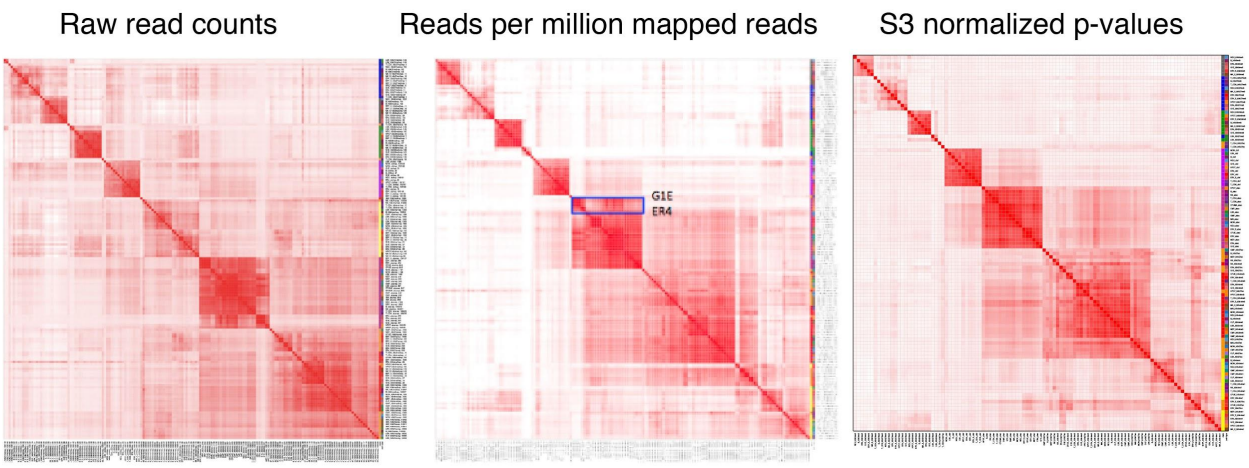
An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis

Supplementary Material

Guanjue Xiang, Cheryl A. Keller, Elisabeth Heuston, Belinda M. Giardine, Lin An, Alexander Q. Wixom, Amber Miller, April Cockburn, Jens Lichtenberg, Berthold Göttgens, Qunhua Li, David Bodine, Shaun Mahony, James Taylor, Gerd A. Blobel, Mitchell J. Weiss, Yong Cheng, Feng Yue, Jim Hughes, Douglas Higgs, Yu Zhang, Ross C. Hardison

Effectiveness of normalizations

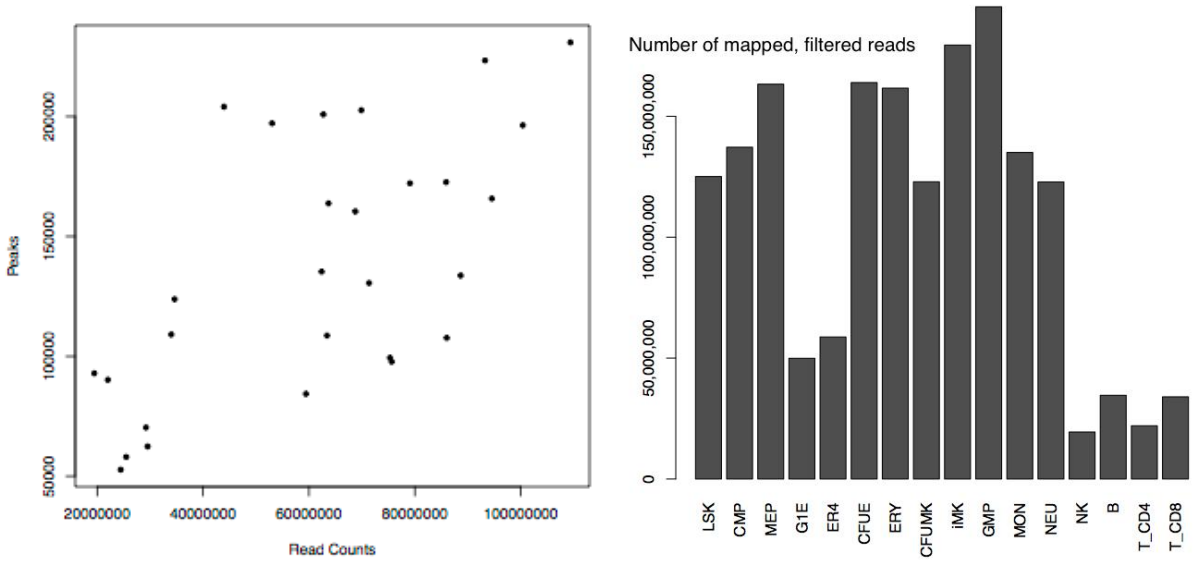
The correlation structure revealed after different types of normalization supports the effectiveness of the normalizations (**Supplementary Figure 1**). The correlation matrix for the initial signal (number of mapped reads per window) showed the clustering by the feature that was emphasized with the normalized data, but it also presented a heterodisperse pattern of off-diagonal correlations and substructure between different features. The off-diagonal correlations were reduced when normalized by sequencing depth, but some clustering within cell types with similar signal-to-noise ratio was observed. Utilization of S3norm to normalize for variation in both sequencing depth and signal-to-noise ratio removed much of the off-diagonal higher correlations, which indicates that with normalization by S3norm effectively removes much of the systematic biases in the data.



Supplementary Figure 1. Correlations across all features and cell types using raw data, data adjusted for sequencing depth, and data after normalization by S3norm.

Accuracy of ATAC-seq peak calls in progenitor cells

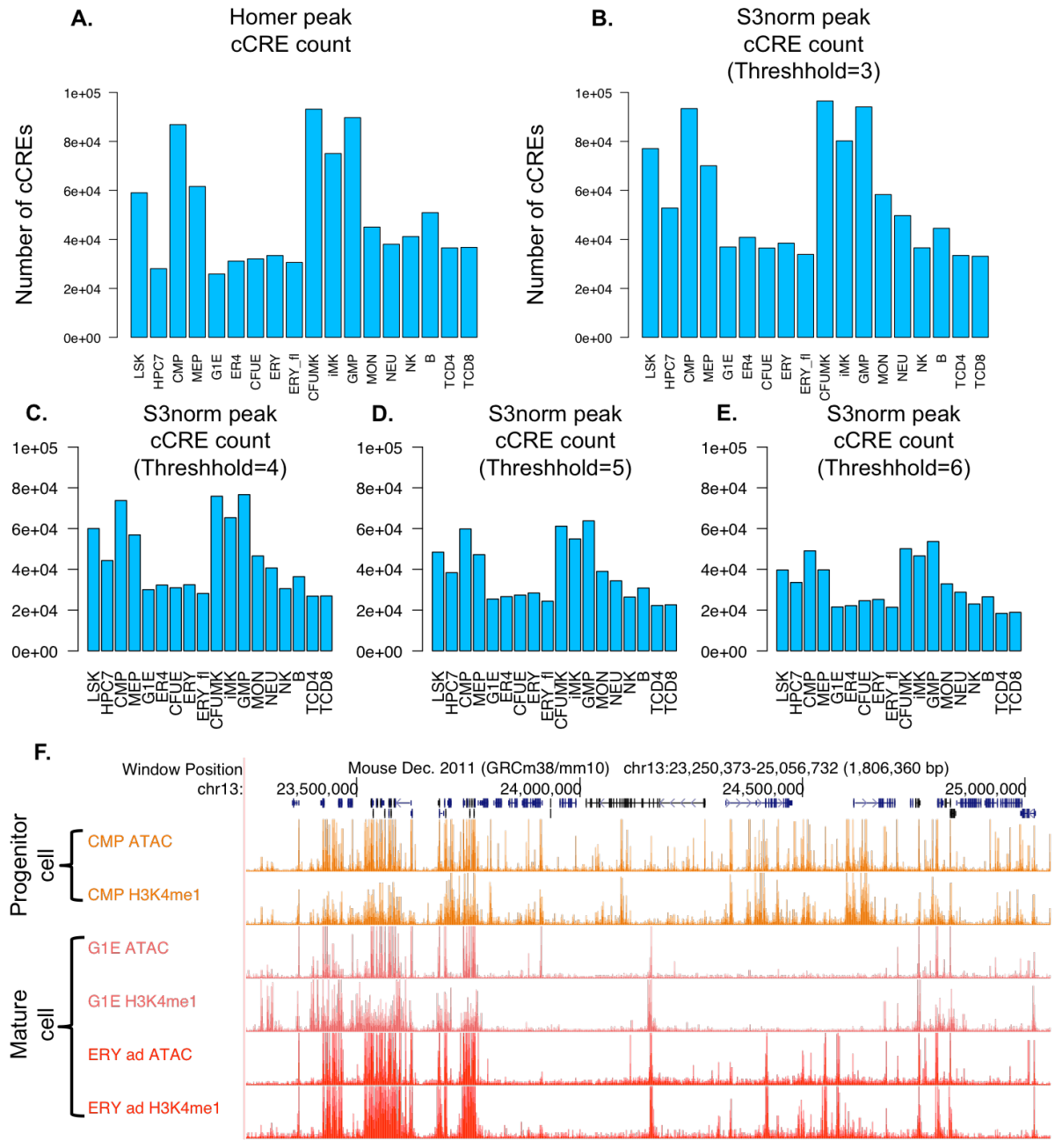
We examined whether differences in sequencing depth could explain the larger numbers of cCREs observed in progenitor and megakaryocytic cells. The number of ATAC-seq peaks was positively associated with sequencing depth (Supplementary Fig. 2A), measured as the number of mapped reads after filtering to remove reads mapping to the mitochondrial chromosome and blacklisted regions. The number of called peaks ranged widely for a given sequencing depth, but the association is positive. While the number of mapped, filtered reads were low for some monolineage cell types (e.g. the lymphoid cells), others such as CFUE, ERY, MON, and NEU were sequenced to a depth equivalent to that for the multilineage progenitor and megakaryocytic cells (Supplementary Fig. 2B). Thus, the differences in sequencing depth do not account fully for the trend observed for decreased numbers of cCREs during differentiation.



Supplementary Figure 2. Read counts and numbers of peak calls in ATAC-seq data. (A) Scatterplot of numbers of called peaks as a function of the numbers of mapped, filtered reads for each cell type. (B) Bar graph showing the numbers of mapped, filtered reads for each cell type.

The decrease in the number of ATAC-seq peaks across differentiation was also observed after normalization of the signals. The S3 normalization method (Xiang et al. 2019b) adjusts signals to account for differences both in sequencing depth and signal-to-noise ratio. Using a simple peak calling threshold on the ATAC-seq and DNase-seq data after S3 normalization generated a profile of peak numbers across the cell types that was similar to that obtained from the peak calls by Homer (Heinz et al. 2010) (Supplementary Figure 3). Specifically, the higher number of ATAC-seq peaks observed in progenitor and megakaryocytic cells was still observed after normalization for differences in sequencing depth and signal-to-noise ratio. The same result was seen over a range of choices for threshold for peaks. Thus, the observed higher numbers of peaks were robust both to normalization for sequencing depth and signal to noise ratio and to changes threshold settings. Examining a histone gene locus as an illustrative example, the ATAC-seq signal track of one progenitor cell (CMP) and two mature cells (G1E and ERY)

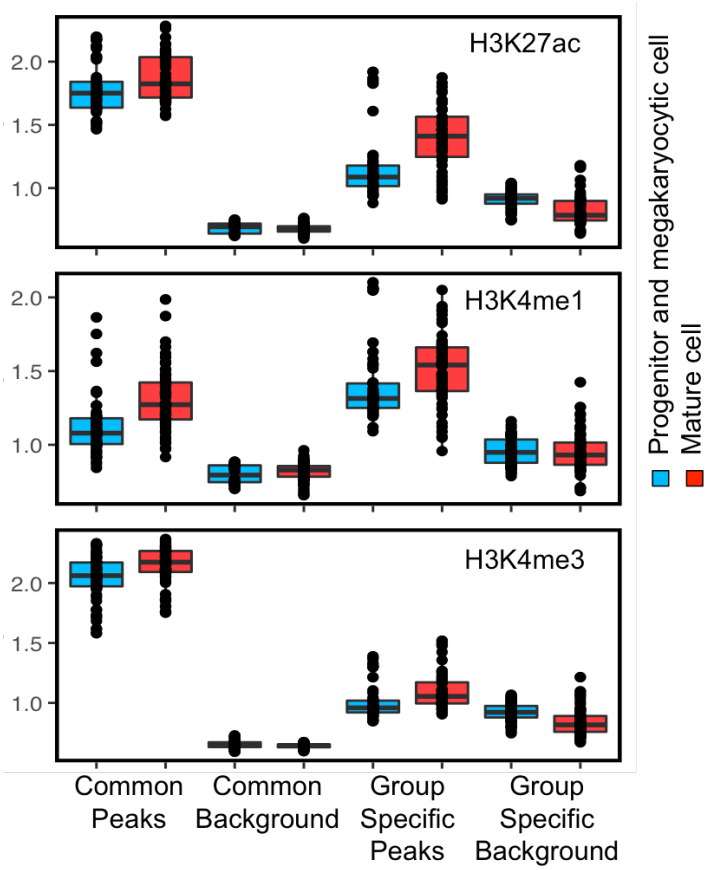
showed more ATAC-seq peaks in the progenitor cell (Supplementary Figure 3F). The signal of these ATAC-seq peaks was also consistent with the H3K4me1 ChIP-seq signal.



Supplementary Figure 3. The ATAC-seq peak number in the hematopoietic cell types. (**A**) Panel **A** plots numbers of ATAC-seq peaks called by Homer. (**B**) Panel **B** plots the numbers of the ATAC-seq peaks called by setting a threshold on the S3norm signal. The DNA intervals with S3norm greater than 3 are

called as peaks. (**C**-**E**) The panels **C**-**E** are the peak numbers of the ATAC-seq peaks called by using different S3norm signal thresholds. (**F**) The signal track of ATAC-seq and H3K4me1 ChIP-seq in progenitor cell (CMP) and mature cells (G1E and ERY).

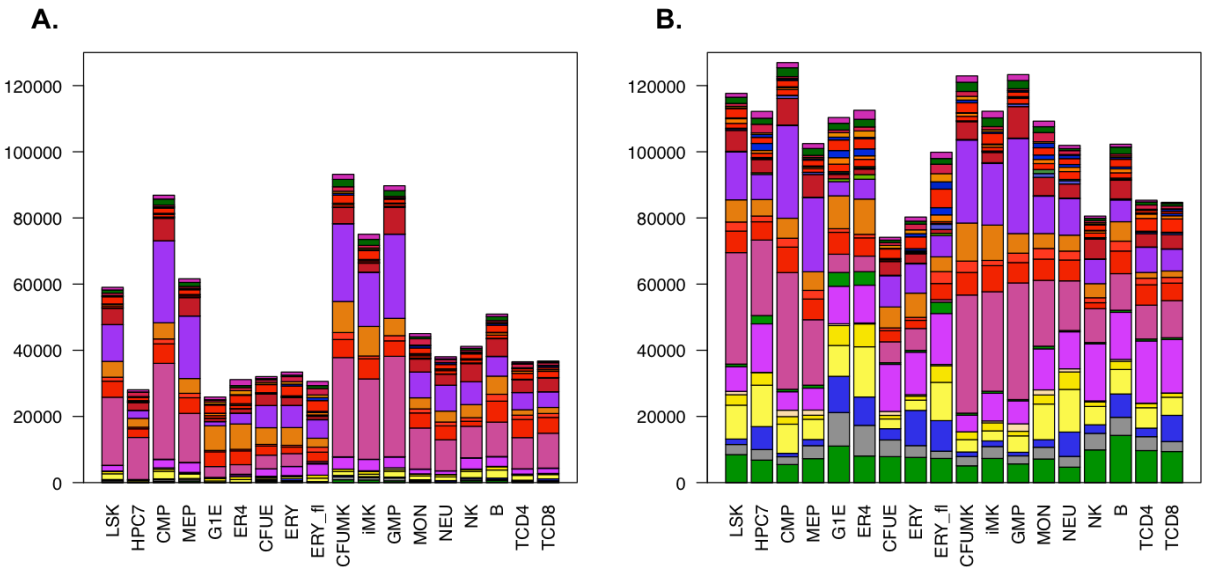
To evaluate more broadly the accuracy of ATAC-seq and DNase-seq peak calls in progenitor and megakaryocytic cells versus mature cells, we leveraged the information about histone modifications in these cell types. We posited that correct calls for peaks in nuclease sensitivity should be enriched in histone modifications associated with gene activation (H3K4me1, H3K4me3, and H3K27ac), and therefore incorrect ATAC-seq peaks (false positives) would be less likely to have histone modifications. The enrichment for such histone modifications was computed in nuclease sensitive peaks between all pairs of cell types, where each member of a pair came from either the (a) progenitor and megakaryocytic cell type group or (b) the mature cell type group (Supplementary Figure 4). The expected enrichment for H3K4me1, H3K27ac, and H3K4me3 was observed in peaks common to each cell type pair but not in DNA intervals that not peaks in either member of a pair (but they are ATAC-seq peaks in at least one other cell type). Importantly, the enrichment for the diagnostic histone modifications was also seen in the nuclease sensitive peaks that are specifically found in the progenitor and megakaryocytic cells but not in the mature cell type for each pair. A similar pattern was seen when the signal strength is used to evaluate the peaks and non-peaks. *These data reject the hypothesis that the nuclease sensitive peaks specifically in progenitor and megakaryocytic cells are false positives.* The lower enrichment for H3K4me3 in group-specific ATAC-seq peaks (progenitor and megakaryocytic cells vs. mature cells) suggests that these are less likely to be promoters. The increase in enrichment of H3K27ac in group-specific peaks in mature cells is consistent with the activation of many poised enhancers during maturation.



Supplementary Figure 4. The enrichment for histone modification (H3K4me1, H3K27ac, H3K4me3) in ATAC-seq peaks. Each box plot summarizes the results of enrichment calculations for histone modifications in the ATAC-seq peaks in groups determined for pairs of cell types, where one member of the pair is in the progenitor and megakaryocytic cell type group (LSK, HPC7, CMP, MEP, CFUE, CFUMK, iMK, GMP) (blue box-plot) and the other member of the pair is in the mature cell type group (G1E, ER4, ERY, ERY_fl, MON, NEU, NK, B, T_CD4, T_CD8) (red box-plot). The ATAC-seq peaks were grouped based on their presence or absence in each cell type of the cell type pair. The 1st and the 2nd box-plots summarize the enrichment for histone modification in the peaks common to both cell types in each pair. The 3rd and the 4th box are the enrichment for histone modification in the cCRE DNA intervals that are not ATAC-seq peaks in the two cell types for each pairwise comparison (common non-peak regions). The 5th is the enrichment for histone modification in the progenitor and megakaryocytic cell group specific-peak regions. The 6th is the enrichment for histone modification in the mature cell type group specific-peak regions. The 7th is the enrichment for histone modification in the progenitor and megakaryocytic cell

type group specific-non-peak regions. The 8th is the enrichment for histone modification in the mature cell type group specific-non-peak regions.

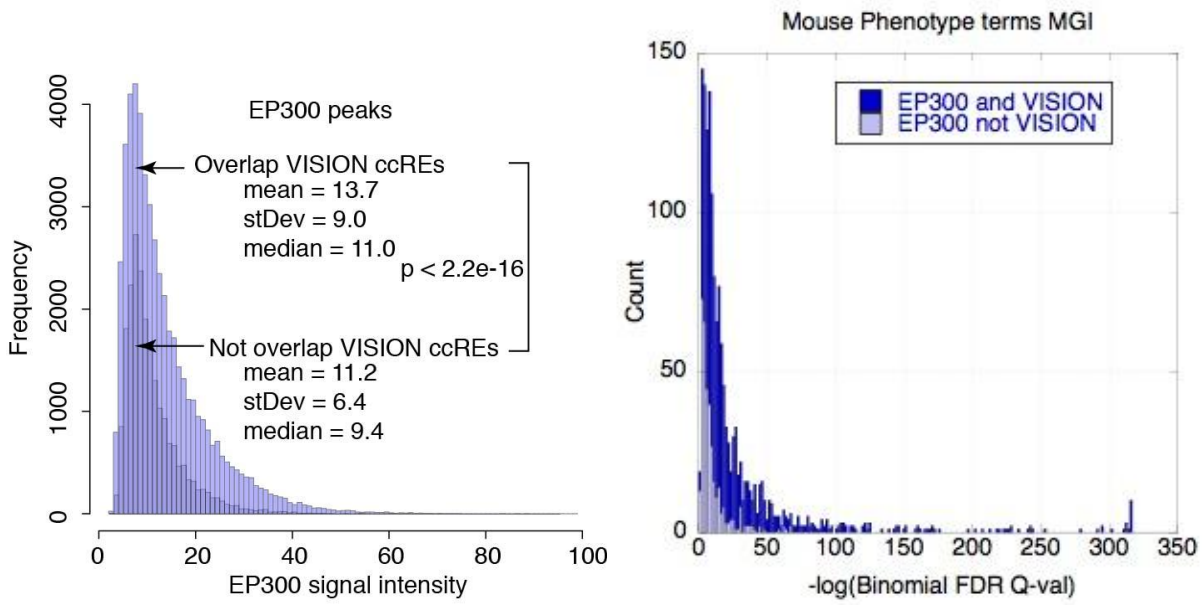
Not only were there fewer ATAC-seq peaks and active cCREs in mature cells, but there were fewer active regions among the cCRE DNA intervals in mature cells (Supplementary Figure 5). The DNA intervals called as a cCRE in any cell type were assigned to their IDEAS epigenetic state in each cell type, without requiring that each cCRE DNA interval also be an ATAC-seq peak or DNase-seq peak. As expected, more cCREs were assigned to a state, but the number of DNA intervals in activity-related states remained higher in progenitor and megakaryocytic cells than in mature cells. Specifically, the group of progenitor and megakaryocytic cell types still had more active states compared with the other mature cells (Supplementary Figure 5B). For the mature (non-iMK) cell types, many of the cCRE intervals were in a quiescent state, a repressed state, the H3K4me1 only state that could reflect a memory of a previously active enhancer, or the H3K36me3 only state associated with transcription (Supplementary Figure 5B).



Supplementary Figure 5. Numbers of cCRE DNA intervals in each IDEAS state at different schema for inclusion. The ~205,000 cCRE DNA regions (with ATAC-seq peaks determined by thresholding on the tracks after S3norm) were assigned to IDEAS states in each cell type using two different rules for inclusion, and the resulting numbers in each state are shown in the bar graphs. **A.** Each cCRE DNA interval was also an ATAC-seq peak in the indicated cell type. **B.** The cCRE DNA interval did not have to be an ATAC-seq peak in the indicated cell type, but the cCRE DNA intervals in the quiescent state were not counted.

EP300 peaks that do or do not overlap with VISION cCREs

While the overlap of the cCRE datasets with the collection of EP300 peaks supported the quality of those datasets, no set of cCREs captured all the EP300 peaks. This lack of full overlap raises the question of whether the EP300 peaks over-estimated the cCREs or the cCRE sets were missing regulatory elements. We examined the EP300 peaks that did or did not overlap with VISION cCREs for features that could distinguish the two groups and thus may shed some light on this issue. The signal strength of the EP300 peaks had a similar distribution for both the set that overlaps with VISION cCREs and the set that does not overlap. However, the set that overlaps VISION cCREs had a significant trend toward higher signals than did the peaks that do not overlap (**Supplementary Figure 6**). Thus we concluded that the VISION cCREs tended to capture the stronger EP300 peaks.



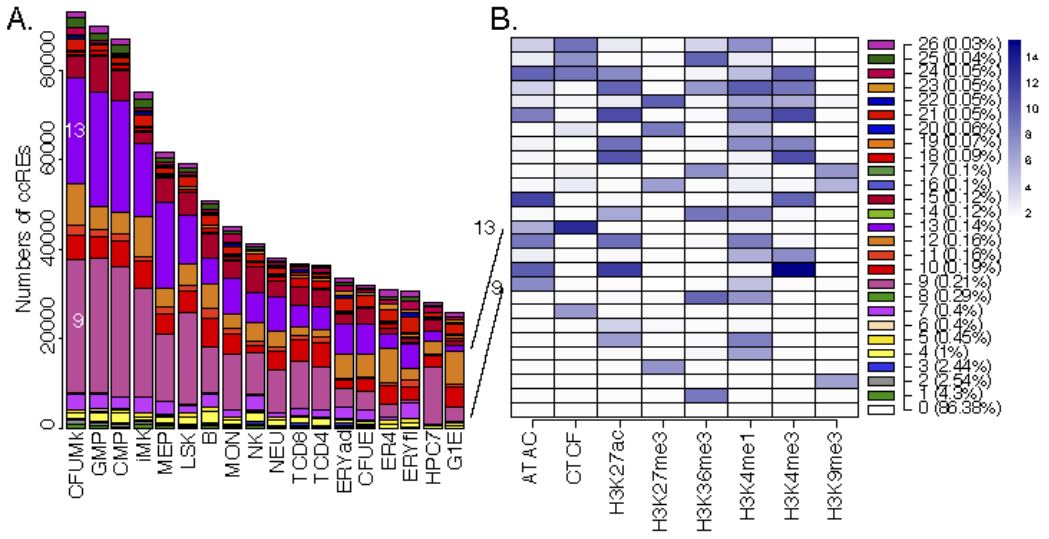
Supplementary Figure 6. Distributions of signal intensity (*left*) and enrichment Q-values for mouse phenotype terms (*right*) for EP300 peaks that did or did not overlap with VISION cCREs.

Both sets of EP300 peaks were enriched for expected functional terms in various ontologies, but the set that overlapped with VISION cCREs was enriched in more terms with a greater level of significance. A subset of 10,000 EP300 peaks were randomly chosen from each set (overlapping VISION cCREs or not), and analyzed for functional term enrichment using the GREAT tool (McLean et al. 2010) Focusing on Mouse Phenotype (MGI) and MSigDB Pathways, lists of functional terms with hundreds to over a thousand terms relevant to hematopoiesis were enriched in both sets of EP300 peaks. However, the EP300 peaks that overlapped with VISION cCREs returned more terms with lower FDR Q-values when compared to the EP300 peaks not overlapping VISION. Using Mouse Phenotype as an example, peaks common to EP300 and VISION returned 1138 terms, many with extremely low Q-values, whereas the peaks only in EP300 returned 361 terms with higher, but still significant, Q-values. These distributions were significantly different (mean for EP300 peaks overlapping VISION cCREs was 33.2, mean for peaks not in VISION cCREs was 10.0; p-value<0.001 for both Student's *t*-test and Wilcoxon

test). These indicators of higher significance suggest that the EP300 peaks overlapping with VISION cCREs may be more intimately involved in hematopoietic regulation than those that do not overlap.

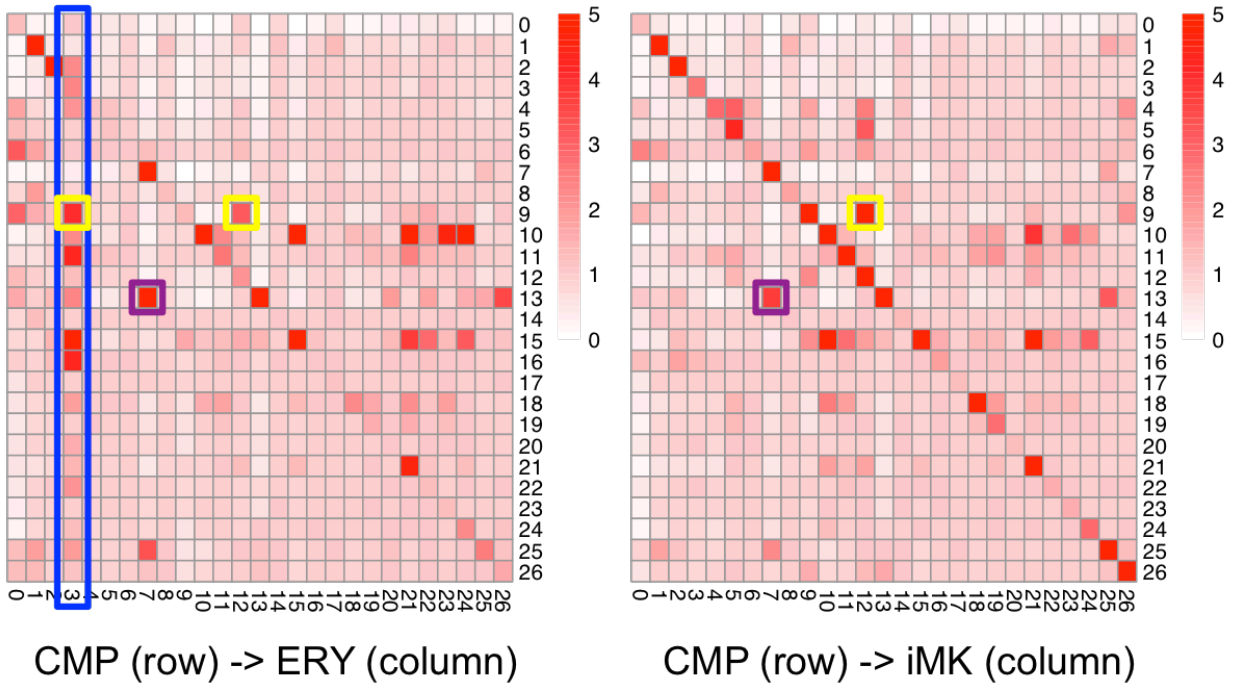
Transitions in epigenetic states of cCREs between cell types

While the numbers of active cCREs tended to decrease during commitment and maturation of lineages (except iMK), the reduction was particularly pronounced for cCREs in a poised enhancer mode (state 9 EN) or in a CTCF-bound nuclease accessible state (state 13 CN) (**Supplementary Figure 7**).



Supplementary Figure 7. Transitions in epigenetic states at cCREs across hematopoietic differentiation. **A.** The numbers of cCREs in each cell type are colored by their IDEAS epigenetic state, in numerical order from bottom to top of each bar. **B.** The composition of each state (heatmap in shades of blue) and the fraction of genome assigned are shown with the states in numerical order to facilitate comparison with panel **A**. States 9 and 13, which are prominent in multilineage progenitor cells, are emphasized.

We then determined the states into which these cCREs tended to transition by computing the enrichment for each state transition between all pairs of cell types, as illustrated for state transitions in cCREs for comparison between CMP and ERY (**Supplementary Figure 8A**) and between CMP and iMK (**Supplementary Figure 8B**). This examination of all state transitions in cCREs between pairs of cells revealed that the decrease in cCREs in state 9 (EN) occurred both through conversion of the cCRE to a different state and via a loss of accessibility (state 0). More specifically, cCREs in the poised enhancer state 9 (EN) in CMP tended to transition either to the active enhancer-like state 12, the polycomb-repressed state 3, or the low signal quiescent state 0 in erythroid cells (**Supplementary Figure 8A**). In contrast, those CMP state 9 cCREs transitioned most frequently to state 12 (active enhancer) in iMK, with less enrichment for transitioning to the quiescent state 0 and almost no enrichment for transitioning to the repressed polycomb state (**Supplementary Figure 8B**). Notably, cCREs in several different states in CMP were enriched for transitions to the polycomb state 3 in ERY (vertical blue box in **Supplementary Figure 8A**). These results illustrate specific mechanisms for the recent report of more substantial changes in epigenomic landscape during differentiation of CMP to ERY than to iMK (Heuston et al. 2018).



Supplementary Figure 8. Transitions between IDEAS epigenetic states for cCREs in CMP after differentiation to ERY (A) or iMK (B). The numbers of cCREs in all state transition pairs were determined, and the enrichment was calculated as observed numbers over those expected given the numbers of cCREs in states across the whole genome in all cell types. The intensity of the red color in each cell reflects the level of enrichment. Boxes around cells emphasize transitions in states; yellow and purple for transitions from state 9 and state 13, respectively, in CMP, and blue for transitions to state 3 in ERY.

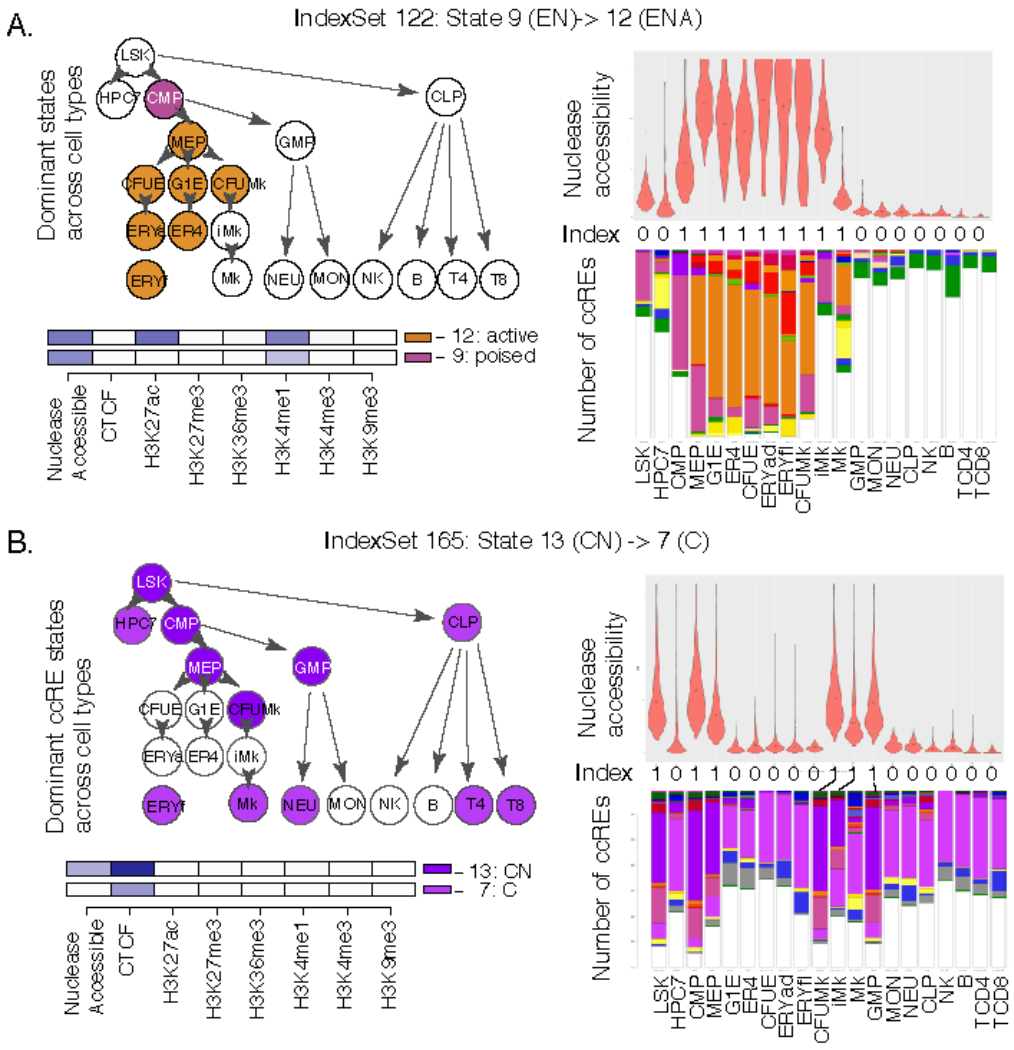
Surprisingly, for another major state in progenitor and megakaryocytic cells, much of the decrease in numbers of cCREs in state 13 (CTCF and nuclease accessible) occurred through a loss of accessibility while retaining occupancy by CTCF (state 7, **Supplementary Figure 8**). This transition from state 13 to state 7 was observed for cCREs overall during differentiation to both ERY and iMK (**Supplementary Figure 8A and B**).

Examination of state transitions in discrete groups of cCREs with the same pattern of appearance across cell types

A complementary approach to studying state transitions examined the transitions within well-defined discrete groups of cCREs, which were clustered by their appearance pattern across cell types. The clusters were generated by an indexing approach that gives discrete categories of presence (1) or absence (0) of a nuclease accessible peak call in each cell type (Xiang et al., 2019 Snapshot). Thus, each cCRE had an 18-character index string, and cCREs with identical indices were placed in a group termed an index set. This clustering by discrete presence-*vs.*-absence calls across cell types gave insights into the history and progression of cCRE states.

The decrease in cCREs in state 9 (EN) occurred both through conversion of the cCRE to a different state and via a loss of accessibility. The former was exemplified by index set 122 (**Supplementary Figure 9A**), in which the poised enhancer-like state 9 became an active enhancer-like state 12 in erythroid cells. Index set 207 showed a loss of nuclease accessibility in maturing cells (**Supplementary Figure 10**).

Another major state in progenitor and megakaryocytic cells is state 13 (CTCF and nuclease accessible), and the decrease in numbers of cCREs occurred through a loss of accessibility while retaining occupancy by CTCF (state 7, **Supplementary Figure 8**). This transition from state 13 to state 7 was observed in specific index sets, such as index set 165 (**Supplementary Figure 9B**).

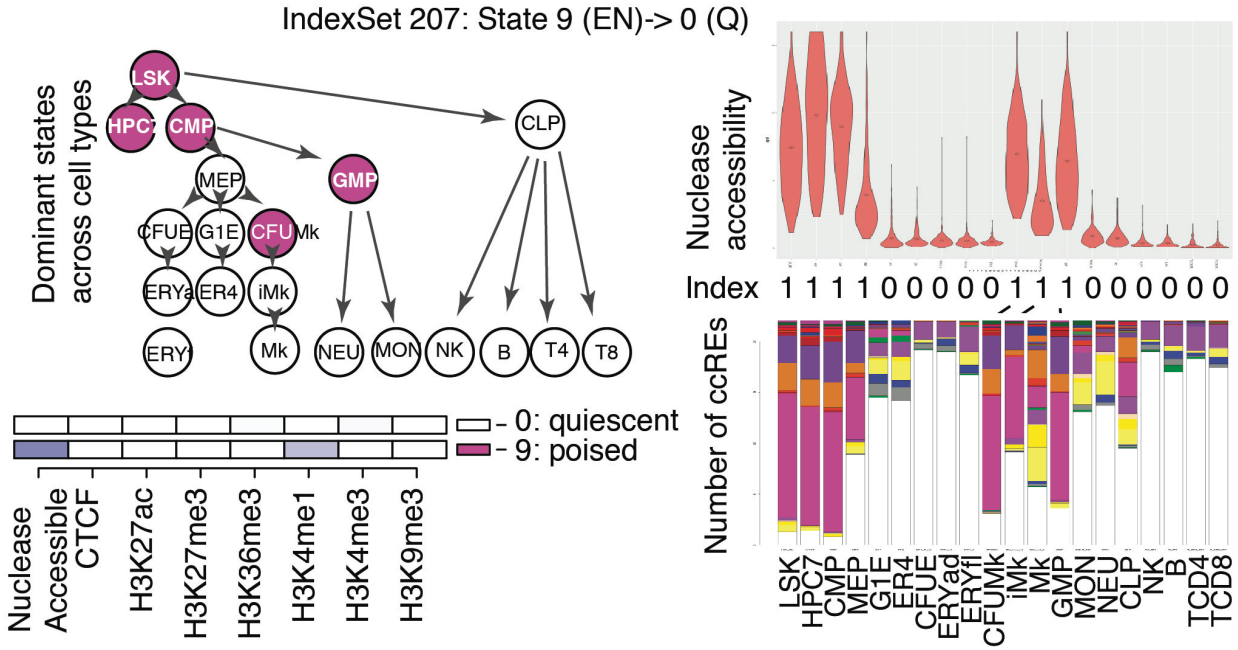


Supplementary Figure 9. State transitions for groups of cCREs related by their patterns of presence or absence across cell types. **A.** The state transitions and changes in nuclease accessibility across cell types are shown for the cCREs in index set 122, which undergo a transition from state 9 (poised enhancer) to state 12 (active enhancer). The index, which is a vector of presence or absence calls, is shown on the *right*, between the violin plot of nuclease accessibility and the bar plot for numbers of cCREs colored by their epigenetic states. The dominant state for this group of cCREs in each cell type is shown by the coloring of the circles in the differentiation tree on the *left*. The state compositions for the initial and final states are shown on the lower *left*. **B.** Results as in panel **A** for the cCREs in index set 13, which are CTCF-bound sites that undergo a transition from nuclease accessible to inaccessible. The results in panels **A** and **B** are from the Snapshot tool (Xiang et al. 2019a). Similar visualizations for all

index sets can be viewed and downloaded both via our VISION website (http://usevision.org) or from GitHub (https://github.com/guanjue/vision_index_set).

Fates of cCREs that are in a poised enhancer state in progenitor cells

A large fraction of cCREs in progenitor cells were in state 9 (EN, similar to a poised enhancer). Some of them transitioned to a state associated with active enhancers, exemplified by cCREs in index set 122 (Fig. 7B, main text). Another major trend for the state 9 cCREs was a loss of histone modifications and nuclease sensitivity to transition to state 0 (quiescent; Supplementary Figure 8). This transition is exemplified by cCREs in index set 207 (Supplementary Figure 10).



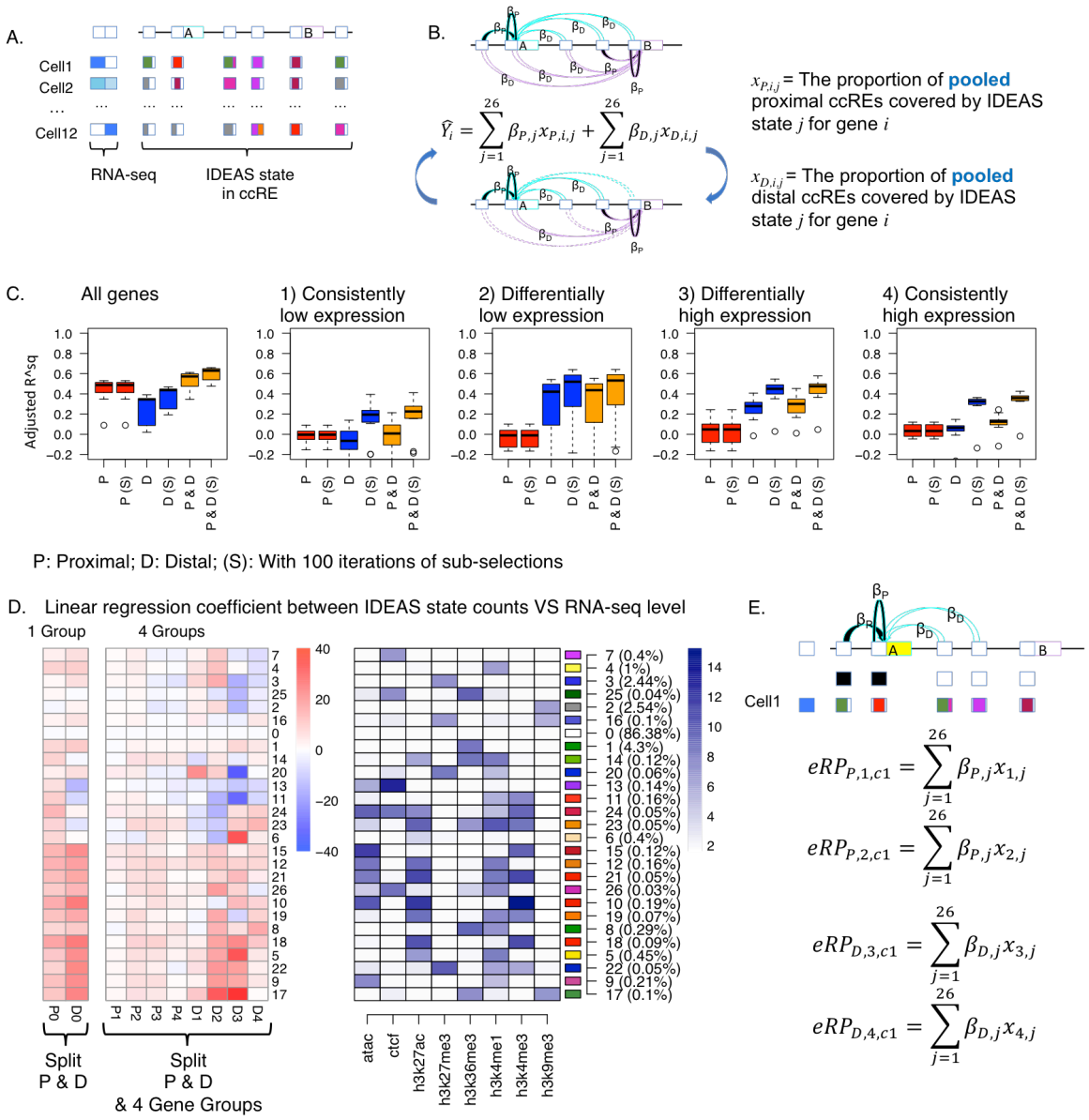
Supplementary Figure 10. Transition from epigenetic state 9 (EN) to state 0 (quiescent). The state transitions and changes in nuclease accessibility across cell types are shown for the cCREs in index set 207. The index, which is a vector of presence or absence calls, is shown on the *right*, between the violin plots of nuclease accessibility and the bar plot for numbers of cCREs colored by their epigenetic states. The dominant state for this group of cCREs in each cell type is shown by the coloring of the circles in the

differentiation tree on the *left*. The state compositions for the initial and final states are shown on the lower *left*. These results are from the Snapshot tool (Xiang et al. 2019a).

Method for calculating epigenetic Regulatory Potential (eRP) scores

This section presents **Supplementary Figure 11**, which has more detailed information on the results of the multivariate regression modeling to estimate the impact of epigenetic states and individual cCREs on expression of potential target genes. A summary of the method is given in the Results, and a detailed description is in the Materials and Methods. Very briefly, we start with our catalog of cCREs, including their epigenetic states across cell types, and expression levels of genes in 12 cell types, specifically the ones in which RNA-seq was done using the same protocol in the same laboratories (**Supplementary Figure 11A**). We then use a multivariate regression approach to relate RNA-seq levels with epigenetic states of cCREs around each gene. The proportion of base pairs in the pooled cCREs in each state for a gene was used as the predictor variable for each state. We hypothesized epigenetic states of cCREs in the proximal cCREs and distal cCREs have different effects on regulating the gene expression. We thus treated them as two distinct components in the model. We also hypothesized the contributions of cCREs on different types of genes are different. We thus trained the multivariate regression models were trained using all genes or genes in four expression categories based on their average expression levels and the variance of expression level across different cell types, specifically those with (1) consistently low, (2) differentially low, (3) differentially high, and (4) consistently high expression across cell types. Within a certain window, there can be a large amount of cCREs. However, it is unlikely that all of them can contribute to the regulation of a specific gene. Thus, a sub-selection iterative routine was used to remove cCREs that contribute little to explaining expression (**Supplementary Figure 11B**). The prediction accuracy in testing data after the sub-selection increased in all of the models (**Supplementary Figure 11C**). It indicates the sub-selection step effectively reduced the

overfitting issue in the models. Each regression coefficient, beta, from this method estimates the contribution of a specific state to gene expression (**Supplementary Figure 11D**). The coefficients were computed using all genes or genes in four expression categories. The contribution of individual cCREs to expression was estimated as a weighted sum of the regression coefficients for the component states, termed an epigenetic Regulatory Potential (eRP) score for each cCRE (**Supplementary Figure 11E**). Each cCRE has a different eRP score for each potential target gene in each category in each cell type.



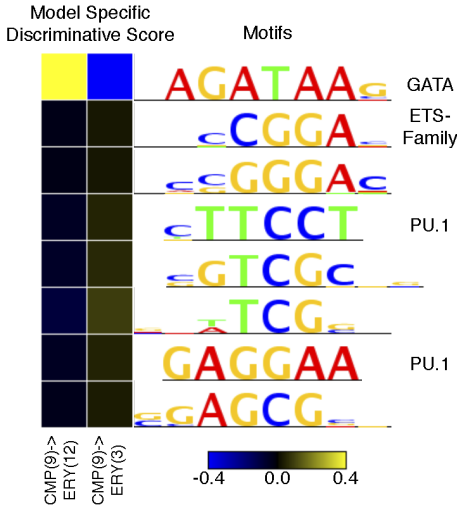
Supplementary Figure 11. Initial estimates of regulatory output and target gene prediction using regression models of IDEASs states in cCREs versus gene expression. **A.** Illustration of cCREs around two potential target genes, showing expression profiles of the genes across cell types (shades of blue, *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.** Diagram illustrating the 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 (indicated by dotted lines for omitted and solid lines for included cCREs in the lower diagram). **C.**

Ability of eRP scores of cCREs to explain levels of expression on chr1-chr19 and chrX in the twelve cell types WITH and WITHOUT (S) sub-selection for all genes and (1-4) in the four categories of genes. 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.** Values of the regression coefficients beta for each epigenetic state for proximal and distal cCREs. 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). These are aligned with a heatmap for the composition of each state, shown as shades of blue. The regressions were conducted for all genes (left three columns) or with genes in four distinct expression groups. **E.** Diagram illustrating the calculation of epigenetic Regulatory Potential (eRP) scores for cCREs as weighted sums of regression coefficients for the states covering each cCRE in each cell type.

Discriminative motif analysis for cCREs with different epigenetic state transitions

In CMP cell, we observed the cCREs with the same poised enhancer state (state 9 EN) can change to different states such as active enhancer state (state 12 ENA) and polycomb-repressed state (state 3 Pc) in ERY cell. To explore the potential factors associated with these different state transitions, we used a machine learning method called SeqUnwinder to analyzed the difference of sequence features in these cCREs (Kakumanu et al. 2017). Specifically, we first extracted the cCREs with poised enhancer state in CMP cell. The cCREs that become active enhancer state in ERY cell were clustered into the first group. The others that become polycomb state in ERY cell were clustered into the second group. We then applied the SeqUnwinder with the default setting to identify the motifs that can distinguish those two groups of cCREs. The results were shown in **Supplementary Figure 12.** Eight DNA binding motifs which include GATA motif, PU.1 motif, and motif of ETS transcription factor family were identified as the discriminative motifs between the two groups of cCREs. The GATA motif has a higher discriminative score for a certain for the first group of cCREs that transition from poised enhancer state to active enhancer state. While the PU.1 motif and the motif of ETS transcription

factor family have higher discriminative scores for the second group of cCREs that transition from poised enhancer state to polycomb state. These results suggested the different binding pattern of these transcription factors could be associated with epigenetic history of poised enhancer during cell differentiation.



Supplementary Figure 12. Discriminative motif analysis of cCREs that transition from poised enhancers state (state 9 EN) in CMP cell to active enhancer state (state 12 ENA) or polycomb-repressed state (state 3 Pc) in ERY cell. The discriminative motifs from SeqUnwinder are shown on the right side of the figure. The heatmap on the left shows the discriminative scores of each motif. The cCRE with the motif that has a higher discriminative score for a certain cCRE group are more likely to be classified as members of that cCRE group.

References

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.

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

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495-501.

Xiang G, Giardine B, An L, Sun C, Keller CA, Heuston E, Bodine DM, Hardison RC, Zhang Y. 2019a. Snapshot: clustering and visualizing epigenetic history during cell differentiation. doi:https://doi.org/10.1101/291880.

Xiang G, Keller CA, Giardine B, An L, Hardison RC, Zhang Y. 2019b. 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.