Interactome Maps of Mouse Gene Regulatory Domains Reveal Basic Principles of Transcriptional Regulation

Kyong-Rim Kieffer-Kwon,1,12 Zhonghui Tang,2,12 Ewy Mathe,1,12 Jason Qian,1,12 Myong-Hee Sung,3,12 Guoliang Li,2 Wolfgang Resch,1 Songjoon Baek,3 Nathanael Pruet,1 Lars Grentved,3 Laura Vian,1 Steevenson Nelson,1 Hossein Zare,4 Ofir Hakim,5 Deepak Reyon,6,7 Arito Yamane,1 Hirotaka Nakahashi,1 Alexander L. Kovalchuk,6 Jizhong Zou,3 J. Keith Joung,5,7 Vittorio Sartorelli,4 Chia-Lin Wei,10 Xiaoan Ruan,2 Gordon L. Hager,3,12 Yijun Ruan,2,12 and Rafael Casellas11,15

1Genomics and Immunity, NIAMS, National Institutes of Health, Bethesda, MD 20892, USA
2The Jackson Laboratory for Genomic Medicine, and Department of Genetic and Development Biology, University of Connecticut, 400 Farmington, CT 06030, USA
3Laboratory of Receptor Biology and Gene Expression, NCI, National Institutes of Health, Bethesda, MD 20892, USA
4Laboratory of Muscle Stem Cells and Gene Regulation, NIAMS, National Institutes of Health, Bethesda, MD 20892, USA
5Bar-Ilan University, Ramat-Gan 5290002, Israel
6Molecular Pathology Unit, Center for Computational and Integrative Biology, and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129, USA
7Department of Pathology, Harvard Medical School, Boston, MA 02115 USA
8Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD 20852, USA
9Laboratory of Stem Cell Biology, NIAMS, National Institutes of Health, Bethesda, MD 20892, USA
10DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
11Center of Cancer Research, NCI, National Institutes of Health, Bethesda, MD 20892, USA
12These authors contributed equally to this work
*Correspondence: rafael.casellas@nih.gov
http://dx.doi.org/10.1016/j.cell.2013.11.039

SUMMARY

A key finding of the ENCODE project is that the enhancer landscape of mammalian cells undergoes marked alterations during ontogeny. However, the nature and extent of these changes are unclear. As part of the NIH Mouse Regulome Project, we have combined DNaseI hypersensitivity, ChIP-seq, and ChIA-PET technologies to map the promoter-enhancer interactomes of pluripotent ES cells and differentiating B lymphocytes. We confirm that enhancer usage varies widely across tissues. Unexpectedly, we find that this feature extends to broadly transcribed genes, including Myc and Pim1 cell-cycle regulators, which associate with an entirely different set of enhancers in ES and B cells. By means of high-resolution CpG methylomes, genome editing, and digital footprinting, we show that these enhancers recruit lineage-determining factors. Furthermore, we demonstrate that the turning on and off of enhancers during development correlates with promoter activity. We propose that organisms rely on a dynamic enhancer landscape to control basic cellular functions in a tissue-specific manner.

INTRODUCTION

Gene expression during development is orchestrated by promoter sequences and a variety of distal cis-regulatory elements. Key among these are enhancers, which associate with promoters to increase the transcriptional output of target genes in a tissue-specific manner (Visel et al., 2009). Enhancers are typically distinguished from nonregulatory DNA by their hypersensitivity to DNase digestion (Sabo et al., 2006) and binding of chromatin modifiers. The CBP/p300 acetyltransferase for instance mediates H3K27 acetylation of chromatin at active enhancers (Creyghton et al., 2010). In addition, enhancers display high levels of H3K4 monomethylation (H3K4me1; Buecker and Wysocka, 2012), and a relative depletion of H3K4me3 (Heintzman et al., 2007) and the histone variant H2AZ (Kouzine et al., 2013). Based on these parameters, ~400,000 genomic sites displaying enhancer-like features were recently discovered, spanning nearly 10% of the human genome (ENCODE Project Consortium et al., 2012). Enhancers control lineage identity by recruiting transcription factors, cofactors, and RNA Polymerase II (PolII) to target genes. They physically interact with promoters resulting in looping out of intervening sequences (Krivenga and Dean, 2012), which in some instances can span over 1 Mb of DNA (Nobrega et al., 2003). In contrast to promoters and insulators, which vary little across cell types, the enhancer landscape changes considerably during development (Thurman et al., 2012). This feature predicts that...
functional connectivity in mammalian cells (1) must display a high degree of tissue specificity and (2) should closely reflect transcriptome changes during cell differentiation. However, these ideas have not been fully explored because of the difficulty of mapping promoter-enhancer connections during development.

In the absence of direct approaches, enhancers have been typically assigned to “cognate” promoters based on linear proximity or shared chromatin states. This strategy has limitations because enhancers do not always regulate nor share chromatin profiles with the nearest promoter. Alternatively, chromosome conformation capture techniques have been used to explore regulatory interactions at predefined genomic loci. However, the resolution of 3C-based techniques alone is insufficient to map promoter-enhancer connectivity in entire genomes (Xie and Ren, 2013). To overcome this challenge, the ChIA-PET protocol was recently developed (Fullwood et al., 2009). ChIA-PET is a ChIP-based method that captures long-range chromatin interactions involving or mediated by a protein of interest such as estrogen receptor $\alpha$ in adenocarcinoma cells (Fullwood et al., 2009) or RNA PolII in human cell lines (Li et al., 2012).

We here introduce the NIH Mouse Regulome Project, an initiative that seeks to define the 3D interplay of gene regulatory domains in developing mouse primary cells. In this first report we compare pluripotent embryonic stem (ES) cells and differentiated B lymphocytes. By combining ChIA-PET, CpG methylomes, DNaseI hypersensitivity, transcriptomes, digital footprinting, and TALEN-mediated genome editing, our studies reveal the dynamics of the mouse regulome during ontogeny.

RESULTS

A Comprehensive Map of Regulatory Domains and Their Interactions in Mouse Primary Cells

To characterize the mouse regulome in primary cells we first applied DNaseI hypersensitivity (DHS) followed by deep-sequencing to CD43+ B lymphocytes activated in the presence of lipopolysaccharide and interleukin-4 (LPS+IL-4). From two independent experiments (268 million aligned reads), we identified 90,015 high-confidence DNaseI domains in B cells. As expected, DNaseI-seq profiles were highly reproducible between biological replicates (Spearman’s $r = 0.99$; Figure S1A available online).

To identify DHS sites associated with gene regulatory domains, we next mapped Nipbl, Med12, and p300 by ChIP-seq (Figure 1A). We chose the cohesin-loading factor Nipbl and the Med12 component of mediator because they demarcate enhancers tethered with core promoters (Kagey et al., 2010). Likewise, the transcription regulator p300 occupies a subset of active promoters and enhancers (Chen et al., 2008). There was a considerable, although incomplete, overlap in the recruitment of these factors in B cells (Figures 1A and S1B). Active promoters were thus identified as Nipbl+$^+$, Med12+$^+$, or p300+$^+$ DHS sites that overlapped with ENSEMBL-annotated transcription start sites (TSSs). Based on this strategy we identified 17,004 DHS promoter elements associated with 16,931 genes, 49,763 enhancer elements, and an additional 23,248 that did not overlap with either (Figure 1B). Promoters were in general H3K4me1$h^+$, H3K4me3$h^+$, H2AZ$h^+$, whereas the opposite signature demarcated enhancer domains (Figure 1A; Kouzine et al., 2013). DHS analysis of mouse ES cells uncovered a similar number of DHS promoters (16,771) and an increased number of enhancers (62,766, Figure 1B).

To directly map the promoter-enhancer interactome, we applied chromatin interaction analysis by paired-end-tag sequencing (ChIA-PET; Fullwood et al., 2009; Zhang et al., 2012), which combines PolII ChIP with 3C technology (Figures 1C, S2A). We generated two independent B cell ChIA-PET libraries, from which ~15 million reads were uniquely aligned and classified into two separate data sets: 5.7 million reads of PolII chromatin occupancy, and 9.2 million reads clustered into 14,247 high-confidence PolII long-range cis interactions or PETs (Figure 1D and Table S1). Both data sets were correlated between replicates (Spearman’s $p > 0.83$, Figures S2B and S2C).

Attesting to the specificity of ChIA-PET, most PolII long-range interactions (13,070, 92%) were linked to at least one gene regulatory domain (Figure S1C). Furthermore, of 16,931 B cell promoters associated with DHS domains, 6,890 were involved in PolII long-range interactions. In general, these genes were transcribed 2-fold higher ($p < 2 \times 10^{-16}$, Figure S1D) and recruited more PolII ($p < 2 \times 10^{-16}$, Figure S1E) than nonanchored ones. We also detected 6,813 DHS enhancer domains involved in PolII interactions. Of these, 71% were active (H3K27Ac$^+$)), whereas up to 60% of nonanchored ones were poised (H3K27Ac$^-$) (Figure S1F). In general, the number of ChIA-PET interactions per regulatory site was proportional to the extent of DNaseI digestion (Figure S1G). Thus, ChIA-PET preferentially detects PolII long-range interactions involving H3K27Ac$^+$ enhancers and transcriptionally active promoters.

As previously shown (Li et al., 2012), PolII interactions fell into four distinct groups: (1) intragenic, connecting promoters to gene bodies; (2) extragenic, connecting promoters to distal regulatory elements; (3) intergenic, tethering promoters from different genes; and (4) enhancer-enhancer interactions (Figure S1H). Examples of these are provided in Figure 1D for the Atsp5-Mir155 gene locus. Consistent with high expression of Mir155 in activated B cells (Kuchen et al., 2010), its promoter was associated with 83 long-range interaction tags (Figure 1D, upper). Of these, 70 were extragenic, involving 5 upstream enhancer domains, while 13 were intragenic, connecting the promoter to downstream sequences. An additional 23 PolII long-range interactions interconnected the 5 enhancers upstream of Mir155. In contrast to B cells, ES cells actively transcribe Mrpl39, Jam2, and Atsp5 but express little Mir155 mRNA (Figure 1D, lower). Consistent with this, we identified 30 intergenic connections between Mrpl39, Jam2, and Atsp5 promoters in ES cells, whereas few connections involved Mir155 (Figure 1D). As in previous ChIA-PET studies, both direct and indirect interactions were considered in our analysis (Figure S1I).

TALEN-Mediated Validation of Promoter-Enhancer Connectivity

ChIA-PET confirmed established connections between gene regulatory domains. For instance, the pluripotent gene Sox2 was associated in ES cells with a series of enhancers recently described by 5C studies (Figure S3A; Phillips-Cremins et al., 2013). Likewise, the immunoglobulin heavy chain ($\lambda$) 3′ Ez
enhancer was found in spatial proximity to transcribed Igµ, Igγ1, and Igε in B cells only (Figure S3B). We also found evidence of Igµ-Igγ1 associations, representing either synapses between the recombining genes (Wuerffel et al., 2007) or fully recombined DNA. At the Igκ locus, the Vκ–5′Ex domain was connected to the previously characterized 3′Ex and Ed enhancers (Figure 2A; Liu

Figure 1. Characterization of Regulatory Domains and Their Interactions in Mouse Primary B and ES Cells
(A) Pax5 locus in activated B cells displaying DNaseI hypersensitivity (DHS); recruitment of Nipbl, Med12, and p300; and chromatin marks H2AZ, H3K4me1, and H3K4me3.
(B) Bar graphs showing the number of DHS islands in B and ES cells overlapping with promoters (TSS+, white), enhancers (TSS-, Nipbl+, or Med12+, or p300+, red), or nonoverlapping (blue).
(C) The ChIA-PET protocol combines PolII ChIP with conformation capturing techniques to map the interaction of active promoters with gene regulatory domains.
(D) Examples of ChIA-PET clusters at the Mir155 locus in activated B cells (red connectors) or ES cells (blue connectors). Each connector links two or more long-range interactions (PETs) separated by <500 bps (Figure S2A). ChIP-seq data are represented as reads per kb per million sequences (RPKM). Promoters (P) and enhancers (E) are boxed and the number of total PETs is provided in parenthesis. Interactions between enhancers and Mir155 are represented by semi-circle connectors. mRNA expression is provided for B and ES cells as RPKM values (+, strand transcription in green; –, strand in blue).
See also Figures S1, S2, and S3, and Table S1.
Figure 2. In Vivo Validation of ChIA-PET by Genome Editing

(A) ChIA-PET at the Igκ locus identifies previously characterized 5′Ex, 3′Ex, and Ed enhancers, as well as new enhancers E4 and E5. Number of PETs associated with each regulatory domain (boxed) are provided in parenthesis. The DHS activated B cell track is also provided (black).

(legend continued on next page)
et al., 2002; Meyer and Neuberger, 1989). Unexpectedly, the analysis uncovered two additional enhancers located 8 kb (E4) and 15 kb (E5) downstream of Ed (Figure 2A).

We also found additional enhancers (E1-E2) associated with the activation induced deaminase (AID) gene Aicda (Figure 2B). The three enhancers previously shown to regulate AID transcription in vivo were also linked by Polll long-range interactions in the analysis (E3-E5, Figure 2B; Crouch et al., 2007; Huong et al., 2013; Sayegh et al., 2003). The Apobec1 promoter and a sixth enhancer located in Apobec1 intron 2 were also clustered (Figure 2B). To validate ChiA-PET associations, we deleted E1 and E2 in CH12 mouse lymphoma cells. We chose this B cell line because upon activation it transcribes high levels of AID and undergoes efficient Igµ-Igα recombination (Nakamura et al., 1996). To facilitate homozygous gene targeting, knockout constructs were cotransfected with enhancer-specific transcription activator-like effector nucleases (TALENs), assembled via a solid-phase high-throughput system (Reyon et al., 2012; Figure S4A). Upon activation, wild-type CH12 cells increased AID mRNA expression ~5-fold and recombined to Igα (15%, Figures 2C and 2D). Deletion of E1 or E2 however markedly reduced AID transcription and Igα expression (Figures 2C and 2D), consistent with the notion that the extent of switching is proportional to AID expression (Takizawa et al., 2008). Transcription of Apobec1 was also impaired in the mutant cells, whereas noninteracting Ezh2 and Cd83 genes were unaffected (Figure 2C). Importantly, E1−− and E2−− cells displayed an overall reduction in Nipbl and PolII occupancy at all regulatory domains within the Aicda locus, including the Apobec1 promoter (Figure 2E and S4B). In contrast, this effect was not observed at the Foxo2-Necap1 locus, ~190 kb downstream of Apobec1 (Figure 2E and S4B). Thus, E1 and E2 regulate AID and Apobec1 transcription by controlling local recruitment of PolII.

To further validate the ChiA-PET results, we targeted additional regulatory elements associated with the Pou2af1 (OCA-B), and Cd79a genes. We uncovered an intronic enhancer (E3) ~15 kb downstream of OCA-B TSS required for transcriptional upregulation upon B cell activation but dispensable for basal transcription in nonstimulated cells (Figure S4C). This activity is consistent with the reported dynamics and signaling requirements of OCA-B expression during B cell differentiation (Casellas et al., 2002; Qin et al., 1998). A similar analysis confirmed the presence of enhancer elements that augment basal Cd79a transcription (Figure S4D). Additional gene targeting experiments within the Pim1 oncogene locus are discussed below (Figure 5). Taken together, these results demonstrate that at least a fraction of PolII long-range interactions, as defined by ChiA-PET, represent functional promoter-enhancer connections in B lymphocytes.

Single- and Higher-Order Gene Clusters in Primary Mouse Cells

Up to 54% of genes recruiting PolII in activated B cells were associated with long-range interactions (6,890 of 12,652). Of these, 1,231 (18%) represented single promoters tethered to at least one enhancer (Figure 3A). These clusters created complex architectures and spanned an average of 78 kb of genomic DNA (Figure S5A). The most elaborate of this group was the Gpr183 promoter, which was connected either directly or indirectly to 12 enhancers via 76 long-range interactions (Figure 3A). Another example was Cd83, which displayed 158 PolII intragenic and extragenic connections (Figure 3B).

Among single-promoter gene clusters, we found examples of the recently dubbed superenhancer domains (Whyte et al., 2013), which were defined based on clustering of gene regulatory domains: e.g., mir290/295 and Sox2 loci (Figures S5B and S3A). However, the vast majority of anchored genes (5,606, 81%) formed higher-order multigene complexes (1,481 B cell clusters, Figure 3C), which could not be easily deduced as interacting based on visual inspection of DHS island distribution. The average span of these clusters was 179 kb (Figure S5A). Its prime example in B cells was the Rela cluster in chromosome 19, which was composed of 66 genes and 398 long-range interactions (Figure 3C). Promoters linked by intergenic connections displayed higher PolII density and mRNA synthesis relative to genes from single-promoter clusters or not anchored to other domains (Figure S5C). Furthermore, families of genes coexpressed during ontogeny were overrepresented in the multiple-promoter gene group (see Experimental Procedures). Among these we found the major histocompatibility complex H2-M3 cluster (Fisher exact test p = 1.3 × 10−14), the Hist1h histone family (p = 7.7 × 10−23), and the lymphoid signaling Gimap cluster (p = 7.6 × 10−14, Figure S5D).

In contrast to promoters, which readily formed higher-order complexes, the vast majority of enhancers (~90%) were linked to a single promoter, and less than 2% of all enhancers were linked to more than two promoters (Table S1). One exception was an enhancer downstream of Gimap6, which was directly linked to seven promoters (Figure S5D).

Transcriptional Correlation between IncRNA and Associated Coding Genes

Long noncoding RNAs (lncRNA) are a new class of RNAs believed to play regulatory functions (Batista and Chang, 2013). ChiA-PET identified hundreds of associations between protein-coding and lncRNA genes in multiple-promoter clusters. For instance, IncRNA E(ENSMUSG)85930 is extensively associated with Clec2d and to a lesser extent with Cd69 (Figure 3D). Other examples involving genes key for B cell development included Ptprcap-E90702, Cd81-E59277, and Bcl11a-E123592 pairs (Figure S5E). lncRNAs are believed to modulate transcription of...
neighboring genes by promoting local topological changes in chromatin (Ponting et al., 2009). Consequently, transcription of IncRNAs and their targeted genes is often coordinated (Guil and Esteller, 2012). To test this idea across the genome we measured expression of IncRNAs and their interacting protein-coding genes as defined by ChIA-PET. We found that genes associated with highly abundant IncRNAs were transcribed at higher levels than those associated with IncRNAs detected at low or trace levels ($p < 0.05$, Figure 3E). These findings are consistent with the proposal that transcription of IncRNAs and their targets can be coordinated. Whether the same scenario applies to promoter-promoter clusters not involving IncRNA remains to be determined.

**Broadly Expressed Genes Are Linked to Cell-Type-Specific Enhancers**

As expected, genes differentially expressed in B and ES cells were linked to tissue-specific regulatory elements. The pluripotency gene Sox2 for instance was associated with ES-cell-specific enhancers (Figure S3A). Conversely, the B-cell-specific Cd79b gene was only anchored to enhancers in the B cell compartment (Figure S6A).

We next turned our attention to genes transcribed in both cell types. Of 6,890 promoters anchored by ChIA-PET in B cells, 4,854 (70%) were also anchored in ES cells (Figure 4A, Venn diagram). As an example, the Hexim1-2 genes were linked to the same downstream enhancer (E1) in B and ES cells (Figure 4B). Surprisingly, most anchored promoters in the two cell types (4,430, 94%) were associated with at least one additional tissue-specific enhancer (Figure 4A, lower pie chart). A striking example was the Myc proto-oncogene, which displayed a completely different enhancer landscape in ES and B cells (Figure 4C). In B cells, Myc was linked to ten enhancers (E5-E14) located near or downstream of exon 3 of the IncRNA Pvt1, whereas in ES cells all enhancers associated with Myc (E1-E4) were found upstream of this site (Figure 4C). Other examples included Tgif1, Smad7, and Malat1, which were preferentially linked to ES-cell-specific enhancers, whereas Swap70, Etv5, and Pim1 were tethered to a greater number of enhancers in B lymphocytes (Figure 4D).

To explore whether changes in the enhancer landscape impacts transcription of this gene group, we measured their expression by calculating mRNA copy numbers per cell. Genes that turned on or off a single tissue-specific enhancer displayed little or no changes in transcription levels in the two cell types (Figure 4E). However, as genes interacted with two or more additional enhancers their expression was significantly different ($p < 9 \times 10^{-10}$, Figure 4E). Myc, for instance, was transcribed ~4 times...
higher in B cells than in ES cells (Table S2). This observation is consistent with the notion that, in general, transcription levels of a given promoter are commensurate with the number of regulatory domains it is regulated by (Li et al., 2012). On the basis of these findings we conclude that (1) broadly expressed genes can be regulated by cell-type-specific enhancers and (2) the turning on and off of enhancers during ontogeny impacts transcription levels.

Dynamic CpG Methylation of Cell-Type-Specific Regulatory Domains

Cellular differentiation is accompanied by changes in DNA methylation at promoters and distal regulatory domains (Shen et al., 2013; Song et al., 2013; Stadler et al., 2011; Ziller et al., 2013). To explore whether the dynamics of DNA methylation correlate with differential enhancer usage, we applied bisulphite sequencing (Bis-Seq) and generated methylome libraries at single-nucleotide resolution. Bis-Seq of activated B cells provided a total of 148 billion mappable methylome bases. We complemented this data set with Bis-Seq libraries from mouse ES cells (Stadler et al., 2011) and calculated the percentage of CpG methylation at gene regulatory domains. With few exceptions, activated B-cell-specific enhancers were highly methylated in ES cells (>80% of CpGs), whereas enhancers common to both cell types displayed a broad range of CpG methylation levels (Figure 5A). High methylation was also observed at ES cell-specific enhancers in activated lymphocytes, whereas B cell enhancers displayed low methylation levels (Figure 5A). Importantly, the level of CpG methylation was lower at active than at poised enhancers (p < 3 × 10^{-16}, Figure 5A). As expected, promoters of silent genes displayed on average higher CpG methylation than active ones (p < 2.2 × 10^{-16}, Figure 5A). Methylation levels were also inversely proportional to the extent of ChIA-PET signals (Figure S6B). Thus, enhancers are highly methylated when inactive, but become demethylated during development concomitant with the presence of tissue-specific PolII interactions.

We explore in Figure 5B transcriptional regulation of the Pim1 oncogene, whose promoter is tethered to an entirely different set of enhancers in activated B cells and ES cells. The analysis shows a direct correlation between CpG demethylation and enhancer usage. For instance, B cell enhancers E2 and E5 and ES cell enhancers E6 and E7 display nearly complete CpG demethylation in a cell-type-specific manner (Figure 5B). To confirm that these enhancers truly promote Pim1 transcription, we targeted E2 and E6 in CH12 B cells and ES cells, respectively. As measured by qPCR, we found a significant decrease in Pim1 mRNA levels in the targeted cells, whereas expression of Brd2, Mtch1, and Gapdh was unaffected (Figure 5C). Attempts to delete E2 and E6 in the cell type where they are inactive were unsuccessful (not shown), likely due to the inability of TALENs to target methylated DNA (Bultmann et al., 2012).

In ES cells, B-cell-specific enhancers, including those linked to broadly expressed genes, were hypermethylated (Figure S6C). To examine at which stage during B lymphopoiesis these regulatory elements become demethylated, we generated methylome libraries from bone marrow hematopoietic stem cells (KSL), B lymphoid precursors (CLP), and peripheral G0 resting B cells. Of 1,518 B-cell-specific enhancers that were linked by long-range interactions and hypomethylated during activation, only 82 (5%) were also hypomethylated in ES cells (Figure 5D). However, in KSL, precursors nearly half (714, 47%) of activated B cell enhancers were already demethylated (Figure 5D). This group included Pim1 enhancers E2, E3, and E5, which displayed nearly identical CpG methylation levels in KSL, CLP, G0 resting, and cycling B cells (Figure S6D). As KSLs develop into CLP B cell precursors, demethylation was observed in 61% (926) of B cell enhancers. In resting G0 B cells, this number increased to 80% (1,207, Figure 5D). At this stage of development, the overall mean methylation was comparable between activated B cell enhancers and those functional both in B and ES cells (Figure S6C). Thus, most activated B cell enhancers, including those associated with broadly expressed genes, are demethylated by the time naive lymphocytes migrate from the bone marrow to the periphery. This finding is consistent with the notion that the genome of G0 lymphocytes is primed for activation and that most genes expressed during the humoral immune response are transcribed at basal levels in the naive compartment (Kouzine et al., 2013; Nie et al., 2012). At the same time, it is important to point out that ~20% of activated B cell enhancers do not become fully demethylated until activation occurs. Among these, we find Pim1 E4 and Aicda E3 and E4 (Figures S6D and S6E).

Digital Genomic Footprinting Characterizes TF Binding in the Mouse Genome

The observation that cell-type-specific enhancers can promote transcription of broadly expressed genes implies that factors driving lineage specification are involved in this regulation. To explore this idea, we sought to comprehensively catalog transcription factor occupancy in mouse B lymphocytes and ES cells. To this end, we took advantage of the fact that transcription factors protect their binding sites from DNaseI cleavage, leaving nucleotide-resolution footprints within DHS islands (Neph et al., 2012). Figure 6A, for instance, shows four DNaseI footprints at the Pold4 gene promoter in G0 and cycling B cells. Importantly, these footprints overlap with recognizable binding motifs for transcription factors PU.1, Ebf1, Egr1, and Sp1 (Figure 6A). By applying an established algorithm (Baek et al., 2012), we detected 706,669 high-confidence (FDR < 5%) footprints within 75,917 B cell DHS domains (70% of total DHS). To link these footprints to known transcription factor recognition sequences, we examined all empirically defined DNA binding motifs, compiled by HOMER, UniPROBE, JASPAR, and similar databases. We found a significant enrichment in transcription factor binding motifs within DHS footprints (p < 1 × 10^{-6}, Figure S7A). Altogether, we linked 247 distinct transcription factor DNA motifs to 122,505 footprints in B cells (Table S1). In addition, de novo motif discovery yielded 18 new binding sites that did not match known recognition sequences (Table S1). A similar analysis linked 306 DNA motifs to 346,284 footprints in ES cells (Table S1).

Figure 6B shows examples of cleavage profiles for transcription factors Irf8, Sp1, Nrf1, PU.1, and CTCF (an extended view of footprints is provided in Figure S7B). Importantly, ChIP-seq analysis showed a correlation between PU.1 and CTCF occupancy, their DNA binding motifs, and cognate footprints.
Figure 4. Tissue-Specific Enhancers Contribute to the Transcriptional Regulation of Broadly Expressed Genes

(A) Venn diagram showing the number of ChIA-PET anchored promoters in B cells (left), ES cells (right), or in both cell types (middle). For the latter group, the pie chart below shows the number of promoters linked to the same (yellow) or to at least one cell-type-specific enhancer (gray).

(B) Aabd4 and Hexim1/2 gene promoters associate with the same downstream enhancer (E1) in B and ES cells.

(C) The Myc oncogene is linked to an entirely different set of enhancers in ES (blue) and B (red) cells. Enhancers are numbered from 1–10 based on proximity to the Myc promoter (P1).

(legend continued on next page)
Figure 6C). Similar results were obtained with Irf8 and Ebf1 (not shown). To confirm that footprinting profiles truly reflect the morphology of DNA-protein interactions, we turned to available transcription factor:DNA cocrystal structures. For instance, consistent with the published Ebf1:DNA structure (Treiber et al., 2010), backbone phosphates in direct contact with Ebf1 amino acids via hydrogen bounds (in red in Figure 6D) were the most protected from DNaseI cleavage, whereas more exposed residues (in cyan) displayed by comparison greater susceptibility to digestion. Similarly, for the cRel (NF-κB) factor DNaseI cleavage paralleled the topology of the protein-DNA interphase with a marked depression at DNA binding motif (Figure 6E; Huang et al., 2001). We conclude that DNaseI cleavage profiles can recapitulate the structural features of transcription factor:DNA interactions, and thus they reflect the occupancy of gene regulatory proteins across the genome.

Lineage-Determining Factors Regulate Transcription of Broadly Expressed Genes

Having validated the genomic footprinting approach, we next turned to the question of whether lineage-specific factors associate with enhancers controlling transcription of broadly expressed genes. To this end we classified B and ES cells enhancers into three groups: (1) those shared between the two cell types and linked to the same promoter as determined by ChIA-PET, (2) cell-type-specific enhancers bound to promoters active both in B and ES cells, and (3) cell-type-specific enhancers linked to cell-type-specific promoters (Figure 7A). Within the B cell compartment, all transcription factors analyzed were associated with the three enhancer groups. Ebf1, Oct2, and E2A footprints for instance showed no significant differences in their distribution regardless of enhancer specificity (Figure 7B). Other factors, such as E2f2 and Foxo1 displayed statistical significant biases for shared and cell-type-specific enhancers, respectively (Figure 7B). However, even in
these cases, footprints were not excluded from any enhancer group. A similar distribution was observed in ES cells (Figure S7C). Thus, lineage specification factors associate both with tissue-specific and broadly active enhancers. Confirming this finding, ChIP-seq analysis of the pluripotent factors Nanog, Oct4, and Sox2 showed occupancy of B and ES cell shared enhancers at the Ube2g1 locus as well as binding to ES-cell-specific enhancers at the Myc locus (Figure 7C). A global analysis of
Figure 7. Distribution of Transcription Factor Binding across the B and ES Cell Genome

(A) Classification of B and ES cell enhancers into those present in both cell types and associated with shared promoters (light blue), cell-type-specific enhancers associated with broadly active promoters (gray), and cell-specific enhancers associated with cell-specific promoters (blue).

(B) Fraction of each enhancer type harboring a particular motif in a footprint.

(C) ChIP-seq analysis of pluripotent factors Nanog, Oct4, and Sox2 at the Ube2g1 and Myc gene loci.

(D) Transcription of broadly expressed genes is driven during development by a relay race type of regulation, where the landscape of active enhancers varies in different cell types concomitant with the dynamics of CpG methylation. See also Figure S7.
Irf4, Irf8, Pu.1, Ebf1, and Stat6 ChIP-seq from B cells also corroborated the results (data not shown). We conclude that transcription factors driving lineage specification associate with regulatory elements of broadly expressed genes. We point out that the functional significance of TF recruitment as defined by digital footprintfing remains to be empirically determined. At the same time, our findings are consistent with the notion that broadly expressed genes can be regulated in a tissue-specific manner.

DISCUSSION

We have here characterized the first interactomes of gene regulatory domains in primary cells. The data provide a wealth of information from the mouse genome with thousands of promoter-enhancer pairs. Even for loci that have been extensively studied, the analysis uncovered new connectivities. At the Igk locus, for instance, we identified novel enhancers linked to 5’Ex, the regulatory domain from which NF-κB was originally isolated 27 years ago (Sen and Baltimore, 1986). ChiA-PET also uncovered three additional enhancers tethered at a long distance (up to 50 kb) to the AID gene promoter; the previously characterized AID enhancers are all located within 15 kb of the TSS. Thus, one clear advantage of ChiA-PET lies in its ability to identify long-range interactions, even when they leapfrog noninteracting genes, as is the case for 65% of all enhancers (Table S1). A striking example is a pair of giant enhancers linked to Pax5 by skipping over ~250 kb of DNA containing the Zcchc7 gene (Figure S7D).

Like other conformation capturing techniques, ChiA-PET does not directly address functionality of chromatin interactions. This can only be determined empirically by other means. Typically, enhancer activity is defined by luciferase-based plasmids or LacZ transgenes. However, these experimental approaches only provide an incomplete view of transcriptional regulation because they either lack proper chromatin structure or the influence of neighboring enhancers, insulators, and silencers is not taken into account. Conversely, genome editing provides a means to measure the impact of enhancer deletion in the physiological context. In most cases, we found that cognate promoter activity was partially reduced following enhancer ablation, supporting the model that the contribution of individual enhancers to gene expression is additive in nature. Examples of this category were enhancers linked to the Cd79a, Pou2af1, and Pim1. On the other hand, the targeting of AID 5’ enhancers E1 or E2 nearly entirely abolished AID expression and activity, a result that is reminiscent of those obtained upon deletion of AID enhancers E3 and E5 in BAC transgenic mice (Crouch et al., 2007; Huong et al., 2013), or by interference with E4 activity in primary B cells (Sayegh et al., 2003). Thus, rather than working in additive fashion, AID gene regulatory elements seem to synergize or act as a cooperative unit. Conceivably, the local topology of the Aicda-Apobec1 locus requires activation of all enhancers for optimal transcription to occur. This strategy may be useful for genes that require tight regulation as AID, whose expression is strictly limited to activated B cells to minimize its well known tumorigenic activity (Casellas et al., 2009). Consistent with this, our methylome analysis indicates that the AID locus is not completely demethylated until B cells are activated.

A direct comparison between the B and ES cell interactomes revealed that up to 95% of genes anchored in both cell types are associated with at least one tissue-specific enhancer. Transcription modularity, the mechanism whereby genes accumulate regulatory elements during ontogeny, is a well-described phenomenon that controls the spatiotemporal expression of developmental genes (Davidson, 2001; Visel et al., 2007). For instance, the cardiac homeobox gene Nkx2-5 is targeted to specific subregions of the developing heart by turning on additional cis-regulatory domains over time (Schwartz and Olson, 1999). Similarly, expression of the human apolipoprotein E gene is triggered in hepatocytes and astrocytes by enhancers only active in those tissues (Grehan et al., 2001). Our studies demonstrate that the turning on and off of enhancers is not a singularity of developmental gene loci. Instead, it is a widespread mechanism that regulates broadly expressed genes involved in basic cellular functions, such as cell-cycle initiation (Myc), signal transduction (β-catenin), and cellular motility (Malat1).

Mechanistically, we show that the changing enhancer landscape in mammalian development results from the unbiased recruitment of lineage-determining factors, which associate with enhancers anchored not only to tissue-specific promoters but also to constitutively active ones. Based on these observations we propose a “relay race” model of transcriptional regulation, whereby broadly active genes make use of tissue-specific cis-regulatory elements and transcription factors as cells progress through development (Figure 7D). For genes that only replace a subset of their regulatory domains, transcription is roughly maintained at comparable levels in different cell types. However, as the number of connected enhancers fluctuates considerably, promoter activity can be significantly altered.

We can think of at least two reasons why higher organisms modulate the enhancer landscape of broadly expressed genes. First, as aforementioned, it enables fine-tuning of protein output, which in turn controls protein activity. Second, it places basic cellular functions under the control of tissue-specific factors. In the B cell compartment, these strategies are perhaps best illustrated during the immune response to invading pathogens. In this microenvironment B lymphocytes move rapidly from a G0, quiescent state to one of the fastest proliferative rates among eukaryotic cells (Liu et al., 1991). Key in this process is Myc which, along with TFIILH, triggers a ~10-fold amplification of the lymphocyte transcriptome concomitant with cell-cycle entry (Kouzine et al., 2013; Nie et al., 2012). Thus, vis-à-vis proliferation B cells differ substantially from continuously dividing ES cells, in that they must rapidly engage a burst of Myc expression and activity during the immune response to cope with fast dividing pathogens. Our studies imply that this unique response is mediated, at least in part, by the large number of cis- and trans-responsive elements that associate with the Myc promoter in the B cell compartment.

EXPERIMENTAL PROCEDURES

Cell Isolation and Culture

Hematopoietic stem cells were isolated from the bone marrow of 26-week-old C57BL6 mice (Jackson Laboratory). Cells were purified following Ema et al.’s protocol (PMID: 17406558). KSL (Kit+, Sca1+, Lin-), IL-7R+ and CLP (Kit+, Sca1+, Lin-, IL-7R+) cells were sorted using MoFlo Legacy (Beckman Coulter) and BD FACSaria III. Resting splenic B cells were isolated from
6- to 8-week-old wild-type C57BL/6J mice with anti-CD43 Microbeads (anti-Ly48; Miltenyi Biotech) and were activated for 48–60 hr with LPS (50 μg/ml; Sigma), IL-4 (5 ng/ml; Sigma) and 0.5 μg/ml of anti-CD180 (RP109) antibody (RP/14; BD Pharmingen). E14 tg2A mouse embryonic stem cells were maintained as described in (PMID:18555785). Switchable IgM+→IgA+ murine CH12-F3 Ly-1+ B cell lymphoma line was maintained and passaged every 2 days in RPMI 1640 supplemented with 10% FBS (ATCC), 1% penicillin/streptomycin (Invitrogen), 55 μM 2-μl mercaptoethanol (Invitrogen). All cells were maintained at 37°C and 5% CO2 in a humidified incubator.

ACCESSION NUMBERS

The Short Read Archive Project Number for the ChiP-seq data in activated B cells (other than CTGF) and CH12 cells, DHS-seq in activated B cells, RNA-seq, whole genome methylation (other than ES cells), and ChiA-PET data reported in this paper is SRP029721.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.11.039.

ACKNOWLEDGMENTS

We thank Kefei Yu for CH12 cells and protocols; J. Simone and J. Lay for cell sorting; G. Gutierrez for technical assistance with Illumina analyzer. This work was supported by the Intramural Research Program of NIAMS and NCI, and internal Jackson Laboratory fund JAX19020120 to Y.R. J.K.J. was supported by NIH grants DP1 GM105378 and P50 HG005550, the Defense Advanced Research Projects Agency grant W911NF-11-2-0006 and The Jim and Ann Orr Massachusetts General Hospital Research Scholar Award. J.Q. was supported by the NIH UGSP program. All animal experiments were performed according to NIH guidelines. High-performance computation was performed using NIH Helix Systems (http://helix.nih.gov). J.K.J. has a financial interest in Transposagen Biopharmaceuticals. J.K.J.’s interests are reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

Received: September 11, 2013
Revised: November 1, 2013
Accepted: November 25, 2013
Published: December 19, 2013

REFERENCES

Meyer, K.B., and Neuberger, M.S. (1989). The immunoglobulin kappa locus in transgenic mouse B cell lymphoma line was maintained and passaged every 2 days in RPMI 1640 supplemented with 10% FBS (ATCC), 1% penicillin/streptomycin (Invitrogen). All cells were maintained at 37°C and 5% CO2 in a humidified incubator.

Cell 155, 1507–1520, December 19, 2013 ©2013 Elsevier Inc. 1519


EXTENDED EXPERIMENTAL PROCEDURES

Genome Editing

TALENs for specific loci were designed and assembled using protocols described in (Reyon et al., 2012). The donor vector included 500-1500 bp homology arms and a loxP-flanked puromycin-T2A-thymidine kinase cassette to select for targeted clones (Figure S3). Activated B or E14 cells were nucleofected with the donor cassette and the TALEN plasmid pair using Nucleofector Kit V according to the manufacturer’s instructions (Lonza). After 72 hr, limiting dilution was performed in media containing 0.5–1 µg/ml puromycin (Sigma) and incubation was continued for 6–8 days. Individual clones were picked and genomic DNA was extracted (Promega). Genotyping was done by nested PCR using locus-specific external and vector internal primers (Table S3) under the following conditions: 98°C for 30 s; 35 × (98°C for 10 s, 63°C for 30 s, 72°C for 30 s); 72°C for 3 min; hold at 4°C. PCR products were run on 1% agarose gel. Positive clones were expanded and nucleofected with a plasmid expressing CMV- or EF1α-driven Cre recombinase (Addgene). At 72 hr, limiting dilution was performed in the presence of 0.5–2 µg/ml ganciclovir (Sigma) for 6–8 days. Once again, single clones were picked and genotyped by PCR using primers amplifying deleted loci (Figure S3).

qPCR

RNA from locus-deleted clones was extracted with RNAqueous-Micro Kit (Ambion). cDNA was synthesized with SuperScript III First-Strand Synthesis SuperMix (Invitrogen). qPCR samples were mixed with SyBrGreener (Invitrogen) and run on BioRad CFX96. qPCR primers are listed in Table S3.

ChiP-Seq

Cultured cells were fixed with 1% formaldehyde (Sigma) for 10’ at 37°C. Fixation was quenched by addition of glycine (Sigma) at a final concentration of 125 mM. Twenty million fixed cells were washed with PBS and resuspended in 1 ml of RIPA buffer (10 mM Tris [pH 7.6], 1 mM EDTA, 0.1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100, 1 × Complete Mini EDTA free proteinase inhibitor [Roche]) or stored at −80°C until further processing. Sonication was performed using Covaris S2 sonicator at duty cycle 20%, intensity 5, cycle/burst 200 for 30 min. Five to ten micrograms of anti-NIPBL (Bethyl, A301-779A), anti-p300 (Santa Cruz, SC-584), anti-Med12 (Bethyl, A300-774A), anti-H3K27Ac (Abcam ab4279-100), anti-H2AZ (Abcam, ab4174-100), anti-H3K4me1 (Abcam ab8895-50), anti-H3K4me3 (Millipore 04-745), and anti-PU1 (Santa Cruz, SC-352) was incubated with 40 µl of Dynabeads Protein A (or G) for 40 min at room temperature. Antibody-bound beads were added to 500 µl of sonicated chromatin, incubated at 4°C overnight, and washed twice with RIPA buffer, twice with RIPA buffer containing 0.3 M NaCl, twice with LiCl buffer (0.25 M LiCl, 0.5% Igepal-630, 0.5% sodium deoxycholate), once with TE (pH 8) plus 0.2% Triton X-100, and once with TE (pH 8.0). Crosslinking was reversed by incubating the beads at 65°C for 4 hr in the presence of 0.3% SDS and 1 mg/ml Proteinase K. ChiP DNA was purified by phenol-chloroform extraction followed by ethanol precipitation. DNA was subsequently blunt-ended with End-It DNA end repair kit (Epicentech), and A-tailed with Taq DNA polymerase (Invitrogen) in the presence of 200 mM of dATP for 40 min at 70°C. Samples were purified by phenol-chloroform extraction after each reaction. Illumina compatible adaptors (Illumina or Bioo Scientific) were then ligated with T4 DNA ligase (Enzymatics), and the reaction was purified once with AMPure XP magnetic beads (Beckman Coulter). Samples were PCR amplified for 18 cycles with KAPA HiFi DNA polymerase mix (KAPA Biosystems) and run on a 2% agarose gel and size-selected at 200–300 bp. Thirty-six or 50 bp of sequencing data (Table S1) were acquired on the Illumina GAII or HiSeq2000 (Illumina).

DHS-Seq

Digital DNase I mapping was performed as described in reference (Sekimata et al., 2009). Briefly, we pelleted 1 × 10⁸ B cells, washed them with cold PBS and resuspended them in Buffer A (15 mM Tris–Cl (pH 8.0), 15 mM NaCl, 60 mM KCl, 1 mM EDTA (pH 8.0), 0.5 mM EGTA (pH 8.0), 0.5 mM spermidine, 0.15 mM spermine) to a final concentration of 2 × 10⁶ cells/ml. Nuclei were isolated by dropwise addition of an equal volume of Buffer A containing 0.04% NP-40, followed by incubation on ice for 10’. Nuclei were centrifuged at 1,000 g for 5 min and then resuspended and washed with 25 ml of cold Buffer A. Nuclei were resuspended in 2 ml of Buffer A at a final concentration of 1 × 10⁷ nuclei/ml. We performed DNase I (Roche, 10–80 U/ml) digests for 3’ at 37°C in 2 ml volumes of DNase I buffer (13.5 mM Tris–HCl pH 8.0, 87 mM NaCl, 54 mM KCl, 6 mM CaCl₂, 0.9 mM EDTA, 0.45 mM EGTA, 0.45 mM Spermidine). Reactions were terminated by addition of an equal volume (2 ml) of stop buffer (1 mM Tris–Cl (pH 8.0), 5 mM NaCl, 20% SDS and 0.5 M EDTA (pH 8.0), 10 µg/ml RNase A, Roche) and incubated at 55°C. After 15’, we added Proteinase K (25 µg/ml final concentration) to each digest reaction and incubated for one hour at 55°C. After DNase I treatment, phenol-chloroform extractions were performed. Control (untreated) samples were processed as above except for the omission of DNase I. DNase I double-cut fragments and sequencing libraries were constructed as described in the ChiP-seq protocol with the exception of the size-selection of 300–400 bp.

RNA-Seq

Total RNA from 10⁶ ES, resting, or activated B cells was isolated by Trizol extraction. To obtain more precise measurements of transcription, RNA spike-ins were used by adding 1 µl of 1/10 dilution of Ambion’s ERCC RNA Spike-in Mix (catalog number 4456740) to...
total RNA, mRNA was then isolated and the standard RNA-Seq library preparation was performed following Illumina’s RNA-Seq protocol v2.

**Bi-Seq**

Genomic DNA was isolated from 5x10^6 cells using QIAGEN DNeasy blood and tissue kit. Libraries were prepared following whole-genome bisulfite sequencing for methylation analysis guide from Illumina (15021861_B) with slight modifications. Briefly, 5 µg of genomic DNA was sheared and blunt-ended with End-It DNA end repair kit (Epicenter) and A-tailed with Taq DNA polymerase (Invitrogen) in the presence of 200 mM of dATP for 40 min at 70°C. Illumina compatible adaptors (5’ P-GATXGGAGAGXGTTXAG XAGGAATGXXGAXG, 5’ AXAXTTTTTTXAXAXGAXGXTXTTXGAXGTXT where X is a methylated cytosine) were then ligated with T4 DNA ligase (Enzymatics). Adaptor-ligated DNA of 275-350 bp was isolated by 2% agarose gel electrophoresis, and sodium bisulfite conversion performed on it using the Epitect Bisulfite kit (QIAGEN). Bisulfite converted DNA was divided in three tubes and PCR amplified for 6 cycles by PfuTurbo Cx hotstart DNA polymerase (Stratagene). The reaction products were purified using the MinElute PCR purification kit (QIAGEN) then separated by 2% agarose gel electrophoresis and purified from the gel using the MinElute gel purification kit (QIAGEN).

**ChIA-PET**

RNA PolII ChIA-PET was performed as previously described (Fullwood et al., 2009; Goh et al., 2012; Li et al., 2012). Briefly, B or ES cells (up to 3x10^6 cells) were treated with 1% formaldehyde at room temperature for 10 min and then neutralized using 0.2 M glycine. The crosslinked chromatin was subjected to fragmentation with an average length of 300 bp by sonication. The anti-PolII monoclonal antibody BWG16 (Covance, MMS-126R) was used to enrich PolII-bound chromatin fragments. A portion of ChIP DNA was eluted off from antibody-coated beads for concentration quantification using PicoGreen fluorometry and for enrichment analysis using quantitative PCR. For ChIA-PET library construction ChIP DNA fragments from two biological replicates were end-repaired using T4 DNA polymerase (NEB) and ligated to either linker A or linker B. Other than four nucleotides in the middle of the linkers that were used as nucleotide barcode, the two linkers share the same nucleotide sequences. After linker ligation, the two samples were combined for proximity ligation in diluted conditions. During the proximity ligation, DNA fragments within the same ChIP complex with the same linker were ligated, which generated the ligation products with homodimer linker composition. However, chimeric ligations between ChIP fragments that are bound in different chromatin complexes could also occur, thus producing ligation products with heterodimer linker composition. These heterodimer linker composition products were used to assess the frequency of nonspecific ligations and were then removed bioinformatically. As shown in Figure S2A, all heterodimer linker ligations are by definition nonspecific. Because random intermolecular associations in the test tube are expected to be comparable for linkers A and B, the frequency of random homo and heterodimer linker ligations must also be equivalent. In our PolII ChIA-PET libraries, only 3.8% of pair-end ligations (PETs) involved heterodimer linkers. Thus, we estimate that less than 5% of total homodimer ligations are nonspecific. Even though this represents but a small number of PETs, we reduced this number even further by discarding singleton PETs. In other words, we only report PolII long-range interactions when two or more pair-end reads create a PET cluster (Figure S2A). The strategy is based upon the fact that random heterodimer ligations rarely form PET clusters. In the PolII ChIA-PET libraries for instance, of 487,981 heterodimer PETs, only 26 PET clusters were obtained. Conversely, 9M homodimer PETs created ~15,000 PET clusters. Thus, while there are ~20 times more homodimer than heterodimer PETs, we obtain ~600 times more homodimer than heterodimer PET clusters. Following proximity ligation, the Paired-End-Tag (PET) constructs were extracted from the ligation products and the PET templates were subjected to paired-end sequencing using Illumina GAII.

**Bioinformatics Software**

- ABI Solid whole transcriptome alignment pipeline (WTP) and Bioscope 1.0
- BatMis 3.0 (Tennakoon et al., 2012)
- Bedtools 2.17.0 (Quinlan and Hall, 2010)
- Bowtie 0.12.8 (Langmead et al., 2009)
- ChIA-PET Tool (Li et al., 2010)
- Cufflinks 2 (Trapnell et al., 2010)
- Cytoscape 2.8.2 (Shannon et al., 2003)
- DNaSe2Hotspots (Baek et al., 2012)
- Fastqc 0.10.0 and 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- ggplot2 (Wickham, 2009)
- gsnap 2012-07-20 (Wu and Nacu, 2010)
- htseq 0.5.3p9 (Simon Anders, http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html)
- Illumina Casava (versions 1.3 to 1.8.2)
- Macs2 2.0.10 (Zhang et al., 2008)
- Samtools 0.1.18 (Li et al., 2009)
Bioinformatics Analyses

**DHS-Seq, ChIP-Seq, and RNA-Seq Processing**

DHS-Seq and ChIP-seq were sequenced using the Illumina Genome Analyzer II and the Illumina HiSeq 2000, following the manufacturer’s instructions. The standard Illumina pipeline (versions 1.3 to 1.8.2) was used for image analysis and base calling. The software Fastqc was applied to the fastq files to ensure that the quality of each read position was no less than 28.

**Alignment of DHS-Seq and ChIP-Seq Reads**

Reads were aligned to the National Center for Biotechnology Information mouse genome data (July 2007; NCBI37/mm9). The alignment software Bowtie version 0.12.8 was used with the following options: --best all-strata -m1 -n2 --[read length]. These options report the reads that align uniquely to the best stratum and allowing 2 mismatches.

**Alignment of RNA-Seq Reads**

Strand-specific B and E14 cell RNA-Seq reads were mapped onto the mouse reference genome (mm9) with SOLiD WTP and analyzed by ABI SOLiD Bioscope (version 1.0) analysis pipeline. Expression values were determined in terms of reads per kilobase per million mapped reads (RPKM).

For increased precision in transcription measurements, ERCC RNA spike-in mix was added to the total RNA before standard RNA-Seq library preparation (as described above). Reads were aligned against the mouse genome (build mm9) with gsnap using only known splice sites obtained from the Refseq annotation as present in the UCSC genome browser database in January 2012 with -novelsplicing g = 0. Subsequently, the same reads were also aligned to the ERCC RNA standard, also with gsnap but not looking for splice sites. The number of reads matching each Refseq gene was then determined using htsq-count while the number of reads matching ERCC standard RNAs was determined by a simple line count. Spike and mRNA counts were then read into R where counts were normalized by library size and exonic size of each gene to obtain RPKM (reads per kb per million aligned reads). A linear model was fitted to the ERCC spike data to relate the known copy number to the measured RPKM and cell type: lm(log10(known copy number) ?log10(RPKM) + cell, data = counts). The linear model was then used to estimate the copy number of each expressed gene based on the cell type and measured RPKM.

**PolII ChIA-PET Processing**

The pipeline for ChIA-PET sequencing processing is described in (Li et al., 2010). Briefly, the redundant sequences were collapsed into a nonredundant PET sequence set based on sequence content. The nonredundant PET sequences were analyzed for linker barcode composition and identified as sequences with homodimer linker derived from specific ligation products, or sequences with heterodimer linker derived from nonspecific ligation products. The linker composition information was used to evaluate the noise in the ChIA-PET library and sequences with heterodimer linkers were removed. After trimming the linkers, PET sequences were mapped to the mouse reference genome (mm9) using customized BatMis 3.0 and only perfectly and uniquely aligned PETs were retained. PETs uniquely aligned and similarly mapped (within ± 2 bp) were merged into one PET. The PET categories were established by evaluating the mapping orientation and genomic span between two tags of a PET. The interligation PETs were further categorized into intrachromosomal PETs, where the two tags from a PET lay on the same chromosome, and interchromosomal PETs, where the two tags lay on different chromosomes. The self-ligation PETs were utilized as a proxy for ChIP fragments since they provide two defined end points, as described above. The coverage of all self-ligated PET sequences across the genome reflects the specific PolII binding sites, which is analogous to ChIP-seq mapping for protein binding sites. The local summits of the sequence coverage were called as potential peaks if there were multiple self-ligation PETs overlapping in that region. Assuming multiple self-ligation PETs would not occur by random chance, the random background was set as the maximum of the global tag density, local tag densities at 10 kb and 20 kb windows around the peak, and the Poisson distribution was applied to estimate the p values for the peaks. P values were then adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate method (B-H FDR method). FDRs less than 0.05 were used as the criteria for final peak calling. Throughout the text, PET clusters with 2 or more PETs are referred to as interactions or connections.

**Bi-Seq Processing**

In silico conversion mimicking complete bisulfite conversion was performed on the National Center for Biotechnology Information mm9 mouse genome. Specifically, in silico C → T and G → A mm9 genomes were constructed by converting all C or Gs into T or As. Sequence reads were C → T converted and aligned to the C → T and G → A genomes, resulting in two alignments per sample. Alignments were performed with Bowtie version 0.12.8 with the options “--best -n2 -k2,” which report the top 2 best alignments. Alignments against the C → T converted genome had the additional option --norc (only alignments against the forward strand were reported) and alignments against the G → A converted genome was performed with the additional option --nofw (only alignments against the reverse strand were reported). The two alignments per sample were merged and sorted using Samtools, and only reads that had a unique alignment with the minimum number of mismatches were retained. Going through all the sorted unique alignments, the number of methylated Cs (C nucleotide in the original read and C nucleotide in the genome) and the number of unmethylated Cs (T nucleotide in the original read and C nucleotide in the genome) were determined for every position covered. The context of each C was determined and each C was classified as CpG, CHH, or CHG, where H is either A, T, or C nucleotide. For the CpGs, base positions that were consecutive and on different strands were merged. To ensure proper quality of the methylation calls, a minimal base quality score of 20 and a minimum of 5 uniquely aligned reads covering a given position were required for
that position to be considered. For each C nucleotide with sufficient quality and coverage then, we calculated the percentage of methylated C nucleotides with respect to the sum of methylated and unmethylated C nucleotides. For each CG, the difference in percent methylation between KSL or CLP or resting B cells or activated B cells and ES cells was calculated.

**Definition of DHS Hotspots and Footprints**

DHS hotspots were detected using the software DNase2Hotspots with the following parameters: background window size 200,000 bp, target window size 250 bp, mergeable gap: 0, z-score threshold 2, FDR 0%. Reads that mapped to chromosome M were removed prior to hotspot detection. The software DNase2Hotspots uses an algorithm that identifies local enrichment of tags in a 250 bp target window relative to a local background window spanning 200 kb. To remove artifacts, reads that overlapped satellites, long interspersed repetitive elements, and short tandem repeats were removed. With the remaining reads, a z-score, \((n – \mu)/\sigma\) is calculated for each target window where \(n\) is the original number of tags observed in the target window, \(\mu\) is the expected number of tags overlapping that target window based on the number of mappable reads in the background window, and \(\sigma\) is the standard deviation of the expectation. A false discovery rate is estimated for each z-score by calculating z-scores using randomly sampled uniquely mapped reads from the observed data set. Hotspots with 0% FDR were selected here. Finally, DHS hotspots located within 1 kb of TSSs were merged into promoter domains, while all other DHS were merged into enhancer domains if they were within 2 kb from each other.

Transcription factor footprints were called by applying ‘DNase2TF’, a program that implements a new footprint detection algorithm (S. Baek, G.L. Hager, M.-H. Sung, unpublished data). Briefly, the algorithm looks for regions of local depletion in DNaseI cutting within each hotspot, taking the enzyme’s dinucleotide bias and mappability of sequence reads into account. Candidate intervals are obtained by considering binomial z-scores that reflect the depletion in cutting relative to a background window 3-fold the width of the candidate region and then by merging consecutive intervals. Candidate intervals whose z-scores correspond to FDR < 5% were retained for subsequent analyses.

**Annotation of DHS Regions into Promoters and Enhancers**

The Bedtools software suite was utilized to map the ChIP-seq reads onto the DHS regions. First, the fragment sizes for each ChIP-seq experiment was estimated using macs2. The 3’ end of the aligned ChIP-seq reads were then lengthened to match these predicted fragment sizes. Next, the read counts overlapping each DHS region was calculated using bedtools intersect. These overlapping read counts were then normalized to the total number of reads for each epigenetic mark and p values based on the negative binomial distribution of the negative controls (a combination of negative pull-downs for B cells and IgG for ES cells) were calculated for each region. DHS regions with a Benjamini-Hochberg FDR-adjusted p value less than 1% were considered positive. DHS regions that were within 1 Kb of a TSS were denoted DHS-promoter. The remaining regions were classified as enhancers if they were p300 or Med12 or Nipbl positive.

**ChIA-PET Cluster Interactions: Definitions and Annotation**

Interligation PETs reflect long-range chromatin interactions. However, there is technical noise from various sources, which should be inevitably considered. To determine if an interligation PET represents a specific interaction event between two DNA fragments that are bound together in close spatial proximity by a PolII protein complex, we reasoned that the multiple interligation PETs would be enriched by the ChIP procedure between the same DNA fragments. To identify such chromatin interactions, both ends of the interligation PETs were extended by 500 bp along the reference genome, and PETs overlapping at both ends were clustered together as one PET cluster. The number of PETs in a PET cluster therefore reflects the frequency of an interaction between two genomic regions. To determine whether the observed number of PETs in a PET cluster was significantly different from background noise or weak interactions represented by singleton interligation PETs, we evaluated p values from a hypergeometric distribution with the tag counts from both anchor regions of PET clusters and the sequencing depth as input. P values were corrected using the B-H FDR method for multiple hypothesis testing and the FDR cutoff is 0.05.

**Classification of PET Clusters**

In this study, we only considered the intrachromosomal PET clusters with span less than 1 Mb on chromosomes and used the defined DHS regions to annotate PET cluster interactions. DHS regions were extended by 1.5 kb in both directions and interaction PET clusters were considered anchored if they overlapped the extended DHS regions by at least 1 base. Interaction PET clusters were then annotated into the following 4 types, according to the type of DHS region they overlapped: intragenic (promoter to gene internal region), extragenic (promoter to enhancer), and intergenic (promoter to distal promoter), and enhancer-enhancer (neither anchor of a PET cluster overlaps with a DHS promoter region). In some cases, interaction PET clusters overlap with both DHS promoters and enhancers. If the interaction PET cluster overlapped with a DHS promoter region and a DHS enhancer region located in the same gene internal region, the interaction was denoted as intragenic interaction. If the interaction PET cluster overlapped with a DHS promoter and a distal DHS enhancer, the interaction was denoted as an extragenic interaction as shown in Figure S1H.

**Defining Transcription Models from Chromatin Interactions**

PET interaction clusters were grouped with other interaction clusters based on what type of DHS region they were anchored to. We then defined three transcription models based on how the gene promoters were involved in these complex chromatin interactions: multigene (MG) interaction model, single gene (SG) interaction model, and basal promoter (BP) model. The MG model comprises intragenic promoter-promoter interactions grouped in an interaction cluster that could also include intragenic and extragenic enhancer-promoter interactions. The SG model consists of single or multiple enhancer interactions with only one gene promoter.
whereas the BP model includes genes with PolII binding but no chromatin interaction. MG and SG interaction cluster models were visualized with Cytoscape.

**Reproducibility of ChIA-PET PolII Peaks and Interactions**
Since the sonicated fragment size of ChIA-PET library is less than 1 kb, PolII peaks between biological replicates were considered overlapping if the distance from peak center to peak center was within 1 Kb. To account for the sonication fragment size anchors of interaction PET clusters were extended 1 kb at both ends. Extended PET clusters were considered overlapping if they shared at least one nucleotide.

**Identification of lncRNAs Associated Genes**
The lncRNA-associated genes were identified by PET clusters, which connected the DHS promoter regions of lncRNA and their target gene. Ensemble version NCBI36.67 was utilized for annotation. When a lncRNA associated with multiple genes, the gene with highest connection frequency was selected as the lncRNA-associated gene. Expression levels of lncRNAs and associated genes were measured as reads per kilobase per million reads (RPKM) from RNA-Seq sequencing by using Cufflinks 2 (Trapnell et al., 2013).

**CpG Methylation Changes across Cell Types**
RPKM values were calculated for genes annotated using Refseq annotation as found in the UCSC genome browser database (January 2012). Using R, a two-component mixture model was fit to the log-transformed RPKM values and genes with expression levels exceeding the 95th percentile were considered active while genes with expression levels below the 1st percentile were considered silent. Methylation levels was averaged over each gene (TSS ± 200 bp) and the distribution of methylation levels in silent and active genes were depicted with violin plots (Figure 5A). To assess global demethylation in enhancers across cellular development (Figure 5D), we considered DHS enhancer regions anchored by PETs in activated B cells only. A subset of these regions were noted as demethylated (average methylation ≤ 40%) in activated B cells and the number of these regions is depicted in Figure 5D.

**Accession Numbers**
All accession numbers for raw data are provided in Table S1. All bed files used in the text are also available upon request.

**SUPPLEMENTAL REFERENCES**


Figure S1. Properties of DHS-Seq and ChIA-PET Data Sets, Related to Figure 1.

(A) Reproducibility of DHS-seq signal at defined hotspots in two biological replicates of activated B cells. Correlation was calculated via Spearman's coefficient ρ.

(B) Overlap in B cell DHS regions vis-à-vis Med12, p300, and Nipbl occupancy.

(C) Bar graph showing the percentage of ChIA-PET interactions overlapping with DHS or non-DHS genomic domains.

(D) Box plot representing the transcription levels of genes associated or not associated with ChIA-PET connections at promoters (p < 2\textsuperscript{-16}).

(E) Box plot representing PolII density (RPKM) at promoters associated or not associated with ChIA-PET connections (p < 2\textsuperscript{-16}).

(F) Percentage of H3K27Ac\textsuperscript{+} (active) enhancers within ChIA-PET anchored and not anchored groups.

(G) DHS signal intensity at enhancers and promoters associated with 0, 2, or more than 5 PolII long-range interactions.

(H) Schematics showing the classification of PolII ChIA-PET interactions (PETs) as intragenic, extragenic, intergenic, or enhancer-enhancer.

(I) Representation of direct and indirect connections between promoters and enhancers as determined by ChIA-PET. Percentages were calculated for B and ES cell ChIA-PET data sets combined.
Figure S2. Analysis of PolII Long-Range Interactions via ChIA-PET, Related to Figure 1

(A) Schematics showing details of the ChIA-PET protocol, including ChIP pull-down and chromosome conformation capturing steps. To facilitate their representation throughout the main text, individual PolII long-range interactions or PETs within 500 bp from each other were grouped into PET clusters. By definition (see Experimental Procedures), two elements are considered to be connected if they are linked by a PET cluster of 2 or more individual PETs. As in previous ChIA-PET publications, singletons were excluded from the analysis.

(B) Reproducibility of PolII ChIA-PET peaks at biological replicates from activated B cells and ES cells.

(C) Overlapping PET clusters from two B cell and ES cell ChIA-PET biological replicates. The panel depicts the reproducibility of interacting PET clusters where each dot represents PET counts from replicates and dot sizes are proportional to the number of overlapping PET clusters.
Figure S3. Regulatory Domain Interactions Determined by ChIA-PET at Sox2 and Igh Gene Loci from ES Cells and LPS+IL-4 Activated B Lymphocytes, Respectively, Related to Figure 1

(A and B) ChIA-PET and DHS data sets in each cell type are provided for both loci. At Igh, the number of PETs anchoring the constant domain and the 3'Eα enhancer are provided. mRNA expression is also provided as RPKM values (+ strand transcription in green, − strand in blue).
Figure S4. Validation of ChIA-PET Connections via Genome Editing, Related to Figure 2
(A) Custom FLASH TALENs were engineered to selectively delete specific DHS enhancer domains in CH12 B cells or ES cells. The donor construct contains a loxP-PGK promoter/puromycin-T2A-thymidine kinase/PolyA-loxP cassette to select positive clones. PCR primers (Table S3) specific for genomic and construct sequences were used to verify the insertion of knockout constructs at the desired targeted genomic locus. The cassette was removed by Cre-mediated recombination and clones were selected with Ganciclovir and verified by PCR.
(B) Targeted deletion of AID enhancer E1 results in a marked decrease in PolII occupancy at AID gene regulatory domains but not at the Foxj2-Necap1 locus (C24 190 kb downstream of Apobec1).
(C) Pou2af1 enhancer E3 was deleted and Pou2af1 expression levels were measured by qPCR (bar graph) in nonactivated (N.A.), or CH12 cells (WT, black bars; and D E3, yellow bars) activated in the presence of IL-6, a CD40, TGFb, and IL4. P values were 0.15 (N.A.), 0.008 (IL6), and 0.03 (a CD40+TGFb+IL4). To demarcate regulatory domains both DHS (black) and H3K4me3 (red) profiles are provided.
(D) Enhancers 1 (E1) and 2 (E2) at the Cd79a locus were deleted and transcription of genes within (Rps19, Cd79a, and Arhgef1) and outside the cluster (Pou2f2 and Rps3) was assessed by qPCR (right bar graph) in WT (black bars), ΔE2 (gray bars), and ΔE1 (yellow bars) CH12 cells. Data represent the mean ± SEM (n = 6). P values were 0.009 (Cd79a, ΔE2), 0.0001 (Cd79a, ΔE1), 0.06 (Rps19, ΔE2), and 0.04 (Arhgef1, ΔE1). Note: We point out that most PolII long-range interactions link E1 to E2, instead of tethering the enhancers to the Cd79a promoter. One interesting possibility is that these interactions help establish the higher-order chromatin structure of the two loci, which in turn may regulate promoter activity as has been shown at the mouse b-globin locus within the context of CTCF binding (Phillips and Corces, 2009). Notably, E1 and E2 enhancers in question display clear CTCF occupancy, which might facilitate their association (depicted with a semi circle). A similar scenario applies at the Pax5 locus (see Figure S7D below).
Figure S5. Characterization of PET Gene Clusters, Related to Figure 3
(A) Distribution of the cluster span (intracluster distance) for single- and multiple-promoter gene clusters. Distances are provided in megabases (Mb).
(B) PET cluster associated with Mir290-295 locus.
(C) Heat map: PolII recruitment at promoters where no ChIA-PET interactions were detected (not-anchored) or at promoters from single- or multiple-gene clusters (for definitions see Figure 3 main text). Bar graph: transcription levels, as measured by RNA-Seq (FPKM values), of the three promoter groups.
(D) Clustering of the lymphoid signaling Gimap gene family. The enhancer located downstream of the Gimap6 gene (boxed) is unique among B cell gene regulatory domains in that it interacts with 7 different Gimap promoters (red arrows).
(E) PET interactions between Bcl11a and lincRNA E123592. DHS islands are provided to delineate regulatory elements.
Figure S6. Changes in CpG Methylation during B Cell Development, Related to Figure 4

(A) Long-range interactions and DHS profiles at loci containing the B-cell-specific Cd79b gene.

(B) Box plot showing CpG methylation levels (y axis) relative to number of PolII long-range interactions in B cells (x axis).

(C) Comparison of methylation levels (%) at DHS enhancer regions present only in ES cells (blue line), activated B cells (red line), and in both cell types (black line).

(D) and (E) Relative demethylation profiles at Pim1 and Aicda gene loci in activated (a) and resting (r) B cells, as well as CLP, and KSL bone marrow progenitors relative to ES cells (the methylome of ES cells was subtracted from that of each cell type as in Figure 5B of the main text). Enhancers and promoters are highlighted on top of each graph.
Figure S7. Distribution of Transcription Factor Footprints at Promoters and Enhancers, Related to Figure 7

(A) Bar graph representing the enrichment of transcription factor DNA motifs within footprints, DHS islands, or genomic DNA not associated with DHS islands in B cells. Motif enrichment was calculated based on the percentage of nucleotides within the particular domain that overlapped with DNA motifs.

(B) Extended view (±1 kb) of digital footprint signatures for Sp1 and Erf1.

(C) Distribution of ES cell transcription factor motifs at the three enhancer groups defined in Figure 7A of the main text (enhancers present in both cell types and associated with shared promoters [light blue], cell-type-specific enhancers associated with promoters active in B and ES cells [gray], and cell-type-specific enhancers associated with cell-type-specific promoters [blue]).

(D) Long-range interactions and DHS profiles at the B-cell-specific Pax5 gene. Two linked enhancers ~250 kb from the Pax5 TSS are highlighted. Note: As discussed within the context of the Cd79a locus (Figure S4D), the Pax5 interacting enhancers are recruit substantial CTCF, which might facilitate their tethering in 3D space (depicted with a semi-circle).