nature genetics

Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation

Alvaro J González[1,2], Manu Setty[1,2] & Christina S Leslie[1]

We carried out an integrative analysis of enhancer landscape and gene expression dynamics during hematopoietic differentiation using DNase-seq, histone mark ChIP-seq and RNA sequencing to model how the early establishment of enhancers and regulatory locus complexity govern gene expression changes at cell state transitions. We found that high-complexity genes—those with a large total number of DNase-mapped enhancers across the lineage—differ architecturally and functionally from low-complexity genes, achieve larger expression changes and are enriched for both cell type–specific and transition enhancers, which are established in hematopoietic stem and progenitor cells and maintained in one differentiated cell fate but lost in others. We then developed a quantitative model to accurately predict gene expression changes from the DNA sequence content and lineage history of active enhancers. Our method suggests a new mechanistic role for PU.1 at transition peaks during B cell specification and can be used to correct assignments of enhancers to genes.

Genome-scale studies of cellular differentiation have observed that many enhancers involved in cell type–specific programs are already established in precursor cells. For example, we recently found that most enhancers involved in the regulatory T (T_{reg}) cell transcriptional program, identified on the basis of their occupancy by the T_{reg} cell master regulator Foxp3, were DNase accessible in $CD4^+$ precursor cells, where they were occupied by other factors that 'place hold' to maintain the potential for T_{reg} cell differentiation[1]. Evidence in support of early enhancer establishment or chromatin poising has also been documented in B cell and macrophage specification[2,3], T cell development[4], early hematopoiesis[5] and commitment of multipotent endoderm cells to liver and pancreas cell fates[6]. Earlier concepts of poising include bivalent domains in embryonic stem cells (ESCs), where the active mark trimethylation of histone H3 at lysine 4 (H3K4me3) and repressive mark trimethylation of histone H3 at lysine 27 (H3K27me3) coincide[7], other poised ESC elements are marked by monomethylation of histone H3 at lysine 4 (H3K4me1) and H3K27me3 (ref. 8), and poised but inactive enhancers are marked by monomethylation of histone H3 at lysine 4 (H3K4me1) but not acetylation of histone H3 at lysine 27 (H3K27ac)[9].

Meanwhile, recent studies have defined the notion of cell type–specific 'super-enhancers'—spatially clustered enhancers occupied by master regulator transcription factors for the cell type—that regulate developmentally important genes[10,11]. Others have used segmentation of histone mark data to identify long (>3-kb) 'stretch enhancers' (ref. 12), associated wide domains of the active mark H3K27ac with high regulatory potential[13] or characterized broad domains of H3K4me3 as 'buffer domains' for important cell type–specific genes[14].

Here we introduce a new definition of regulatory locus complexity based on the multiplicity of DNase I–hypersensitive sites (DHSs) regulating a gene across a lineage. We investigate how locus complexity and early enhancer establishment during hematopoietic differentiation work together to shape transcriptional programs and quantitatively determine gene expression changes during cell state transitions. Through an integrative DHS-centric analysis of chromatin state and gene expression across ESCs and five primary hematopoietic cell types and predictive modeling of gene expression changes in cell fate specification, we propose that both regulatory complexity and early enhancer establishment contribute to achieving large expression changes during differentiation and strong cell type–specific expression patterns for important cell identity genes.

RESULTS

A lineage DHS atlas defines gene regulatory complexity

We carried out an integrative analysis of DHS sequencing (DNase-seq) data, histone modification chromatin immunoprecipitation and sequencing (ChIP-seq) data for multiple marks (H3K27ac, H3K27me3, H3K4me1 and H3K4me3) and RNA sequencing (RNA-seq) data to link enhancer dynamics and spatial organization to gene expression changes during hematopoietic differentiation. We focused on six cell types characterized by the Roadmap Epigenomics project[15–17] (**Supplementary Table 1**): human embryonic stem cells (hESCs), hematopoietic stem and progenitor cells ($CD34^+$ HSPCs), one myeloid cell type ($CD14^+$ monocytes) and three lymphoid lineages ($CD19^+$ B cells, $CD3^+$ T cells and $CD56^+$ natural killer (NK) cells). We first performed peak calling on DNase-seq profiles, using three biological replicates per cell type to control for irreproducible discovery rate (IDR)[18], and assembled an atlas of over 120,000 reproducible DHSs (median width = 456 bp; Online Methods and

[1]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to C.S.L. (cleslie@cbio.mskcc.org).

Received 26 February; accepted 19 August; published online 21 September 2015; doi:10.1038/ng.3402

Supplementary Fig. 1). We initially assigned each DHS in the atlas to the nearest gene, and we defined the regulatory locus complexity of a gene as the total number of DHSs, over all cell types, assigned to it in the atlas. Enhancer assignment to nearest genes can be prone to errors, especially in gene-dense regions or, conversely, for distal intergenic enhancers. However, 58% of DHSs in the atlas reside within the transcription unit of their assigned target gene (from 2 kb upstream of the transcription start site (TSS) to 2 kb downstream of the last annotated 3′ end), and an additional 10% lie within 10 kb of the target gene. Therefore, in a majority of cases, we assigned the DHS to the encompassing or local gene.

Indeed, roughly half of the non-promoter DHSs in the atlas, as well as in individual cell types, are intronic (**Supplementary Table 2**). Several studies have characterized conserved intronic enhancers that reside in developmentally important lymphoid lineage genes, including *Foxp3* and *Ikzf1*, and demonstrated that they indeed regulate the genes by which they are encompassed[19,20]. Furthermore, recent high-resolution promoter capture Hi-C (genome-wide chromosome conformation capture) analysis found that transcriptionally active promoters asymmetrically interact with the gene body more than with local upstream sequence, suggesting an activating role for intronic enhancers[21]. However, intronic enhancers sometimes regulate a nearby gene rather than the gene encompassing them[22–24], and our assignment of intronic DHSs is therefore imperfect. We return to the problem of improving gene assignments for enhancers later.

We next grouped genes into three equally sized classes (tertiles) on the basis of their regulatory complexity in the atlas: low-complexity genes, with a total assignment of zero to two DHSs; medium-complexity genes, with three to seven DHSs; and high-complexity genes, with eight or more DHSs. Examples of a low-complexity gene (*NUP54*, a housekeeping gene that encodes a component of the nuclear pore complex) and a high-complexity gene (*RUNX1*, encoding a developmentally important hematopoietic transcription factor) are shown in **Figure 1a**. *NUP54* had a single DHS assigned to it, a promoter peak that is accessible and has active histone marks (H3K4me3 and H3K27ac) across all cell types; contains no additional DHSs in its short transcription unit; and displays a high constitutive level of gene expression with little variation across cell types. By contrast, *RUNX1* was assigned 43 DHSs, including several promoter peaks and a large number of intronic enhancers across its very long transcription unit (only the first 4 introns are shown); its enhancers display dynamic patterns of accessibility and chromatin marks across cell types achieves dramatic expression changes in specific cell state transitions.

Complexity tied to gene architecture, function and expression

In fact, the differences in gene architecture, function and expression dynamics between *NUP54* and *RUNX1* were characteristic of the differences between low- and high-complexity genes. As one might expect, the distribution of log-transformed lengths of the transcription units for high-complexity genes was significantly greater than that for medium-complexity genes, which were longer than low-complexity genes ($P < 1 \times 10^{-16}$, one-sided Kolmogorov-Smirnov (KS) test for both comparisons; **Fig. 1b**). Furthermore, the fraction of the transcription unit consisting of intronic sequence was significantly higher in high- versus medium-complexity genes and in medium- versus low-complexity genes ($P < 1 \times 10^{-16}$, one-sided KS tests; **Fig. 1c**), as genes with higher complexity had longer introns (**Supplementary Fig. 2**). High-complexity genes that were highly expressed (in the top 10% of the expression distribution) in a cell type were strongly enriched for cell type–specific functions such as hematopoiesis and leukocyte activation (HSPCs), lymphocyte activation (B cells) and immune response (monocytes), whereas highly expressed low-complexity genes were enriched for similar housekeeping functions, such as translation and ubiquitination, in all cell types (Online Methods and **Supplementary Table 3**). The distribution of genes over the complexity classes and the properties of these classes also held for the top 10% of expressed genes in each cell type (**Supplementary Fig. 3**). Interestingly, translation elongation, RNA splicing and mRNA processing terms were consistently enriched in the medium-complexity genes.

Notably, genes in the low-, medium- and high-complexity classes showed a significantly increasing dynamic range of gene expression across cell types ($P < 1 \times 10^{-4}$, one-sided KS tests for low- versus medium-complexity genes and medium- versus high-complexity genes; **Fig. 1d**). Moreover, when examining highly expressed genes in cell state transitions such as the specification of HSPCs to monocytes, high-complexity genes achieved the largest increases in expression, with medium- and low-complexity genes showing progressively less upregulation ($P < 2 \times 10^{-7}$ for high- versus medium-complexity genes and $P < 4 \times 10^{-6}$ for medium- versus low-complexity genes, one-sided KS tests; **Fig. 1e**; see **Supplementary Fig. 4** for data from other cell state transitions). These results suggest that high regulatory complexity may help to achieve large changes in expression during differentiation.

Given our assignments of DHSs to nearest genes, we also investigated how gene density interacted with our complexity definitions. For each gene, we computed the number of genes on either strand overlapping a 1-Mb window centered on the gene (**Fig. 1f**) or in two 500-kb windows flanking each gene (**Supplementary Fig. 5a**). Both analyses showed that high-complexity genes occurred in low-density regions, whereas low-complexity genes occurred in gene-dense regions. We considered whether tightly packed low-complexity genes could share their DHSs with each other in a dense network of chromatin loops. However, we found that a large fraction (61%) of DHSs at low-complexity genes and present in at least one differentiated cell type corresponded to promoter or ubiquitous peaks and that the percentage of promoter and ubiquitous peaks increased with gene density (**Fig. 1g**). If there does exist a network of interacting DHSs in our low-complexity class, it would appear to be constitutive rather than dynamic, consistent with the housekeeping functions and limited expression dynamics of low-complexity genes.

High-complexity genes are enriched for transition DHSs

We next examined the patterns of DHS accessibility across cell types and their association with the gene complexity classes. The heat map in **Figure 2a** shows the most frequent patterns of accessibility found among ~48,000 DHSs present in monocytes (minor patterns omitted). Major accessibility patterns included promoter DHSs that were constitutively open across all cell types (11.1% of monocyte DHSs); ubiquitous intergenic and intronic DHSs (19.2%), non-promoter peaks open in all cell types; hematopoietic DHSs (4.4%), absent in hESCs but present in all hematopoietic cell types; and cell type–specific DHSs (35.6%), enhancers that were present only in monocytes. Finally, another major accessibility pattern we identified consisted of transition peaks (12.6%), enhancers that were established in HSPCs and maintained in monocytes but lost in other cell fates. Similar early enhancer establishment was suggested by histone mark analysis in early hematopoiesis[5]. In contrast to promoter and ubiquitous peaks, transition and cell type–specific peaks were preferentially found in regions with low gene density (**Supplementary Fig. 5b**). Examination of lymphoid cell types yielded similar major patterns, with the addition of a lymphoid-restricted accessibility pattern (**Supplementary Figs. 6–8**). In particular, each differentiated cell type had a distinct set of transition enhancers, established in HSPCs and maintained upon specification to that cell fate but lost in others. Because our analysis identified a substantial number of DHSs shared by B cells and monocytes and by T cells and NK cells, we allowed sharing of peaks by these two pairs of cell types when defining cell type–specific and transition peaks, provided that they were not found in other differentiated cell types (Online Methods and **Supplementary Table 4**).

Meta-peak visualizations of the major DNase accessibility patterns for monocyte peaks, together with average chromatin signals across

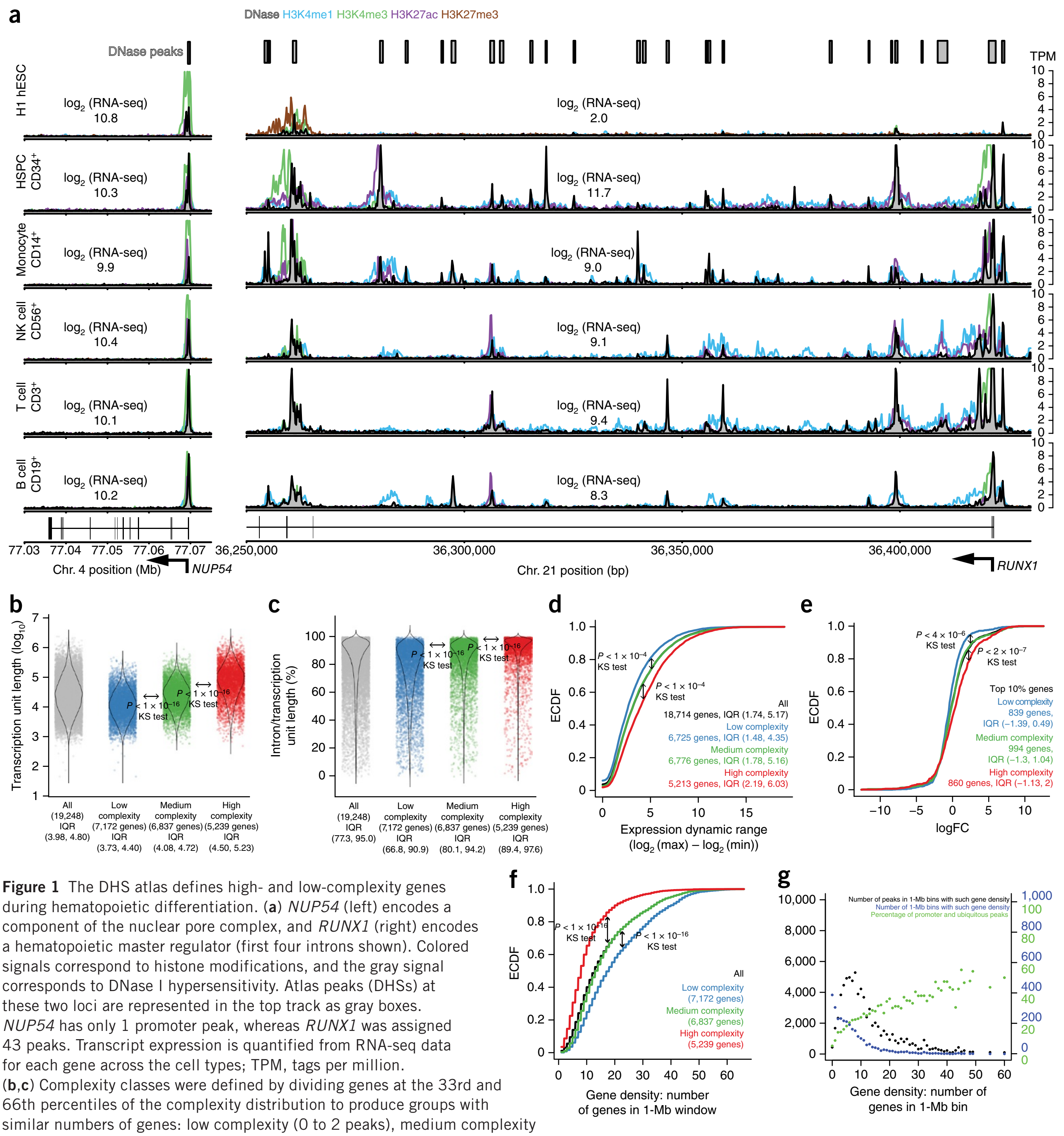


Figure 1 The DHS atlas defines high- and low-complexity genes during hematopoietic differentiation. (**a**) *NUP54* (left) encodes a component of the nuclear pore complex, and *RUNX1* (right) encodes a hematopoietic master regulator (first four introns shown). Colored signals correspond to histone modifications, and the gray signal corresponds to DNase I hypersensitivity. Atlas peaks (DHSs) at these two loci are represented in the top track as gray boxes. *NUP54* has only 1 promoter peak, whereas *RUNX1* was assigned 43 peaks. Transcript expression is quantified from RNA-seq data for each gene across the cell types; TPM, tags per million. (**b**,**c**) Complexity classes were defined by dividing genes at the 33rd and 66th percentiles of the complexity distribution to produce groups with similar numbers of genes: low complexity (0 to 2 peaks), medium complexity (3 to 7 peaks) and high complexity (8 or more peaks). High-complexity genes have longer transcription units (**b**) with higher fractions of intronic sequence (**c**). *P* values were computed using one-sided KS tests. IQR, interquartile range. (**d**,**e**) Gene expression variation correlates with regulatory complexity. (**d**) Distributions of the expression dynamic range for the three complexity groups. (**e**) Log-transformed fold changes in expression (logFC) during the transition from CD34+ HSPCs to CD14+ monocytes for the complexity groups, based on data for the top 10% most highly expressed genes in each cell type. ECDF, empirical cumulative distribution function. (**f**,**g**) Gene density analysis for genes in the complexity classes, based on gene counts in 1-Mb genomic bins centered on the transcription units of the genes. (**f**) Low-complexity genes are enriched in gene-dense regions, whereas high-complexity genes are found in gene-poor regions. (**g**) Regions of high gene density contain predominantly promoter and ubiquitous DNase peaks (green dots).

cell types, are shown in **Figure 2b–f**. DNase sensitivity and histone mark distributions within each pattern (**Supplementary Figs. 9–15**) showed that accessibility was highest in promoter peaks and lowest in cell type–specific peaks and that H3K4me1 levels were highest in transition peaks, followed by both cell type–specific and hematopoietic peaks. Ubiquitous peaks displayed a bimodal distribution for the active marks H3K4me3 and H3K27ac (**Supplementary Figs. 10, 11, 13** and **14**), suggesting that a subpopulation of these DHSs are not active transcriptional enhancers. When we overlaid CTCF ChIP-seq peaks from the Encyclopedia of DNA Elements (ENCODE) lymphoblastoid

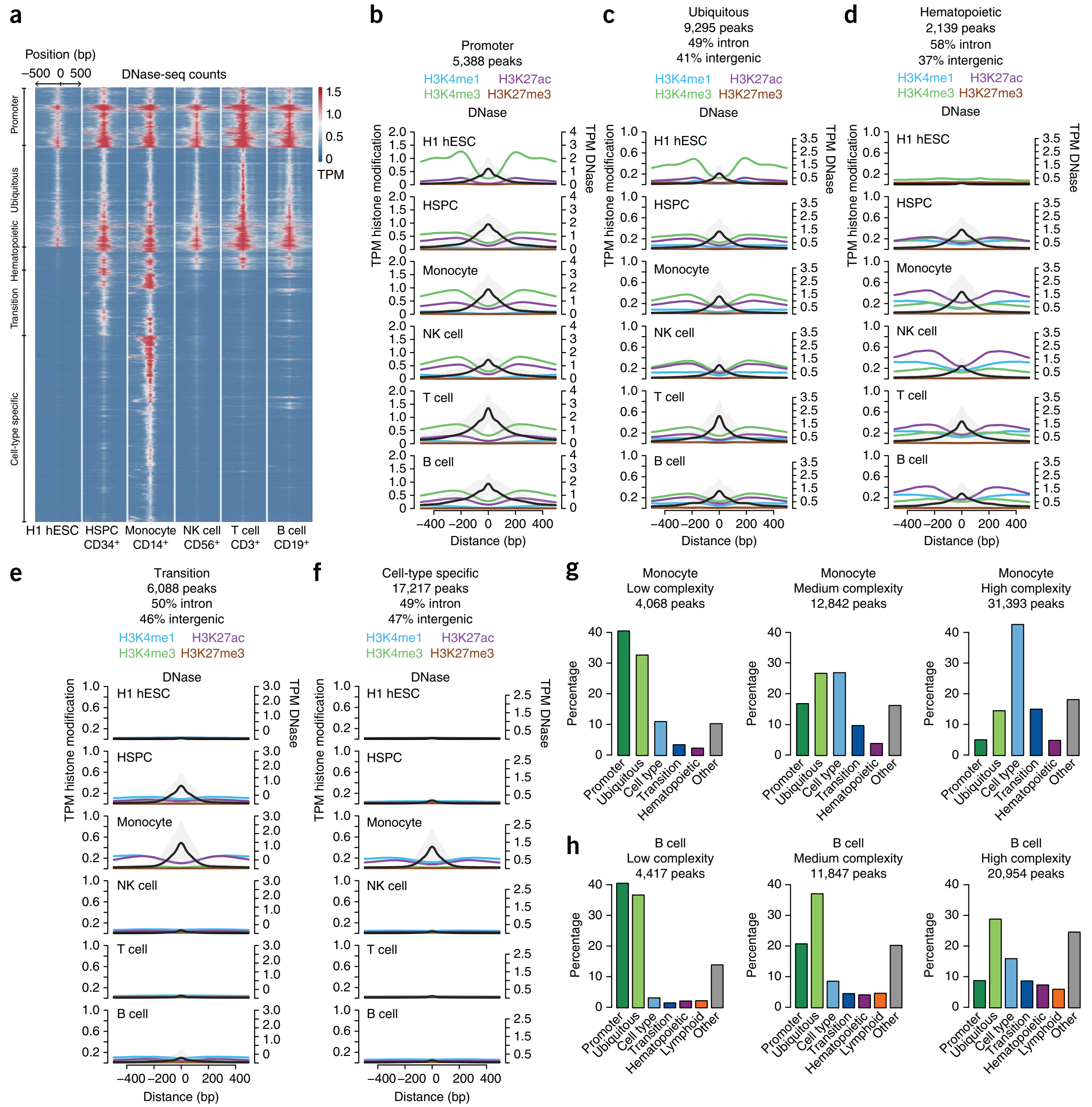


Figure 2 High-complexity genes contain enhancers with distinct dynamics. (**a**) Heat map showing DNase accessibility at peaks for hESCs and five primary hematopoietic cell types. Each row represents one of the 48,303 DNase peaks present in CD14$^+$ monocytes. DNase read counts are displayed over a 1,000-bp window, binned at 10 bp, and normalized (TPM) in each cell type. We find that 83% of CD14$^+$ peaks can be grouped into one of the following accessibility patterns: promoter, ubiquitous (non-promoter), hematopoietic, transition and cell type–specific peaks. (**b**–**f**) Each DNase accessibility pattern displays characteristic chromatin signals. DNase and histone modification signals are shown as meta-peaks, plotting the average of each signal across all peaks in the category, for promoter (**b**), ubiquitous (**c**), hematopoietic (**d**), transition (**e**) and cell type–specific (**f**) peaks. For DNase, the 10th and 90th TPM percentiles at each position are represented by gray shading around the mean signal. (**g**,**h**) Genes in different complexity classes show distinct enrichments for accessibility patterns at DHSs. The majority of DHSs associated with low-complexity genes are promoter or ubiquitous peaks, whereas DHSs assigned to high-complexity genes are strongly enriched for cell type–specific and transition peaks, in both monocytes (**g**) and B cells (**h**). The lymphoid pattern is defined by DHSs accessible in B cells, T cells and NK cells. "Cell type" is short for cell type specific.

cell line GM12878 with ubiquitous DHSs, we saw dramatic enrichment of CTCF signal at low-H3K4me3 ubiquitous peaks (Online Methods and **Supplementary Fig. 16**), suggesting a role for these peaks in three-dimensional chromatin structure. An early ENCODE study of 1% of the human genome across six unrelated cell types found that almost all ubiquitous DNase peaks were promoter peaks (86%), and most of the remainder were bound by CTCF (10%)[25]; our lineage-based analysis identifies a larger class of ubiquitous, non-promoter DHSs with active histone marks and without CTCF occupancy. Similar observations could be made for analogous enhancer types in other differentiated cells (**Supplementary Figs. 9–15**).

Strikingly, DHSs assigned to genes in different complexity classes were enriched for distinct lineage-wide DNase accessibility patterns. The DHSs of low-complexity genes were dominated by promoter and ubiquitous peaks in monocytes (**Fig. 2g**) and B cells (**Fig. 2h**), as well as other cell fates (**Supplementary Fig. 17**). In contrast, the DHSs of high-complexity genes were strongly enriched for both cell type–specific and transition peaks ($P < 2 \times 10^{-107}$ and $P \sim 0$ for transition and monocyte-specific peaks, Fisher's test (**Fig. 2g**); $P < 5 \times 10^{-85}$ and $P < 2 \times 10^{-153}$ for transition and B cell–specific peaks (**Fig. 2h**)). Therefore, consistent with their greater dynamic range of expression, high-complexity genes are also enriched for enhancers with dynamic accessibility, including both transition and cell type–specific enhancers.

Upregulated high-complexity genes gain active enhancers

We next examined the gain and loss of active enhancers in cell state transitions from HSPCs to differentiated cell types. Whereas cell type–specific enhancers changed their accessibility, transition peaks displayed a change in chromatin marks. Two-dimensional density plots of the active mark H3K27ac (ref. 9) versus the mark H3K4me1 for all DNase peaks in HSPCs and monocytes are shown in **Figure 3a**,**b** (see **Supplementary Fig. 18** for other cell types). Gaussian mixture modeling defined three overlapping subpopulations (Online Methods): peaks with low activity and moderate H3K4me1 levels, peaks with moderate activity and high H3K4me1 levels, and peaks with high activity and moderate H3K4me1 levels. In HSPCs, monocyte transition peaks were strongly enriched in both the low-activity/moderate-H3K4me1 and moderate-activity/high-H3K4me1 subpopulations. Upon specification to monocytes, the transition peaks largely moved to the moderate-activity/high-H3K4me1 region of the monocyte landscape (**Fig. 3b**) and displayed a strong increase in H3K27ac levels (**Supplementary Fig. 19**); the transition peaks for other cell fates showed the same patterns (**Supplementary Figs. 20–22**). A partial analogy can be made to the bivalent domains in hESCs[7], some of which coincided with DNase peaks and formed a subpopulation of DHSs with moderate levels of the active promoter mark H3K4me3 and high levels of the repressive mark H3K27me3 in the corresponding two-dimensional hESC landscape (**Supplementary Fig. 23**) but resolved to active or inactive DHSs upon specification to HSPCs (**Supplementary Figs. 24–26**).

We then examined the 10% most highly expressed genes in HSPCs and monocytes and identified the high- and low-complexity genes with the largest expression changes in HSPC to monocyte specification. Although a handful of low-complexity genes appeared to achieve large changes in expression without any gain or loss of DHSs (**Fig. 3c** and **Supplementary Figs. 27–29**), large increases or decreases in expression at high-complexity genes were consistently associated with gains and losses of active cell type–specific enhancers, almost always accompanied by gain or loss of H3K4me1 or H3K27ac at one or more transition enhancers (**Fig. 3d**). Indeed, highly expressed high-complexity genes with at least one transition peak experienced greater chromatin remodeling than those without a transition peak, gaining significantly more cell type–specific peaks ($P < 1 \times 10^{-16}$, KS test) while losing more HSPC-specific peaks, and achieved higher log-transformed fold changes in expression during monocyte specification (**Supplementary Fig. 30a**–**c**). These results suggest that regulatory complexity is important for achieving cell type–specific gene expression; however, complex locus control regions are not typically established in a single cell state transition but rather are primed through early enhancer establishment.

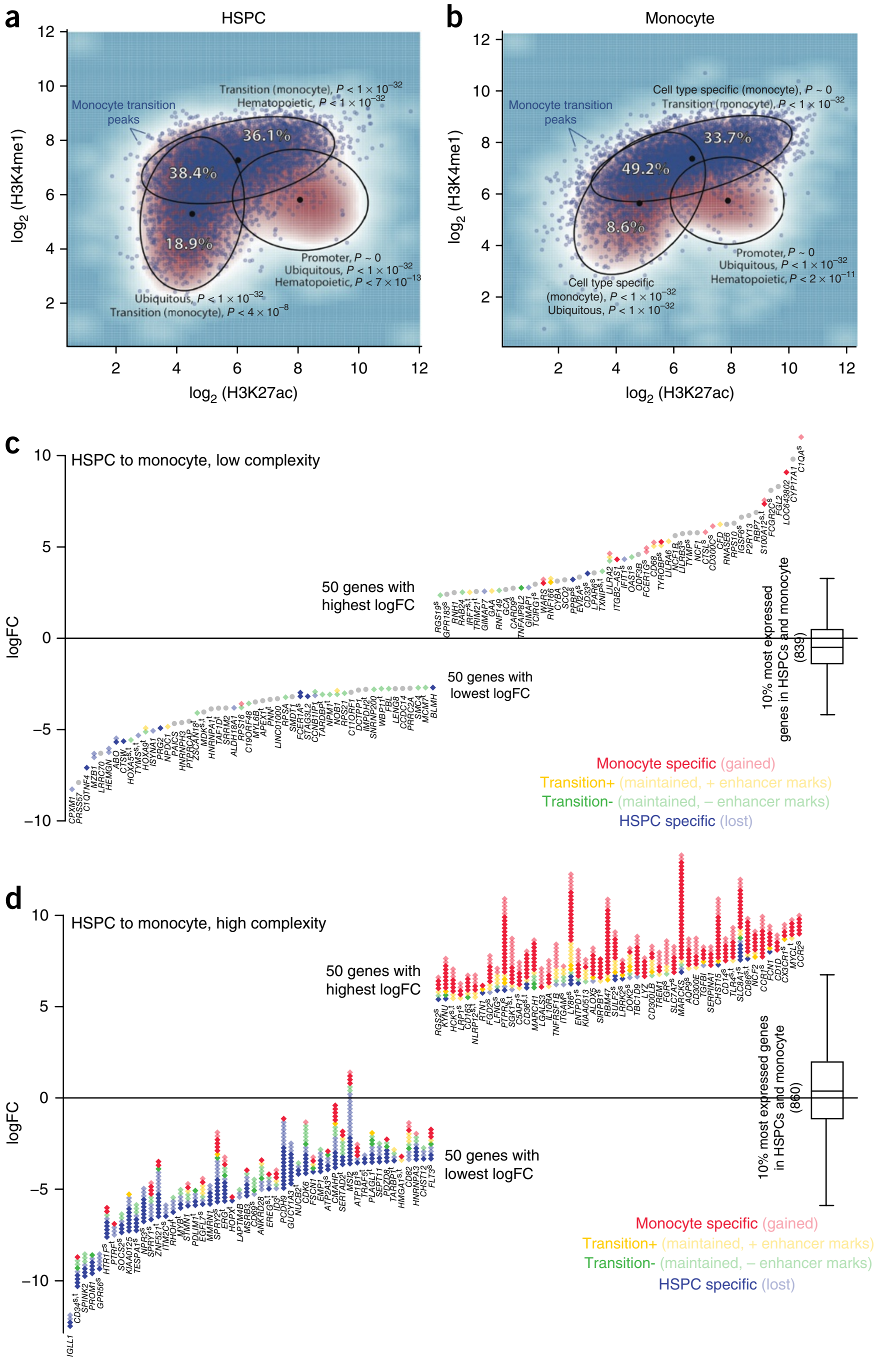
SeqGL predicts transcription factor binding signals from cell-type DHSs

We next asked whether we could predict the expression changes of genes in cell state transitions from the DNA sequence signals and lineage history of their active enhancers. To dissect sequence motifs in enhancers, we applied SeqGL, a group lasso algorithm we developed to identify multiple DNA signals *de novo* from DNase data using *k*-mer patterns[26]. SeqGL first computes a count matrix of occurrences of *k*-mers in DNase peak sequences and flanking regions and then clusters *k*-mers by their count vectors to identify *k*-mer groups. Part of the clustered count matrix obtained by applying SeqGL to B cell DNase-seq data is shown in **Figure 4a**. Each group tends to contain sequence-similar *k*-mers that may represent variants of the DNA recognition signal of a single transcription factor, found in a subset of training examples. To obtain a scoring function for each *k*-mer group, SeqGL learns weights over *k*-mers by training a binary logistic regression model to discriminate between peaks and flanking regions[26] (Online Methods). Groups with primarily positive weights represent predicted transcription factor binding signals found in peaks (**Fig. 4a**), and, for each of these groups, we could identify highly scored peak regions, generate a binding motif and compare to existing databases to identify the corresponding transcription factor, if present (**Fig. 4a** and Online Methods).

In a previous analysis of ENCODE DNase-seq data, SeqGL was more sensitive than traditional motif discovery methods in identifying transcription factor binding signals supported by ChIP-seq data in the same cell type[26]. Running SeqGL on DNase-seq data for our 5 hematopoietic cell types produced ~50 predicted transcription factor binding signals mapping to 25–34 distinct known transcription factor motifs per cell type (**Supplementary Table 5**). Although we do not have parallel transcription factor ChIP-seq data for these cell types, SeqGL retrieved many transcription factor binding signals that are supported by the literature and previous ChIP-seq studies, including signals for PU.1, RUNX and AP-1 in HSPCs as well as differentiated myeloid and lymphoid cells[27]; T-BOX/TBET and NF-κB signals in NK cells, T cells and monocytes[28]; myeloid zinc-factor MZF signals in HSPCs and monocytes[29]; CEBP signals in monocytes[30]; and STAT signals exclusively in T cells[31]. However, because SeqGL extracts DNA sequence signals from DNase-seq data and relies on motif databases to associate signals with transcription factors, it cannot distinguish between transcription factors with nearly identical binding motifs.

An example of SeqGL predictions of transcription factor binding sites for DHSs in the *PAX5* locus after training the model on DNase peaks in CD19$^+$ B cells is shown in **Figure 4b**. We analyzed the transcription factor score distributions from SeqGL across different categories of DHSs with active histone marks (H3K4me1 or H3K27ac). We refined the previous DNase accessibility categories by performing pairwise differential read count analysis on DHSs to identify cell type–specific or shared events[32] (Online Methods), obtaining four categories present in B cells: promoter, ubiquitous (non-promoter),

Figure 3 Gain of active enhancers in cell state transitions correlates with increased expression. (**a**,**b**) Two-dimensional topographical plots show the density of all accessible DNase peaks in CD34$^+$ HSPCs (**a**) and CD14$^+$ monocytes (**b**) in the landscape defined by the active mark H3K27ac (*x* axis) and the mark H3K4me1 (*y* axis). Gaussian mixture modeling identified three subpopulations in each plot, shown by the mean of each mixture component with ellipses corresponding to 90% probability curves. Enrichments of accessibility patterns in each subpopulation are indicated; all enrichment *P* values were obtained by Fisher's test. Percentages indicate the proportion of all the monocyte transition peaks (shown in blue) that lie in each region of the ellipses. (**c**,**d**) Gene expression and enhancer changes are shown for low-complexity (**c**) and high-complexity (**d**) genes in the transition from HSPCs to monocytes for the 100 genes with the largest absolute log-transformed fold change in expression in each class. Each gene is illustrated by a stack of diamonds, where each diamond represents a DHS associated with the gene. The bottom of the stack corresponds to the log-transformed fold change in expression for the gene (*y* axis). Red (blue) diamonds are active peaks gained (lost) in the transition. Peaks maintained in the transition and that have gains (losses) in enhancer marks (H3K4me1 or H3K27ac) are represented by yellow (green) diamonds (Online Methods). Gray dots represent peaks present in both cell types without changes in histone marks. Box-and-whisker plots show the distribution of log-transformed fold change in expression for all the genes, with the box showing the interquartile range (IQR) and the whiskers placed at a distance of 1.5 × IQR from the corresponding quartile. Gene names annotated with "t" and "s" have gene ontology annotations for transcriptional and signaling processes, respectively.



transition and cell-type specific (**Fig. 4c**). As expected, the NFY sequence signal was enriched in B cell promoter peaks (**Fig. 4d**), CTCF signal was enriched in ubiquitous as well as promoter peaks (**Fig. 4e**) and IRF signal was enriched in B cell–specific peaks (**Fig. 4f**). Interestingly, the PU.1 sequence signal was enriched not only in cell type–specific peaks but also in transition peaks (**Fig. 4g**). Although PU.1 is known to have a key role in both HSPC and B cell function[2,33], the binding of PU.1 at transition peaks and its impact on gene regulation have not been characterized. We therefore developed a regression framework to model the influence of PU.1 and other factors on gene expression changes in cell fate specification.

Regression model accurately predicts expression changes

We learned global regression models to predict expression changes from DNA signals in active enhancers by training on high- and low-complexity genes in differentiation from HSPCs to distinct cell fates. We trained a separate regression model for each cell fate and used SeqGL to derive transcription factor sequence signals from DNase-seq data in HSPCs and the relevant differentiated cell type. In the model for B cell specification, each gene was represented by a set of SeqGL-predicted transcription factor binding scores (features) computed from its active DHSs (marked by H3K4me1 or H3K27ac) in HSPCs and B cells (**Fig. 5a** and Online Methods). The model learns a regression coefficient for each transcription factor signal or pair of signals in each category of DHSs (promoter, ubiquitous, HSPC specific, B cell specific and transition) to allow these signals to have distinct roles in predicting expression changes. For example, transcription factor signals in HSPC- and B cell–specific peaks represent regulatory inputs that are lost and gained, respectively, in the cell state transition; transition peaks may be bound by transcription factors in HSPCs that act as place holders for later binding by other transcription factors in B cells or may be continuously occupied by the same transcription factors in both cell types. The contribution of transcription factors in these different roles to

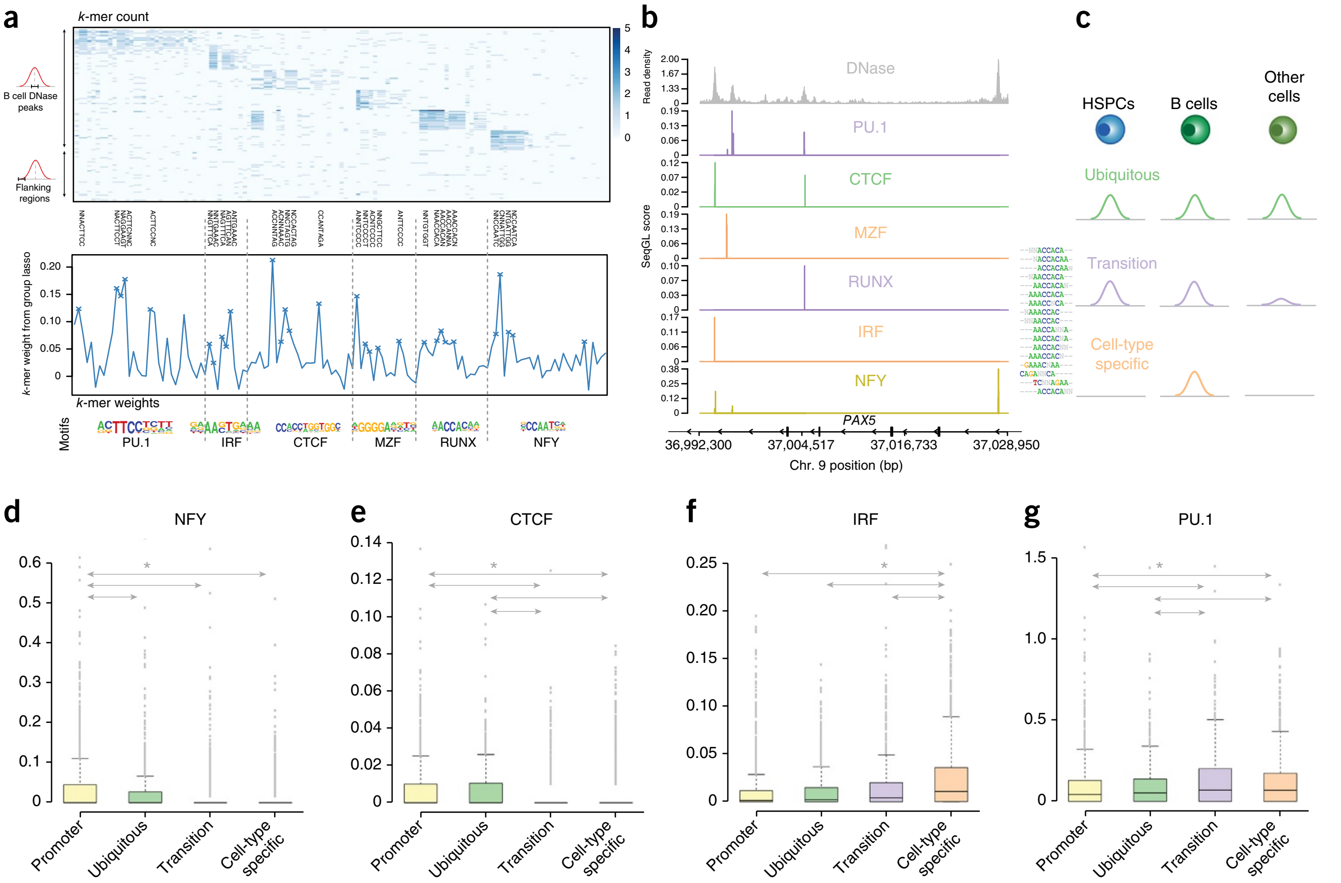


Figure 4 SeqGL identifies multiple transcription factor sequence signals in B cell DNase peaks. (**a**) SeqGL identifies multiple DNA sequence signals *de novo* from DNase-seq data using *k*-mer patterns that can discriminate peaks and flanking regions. SeqGL first computes a *k*-mer count matrix from the set of input sequences and clusters co-occurring *k*-mers into groups. A subset of this matrix for B cells is shown in the top panel. SeqGL then learns weighting over *k*-mers in each group that discriminates peaks from flanking regions. The bottom panel shows the inferred weights and top-ranking *k*-mers for each group. Groups with positive weights represent predicted transcription factor binding signals. (**b**) The *PAX5* locus is shown, highlighting the top transcription factors identified by training SeqGL on B cell DNase peaks. The first row shows DNase read density, and the remaining rows show group *k*-mer scores for SeqGL-predicted transcription factor binding signals. For RUNX, the *k*-mer patterns contributing to the scoring model are shown. (**c**) Active B cell peaks were divided into four categories: (i) promoter, peaks defined in the promoter; (ii) ubiquitous, peaks with similar accessibility across cell types; (iii) transition, peaks retained specifically in the transition from HSPCs to B cells; and (iv) cell-type specific, peaks that are significantly higher in B cells. (**d**–**g**) Bar plots showing binding signal enrichment in different peak categories. NFY binding is strongly enriched in promoter peaks (**d**), and CTCF binding is enriched in both ubiquitous and promoter peaks (**e**). The peak for the B cell factor IRF is significantly higher in cell type–specific peaks (**f**), whereas PU.1 binding is enriched in both cell type–specific and transition peaks (**g**). $*P < 0.01$, Wilcoxon rank-sum test. Error bars, s.d.

orchestrating gene expression changes is captured by the sign and magnitude of the learned regression weights (**Fig. 5a**).

Restricting to the top 10% most highly expressed genes in HSPCs and the differentiated cell type, we found that the predicted expression changes for held-out genes (genes not used for training) in the B cell transition were remarkably accurate (Spearman correlation $\rho = 0.666$), especially for high-complexity genes ($\rho = 0.742$), which had the largest expression changes (**Supplementary Fig. 31**). Low-complexity genes had more modest expression changes, and the model behaved appropriately for these genes by predicting smaller log-transformed fold changes in expression (**Supplementary Fig. 31**). Similarly strong prediction performance was obtained for transitions to other cell fates (**Supplementary Table 6**).

Analysis of the regression models highlights the specialized roles of different kinds of enhancers (**Supplementary Fig. 32**). The first heat map in **Figure 5b** shows the contribution of transcription factors residing in different DHS categories in predicting the expression changes for high-complexity genes in B cell specification, and the remaining heat maps show results for other cell state transitions. In some cases, several *k*-mer groups were associated with the same transcription factor by SeqGL and led to multiple columns labeled with the same transcription factor, sometimes within the same DHS category. Some of these groups with multiple *k*-mers for a transcription factor might represent subtle differences in the underlying motifs and perhaps merit experimental examination, whereas others are likely a result of overclustering (**Supplementary Fig. 33**), which nonetheless does not harm the predictive performance of the regression model. Using second-order features significantly improved performance over first-order features alone ($P < 5 \times 10^{-7}$, Wilcoxon signed-rank test) but only fine-tuned the predictions. The model identified known HSPC factors such as GATA and RUNX[27] in HSPC-specific peaks as predictors of negative log-transformed fold changes in expression for genes specifically expressed in HSPCs; similarly, known B cell factors such as IRF[27] in B cell–specific peaks predicted positive log-transformed fold changes in expression in genes specifically expressed in B cells. Interestingly, binding signals for PU.1 in HSPC- and B cell–specific peaks were associated with downregulation of HSPC genes and upregulation of B cell genes, respectively; moreover, a subset of B cell genes harbored PU.1 signals in transition peaks that strongly predicted their changes in expression. This finding suggests a new mechanistic role for the binding of PU.1 in HSPCs at transition enhancers to poise genes for B cell specification.

We therefore defined three sets of high-complexity, highly expressed genes in B cells: (i) genes with predictive signals only in transition peaks; (ii) genes with predictive signals only in B cell–specific peaks; and (iii) poised genes with predictive signals in both B cell–specific peaks and transition peaks. Strikingly, poised genes displayed the strongest upregulation upon differentiation into B cells (**Fig. 5c**) as well as the strongest enrichment of B cell differentiation and activation ontology terms (**Fig. 5d**), suggesting that key B cell identity genes are regulated by transcription factor binding to both B cell–specific and transition enhancers. Among the poised genes were a number with important B cell functions, such as the antiapoptotic and regulatory genes *BCL2*, *IKZF1*, *KLF6*, *TCF3* and *IRF8*, as well as the immune signaling genes *PLCG2*, *SYK*, *IL4R*, *INPP5D*, *IFNGR1* and *IL16*.

We confirmed that the transition peaks of poised genes had significantly higher DNase accessibility and lower levels of H3K27me3 repressive marks in comparison to counterpart B cell–specific peaks in HSPCs ($P < 3 \times 10^{-93}$, Wilcoxon rank-sum test) and were close in chromatin state to the HSPC-specific peaks of HSPC genes ($P < 4 \times 10^{-3}$, Wilcoxon rank-sum test; **Supplementary Fig. 34**). This suggests that PU.1 binding at transition enhancers—the main predictive transition peak signal—might help maintain the chromatin in an open state. However, poised B cell genes had much lower expression in HSPCs than HSPC-specific genes ($P < 3 \times 10^{-93}$, Wilcoxon rank-sum test; **Supplementary Fig. 34**). To explain this, we confirmed that poised genes had reduced signals for key HSPC-specific factors such as GATA in their HSPC and B cell active enhancers while harboring strong B cell–specific transcription factor signals such as ones for IRF in their B cell active enhancers (**Fig. 5e** and Online Methods). Therefore, although poised genes lack HSPC sequence signals to achieve high expression in precursor cells, chromatin poising by PU.1 may help promote their upregulation once B cell factors are expressed.

Interestingly, our modeling of monocyte specification also highlighted a role for PU.1 poising, but here the PU.1 signal arose in promoter peaks and might maintain open chromatin for later binding by the CEBP promoter-associated factor, a key regulator of monocyte/macrophage fate (**Supplementary Figs. 35** and **36**).

Regression model nominates gene-enhancer reassignments

Although overall we observed high rank correlation between measured and predicted expression changes for highly expressed genes in cross-validation (**Fig. 6a**), predictions for a small number of genes had high error (**Supplementary Fig. 37**). We hypothesized that these errors might arise from misassignment of enhancers to nearby genes. Therefore, we implemented an iterative reassignment procedure where we trained regression models using cross-validation, flagged held-out genes with large prediction errors, scanned for marginally expressed genes neighboring the flagged genes and reassigned the active enhancers of the silent genes to the poorly predicted genes (Online Methods). In almost all cases, the prediction errors for expression changes of flagged genes decreased (**Supplementary Fig. 38**), and overall rank correlation in cross-validation also improved ($\rho = 0.78$; **Supplementary Fig. 37**). For an independent validation of the enhancer reassignment method, we further found that B cell enhancer reassignments significantly reduced prediction error for the affected genes in the monocyte, NK cell and T cell regression models ($P < 2 \times 10^{-3}$, 1×10^{-9} and 3×10^{-7}, respectively, Wilcoxon signed-rank test; **Fig. 6b** and **Supplementary Fig. 39**). Similarly, using reassignments of enhancers to genes from other regression models led to significantly improved regression performance in independent cell types in all but one case (**Supplementary Fig. 39** and **Supplementary Table 7**). As an example, the B cell marker gene *MS4A1* (*CD20*) resides in a gene-dense region where no nearby genes are expressed in B cells. Originally assigned two active DNase peaks, it acquired 13 actively marked peaks through enhancer reassignment to achieve a notable improvement in regression prediction (**Supplementary Fig. 40**).

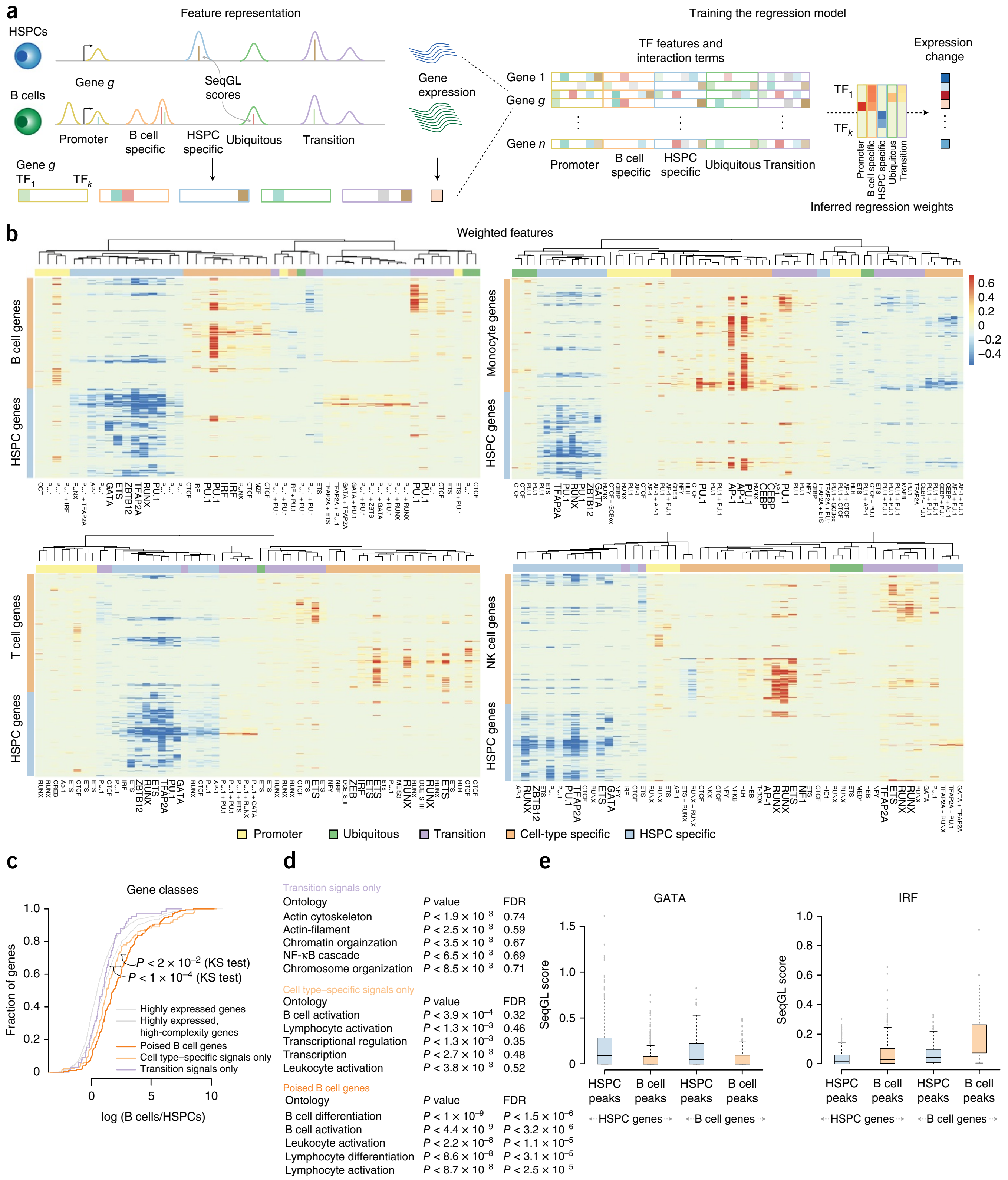
DISCUSSION

Recent studies have proposed various epigenetic signals to define highly cell type–specific super-enhancers that mark cell identity genes[10–12,14]. We found that the highly expressed, high-complexity genes in our cell types are indeed enriched for genes proximal to super-enhancers defined by broad ChIP-seq domains of Mediator (for hESCs) or H3K27ac (for hematopoietic cell types[10]), although a majority of previously defined super-enhancer genes do not overlap with these genes (**Supplementary Fig. 41**). Moreover, about half of previously defined hESC super-enhancers contain no or one DHS, whereas ~41% of HSPC super-enhancers and >30% of lymphoid cell super-enhancers contain two or fewer DHSs (**Supplementary Fig. 42**). Therefore, current definitions of super-enhancers do not always include complex locus control regions with many individual enhancers, as others have also observed[34]. Rather than defining a separate repertoire of super-enhancers for each cell type, we define regulatory complexity as a property of individual genes based on DHS multiplicity across a lineage. This definition may be more natural from a mechanistic and evolutionary standpoint by characterizing a gene's

Figure 5 Regression model suggests a role for PU.1 in the early establishment of B cell enhancers. (**a**) Regression framework to model the effect of transcription factor signals on gene expression changes in the transition from HSPCs to B cells. Transcription factor (TF) binding scores in both HSPC and B cell peaks (categorized as in **Fig. 4c**) form the features for each gene. The weights of different transcription factors across peak categories, representing their contribution to expression changes, are inferred using ridge regression. (**b**) Heat maps showing the predicted contribution of each transcription factor across peak categories to gene expression changes in cell state transitions. Each row represents a high-complexity gene, and each column represents a transcription factor signal or pair of transcription factor signals for a particular enhancer category; the value in each cell is the SeqGL-derived binding score weighted by the regression coefficient, so that we can visualize whether the transcription factor signal contributes positively (orange) or negatively (blue) to the overall predicted log-transformed change in expression. PU.1 signals from multiple peak classes occur in all models. High expression in HSPCs, monocytes, B cells, T cells and NK cells is strongly predicted by corresponding cell type–specific factors. Changes in many B cell genes are predicted not only by signals in B cell peaks but also by transition peak signals, primarily for PU.1. Key transcription factors with known roles in cell type specification are highlighted with a larger font. (**c**) The B cell–specific genes in **b** were classified into three categories: genes with predictive signals from (i) transition peaks alone; (ii) B cell–specific peaks alone; and (iii) both B cell–specific and transition peaks, termed poised genes. Poised genes are significantly upregulated in comparison to other categories ($P < 2 \times 10^{-2}$, category (i) and $P < 1 \times 10^{-4}$, category (ii), one-sided KS tests). (**d**) Functional enrichment through gene ontology analysis shows that poised genes are significantly more enriched for ontologies related to B cell functions than other genes. FDR, false discovery rate. (**e**) HSPC-specific signals such as ones for GATA are significantly higher in active HSPC peaks for HSPC genes than those for poised B cell genes ($P < 0.03$, Wilcoxon rank-sum test) and are low in B cell peaks. Similarly, signals for IRF, a B cell factor, are significantly higher in active peaks of poised B cell genes than in HSPC genes ($P < 3 \times 10^{-4}$) and are low in HSPC peaks. Error bars, s.d.

regulatory potential and the cell type–dependent constraints on that potential. Although we divide genes into low-, medium- and high-complexity classes to simplify our analysis, there may be no intrinsic threshold dividing super-enhancers from typical enhancers.

Our analysis also demonstrates the value of examining the lineage dynamics and DNA sequence content of enhancers in addition to their spatial clustering, proposing a distinctive role for transition enhancers with PU.1 signal in B cell specification. We found that the presence of a transition peak in high-complexity genes is associated with higher gain in cell type–specific DHSs and greater enhancer remodeling in the cell state transition. A potential model is that the transition enhancer nucleates the establishment of enhancer-promoter DNA



Transition signals only

Ontology	P value	FDR
Actin cytoskeleton	$P < 1.9 \times 10^{-3}$	0.74
Actin-filament	$P < 2.5 \times 10^{-3}$	0.59
Chromatin orgainzation	$P < 3.5 \times 10^{-3}$	0.67
NF-κB cascade	$P < 6.5 \times 10^{-3}$	0.69
Chromosome organization	$P < 8.5 \times 10^{-3}$	0.71

Cell type–specific signals only

Ontology	P value	FDR
B cell activation	$P < 3.9 \times 10^{-4}$	0.32
Lymphocyte activation	$P < 1.3 \times 10^{-3}$	0.46
Transcriptional regulation	$P < 1.3 \times 10^{-3}$	0.35
Transcription	$P < 2.7 \times 10^{-3}$	0.48
Leukocyte activation	$P < 3.8 \times 10^{-3}$	0.52

Poised B cell genes

Ontology	P value	FDR
B cell differentiation	$P < 1 \times 10^{-9}$	$P < 1.5 \times 10^{-6}$
B cell activation	$P < 4.4 \times 10^{-9}$	$P < 3.2 \times 10^{-6}$
Leukocyte activation	$P < 2.2 \times 10^{-8}$	$P < 1.1 \times 10^{-5}$
Lymphocyte differentiation	$P < 8.6 \times 10^{-8}$	$P < 3.1 \times 10^{-5}$
Lymphocyte activation	$P < 8.7 \times 10^{-8}$	$P < 2.5 \times 10^{-5}$

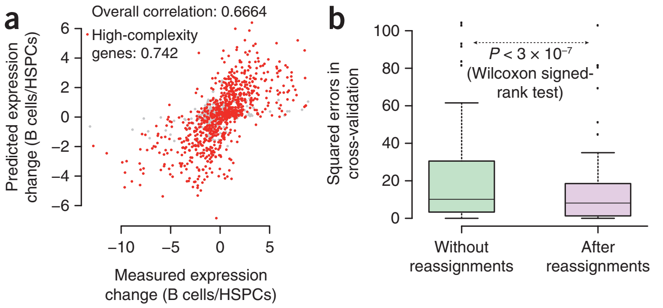


Figure 6 Regression model proposes reassignment of enhancers to genes. (**a**) The regression performs very well on high-complexity genes (Spearman rank correlation $\rho = 0.74$, tenfold cross-validation). A number of genes that are highly expressed in either B cells or HSPCs had high regression errors (**Supplementary Fig. 37**), often because their large expression changes were not associated with corresponding high complexity in the assignment of enhancers to genes. To correct the potential misassignment of enhancers to genes, an iterative reassignment scheme was implemented: for each gene with high regression error and high expression in either B cells or HSPCs, the nearest marginally expressed gene (expression <75th percentile) neighboring the current regulatory locus was masked, and the peaks associated with the marginally expressed gene were reassigned to the gene with regression error; the regression model was then retrained. This procedure was iterated until convergence of regression coefficients. After iterative enhancer reassignment, the regression model shows improved performance with significant reduction in prediction errors for affected genes in both cell types ($P < 3 \times 10^{-3}$, Wilcoxon signed-rank test; **Supplementary Fig. 38**). (**b**) For an independent evaluation of the enhancer reassignment procedure, the enhancer reassignments generated in B cells were used to predict gene expression changes in the transition to T cells. Squared errors using these reassignments were significantly reduced in comparison to errors without any reassignments ($P < 3 \times 10^{-7}$, Wilcoxon signed-rank test). Similarly, using enhancer reassignments from any of the other regression models led to improved regression performance in independent cell types (**Supplementary Fig. 39**).

looping in the new cell state. New genome editing tools should allow this model to be tested experimentally. We lacked the data to examine the sequential establishment of complex locus control regions over multiple steps in differentiation. However, we did find a sizeable set of non-promoter, non-ubiquitous DHSs that are established as accessible in hESCs and maintained until a differentiated cell fate is achieved; like transition peaks, these DHSs are enriched in high-complexity genes (**Supplementary Fig. 43**). Although limited by the data available, we have shown proof-of-principle results that complex locus control and early establishment of enhancers have important roles in transcriptional programs during differentiation. Notably, our study is not limited to a traditional examination of stable end states in development and the repertoire of active regulatory elements in these end states. Rather, our analysis considers the asynchronous establishment and activity of enhancers that collaborate to mediate large developmental shifts in gene expression.

Our regulatory model starts by simply assigning DHSs to the nearest gene. There is disagreement in the literature about how often this rule results in incorrect assignment. An early ENCODE 5C analysis on 1% of the human genome concluded that only ~7% of looping interactions are with the nearest gene[35]. By contrast, recent cohesin ChIA-PET interaction data in mouse embryonic stem cells supported the assignment of enhancers to the proximal active gene in a large majority of cases (83% of super-enhancers and 87% of typical enhancers) and also largely supported assignment of enhancers to single rather than multiple genes[36], and promoter capture Hi-C analysis of an ENCODE lymphoblastoid cell line found that interactions with promoters are directed at the nearest promoter in almost two-thirds of cases, whereas the remaining interactions pass over intervening active or inactive promoters[21]. These studies are largely consistent with our initial assumptions, although definitive interaction data are not yet available and the statistical analysis of 3C sequencing data is challenging.

Computational methods for predicting enhancer-gene associations, often based on correlating the accessibility or active marks of DNase-mapped enhancers and promoters across many cell types[37–39], have been effectively deployed in large-scale analyses. However, our results suggest that such strategies may not work in all cases: high-complexity genes can have constitutively accessible and active promoters across a lineage but cell type–specific enhancers. Other approaches have correlated the activity of enhancers with target gene expression across unrelated cell types without using a gene regulation model[37,40,41]. Despite assigning enhancers to nearest genes, our regression model predicts gene expression changes with high accuracy, and larger prediction errors for a minority of genes suggested enhancer reassignments that reduced prediction error in independent cell types. As new technologies enable the profiling of fine-grained cell populations using small cell numbers[5,42], our approach may point the way forward to improved enhancer annotation.

URLs. Roadmap Epigenomics data portal, http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/; supplementary website, http://cbio.mskcc.org/public/Leslie/Early_enhancer_establishment/; SegGL source code repository, https://bitbucket.org/leslielab/seqgl.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank A. Kundaje for extensive advice on the processing of Roadmap Epigenomics data sets, and we thank A. Arvey for helpful discussions at early stages in the project. This work was supported by US National Institutes of Health grants R01-HG006798, U01-HG007893 and U01-HG007033.

AUTHOR CONTRIBUTIONS

A.J.G. performed computational analyses to construct the DHS atlas, characterize gene complexity classes, describe histone modifications at enhancer classes, and quantify gain and loss of active DHSs with gene expression changes and contributed to writing the manuscript. M.S. developed the DNase peak calling pipeline and the SeqGL tool, performed the regression analysis and iterative reassignment of enhancers, and contributed to writing the manuscript. C.S.L. conceived the project, advised on the analysis and algorithm development, supervised the research and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Samstein, R.M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–166 (2012).
2. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
3. Smale, S.T. Pioneer factors in embryonic stem cells and differentiation. *Curr. Opin. Genet. Dev.* **20**, 519–526 (2010).
4. Rothenberg, E.V. The chromatin landscape and transcription factors in T cell programming. *Trends Immunol.* **35**, 195–204 (2014).
5. Lara-Astiaso, D. *et al.* Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).

6. Xu, C.R. *et al.* Chromatin "prepattern" and histone modifiers in a fate choice for liver and pancreas. *Science* **332**, 963–966 (2011).
7. Bernstein, B.E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
8. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
9. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
10. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
11. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
12. Parker, S.C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926 (2013).
13. Wang, H. *et al.* NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proc. Natl. Acad. Sci. USA* **111**, 705–710 (2014).
14. Benayoun, B.A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673–688 (2014).
15. Stergachis, A.B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).
16. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
17. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
18. Li, Q.H., Brown, J.B., Huang, H.Y. & Bickel, P.J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
19. Zheng, Y. *et al.* Role of conserved non-coding DNA elements in the *Foxp3* gene in regulatory T-cell fate. *Nature* **463**, 808–812 (2010).
20. Yoshida, T. *et al.* Transcriptional regulation of the *Ikzf1* locus. *Blood* **122**, 3149–3159 (2013).
21. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
22. Kieffer-Kwon, K.R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507–1520 (2013).
23. Anderson, E. & Hill, R.E. Long range regulation of the sonic hedgehog gene. *Curr. Opin. Genet. Dev.* **27**, 54–59 (2014).
24. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
25. Xi, H. *et al.* Identification and characterization of cell type–specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3**, e136 (2007).
26. Setty, M. & Leslie, C.S. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLoS Comput. Biol.* **11**, e1004271 (2015).
27. Wickrema, A. & Kee, B. Molecular Basis of Hematopoiesis. (Springer, 2009).
28. Lazarevic, V., Glimcher, L.H. & Lord, G.M. T-bet: a bridge between innate and adaptive immunity. *Nat. Rev. Immunol.* **13**, 777–789 (2013).
29. Perrotti, D. *et al.* Overexpression of the zinc finger protein MZF1 inhibits hematopoietic development from embryonic stem cells: correlation with negative regulation of CD34 and c-*myb* promoter activity. *Mol. Cell. Biol.* **15**, 6075–6087 (1995).
30. Pan, Z., Hetherington, C.J. & Zhang, D.E. CCAAT/enhancer-binding protein activates the CD14 promoter and mediates transforming growth factor β signaling in monocyte development. *J. Biol. Chem.* **274**, 23242–23248 (1999).
31. Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* **151**, 981–993 (2012).
32. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
33. Mak, K.S., Funnell, A.P., Pearson, R.C. & Crossley, M. PU.1 and haematopoietic cell fate: dosage matters. *Int. J. Cell Biol.* **2011**, 808524 (2011).
34. Pott, S. & Lieb, J.D. What are super-enhancers? *Nat. Genet.* **47**, 8–12 (2015).
35. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
36. Dowen, J.M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
37. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
38. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
39. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
40. Malin, J., Aniba, M.R. & Hannenhalli, S. Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers. *Nucleic Acids Res.* **41**, 6828–6838 (2013).
41. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
42. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

ONLINE METHODS

Data and preprocessing. Aligned DNase-seq BAM files were downloaded from the Roadmap Epigenomics data portal (see URLs). RNA-seq data were downloaded as fastq files and aligned to hg19 using STAR[43] (see **Supplementary Table 1** for accession numbers and links to each library). Processed data files have also been made available at the supplementary website for the paper (see URLs).

The Bioconductor package GenomicFeatures[44] was used to determine the number of reads mapping to each gene. The reads per kilobase (RPK) value was calculated for each gene, and RPK values were quantile normalized across all the cell types to obtain gene expression levels.

DNase peak calling. The peak calling pipeline is outlined in **Supplementary Figure 1**. Peak calling was performed on each cell type individually: first, the reads from different replicates were pooled, and the MACS peak caller[45] was then used to identify peaks with a permissive threshold ($P < 2 \times 10^{-3}$). MACS identified broad regions of DNase accessibility. Therefore the PeakSplitter tool[46] was applied to identify constituent narrow peaks, each potentially representing a specific binding event. Finally, IDR[18] was used to identify reproducible peaks for each cell type (IDR $< 1 \times 10^{-2}$).

After identification of reproducible peaks, a simple heuristic was used to combine the peaks from multiple cell types to build a common atlas comprising peaks from all cell types. For simplicity, first, suppose that there are only two cell types. In this case, the set of non-overlapping peaks from the two cell types is first added to the atlas. Then the following heuristic is used to combine overlapping peaks: if the overlap between the peaks is >75%, the non-overlapping portions of the peaks are removed to create a single unified peak that is added to the atlas; if the overlap is <75%, the overlapping portions are removed to create two separate peaks that are both added to the atlas. This procedure can be extended to multiple cell types by first building the set of peaks for any two cell types and then combining this atlas with the third cell type, and so on. This procedure was extended to all the cell types to create the atlas of DNase peaks for subsequent analysis.

Assignment of DNase peaks to genes. The June 2013 RefSeq transcript annotation of the hg19 version of the human genome was used for the genomic location of transcription units. For genes with multiple gene models, the longest transcription unit was used for the gene locus definition. DNase peaks located in the body of the transcription unit, together with the 2-kb regions upstream of the TSS and downstream of the 3′ end, were assigned to the gene. If a peak was found in the overlap of the transcription units of two genes, one of the genes was chosen arbitrarily. Intergenic peaks were assigned to the gene whose TSS or 3′ end was closest to the peak. In this way, each peak was unambiguously assigned to one gene. This approach to associating enhancers with target genes has been used previously[47].

DNase peak atlas and enhancer categories. We found a total of 120,583 DNase peaks (reproducible across replicates), of which 51.4% were present in hESCs, 45.4% were present in HSPCs, 40.1% were present in monocytes, 30.9% were present in B cells, 30% were present in NK cells and 29.2% were present in T cells. Examining genomic locations, 44.6% of the peaks were found in introns, 41.9% were found in intergenic regions, 8.7% were found in promoters (5′ UTR and 2-kb upstream of the TSS), 2.9% were found in 3′ UTRs and 1.9% were found in coding sequences. The peaks were assigned to enhancer classes (DNase accessibility patterns) according to the cell types where they appeared accessible as well as their genomic locations (promoter and non-promoter). In this way, we found the largest classes to be hESC specific (31,119 peaks), monocyte specific (16,327 peaks) and ubiquitous (14,683 peaks).

We looked at the distribution of DNase accessibility patterns at different types of genomic loci and noticed that, of the 10,520 peaks that were located in the promoters of genes, 58% were ubiquitous (present in all cell types and in all cell types but monocytes; see note below) and 18% were accessible only in hESCs or HSPCs. Therefore, the overwhelming majority of promoter peaks that appear in a differentiated cell type (the focus of this study) are ubiquitous, and for this reason we included the ubiquitous promoter peaks as a DNase accessibility pattern. Non-promoter peaks had a much wider range of accessibility patterns, of which the most abundant ones constituted the major classes that were defined in this study: cell type–specific peaks (present in a differentiated cell type; 26,933), transition peaks (present in HSPCs and a differentiated cell type; 9,506), ubiquitous peaks (present in all cell types and in all cell types but monocytes; 11,997; see note below), hematopoietic peaks (present in all cell types but hESCs; 2,139) and lymphoid peaks (present in T cells, B cells and NK cells; 1,917). In the transition and cell type–specific classes, we included peaks that were present in both monocytes and B cells and also peaks present in both T cells and NK cells, as these categories have a considerable number of peaks. See **Supplementary Table 4** for numbers on the sharing of peaks by cell types.

We note that, in the ubiquitous classes (both promoter and non-promoter), we included peaks present everywhere but in monocytes because, upon reexamination of DNase-seq data, they exhibited non-negligible levels of DNase accessibility in monocytes (despite the lack of called reproducible peaks), and their histone modification signals were highly similar to the signals in ubiquitous peaks.

Histone modification ChIP-seq processing. Aligned histone modification ChIP-seq BAM files were downloaded from the Roadmap Epigenomics data portal (see **Supplementary Table 1** for accession codes and links to each library). Each data set was subjected to cross-correlation analysis for quality control, as described previously[48]. Briefly, it is assumed that a high-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest and that the sequence tag density accumulates to the left of the bound protein on the forward strand and to the right of the bound protein on the reverse strand, making the two strands appear shifted around the binding event. Therefore, one can compute the Pearson linear correlation between the plus and minus strands, after shifting the plus strand by k bp and expect to observe two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length ('phantom' peak). Metrics obtained from this plot and the presence/absence of these two peaks were used to flag and discard some data sets and also to experimentally estimate the fragment length in each library. In the end, we were able to collect data sets with high signal-to-noise ratio for the histone modifications H3K4me1, H3K4me3, H3K27ac and H3K27me3 in the six cell types of interest, although only in a few cases were we able to keep more than one replicate. When replicates remained available, they were pooled into one data set. These signals, after having been shifted by half the estimated fragment length, were used for counting enrichment of histone modifications at DNase peaks (400 bp in the flanking regions of the peak were included for counts) and for generating signal tracks (**Figs. 1** and **2**). Counts at DNase peaks were normalized to RPK values (to normalize for peak width) and then quantile normalized across different cell types. For signal tracks, counts of reads in bins of 200 bp in length were used in **Figure 1** and normalized to tags per million, and bins of 10 bp in length were used for the meta-peak signals in **Figure 2** and also normalized to tags per million.

DNase and histone modification heat maps. The DNase heat maps in **Figure 2** and **Supplementary Figures 6–8** were created by pooling replicates of DNase data sets and binning the 1-kb region around the peak summit in bins of 10 bp in length. Read counts at bins were normalized to tags per million, and these vectors were hierarchically clustered (for each enhancer class independently) using the hclust function of R with Euclidean distance. Heat maps were drawn using the function aheatmap from the package NMF[49]. For histone modification heat maps (**Supplementary Figs. 25** and **26**), a similar procedure was followed, but, for improved visualization, the dynamic range of the histone modification counts was linearly transformed to have the same dynamic range of DNase counts.

Gaussian mixture models. Two-dimensional Gaussian mixture models were used to identify subpopulations of DNase peaks according to their histone modification counts in each cell type. One analysis focused on H3K27ac versus H3K4me1 to study the poising of enhancers (**Fig. 3a**,**b** and **Supplementary Fig. 19**), and a second analysis focused on H3K4me3 versus H3K27me3 to study bivalent peaks and their fates in differentiation (**Supplementary Figs. 23–26**). In each case, log-transformed counts were used to fit a Gaussian

doi:10.1038/ng.3402

mixture model with three components using the R package mixtools[50]. The function mvnormalmixEM, which runs the expectation-maximization algorithm for fitting, was used with random initialization.

In the case of H3K4me3 versus H3K27me3 in hESCs (**Supplementary Fig. 23**), we noticed that the three Gaussians captured two subpopulations of interest (one with both signals low and one with high H3K4me3 and low H3K27me3) and a third subpopulation that encompassed the background. Therefore, to capture bivalent peaks, we ran the expectation-maximization fit again with four components, initializing it with the three components learned in the first pass and a fourth centered around bivalent peaks (by visual inspection). Once the bivalent component was captured, bivalent peaks (blue dots in **Supplementary Fig. 23**) were defined as those having posterior probability greater than 0.50 in this component. In all the fits, probability level curves with $P = 0.90$ were used for visualizing the different subpopulations. These curves enclose all the points $\vec{x}$ that satisfy:

$$(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu}) \le \chi_k^2(p)$$

where $\vec{\mu}$ is the mean of the subpopulation, Σ is the covariance matrix and $\chi_k^2(p)$ is the quantile function for probability p (in this case, 0.90) of the χ^2 distribution with k degrees of freedom, where k is the number of dimensions. Because here k is 2, the χ^2 distribution simplifies to an exponential distribution with a mean of 2.

Detailed description of diamond plots. In **Figure 3c**,**d**, we illustrate how up (down) expression changes in cell state transitions are accompanied by gains (losses) of active DHSs in the regulatory loci of genes. Both panels show gene expression changes and enhancer changes in low-complexity genes (**Fig. 3c**) and high-complexity genes (**Fig. 3d**) for the transition from HSPCs to monocytes. Each gene is illustrated by a stack of diamonds, where a diamond represents a DHS associated with the gene. The bottom of the stack corresponds to the log-transformed fold change in expression for the gene (y axis). Red diamonds are active peaks gained in the transition, with a significant change in DNase accessibility (FDR < 1%, 0 < log(fold change)) and the log-transformed fold change in H3K4me1 or H3K27ac levels greater than 1. Blue diamonds are peaks lost in the transition, similarly defined as having a significant change in DNase accessibility (FDR < 1%, log(fold change) < 0) and log-transformed fold change in H3K4me1 or H3K27ac levels smaller than −1. Peaks that are maintained in the transition are represented by yellow and green diamonds, defined as DNase peaks present in both cell types (no significant change at FDR of 2%) with a log-transformed fold change in either H3K4me1 or H3K27ac levels greater than 1 (yellow diamonds) or smaller than −1 (green diamonds). Gray dots represent peaks present in both cell types without changes in histone marks. In solid colors are peaks that satisfy the previous conditions but also are assigned to different monocyte enhancer types: cell type specific (red), HSPC specific (blue) or transition (yellow and green). Restricting to the top 10% most expressed genes in both cell types, **Figure 3c**,**d** shows for each complexity group only the 100 genes with the largest absolute log(fold change).

CTCF ChIP-seq processing. We observed that non-promoter ubiquitous peaks displayed a markedly bimodal distribution of H3K4me3 in all the differentiated cell types (**Supplementary Figs. 13** and **14**); H3K27ac had a similar behavior but to a smaller extent (**Supplementary Figs. 10** and **11**). We hypothesized that the subpopulation with low active signals should be enriched for structural DHSs, and, to test this possibility, we measured CTCF signals at these loci. By visual inspection of the monocyte H3K4me3 density plot at ubiquitous peaks, we defined the 'structural' group as the peaks with $\log_2$-transformed H3K4me3 signal between 3.8 and 4.2 (873 peaks) and the 'active' group as the peaks with signal between 8.8 and 9.2 (721 peaks). We downloaded aligned BAM files for two biological replicates of CTCF ChIP-seq in the lymphoblastoid cell line GM12878 from the ENCODE web portal (accession ENCSR000DRZ), counted reads at peaks and transformed to RPK values to normalize for peak width. The signals for the two groups of peaks were compared with empirical cumulative distribution functions (**Supplementary Fig. 16**).

SeqGL overview. SeqGL is a novel group-lasso–based *de novo* motif discovery algorithm to extract multiple transcription factor binding signals from ChIP-seq, DNase-seq and ATAC-seq profiles[26]. SeqGL identifies these signals by training a discriminative model to differentiate between sequences in peaks versus flanking sequences using a k-mer feature representation. The clustering of these k-mer features across training examples showed a block structure, and this block structure is encoded as a group lasso constraint in the binary classification problem. Formally, this can be written as:

$$\min_{\mathbf{w}} \sum_i \log(1+\exp(-y_i \mathbf{w}\cdot\mathbf{x}_i)) + \lambda_1 \sum_g \|\mathbf{w}_g\|_2 + \lambda_2 \sum_m |w^m|$$

where $\mathbf{x}_i$ represents the k-mer count vector for example i and y_i are the labels: +1 for peaks and −1 for flanking regions. The first summation is over all the examples (peaks and flanking regions) and defines the logistic loss function. The second summation encodes the group lasso constraints across all k-mer weights $\mathbf{w}_g$ for all groups g. Here $\mathbf{w}_g$ is a vector of k-mer weights that belong to group g. The third constraint encodes the sparsity constraints over all k-mers and sets the weight of non-informative k-mers to zero. The two regularization parameters λ_1 and λ_2 control sparsity at the group level and k-mer level, respectively. The prediction scores of flanking regions are used as an empirical null distribution to identify the subset of peaks best predicted by a group that discriminates peaks from flanking regions. HOMER is then run on this subset of peaks to associate a transcription factor with each group.

Benchmarked on over 100 ChIP-seq experiments, SeqGL outperformed traditional motif discovery tools in discriminative accuracy. SeqGL also successfully scaled to DNase-seq or ATAC-seq maps, identifying numerous transcription factor signals confirmed by ChIP, including those missed by conventional motif finders[26]. SeqGL was run by sampling 40,000 subpeaks of the cell type identified using PeakSplitter. Two hundred groups were used for the analysis. Group scores and motifs were identified using the scores of the flanking regions as empirical null distributions.

Peak category definitions for regression analysis. To refine peak categories for regression analysis, edgeR[32] was performed on all pairs of cell types using only peaks defined in either of the two cell types under consideration. A peak was identified as significantly differential if it satisfied FDR-corrected $P < 1 \times 10^{-2}$ and absolute DNase fold change >1. A number of DNase peaks were observed to be shared by (i) B cells and monocytes and (ii) T cells and NK cells (**Supplementary Table 4**). Hence, this information was used in defining the peak categories (**Fig. 4b**), as described below.

The B cell peak categories were defined as follows: (i) promoter, peaks in the promoter regions of genes, defined as the 2-kb window up- and downstream of the TSS; (ii) ubiquitous, peaks with no significant change in all comparisons of B cells versus HSPCs, T cells and NK cells; (iii) transition, peaks with a significant change in all B cells versus T cells and NK cells; and (iv) cell type–specific: peaks with a significant change in all B cells versus HSPCs, T cells and NK cells. A similar logic was used to categorize peaks in the other cell types.

Predictive model of gene expression changes. SeqGL scores were computed for all the subpeaks. Broad peak scores were computed by summing the constituent subpeak scores. The number of nonzero k-mer groups identified in each cell type is an outcome of the SeqGL optimization algorithm. In particular, HSPC SeqGL scores for 53 nonzero k-mer groups were used for HSPC-specific peaks and transition peaks, and B cell SeqGL scores for 48 groups were used for cell type–specific, transition, ubiquitous and promoter peaks. Similarly, ~50 groups were identified by SeqGL in each other cell type. These scores were used to learn the weights of transcription factor signals occurring in different peak categories using a regression model.

Each category of peak (promoter, ubiquitous, cell-type specific, HSPC specific and transition) was represented by all SeqGL transcription factor features from the relevant cell types (that is, for B cell specification, B cell features were used for all peak categories except HSPC-specific peaks, and HSPC features were used for all peak categories except B cell–specific peaks). For each transcription factor signal and peak category, all the SeqGL scores for this signal across all the 'active' peaks belonging to the category and assigned to a gene were summed up, and this sum was used as the

doi:10.1038/ng.3402

corresponding feature value for the gene; active peaks were those marked with either H3K4me1 or H3K27ac (normalized signal above the 75th percentile among all DHSs in the cell type). Therefore, the extent of DNase accessibility was not considered in the model; all reproducible DNase peaks with active histone marks in a particular category were treated equally. The multiplicity of peaks containing a transcription factor signal did figure into the model, for example, if a gene gained many B cell–specific peaks, all containing an IRF signal, than the corresponding feature value (IRF-binding signal in B cell–specific peaks) would be high because it would be the sum of multiple IRF SeqGL scores.

Transcription factor binding site identification was used to turn each gene's set of assigned (active) DNase peaks into a feature vector of binding signals. As described above (and as we illustrate in **Fig. 5a**), binding scores for each transcription factor were summed across each category of peaks. Therefore, for example, the information in the set of ubiquitous peaks assigned to a gene was summarized as a feature vector of TF scores, one for each SeqGL TF signal. If there are multiple peaks in a category, information about which specific peak the signal came from was not retained. Similarly, while we use second-order features in the model (for example, PU.1 signals and GATA signals in HSPC-specific peaks), this is the product of the summary features for PU.1 and for GATA in HSPC-specific peaks rather than co-occurrences of these signals in individual peaks. Retaining more information about the sequence signals in individual peaks and interactions between peaks could be a direction for future work, perhaps by integrating high-resolution Hi-C data. For now, the model is sufficient to learn the major transcription factors that explain gene expression changes and the categories of peaks through which they act. Ridge regression models were trained using the SPAMS package[51], and a regression model was trained separately for the transition from HSPCs to each differentiated cell type.

DNA signals in poised B cell and HSPC genes. Active HSPC and B cell peaks (including cell type–specific, promoter, transition and ubiquitous) peaks were first identified in both poised B cell and HSPC genes. The B cell and HSPC SeqGL models were used to predict sequence signals in these peaks.

Gene ontology analysis. Gene ontology analysis was performed using the DAVID tool[52]. All human genes were used as background, and analysis was performed using the biological process ontology terms.

Iterative reassignment of enhancers. Genes with high cross-validation errors were first identified (error >85th percentile). For each high-error gene, the nearest marginally expressed gene (expression <75% percentile, normalized RPM threshold = 9.5) neighboring the high-error gene's currently defined regulatory locus was identified, and the enhancers of the non-expressed gene were assigned the expressed gene. For high-error genes with positive fold change, differentiated cell expression was used to identify non-expressed neighbors; HSPC expression was used to identify non-expressed genes for high-error genes with negative fold change. Tenfold cross-validation was then performed with the new enhancer assignments. This process was repeated until the improvement in cross-validation was less than 1×10^{-2}.

43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
45. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
46. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11**, 415 (2010).
47. McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
48. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
49. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
50. Benaglia, T., Chauveau, D., Hunter, D.R. & Young, D.S. mixtools: an R package for analyzing mixture models. *J. Stat. Soft.* **32**(6), 1–29 (2009).
51. Mairal, J., Bach, F., Ponce, J. & Sapiro, G. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.* **11**, 19–60 (2012).
52. Huang, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

doi:10.1038/ng.3402