2016; Bao et al. 2019).

Methods

Cell populations and sources of epigenomic and transcriptomic data

Detailed information about the cell populations and cell lines analyzed is in the [Supplemental Material, section 1](#). The ChIP-seq

and ATAC-seq procedures followed previously published methods (Wilson et al. 2010; Buenrostro et al. 2013; Wu et al. 2014; Heuston et al. 2018). Detailed information about the experimental methods, sources of data sets, bioinformatic pipelines, and quality assessments are in the [Supplemental Material, section 2](#), and the [Supplemental Tables](#).

Data normalization and comparison

A novel method for normalization, called S3norm (Xiang et al. 2020), was used to produce comparable peak signals without inflating background regions. This method is described in more detail in sections 3 and 5 of the [Supplemental Material](#), and the pipeline is deposited at GitHub (<https://github.com/guanjue/S3norm>). The methods for comparing epigenetic signals across cell types are described in [section 4](#) of the [Supplemental Material](#).

Integrative analysis and cCRE calls

The implementation of IDEAS (Zhang and Hardison 2017; Zhang et al. 2016) for the mouse hematopoietic cell data sets is described in the [Supplemental Material, section 6](#), and the software is available from GitHub (https://github.com/guanjue/IDEAS_2018). The method for calling cCREs is in the [Supplemental Material, section 8](#). The methods for comparing signals in peaks of nuclease sensitivity and in transcriptomes across cell types are in [section 10](#) of the [Supplemental Material](#).

Estimating impact of cCREs on candidate target genes

The methodology for estimating the output of individual cCREs based on their epigenetic states and correlations with expression of candidate target genes is presented in [section 16](#) of the [Supplemental Material](#).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE143271 and to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA599438. Data and servers for visualization also are available at the VISION Project website (<http://usevision.org>). All scripts for data analyses and data visualization can be found at GitHub (https://github.com/rosshardison/VISION_mouseHem_code) and as [Supplemental Code](#).

Competing interest statement

The authors declare no competing interests.

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, and NHGRI U54HG006998.

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. doi:10.1038/nbt.2153

- 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. doi:10.1182/blood-2014-01-548305
- An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, Li Q, Zhang Y. 2019. OnTAD: Hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol* **20**: 282. doi:10.1186/s13059-019-1893-y
- Bao EL, Cheng AN, Sankaran VG. 2019. The genetics of human hematopoiesis and its disruption in disease. *EMBO Mol Med* **11**: e10316. doi:10.15252/emmm.201910316
- Baron MH, Barminko J. 2016. Chromatin condensation and enucleation in red blood cells: an open question. *Dev Cell* **36**: 481–482. doi:10.1016/j.devcel.2016.02.014
- 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. doi:10.1038/nmeth.2688
- Cantor A, Orkin S. 2002. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**: 3368–3376. doi:10.1038/sj.onc.1205326
- 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. doi:10.1101/gr.098921
- 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. doi:10.1038/nature13985
- 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. doi:10.1038/ng.3646
- 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. doi:10.1186/s13072-015-0009-5
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. doi:10.1038/nature11247
- The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopedias of DNA elements in the human and mouse genomes. *Nature* (in press).
- Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817–825. doi:10.1038/nbt.1662
- Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216. doi:10.1038/nmeth.1906
- Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**: 364–376. doi:10.1038/nbt.3157
- 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. doi:10.1016/j.molcel.2009.11.001
- Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**: 587–594. doi:10.1038/nature08533
- 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. doi:10.1093/bioinformatics/bty575
- Haussler 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. doi:10.1093/nar/gky1095
- Hardison RC, Taylor J. 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**: 469–483. doi:10.1038/nrg3242
- 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. doi:10.1016/j.molcel.2010.05.004
- Heuston EF, Keller CA, Lichtenberg J, Giardine B, Anderson SM, NIH Intramural Sequencing Center, Hardison RC, Bodine DM. 2018. Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. *Epigenetics Chromatin* **11**: 22. doi:10.1186/s13072-018-0195-z
- Higgs DR. 2013. The molecular basis of α -thalassemia. *Cold Spring Harb Perspect Med* **3**: a011718. doi:10.1101/cshperspect.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. doi:10.1038/nmeth.1937
- 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. doi:10.1093/nar/gks1284
- 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.e7. doi:10.1016/j.molcel.2017.02.027
- Huang H, Cantor AB. 2009. Common features of megakaryocytes and hematopoietic stem cells: What's the connection? *J Cell Biochem* **107**: 857–864. doi:10.1002/jcb.22184
- Huang J, Liu X, Li D, Shao Z, Cao H, Zhang Y, Trompouki E, Bowman TV, Zou LI, Yuan GC, et al. 2016. Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Dev Cell* **36**: 9–23. doi:10.1016/j.devcel.2015.12.014
- Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that discriminate between overlapping regulatory annotations. *PLoS Comput Biol* **13**: e1005795. doi:10.1371/journal.pcbi.1005795
- 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. doi:10.1146/annurev.immunol.21.120601.141007
- 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. doi:10.1016/j.molcel.2011.12.021
- Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretzky 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. doi:10.1126/science.1256271
- Laurenti E, Göttgens B. 2018. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**: 418–426. doi:10.1038/nature25022
- Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* **152**: 1237–1251. doi:10.1016/j.cell.2013.02.014
- 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. doi:10.1038/s41592-018-0218-5
- Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* **20**: 1983–1992. doi:10.1109/TVCG.2014.2346248
- Ling T, Crispino JD. 2020. GATA1 mutations in red cell disorders. *IUBMB Life* **72**: 106–118. doi:10.1002/iub.2177
- Long HK, Prescott SL, Wysocka J. 2016. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**: 1170–1187. doi:10.1016/j.cell.2016.09.018
- 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 Rep* **27**: 3228–3240.e7. doi:10.1016/j.celrep.2019.05.046
- Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte 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. doi:10.1038/ncb2709
- Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG, Göttgens B. 2016. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**: e20–e31. doi:10.1182/blood-2016-05-716480
- Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* **11**: 1–23. doi:10.1146/annurev-genom-082509-141651
- Orkin SH, Zou LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**: 631–644. doi:10.1016/j.cell.2008.01.025
- 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. doi:10.1093/hmg/ddx306
- 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–e148. doi:10.1182/blood-2011-05-355107
- 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. doi:10.1101/gr.164178.113
- 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* **111**: 4103–4108. doi:10.1073/pnas.1401598111
- 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. doi:10.1186/s13059-016-0939-7
- Rekhtman N, Radparvar F, Evans T, Skoultschi AI. 1999. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev* **13**: 1398–1411. doi:10.1101/gad.13.11.1398
- Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* **414**: 105–111. doi:10.1038/35102167
- Roadmap Epigenomics Consortium, 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. doi:10.1038/nature14248
- Sánchez-Castillo M, Ruau D, Wilkinson AC, Ng FS, Hannah R, Diamanti E, Lombard P, Wilson NK, Göttgens B. 2015. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* **43**: D1117–D1123. doi:10.1093/nar/gku895
- 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. doi:10.1038/nature12495
- Schreiber J, Bilmes J, Noble WS. 2020. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol* (in press).
- Song J, Chen KC. 2015. Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol* **16**: 33. doi:10.1186/s13059-015-0598-0
- Stunnenberg HG, International Human Epigenome Consortium, Hirst M. 2016. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**: 1145–1149. doi:10.1016/j.cell.2016.11.007
- 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. doi:10.1074/jbc.M112.413260
- Sykes SM, Scadden DT. 2013. Modeling human hematopoietic stem cell biology in the mouse. *Semin Hematol* **50**: 92–100. doi:10.1053/j.seminhematol.2013.03.029
- 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. doi:10.1101/gr.4537706
- 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. doi:10.1016/j.devcel.2011.04.008
- Till JE, McCulloch EA. 1961. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res* **14**: 213–222. doi:10.2307/3570892
- 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. doi:10.1016/j.cell.2016.04.048
- 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. doi:10.1101/gr.5353806
- 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. doi:10.1186/s13059-018-1519-9
- 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. doi:10.1016/j.cell.2014.08.009
- 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. doi:10.1182/blood-2008-08-078220
- Wilson NK, Foster SD, Wang X, Knezevic K, Schütte 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. doi:10.1016/j.stem.2010.07.016
- Wilson NK, Schoenfelder S, Hannah R, Sánchez Castillo M, Schütte 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* **127**: e12–e23. doi:10.1182/blood-2015-10-677393
- 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–e138. doi:10.1182/blood-2011-03-341404
- 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. doi:10.1101/gr.125088.111
- 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. doi:10.1101/gr.164830.113
- Xiang G, Keller CA, Giardine B, An L, Li Q, Zhang Y, Hardison RC. 2020. S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *Nucleic Acids Res* (in press). doi:10.1093/nar/gkaa105
- 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.e20. doi:10.1016/j.cell.2018.12.036
- 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. doi:10.1016/j.molcel.2009.11.002
- 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. doi:10.1038/nature13992
- 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. doi:10.1093/nar/gkx659
- Zhang Y, Mahony S. 2019. Direct prediction of regulatory elements from partial data without imputation. *PLoS Comput Biol* **15**: e1007399. doi:10.1371/journal.pcbi.1007399
- Zhang P, Behre G, Pan J, Iwama A, Wara-Aswapati N, Radomska HS, Auron PE, Tenen DG, Sun Z. 1999. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc Natl Acad Sci* **96**: 8705–8710. doi:10.1073/pnas.96.15.8705
- Zhang Y, An L, Yue F, Hardison RC. 2016. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* **44**: 6721–6731. doi:10.1093/nar/gkw278
- Zheng X, Zheng Y. 2018. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics* **34**: 1568–1570. doi:10.1093/bioinformatics/btx802
- Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931–934. doi:10.1038/nmeth.3547

Received August 9, 2019; accepted in revised form February 21, 2020.



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

Guanjue Xiang, Cheryl A. Keller, Elisabeth Heuston, et al.

Genome Res. 2020 30: 472-484 originally published online March 4, 2020

Access the most recent version at doi:[10.1101/gr.255760.119](https://doi.org/10.1101/gr.255760.119)

Supplemental Material

<http://genome.cshlp.org/content/suppl/2020/03/18/gr.255760.119.DC1>

References

This article cites 80 articles, 18 of which can be accessed free at:
<http://genome.cshlp.org/content/30/3/472.full.html#ref-list-1>

Open Access

Freely available online through the *Genome Research* Open Access option.

Creative Commons License

This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>
