Published in final edited form as:

Nat Genet. 2015 November ; 47(11): 1249–1259. doi:10.1038/ng.3402.

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

Alvaro J. González[1,2], **Manu Setty**[1,2], and **Christina S. Leslie**[1]

[1]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY

Abstract

We carried out an integrative analysis of enhancer landscape and gene expression dynamics in hematopoietic differentiation using DNase-seq, histone mark ChIP-seq, and RNA-seq to model how early establishment of enhancers and regulatory locus complexity govern gene expression changes at cell state transitions. We found that high complexity genes – i.e. those with 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 novel mechanistic role for PU.1 at transition peaks in B cell specification and can be used to correct enhancer-gene assignments.

Introduction

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 (Treg) cell transcriptional program – based on their occupancy by the Treg cell master regulator Foxp3 – were DNase accessible in CD4+ precursor cells, occupied by other factors that "place-hold" to maintain the potential for Treg cell differentiation[1]. Evidence in support of early enhancer establishment or chromatin poising has also been documented in B cell and

Correspondence should be addressed to: C.S.L. (cleslie@cbio.mskcc.org).

[2]These authors contributed equally to this work.

URLs

Supplementary website: http://cbio.mskcc.org/public/Leslie/Early_enhancer_establishment/

SegGL source code repository: https://bitbucket.org/leslielab/seqgl

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/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 of the project, advised on the analysis and algorithm development, supervised the research, and wrote the manuscript.

Competing Financial Interests Statement

The authors have no competing financial interests.

macrophage specification[2,3], T cell development[4], early hematopoiesis[5], and multipotent endoderm cells at enhancers associated with liver and pancreas cell fates[6]. Earlier concepts of poising include bivalent domains in embryonic stem cells (ESCs), where the active mark H3K4me3 and repressive mark H3K27me3 coincide[7]; other poised ESC elements marked by H3K4me1 and H3K27me3[8]; and poised/inactive enhancers marked with H3K4me1 but not H3K27ac[9].

Meanwhile, recent studies have defined the notion of cell-type specific "super-enhancers" – spatially clustered enhancers, occupied by master regulator transcription factors (TFs) for the cell type – that regulate developmentally important genes[10,11]. Others have used segmentation of histone mark data to identify long (>3kbp) "stretch enhancers"[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 DNaseI hypersensitive sites (DHSs) regulating a gene across a lineage. We investigate how locus complexity and early enhancer establishment in hematopoietic differentiation work together to shape transcriptional programs and quantitatively determine gene expression changes in 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 DNase-seq, histone modification ChIP-seq for multiple marks (H3K27ac, H3K27me3, H3K4me1, H3K4me3), and RNA-seq data in order to link enhancer dynamics and spatial organization to gene expression changes in hematopoietic differentiation. We focused on six cell types characterized by the Roadmap Epigenomics project[15–17] (Supplementary Table 1): human embryonic stem cells (hESC), hematopoietic stem and progenitor cells (CD34+ HSPC), one myeloid cell type (CD14+ monocytes), and three lymphoid lineages (CD19+ B cells, CD3+ T cells, CD56+ 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 120K reproducible DNase hypersensitive sites (DHSs, median width = 456bp; Supplementary Fig. 1, Online Methods). 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 atlas DHSs, over all cell types, assigned to it. Nearest-gene enhancer assignment can incur 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 2Kbp upstream of the TSS to 2Kbp downstream of the last annotated 3′ end) and an additional 10% lie within 10Kbp of the target gene. Therefore, in a majority of cases, we assign a 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*[19,20], and demonstrated that they indeed regulate these genes. Furthermore, recent high-resolution capture Hi-C analysis found that transcriptionally active promoters asymmetrically interact with the gene body more than with the local upstream sequence, suggesting an activating role for intronic enhancers[21]. However, intronic enhancers sometimes regulate a nearby gene rather than the encompassing gene[22–24], and therefore our assignment of intronic DHSs is imperfect. We will return to the problem of improving enhancer-gene assignments later.

We next grouped genes into three equally sized classes (tertiles) based on their regulatory complexity: low complexity genes, with a total assignment of 0 to 2 atlas DHSs; medium complexity genes, with 3 to 7 DHSs; and high complexity genes, with 8 or more DHSs. Fig. 1a shows an example of a low complexity gene, *NUP54*, a housekeeping gene that encodes a component of the nuclear pore complex; and a high complexity gene, *RUNX1*, a developmentally important hematopoietic transcription factor. *NUP54* has a single DHS assigned to it, a promoter peak that is accessible and has active histone marks (H3K4me3 and H3K27ac) across all cells; 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* is assigned 43 DHSs, including several promoter peaks and a large number of intronic enhancers across its very long transcription unit (only the first four introns are shown); its enhancers display dynamic patterns of accessibility and chromatin marks across cell types; and it achieves dramatic expression changes in specific cell state transitions.

Complexity tied to gene architecture, function, expression

In fact, the differences in gene architecture, function, and expression dynamics between *NUP54* and *RUNX1* are characteristic of the differences between low and high complexity genes. As one might expect, the log length distribution of the transcription unit for high complexity genes is significantly greater than that of medium complexity genes, which are longer than low complexity genes (Fig. 1b; $P < 1\times10^{-16}$, one-sided Kolmogorov-Smirnov test for both comparisons). Furthermore, the fraction of the transcription unit consisting of intronic sequence is significantly higher in high versus medium complexity genes and in medium versus low complexity genes (Fig. 1c; $P < 1\times10^{-16}$, one-sided KS tests) since higher complexity genes have longer introns (Supplementary Fig. 2). High complexity genes that are highly expressed (top 10% of expression distribution) in a cell type are strongly enriched for cell type specific functions such as hemopoiesis and leukocyte activation (HSPCs), lymphocyte activation (B cells), and immune response (monocytes), while highly expressed low complexity genes are enriched for similar housekeeping functions, such as translation and ubiquitination, in all cell types (Supplementary Table 3; Online Methods). The distribution of genes over complexity classes and properties of these classes also hold for the top 10% 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.

Importantly, genes in low, medium, and high complexity classes show significantly increasing dynamic range of gene expression across cell types (Fig. 1d; $P < 1\times10^{-4}$, one-sided KS tests for low vs. medium and medium vs. high). Moreover, when examining highly expressed genes in cell state transitions such as the specification of HSPCs to monocytes, high complexity genes achieve the largest increases in expression, with medium and low complexity genes showing progressively less upregulation (Fig. 1e; $P < 2\times10^{-7}$ for high vs. medium, $P < 4\times10^{-6}$ for medium vs. low, one-sided KS tests; see Supplementary Fig. 4 for other cell state transitions). These results suggest that high regulatory complexity may help to achieve large expression changes in differentiation.

Given our nearest-gene DHS assignments, we also investigated how gene density interacts with our complexity definitions. For each gene, we computed the number of genes on either strand overlapping a 1M window centered at the gene (Fig. 1f) or in two 500KB windows flanking each gene (Supplementary Fig. 5a). Both analyses showed that high complexity genes occur in low density regions, while low complexity genes occur in gene dense regions. We considered if 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 are promoter peaks or ubiquitous peaks, and that the percentage of promoter/ubiquitous peaks increases 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 gene complexity classes. The heatmap in Fig. 2a shows the most frequent patterns of accessibility found among ~48K DHSs present in monocytes (minor patterns omitted). Major accessibility patterns include promoter DHSs that are 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 are present only in monocytes. Finally, another major accessibility pattern we identified consisted of “transition peaks” (12.6%), enhancers that are established in HSPCs and maintained in monocytes but lost in other fates. Similar early enhancer establishment was suggested through histone mark analysis in early hematopoiesis[5]. In contrast to promoter and ubiquitous peaks, transition and cell-type specific peaks are preferentially found in regions with low gene density (Supplementary Fig. 5b, red dots). Examination of lymphoid cell types yielded similar major patterns, with the addition of a lymphoid-restricted accessibility pattern (Supplementary Fig. 6–8). In particular, each differentiated cell type had a distinguished set of transition enhancers, established in HSPCs and maintained upon specification to that cell fate but lost in others. Since our analysis identified a substantial number of DHSs shared between B cells and monocytes, and between T and NK cells, we allowed sharing of peaks between these two pairs of cell types when defining “cell-type

specific" and "transition" peaks, provided they were not found in other differentiated cell types (Online Methods, Supplementary Table 4).

Meta-peak visualizations of the major DNase accessibility patterns for monocyte peaks, together with average chromatin signals across cell types, are shown in Fig. 2b–f. DNase and histone mark distributions within each pattern (Supplementary Fig. 9–15) show that accessibility is highest in promoter peaks and lowest in cell-type specific peaks, and that H3K4me1 is highest in transition peaks, followed by both cell-type specific and hematopoietic peaks. Ubiquitous peaks display a bimodal distribution of the active marks H3K4me3 and H3K27ac (Supplementary Fig. 10–11, 13–14), suggesting that a subpopulation of these DHSs are not active transcriptional enhancers. When we overlaid CTCF ChIP-seq from the ENCODE lymphoblastoid cell line GM12878 with ubiquitous DHSs, we saw dramatic enrichment of CTCF signal at low H3K4me3 ubiquitous peaks (Supplementary Fig. 16, Online Methods), suggesting their a role in 3D chromatin structure. An early ENCODE study of 1% of the human genome across 6 unrelated cell types found that almost all ubiquitous DNase peaks were promoter peaks (86%), and most of the remainder were CTCF bound (10%)[25]; our lineage-based analysis reveals a larger class of ubiquitous non-promoter DHSs with active histone marks and without CTCF occupancy. Similar observations can be made for analogous enhancer types in other differentiated cells (Supplementary Fig. 9–15).

Strikingly, DHSs assigned to genes in different complexity classes are enriched for distinct lineage-wide DNase accessibility patterns. DHSs of low complexity genes are dominated by promoter and ubiquitous peaks in monocytes (Fig. 2g) and B cells (Fig. 2h) as well as other cell fates (Supplementary Fig. 17). Meanwhile, DHSs of high complexity genes are strongly enriched for both cell-type specific and transition peaks (Fig. 2g, $P < 2\times10^{-107}$ and $P \sim 0$ for transition and monocyte specific peaks, Fisher's test; Fig. 2h, $P < 5\times10^{-85}$ and $P < 2\times10^{-153}$ for transition and B cell specific peaks). 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. While cell-type specific enhancers change their accessibility, transition peaks display a change in chromatin marks. Fig. 3a, b show 2D density plots of the active mark H3K27ac[9] vs. the mark H3K4me1 for all DNase peaks in HSPCs and monocytes (see Supplementary Fig. 18 for other cell types). Gaussian mixture modeling defines three overlapping subpopulations (Online Methods): peaks with low activity/moderate H3K4me1, moderate activity/high H3K4me1, and high activity/moderate H3K4me1. Transition peaks for monocyte specification are indicated in blue. In HSPCs, monocyte transition peaks are strongly enriched in both the low activity/moderate H3K4me1 and moderate activity/high H3K4me1 subpopulations. Upon specification to monocytes, the transition peaks largely move to the moderate activity/high H3K4me1 region of the monocyte landscape (Fig. 3b) and display a strong increase in H3K27ac (Supplementary Fig. 19); transition peaks for other cell fates show the same patterns (Supplementary Fig.

20–22). A partial analogy can be made to bivalent domains in hESC cells[7], some of which coincide with DNase peaks and form a subpopulation of DHSs with moderate active promoter mark H3K4me3/high repressive mark H3K27me3 in the corresponding 2D hESC landscape (Supplementary Fig. 23) but resolve to active or inactive DHSs upon specification to HSPCs (Supplementary Fig. 24–26).

We then examined the 10% most highly expressed genes in HSPCs and monocytes and identified the low and complexity genes with largest expression changes in HSPC to monocyte specification. While a handful of low complexity genes appeared to achieve large expression changes without any gain/loss of DHSs (Fig. 3c and Supplementary Fig. 27–29), large increases/decreases in expression at high complexity genes were consistently associated with gains/losses in active cell-type specific enhancers, almost always accompanied by gain/loss of H3K4me1/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 while losing more HSPC peaks, and achieved higher log fold changes in monocyte specification (Supplementary Fig. 30a, b, c). These results suggest that regulatory complexity is important for achieving cell-type specific gene expression; however, these complex locus control regions are not typically established in a single cell state transition but rather are primed through early enhancer establishment.

SeqGL predicts TF binding signals from cell type DHSs

We next asked whether we could predict 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 flanks and clusters *k*-mers by their count vectors to identify *k*-mer groups. Fig. 4a shows part of the clustered count matrix obtained by applying SeqGL to B cell DNase-seq data. Each group tends to contain sequence-similar *k*-mers that may represent variants of the DNA recognition signal of a single TF, 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 flanks[26] (Online Methods). Groups with primarily positive weights represent predicted TF binding signals found in peaks (Fig. 4a), and for each of these groups, we can identify highly scored peak regions, generate a binding motif, and compare to existing databases to identify the corresponding TF, if found (Fig. 4a, Online Methods).

In a previous analysis of ENCODE DNase-seq data, SeqGL was more sensitive than traditional motif discovery methods at identifying TF 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 TF binding signals mapping to 25–34 distinct known TF motifs per cell type (Supplementary Table 5). While we do not have parallel TF ChIP-seq data for these cell types, SeqGL retrieved many TF binding signals that are supported by the literature and previous ChIP-seq studies, including 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, T cells and monocytes[28]; myeloid zinc factor MZF in HSPCs and monocytes[29]; CEBP in monocytes[30]; and STAT exclusively in T cells[31]. However, since SeqGL extracts DNA sequence signals from DNase-seq and relies on motif databases to associate signals to TFs, it cannot distinguish between TFs with nearly identical binding motifs.

Fig. 4b gives an example of SeqGL TF binding predictions for DHSs in the *PAX5* locus after training the model on DNase peaks in CD19+ B cells. Each track below the DNase signal track shows the SeqGL group *k*-mer score for the indicated TFs; for RUNX, *k*-mer patterns contributing to the scoring model are shown. We analyzed SeqGL's TF score distributions 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), transition, and cell-type specific (Fig. 4c). As expected, the NFY sequence signal is enriched in B cell promoter peaks (Fig. 4d), CTCF in ubiquitous as well as promoter peaks (Fig. 4e), and IRF in B cell specific peaks (Fig. 4f). Interestingly, PU.1 is enriched not only in cell-type specific peaks but also in transition peaks (Fig. 4g). While PU.1 is known to play a key role in both HSPC and B cell function[2,33], the binding of PU.1 in 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 TF sequence signals from DNase-seq data in HSPCs and the differentiated cell type. In the model for B cell specification, each gene is represented by a set of SeqGL-predicted TF binding scores (features) computed from its active (H3K4me1 or H3K27ac marked) DHSs in HSPCs and B cells (Fig. 5a, Online Methods). The model learns a regression coefficient for each TF signal or pair of signals in each category of DHS (promoter, ubiquitous, HSPC specific, B cell specific, transition) to allow these signals to play distinct roles in predicting expression changes. For example, TF signals in HSPC and B-cell specific peaks represent regulatory inputs that are lost or gained, respectively, in the cell state transition; transition peaks may be bound by TFs in HSPCs that act as "place-holders" for later binding by other TFs in B cells or may be continuously occupied by the same TFs in both cell types. The contribution of TFs in these different roles to 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 in the B cell transition were remarkably accurate (Spearman correlation $\rho = 0.666$), especially for the high complexity genes ($\rho = 0.742$) which have the largest expression changes

(Supplementary Fig. 31). Low complexity genes have more modest expression changes, and the model behaved appropriately for these genes by predicting smaller log fold changes (Supplementary Fig. 31). Similarly strong prediction performance was obtained for transitions to other cell fates (Supplementary Table 6).

Analysis of the regression models reveals the specialized roles of different kinds of enhancers (Supplementary Fig. 32). The first heatmap in Fig. 5b shows the contribution of TFs residing in different DHS categories for predicting gene expression changes of high complexity genes in B cell specification, and the remaining heatmaps show other cell state transitions. In these heatmaps, each row represents a high complexity gene and every column represents a TF signal or pair of TF 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 if it contributes positively (orange) or negatively (blue) to the overall predicted log expression change. In some cases, several *k*-mer groups are associated with the same TF by SeqGL and can lead to multiple columns labeled with the same TF, sometimes within the same DHS category. Some of these multiple *k*-mer groups for a TF may represent subtle differences in the underlying motifs and perhaps merit experimental examination, while others are likely a result of over-clustering (Supplementary Fig. 33), which nonetheless does not harm the predictive performance of the regression model. Using second order features significantly improves performance over first order features alone ($P < 5\times10^{-7}$, Wilcoxon signed rank test) but only fine-tunes the predictions. The model identifies known HSPC factors like GATA and RUNX[34] in HSPC specific peaks as predictors of negative log fold change for genes specifically expressed in HSPCs; similarly, known B cell factors like IRF[34] in B cell specific peaks predict positive log fold changes in B cell genes. Interestingly, binding signals for PU.1 in HSPC and B cell peaks are associated with downregulation of HSPC genes and upregulation of B cell genes, respectively; moreover, a subset of B cell genes harbor PU.1 signals in transition peaks that strongly predict their expression changes. This finding suggests a novel mechanistic role for 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 to 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 TF binding to both B cell specific and transition enhancers. Among the poised genes were a number with important B cell functions, such as the anti-apoptotic 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 have significantly higher DNase accessibility and lower H3K27me3 repressive marks compared to their B cell specific peaks in HSPCs ($P < 3\times10^{-93}$, Wilcoxon rank sum test) and are close in chromatin state to the HSPC specific peaks of HSPC genes (Supplementary Fig. 34; $P < 4\times10^{-3}$, Wilcoxon rank sum test). This suggests that PU.1 binding at transition enhancers – the main predictive

transition peak signal – may help maintain the chromatin in an open state. However, poised B cell genes have much lower expression in HSPCs than HSPC genes (Supplementary Fig. 34; $P < 3\times10^{-93}$, Wilcoxon rank sum test). To explain this, we confirmed that poised genes have reduced signals for key HSPC-specific factors like GATA in their HSPC and B cell active enhancers while harboring strong B cell specific TF signals like IRF in their B cell active enhancers (Fig. 5e, Online Methods). Therefore, while 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 reveals a role for PU.1 poising, but here the PU.1 signal arises in promoter peaks and may maintain open chromatin for later binding by the promoter-associated factor CEBP, a key regulator of monocyte/macrophage fate (Supplementary Fig. 35–36).

Regression model nominates gene-enhancer reassignments

While 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 may 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 ($P < 2\times10^{-3}$, 1×10^{-9}, 3×10^{-7}, respectively, Wilcoxon signed rank test; Fig. 6b, Supplementary Fig. 39). Similarly, using enhancer-gene reassignments from other regression models led to significantly improved regression performance in independent cell types in all but one case (Supplementary Fig. 39; Supplementary Table 7). As an example, the B cell marker gene *CD20* resides in a gene dense region where no nearby genes are expressed in B cells. Originally assigned 2 active DNase peaks, it acquires 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 domains of Mediator ChIP-seq (for hESCs) or H3K27ac (for hematopoietic cell types[10]), though 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 0 or 1 DHSs, while ~41% of HSPC super-enhancers and >30% of lymphoid cell super-enhancers contain 2 or fewer DHSs (Supplementary Fig. 42). Therefore, current definitions of super-enhancers do not always identify complex locus control regions with many individual enhancers, as others have also observed[35]. 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 their DHS multiplicity across a lineage. This definition may be more natural from a mechanistic and evolutionary standpoint by characterizing a gene's regulatory potential and the cell-type dependent constraints on that potential. While we divide genes into low, medium, and high complexity classes to simplify our analysis, there may be no intrinsic threshold dividing super-from non-super-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 distinguished role of PU.1 transition enhancers in B cell specification. We found that presence of a transition peak in high complexity genes is associated with higher gain of 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 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; like transition peaks, these DHSs were enriched in high complexity genes (Supplementary Fig. 43). While limited by available data, we have shown proof-of-principle results that complex locus control and early establishment of enhancers play important roles in transcriptional programs in differentiation. Importantly, 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 expression.

Our regulatory model starts by simply assigning DHSs to the nearest gene. There is disagreement in the literature about how often this rule is incorrect. An early ENCODE 5C analysis on 1% of the human genome concluded that only ~7% of looping interactions are with the nearest gene[36]. By contrast, recent Cohesin ChIA-PET interaction data in mouse ES cells supported the assignment of enhancers to the proximal active gene in a large majority of cases (83% of super-enhancers, 87% of typical enhancers) and also largely supported assignment of enhancers to single rather than multiple genes[37], 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, while the remaining interactions pass over intervening active or inactive promoters[21]. These studies are largely consistent with our initial assumptions, although definitive interaction data is 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[38–40] – 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 activity of enhancers with target gene expression across unrelated cell types without using a gene regulation model[38,41,42]. Despite nearest-gene enhancer assignment, our regression model predicts gene expression changes with high accuracy, and larger prediction errors for a minority of genes suggested enhancer reassignments that reduced the loss in independent cell types. As new technologies enable profiling of fine-grained cell populations using small cell numbers[5,43], our approach may point the way forward to improved enhancer annotation.

Online Methods

Data and preprocessing

Aligned DNase-seq bam files were downloaded from the Roadmap Epigenomics data portal (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/). RNA-seq data was downloaded as fastq files and aligned to hg19 using STAR[44]. 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[45] was used to determine the number of reads mapping to each gene. Reads per kilobase (RPK) was calculated for each gene, and the 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 Fig. 1. Peak calling was performed on each cell type individually: first the reads from different replicates were pooled, and then the MACS peak caller[46] was used to identify peaks with a permissive threshold ($P < 2\times10^{-3}$). MACS identified broad regions of DNase accessibility. Therefore the PeakSplitter tool[47] 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 < 1e-2).

After identification of reproducible peaks, a simple heuristic was used to combine 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 are 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 which is added to the atlas; if the overlap is < 75%, the overlapping portions are removed to create two separate peaks which 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 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 2kb regions upstream of the TSS and downstream of the 3′ end, were assigned to the gene. If a peak is found in the overlap of the transcription units of two genes, one of the genes is chosen arbitrarily. Intergenic peaks were assigned to the gene whose TSS or 3′ end is closest to the peak. In this way, each peak was unambiguously assigned to one gene. This approach to associating enhancers to target genes has been used previously[48].

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 hESC cells, 45.4% in HSPCs, 40.1% in monocytes, 30.9% in B cells, 30% in NK cells and 29.2% in T cells. Examining genomic locations, 44.6% of the peaks were found in introns, 41.9% in intergenic regions, 8.7% in promoters (5′UTR and 2kb upstream of TSS), 2.9% in 3′UTRs and 1.9% 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 are 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 hESC cells 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 hESC, 2,139) and lymphoid peaks (present in T, B 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 and NK cells, since these categories have a considerable number of peaks. See Supplementary Table 4 for numbers on sharing of peaks between cell types.

Note: 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 modifications 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 numbers and links to each library. Each data set was subjected to cross-correlation analysis for quality control, as described previously[49]. 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 forward strand, and to the right 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* base pairs, 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/me3 and H3K27ac/me3 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 (Fig. 1 and Fig. 2). Counts at DNase peaks were normalized to reads per kilobase (to normalize for peak width) and then quantile normalized across different cell types. For signal tracks, counts of reads in bins of length 200bp were used in Fig. 1 and normalized to tags per million; and bins of length 10bp were used for the meta-peak signals in Fig. 2 and also normalized to tags per million.

DNase and histone modification heat maps

The DNase heatmaps in Fig. 2 and Supplementary Fig. 6–8 were created by pooling replicates of DNase data sets and binning the 1kbp region around the peak summit in bins of length 10bp. 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. Heatmaps were drawn using the function "aheatmap" from the package "NMF"[50]. For histone modification heatmaps (Supplementary Fig. 25–26), 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 vs. H3K4me1 to study the poising of enhancers (Fig. 3a, b and Supplementary Fig. 19), and a second analysis focused on H3K4me3 vs. H3K27me3 to study bivalent peaks and their fates in differentiation (Supplementary Fig. 23–26). In each case, log transformed

counts were used to fit a Gaussian mixture model with 3 components using the R package "mixtools"[51]. The function "mvnormalmixEM", which runs the EM algorithm for fitting, was used with random initialization.

In the case of H3K4me3 vs. H3K27me3 in hESCs (Supplementary Fig. 23), we noticed that the three Gaussians capture two subpopulations of interest (one with both signals low, and one with high H3K4me3 and low H3K27me3) and a third subpopulation that encompasses the background. Therefore, to capture bivalent peaks, we ran the EM fit again with four components, initializing it with the 3 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:

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

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 chi-squared distribution with k degrees of freedom, where k is the number of dimensions. Since here k is 2, the chi-squared distribution simplifies to an exponential distribution with mean 2.

Detailed description of diamond plots

In Fig. 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 (c) low complexity genes and (d) high complexity genes for the transition from HSPCs to monocytes. Each gene is illustrated by a stack of diamonds, where a diamond represents a DHS associated to the gene. The bottom of the stack corresponds to the logFC of the gene (y-axis). Red diamonds are active peaks gained in the transition, with a significant change in DNase accessibility (FDR < 1%, 0 < logFC) and logFC of H3K4me1 or H3K27ac greater than 1. Blue diamonds are peaks lost in the transition, similarly defined as having significant change in DNase accessibility (FDR < 1%, logFC < 0) and logFC of H3K4me1 or H3K27ac 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 logFC in either H3K4me1 or H3K27ac greater than 1 (yellow diamonds) or smaller than −1 (green diamonds). Grey 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, panels (c) and (d) show for each complexity group only the 100 genes with largest absolute logFC. The box-and-whisker plots on the right show the logFC distributions for all the genes; the box represents the interquartile range (IQR), and the whiskers are placed at distance 1.5*IQR from the corresponding quartile. Gene names have been annotated with [t]

and [s] whenever the gene has annotation terms associated with transcription or signaling, respectively. In the high complexity group, greater changes in gene expression are associated with large numbers of gained/lost peaks, typically in the presence of transition peaks.

CTCF ChIP-seq processing

We observed that non-promoter ubiquitous peaks display a markedly bimodal distribution of H3K4me3 in all the differentiated cell types (Supplementary Fig. 13–14); H3K27ac has a similar behavior but to a smaller extent (Supplementary Fig. 10–11). 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 log2 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 reads per kilobase 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 TF binding signals from ChIP-, DNase- and ATAC-seq profiles[26]. SeqGL identifies these signals by training a discriminative model to differentiate between sequences in peaks vs. flanking sequences using a *k*-mer feature representation. The clustering of these *k*-mer features across training examples revealed 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:

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

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 flanks. The first summation is over all the examples (peaks and flanks) 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 flanks is used as an empirical null distribution to identify the subset of peaks best predicted by a group that discriminates peaks from flanks. HOMER is then run on this subset of peaks to associate a TF 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- or ATAC-seq maps, identifying numerous TF 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. 200 groups were used for the analysis. Group scores and motifs were identified using the scores of the flanks 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 an FDR-corrected $P < 1\times10^{-2}$ and absolute DNase fold change > 1. A number of DNase peaks were observed to be shared between (i) B cells and monocytes and (ii) T cells and NK cells (see 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: (1) promoter: peaks in the promoter regions genes, defined as 2kb window up- and downstream of the TSS; (2) ubiquitous: peaks with no significant change in all comparisons of B cell vs. HSPCs, T cells and NK cells; (3) transition: peaks with significant change in all of B cell vs. T cells and NK cells; and (4) cell-type specific: peaks with significant change in all of B cell vs. HSPCs, T cells and NK cells. A similar logic is 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 non-zero *k*-mer groups identified in each cell type is an outcome of the SeqGL optimization algorithm. In particular, HSPCs SeqGL scores for 53 non-zero *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 are used to learn the weights of TF signals occurring in different peak categories using a regression model.

Each category of peaks (promoter, ubiquitous, cell-type specific, HSPC-specific, and transition) is represented by all SeqGL TF features from the relevant cell types (i.e. for B cell specification, B cell features are used for all peak categories except HSPC-specific, and HSPC features are used for all peak categories except B cell specific). For each TF 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 are summed up, and this sum is used as the corresponding feature value for the gene; “active” peaks are those marked with either H3K4me1 or H3K27ac. Therefore, the extent of DNase accessibility is not considered in the model; all reproducible DNase peaks with active histone marks in a particular category are treated equally. The multiplicity of peaks containing a TF signal does figure into the model, e.g. if a gene gains many B cell specific peaks all containing an IRF signal, than the

corresponding feature value (IRF binding signal in B cell specific peaks) will be high because it will be the sum of multiple IRF SeqGL scores.

The TF binding site identification is 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 TF are summed across each category of peaks. Therefore, e.g., the information in the set of ubiquitous peaks assigned to a gene is 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 is not retained. Similarly, while we use second order features in the model – e.g. "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 TFs that explain gene expression changes and the categories of peaks through which they act. Ridge regression models were trained using the SPAMS package[52], and a regression model was trained separately for the transition from HSPCs to each differentiated cell type.

DNA signals in poised B cells 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[53]. 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. 10-fold cross-validation was then performed with the new enhancer assignments. This process was repeated until the improvement in cross validation was less than 1e-2.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

1. Samstein RM, et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell. 2012; 151:153–66. [PubMed: 23021222]
2. Heinz S, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell. 2010; 38:576–589. [PubMed: 20513432]
3. Smale ST. Pioneer factors in embryonic stem cells and differentiation. Curr Opin Genet Dev. 2010; 20:519–26. [PubMed: 20638836]
4. Rothenberg EV. The chromatin landscape and transcription factors in T cell programming. Trends Immunol. 2014; 35:195–204. [PubMed: 24703587]
5. Lara-Astiaso D, et al. Immunogenetics. Chromatin state dynamics during blood formation. Science. 2014; 345:943–9. [PubMed: 25103404]
6. Xu CR, et al. Chromatin “prepattern” and histone modifiers in a fate choice for liver and pancreas. Science. 2011; 332:963–6. [PubMed: 21596989]
7. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–26. [PubMed: 16630819]
8. Rada-Iglesias A, et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011; 470:279–83. [PubMed: 21160473]
9. Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010; 107:21931–6. [PubMed: 21106759]
10. Hnisz D, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–47. [PubMed: 24119843]
11. Whyte WA, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153:307–19. [PubMed: 23582322]
12. Parker SC, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A. 2013; 110:17921–6. [PubMed: 24127591]
13. Wang HF, et al. NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. Proc Natl Acad Sci U S A. 2014; 111:705–710. [PubMed: 24374627]
14. Benayoun BA, et al. H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. Cell. 2014; 158:673–88. [PubMed: 25083876]
15. Stergachis AB, et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell. 2013; 154:888–903. [PubMed: 23953118]
16. Zhu J, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell. 2013; 152:642–54. [PubMed: 23333102]
17. Roadmap Epigenomics C et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–30. [PubMed: 25693563]
18. Li QH, Brown JB, Huang HY, Bickel PJ. Measuring Reproducibility of High-Throughput Experiments. Annals of Applied Statistics. 2011; 5:1752–1779.
19. Zheng Y, et al. Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate. Nature. 2010; 463:808–12. [PubMed: 20072126]
20. Yoshida T, et al. Transcriptional regulation of the Ikzf1 locus. Blood. 2013; 122:3149–59. [PubMed: 24002445]
21. Mifsud B, et al. Mapping long-range promoter contacts in human cells with high-resolution capture HiC. Nat Genet. 2015

22. Kieffer-Kwon KR, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell. 2013; 155:1507–20. [PubMed: 24360274]
23. Anderson E, Hill RE. Long range regulation of the sonic hedgehog gene. Curr Opin Genet Dev. 2014; 27:54–9. [PubMed: 24859115]
24. Schoenfelder S, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015; 25:582–97. [PubMed: 25752748]
25. Xi H, et al. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet. 2007; 3:e136. [PubMed: 17708682]
26. Setty M, Leslie CS. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. PLoS Comput Biol. 2015; 11:e1004271. [PubMed: 26016777]
27. Wickrema, A.; Kee, B., editors. Molecular Basis of Hematopoiesis. Springer; 2009.
28. Lazarevic V, Glimcher LH, Lord GM. T-bet: a bridge between innate and adaptive immunity. Nat Rev Immunol. 2013; 13:777–89. [PubMed: 24113868]
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. 1995; 15:6075–87. [PubMed: 7565760]
30. Pan Z, Hetherington CJ, Zhang DE. CCAAT/enhancer-binding protein activates the CD14 promoter and mediates transforming growth factor beta signaling in monocyte development. J Biol Chem. 1999; 274:23242–8. [PubMed: 10438498]
31. Vahedi G, et al. STATs shape the active enhancer landscape of T cell populations. Cell. 2012; 151:981–93. [PubMed: 23178119]
32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]
33. Mak KS, Funnell AP, Pearson RC, Crossley M. PU.1 and Haematopoietic Cell Fate: Dosage Matters. Int J Cell Biol. 2011; 2011:808524. [PubMed: 21845190]
34. Wickrema, A.; Kee, BL. Molecular basis of hematopoiesis. Vol. viii. Springer; New York, NY: 2009. p. 2576 p. of plates
35. Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2014; 47:8–12. [PubMed: 25547603]
36. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489:109–13. [PubMed: 22955621]
37. Dowen JM, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell. 2014; 159:374–87. [PubMed: 25303531]
38. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–9. [PubMed: 21441907]
39. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]
40. Shen Y, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012; 488:116–20. [PubMed: 22763441]
41. Malin J, Aniba MR, Hannenhalli S. Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers. Nucleic Acids Res. 2013; 41:6828–38. [PubMed: 23700312]
42. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–12. [PubMed: 19295514]
43. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10:1213–8. [PubMed: 24097267]
44. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]
45. Lawrence M, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9:e1003118. [PubMed: 23950696]
46. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

47. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics. 2010; 11:415. [PubMed: 20691053]

48. McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]

49. Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–31. [PubMed: 22955991]

50. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010; 11:367. [PubMed: 20598126]

51. Benaglia T, Chauveau D, Hunter DR. mixtools: An R Package for Analyzing Mixture Models. Journal of Statistical Software. 2009; 32

52. Mairal JFB, Ponce J, Sapiro G. Online Learning for Matrix Factorization and Sparse Coding. Journal of Machine Learning Research. 2012; 11:19–60.

53. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13. [PubMed: 19033363]

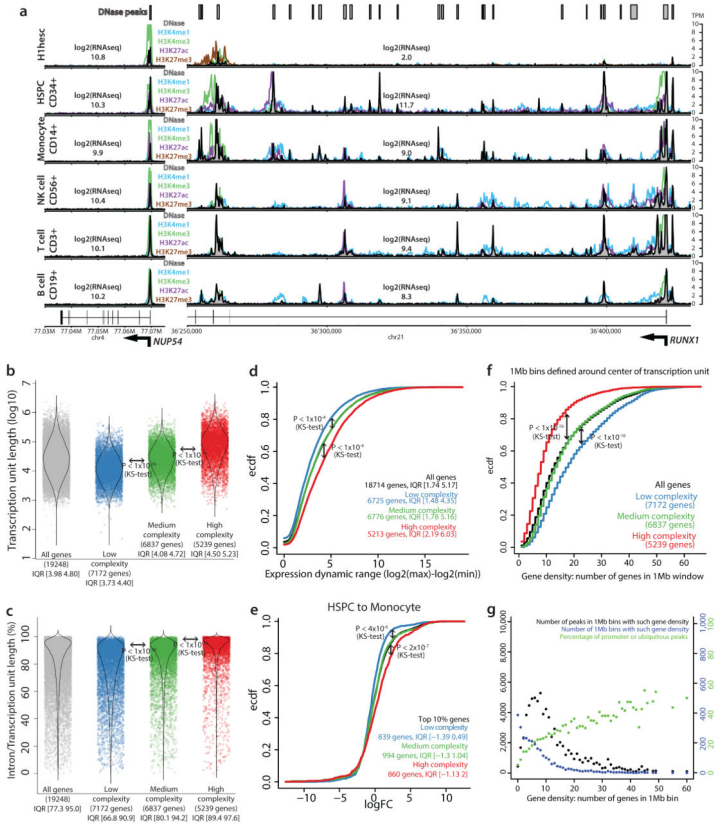

Figure 1. DHS atlas defines high and low complexity genes in hematopoietic differentiation

(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 grey signal to DNaseI hypersensitivity. Atlas peaks (DHSs) at these two loci are represented in the top track as grey boxes. *NUP54* has only one promoter peak while *RUNX1* was assigned 43 peaks. Transcript expression is quantified from RNA-seq for each gene across cell types.

(b–c) Complexity classes were defined by taking the 33 and 66 percentiles of the complexity distribution to produce groups with similar number of genes: low complexity (0 to 2 peaks), medium (3 to 7), and high (8 or more). High complexity genes have (b) longer transcription units with (c) higher fractions of intronic sequence. *P*-values are computed using one-sided Kolmogorov-Smirnov tests. IQR: interquartile range.

(d–e) Gene expression variation correlates with regulatory complexity. (d) Distribution of expression dynamic range for the three complexity groups. (e) Expression log fold changes (logFC) in the transition from CD34+ HSPCs to CD14+ monocytes for the complexity groups, based on the top 10% most highly expressed genes in each cell type.

(f–g) Gene density analysis for genes in complexity classes, based on gene counts in 1MB genomic bins. (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).

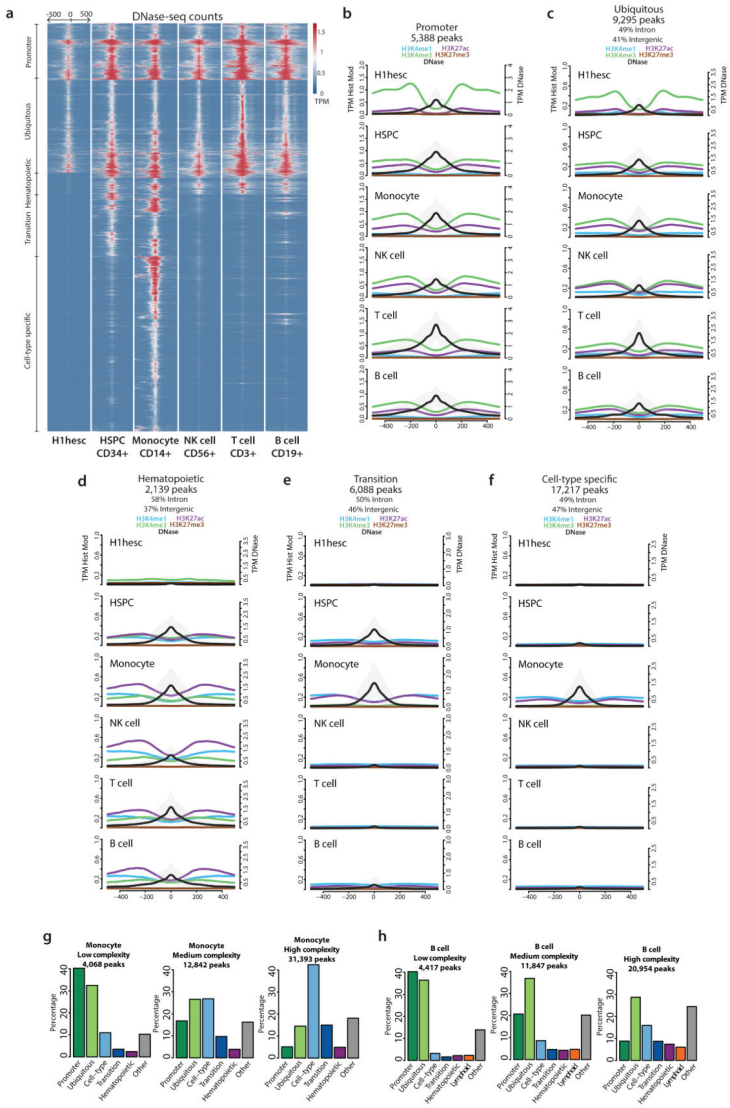

Figure 2. High complexity genes contain enhancers with distinct dynamics

(a) Heatmap showing DNase accessibility at peaks for hESC cells and 5 primary hematopoietic cell types. Each row represents one of the 48,303 DNase peaks present in CD14+ monocytes. DNase read counts are displayed in a 1,000 base window, binned at 10 bases, and normalized (TPM) in each cell type. 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 metapeaks, plotting the average of each signal across all peaks in the category. For DNase, the 10 and 90 TPM percentiles at each position are represented as a grey shadow 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, while DHSs assigned to high complexity genes are strongly enriched

for cell-type specific and transition peaks. The lymphoid pattern is defined as DHSs accessible in B, T and NK cells. “Cell-type” is short for cell-type specific.

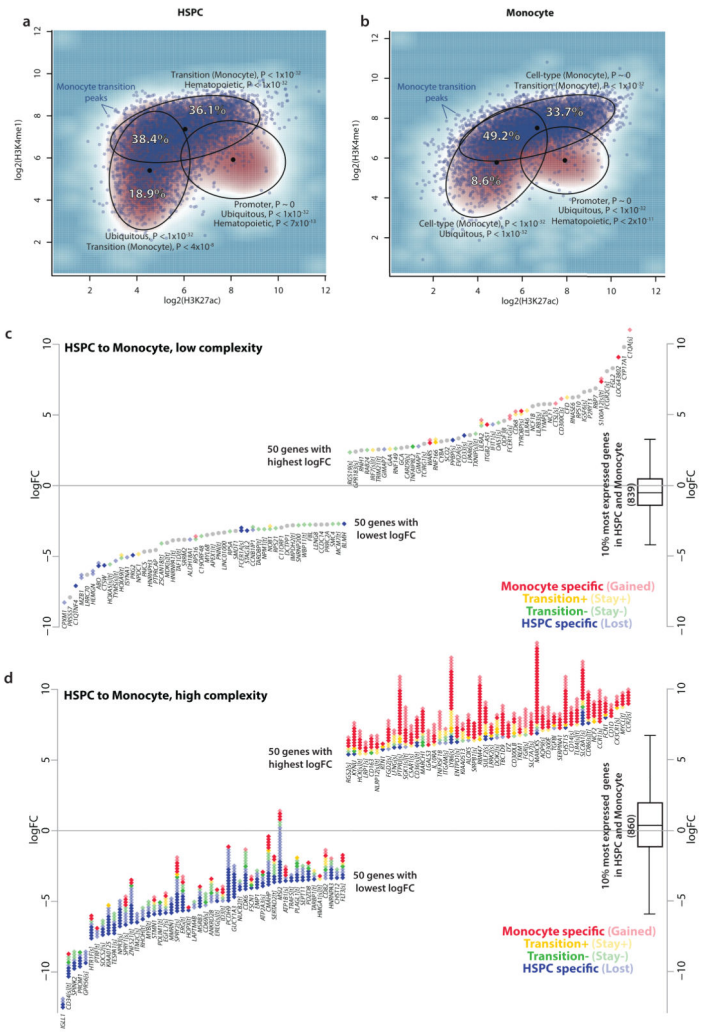

Figure 3. Gain of active enhancers in cell state transitions correlates with increase in expression

(a–b) The 2D topographical plots show the density of all accessible DNase peaks in (a) CD34+ HSPCs and (b) CD14+ monocytes in the landscape defined by the active mark H3K27ac (x-axis) and the mark H3K4me1 (y-axis). Gaussian mixture modeling identifies three subpopulations in each plot, shown by the mean of each mixture component and the ellipse corresponding to the 90% probability level curve. Enrichments of accessibility patterns in each subpopulation are indicated in figure; all enrichment *P*-values by Fisher's test. Percentages indicate proportion of all the monocyte transition peaks that lie in each region of the ellipses.

(c–d) Panels show gene expression changes and enhancer changes in (c) low complexity and (d) high complexity genes for the transition from HSPCs to monocytes for the 100 genes with largest absolute logFC in each class. Each gene is illustrated by a stack of diamonds, where a diamond represents a DHS associated to the gene. The bottom of the stack corresponds to the logFC of 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). Grey dots represent peaks present in both cell types without changes in histone marks. Box-and-whisker plots show the logFC distributions for all the genes. Gene names annotated with [t]/[s] have GO annotations for transcriptional/signaling processes.

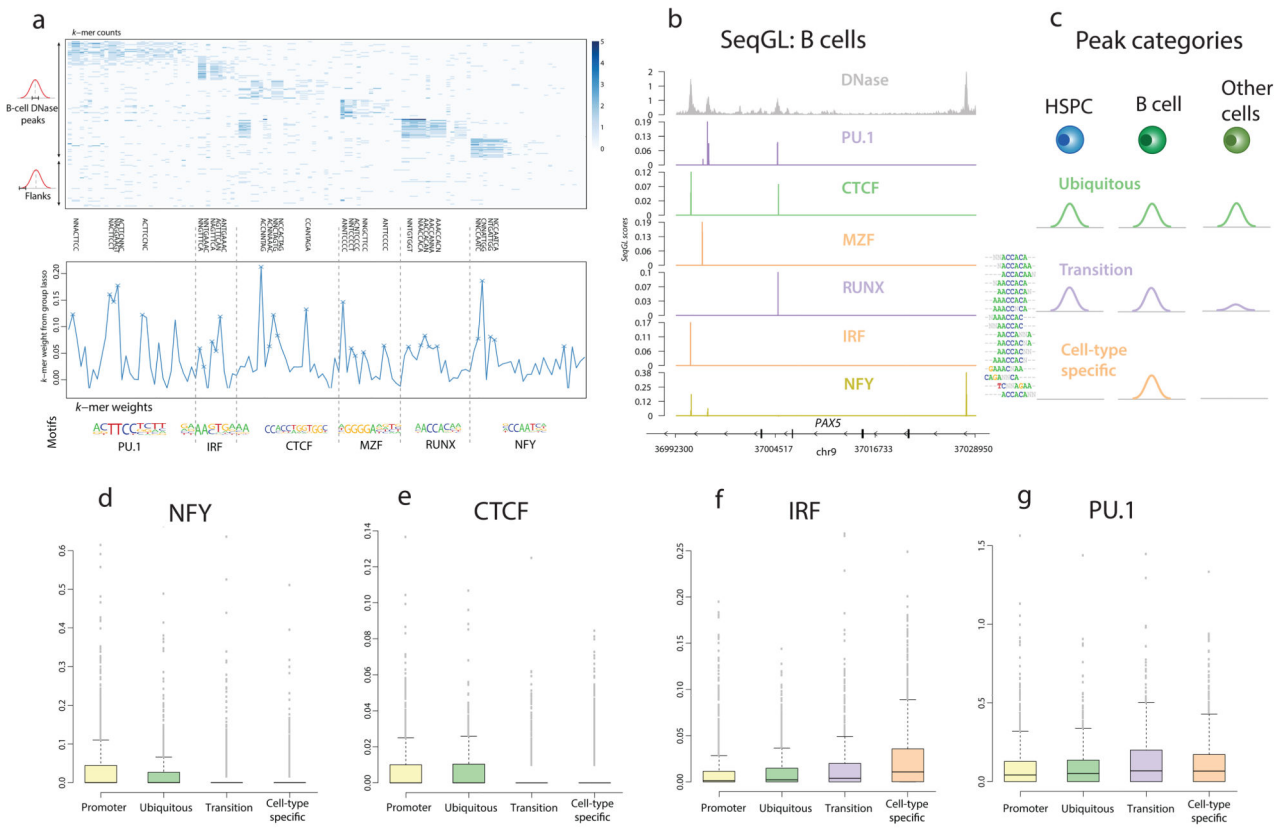

Figure 4. SeqGL identifies multiple TF sequence signals in B cell DNase peaks

(a) SeqGL identifies multiple DNA sequence signals *de novo* from DNase-seq data using *k*-mer patterns that can discriminate between peaks and flanks. 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 a weighting over *k*-mers in each group that discriminates peaks from flanks. The bottom panel shows the inferred weights and top ranking *k*-mers of each group. Groups with positive weights represent predicted TF binding signals.

(b) The *PAX5* locus is shown highlighting the top TFs identified by training SeqGL on B cell DNase peaks. The first row shows the DNase read density and the remaining rows show SeqGL predicted TF binding signals.

(c) The active B cell peaks were divided into four categories: (1) promoter: peaks defined in the promoter; (2) ubiquitous: peaks with similar accessibility across cell types; (3) transition: peaks retained specifically in the transition from HSPC to B cells; (4) cell-type specific: peaks that are significantly higher in B cells.

(d–g) Barplots showing binding signal enrichment in different peak categories. NFY is strongly enriched in promoters, and CTCF in both ubiquitous and promoter peaks. The B cell factor IRF is significantly higher in cell-type specific peaks, while PU.1 is enriched in both cell-type specific and transition peaks. “*” indicates $P < 0.01$ by Wilcoxon rank sum test. Error bars represent standard deviation.

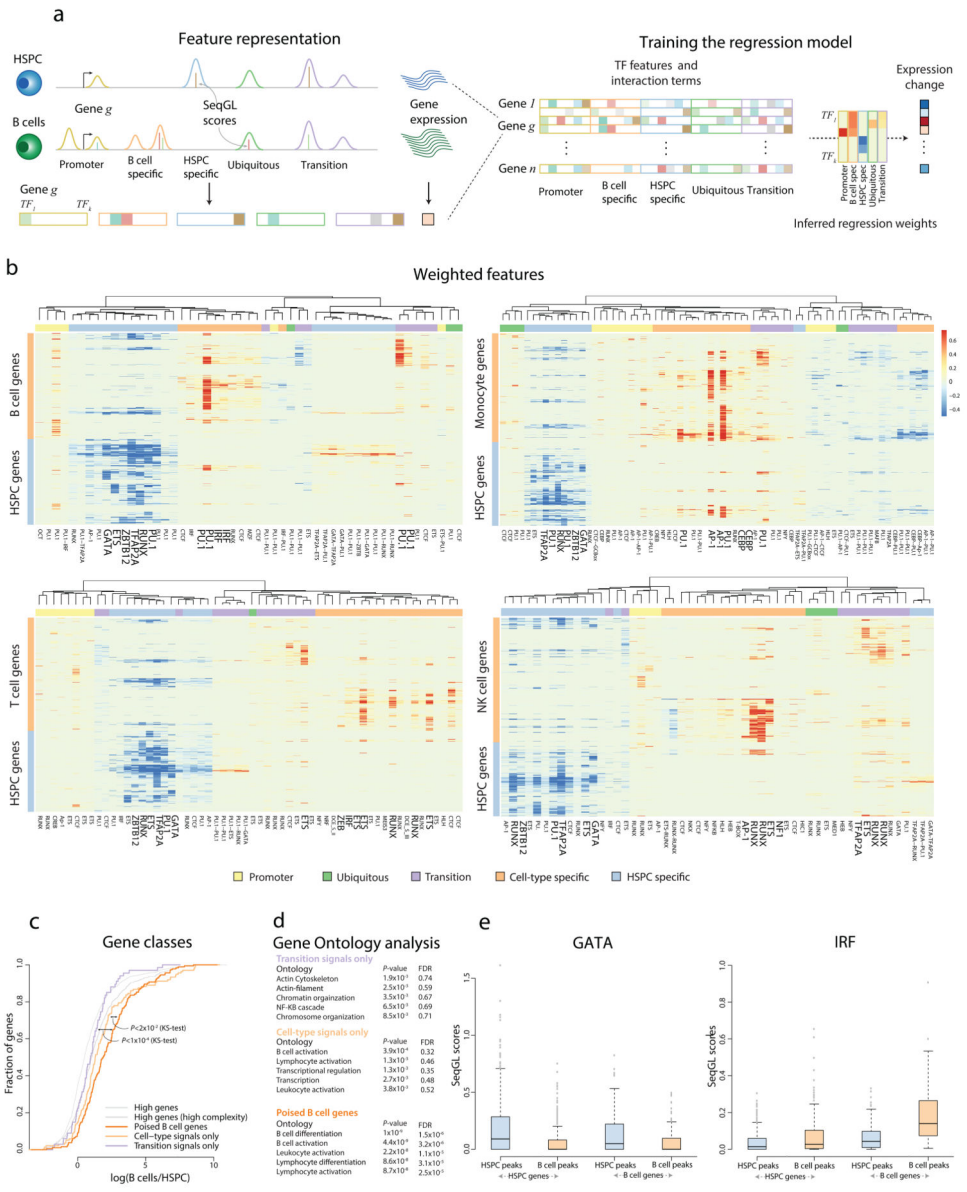

Figure 5. Regression model suggests a role for PU.1 in early establishment of B cell enhancers

(a) Regression framework to model the effect of TF signals on gene expression changes in the transition from HSPCs to B cells. 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 TFs across peak categories, representing their contribution to expression changes, are inferred using ridge regression.

(b) Heatmaps showing the predicted contributions of each TF across peak categories to gene expression changes in cell state transitions. PU.1 signals from multiple peak classes occur in all models. High expression in HSPC, monocytes, B, T 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 PU.1.

(c) B cells genes in Fig. 5b 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 compared to other categories ($P < 0.02$, category (i) and $P < 1 \times 10^{-4}$, category (ii), one-sided KS test).
(d) Functional enrichment through gene ontology analysis shows that “poised” genes are significantly more enriched ontologies related to B cell functions compared to other genes.
(e) HSPC specific signals like GATA are significantly higher in active HSPC peaks of HSPC genes compared to those of 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 compared to HSPC genes ($P < 3 \times 10^{-4}$) and are low in HSPC peaks. Error bars represent standard deviation.

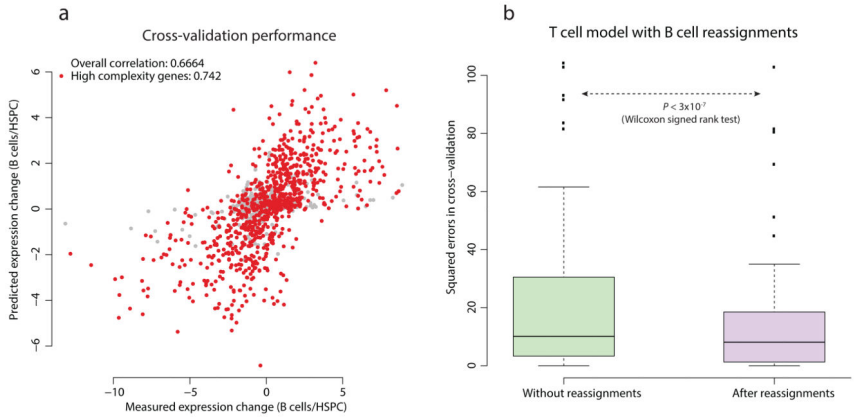

Figure 6. Regression model proposes reassignment of enhancers to genes

(a) The regression performs very well on high complexity genes (Spearman rank correlation ρ=0.74, 10-fold cross-validation.). A number of genes that are highly expressed in either B cells or HSPC had high regression errors (Supplementary Fig. 37), often because their large expression change was not associated with corresponding high complexity in the enhancer-gene assignment. 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 HSPC, 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 of prediction errors for affected genes in both cell types (Supplementary Fig. 38).

(b) For an independent evaluation of the enhancer reassignment procedure, the enhancer reassignments identified in B cells were used to predict gene expression changes in the T cell transition. Squared errors using these reassignments were significantly reduced compared to errors without any reassignments ($P < 3 \times 10^{-7}$, Wilcoxon signed rank test). Similarly, using enhancer-gene reassignments from any of the other regression models led to improved regression performance in independent cell types (Supplementary Fig. 39).