Spring 5-15-2017

Epigenetic Regulation of Lymphocyte Development and Transformation

Yue Huang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Eugene Oltz, Chair
Douglas Chalker
Barak Cohen
John Edwards
Jacqueline Payton
Barry Sleckman

Epigenetic Regulation of Lymphocyte Development and Transformation

by
Yue Huang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I'm thankful to my Ph.D. advisor and dissertation mentor, Gene Oltz, for the opportunity to work in his lab and his mentorship over the past seven years. Gene holds a high standard for scientific research. Towards this goal, he has trained me diligently with the passion for new discoveries, continuing encouragement, constant critiques and patience. I also appreciate his training for reading literature, being critical, writing, presenting and collaborating.

I'm grateful to all the members of the Oltz lab, both past and present, for scientific and emotional support over the years. Especially, I would like to thank Olivia Koues and Suhasni Gopalakrishnan, who have provided tremendous technical guidance and scientific suggestions for my thesis projects.

A thank you to my dissertation committee for guiding my dissertation progress. Special thanks to Jackie Payton, who has served almost as my co-mentor. I wouldn't have accomplished this much without many discussions and valuable insights from her. I would also like to thank Barak Cohen, who often gives different perspectives and reminds me to go back to the big picture.

I'm thankful to the Division of Biology and Biomedical Sciences, and the Washington University, for outstanding classmates and colleagues, prestigious and collaborative scientific environment, healthcare services, career development and logistics.

I'm indebted to my friends, especially in the Faith Hope Love Fellowship and St. Louis Chinese Christian Church. They have provided a second home for me in St. Louis, in which we share our tears and laughter, listen, give and accept one another.

A sincere thank you to my parents, for raising me up and always supporting me with unconditional love. Thanks also go to my parent-in-laws for their constant care and love. The completion of my graduation work will gratify and comfort their hearts.

I cannot thank enough to my husband, Yin, for all our time together and all he has done for me. Being in long-distance relationship for over a decade is tough. Both of us being in graduate school makes it tougher. I sincerely thank my husband for being my carer, helper, comforter, listener, counselor, advisor, teacher, and friend. Above all, thank you for being my husband with abundant love.

Finally, I appreciate the opportunity to come to the US, and become a Christian in St. Louis. Thank you, Jesus for your love, comfort and guidance, letting me know who I am, and giving me hope for future.

Yue Huang

Washington University in St. Louis

May 2017

ABSTRACT OF THE DISSERTATION

Epigenetic Regulation of Lymphocyte Development and Transformation

by

Yue Huang

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2017

Dr. Eugene M. Oltz, Chair

Cell identity and function rely on intricately controlled programs of gene regulation, alterations of which underlie many diseases, including cancer. Epigenetic analyses of normal and diseased cells have started to elucidate different facets of epigenetic mechanisms for gene regulation. These include changes in nucleosome density, histone modifications, factor binding and chromosomal architecture. All of these aspects contribute to the activities of regulatory elements conferring promoter, enhancer and insulator functions and the *cis*-regulatory circuits formed by these elements. Despite this progress, an urgent need remains to profile these features and to study how they cooperatively function in normal and pathogenic settings. Here, using the mouse *T cell receptor beta* locus as a model, we first quantified 13 distinct features, including transcription, chromatin environment, spatial proximity, and predicted qualities of recombination signal sequences (RSS), to assess their relative contributions in shaping recombination frequencies of Vβ gene segments. We found that the most predictive parameters are chromatin modifications associated with transcription, but recombination efficiencies are largely independent of spatial proximity. These findings enabled us to build a novel computational

model predicting Vβ usage that uses a minimum set of five features. Expanding on these results, we applied chromatin profiling and computational algorithms to other mouse antigen receptor loci, to classify and identify novel regulatory elements. We defined 38 chromatin states that reflect distinct regulatory potentials. One of these states corresponded to known enhancers and also identified new enhancer candidates in *immunoglobulin* loci. Indeed, all four candidate elements exhibited enhancer activity in B cells when subjected to functional assays, validating that our chromatin profiling and computational analyses successfully identified enhancers in antigen receptor loci. Finally, we translated these approaches to human B cell lymphoma to predict pathogenic *cis*-regulatory circuits composed of dysregulated enhancers and target genes. We then selected and functionally dissected a pathogenic *cis*-regulatory circuit for the mitosis-associated kinase, *NEK6*, which is overexpressed in human B cell lymphoma. We found that only a subset of predicted enhancers is required to maintain elevated *NEK6* expression in transformed B cells. Surprisingly, a B cell-specific super-enhancer is completely dispensable to maintain *NEK6* expression and chromatin architecture within its chromosomal neighborhood. Moreover, we showed that a cluster of binding sites for the CTCF architectural factor serves as a chromatin boundary, blocking the functional impact of a *NEK6* regulatory hub on neighboring genes. These results emphasize the necessity to test predicted *cis*-regulatory circuits, especially the roles of enhancers and super-enhancers, when prioritizing elements as targets for epigenetic-based therapies. Our findings collectively pave the way for future investigations into the roles of *cis*-regulatory and architectural elements in regulating gene expression programs during normal development or pathogenesis.

Chapter 1 : Introduction

1.1 Mechanisms of Gene Regulation

Stringent regulation of gene expression is a pivotal process in biology, relying on the interplay among *cis*-regulatory elements with promoter, enhancer and insulator functions. Recent studies on gene regulatory mechanisms have started to reveal regulatory elements of different types by genome-wide chromatin profiling. In particular, tens of thousands of enhancers are predicted to be active in any human cell type (Hnisz et al., 2016a). However, the functional impact of most predicted enhancers on gene expression remains unclear. Researchers have begun to predict target genes of enhancers using computational approaches, which build *cis*-regulatory circuits composed of enhancers and associated gene promoters (Koues et al., 2015; Maurano et al., 2012). Correlative chromatin patterns at promoters and enhancers are frequently used to connect these regions *in silico*. In addition, chromatin structures formed by insulator elements are applied to circuitry generating methods, because genome architecture has emerged as a fundamental aspect to regulate promoter-enhancer interactions. Identification and functional dissection of *cis*-regulatory elements and circuits will improve our understanding of gene regulatory mechanisms in normal development, as well as in pathogenesis, because perturbations of enhancers and insulators contribute to dysregulated gene expression programs in many diseases (Koues et al., 2015; Maurano et al., 2012; Hnisz et al., 2016b; Lupiáñez et al., 2015).

Conventional enhancers

First discovered 30 years ago, enhancers are characterized as short DNA sequences that activate transcription in a location-, orientation-, and promoter-independent manner (Bulger and Groudine, 2011). Enhancers are located distal to gene promoters and can act over long distances within the same chromosome. For instance, the limb bud enhancer of *sonic hedgehog* (*Shh*) gene is located 1 Mb upstream of *Shh* transcription start site, and is required for *Shh* transcription in limb bud (Lettice et al., 2003; Sagai et al., 2005). A dominant model for enhancer function is that they interact with promoters via looping, which has been extensively supported in different systems. Empirically, enhancers are characterized as nucleosome-free regions bound by transcription factors (TFs), as well as the transcription coactivator EP300, and are marked by histone H3 lysine 27 acetylation (H3K27ac) (Bulger and Groudine, 2011). Proper enhancer function is crucial for all cellular processes, and dysregulation of enhancer activity leads to diseases, including different types of cancers (Sur and Taipale, 2016). Enhancer misregulation mainly arises from two sources: inherited or acquired sequence variants that alter enhancer activities *in cis*, and mutations in TFs or histone modifiers that affect enhancer activities *in trans*. Indeed, genetic variants that predispose to cancer are enriched at enhancers, as revealed by analyses of genome-wide association studies (GWAS). In addition, tumor-specific enhancers have been identified by epigenome comparisons of tumor cells and their normal counterparts (Akhtar-Zaidi et al., 2012; Koues et al., 2015; Maurano et al., 2012). Therefore, it remains an important goal to identify enhancers across the genome in different cell types, especially in diseased cells. This information will facilitate epigenetic-based therapeutic applications, in which key dysregulated enhancers are targeted to reverse the expression of associated pathogenic genes.

Super-enhancers

In addition to conventional enhancers, a special class of regulatory elements, coined super-enhancers (SEs), have recently been revealed from epigenome analyses (Whyte et al., 2013). SEs encompass large hyperacetylated clusters of conventional enhancers (CEs), which bind lineage-restricted TFs. SEs co-localize with a limited set of genes that are most essential for controlling cell identity. SEs are hotspots of disease-associated variants, which are thought to destroy TF binding sites and abolish SE function, as well as the expression of associated genes (Hnisz et al., 2013; Koues et al., 2016). Furthermore, SEs are amplified or acquired *de novo* near oncogenes and contribute to several classes of cancer (Hnisz et al., 2013; Mansour et al., 2014). Therefore, therapeutic applications targeting SEs may effectively reverse expression of key pathogenic genes. For example, SEs and associated oncogenes may be preferentially inhibited in some solid tumor types by a BET-bromodomain inhibitor targeting the transcriptional coactivator, BRD4 (Chapuy et al., 2013; Lovén et al., 2013). Although SEs may be high priority therapeutic targets, very few of these regulatory regions, which are simply identified by computational algorithms, have been experimentally verified, and their roles in controlling gene expression are mainly based on correlations with nearest genes. Therefore, future research on super-enhancers requires more experimental validations of predicted SEs and better algorithms of assigning SEs to target genes.

Protein factors regulating gene expression

Regulatory DNA elements control transcription by virtue of their association with different sets of regulatory factors, including RNAPII (RNA polymerase II), various TFs and global structural proteins. During gene activation, enhancers interact with their target promoters

and initiate transcription in a temporal fashion. First, sequence-specific enhancers recruit activator proteins and co-activators, including histone modifying enzymes and ATP-dependent chromatin remodeling complexes. The loaded enhancers interact with target promoters via the transcriptional Mediator complex, which associates with RNAPII and TFs to facilitate enhancer-promoter looping and transcription (Allen and Taatjes, 2015). Finally, the promoter-enhancer complex recruits general transcription factors and RNAPII to activate transcription (Ong and Corces, 2011).

In addition, recent findings have highlighted the roles of CCCTC-binding factor (CTCF) and cohesin in regulating gene function and constructing global chromatin architecture. CTCF is a highly conserved zinc finger protein with widespread regulatory functions, including transcription activation and repression, insulation and global organization (Ong and Corces, 2014). Cohesins are protein complexes that function in sister chromatid cohesion, chromosome segregation, DNA repair, long-range looping and gene regulation, through forming a ring structure and encircling DNA (Dorsett, 2011). The fundamental roles of CTCF and cohesin in gene regulation and genome organization will be elaborated in the next section.

Chromosomal architecture

In addition to regulatory elements and protein factors, chromosomal architecture fundamentally contributes to diverse cellular processes, including transcription, recombination and DNA repair (Schoenfelder et al., 2010). Recent studies have shown that the mammalian genome is compartmentalized into topologically associated domains (TADs) (Dixon et al., 2016). TADs serve as building blocks of the genome architecture, facilitating interactions among loci within the same TAD and prohibiting interactions across TADs. Consistent with this role,

TADs are highly conserved among different cell types and even species (Dixon et al., 2012). At a biochemical level, TADs are chromatin loops created through dimeric interactions between CTCF proteins, when bound at two boundary elements in a convergent orientation, and are stabilized by association with cohesin (Hnisz et al., 2016a). Each TAD is divided into sub-structures called sub-TADs, insulated neighborhoods, or contact domains, which differ, at least partially, among cell types and developmental stages (Dixon et al., 2016; Hnisz et al., 2016a). On a fundamental level, these contact domains are composed of structural and regulatory loops. Structural loops are formed between two boundary elements by a CTCF-CTCF homodimer and the cohesion complex. These basic structures segregate active and inactive chromatin regions, restrict inappropriate interactions across regions, and promote enhancer associations with target genes between the two loop anchors. In contrast, regulatory loops are generated between promoters and enhancers by cohesin and the Mediator complex, activating transcription of associated promoters (Kagey et al., 2010). These chromosome structures are essential for gene expression in normal cells, and are frequently perturbed in pathogenesis. For example, CTCF binding sites are enriched for disease-associated variants, which abolish CTCF binding motifs and boundary functions, leading to abnormal associations between enhancers and alternative promoters in diseases (Hnisz et al., 2016b; Lupiáñez et al., 2015).

With emerging knowledge regarding chromosomal architecture, several important questions need to be addressed: What are the chromatin interactome patterns in different cell types? Does chromosomal interaction between two loci indicate correlative or causal relationships? Are TADs, sub-TADs, contact domains and insulated neighborhoods fundamentally different or do they represent the same feature identified from different data sources and algorithms? How do structures of sub-TADs form and vary during development,

differentiation and transformation? Answers to these questions will shed light on underlying mechanisms of gene regulation, and provide insights into causes of many human diseases.

1.2 Antigen Receptor Genes

Antigen receptor (AgR) genes are an excellent model to study gene regulatory mechanisms, regarding enhancers, transcription factors and chromosomal architecture. AgR genes exist in a non-functional germline configuration: the 5' portion of the genes encoding antigen recognition domains are composed of arrays of variable (V), diversity (D, only in some loci) and joining (J) gene segments (Schatz and Ji, 2011). For example, the germline mouse *Igh* gene contains 150 variable (V_H), 9 diversity (D_H), and 4 joining (J_H) gene segments (Figure 1.1). During early lymphocyte development, the V, D and J regions are randomly assembled into a functional antigen receptor gene, a process called V(D)J recombination (Schatz and Ji, 2011). By this means, lymphocytes generate a large repertoire of *immunoglobulins* and *T cell receptors* with different antigen specificities. V(D)J recombination is initiated by the Recombination activating gene 1 (RAG1) and RAG2 proteins, which form a recombinase complex that binds to and cleaves conserved recombination signal sequences (RSS) flanking all V, D, and J gene segments. Cleaved ends on genomic DNA are subsequently rejoined by non-homologous end

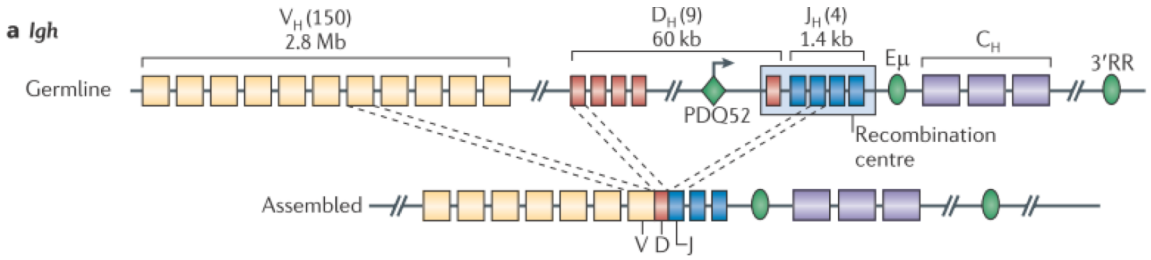


Figure 1.1: V(D)J recombination of the *immunoglobulin heavy chain* (*Igh*) locus

joining (NHEJ) pathway.

V(D)J recombination is a lineage-specific event and under tight developmental regulation, which provides a unique system to study gene regulation, development, and cellular differentiation. After the common lymphoid progenitor (CLP) differentiates into progenitor-B (pro-B) cells, pro-B cells undergo D_H and J_H segment rearrangement on both *Igh* alleles, and then perform V_H to DJ_H recombination to generate one functional *Igh* gene (Figure 1.2). Functional IgH protein is produced and associates with a surrogate immunoglobulin light chain (IgL) protein and forms the precursor-B cell receptor (pre-BCR), which signals the cells to proliferate and differentiate into precursor-B (pre-B) cells. Next, *immunoglobulin light chain* genes, *immunoglobulin kappa* (*Igk*) and *immunoglobulin lamda* (*Igl*), undergo recombination sequentially until one functional light chain gene is generated. Finally, the pre-B cells differentiate into immature B cells, functional IgH and IgL proteins are expressed and assembled into effective immunoglobulin M (IgM isotype). The tight control of V(D)J recombination ensures that each mature B cell expresses only one functional antigen receptor, minimizes wasteful recombination events, and prevents chromosomal aberrations leading to lymphoid malignancies. In a similar sequential process, *T cell receptor* genes undergo V(D)J

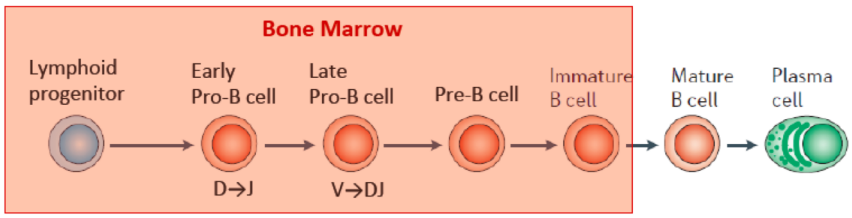


Figure 1.2: immunoglobulin gene assembly during B cell development
(Modified from Nagasawa 2006)

recombination in developing thymocytes. After the expression of precursor B or T cell receptor, the progenitor-B or T cells shut down further rearrangement at the other *Igh* or *T cell receptor β* (*Tcrb*) allele to ensure the monospecificity on mature lymphocytes (Thomas et al., 2009).

V(D)J recombination is regulated by chromatin accessibility (Yancopoulos and Alt, 1985), correlating with an open chromatin state characterized by active DNA and histone modifications, nuclease accessibility, and germline transcription (GLT) (Schatz and Ji, 2011; Thomas et al., 2009). GLT denotes transcription of unrearranged AgR genes before or during recombination. With regard to chromatin accessibility, multiple promoters, enhancers and protein factors play key roles in establishing open chromatin marks, recruiting RNA polymerase II (RNAPII) and maintaining GLT during locus activation (Chakraborty et al., 2009; Subrahmanyam and Sen, 2010).

The nuclear position and architecture of AgR genes are also important for regulating V(D)J recombination. AgR loci undergo three forms of chromosomal movement during rearrangement. First, in early pro-B cells, both *Igh* alleles move from the repressive nuclear periphery to a more central location. Shortly after this movement, *Igh* undergoes locus contraction, in which V_H gene segments are brought into spatial proximity to the D_H-J_H cluster and V(D)J recombination occurs. After the production of a functional *Igh* gene, the other non-functional *Igh* allele de-contracts and relocates to pericentric heterochromatin to prevent further recombination (Jhunjhunwala et al., 2008; Roldán et al., 2005; Thomas et al., 2009). Similar processes occur in *T cell receptor* genes in developing thymocytes. Repositioning and contraction of AgR loci require orchestration among multiple *cis*-regulatory elements and transcription factors (Guo et al., 2011). In conclusion, V(D)J recombination is controlled by

multiple *cis*-regulatory DNA elements, chromatin accessibility, locus contraction and conformation, serving as an excellent model to study gene regulatory mechanisms.

1.3 *cis*-Regulatory Circuitry in B Cell Lymphoma

In addition to their roles in governing normal development, gene regulatory mechanisms underlie pathogenic gene expression programs of diverse malignancies. It is crucial to translate findings of gene regulatory determinants in normal cells to primary human tumors. In this respect, one high-priority cancer type to focus on is Non-Hodgkin lymphoma (NHL), in which altered activities of *cis*-regulatory elements and TFs contribute to pathogenic gene expression programs. NHL is the most prevalent blood cancer and the fifth most common form of malignancy diagnosed in the US, striking >70,000 Americans annually. NHLs are characterized by deregulated expression of large gene cohorts that mediate unchecked cell growth, and are classified into distinct subtypes by gene expression profiles (Alizadeh et al., 2000). Follicular lymphoma (FL), the second most common NHL, is an incurable malignancy that exhibits an indolent clinical course, but often transforms to a more aggressive lymphoma type (Lenz and Staudt, 2010). The molecular basis of altered gene expression patterns in NHL has been revealed by recent studies. ~30% NHL harbors recurrent somatic mutations in chromatin modifier genes (*EZH2*, *MLL2*, *CREBBP*, *EP300*), alterations of which are thought to result in a globally repressed chromatin state and gene expression pattern (Morin et al., 2010). The pathogenic gene expression programs in FL are coordinated by the dysregulation of TFs and their targeted enhancers when comparing tumor B cells with their normal counterparts, termed centrocytes (CCs) (Koues et al., 2015). This evidence suggests that epigenetic dysregulation is a common mechanism for widespread gene expression changes in NHL.

With emerging transcriptome and epigenome data in different cell types, the next step is to connect regulatory elements with target genes and construct the *cis*-regulatory circuitry. Different computational methods have been utilized to connect an enhancer to targets. These include, assigning an enhancer to: (1) the nearest gene, (2) genes within an arbitrary genomic distance, e.g. <1 Mb (Rödelsperger et al., 2011), (3) genes located in the same CTCF block designated by two adjacent CTCF binding sites (Heintzman et al., 2009), (4) promoters in the same chromosomal region defined by correlative chromatin marks (Shen et al., 2012), (5) promoters within 250 kb that have correlative DNase hypersensitivity signals with those at enhancers (Thurman et al., 2012), (6) genes that have conserved synteny with the enhancer among species (Rödelsperger et al., 2011), (7) genes located in the same topological domain (Dixon et al., 2012), (8) genes showing significant interaction frequencies with the enhancer in genome-wide interactome data (Mifsud et al., 2015; Sanyal et al., 2012), (9) genes with combinatorial patterns of features, including transcription, factor binding and chromatin marks, identified by machine learning algorithms (Whalen et al., 2016). These methods are starting to generate paradigms for building *cis*-regulatory circuits. Similar pathogenic *cis*-regulatory circuitries have been constructed in FL by integrative transcriptome and epigenome analyses and correlation-based methods, composed of predicted connections between altered enhancers and promoters of dysregulated genes (Koues et al., 2015). However, based on various assumptions, these circuit-building methods either simplify the complexity of regulatory networks or include many redundant and false connections between enhancers and genes.

Now large sets of transcriptome and epigenome data have accumulated from numerous cell types in scarce populations. Using these data, theoretical *cis*-regulatory circuits have been tentatively constructed, connecting potential enhancers and their putative target genes in normal

and diseased cell types. Specifically targeting pathogenic *cis*-regulatory circuits is a promising direction to achieve the goal of precision medicine. Indeed, sequence-specific epigenetic manipulations have successfully altered the expression levels of target genes using engineered TFs linked with Zinc finger, TALEN or Cas9 (Luo et al., 2017; Vora et al., 2016). The huge advantage of this approach lies in that tumor-specific dysregulated enhancers can be targeted and reversed, restoring normal expression of associated genes in malignant cells, while leaving normal cells intact. These therapeutic applications require selecting key regulatory elements with stringent experimental validations. However, most of current predictions regarding *cis*-regulatory circuits remain untested at a functional level. Thus, functional dissection of predicted circuits will enable us to test these predictions and begin to translate research findings into clinical applications.

1.4 Scope of Thesis

In this dissertation, I focus on gene regulatory mechanisms in normal development, specifically for antigen receptor loci in developing lymphocytes, and in disease, focused on B cell lymphoma. In Chapter 2, I describe our studies on the contributions of different features in shaping Vβ usage in primary mouse *Tcrb* repertoires (Gopalakrishnan et al., 2013). We built a unifying computational model, finding that chromatin modifications and transcription, but not spatial proximity, determine recombination frequencies of Vβ gene segments. This strategy will help predict immune cell repertoires in normal and altered antigen receptor loci. In Chapter 3, I describe our efforts to classify and identify regulatory elements in other mouse antigen receptor loci using computational algorithms to analyze chromatin profiles (Predeus et al., 2014). We successfully classified known enhancers and identified novel enhancers in mouse

immunoglobulin loci, which were further validated by functional assays. These data will facilitate future studies in immune receptor regulation. In Chapter 4, I describe the functional dissection of a pathogenic *cis*-regulatory circuit in human FL, composed of over a dozen augmented enhancers and multiple dysregulated genes, including one encoding a mitosis-associated kinase, *NEK6*. We found that only a minor subset of predicted enhancers, excluding a super-enhancer, is required to maintain *NEK6* expression. We also discovered the boundary function of a CTCF cluster in segregating the *NEK6* regulatory hub. This work emphasizes the need to rigorously validate predictions regarding enhancers, super-enhancers and *cis*-regulatory circuits assigned by computational algorithms.

1.5 References

Akhtar-Zaidi, B., Cowper-Sal-lari, R., Corradin, O., Saiakhova, A., Bartels, C.F., Balasubramanian, D., Myeroff, L., Lutterbaugh, J., Jarrar, A., Kalady, M.F., et al. (2012). Epigenomic enhancer profiling defines a signature of colon cancer. Science *336*, 736–739.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat. Rev. Mol. Cell Biol. *16*, 155–166.

Baker, M. (2011). Genomics: Genomes in three dimensions. Nature *470*, 289–294.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.

Chakraborty, T., Perlot, T., Subrahmanyam, R., Jani, A., Goff, P.H., Zhang, Y., Ivanova, I., Alt, F.W., and Sen, R. (2009). A 220-nucleotide deletion of the intronic enhancer reveals an epigenetic hierarchy in immunoglobulin heavy chain locus activation. J. Exp. Med. *206*, 1019–1027.

Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G.M., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. (2013). Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell *24*, 777–790.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell *62*, 668–680.

Dorsett, D. (2011). Cohesin: genomic insights into controlling gene transcription and development. Curr. Opin. Genet. Dev. *21*, 199–206.

Ferrai, C., de Castro, I.J., Lavitas, L., Chotalia, M., and Pombo, A. (2010). Gene Positioning. Cold Spring Harb. Perspect. Biol. *2*, a000588–a000588.

Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O.I., Verma-Gaur, J., Loguercio, S., Su, A.I., Feeney, A.J., Artyomov, M.N., et al. (2013). Unifying model for molecular determinants of the preselection Vβ repertoire. Proc. Natl. Acad. Sci. U. S. A. *110*,

E3206-15.

Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., Oltz, E.M., and Sen, R. (2011). Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. Cell *147*, 332–343.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hnisz, D., Day, D.S., and Young, R.A. (2016a). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016b). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J.M., Grosveld, F.G., Knoch, T.A., and Murre, C. (2008). The 3D Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions. Cell *133*, 265–279.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Koues, O.I., Kowalewski, R.A., Chang, L.W., Pyfrom, S.C., Schmidt, J.A., Luo, H., Sandoval, L.E., Hughes, T.B., Bednarski, J.J., Cashen, A.F., et al. (2015). Enhancer Sequence Variants and Transcription-Factor Deregulation Synergize to Construct Pathogenic Regulatory Circuits in B-Cell Lymphoma. Immunity *42*, 186–198.

Koues, O.I., Collins, P.L., Cella, M., Robinette, M.L., Porter, S.I., Pyfrom, S.C., Payton, J.E., Colonna, M., and Oltz, E.M. (2016). Distinct Gene Regulatory Pathways for Human Innate versus Adaptive Lymphoid Cells. Cell *165*, 1134–1146.

Lenz, G., and Staudt, L.M. (2010). Aggressive lymphomas. N. Engl. J. Med. *362*, 1417–1429.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Luo, H., Schmidt, J.A., Lee, Y.-S., Oltz, E.M., and Payton, J.E. (2017). Targeted epigenetic repression of a lymphoma oncogene by sequence-specific histone modifiers induces apoptosis in DLBCL. Leuk. Lymphoma *58*, 445–456.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., et al. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science *346*, 1373–1377.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Morin, R.D., Johnson, N.A., Severson, T.M., Mungall, A.J., An, J., Goya, R., Paul, J.E., Boyle, M., Woolcock, B.W., Kuchenbauer, F., et al. (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat. Genet. *42*, 181–185.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Predeus, A. V, Gopalakrishnan, S., Huang, Y., Tang, J., Feeney, A.J., Oltz, E.M., and Artyomov, M.N. (2014). Targeted chromatin profiling reveals novel enhancers in Ig H and Ig L chain Loci. J. Immunol. *192*, 1064–1070.

Rödelsperger, C., Guo, G., Kolanczyk, M., Pletschacher, A., Köhler, S., Bauer, S., Schulz, M.H., and Robinson, P.N. (2011). Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. Nucleic Acids Res. *39*, 2492–2502.

Roldán, E., Fuxa, M., Chong, W., Martinez, D., Novatchkova, M., Busslinger, M., and Skok,

J.A. (2005). Locus “decontraction” and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. Nat. Immunol. *6*, 31–41.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797–803.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. Nat. Rev. Immunol. *11*, 251–263.

Schoenfelder, S., Clay, I., and Fraser, P. (2010). The transcriptional interactome: gene expression in 3D. Curr. Opin. Genet. Dev. *20*, 127–133.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Subrahmanyam, R., and Sen, R. (2010). RAGs’ eye view of the immunoglobulin heavy chain gene locus. Semin. Immunol. *22*, 337–345.

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nat. Rev. Cancer *16*, 483–493.

Thomas, L.R., Cobb, R.M., and Oltz, E.M. (2009). Dynamic regulation of antigen receptor gene assembly. Adv. Exp. Med. Biol. *650*, 103–115.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Vora, S., Tuttle, M., Cheng, J., and Church, G. (2016). Next stop for the CRISPR revolution: RNA-guided epigenetic regulators. FEBS J. *283*, 3181–3193.

Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Yancopoulos, G.D., and Alt, F.W. (1985). Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell *40*, 271–281.

Chapter 2 : Unifying Model for Molecular Determinants of the Pre-selection Vβ Repertoire

This paper has been published in Proceedings of the National Academy of Sciences:

Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O.I., Verma-Gaur, J., Loguercio, S., Su, A.I., Feeney, A.J., Artyomov, M.N., and Oltz, E.M. (2013). Unifying model for molecular determinants of the preselection Vβ repertoire. Proc. Natl. Acad. Sci. U. S. A. 110, E3206-15. http://www.ncbi.nlm.nih.gov/pubmed/23918392

Author contributions: S.G. and E.M.O. designed research; S.G., K.M., Y.H., J.V.-G., and M.N.A. performed research; S.G., K.M., A.P., Y.H., O.I.K., J.V.-G., S.L., A.I.S., A.J.F., M.N.A., and E.M.O. analyzed data; and M.N.A. and E.M.O. wrote the paper.

2.1 Abstract

The primary antigen receptor repertoire is sculpted by the process of V(D)J recombination, which must strike a balance between diversification and favoring gene segments with specialized functions. The precise determinants of how often gene segments are chosen to

complete variable region coding exons remain elusive. We have quantified Vβ usage in the pre-selection *Tcrb* repertoire and report relative contributions of 13 distinct features that may shape their recombination efficiencies, including transcription, chromatin environment, spatial proximity to their DβJβ targets, and predicted quality of recombination signal sequences (RSSs). We show that, in contrast to functional Vβ gene segments, all pseudo-Vβ segments are sequestered in transcriptionally silent chromatin, which effectively suppresses wasteful recombination. Importantly, computational analyses provide a unifying model, revealing a minimum set of five parameters that are predictive of Vβ usage, dominated by chromatin modifications associated with transcription, but largely independent of precise spatial proximity to DβJβ clusters. This learned model-building strategy may be useful in predicting the relative contributions of epigenetic, spatial, and RSS features in shaping pre-selection V repertoires at other antigen receptor loci. Ultimately, such models may also predict how designed or naturally occurring alterations of these loci perturb the pre-selection usage of variable gene segments.

2.2 Introduction

Gene activity is regulated at multiple levels to coordinate expression during development. At a most basic level, the collection of *cis*-acting elements for a genetic locus recruits transcription factors that alter its chromatin environment to either induce or repress gene activity. Emerging studies indicate that the three-dimensional (3D) conformation of a locus also plays an important role in the regulation of its composite genes (Dekker, 2008). At most genes, many levels of control are integrated to achieve the requisite gene expression state. For example, transcriptional promoters interact with their cognate enhancers over considerable distances in the

linear genome to generate "hubs" where the two *cis*-elements are in spatial proximity (Dekker, 2008; Shih et al., 2012).

All of these regulatory strategies are employed to generate functional immunoglobulin (*Ig*) and T cell receptor (*Tcr*) genes during lymphocyte development (Bossen et al., 2012). Each antigen receptor (AgR) locus is composed of multiple variable (V), joining (J), and sometimes diversity (D) gene segments that are assembled by the process of V(D)J recombination, creating a potential variable region exon (Bassing et al., 2002). Recombination is mediated by the RAG-1/2 enzymatic complex, which is expressed in all developing lymphocytes and recognizes semi-conserved recombination signal sequences (RSSs) flanking all AgR gene segments (Schatz and Ji, 2011). Upon selection of two compatible gene segments by RAG-1/2, recombination proceeds via a DNA break/repair mechanism, ultimately fusing the two selected segments (Bassing et al., 2002; Schatz and Ji, 2011).

The assembly of AgR genes is strictly regulated despite a common collection of genomic RSS targets and expression of recombinase in all resting (G0/G1) lymphocyte precursors (Cobb et al., 2006). The most obvious level of regulation is lineage specificity. The RAG-1/2 complex assembles *Tcr* genes in precursor T cells, whereas *Ig* genes are targeted in precursor B cells. Even within an AgR locus, gene segment recombination is ordered, with D–J rearrangements preceding V–DJ. Numerous studies support a key role for chromatin accessibility in determining the recombination potential of gene segments (Feeney, 2009). The primary RAG-1/2 targets in a given cell type are transcriptionally active and DNAse hypersensitive, two hallmarks of accessible chromatin. Indeed, RAG-2 binds directly to a histone modification that accompanies transcription (tri-methylated histone H3 lysine 4, H3K4me3), providing a link between chromatin and recombinase targeting (Liu et al., 2007b; Matthews et al., 2007). At all AgR loci,

activation of (D)J clusters is dependent upon communication between at least one distal enhancer and a proximal promoter, which triggers transcription of the unrearranged (D)J segments (Oestreich et al., 2006). Recent studies indicate that the high transcriptional activity focuses RAG-1/2 binding at (D)J clusters, forming "recombination centers" into which V gene segments must be brought (Ji et al., 2010).

Although chromatin accessibility explains most aspects of RAG-1/2 deposition at recombination centers, this feature is not sufficient to ensure rearrangement of the distant V segments. Insertion of a powerful *Tcra* enhancer (Ea) into *Tcrb* maintains chromatin accessibility at nearby Vβ gene segments but does not facilitate their recombination at a stage of thymocyte development in which only *Tcra* genes rearrange (Jackson et al., 2005). Subsequent studies have shown that long-range recombination of V segments requires changes in the 3D structure of an AgR locus, bringing the V cluster into spatial proximity with (D)J recombination centers located up to 3.2 Mb away (Guo et al., 2011a; Jhunjhunwala et al., 2008; Skok et al., 2007). Long-range interactions and locus conformations are determined in large part by CTCF and cohesin, factors that bind numerous sites throughout the mammalian genome forming loops containing the intervening DNA (Rubio et al., 2008). With regard to AgR loci, deletion of CTCF, its binding sites, or essential cohesin subunits disrupt spatial interactions at *Igk*, *Igh* and *Tcra*, respectively, and perturb V to (D)J recombination (Guo et al., 2011b; Ribeiro de Almeida et al., 2011; Seitan et al., 2011; Xiang et al., 2011).

In addition to lineage-, stage-, and allele-specificity, it is also likely that the relative usage of gene segments is regulated to shape the primary repertoire of V(D)J rearrangements in precursor lymphocyte populations. During subsequent stages of lymphocyte development, V gene segment usage is an important component of positive/negative selection and, in some cases,

is a primary determinant of functional subsets within a lineage (e.g., TRVB13-2 for iNKT cells) (Godfrey et al., 2000). As such, each species may have evolved toward a unique frequency profile for V usage at each AgR locus, balancing requirements for receptor diversity, production of functional subsets, and efficacy of given V segments for antigens expressed by common pathogens. The mechanisms that sculpt pre-selection V repertoires likely incorporate a combination of the chromatin and spatial features described above. However, their relative contributions to the efficiency of long-range V to (D)J recombination at any AgR locus remain unknown.

We now address this basic question in adaptive immunity, beginning with the molecular determinants that shape Vβ usage in pre-selection thymocytes. The *Tcrb* locus is an attractive starting point for building such models because it contains a manageable set of 35 Vβ segments for molecular analysis; the *cis*-elements controlling recombination also are well-defined (Fig. 2.1A). New experimental data for chromatin profiles, spatial proximity and transcription, as well as predictions of RSS quality were incorporated into a computational analysis that weights each of these features in determining Vβ recombination frequencies. Our new data and analyses indicate that *Tcrb* adopts a 3D structure in which the relative proximity of each Vβ gene segment to DβJβ clusters is not a significant determinant in its recombination frequency. Instead, each Vβ gene segment has sufficient spatial access to the DβJβ recombination center, and usage is fine-tuned by local Vβ chromatin environments, with a particular emphasis on transcription-dependent histone modifications. Indeed, these chromatin features are absent at non-functional Vβ gene segments regardless of their RSS quality or precise proximity to DβJβ clusters. This model-building approach should help unravel the primary determinants of pre-selection V usage

at other AgR loci and in predicting how natural alterations of large V clusters may impact immune receptor repertoires.

2.3 Results

Preselection Tcrb repertoire

Recent deep sequencing studies of mRNA corresponding to VβDβJβ combinations expressed in peripheral $CD4^+$ T lymphocytes have provided an approximation of the post-selection *Tcrb* repertoire (Ndifon et al., 2012). However, our goal is to understand variables that impact the efficiency of long-range Vβ to DβJβ recombination, which shapes the pre-selection *Tcrb* repertoire. Accordingly, these analyses must be performed on primary thymocytes prior to their positive- or negative-selection, which may alter the Vβ repertoire. Preferably, a DNA-based assay should be used to quantify Vβ usage because mRNA expression of VβDβJβ rearrangements may be influenced by promoter strength or message stability. We developed the requisite assay (see below), which was applied to genomic DNA (gDNA) from sorted DN3 cells (> 95% purity; CD4-, CD8-, $CD25^{high}$, $CD44^{low}$), a developmental stage in which Vβ to DβJβ recombination occurs at a high frequency, but the vast majority of cells have yet to undergo *Tcrb*-dependent selection (Cobb et al., 2006). We reasoned that the relative frequency of rearrangements in this cell population involving a particular Vβ segment, regardless of whether the joins are productive or out-of-frame, accurately reflects its recombination potential.

Initially, we deep-sequenced products of a multiplex PCR amplification that incorporates primers for each mouse Vβ and Jβ gene segment, analogous to an approach described previously for analysis of human *Tcrb* repertoires (Robins et al., 2009). However, when applied to our DN3

thymocyte samples, a small subset of the mouse Vβ primers exhibit amplification biases in the multiplexing platform, limiting their usefulness for establishing relative Vβ frequencies. In contrast, this approach yields a relative Jβ usage similar to that observed in prior studies suggesting no significant bias in the Jβ primers (Fig. S2.1A) (Ndifon et al., 2012). In keeping with this, we noticed that the collection of VβDβJβ rearrangements for each Jβ segment has a nearly identical Vβ distribution. For example, TRBV16 is used in 8.6% of all rearrangements involving Dβ1Jβ1.1. A nearly identical percentage of Dβ1Jβ1.2 rearrangements, or any other Dβ-Jβ combination, use the TRBV16 gene segment (7.5-8.6%). The Jβ-independent frequency of Vβ usage held true for all Vβ gene segments (Figs. 2.1B, S2.1B). Moreover, recent studies have reported similar Vβ usage for rearrangements involving either Dβ1 or Dβ2 (Ndifon et al., 2012). Thus, an accurate depiction of Vβ usage can be established from a simplified approach in which levels of Vβ rearrangements to a single Jβ gene segment are measured quantitatively.

Accordingly, we designed Taqman PCR assays to independently measure rearrangements between Jβ1.1 and each of the 35 Vβ gene segments that undergo V to DJ recombination (Fig. 2.1A). We also prepared control plasmids containing each of the Vβ-Jβ1.1 combinations to serve as templates for standard curves. Initial experiments verified that all Vβ-Jβ1.1 plasmids amplified with comparable efficiencies ($\pm$5%) using Vβ-specific primers with a Jβ1.1 primer/probe combination. Control PCR assays revealed no significant cross-reactivity of Vβ-specific primers with off-target Vβ segments. Standard curves were used to quantify levels of each Vβ-Dβ1Jβ1.1 recombination product in gDNA from sorted DN3 thymocytes. The relative frequencies of Vβ usage were consistent in three biological replicates and averaged values are shown in Fig. 2.1C. Similar Vβ frequencies were observed in assays measuring a subset of Vβ-Dβ2Jβ2.1 rearrangements (Fig. 2.1D), confirming the Jβ-independence of Vβ usage. Consistent

with previous observations, analysis of gDNA from DN-depleted thymocytes revealed only a few modest differences in Vβ usage, indicating that the pre- and post-selection Vβ repertoires in mouse are largely comparable (Fig. S2.1C) (Wilson et al., 2001). In contrast, deep sequencing of 5'-RACE library from two DN3 samples yielded a distribution that differed at a subset of Vβ segments when compared with our quantitative gDNA-based assay (Fig. 2.1E). These findings suggest that mRNA levels corresponding to rearrangements involving some Vβ gene segments may not accurately reflect their recombination frequency in pre-selection thymocytes.

Overall, we observe a >10-fold range in relative Vβ usage. Only TRBV13-2 (formerly Vβ8.2) and TRBV19 (formerly Vβ6) are significantly over-represented in the primary repertoire of *Tcrb* rearrangements. The preponderance of TRBV13-2 is consistent with analyses using a restricted set of Vβ-specific antibodies from T cell populations (Wilson et al., 2001). In contrast, rearrangements were undetectable for 11 of the 35 Vβ segments. Five of these 11 "inert" gene segments are predicted to have non-functional RSSs (asterisks, Fig. 2.1C and see below), crippling their recognition by the RAG-1/2 recombinase. Six of the remaining inert gene segments have functional RSSs, but are pseudo-gene segments due to disruptions in their coding potentials (ψ, Fig. 2.1C). A lack of VβDβJβ rearrangements involving these six pseudo-gene segments flanked by functional RSSs indicates that other factors influence their recombination efficiencies (see below). Only two functional Vβs, TRBV15 and TRBV30, were under-utilized compared with the remaining 22 functional segments, which displayed only a modest variability in their usage (~3-fold range). These repertoire data suggest that *Tcrb* has evolved to normalize usage of nearly all functional Vβ segments, perhaps by modulating the three determinants of long-range recombination efficiency – RSS quality, spatial proximity, and chromatin environment.

Spatial access of Vβ gene segments to the DβJβ recombination center

Long-range recombination of V gene segments at all *Ig* and *Tcr* loci is facilitated by a contraction process, which places the V cluster into spatial proximity with distal (D)J targets located 0.1-3.2 Mb away in the linear genome (Bossen et al., 2012; Kosak et al., 2002). Deletion of transcription factors or *cis*-elements that disrupt locus contraction significantly impair V to (D)J recombination, supporting a functional link between these processes (Fuxa et al., 2004; Guo et al., 2011a; Liu et al., 2007a; Reynaud et al., 2008). Additional evidence indicates that V clusters fold into a compact rosette-like structure, which may permit extensive interactions between a recombination center and many or all of its upstream V segments (Jhunjhunwala et al., 2008). Alternatively, the spatial architecture of V clusters may sculpt the repertoire by positioning a subset of V segments closer to their (D)J targets (efficient rearrangement), while spatially excluding others (inefficient rearrangement). Indeed, emerging studies at *Igk* suggest that Vκ pseudo-gene segments may be spatially excluded from interactions with Jk substrates, perhaps minimizing their recombination potential (Lin et al., 2012).

To test whether spatial proximity is a key determinant in shaping the pre-selection *Tcrb* repertoire, we measured interaction frequencies between restriction fragments spanning each Vβ segment and fragments spanning either of the two DβJβ clusters using chromosome conformation capture (3C) (Dekker, 2008). In the linear genome, the distance between these restriction fragments range from 250-700 kb (except for TRBV31, which is ~3 kb downstream of Eβ and rearranges by inversion). 3C assays were performed on cross-linked chromatin from RAG1-deficient thymocytes, a predominantly DN3 cell population in which *Tcrb* is in an active germline conformation. The use of RAG-deficient thymocytes circumvents complications in data

analysis that arise from active *Tcrb* rearrangement. Although we cannot rule out a role for RAG-1 in defining the precise 3D conformation of *Tcrb* (Chaumeil et al., 2013), prior studies demonstrate that RAG proteins are dispensable for locus contraction (Skok et al., 2007).

We measured the cross-linking efficiency of each Vβ-containing Hind III fragment to three downstream vantage points within the *Tcrb* recombination center. Specifically, we probed Vβ cross-linking to Hind III fragments containing either of its two substrates (Dβ1 or Dβ2), or the transcriptional enhancer Eβ, which generates active chromatin over the DβJβ clusters (Oestreich et al., 2006; Spicuglia et al., 2000). Regardless of the vantage point, nearly all Vβ gene segments interact more frequently with the DβJβ recombination center in DN thymocytes when compared to $CD19^{+}$ pro-B cells purified from RAG-deficient bone marrow (Figs. 2.2A, S2.2A, S2.2B). These data verify and extend previous analyses showing that *Tcrb* adopts a T cell-specific conformation, juxtaposing the Vβ cluster with its DβJβ targets (Skok et al., 2007).

Of particular note, interaction levels measured from a given vantage point (e.g., Dβ1) display significant differences across the collection of Vβ segments (Fig. 2.2A). There were also differences in interactions between specific Vβ segments and two vantage points. For example, the fragments spanning TRBV1 or TRBV18/19 both interact with Dβ1 at a much higher frequency than with Dβ2 (Figs. 2.2A, S2.2A). Conversely, TRBV17 displays a greater interaction with Dβ2 (Figs. 2.2A, S2.2A). Despite these differences, the TRBV1 and TRBV19 segments are utilized with indistinguishable frequencies in recombination products involving either Dβ1 or Dβ2 (Fig. 2.1D). In contrast to preliminary findings at *Igk* (Lin et al., 2012), a group of pseudo-gene segments spanning TRBV6-TRBV11 each interact with DβJβ clusters at a relatively high frequency, but these gene segments are absent from the pre-selection *Tcrb*

repertoire despite having functional RSSs. These findings suggest that relative Vβ usage in the pre-selection *Tcrb* repertoire cannot be fully explained by differences in their spatial proximity to the DβJβ regions.

To more rigorously investigate the relationship between spatial proximity and long-range recombination, we performed Spearman ranking correlations for 3C and Vβ repertoire data. Because the absolute values of 3C data cannot be quantitatively compared between the three assays, we first ranked cross-linking efficiencies of the Vβ segments within each vantage point (Supplementary Table 2.1). No significant correlations between 3C ranking and TRBV rearrangement are observed for any of the three individual viewpoints within the DβJβ recombination center. We also calculated the average ranking for each Vβ segment over the three assays (Dβ1, Dβ2 and Eβ) and compared these values with relative usage in VβDβJβ joins (Supplementary Table 2.1). As shown in Fig. 2.2B, there is an absence of significant correlation between Vβ usage and its average rank for interactions with the DβJβ recombination center. Consistent with this finding, we also observe no obvious correlation between the recombination frequency of a Vβ segment and its proximity to CTCF binding. We conclude that, although gross locus contraction is important to bring the entire Vβ cluster into spatial proximity with its Dβ substrates, the precise magnitude of each Vβ–Dβ interaction is not a primary determinant of recombination efficiency. Instead, our 3C and repertoire data indicate that once *Tcrb* is contracted in DN thymocytes, the large Vβ cluster adopts a conformation in which spatial access of Vβ segments to the recombination center is not limiting.

Role of RSS quality in determining Vβ use

Despite general conservation of the heptamer-spacer-nonamer configuration, RAG-1/2 substrates exhibit substantial variation compared with the consensus RSS sequence: (CACAGTG) – 12 or 23 bp spacer – (ACAAAAACC) (Hesse et al., 1989; Livak, 2003). In vivo replacement or natural variants of RSSs can alter the usage of gene segments, including those within the *Tcrb* recombination center (Nadel et al., 1998; Posnett et al., 1994; Wu et al., 2003). In vitro studies using plasmid substrates have defined the effects of positional substitutions within the consensus RSS on recombination efficiency (Feeney et al., 2000; Hesse et al., 1989; Jung et al., 2003). Thus, one component of non-random Vβ usage is likely the quality its flanking RSS.

To examine this possibility, we took advantage of an algorithm (http://www.itb.cnr.it/rss/) that predicts the RSS quality of any given sequence (Cowell et al., 2003). In brief, this algorithm calculates the theoretical recombination potential of an RSS using a statistical model that assigns a score based on the contribution of each nucleotide within the heptamer-spacer-nonamer sequence. The algorithm output is a "Recombination signal Information Content" (RIC) score, which predicts the quality of an input RSS with a reasonable degree of accuracy based on data from plasmid recombination substrates (Lee et al., 2003). For *Tcrb*, six of the 35 Vβ gene segments are flanked by non-functional RSSs with a RIC score of < -58.5, the threshold defined by Cowell et. al, (Cowell et al., 2003) (TRBV8, 12-3, 18, 21, 27, and 28). The remaining 29 Vβ segments have a substantial range in predicted RSS quality, with RIC scores between -29 (TRBV4) and -58.2 (TRBV11). Recombination is undetectable for five of the six Vβ segments flanked by RSSs that score below the functional threshold (Fig. 2.1C). The exception is TRBV21, which rearranges at a detectable level, but is predicted to have a

marginally non-functional RSS (RIC score -58.6) consisting of a consensus heptamer and a 22 bp rather than 23 bp spacer.

The correlation between RIC scores and Vβ usage is shown in Fig. 2.3. Although a positive correlation is apparent, the magnitude of Vβ usage diverges significantly from linearity when compared with predicted RSS quality. In general, Vβ RSSs with lower quality (RIC scores -45 to -58) are either inert or rearrange at a level below the average frequency. RSSs with RIC scores >-45 exhibit a broad range of Vβ recombination frequencies, as highlighted by the following examples: (1) TRBV13-2 is the most frequently used segment but shares a nearly identical RIC score with TRBV14, which rearranges at an average frequency (2) Six Vβ segments (TRBV7, 15, 16, 20, 24 and 26) have nearly indistinguishable RIC scores (-41 to -42), but one Vβ is recombinationally inert (TRBV7) and the remaining five display an eight-fold range in their utilization. We cannot rule out the possible contribution of coding sequences adjacent to each RSS in altering its quality as a RAG-1/2 substrate. Inspection of coding flanks revealed only a small subset with features predicted to attenuate RAG cleavage (e.g., AT or pyrimidine stretches for TRBV12-1, 12-2, 14, 17, and 29) (Cuomo et al., 1996; Gerstein and Lieber, 1993; Olaru et al., 2003; Yu and Lieber, 1999). However, as shown below, the recombination frequency of these gene segments correlate best with features of associated chromatin. Together, our data indicate that, although predicted RSS qualities contribute to the formation of a pre-selection *Tcrb* repertoire, other levels of control clearly impact Vβ usage.

Role of chromatin environment in determining Vβ recombination potential

Chromatin accessibility at gene segments has been studied extensively as a determinant of the tissue- and stage-specific mechanisms controlling V(D)J recombination (Cobb et al., 2006;

Feeney, 2009). Germline transcription of gene segments leads to the deposition of H3K4me3, a histone modification that is recognized by RAG2 and augments endonuclease function of the RAG complex (Liu et al., 2007b; Matthews et al., 2007; Shimazaki et al., 2009). As such, levels of chromatin accessibility and transcription at each Vβ segment may help determine its usage in the pre-selection *Tcrb* repertoire.

The emerging approach of "chromatin profiling" uses combinatorial patterns of histone modifications, nucleosome density, and factor binding to assess the epigenetic status of genomic regions (Ernst et al., 2011). To compare epigenetic landscapes at the 35 Vβ segments, we generated new chromatin profiling data from RAG1-deficient thymocytes using chromatin immunoprecipation (ChIP) assays in combination with *Tcrb* microarrays (ChIP-chip) or deep sequencing. We also performed Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), which identifies nucleosome-depleted regions in the genome (Giresi and Lieb, 2009). The new ChIP-chip (P300, H3K27ac, H3K4me2), ChIP-seq (H3ac, H3K4me3, and CTCF), and FAIRE-Chip data from RAG-deficient thymocytes were combined with epigenomic data available in public repositories (H3K4me1, RNA Pol II and H3K9me2) from RAG-deficient thymocytes (Pekowska et al., 2011). We employed a published methodology to integrate cross-platform data derived from ChIP-chip and Chip-seq (Chen et al., 2011). In addition to nucleosome depletion (FAIRE), the analyzed features characterize active promoter regions (transcription, RNA Pol II, H3K4me3, and H3ac), active regulatory elements (H3K4me1, H3K27ac, and P300), poised chromatin (H3K4me2), insulators (CTCF), and silent chromatin (H3K9me2).

Relative intensities for each feature at the 35 Vβ segments ($\pm$ 1 kb) are represented as a heat map in Fig. 2.4A. Examples of several features for selected gene segments in chromatin

environments ranging from highly active to silent are depicted in Fig. 2.4B. Overall, most of the Vβ segments that participate in Vβ to DβJβ recombination exhibit higher levels of active chromatin features than the inert Vβ elements (H3K4me, RNA Pol II/transcription, and histone acetylation). In contrast, the repressive H3K9me2 modification was enriched over many of the inert Vβ segments. One region within the Vβ cluster containing the TRBV12-2 and 13-2 gene segments is conspicuously active (Fig. 2.4B), with high levels of germline transcripts and other features associated with open chromatin, including one of the few discernible P300 peaks. As noted above, TRBV13-2 is also the most frequently rearranged gene segment in DN3 thymocytes, suggesting a dominant correlation between open chromatin and long-range recombination efficiency. Consistent with this possibility, many of the pseudo-gene segments, even those containing functional RSSs, are expressed at a low level and are associated with chromatin that lacks activating histone marks (asterisks, Fig. 2.4A). *In-silico* analysis of Vβ upstream sequences (-1 kb to leader) for predicted transcription factor binding profiles (TRASFAC/JASPAR databases) revealed no distinguishable differences between functional and pseudo-Vβ gene segments. Promoter activity as measured by luciferase assays in a transfected pre-T cell line show that all tested upstream Vβ regions from recombinationally active gene segments (11/11) are functional promoters. In contrast, only some of the tested regions upstream of pseudo-gene segments (4/8) exhibit promoter activity (ψ, Fig. S2.3), indicating no clear correlation between Vβ utilization and promoter strength. Thus, it appears that the mouse Vβ cluster has evolved multiple strategies to silence chromatin at non-functional gene segments.

A reasonable concordance was observed between chromatin environments and recombination efficiencies when comparing Vβ segments with equivalent RIC scores. For example, TRBV15 and TRBV16 are predicted to have RSSs of nearly identical qualities but

reside in distinct chromatin environments. The elevated levels of transcription and activating histone marks at TRBV16 correspond to an elevated level of recombination (Fig. 2.4C). In some cases, both the predicted RSS quality and chromatin environment apparently contribute to Vβ usage. For example, TRBV23 and TRBV24 are both transcriptionally active and have comparable chromatin features (see heatmap, Fig 2.4A); however, the lower predicted RSS quality for TRBV23 (-48.6) when compared to TRBV24 (-41.2) correlates with an attenuated level of recombination. We also noted that contributions of chromatin to rearrangement frequencies may derive from different combinations of features. TRBV20 and TRBV26 exhibit nearly identical usage (2.7 and 2.9%) and RIC scores (-41.5 and -41.1), but patterns of specific chromatin features at these gene segments differ significantly (see heat map, Fig. 2.4A). To further validate these comparisons, we performed semi-quantitative assays to measure the qualities of eight Vβ-RSSs using plasmid-based substrates (including the six Vβ-RSSs mentioned above). The relative qualities of these RSSs, tested in conjunction with a natural target (5'Dβ1-RSS), are in line with predictions from RIC scores (Fig. 2.4D and legend), further supporting our conclusions. Together, these profiling studies indicate a strong contribution of chromatin environment to Vβ recombination frequencies, but also suggest that individual parameters of chromatin accessibility may affect substrate usage in a weighted manner.

Computational analysis of Vβ use determinants

Our data indicate that predicted RSS qualities and chromatin landscapes likely contribute in a combinatorial manner to the efficiency of long-range *Tcrb* assembly. To examine these combinatorial relationships, we employed classification and regression analyses comparing chromatin features and predicted RSS quality with Vβ usage. These analyses were guided by

recent computational strategies devised to predict gene expression levels based on patterns of histone modifications (Dong et al., 2012; Karlic et al., 2010). We applied one validated approach (Dong et al., 2012) to study whether chromatin features, predicted RSS quality, and spatial proximity are predictive of the observed Vβ repertoire.

The chosen computational approach takes into account: (i) the signal intensity of each chromatin feature, (ii) levels of germline transcription, (iii) RIC scores, and (iv) spatial proximity based on the average 3C rank-score. With regard to chromatin features, distinct positional profiles are observed for various histone marks. For example, H3K4me3 is enriched over active promoters and progressively wanes along gene bodies. Accordingly, we divided the regions spanning each Vβ segment into three bins – the Vβ segment itself (leader to RSS), its upstream promoter region (1 kb 5' of leader), and its downstream region (1 kb 3', including the RSS). For each feature, we computed Pearson correlation coefficients for the three bins versus Vβ recombination frequencies (Fig. 2.5). We find the best correlation for a majority of histone modifications in the upstream/promoter bin (H3K4me1, H3K4me2, H3K4me3, P300, H3ac, and H3K27ac). In contrast, repression by H3K9me2 was most correlative in the bin that contains Vβ segments. FAIRE and RNA Pol II signals have very similar predictive abilities over both the Vβ and its downstream bins. These findings are strikingly similar to correlations observed between chromatin features and gene expression (Dong et al., 2012; Karlic et al., 2010), further underscoring the relationship between transcriptional activity and Vβ recombination frequencies. A particularly satisfying outcome of this analysis is the correlation between FAIRE signals and the bins flanking RSSs, presumably reflecting a requirement for nucleosome depletion at RAG-1/2 targets (Kwon et al., 2000; Osipovich et al., 2007).

Next, we identified features that are most predictive of whether a Vβ segment will rearrange at any frequency or will remain inert. For this and the remaining analyses, we used signal intensities only from bins exhibiting the highest correlation between each chromatin mark and Vβ usage (Fig. 2.5, asterisks). A computational approach called Random Forest was employed (Dong et al., 2012), which randomly tests combinations of binned features for their predictive abilities to classify gene segments as "active" or "inert" (Fig. 2.6A). This analysis revealed that three features – predicted RSS quality, FAIRE and RNA Pol II signals – are sufficient to classify the recombination potential of a given Vβ segment with a high level of confidence. The classifications are also evident from linear regression analysis on these three features relative to Vβ recombination frequencies (Fig. 2.6B, 30/35 segments predicted correctly). When we used the Random Forest algorithm, but focused only on values for RIC score, FAIRE and RNA Pol II signals, 32/35 Vβ segments classified correctly as active versus inert (see methods). The three exceptions common to both Random Forest- and linear regression-based classifications are TRBV15, 21, and 22; segments predicted to be inert but exhibiting detectable levels of recombination. These outliers could reflect partial compensation by chromatin features other than the factors determined by our algorithms. Notwithstanding, the most important predictive features of recombinational competency are linked mechanistically to RAG substrate quality (RIC score), substrate accessibility (nucleosome depletion), and RNA Pol II association.

We next moved beyond "black-and-white" classifications to analyze the relative importance of Vβ features in fine-tuning recombination frequencies of the 23 active gene segments. For this purpose, we performed linear regression on the selected bins for each feature versus frequency values. As shown in Fig. 2.6C, the features that correlate most significantly

with Vβ usage are H3K4 methylation, H3Ac, and RNA Pol II occupancy, which normally associate with transcriptionally active regions. The repressive H3K9me2 mark correlates negatively with levels of Vβ recombination. In contrast to its dominant role as a determinant for recombinational competence, RIC scores for the 23 active Vβ gene segments correlate poorly with their relative levels of rearrangement. A similar discordance between recombination frequencies and RSS qualities for a limited set of mouse VH and Vκ gene segments has been described previously (Aoki-Ota et al., 2012; Williams et al., 2001). These findings suggest that chromatin environment, rather than predicted RSS quality, is the dominant feature for fine-tuning Vβ usage in long-range recombination.

We next investigated whether various combinations of the 13 features included in this study are predictive of Vβ recombination efficiencies. As a starting point, we examined the predictive capacity of all 13 features using linear regression (Fig. S2.4A). This analysis yielded a correlation coefficient for best fit of 0.78, which was statistically insignificant (*P-value* > 0.05). We next tested whether a subset of these 13 features correlate in a significant manner with observed frequencies of Vβ usage. For this purpose, we examined various subsets of features, ranging from a single feature to 12 of the 13 variables in all possible combinations. This combinatorial analysis yielded a set of five features that correlate significantly with Vβ usage (Fig. 2.6D, Pearson correlation coefficient = 0.69, *P-value* = 0.03). In descending order of contribution to the fitted model, the identified features were H3K4me3, H3K4me2, transcription, P300, and CTCF. The first four features largely determine the efficiency for most TCRBV segments, while the remaining feature, CTCF proximity, improves the fit for several outliers that are poorly predicted by H3K4me3, H3K4me2, transcription, and P300. When further analyzed by clustering, we found that the four chromatin features (H3K4me3, H3K4me2, P300, and

CTCF) in this set of five core parameters represent four classes of related marks that share a significant portion of epigenetic information (Fig. 2.6E). For example, H3K4me3 correlates strongly with H3ac and RNA Pol II occupancy, three features enriched near active promoters, in essence encapsulating the information content of the entire class. The relative contributions of the five core features to the accuracy of fit and the corresponding linear regression formula are provided in Fig. S2.4B.

Together, the computational analyses derive a two-tiered model for predicting Vβ usage in the pre-selection *Tcrb* repertoire. First, RIC scores in combination with nucleosome and RNA Pol II densities discriminate active from inert substrates. The recombination frequency of the active Vβ set can be discerned from values for the five core parameters identified by statistical correlations. Moreover, this basal set of five parameters may be useful in future studies to predict the impact on pre-selection Vβ repertoires of naturally occurring or engineered perturbations at *Tcrb*.

2.4 Discussion

We took an integrative approach to define the molecular determinants of Vβ recombination frequencies, an important component of the pre-selection *Tcrb* repertoire. Prior studies have examined the independent effects of RSS quality, 3D architecture, transcription, or chromatin accessibility on recombination of specified gene segments. However, to our knowledge, this is the first unified analysis of how these features impact the efficiency of long-range V to (D)J recombination at an endogenous AgR locus. Using several independent computational approaches, we find that: (i) RSS quality and nucleosome density are the major determinants of whether a given Vβ segment will participate in *Tcrb* gene assembly, (ii) the

relative usage of a Vβ segment is fine-tuned by its chromatin environment, (iii) the optimal epigenetic landscape for Vβ recombination is a blend of transcriptional activation marks, nucleosome depletion, and a lack of the repressive H3K9me2 mark, and (iv) the precise magnitude of spatial proximity between a Vβ segment and the DβJβ recombination center does not significantly influence its relative utilization. Collectively, we find that a minimum set of five features can be measured to predict the recombination frequency of a competent Vβ segment with a high degree of accuracy.

A critical component of our study was a determination of the pre-selection Vβ repertoire. The relative usage of Vβ segments may have important consequences with regard to AgR-mediated thymic selection, the production of functional T cell subsets that employ specific Vβ segments, or the baseline antigenic profile recognized by emerging T lymphocytes. We used a DNA-based approach to directly quantify rearrangement levels of the 35 Vβ segments in sorted DN3 thymocytes. This approach avoids two caveats of prior repertoire analyses – biases introduced by thymocyte selection or by mRNA expression differences, both of which were observed in our companion assays. We find that only a few functional Vβ segments are either over- or under-utilized in the pre-selection *Tcrb* repertoire. One of the over-utilized Vβ segments, TRBV13-2 (formerly Vβ8.2), is enriched in iNKT cells, a subset of lymphocytes that respond to lipid antigens and produce a robust cytokine response. We postulate that the ideal chromatin environment encompassing TRBV13-2 has evolved to augment its rearrangement efficiency, ensuring a sufficient production of iNKT cells, which provide a rapid cellular immune response to numerous foreign antigens. Notwithstanding, rearrangement levels for the vast majority of functional Vβ gene segments (18/22) fall within a three-fold range. The

relatively limited range of distribution likely reflects a requirement to maximize *Tcrb* diversity prior to its pairing with *Tcra* for subsequent selection by MHC-peptide complexes.

As shown here, the "normalization" of Vβ usage results predominantly from the chromatin environment encompassing each gene segment, with perhaps a minor contribution from its RSS quality. The dominance of chromatin in fine-tuning Vβ usage was evident from several "outlier" gene segments. The TRBV15 and TRBV30 segments are under-utilized compared with all of the other functional Vβ elements, likely because they are poorly transcribed or lack most features of active chromatin. Likewise, nearly all of the pseudogene segments that are flanked by functional RSSs reside in a repressive chromatin environment. For the latter category, we provide evidence that some, but not all, germline promoters associated with pseudo-Vβ segments have been incapacitated, despite their retention of potential factor binding sites found in functional Vβ promoters. Another potential mechanism for pseudogene suppression could be their localization to the nuclear periphery or lamina (Reddy et al., 2008). However, the precise underlying mechanisms that sequester these pseudogene segments in repressive chromatin, preventing wasteful recombination, remain to be defined.

With regard to the collection of rearranging Vβ segments, the dominant chromatin features in determining their relative usage are associated with active transcription. The strongest correlations exist between recombination efficiencies, histone acetylation (H3ac), H3K4 methylation, nucleosome depletion, and RNA Pol II occupancy. Although a link between this transcriptional epigenetic state and recombination has long been appreciated, its dominant role in sculpting the primary repertoire of antigen receptors is a novel finding of our study. One likely mechanism for this relationship is the affinity of RAG complexes for chromatin bearing the H3K4me3 mark. Prior ChIP-seq studies demonstrate that RAG-1/2 is bound to the DβJβ

recombination center in DN thymocytes but is relatively absent from the Vβ cluster (Ji et al., 2010). This reflects the extremely high levels of H3K4me3 on DβJβ chromatin compared with Vβ segments (~10-fold difference) (Ji et al., 2010). Based on our integrative model, we suggest that after *Tcrb* contracts, pre-bound RAG-1/2 complexes at the DβJβ recombination center may preferentially target Vβ segments that are most enriched for transcription-associated marks, including H3K4me3. Thus, the strength of each Vβ promoter within its native chromosomal context may be a dominant feature for shaping the pre-selection *Tcrb* repertoire.

One important aspect of our study is that the precise magnitude of association between a Vβ segment and DβJβ clusters, as measured by 3C, does not contribute discernibly to its level of usage. Clearly, general locus contraction is an important mechanism for bringing V segments into spatial proximity with their distant (D)J substrates (Bossen et al., 2012). However, the spatial architecture adopted by the large Vβ cluster in DN thymocytes must provide sufficient access to all of its composite gene segments by RAG-1/2 bound at the DβJβ recombination center. Recent studies of *Igk* suggest that most V segments within this locus also may have similar spatial access to their target J segments (Lin et al., 2012). Given the 10-fold range in cross-linking efficiencies between various Vβ segments and the two DβJβ clusters, we conclude that spatial constraints on long-range Vβ to DβJβ recombination are binary rather than digital, requiring only that target gene segments cross a threshold of spatial proximity. Presumably, this spatial threshold is surpassed via a combination of locus contraction and folding of the Vβ cluster into a more compact structure.

In conclusion, a combination of epigenetic, spatial, transcriptional, and RSS features were used to identify the dominant determinants for sculpting the pre-selection Vβ repertoire. We concede that a model for Vβ usage may not completely apply to all other AgR loci. Indeed,

pseudo-Vκ segments interact inefficiently with their target Jκ cluster, perhaps suppressing their recombination (Lin et al., 2012). In contrast, pseudo- and functional Vβ segments interact indistinguishably with their DβJβ substrates. Recombination of pseudo-Vβ segments is, instead, suppressed by sequestration into inactive chromatin. This distinction may reflect a more dominant role for spatial constraints at the much larger *Igk* locus. Notwithstanding, much of the relevant epigenetic and RSS quality data necessary to build predictive models for other AgR loci are available publicly. In most cases, the lacking features are reliable DNA-based analysis of V usage and complete sets of 3C data covering V clusters. We suspect that as multiplex PCR approaches improve, eliminating primer bias, comprehensive pre-selection repertoires for all AgR loci will emerge. Current methods for quantifying spatial proximity on a global scale lack the resolution of focused 3C assays; however, technical improvements and increased sequencing depths may soon overcome these obstacles. The learned model-building strategy employed here should be a valuable guide for defining relative contributions of epigenetic, spatial, and RSS features in shaping pre-selection V repertoires. Ultimately, these models should also be valuable for predicting how designed or naturally occurring alterations of AgR loci perturb the pre-selection V repertoire. These alterations could range from targeted RSS and promoter substitutions to natural variant AgR alleles that lack portions of the large V clusters, creating "holes" in the immune repertoire. Indeed, a striking parallel exists between the usage of several mouse and human Vβ orthologues (Livak, 2003), underscoring the potential utility of our model to predict the effects of human *TCRB* polymorphisms on primary repertoire formation.

2.5 Materials and Methods

Cell Purification and Antibodies. Thymocytes from C57BL/6 mice (4-6 weeks) were depleted of $CD4^+$ and $CD8^+$ cells using MACS (Miltenyi Biotec, CA). The remaining DN cells were stained and sorted for the $CD25^{hi}/CD44^{low}$ DN3 population, yielding a >95% purity. $CD19^+$ bone marrow cells from RAG-deficient mice were purified using MACS in conjunction with CD19 microbeads (Miltenyi Biotec, CA), providing a >90% pure population of pro-B cells. CD4-FITC (561835), CD8-FITC (553031), CD4-biotin (553044), CD8a-biotin (553028), CD44-PE (553134), CD25-APC (557192) antibodies were purchased from BD Biosciences (CA) and used for cell staining and sorting. H3K4me2 (07-030, Millipore), H3K27ac (ab4729), and P300 (C-20) (SC-585X) for H3ac (06-599, Millipore), H3K4me3 (39159, Active Motif) and CTCF (07-729, Millipore) antibodies were purchased and used for ChIP experiments.

High throughput sequencing of *Tcrb* rearrangement. gDNA from sorted DN3 cells was amplified by multiplex PCR for Vβ-Dβ-Jβ rearrangements and the amplicons were deep sequenced by Adaptive Biotechnologies (WA). The gene segment usage were analyzed using ImmunoSEQ™ Analyzer software.

5' RACE. Total RNA (0.5 μg) from DN3 thymocytes was converted to cDNA and 5' RACE was performed using a Cb primer (5'-AGCTCCACGTGGTCAGGGAAGAA-3') following manufacturer's protocol (Ambion, CA). The RACE product were blunted, concatemerized, and sonicated to an average size of 175 bp. The sheared fragments were ligated with Illumina adapters and sequenced using an Illumina HiSeq-2000 to provide paired-end reads extending 101 bases. Raw reads were de-multiplexed and unique FASTA reads obtained using the FASTX tool kit (http://hannonlab.cshl.edu/fastx_toolkit). For quality control, a portion of the 5' RACE

product was TOPO cloned and individual clones were sequenced. Sequences were analyzed using IMGT High-V quest (http://www.imgt.org) (Lefranc et al., 2009).

Quantitative PCR for VβDβJβ Rearrangements. We designed a panel of Taqman PCR assays using probes and primers specific for either Jβ1.1 or Jβ2.1 gene segments in combination with a primer specific for each of the 35 Vβ segments. We also generated a collection of plasmids containing each Vβ cloned directly upstream of either Jβ1.1 or Jβ2.1 in an orientation that mimics the corresponding V-D-J rearrangement product. For this purpose, Jβ1.1 or Jβ2.1 segments were amplified by PCR from mouse gDNA and cloned into the NotI/BamHI sites of pBS-KSII. Subsequently, Vβ segments were amplified and cloned upstream of the Jβ region. The specificity of Vβ primers was confirmed by BLAST searches and a panel of PCR assays showing that amplification of control plasmids containing other Vβ segments was detected at <1% compared with the bona fide target. Template plasmids were used to generate standard curves, allowing us to correct for minor differences in PCR efficiency between each of the assays. Total Vβ-DbJb1.1 or Vβ-DbJb2.1 rearrangement product (alleles) was quantified relative to amounts of an unrearranged region within the genome (b2-microglobulin) using the formula $E^{-Ct(V\text{-}J\beta)}/E^{-Ct(B2M)}$, where E is the primer efficiency. The list of primers and probes used is given in Supplementary Table 2.2.

Chromosome Conformation Capture. 3C assays were performed on 10^7 RAG1-deficient C57BL/6 DN thymocytes or $CD19^+$ pro-B cells using Hind III as described in Hagège *et. al.* (Hagege et al., 2007). Primers and probes designed for Hind III fragments corresponding to each vantage point in the recombination center (Dβ1, Dβ2 and Eβ) were used in Taqman assays with primers specific for each Vβ gene containing fragment. Standard curves were generated for these

Taqman assays using Hind III-digested BACs spanning the entire *Tcrb* locus, which were then ligated to yield a library of all possible products. Interaction between nearest neighbor fragments in ERCC3 gene was set as 1. Cross-linking frequencies were calculated as described in Hagège *et. al.* (Hagege et al., 2007). List of primers, probe sequences and BAC clones are provided in Supplementary Table 2.3.

Chromatin Immunoprecipitation and FAIRE. Chromatin immunoprecipitation (ChIP) experiments for H3K4me2, H3K27ac, and P300 were performed with chromatin from RAG-deficient thymocytes (C57BL/6) as described previously (Degner et al., 2011). The ChIP DNA was purified using a Qiagen DNA purification kit and subjected to whole genome amplification (Sigma, MO), labeled, and hybridized to custom Nimblegen microarrays according to the manufacturer's protocol by Mogene Inc, St. Louis. Total input DNA was used as the hybridization control. A subset of ChIP-Chip data were verified at various locations throughout *Tcrb* using qPCR (not shown). FAIRE was performed on cross-linked nuclei from RAG-deficient DN thymocytes and purified pro-B cells using published methods (Giresi and Lieb, 2009). Purified FAIRE-DNA was used for subsequent analyses by q-PCRs or array hybridization. DNA from non-crosslinked cells, processed in parallel, was used as reference samples. Model-based Analysis of 2-Color Arrays (MA2C, version 1.4.1) was used to normalize the microarray data, detect peaks and generate UCSC wiggle (WIG) files.

ChIP-seq experiments were performed as above using chromatin from RAG-deficient thymocytes (C57BL/6) for H3ac, H3K4me3 and CTCF. ChIP seq data for RNA Pol II, H3K4Me1 and ChIP-Chip data for H3K9me2 from RAG-deficient thymocytes were downloaded from http://www.comline.fr/ciml/ (Pekowska et al., 2011). The ChIP-Seq raw data were aligned to the mouse reference genome (mm9) using Bowtie 0.12.8. The resulting BAM files were used

to generate UCSC wiggle (WIG) files and peaks called using Model-based Analysis of ChIP-Seq software (MACS, version 1.4.2). The list of antibodies used in ChIP experiments is given in supplementary methods.

RNA-seq. Total RNA from RAG-deficient DN thymocytes was extracted using an Ambion (CA) Ribopure kit. Ribosomal RNA was removed using Ribo-ZERO (EpiCentre, CA). mRNA was fragmented and reverse-transcribed to yield double-stranded cDNA, which was sequenced on an Illumina HiSeq-2000 using paired-end reads extending 101 bp. Raw data were de-multiplexed and aligned to the mouse reference genome (mm9) using TopHat 1.4.1. Transcript abundances were estimated from the alignment files using Cufflinks.

Luciferase Assays. The Eβ enhancer was amplified and cloned into the Bam HI site of pGL3 (Promega, WI). Each tested upstream Vβ region (300-500 bp) was amplified and cloned into the Xho I/Hind III sites of the Eβ-containing vector. T3 cells (Ferrier et al., 1990) were transfected transiently with firefly (4μg) and Renilla (40ng) luciferase plasmids using electroporation. After 24 hours, the transfected cells were assayed for firefly and Renilla activities. List of primers are provided in Supplementary Table 2.4.

V(D)J Recombination Substrates. A Dβ1-Jβ1.1 rearrangement that includes the 5′ Dβ1-RSS was amplified from thymus DNA and cloned into pCDNA3.1. Each recombination substrate includes the specified Vβ-RSS together with its upstream and downstream flanking sequences (80 and 130 bp, respectively), which were cloned 5' to the DβJβ1.1 join (deletion substrates). An inert YFP coding sequence was inserted as a stuffer between the Vβ and Dβ1-Jβ1.1 elements. A list of Vβ-specific primers is provided in Supplementary Table 2.5.

Recombination Substrate Assays. Human embryonic kidney 293T cells were transfected with an equimolar mixture of eight recombination substrates (TRBV1, 15, 16, 18, 20, 23, 24, 26), pEBB-RAG1, and pEBB-RAG2, using Trans-IT 293 (Mirus) (40). Plasmid substrates were recovered 48 hours post-transfection and digested with Not I to minimize unrearranged PCR products and Dpn I to cut untransfected substrates (40). The digested DNA mixture was amplified with primers that are common to all substrates -- one that recognizes plasmid sequence upstream of the Vβs and one specific for Jβ1.1, (dsT7-CAAGCTGGCTAGCGTTTAAAC and J1.1TR-CTCGAATATGGACACGGAGGACATGC). PCR was performed for 30 cycles on serial 4-fold dilutions of recovered substrates. The products were separated on 1% agarose gels, transferred to Zetaprobe (BioRad), and probed with labeled Vβ-specific oligonucleotides.

Computational Analysis. Regression analysis was performed following a two-step procedure that is a simplified version of the protocol described previously (Dong et al., 2012). **Step 1**. For each of the chromatin features analyzed, the region spanning Vβ segments was divided into three bins - *j*-th V_j segment itself, 1 kb immediately upstream (U_j), and 1 kb immediately downstream of the V segment (D_j). The signal intensity of each bin (3 bins x 35 Vβs, 105 total bins) was measured from the UCSC wig files containing either read counts (ChIP-Seq) or MA2C scores (ChIP-chip) using BEDtools. The signal intensities were then converted to the natural logarithm of their values. To eliminate any ln(0) values in the computational analyses, a pseudo-count of one was added to the read counts. Pearson's correlation coefficients were then used to define which of the three bins (V_j, U_j, D_j) correlate best with V recombination frequencies. The bin for each feature with the highest correlation coefficient was used in further analyses. Recombination frequencies f_j for V_j regions (expressed in % of overall usage) were transformed into their natural logarithm values ($\ln(f_j+0.01)$, where 0.01 is an added pseudo-count). The Vβ gene segments

were then classified as “rearranging” or “non-rearranging”, and Random Forest classification was used to determine which of the features distinguish best between rearranging and inert Vβ gene segments (R package, RandomForest). **Step 2.** Linear regression analysis was performed for thirteen variables using data corresponding to only the subset of 23 rearranging Vβ segments (non-zero recombination frequency) using R package (leaps) to identify the most important regressors for recombination levels. The analysis was further refined to determine a reduced set of variables that attains statistical significance.

2.6 Figures

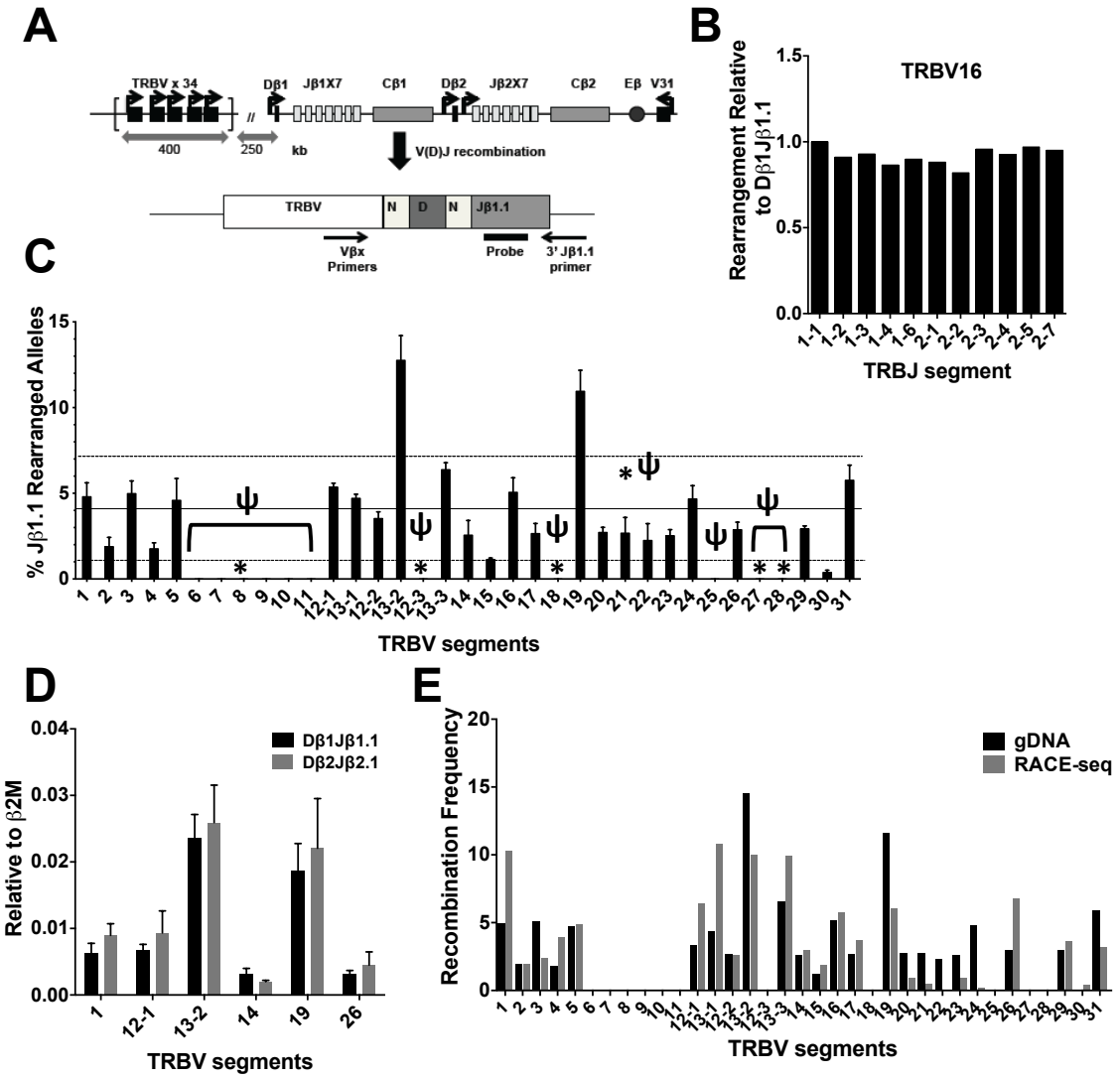


Figure 2.1: Preselection *Tcrb* V repertoire

(A) Schematic representation of the murine *Tcrb* locus (top) and Taqman assay (bottom) used to quantify VβDβ1Jβ1.1 recombination products. Bold arrows near gene segments denote promoters (top). "N" stands for N-regions (non-templated regions of diversification), locations of primers and probes for Taqman assays are shown (bottom). (B) Distribution of V(D)J rearrangements from high throughput sequencing involving select Vβ segments and each of the 11 functional Jβ segments. The distribution for a given Vβ-Jβ combination is calculated as the number of unique reads for that combination divided by the total number of unique reads for the corresponding Jβ element (Jβ1.1: 8.6% for TRBV1, 4.1% for TRBV12-2, 7.6% for V13-1 and 8.6% for TRBV16). Data are represented relative to the distribution of Vβ-Jβ1.1, where percent total Vβ-Jβ1.1 is set to a value of 1. (C) Pre-selection Vβ repertoire. Taqman real-time PCR quantification of VβDβ1Jβ1.1 rearrangements was performed on gDNA from DN3 thymocytes. Signals from each assay were normalized to values obtained from an assay for the invariant β2M gene. Average levels from three independent DN3 preparations are shown (n = 3, $\pm$ SEM). Recombination frequencies are shown as the percent contribution of a given Vβ segment to the total level of Jβ1.1 rearrangement. Pseudo-genes are denoted by ψ and gene segments with non-functional RSSs are marked with an

asterisk. The average Vβ usage and standard deviation are denoted by dotted black lines. (D) Taqman real-time PCR assays measuring VβDβ1Jβ1.1 versus VβDβ2Jβ2.1 rearrangements in DN3 thymocytes were quantified as described in C. (E) Comparison of Vβ usage values in DN3 thymocytes using gDNA- versus mRNA-based methods Average values from gDNA assay (n=3) and RNA-5'RACE seq (n = 2) are shown.

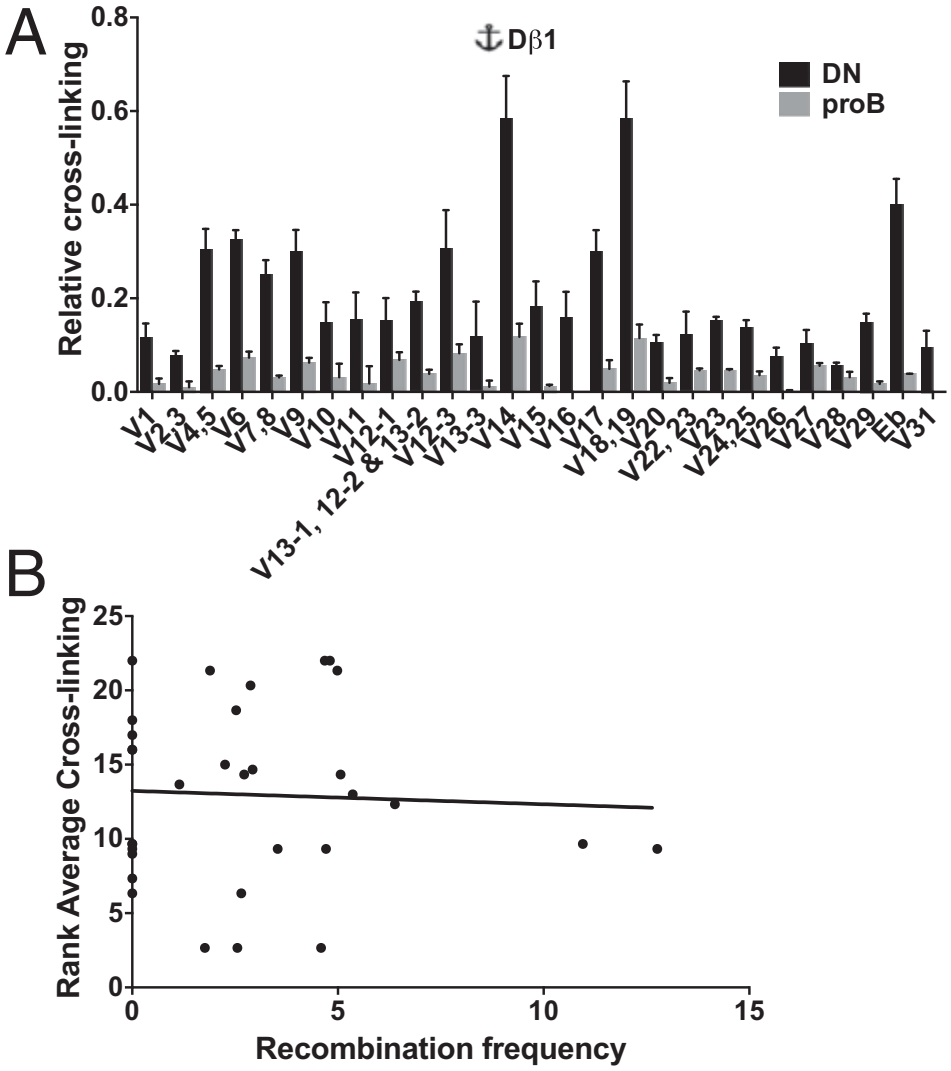


Figure 2.2: Role of Vβ spatial proximity in shaping the *Tcrb* repertoire
(A) 3C analysis of RAG-deficient thymocytes showing relative cross-linking frequencies between a Dβ1 anchor and Hind III fragments spanning Vβ gene segments. Data are presented as mean ± SEM (n = 3). (B) Spearman correlation of Vβ usage and average ranked values for 3C cross-linking frequency from three viewpoints within the recombination center (Dβ1, Dβ2 and Eβ). The Spearman correlation coefficient shows no significance ($r_s = 0.035$, *P-value* =0.85).

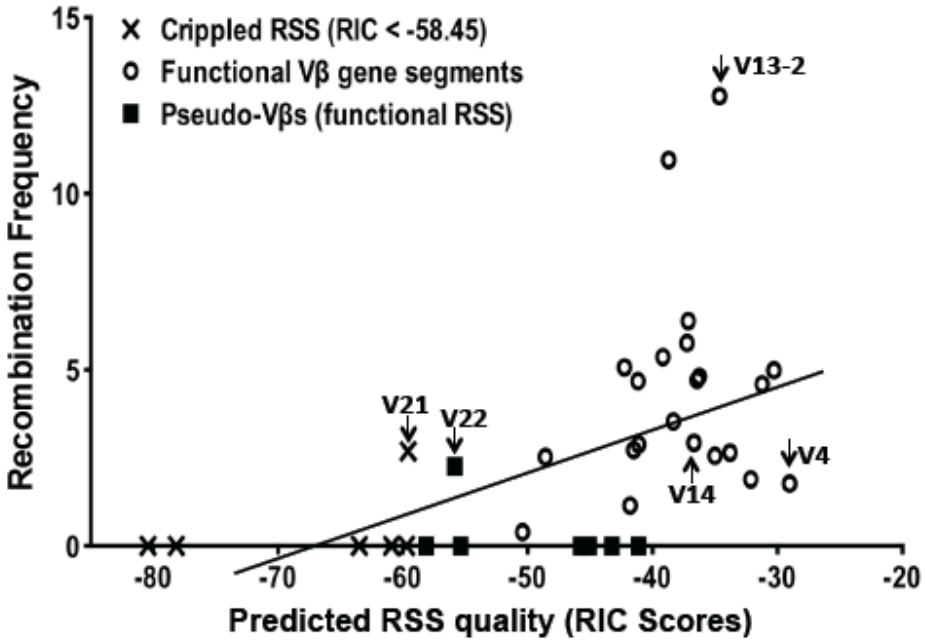


Figure 2.3: Correlation between Vβ utilization and predicted RSS quality
The correlation between predicted Vβ RIC_{23} scores and observed Vβ recombination frequencies (Fig. 2.1B), yielding a Spearman's rank correlation coefficient r_s=0.6456, *P-value* <0.0001.

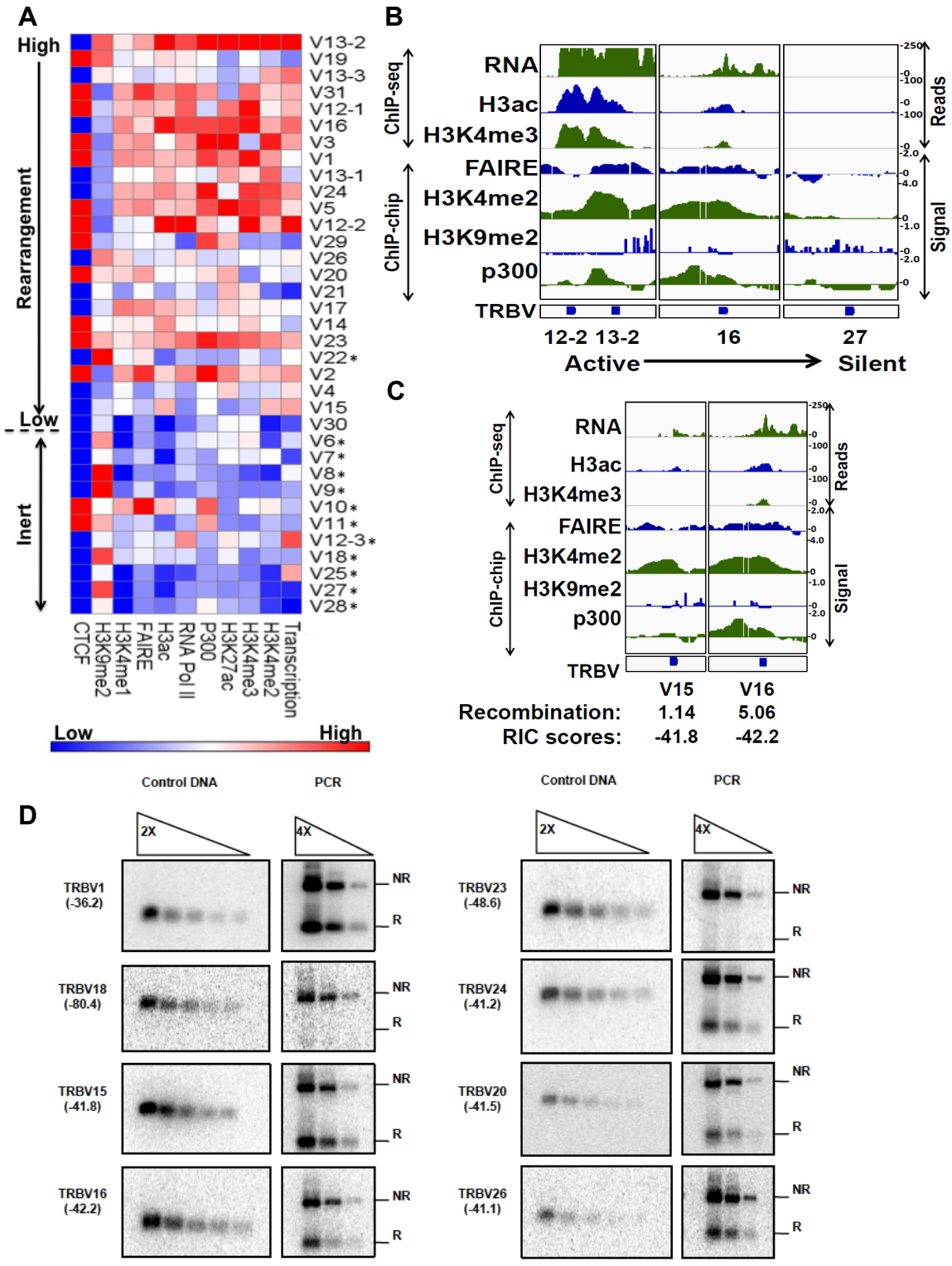


Figure 2.4: Role of chromatin landscape in Vβ usage

(A) Relative intensities of various chromatin features (transcription, RNA Pol II, P300, histone modification signals and proximal CTCF sites) at the 35 Vβ segments are represented as a heatmap. The $\log_2$ values of ChIP-Seq or ChIP-Chip signal intensities at the Vβ segment ($\pm$ 1 kb) for each of the above features were quantified using

BEDtools and the relative intensity for each feature was plotted as a heatmap. CTCF intensities are represented as binary values of one or zero being assigned for presence or absence of CTCF within 1 kb of the Vβ segment. Asterisks denote pseudo V gene segments. (B) Profiles for transcription (RNA), nucleosome depletion (FAIRE), P300, and indicated histone modifications are shown at select Vβ segments. RNA seq data for transcription, ChIP-Seq data for H3ac and H3K4me3, ChIP-chip data (Signal = log2 ratio of ChIP DNA/input DNA) for H3K4me2, P300 and FAIRE are displayed. See text and methods for sources of epigenomic data. (C) Epigenetic profiles at Vβ segments highlighting the influence of chromatin landscapes on gene segment usage. (D) An equimolar mixture of the eight indicated Vβ 23-RSS deletion substrates were assayed for rearrangement in conjunction with the 5′Dβ1 12-RSS following transfection into 293T cells with RAG-1/2 expression vectors (40). Rearrangements were detected by PCR using primers shared by all of the substrates (NR = not rearranged, R = Vβ rearranged to Dβ1). RIC scores for each TRBV-RSS are shown in parentheses. Rearrangements for each substrate were detected using probes specific to the given Vβ segment (specificity controls not shown). A semi-quantitative measure of rearrangement efficiencies was obtained by comparing two-fold dilutions of Vβ plasmid inserts (32-500 ng, left panels) with four-fold dilutions of the PCR product (right panels). Shown are data from one representative PCR amplification out of four independent transfections. Control DNA and PCR products for each Vβ substrate are on the same blot. The TRBV15, 16, 20, 24, and 26 RSSs exhibit similar recombination efficiencies based on this semi-quantitative assay (RIC scores all approx. -42), whereas the TRBV18 and 23 RSSs exhibit minimal rearrangement (lower RIC scores) and TRVB1 rearranges most efficiently (best RIC score).

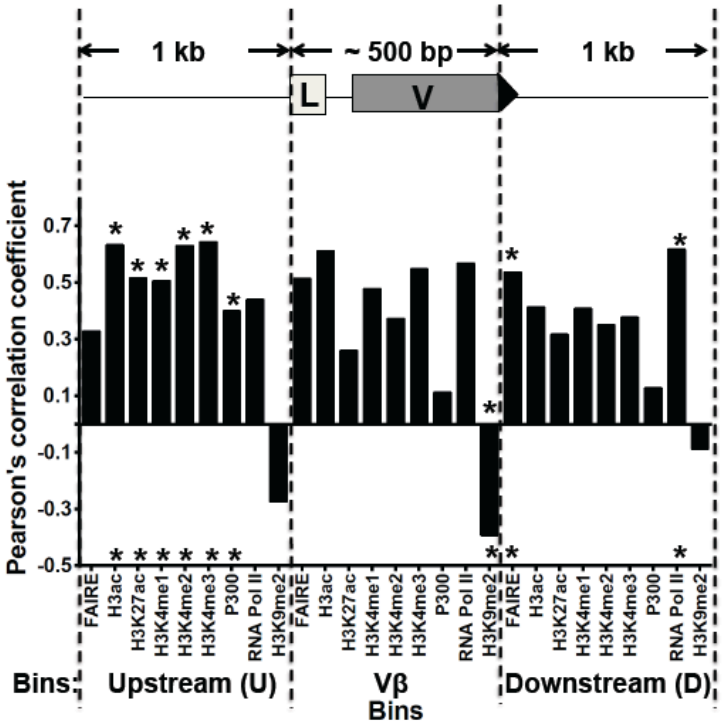


Figure 2.5: Spatial distribution of chromatin features and predictive potential for Vβ usage

The regions surrounding each Vβ segment were divided into three bins (see schematic), U = upstream (1 kb), V = Vβ gene body, and D = downstream (1 kb). Signal densities for each chromatin feature in the spatial bins were correlated with recombination frequencies, yielding a Pearsons's correlation coefficient for each bin. The coefficients were used to determine the best bin, which are denoted by asterisks.

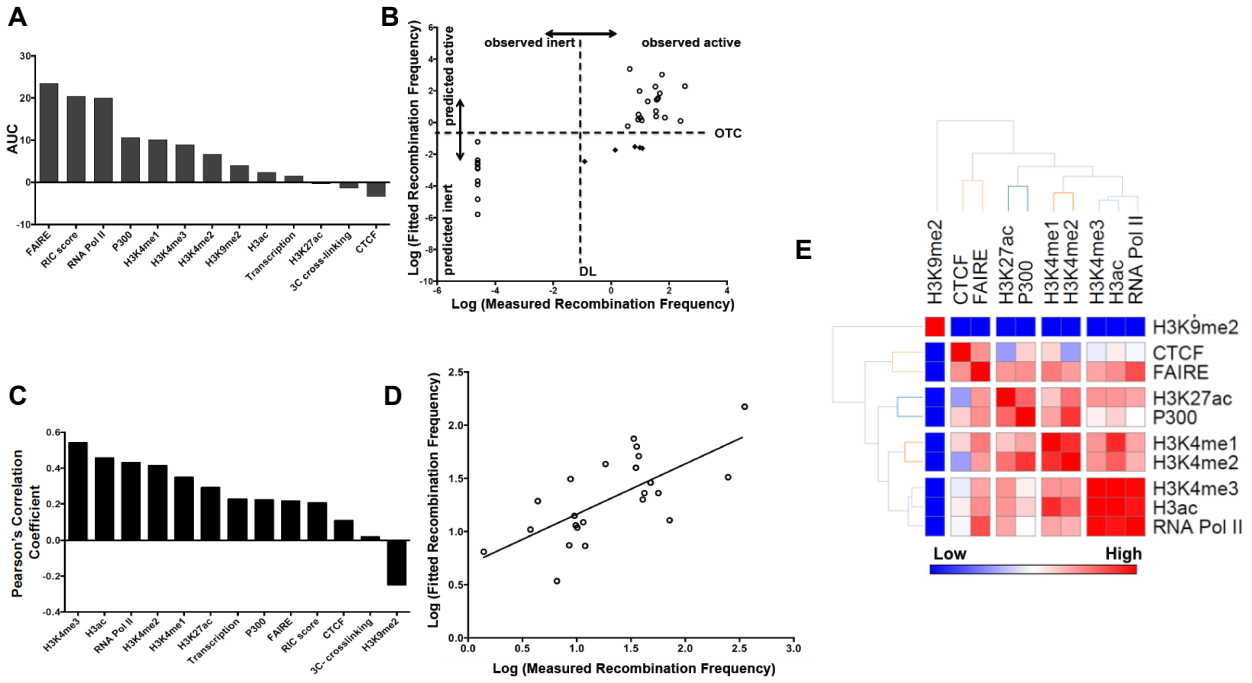


Figure 2.6: Computational analysis of Vβ usage determinants

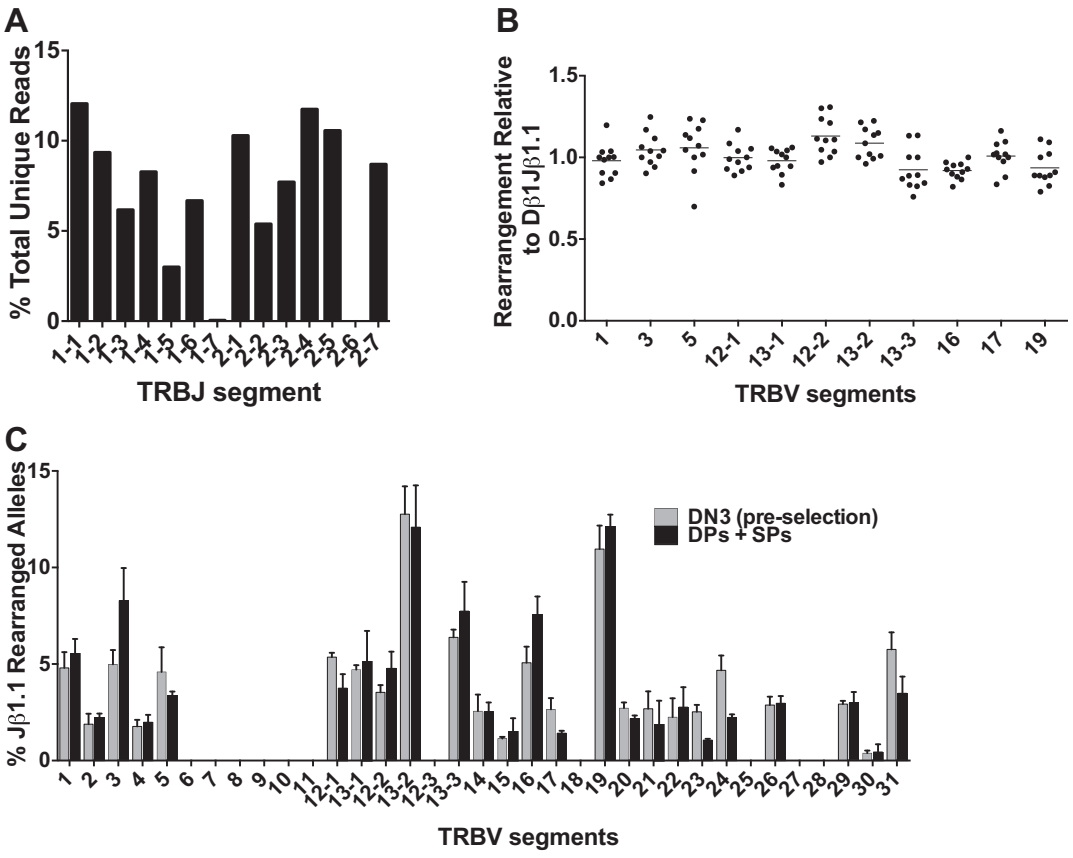
(A) Features that distinguish rearranging from inert Vβ segments (classifier step, see methods). Random forest analysis was performed on the shown features to classify Vβ segments. AUC = Area under the curve, which represents the relative contribution of each feature to the learned classification scheme. (B) Scatter plot representing the classifier step in the two-step model. Linear regression between observed and fitted frequencies using the three most discriminative features for recombining versus inert Vβ gene segments (RIC scores, FAIRE signal and RNA Pol II occupancy). Each symbol represents a Vβ gene segment. Data were generated from the natural logarithm values of recombination frequencies (observed and fitted). The dashed horizontal line represents the optimal threshold for classifying (OTC) rearranging from non-rearranging segments based on the linear combination of the three features. The dashed vertical line represents the detection limit (DL) of Taqman assays used for measuring recombination. Open circles correspond to Vβ segments predicted accurately; black diamonds correspond to outliers. Two of these five exceptions were resolved when the Random Forest algorithm was applied using the three classification features (RIC score, FAIRE, RNA Pol II). (C) Pearson correlation to rank factors that fine-tune Vβ usage in the two-step model (regressor step, see methods). (D) Scatter plot of overall correlation between natural log values of observed and fitted (predicted) frequencies using the five core parameters (H3K4me3, H3K4me2, transcription, P300 and CTCF). Each circle represents one rearranging Vβ segment. The line indicates the best fit between measured and fitted rearrangement frequencies and reflects a strong correlation (Pearson correlation coefficient (0.69, *P-value* = 0.03). (E) Cluster analysis highlights similarities in epigenetic information provided by individual chromatin features.

2.7 Acknowledgements

We thank Drs. Barry Sleckman, Baeck-Seung Lee, and David Schatz for valuable comments and reagents. We thank the Genome Technology Access Center in the Department of Genetics at

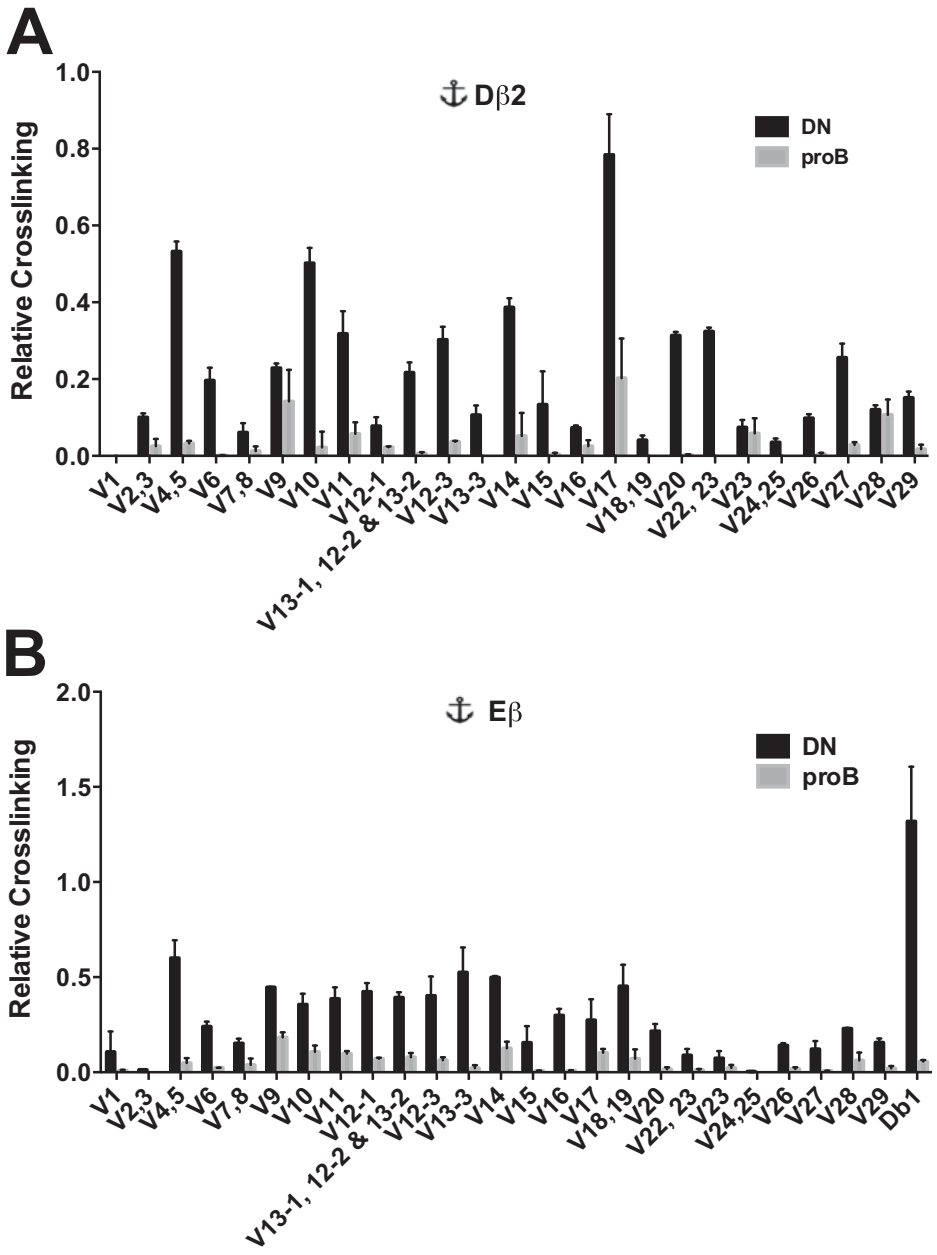
Washington University School of Medicine in St. Louis for help with genomic analyses. The Center is partially supported by National Cancer Institute Cancer Center Support Grant P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant UL1RR024992 from the National Center for Research Resources, a component of the National Institutes of Health (NIH), and the NIH Roadmap for Medical Research. This research was supported by NIH Grants AI 079732, AI 081224 and CA 156690 (to E.M.O.) and AI 082918 (to A.J.F.).

2.8 Supplemental Figures and Tables

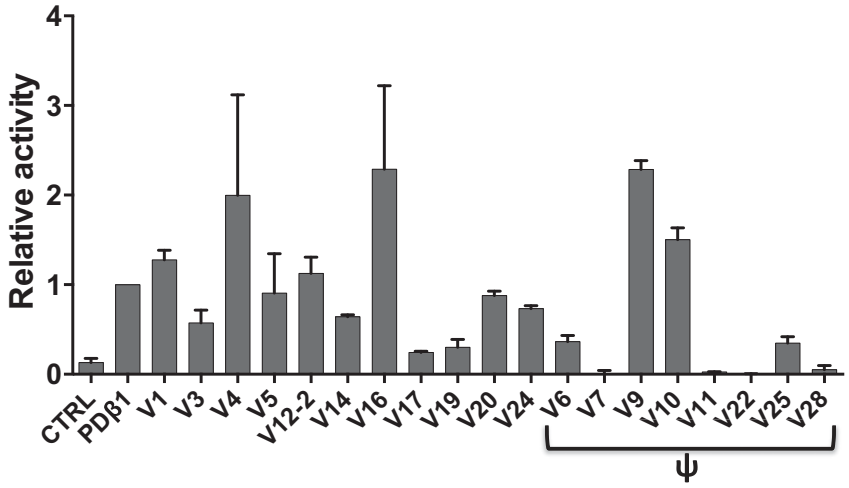


Supplemental Figure 2.1: Vβ repertoire comparisons
(A) Jβ usage profile from high throughput sequencing (mean (n=3),15,000-20,000 unique reads per sample). (B) Distribution of rearrangements from high throughput sequencing involving Vβ segments and each of the 11 functional Jβ segments. Shown are distributions for rearrangements of Vβ segments yielding at least 1000 unique reads. Data are represented relative to the distribution of Vβ-Jβ1.1, where percent total Vβ-Jβ1.1 is set to a value of

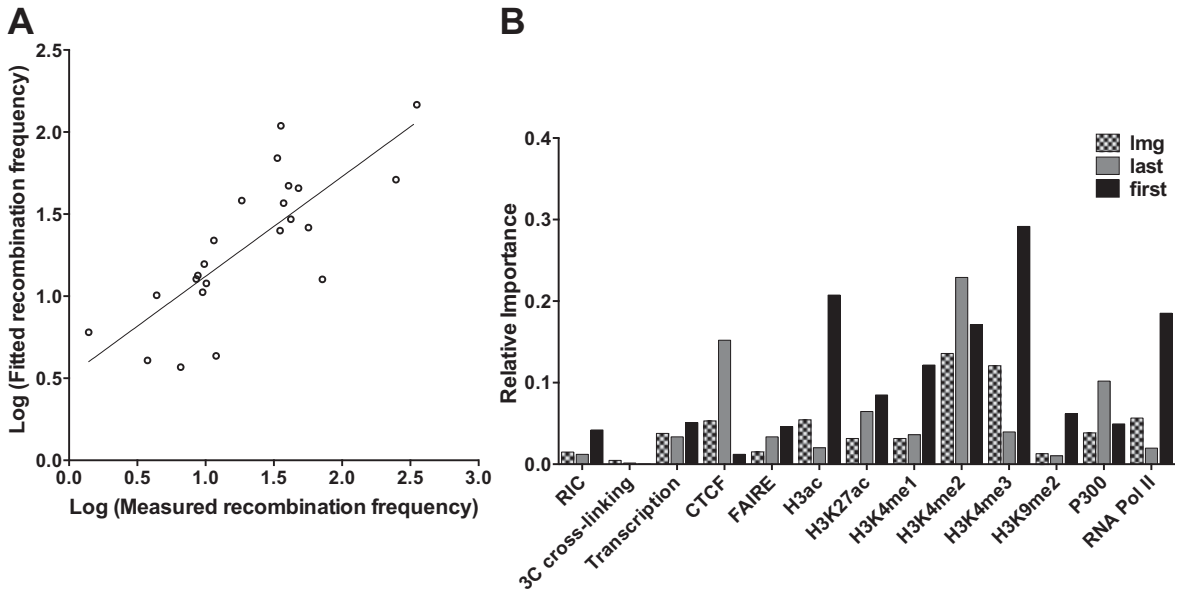
1 (see Fig. 2.1 legend). Each circle represents a data point for a given Jβ segment. (C) Comparison of Vβ usage in pre-selection and post-selection thymocytes measured by the gDNA assay described in Fig. 2.1C (mean ± SEM, *n* = 3)



Supplemental Figure 2.2: Role of spatial proximity in shaping the *Tcrb* repertoire
(A) 3C analysis of Rag-deficient DN thymocytes showing relative cross-linking frequencies between a Dβ2 anchor and Hind III fragments spanning Vβ gene segments. Data are presented as mean ± SEM (n = 3). (B) 3C analysis using an Eβ anchor.



Supplemental Figure 2.3: Luciferase assays
Promoter activity assay for upstream regions of Vβ regions. Relative promoter strengths for select upstream Vβ regions were assessed using luciferase reporter constructs. pGL3 constructs containing the Eβ enhancer and the respective Vβ promoter regions were transfected into a pre-T cell line (T3). Promoter activities were assessed at 24 hours post-transfection and normalized to Renilla (RLU). Data are represented as averages ± SEM (n=3) relative to the control PDβ1 promoter. ψ denotes pseudo Vβ gene segments. CTRL denotes promoterless construct.



Supplemental Figure 2.4: Computational analysis of Vβ usage determinants

(A) Scatter plot of overall correlation between natural log values of observed and fitted frequencies using the complete set of 13 features. Each circle represents one rearranging Vβ segment. The line indicates the best fit between measured and fitted rearrangement frequencies reflect a strong correlation (Pearson correlation coefficient 0.779 (*P-value* = 0.47). (B) Relative contribution of the minimal eight features to the accuracy of fit as computed by three different approaches (lmg, last, first) and the corresponding linear regression coefficients. The best fit formula is as follows:

$$\sum_{i=1}^{8} \text{Coefficient}_i \times \text{Feature}_i$$

The raw values and coefficients corresponding to each feature are provided in Tables S2.5-S2.8).

Supplemental Table 2.1: 3C ranks and rearrangement frequencies

3C HindIII fragments	Rank Eβ	Rank Dβ1	Rank Dβ2	Average Rank	% Recombination
V1	21	20	25	22	4.80392
V2	24	23	17	21.3	1.890142
V3	24	23	17	21.3	4.983436
V4	1	5	2	2.67	1.766591
V5	1	5	2	2.67	4.590954
V6	13	3	12	9.33	0
V7	18	8	22	16	0
V8	18	8	22	16	0
V9	5	7	10	7.33	0
V10	10	16	3	9.67	0
V11	9	12	6	9	0
V12-1	6	14	19	13	5.360448
V13-1	8	9	11	9.33	4.710525
V12-2	8	9	11	9.33	3.534854
V13-2	8	9	11	9.33	12.76473
V12-3	7	4	8	6.33	0
V13-3	2	19	16	12.3	6.387671
V14	3	1	4	2.67	2.558942
V15	17	10	14	13.7	1.14465
V16	11	11	21	14.3	5.066513
V17	12	6	1	6.33	2.652322
V18	4	2	23	9.67	0
V19	4	2	23	9.67	10.95472
V20	15	21	7	14.3	2.722579
V22	22	18	5	15	2.254725
V23	23	13	20	18.7	2.528768
V24	25	17	24	22	4.680041
V25	25	17	24	22	0
V26	19	24	18	20.3	2.878938
V27	20	22	9	17	0
V28	14	25	15	18	0
V29	16	15	13	14.7	2.927078

Supplemental Table 2.2: Primers and probes for Vβ utilization assay

Taqman Probes (5'FAM and 3' TAMRA from Sigma Life Sciences)

Jβ 1.1 Probe	5'FAM-TGTGAGTCTGGTTCCTTTACCAA-3'TAMRA
Jβ 2.1 Probe	5'HEX-TAGGACGGTGAGTCGTGTCC-3'TAMRA

Primers for Cloning VβJβ Template Plasmids

Jβ 1.1 F	5'-GACAGACGGATCCTGGCACTGTGCAAACACAGAAGTC-3'
Jβ 1.1 R	5'-TACATCGCGGCCGCACTCGAATATGGACACGGAGGACA-3'
Jβ 2.1 F	5'GACAGACGGATCCGTAACTATGCTGAGCAGTTCTTCGGACC-3'
Jβ 2.1 R	5'-TACATCGCGGCCGCAGTCCTGGAAATGCTGGCACAAAC-3'
V1-F	5'-TATCTCGAGCTGGAGCAAAACCCAAGGTGG-3'
V1-R	5'-CGAGAAGCTTTGCAGTACAAGGTTCTGCCCT-3'
V2-F	5'-TATCTCGAGCGAAAATTATCCAGAAACCAA-3'
V2-R	5'-CGAGAAGCTTGCACAGAAGTATGTGGCCGAG-3'
V3-F	5'-TATCTCGAGCAGATGGTGACCCTCAATTGT-3'
V3-R	5'-TAGCGAAGCTTTAAGCTGCTGGCACAGAAG-3'
V4-F	5'-TATCTCGAGGACGGCTGTTTTCCAGACTC-3'
V4-R	5'-CGAGAAGCTTTGGCACAGAGATACACAGCAG-3'
V5-F	5'-TATCTCGAGGATATCTAATCCTGGGAAGAGC-3'
V5-R	5'-CGAGAAGCTTCTGCCGTGGATCCAGAAGACT-3'
V6-F	5'-TATCTCGAGGTTACAGACATGGGACAGAATGTCA-3'
V6-R	5'-CGAGAAGCTTAGCTGCTGGCATACATAGTGGAGT-3'
V7-F	5'-TATCTCGAGAGCAGGCTCTGTCTTCTGACTTGT-3'
V7-R	5'-CGAGAAGCTTAGAACAGTGCAGAGTCCTTTGGCT-3'
V8-F	5'-TAGCCTCGAGCATTCAGACTCCCAAATCAT-3'
V8-R	5'-TAGCGAAGCTTTCTGTGCATGATCTGGAGAC-3'
V9-F	5'-TATCTCGAGGTGACACAATTTCTGGTCCTACTGG-3'
V9-R	5'-CGAGAAGCTTCTTCTGGCACAGAGATAGATGCCT-3'

V10-F	5'-TATCTCGAGGGTGGAATCACCCAGACACCTAGATA-3'
V10-R	5'-CGAGAAGCTTAGTACATGGAGGTCTGGTTGGAACTG-3'
V11-F	5'-TATCTCGAGAGGCACTTCTGATATGTGGCCTCT-3'
V11-R	5'-CGAGAAGCTTAGTTAGAAACCATGGCTCTTGCCC-3'
V12-1-F	5'-TATCTCGAGCTGACGTGTATTCCCATCTCT-3'
V12-1-R	5'-TAGCGAAGCTTTCCAGTTCCAAGGCACTCATG-3'
V13-1-F	5'-TATCTCGAGTGGTTAGCCCAAGTGTGCTTCTCT-3'
V13-1-R	5'-CGAGAAGCTTAAGCCAATTCCAGCAGGAGGAAGA-3'
V12-2-F	5'-TATCTCGAGCATTGCTGCTGCTGCTGCTGC-3'
V12-2-R	5'-TAGCGAAGCTTACACGGCAGAGTCCTCTAG-3'
V13-2-F	5'-TATCTCGAGTCCTGTGTTCAAGTGAGTGCTGGT-3'
V13-2-R	5'-CGAGAAGCTTTTGGTCTGGAGGCCTTGTATCCAT-3'
V12-3-F	5'-TAGCCTCGAGCCTTCTCCCCAGGTTCAGC-3'
V12-3-R	5'-TAGCGAAGCTTCACAGTAAAGTCCTCTAGGTCC-3'
V13-3-F	5'-TATCTCGAGGACGATATGATCAGGCTTTG-3'
V13-3-R	5'-TAGCGAAGCTTAGAAATATACAGCTGTCTGAG-3'
V14-F	5'-TATCTCGAGTATGCAGTCCTACAGGAAGGGCAA-3'
V14-R	5'-CGAGAAGCTTAAACTGCTGGCACAGAGATAGGTG-3'
V15-F	5'-TATCTCGAGCAGACACCCAGACATGAGGT-3'
V15-R	5'-CGAGAAGCTTACAGCTGAGTCCTTGGGTTCTG-3'
V16-F	5'-TATCTCGAGCACCTAGGCACAAGGTGACA-3'
V16-R	5'-TAGCGAAGCTTCAGGACTCAGCGGTGTATCT-3'
V17-F	5'-TATCTCGAGGGATACTACGGTTAAGCAGAAC-3'
V17-R	5'-TAGCGAAGCTTAGCACAGAGGTACATGGCAG-3'
V18-F	5'-TAGCCTCGAGGCTGGTGTCACCACGAACCT-3'
V18-R	5'-TAGCGAAGCTTCTCTGCATCTTCCAGATCTGC-3'
V19-F	5'-TATCTCGAGCTCAGACACCCAAATTCCTGA-3'

V19-R	5'-CGAGAAGCTTGCTATACTGCTGGCACAGAGA-3'
V20-F	5'-TATCTCGAGCGTCTATCAATATCCCAGAAG-3'
V20-R	5'-CGAGAAGCTTAGCACCACAGAGATATAAGCC-3'
V21-F	5'-TAGCCTCGAGGTTGTCCAGAATCCTAGACAT-3'
V21-R	5'-TAGCGAAGCTTGTACACAGCTGAATCTGTTAG-3'
V22-F	5'-TATCTCGAGCCAAGTTATCCAGACTCCAT-3'
V22-R	5'-TAGCGAAGCTTATAACACTGAGTCTCCAGCCTC-3'
V23-F	5'-TATCTCGAGGAAAGGCCAGGAAGCAGAGAT-3'
V23-R	5'-CGAGAAGCTTGCTGGAGCACAAGTACAGTGC-3'
V24-F	5'-TATCTCGAGGAGTAACCCAGACTCCACGAT-3'
V24-R	5'-CGAGAAGCTTGACTGCTGGCACAGAGCTACA-3'
V25-F	5'-TAGCCTCGAGCTAGCTTCAAGGCTCTTCTA-3'
V25-R	5'-TAGCGAAGCTTATGTAGAATCTCCTGCTTCT-3'
V26-F	5'-TATCTCGAGCAGACTCCAAGATATCTGGTG-3'
V26-R	5'-CGAGAAGCTTCTGCTGGCACAGAGGTACAGT-3'
V27-F	5'-TAGCCTCGAGCTCCAAAGTACTCTATTATG-3'
V27-R	5'-TAGCGAAGCTTGAGGTAGGATTCATTCTCTG-3'
V28-F	5'-TAGCCTCGAGCATCCAAATCGCAAGACACC-3'
V28-R	5'-TAGCGAAGCTTAGGTGCACACATGCCTGGTCG-3'
V29-F	5'-TATCTCGAGCTGATCAAAAGAATGGGAGAG-3'
V29-R	5'-CGAGAAGCTTCTAGCACAGAAGTACACAGATG-3'
V30-F	5'-TATCTCGAGTGCTTGCCTCATGGATCTCTGTCT-3'
V30-R	5'-CGAGAAGCTTGAACTACAGAAATAGATACTGC-3'
V31-F	5'-TAGCCTCGAGCTGAGACTGATTACATGTAA-3'
V31-R	5'-TAGCGAAGCTTAGAAGCCAGAGTGGCTGAGA-3'

qPCR Primers for Taqman assay

qV1F	5'-GCCACACGGGTCACTGATAC-3'
qV2F	5'-GTTCAAAGAAAAACCATTTAG-3'
qV3F	5'-GATGGTTCATATTTCACTCT-3'
qV4F	5'-CAGATAAAGCTCATTTGAAT-3'
qV5F	5'-GCCCAGACAGCTCCAAGCTAC-3'
qV6F	5'-CAGAGATGCCTGATGGATTGTT-3'
qV7F	5'-CAGCACACCAATTTGGTGACT-3'
qV8F	5'-GAGGTCTCTAAGGGGTAC-3'
qV9F	5'-CTTCTCCATGTTGAAGAGCCAA-3'
qV10F	5'-AGAAATGAGATACAGAGCTTTCC-3'
qV11F	5'-AGTTAGAAACCATGGCTCTTGC-3'
qV12-1F	5'-'TAGCAATGTGGTCTGGTACCAG-3'
qV13-1F	5'-GGTACAAGGCCACCAGAACA-3'
qV12-2F	5'-TCTCTCTGTGGCCTGGTATCAA-3'
qV13-2F	5'-GCTGGCAGCACTGAGAAAGGA-3'
qV12-3F	5'-CCTGAGTGCCTTGGACCT-3'
qV13-3F	5'-TTCCCTTTCTCAGACAGCTGTA-3'
qV14F	5'-TATCAGCAGCCCAGAGACCAG-3'
qV15F	5'-CACTCTGAAGATTCAACCT-3'
qV16F	5'-CTCAGCTCAGATGCCCAAT-3'
qV17F	5'-CAATCCAGTCGGCCTAACA-3'
qV18F	5'-CCACGAACCTAAGATACAT-3'
qV19F	5'-CTCGAGAGAAGAAGTCATCT-3'
qV20F	5'-CAGTCATCCCAACTTATCCT-3'
qV21F	5'-GCTAAGAAACCATGTACCAT-3'
qV22F	5'-CAGTTCCTCTGAGGCTGGA-3'
qV23F	5'-CTGTGTGCCCCTCCAGCTCA-3'

qV24F	5'-CTCAGCTAAGTGTTCCTCGA-3'
qV25F	5'-CTATGTGGCATATTACTGGT-3'
qV26F	5'-CCTTCAAACTCACCTTGCAGC-3'
qV27F	5'-CATTGTTCATATGGCATT-3'
qV28F	5'-CTCTGATAGATATATCAT-3'
qV29F	5'-CTGATTCTGGATTCTGCTA-3'
qV30F	5'-CAATGCAAGGCCTGGAGACA-3'
qV31F	5'-AAATCAAGCCCTAACCTCTAC-3'
qJβ1.1R	5'-CTCGAATATGGACACGGAGGACATGC-3'
qJβ2.1R	5'-CCTGATACAGGGCCTTGGATAGTTA-3'

Supplemental Table 2.3: Primers and probes for 3C assay

3C Anchor Primers and Taqman Probes (5'FAM and 3' TAMRA from Sigma Life Sciences)

Dβ_1 Hind III probe	5'-AAGGCATTGTTGCATGATCCT-3'
Dβ_2 Hind III probe	5'-AAATGCTGGGCCTCTGTAGA-3'
Eβ_ Hind III probe	5'-CATAAGCATTGTCATGTTTGTGACA-3'
ERCC3 Hind III probe	5'-AAAGCTTGCACCCTGCTTTAGTGGCC-3'
Dβ_1 Hind III primer	5'-TGAAATTTTTCTGCCGAAAGGAC-3'
Dβ_2 Hind III primer	5'-GCGGGATCCAAGAGAACTCA-3'
Eβ_ Hind III primer	5'-GAAAATTGGCATCGGTTTGC-3'

Hind III Primers

V1	5'-TATCTCTGTGGGGCATGCAG-3'
V2	5'-TTTCATTCACAGCCGACCAG-3'
V3	5'-TTTCATTCACAGCCGACCAG-3'

V4	5’-AGCTCGACACAGAAAGCAAGTT-3’
V5	5’-AGCTCGACACAGAAAGCAAGTT-3’
V6	5’-GGTTCCCTTCACTTCCCACA-3’
V7	5’-GTCCGCTAGCAGCCAGAGTT-3’
V8	5’-GTCCGCTAGCAGCCAGAGTT-3’
V9	5’-ACCAGAGGGCAGCTGAAAAT-3’
V10	5’-GTGCCTGTACCATGCTGTGG-3’
V11	5’-TTCAGCAAGTAGGTGCGAAGA-3’
V12-1	5’-TGGTGGGATCCTGACAGCTTATA-3’
V13-1	5’-CCATCTGCATGAACACCTTCTT-3’
V12-2	5’-CCATCTGCATGAACACCTTCTT-3’
V13-2	5’-CCATCTGCATGAACACCTTCTT-3’
V12-3	5’-GGATCTTGGTCTCGGGAGGT-3’
V13-3	5’-CTCAGCTGCACCCTCACAAC-3’
V14	5’-CAGGCTTTTGAGTGGCATGT-3’
V15	5’-AGGCAGGAGGTGAGTCTTGG-3’
V16	5’-TATCATGCCCAGCTGCATTC-3’
V17	5’-GTTAGGCCGACTGGATTGGA-3’
V18	5’-GGCAGTGTTACAGAACCCAGTG-3’
V19	5’-GGCAGTGTTACAGAACCCAGTG-3’
V20	5’-TGTGATGGGTTGTCATCTGGA-3’
V22	5’-CCAAGGGATGATGTCACAGG-3’
V23	5’-TACACCGGCCAGGAGAGACT-3’
V24	5’-ACTAGGCCAGCAGAGGATGC-3’
V25	5’-ACTAGGCCAGCAGAGGATGC-3’
V26	5’-AGCATAGGATTGGGCCTCAG-3’
V27	5’-CATCACTGCGCCTAGCAATC-3’

V28	5’-GCGTGTGCCACGTTTTTGTA-3’
V29	5’-CTCTAGCAATCCCCCTGTGC-3’
V31	5’-AAGGAGAGAGCAGGCCACAG-3’
Dβ_1	5’-AAGGCATTGTTGCATGATCC-3’
Dβ_2	5’-TGGGGCCCTCACTTTTCTTA-3’
Eβ	5’-TCCTAAGGAGAGGCAGAGTGG-3’
ERCC3	5’-GACTTCTCACCTGGGCCTACA-3’

Supplemental Table 2.4: Luciferase cloning primers

Eβ-F	5’-ATTGGATCCGTTAACCAGGCACAGTAGGACC-3’
Eβ-R	5’-ATTGGATCCCCATGGTGCATACTGAAGGCTTC-3’
Pro-V1F	5’-TAGCCTCGAGGAGTGACTAGTTACTTCTGC-3’
Pro-V1R	5’-TAGCGAAGCTTCTCTGAGACCTCAGGTTCTC-3’
Pro-V3-F	5’-TATCTCGAGGGGACTCAGTTCAGTAGTC-3’
Pro-V3-R	5’-CGAGAAGCTTAGTAGGGTCACGGCAGGAA-3’
Pro-V4F	5’-TAGCCTCGAGTGTGCTAAGGGCACCAATGAAT-3’
Pro-V4R	5’-TAGCGAAGCTTGTTGGGTCAAGGCAGGGCAAAT-3’
Pro V5-FX	5’-TAGCCTCGAGTATCCATTGTATGCTCTGTTTG-3’
Pro V5-RH	5’-TAGCGAAGCTTGGTGGAATCAGGCTCCAGACG-3’
Pro V6-FX	5’-TAGCCTCGAGCTACAAGCTCCCAAGAGAGAG-3’
Pro V6-RH	5’-TAGCGAAGCTTCTCTGGAGAAGACAGAGGAC-3’
Pro-V7F	5’-TAGCCTCGAGGCTGCTGAATAGCAAGTTTCCAG-3’
Pro-V7R	5’-TAGCGAAGCTTTTGGAGGTTTGGATCTGTAGTCT-3’
Pro V9-FX	5’-TAGCCTCGAGGGAACTTTCATGTGAGGAGA-3’
Pro V9-RH	5’-TAGCGAAGCTTCTGCAAAAATATAAGTTGTGAACAG-3’
Pro V10-FX	5’-TAGCCTCGAGGGGATATCTCTATGCTTTAATG-3’

Pro V10-RH	5’-TAGCGAAGCTTCTGGAGAAGGAGGCATAAGGA-3’
Pro-V11F	5’-TAGCCTCGAGTTCCCTACAGTGTCAAGGGCTG-3’
Pro-V11R	5’-TAGCGAAGCTTTGTACCCACAGGGTTGTTCTCA-3’
Pro-V12-2-F	5’-TAGCCTCGAGCAACTGACTCAGAGAAAAAC-3’
Pro-V12-2-F	5'-TAGCGAAGCTTTCCTCTCAGGATACTGGTCTCT-3'
Pro-V14F	5’-TACATCGCTAGCCATTTATGTGTACCATAATAAT-3’
Pro-V14R	5’-TAGCCTCGAGGGCAGATTGAGGGCAGAGGAG-3’
Pro-V16F	5’-TAGCCTCGAGTTGCAATCTACCTCTGCTGCTC-3’
Pro-V16R	5’-TAGCGAAGCTTTTGTGATGACACCACTGTCTCCG-3’
Pro V17-FX	5’-TAGCCTCGAGGCAGGTGTGACCTACGATAAC-3’
Pro V17-RH	5’-TAGCGAAGCTTGGATGGTCCAGAACAGGAAA-3’
Pro-V19-F	5’-TATCTCGAGCATTTGAGAAAGACAACAA-3’
Pro-V19-R	5’-CGAGAAGCTTAGTTTGGAGGGACTTTCTT-3’
Pro-V20F	5’-TAGCCTCGAGGATAAGGTAACTGAAGCGGGA-3’
Pro-V20R	5’-TAGCGAAGCTTCTTCAGTGTTGACTTCACACC-3’
Pro-V22F	5’-TAGCCTCGAGGATGAAATATGGTAACAAGG-3’
Pro-V22R	5’-TAGCGAAGCTTAGGAGATAAAGGGCTACATA-3’
Pro-V24F	5’-TACATCGCTAGCCCAATGATATGTGCAGAGATGA-3’
Pro-V24R	5’-TAGCCTCGAGGATCACACTAGGCCAGCAGAG-3’
Pro-V25F	5’-TAGCCTCGAGCAATTGGGCCATCTTCTGCCAC-3’
Pro-V25R	5’-TAGCGAAGCTTCAGGTGGATACTTCATTCC-3’
Pro-V28F	5’-TAGCCTCGAGAGTTGTCTTGTGGGCAACTCTG-3’
Pro-V28R	5’-TAGCGAGATCTGCTAGATAGCCTCAAGGCTGCAAA-3’

Supplemental Table 2.5: Recombination substrate oligos

Primer name	Sequences

RS V1F	TAGCCTCGAGATACGGAGCTGAGGCTGCAAG
RS V1R	TACATCGCGGCCGCAGTCACCTTATAACTCATGCA
RS V15F	TAGCCTCGAGCCTTCTCCACTCTGAAGATTC
RS V15R	TACATCGCGGCCGCTTCCACCCAAGATTTCTTAA
RS V16F	TAGCCTCGAGACTCAACTCTGAAGATCCAGA
RS V16R	TACATCGCGGCCGCTAATGTAATACTCGTTACCAT
RS V18F	TAGCCTCGAGCCCAACATCCTAAAGTGGG
RS V18R	TACATCGCGGCCGCTTCCTCCGTAAGCATGGTG
RS V20F	TAGCCTCGAGCAGTCATCCCAACTTATCCT
RS V20R	TACATCGCGGCCGCCTCCTGGGTACCCTCCCATTTC
RS V23F	TAGCCTCGAGCACTCTGCAGCCTGGGAATC
RS V23R	TACATCGCGGCCGCTGACTTGGTCTGGGTGTGCTG
RS V24F	TAGCCTCGAGAGTGCATCCTGGAAATCCTAT
RS V24R	TACATCGCGGCCGCAGACCTGGCCTGTTTCTCATG
RS V26F	TAGCCTCGAGCAAGAAGTTCTTCAGCAAATA
RS V26R	TACATCGCGGCCGCGATACAGGTTTCAGTTAGTT

Supplemental Table 2.6: Computational analysis coefficients for determinants of Vβ frequencies (all Tcrb V gene segments): Classifer step, three features

	Estimate	Std Err	t	Pr(>|t|)
(Intercept)	1.09059	1.52205	0.717	0.47903
RIC score	0.08803	0.02619	3.362	0.00207
FAIRE	0.03185	0.01639	1.944	0.06105
RNA Pol II	0.65913	0.26654	2.473	0.01909

Supplemental Table 2.7: Computational analysis coefficients for determinants of Vβ frequencies (all Tcrb V gene segments): Combinatorial analysis of 13 features and their correlation to recombination frequency

Number of Features	Pearson Correlation Coefficient	P-value

13	0.77954	0.4707
8	0.74191	0.1015
7	0.72604	0.07434
6	0.71277	0.04925
5	0.68818	0.03779
4	0.66405	0.0265
3	0.64982	0.01359
2	0.60304	0.01089
1	0.53998	0.00782

Supplemental Table 2.8: Coefficients for determinants of Vβ frequencies (rearranging Vβ segments)

All *Tcrb* V gene segments (Regressor step, 13 features)

	Estimate	Std Err	t	Pr(>|t|)
(Intercept)	0.08707	5.81E+00	0.015	0.9882
RIC score	0.08817	3.72E-02	2.373	0.0273
3C crosslinking	-2.17745	3.14E+00	-0.693	0.4961
Transcription	-0.1299	1.56E-01	-0.83	0.4156
CTCF	0.9394	1.24E+00	0.756	0.4579
FAIRE	0.01538	2.46E-02	0.625	0.5384
H3ac	-0.31847	5.05E-01	-0.63	0.5353
H3K27ac	0.03124	3.86E-02	0.81	0.4271
H3K4me1	0.16488	6.88E-01	0.24	0.813
H3K4me2	0.0194	1.67E-02	1.159	0.2595
H3K4me3	-0.08483	3.68E-01	-0.231	0.8197
H3K9me2	0.74873	1.27E+00	0.59	0.5618
P300	-0.03168	3.03E-02	-1.047	0.3069

RNA Pol II	1.10351	5.21E-01	2.119	0.0462

All *Tcrb* V gene segments (Regressor step, 5 features)

	Estimate	Std Err	t	Pr(>|t|)
(Intercept)	0.62866	0.35047	1.794	0.0907
Transcription	-0.0613	0.0467	-1.313	0.2066
CTCF	0.31418	0.25634	1.226	0.2371
H3K4me2	0.00779	0.00365	2.137	0.0475
H3K4me3	0.16139	0.0666	2.423	0.0268
P300	-0.0066	0.00646	-1.027	0.319

2.9 References

Aoki-Ota, M., Torkamani, A., Ota, T., Schork, N., and Nemazee, D. (2012). Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. J Immunol *188*, 2305-2315.

Bassing, C.H., Swat, W., and Alt, F.W. (2002). The mechanism and regulation of chromosomal V(D)J recombination. Cell *109 Suppl*, S45-55.

Bossen, C., Mansson, R., and Murre, C. (2012). Chromatin topology and the regulation of antigen receptor assembly. Annu Rev Immunol *30*, 337-356.

Chaumeil, J., Micsinai, M., Ntziachristos, P., Deriano, L., Wang, J.M., Ji, Y., Nora, E.P., Rodesch, M.J., Jeddeloh, J.A., Aifantis, I.*, et al.* (2013). Higher-order looping and nuclear organization of Tcra facilitate targeted rag cleavage and regulated rearrangement in recombination centers. Cell Rep *3*, 359-370.

Chen, Y., Meyer, C.A., Liu, T., Li, W., Liu, J.S., and Liu, X.S. (2011). MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-chip or ChIP-seq data. Genome Biol *12*, R11.

Cobb, R.M., Oestreich, K.J., Osipovich, O.A., and Oltz, E.M. (2006). Accessibility control of V(D)J recombination. Adv Immunol *91*, 45-109.

Cowell, L.G., Davila, M., Yang, K., Kepler, T.B., and Kelsoe, G. (2003). Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. J Exp Med *197*, 207-220.

Cuomo, C.A., Mundy, C.L., and Oettinger, M.A. (1996). DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. Mol Cell Biol *16*, 5683-5690.

Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D.*, et al.* (2011). CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc Natl Acad Sci U S A *108*, 9566-9571.

Dekker, J. (2008). Gene regulation in the third dimension. Science *319*, 1793-1794.

Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigo, R., Birney, E.*, et al.* (2012). Modeling gene expression using chromatin features in various cellular contexts. Genome Biol *13*, R53.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M.*, et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Feeney, A.J. (2009). Genetic and epigenetic control of V gene rearrangement frequency. Adv Exp Med Biol *650*, 73-81.

Feeney, A.J., Tang, A., and Ogwaro, K.M. (2000). B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. Immunol Rev *175*, 59-69.

Ferrier, P., Krippl, B., Blackwell, T.K., Furley, A.J., Suh, H., Winoto, A., Cook, W.D., Hood, L., Costantini, F., and Alt, F.W. (1990). Separate elements control DJ and VDJ rearrangement in a transgenic recombination substrate. EMBO J *9*, 117-125.

Fuxa, M., Skok, J., Souabni, A., Salvagiotto, G., Roldan, E., and Busslinger, M. (2004). Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. Genes Dev *18*, 411-422.

Gerstein, R.M., and Lieber, M.R. (1993). Coding end sequence can markedly affect the initiation of V(D)J recombination. Genes Dev *7*, 1459-1469.

Giresi, P.G., and Lieb, J.D. (2009). Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods *48*, 233-239.

Godfrey, D.I., Hammond, K.J., Poulton, L.D., Smyth, M.J., and Baxter, A.G. (2000). NKT cells: facts, functions and fallacies. Immunol Today *21*, 573-583.

Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., Oltz, E.M., and Sen, R. (2011a). Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. Cell *147*, 332-343.

Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.L., Hansen, E., Despo, O., Bossen, C., Vettermann, C.*, et al.* (2011b). CTCF-binding elements mediate control of V(D)J recombination. Nature *477*, 424-430.

Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc *2*, 1722-1733.

Hesse, J.E., Lieber, M.R., Mizuuchi, K., and Gellert, M. (1989). V(D)J recombination: a functional definition of the joining signals. Genes Dev *3*, 1053-1061.

Jackson, A., Kondilis, H.D., Khor, B., Sleckman, B.P., and Krangel, M.S. (2005). Regulation of T cell receptor beta allelic exclusion at a level beyond accessibility. Nat Immunol *6*, 189-197.

Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J., Grosveld, F.G., Knoch, T.A., and Murre, C. (2008). The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. Cell *133*, 265-279.

Ji, Y., Resch, W., Corbett, E., Yamane, A., Casellas, R., and Schatz, D.G. (2010). The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell *141*, 419-431.

Jung, D., Bassing, C.H., Fugmann, S.D., Cheng, H.L., Schatz, D.G., and Alt, F.W. (2003). Extrachromosomal recombination substrates recapitulate beyond 12/23 restricted VDJ recombination in nonlymphoid cells. Immunity *18*, 65-74.

Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A *107*, 2926-2931.

Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le Beau, M.M., Fisher, A.G., and Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science *296*, 158-162.

Kwon, J., Morshead, K.B., Guyon, J.R., Kingston, R.E., and Oettinger, M.A. (2000). Histone acetylation and hSWI/SNF remodeling act in concert to stimulate V(D)J cleavage of nucleosomal DNA. Mol Cell *6*, 1037-1048.

Lee, A.I., Fugmann, S.D., Cowell, L.G., Ptaszek, L.M., Kelsoe, G., and Schatz, D.G. (2003). A functional analysis of the spacer of V(D)J recombination signal sequences. PLoS Biol *1*, E1.

Lefranc, M.P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J.*, et al.* (2009). IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res *37*, D1006-1012.

Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K., and Murre, C. (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat Immunol.

Liu, H., Schmidt-Supprian, M., Shi, Y., Hobeika, E., Barteneva, N., Jumaa, H., Pelanda, R., Reth, M., Skok, J., and Rajewsky, K. (2007a). Yin Yang 1 is a critical regulator of B-cell development. Genes Dev *21*, 1179-1189.

Liu, Y., Subrahmanyam, R., Chakraborty, T., Sen, R., and Desiderio, S. (2007b). A plant homeodomain in RAG-2 that binds Hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. Immunity *27*, 561-571.

Livak, F. (2003). Evolutionarily conserved pattern of gene segment usage within the mammalian TCRbeta locus. Immunogenetics *55*, 307-314.

Matthews, A.G., Kuo, A.J., Ramon-Maiques, S., Han, S., Champagne, K.S., Ivanov, D., Gallardo, M., Carney, D., Cheung, P., Ciccone, D.N.*, et al.* (2007). RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature *450*, 1106-1110.

Nadel, B., Tang, A., Lugo, G., Love, V., Escuro, G., and Feeney, A.J. (1998). Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. J Immunol *161*, 6068-6073.

Ndifon, W., Gal, H., Shifrut, E., Aharoni, R., Yissachar, N., Waysbort, N., Reich-Zeliger, S., Arnon, R., and Friedman, N. (2012). Chromatin conformation governs T-cell receptor Jbeta gene segment usage. Proc Natl Acad Sci U S A *109*, 15865-15870.

Oestreich, K.J., Cobb, R.M., Pierce, S., Chen, J., Ferrier, P., and Oltz, E.M. (2006). Regulation of TCRbeta gene assembly by a promoter/enhancer holocomplex. Immunity *24*, 381-391.

Olaru, A., Patterson, D.N., Villey, I., and Livak, F. (2003). DNA-Rag protein interactions in the control of selective D gene utilization in the TCR beta locus. J Immunol *171*, 3605-3611.

Osipovich, O., Cobb, R.M., Oestreich, K.J., Pierce, S., Ferrier, P., and Oltz, E.M. (2007). Essential function for SWI-SNF chromatin-remodeling complexes in the promoter-directed assembly of Tcrb genes. Nat Immunol *8*, 809-816.

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., Imbert, J., Andrau, J.C., Ferrier, P., and Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J *30*, 4198-4210.

Posnett, D.N., Vissinga, C.S., Pambuccian, C., Wei, S., Robinson, M.A., Kostyu, D., and Concannon, P. (1994). Level of human TCRBV3S1 (V beta 3) expression correlates with allelic polymorphism in the spacer region of the recombination signal sequence. J Exp Med *179*, 1707-1711.

Reddy, K.L., Zullo, J.M., Bertolino, E., and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature *452*, 243-247.

Reynaud, D., Demarco, I.A., Reddy, K.L., Schjerven, H., Bertolino, E., Chen, Z., Smale, S.T., Winandy, S., and Singh, H. (2008). Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros. Nat Immunol *9*, 927-936.

Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E.*, et al.* (2011). The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. Immunity *35*, 501-513.

Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H., and Carlson, C.S. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood *114*, 4099-4107.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A *105*, 8309-8314.

Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. Nat Rev Immunol *11*, 251-263.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K.*, et al.* (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467-471.

Shih, H.Y., Verma-Gaur, J., Torkamani, A., Feeney, A.J., Galjart, N., and Krangel, M.S. (2012). Tcra gene recombination is supported by a Tcra enhancer- and CTCF-dependent chromatin hub. Proc Natl Acad Sci U S A.

Shimazaki, N., Tsai, A.G., and Lieber, M.R. (2009). H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. Mol Cell *34*, 535-544.

Skok, J.A., Gisler, R., Novatchkova, M., Farmer, D., de Laat, W., and Busslinger, M. (2007). Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes. Nat Immunol *8*, 378-387.

Spicuglia, S., Payet, D., Tripathi, R.K., Rameil, P., Verthuy, C., Imbert, J., Ferrier, P., and Hempel, W.M. (2000). TCRalpha enhancer activation occurs via a conformational change of a pre-assembled nucleo-protein complex. EMBO J *19*, 2034-2045.

Williams, G.S., Martinez, A., Montalbano, A., Tang, A., Mauhar, A., Ogwaro, K.M., Merz, D., Chevillard, C., Riblet, R., and Feeney, A.J. (2001). Unequal VH gene rearrangement frequency within the large VH7183 gene family is not due to recombination signal sequence variation, and mapping of the genes shows a bias of rearrangement based on chromosomal location. J Immunol *167*, 257-263.

Wilson, A., Marechal, C., and MacDonald, H.R. (2001). Biased V beta usage in immature thymocytes is independent of DJ beta proximity and pT alpha pairing. J Immunol *166*, 51-57.

Wu, C., Bassing, C.H., Jung, D., Woodman, B.B., Foy, D., and Alt, F.W. (2003). Dramatically increased rearrangement and peripheral representation of Vbeta14 driven by the 3'Dbeta1 recombination signal sequence. Immunity *18*, 75-85.

Xiang, Y., Zhou, X., Hewitt, S.L., Skok, J.A., and Garrard, W.T. (2011). A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. J Immunol *186*, 5356-5366.

Yu, K., and Lieber, M.R. (1999). Mechanistic basis for coding end sequence effects in the initiation of V(D)J recombination. Mol Cell Biol *19*, 8094-8102.

Chapter 3 : Targeted Chromatin Profiling Reveals Novel Enhancers in Ig H and Ig L Chain Loci

This paper has been published in Journal of Immunology:

Predeus, A. V, Gopalakrishnan, S., Huang, Y., Tang, J., Feeney, A.J., Oltz, E.M., and Artyomov, M.N. (2014). Targeted chromatin profiling reveals novel enhancers in Ig H and Ig L chain Loci. J. Immunol. 192, 1064–1070. http://www.ncbi.nlm.nih.gov/pubmed/24353267

Author contributions: A.V.P., S.G., E.M.O. and M.N.A. designed research; A.V.P., S.G., Y.H. performed research; All authors analyzed data; and A.V.P. and M.N.A. wrote the paper with input from E.M.O.

3.1 Abstract

The assembly and expression of mouse antigen receptor genes is controlled by a collection of *cis*-acting regulatory elements, including transcriptional promoters and enhancers. Although many powerful enhancers have been identified for immunoglobulin (*Ig*) and T cell receptor (*Tcr*) loci, it remained unclear whether additional regulatory elements remain undiscovered. Here, we use chromatin profiling of pro-B cells to define 38 epigenetic states in mouse antigen receptor loci, each of which reflects a distinct regulatory potential. One of these

chromatin states corresponds to known transcriptional enhancers and identifies a new set of candidate elements in all three *Ig* loci. Four of the candidates were subjected to functional assays and all four exhibit enhancer activity in B but not in T lineage cells. The new regulatory elements identified by focused chromatin profiling likely have important functions in the creation, refinement, and expression of *Ig* repertoires.

3.2 Introduction

Many of the strategies employed by developing lymphocytes to regulate gene expression share features with mechanisms that control the stepwise assembly of antigen receptor (AgR) loci (Osipovich and Oltz, 2010). Both processes require highly orchestrated interfacing between *cis*-regulatory elements, transcription factors, covalent modification of histones, changes in chromatin accessibility, and recruitment of machinery that drives transcription or recombination. In AgR loci, enhancer and promoter elements play crucial roles in modulating chromatin associated with variable (V), diversity (D), and joining (J) segments to control their recombination potential at each stage of lymphocyte development (Degner-Leisso and Feeney, 2010). Accordingly, most of the *cis*-elements associated with AgR loci are lineage- and stage-specific.

In addition to classical enhancers, recent studies identified a novel class of elements, termed “super-enhancers”, which are thought to regulate the expression of genes that serve as primary determinants of cell identity (Loven et al., 2013; Whyte et al., 2013). Super-enhancers are focal points for lineage-specifying transcription factors and for the ubiquitous mediator complex, which is required for activator-dependent gene expression. Moreover, super-enhancers are centered within unusually large stretches of activating histone modifications, such as

acetylation of histone H3 at the lysine 27 position (H3K27Ac). Three regions harboring super-enhancers have been identified within the *Igh* locus, including the classical enhancers, termed Eμ, and the 3' regulatory region (3'RR) (Whyte et al., 2013). However, the collection of *cis*-elements, known as the cistrome, which govern AgR gene assembly and expression during the early stages of lymphocyte development remains incomplete. Here, we identify novel enhancers within all three *Ig* loci, which exhibit activity specific for precursor B-lymphocytes, using focused computational analyses of publically available and new chromatin data.

3.3 Results

Epigenetic landscapes of AgR loci

Genome-wide patterns of histone modifications have been characterized for numerous cell types using chromatin immunoprecipitation (ChIP), followed by high-throughput sequencing (ChIP-Seq) (Barski et al., 2007). Bioinformatic integration of these data has emerged as a powerful method for the functional assignment of genomic regions, including the identification of promoters, enhancers, and microRNA sites (Abeel et al., 2009; Yip et al., 2012; Zhang et al., 2011). However, histone modifications can play additional, specialized roles at genetic loci. For example, the H3K4me3 modification is a hallmark of active promoters but, at AgR loci, this epigenetic mark also enhances binding to RAG-2, an essential component of the V(D)J recombination machinery (Matthews et al., 2007). The specialized roles of histone modifications at certain loci may produce unique epigenetic patterns that are impossible to unravel with supervised segmentation methods. To decipher such novel patterns, unsupervised algorithms have been used (Ernst and Kellis, 2012; Hoffman et al., 2012). For example, the epigenome of $CD4^+$ T lymphocytes was segmented into 53 functional states, including active and repressed

promoters, enhancers, and gene bodies (Ernst and Kellis, 2010). These approaches rely on statistical enrichment of specific combinations of chromatin marks throughout the genome to identify reproducible patterns. However, because genes represent the major organizational unit of the genome, the most robust patterns identified with current approaches correspond to promoter/gene bodies.

The unique segmented organization of AgR loci, coupled with genome-scale statistical analyses, could potentially mask AgR-specific patterns that are rare or non-existent in the remaining epigenome. To circumvent these potential complications in our search for new regulatory elements, we restricted combinatorial analysis of chromatin features to data covering only the seven AgR loci (*Igh*, *Igk*, *Igl*, *Tcra/d*, *Tcrb*, and *Trcg*). All data sets were obtained from purified pro-B cells harboring germline AgR loci (RAG-deficient), with the exceptions of Med1 and PU.1 association, which correspond to ChIP-Seq data from a transformed pro-B cell line (Whyte et al., 2013).

We first calculated the coverage of individual features at the seven AgR loci (histone modifications, factor binding, transcription) compared with the entire genome. As shown in Fig. 3.1, the epigenetic landscape of AgR loci is distinguished from the rest of the genome in several important respects. First, *Ig* and *Tcr* loci display a much lower density of the repressive H3K27me3 modification in pro-B cells relative to the entire genome. The dearth of this epigenetic mark suggests that Polycomb-mediated repression is less pronounced at AgR loci, even when a locus is silent for transcription/recombination. Second, the density of H3K36me3, a modification associated with transcriptional elongation, as well as transcripts themselves (RNA-Seq), are substantially decreased in AgR loci relative to the whole genome. This finding likely reflects the predominance of gene segments, rather than conventionally expressed genes in AgR

loci, as well as the limited amounts of transcription arising from the four *Tcr* loci in pro-B cells. Third, despite lower overall transcription, signals for the mediator component, Med1, and the transcription factors PU.1 and E2A are increased several-fold relative to the entire epigenome, suggesting a higher density of regulatory sites, potentially corresponding to enhancers. Finally, transcription factors c-Myc and EBF, which have important functions in pro-B cells, show substantially lower peak densities compared to the whole genome. This implies that the binding sites for these factors are mostly located outside of AgR loci. Overall, these initial analyses indicate that the distribution of important chromatin features in AgR loci differs substantially from the remainder of the genome. Therefore, identification of novel regulatory regions within AgR loci will benefit from a more focused computational analysis of chromatin states tailored to these regions.

Chromatin profiling of AgR loci in pro-B cells

The complex structure of *Ig* and *Tcr* loci requires advanced computational analysis to identify major AgR-specific chromatin patterns. For this purpose, we applied the ChromHMM algorithm (Ernst and Kellis, 2012), focusing on only the seven AgR loci. The resulting chromatin states may then be used to identify active and poised regulatory elements in an unbiased manner. Briefly, ChromHMM utilizes a hidden Markov model that captures two types of information -- the co-occurrence frequency of individual features at either the same location (emission probabilities) or at adjacent locations (transition probabilities) -- yielding patterns of chromatin features defined as characteristic states.

In total, we considered 19 distinct chromatin features in pro-B cells for this analysis (Fig. 3.2A), derived from published or new data sets, including histone modifications, key

transcription factors, nucleosome density, and transcription (Supplemental Table S3.1). Using ChromHMM, we compared individual emission probabilities for models in which the combinatorial number of states was varied from 20 to 40 and found that 38 states optimally described the epigenetic landscape of AgR loci. A higher model dimensionality produced redundancies, whereas distinct states, corresponding to active or poised regulatory elements, were merged when fewer than 38 were considered. Each state in the model corresponds to either a single feature or combinations of features, yielding an unbiased description of the AgR epigenetic landscape. A full list of states in the optimized model can be found in Supplemental Table S3.3. The probabilistic relationship between chromatin features and an individual state is important to note. For example, state 13 is defined by simultaneous presence of H3K36me3, H3Ac, DNAse, E2A, Med1 and some other marks, all with probabilities of nearly 1 (dark blue, Fig. 3.2A), indicating that all state 13 regions have these chromatin features. However, the probability of observing p300 in state 13 is intermediate (0.4, light blue), reflecting the fact that only some state 13 regions associate with p300 (e.g. Eμ), whereas others do not.

For the 38-state model (Fig. 3.2A), AgR chromatin can be divided broadly into 3 non-redundant categories. The first category includes twelve chromatin states that are defined by the presence of a single feature, such as H3K4me1, H3K27me3 or Pax5, suggesting a limited regulatory potential for these regions in pro-B cells (states 15, 29, and 34, respectively). A second category corresponds to states associated with only two or three chromatin features, which may reflect a partially active, or poised, configuration (e.g., states 1, 2, 20, and 26). Finally, six of the 38 chromatin states show strong enrichment for multiple histone modifications or other features of active chromatin (states 3, 4, 5, 7, 8, and 13). Regions assigned to these chromatin states likely harbor active regulatory elements since they are also nucleosome poor

(DHS peaks) and have other modifications that characterize promoters or enhancers (H3K4me3, H3ac etc). Notably, state 4 is characterized by its robust enrichment for 10 chromatin features (hereafter, enrichment indicates that the probability for a feature is >0.5), including association with the transcription factors p300, PU.1, and Med1. Given these characteristics, regions within AgR loci categorized as state 4 likely encompass *cis* elements with a high regulatory potential. Indeed, the chromatin states most highly enriched for activation features (states 3, 4, 5, and 13) are predominantly localized to Ig loci, particularly to *Igh*, which are more active (or poised) in pro-B cells compared with *Tcr* (Fig. 3.2B). Together, focused epigenetic analysis of AgR loci defines a unique set of chromatin states, some of which likely reflect functionality in the context of gene regulation and recombination.

Chromatin state functions

To garner functional insights, we first assessed whether classes of known AgR elements segregate into different chromatin states. As shown in Fig. 3.2C, we parsed the AgR loci into seven functional categories corresponding to the following annotated regions (± 500 bp): (i) *Igh* V segments (including upstream promoters), (ii) *Igk* V segments plus promoters, (iii) all D segments (*Ig* and *Tcr*), (iv) all J segments, (v) all constant regions, (vi) all known enhancers, and (vii) Pax5-activated intergenic repeat (PAIR) elements, a set of promoters that direct anti-sense transcription within specific regions of the *Igh* V cluster (Ebert et al., 2011; Verma-Gaur et al., 2012). Strikingly, most of the annotated AgR regions segregate from one another into a small number of individual chromatin states, likely reflecting the relationships between epigenetic features and their functionality. For example, *Igh* V regions, whose associated promoters exhibit varying degrees of activity in pro-B cells (Choi et al., 2013), belong to five different chromatin

states. Conversely, *Igk* V regions belong to only two states (21 and 22), distinct from those of *Igh* Vs, which display a highly restricted set of chromatin features, presumably reflecting their poised status in pro-B cells (H3K4me2 and PU.1). Moreover, most of the PAIR anti-sense promoters belong to chromatin state 3, which recapitulates known features of these regulatory elements, including their simultaneous association with Pax5, CTCF, and Rad21 (Figs. 2A and 2C). Most notably, eight of the twelve known AgR enhancers belong to chromatin state 4, which is most enriched for activating features (Fig. 3.3 and see below). One exception, Eμ, belongs to state 13 (Fig. 3.3A), likely because of its dual function as a strong promoter. A second exception, Eγ4, should be excluded from consideration since its epigenetic profile is masked by a proximal gene (*Stard3nl*) that is highly expressed in pro-B cells (Fig. 3.3F).

Notwithstanding, the vast majority of known AgR enhancers, whether active (3'RR, Fig. 3.5A) or inactive in pro-B cells (Eβ, Eα, Eγ2, 3'Eκ, Eλ31, Eλ24, Figs. 3 and 4), were assigned to state 4 using this unbiased analysis. Based on the aforementioned characteristics, chromatin state 4 provides a focused set of candidates for novel regulatory elements in the AgR cistrome. Overall, state 4 is limited to 41 segments, spanning 42 kb of the 9.2 Mb that encompasses all AgR loci (~0.4%). The priority status of state 4 as an identifier of enhancers is further supported by its enrichment for the p300 histone acetyltransferase, which is considered to be a general feature of enhancers, as well as its enrichment in three key transcription factors for pro-B cell gene expression programs (PU.1, E2A, and Pax5). Accordingly, most of the putative *cis*-elements belonging to state 4 are situated in active (46% in *Igh*) or poised loci (36% in *Igk* and *Igl*), while only 17% of such regions are located in the four *Tcr* loci (Fig 2B). Additionally, most of the state 4 regions are well-conserved at the level of DNA sequence (average conservation score is 0.682 for 4, see Methods and Supporting Tables S3.3 and S3.4). We conclude that

chromatin state four, which encompasses the most of the known AgR enhancers whether active, poised, or inactive in pro-B cells, will provide a rich source of candidate elements for functional analyses.

Characterization of novel enhancers in Ig L chain loci

Leveraging the predictive power of our AgR chromatin analyses, we selected three state 4 regions from *Igk or Igl* for functional assays. These light chain loci exhibit only modest transcriptional activity in primary pro-B cells and mostly likely reside in a "poised" chromatin configuration (Mercer et al., 2011). We first tested a state 4 element, situated in the large cluster of Vκ gene segments, for potential enhancer function (Fig. 3.4A, κRE1). Expression of luciferase reporters harboring the SV40 promoter was robustly augmented (7-fold) in a pro-B cell line (63-12) upon inclusion of κRE1 (Fig. 3.4C). In contrast, this region was devoid of enhancer activity when tested in pro-T or plasma cell lines (P5424 and J558L, respectively). Thus, the new *Igk cis*-element is a stage- and cell type-specific enhancer, suggesting a role in controlling the recombination potential of some mouse Vκ gene segments.

The mouse *Igl* locus has two highly conserved enhancers, termed Eλ13 and Eλ24, located distal to each of the VλJλ cassettes (Hagman et al., 1990). In addition, we identified two state 4 regions lying even more distal to the VλJλ cassettes (Fig. 3.4B, λRE1 and λRE3), which are highly conserved (1.054 and 0.706, respectively). These regions may represent "shadow" enhancers, which are suspected to serve as booster or redundancy elements for the regulation of many genes (Hobert, 2010). Indeed, in conjunction with the Vλ1 promoter, each of these regions

augment reporter gene expression in pro-B and plasma cells, but not in a mouse pro-T cell line (Fig. 3.4D). In the J558L plasmacytoma, λRE1 also boosts the function of its nearby enhancer, Eλ31, in an additive manner. As a control, the λRE2 region (Fig. 3.4B), which associates with PU.1 but identifies with chromatin states 20 and 21, fails to augment reporter gene expression in either pro-B or plasma cells (Fig. 3.4D). Taken together, assignment as chromatin state 4 accurately predicts the location of novel enhancers in *Ig* L chain loci.

Characterization of a super-enhancer in Igh

Transcription, V(D)J recombination, and *Igh* class switching are controlled by a set of enhancers and promoters, most of which have presumably been uncovered (Medvedovic et al., 2013). Our chromatin analysis assigned the two classical *Igh* enhancers as states 4 (3'RR) and 13 (Eμ) (Fig. 3.5A). These enhancers have distinct, but important functions in the B cell lineage, including transcription and recombination of adjacent DHJH gene segments (Eμ) and control of class-switch recombination (Eμ and 3'RR) (Perlot and Alt, 2008). A third stretch of *Igh*, embedded between Cγ1 and Cγ2b, was recently described as a super-enhancer (Whyte et al., 2013). Young and colleagues define super-enhancers using several parameters, including an exaggerated intensity of Med1 and PU.1 binding relative to other ChIP-Seq peaks, implicating these regions as key regulatory elements controlling cell identity genes (Whyte et al., 2013). To identify super-enhancers, the authors find regions of overlap for "master" transcription factors, such as PAX5 and PU.1 in pro-B cells, which also co-localize with the most intense and broadest peaks for the general transcription factor, Med1. Accordingly, we performed an unbiased analysis of Med1 distributions using SICER (Zang et al., 2009), which allows accurate peak-calling for broad chromatin features (see Materials and Methods). We discovered that the 3'RR,

Eμ, and the new super-enhancer, heretofore called Igh-SE, all appear as outliers in both width and read density for Med1, when compared with the entire epigenome (Fig. 3.5B).

Importantly, our focused AgR chromatin analysis splits the Igh-SE into two active regions that belong to states 4 and 5 (Fig. 3.5A, hRE1 and hRE2, respectively). As shown in Fig. 3.5C, only the hRE1 region functions as an enhancer in pro-B cells when monitored by luciferase reporters, but is devoid of enhancer activity in pro-T or plasma cell lines. The other region, hRE2, belongs to state 5 and likely corresponds to the Cγ2b germline promoter, which is active in pro-B cells based on its enrichment for H3K4me3 and the presence of sterile Iγ2b transcripts (Fig. 3.5A). The new hRE1 enhancer region is highly conserved (0.723) and interacts physically with other *Igh* regulatory elements (Medvedovic et al., 2013), strongly suggesting an important, but unknown function during the early stages of B cell development.

3.4 Discussion

We have used tailored computational approaches to assign chromatin states throughout all seven AgR loci in pro-B cells. Although the functional significance of many chromatin states remains to be defined, state 4 was found to accurately predict sites corresponding to AgR regulatory regions, both known and novel. The set of potential regulatory regions identified by state 4 also includes AgR enhancers that are inactive or only poised in pro-B cells (e.g., Eβ and Eλ24, respectively), broadening the scope of this chromatin-guided approach for enhancer discovery. Indeed, all state 4 regions tested in this study (4/4), which were derived from each of the three *Ig* loci, have enhancer activity in pro-B cells.

Strikingly, the only tested region that was found to be inactive, hRE2, was assigned to a separate chromatin state that is enriched for *Igh* V region promoters (state 5). This region likely corresponds to the germline Cγ2b promoter located near the hRE1 enhancer. We suspect that hRE1 plays a role in stabilizing the active conformation of *Igh* required for V(D)J recombination (Medvedovic et al., 2013) or in class-switch recombination (CSR), a process that occurs in both precursor and mature B cells (Han et al., 2007). The primary enhancer region directing CSR in activated B cells is thought to be the 3'RR. However, deletion of 3'RR abrogates recombination to all *Igh* isotypes except Cγ1, the constant region lying most proximal to the hRE1 enhancer (Vincent-Fabert et al., 2010). As such, our chromatin state analysis of pro-B cells provides at least four new enhancer elements, including hRE1, which can now be studied in vivo for their roles in *Ig* gene assembly, expression, isotype switching, and somatic hypermutation.

In summary, we have developed an unbiased epigenome-based approach to define the regulomes of AgR and other complex loci, such as those encoding NK cell receptors or MHC molecules (Shiina et al., 2004). While our functional validation of new enhancers focused on regions belonging to chromatin state 4, other states may also harbor important regulatory elements. These include state 5, which was enriched for promoters, and state 13, which spans Eμ and a flanking portion of the 3'RR. Future validations, including targeted disruption of these elements, will produce a more complete picture of AgR regulomes in the context of lymphocyte development and activation.

3.5 Materials and Methods

Data collection and processing. We considered 16 different epigenetic modifications that can be classified into four groups: 1) histone modifications (H3K4me1, H3K4me2, H3K4me3,

H3K27ac, H3K27me3, H3K36me3, H3K9ac/K14ac), 2) key transcription factors (p300, PU.1, Med1, c-Myc, Rad21, CTCF, EBF, E2A, Pax5), 3) nucleosome-poor, transcribed regions (DNase I hypersensitivity (DHS) and RNA Pol II occupancy), and 4) mature transcriptional signal from RNA-Seq experiments. Fourteen genome-wide experiments were available in public databases. For RNA Pol II and H3K27ac, new chromatin immunoprecipitation (ChIP) analyses were performed on a custom-made microarray covering all AgR loci (ChIP-Chip, see below). Supplemental Table S3.1 summarizes the sources of all experimental data.

All ChIP-Seq and DHS experiments were processed starting from SRA files. The binary SRA archives were converted into FASTQ files using the SRA toolkit, then aligned with Bowtie (Langmead et al., 2009) (version 0.12.7) using "-m 1 -v3 --best --strata" options. The resulting alignment SAM file was converted into read BED files using BEDTools. RNA-Seq data were aligned with TopHat (Trapnell et al., 2009) (version 1.4.1.1) using "--prefilter-multihits --max-multihits 15 --segment-length 20" options, and GenBank annotated mRNAs as an alignment reference (--GTF option).

Peak calling. We applied the SICER (Zang et al., 2009) (v1.1) algorithm to reads BED files and call peaks for all ChIP-Seq and DHS experiments. We used settings for narrow peaks (200 bp window size, 200 bp gap size, and FDR of 0.01) in all cases except for H3K27me3 and H3K36me3, which have broad signal distributions (200 bp window size, 600 bp gap size, and 0.01 FDR). Peak identification for RNA-Seq was performed by transcriptome assembly with Cufflinks (Trapnell et al., 2012) using no reference transcriptome, and exons of assembled transcripts with FPKM >2 were considered as peaks. For ChIP-Chip experiments, peak calling was performed with MA2C (Song et al., 2007) using a p-value of 0.01.

Genome segmentations. BED files obtained after peak calling were binarized using BEDTools (Quinlan and Hall, 2010). Genome-wide files were prepared with 200 bp windows and the overlaps of peak BED files and window files were calculated. If overlap constituted more than 50%, the bin was assigned 1. The exact regions of mouse genome (mm9 assembly) that were used for the analysis of AgR loci: chr6: 40838000 - 40845000, chr6: 40986000 - 41250000, chr6: 41476000 - 41555000 (*Tcrb*), chr13 : 19245000 - 19449000 (*Tcrg*), chr14 : 52962000 - 54855000 (*Tcra/d*), chr12 : 114435000 - 117280000 (*Igh*), chr6 : 67490000 - 70715000 (*Igk*), chr16 : 18971000 - 19285000 (*Igl*). Values for the genome outside of AgR boundaries were automatically set to 0, thus excluding all conventional genes from the segmentation. The resulting binarized input was then used in ChromHMM segmentation software (v1.10) (Ernst and Kellis, 2012) to generate hidden Markov models with the number of states ranging from 20 to 40, generating emission and transition probabilities, as well as segmentation BED files and HTML output. Corresponding BED files are available online at https://artyomovlab.wustl.edu/publications/supp_materials/AgR_2013/ .

Conservation analysis. Individual states were overlapped with *phyloP30WayPlacental* track from UCSC table browser (downloadable as a WIG file; a complete description of how the conservation score was generated is provided at http://hgdownload-test.cse.ucsc.edu/goldenPath/mm9/multiz30way/ multiz30way.html), and maximum conservation score for each interval was obtained using an in-house script by picking the highest value within each genomic interval. After this, the average maximum conservation score was calculated for each state by summing individual scores and dividing them by the number of intervals in the state.

ChIP-Chip experiments. Pro-B cells from RAG-deficient mice (C57BL/6, 4-6 weeks) were purified using MACS in conjunction with CD19 microbeads (Miltenyi Biotec, CA). ChIP experiments for H3K27ac and RNA Pol II were performed as described (Gopalakrishnan et al., 2013) using the following antibodies: H3K27ac (Abcam, ab4729) and Pol II (Abcam, ab5131). ChIP-DNA was purified using a Qiagen kit and subjected to whole genome amplification (Sigma, MO), labeled, and hybridized to custom Nimblegen microarrays according to the manufacturer's protocol by Mogene Inc., St. Louis. Total input DNA was used as the hybridization control.

Luciferase assays. The following cell lines were used: P5424 (RAG-$1^{-/-}$, $p53^{-/-}$ pro-T cell line), 63-12 (RAG-$2^{-/-}$ A-MuLV transformed pro-B cell line), and J558L (B myeloma cell line). All cell lines were cultured at 37°C with 5% CO_2 in RPMI 1640 supplemented with 10% FCS, 2mM L-Glutamine, 1% Penicillin/Streptomycin and 50uM b-mercaptoethanol. For transient transfection, cells were centrifuged at 100g for 5 minutes at room temperature, resuspended in serum-free RPMI 1640 at 10^7/ml. After this, $3x10^6$ cells were mixed with 3ug respective Firefly plasmid and 30ng Renilla control plasmid pRL-CMV (Promega), electroporated at 250V/960uF, transferred into 5ml pre-incubated media and cultured for 24 hours. Then Firefly and Renilla activities were measured using a dual assay kit, and the fold changes were calculated following the technical manual (Promega E2920).

Candidate regulatory elements were amplified using PCR. A full list of cloning primers is provided in Supplemental Table S3.2. The *Igl* enhancers Eλ24 and Eλ31 were amplified and cloned into the Bam HI site of pGL3 (Promega, WI). The regions of interest upstream of canonical enhancers, denoted λRE1, λRE2, and λRE3, were cloned individually in the Sal I site of pGL3. The Vλ1 promoter was cloned into the Xho I/Hind III sites of the Igλ enhancer-

containing luciferase constructs. The hRE1, hRE2, and κRE1 regions were amplified and blunt-end cloned into the Bam HI site of pGL3-Promoter, which contains an SV40 promoter. Cells were co-transfected with a Renilla expression plasmid for normalization and analyzed as described previously (Gopalakrishnan et al., 2013).

3.6 Figures

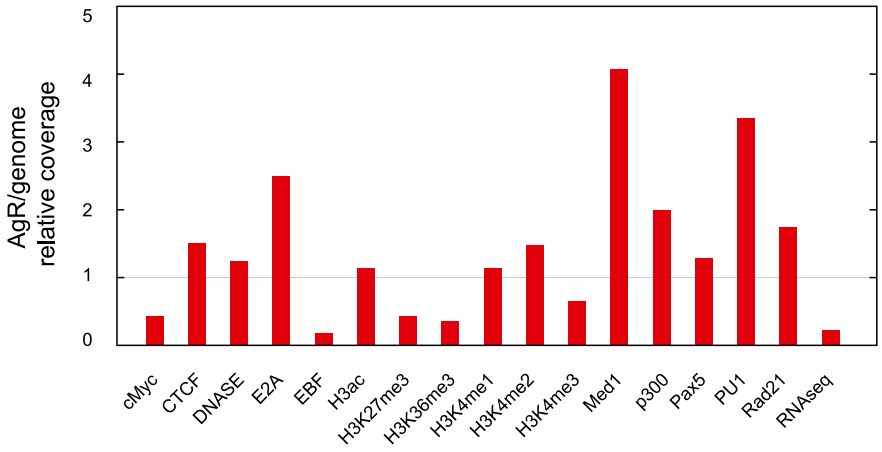


Figure 3.1: Unique epigenetic characteristics of mouse antigen receptor (AgR) loci
The y-axis represents the ratio of DNA space covered by a feature within AgR loci relative to the entire genome. A value of 1.0 corresponds to an equal distribution of that chromatin feature in AgR loci and the entire genome. Seventeen genome-wide ChIP-Seq, RNA-Seq, and DHS experiments were incorporated into the computational analyses.

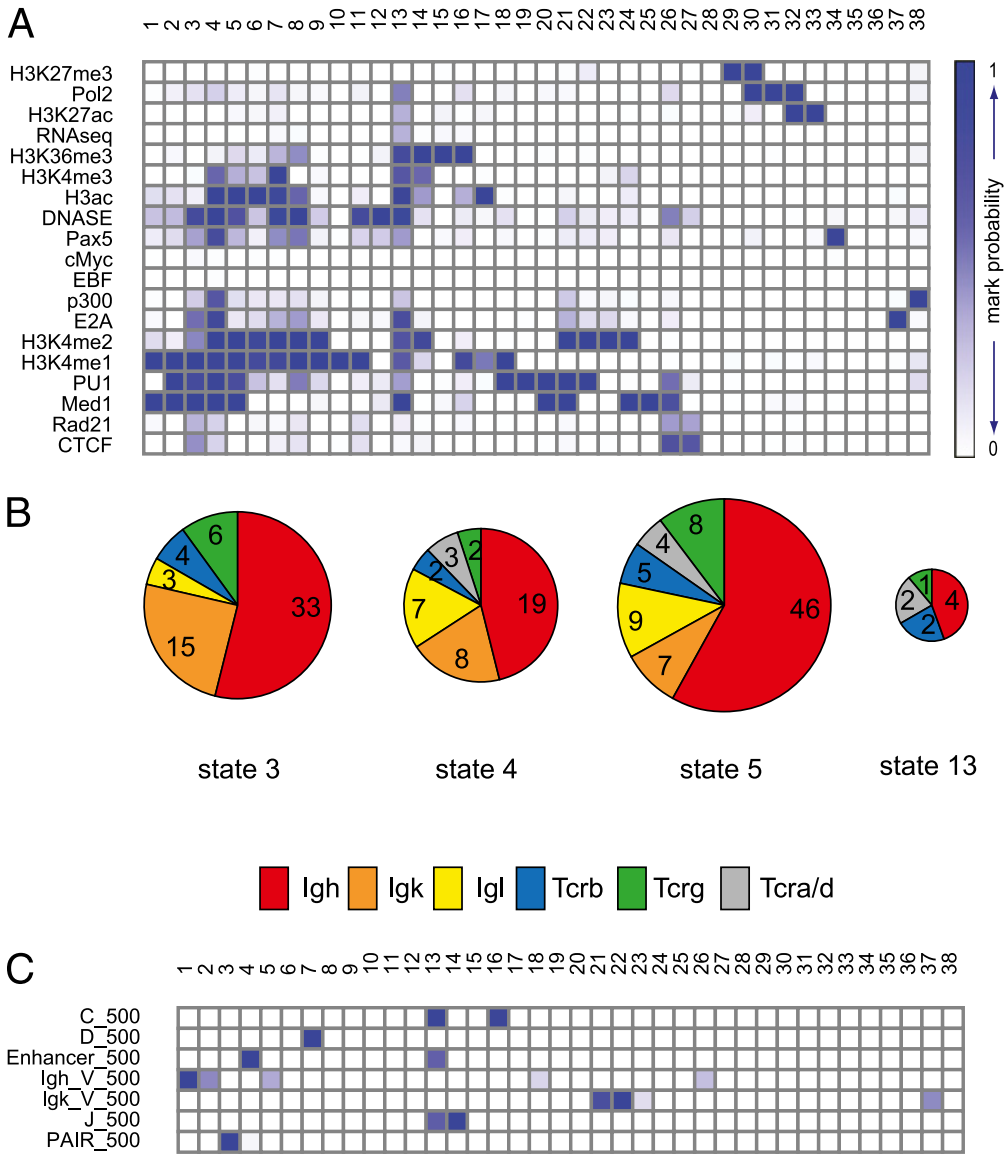


Figure 3.2: Unbiased characterization of the AgR epigenetic landscape
(A) The 38-state model of chromatin for AgR loci in pro-B cells. The hidden Markov model was based on the distribution of 19 chromatin features over all seven AgR loci: 17 genome-wide features shown on Fig. 3.1 were narrowed to AgR and two features, Pol II and H3K27Ac, profiled by ChIP-Chip of AgR loci. The shades of blue represent numerically determined emission probabilities that range from 0.0 to 1.0 and describe the precise "composition" of each state, or the probability to find a certain chromatin mark or transcription factor in the region defined as a particular state. A mark is considered "enriched" in a particular state if its emission probability in the model is >0.50. (B) Distribution across AgR loci for states with the highest regulatory potentials. Pie charts are scaled to the total numbers of regions corresponding to each state. Significant enrichment of these chromatin states is observed for *Ig* loci, suggesting lineage-specific activities for these regulatory states. (C) Enrichment of individual states for specific AgR elements. Each chromatin state was evaluated for its composition with respect to the indicated elements (± 500 bp from their annotated borders). Shades of blue correspond to hypergeometric probability of enrichment compared to random distribution across the entire collection of elements.

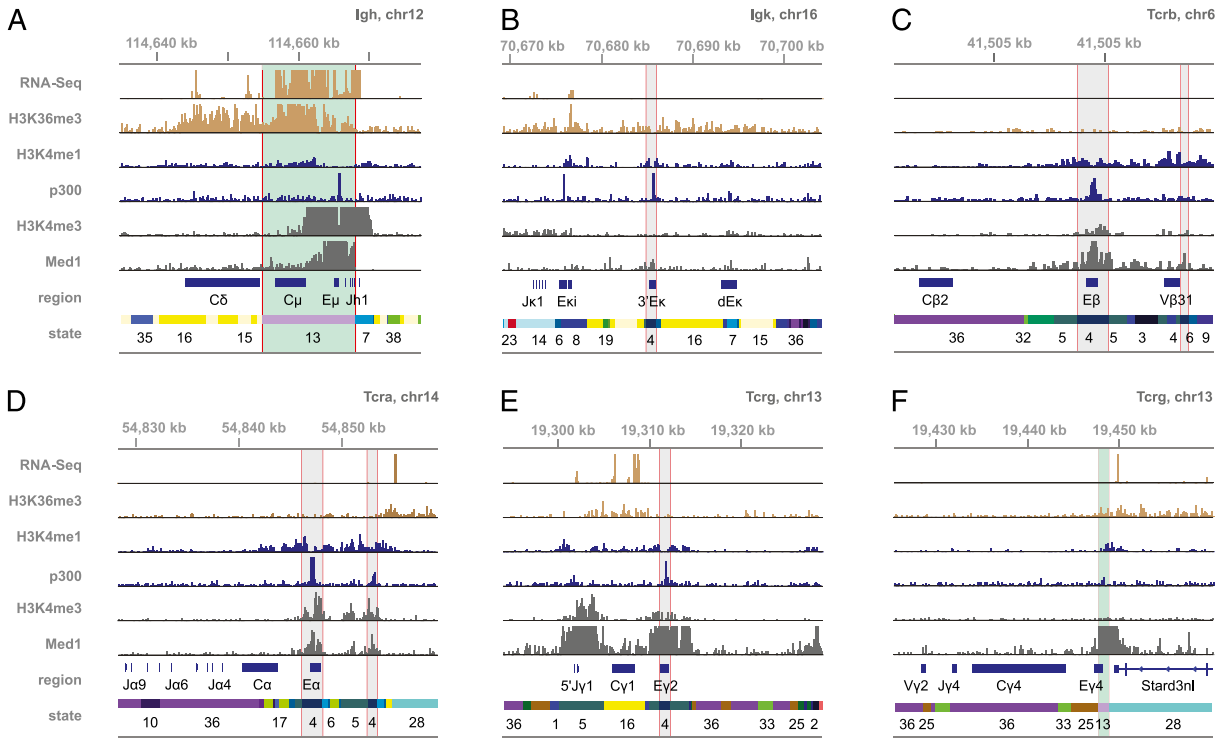
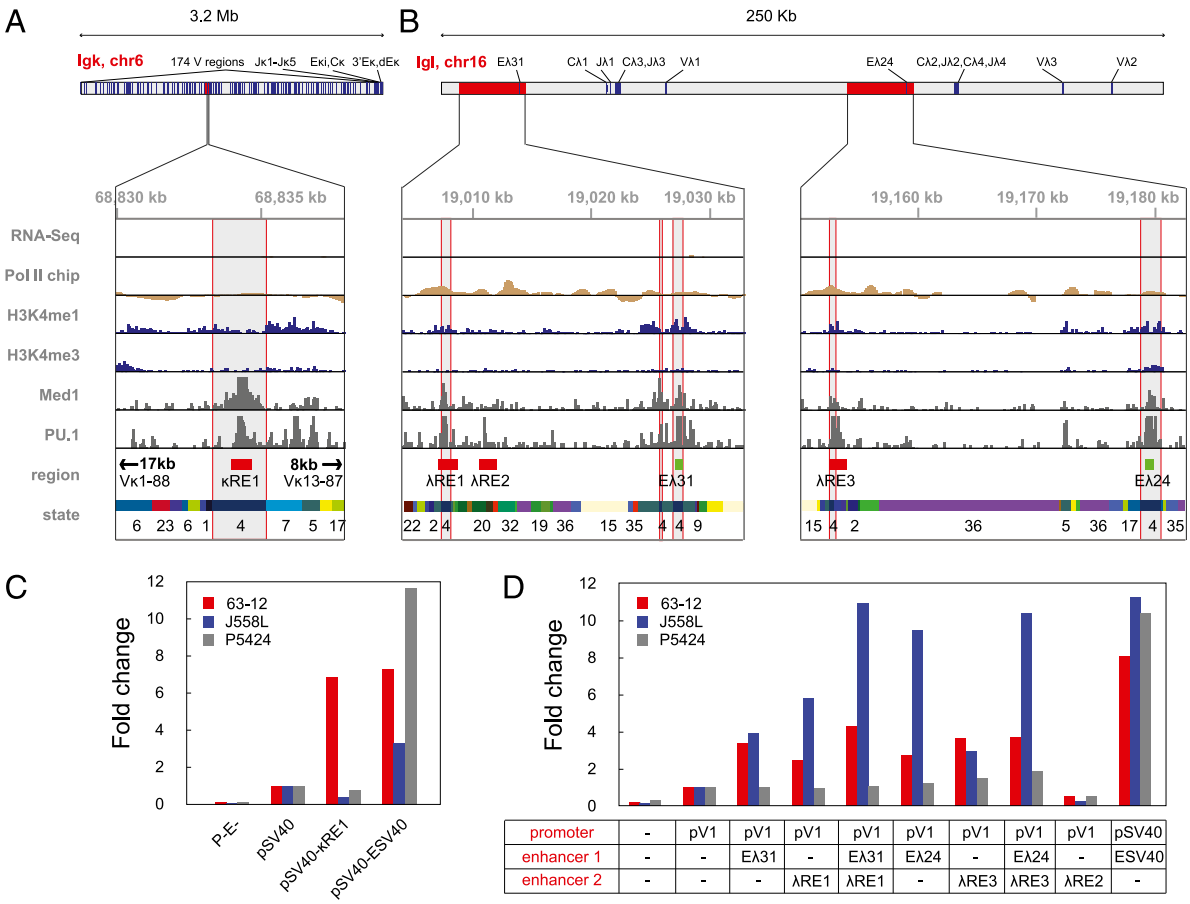


Figure 3.3: Chromatin states for selected regions of *Ig* and *Tcr* loci
Tracks for the indicated epigenetic features (ChIP-Seq) or transcription (RNA-Seq) as visualized in the IGV browser. Annotations for known elements and their corresponding chromatin states are shown in the bottom two tracks. Genomic coordinates are shown above the tracks (build mm9). State 13 is characteristic of actively transcribed elements (highlighted in light green) and harbors two enhancers, Eμ (A) and Eγ4 (F). State 4, which is characterized by a lack of transcription, the presence of activating chromatin marks, and binding by E2A, Pax5, PU.1, p300, and Med1, coincides with most known AgR enhancers, including 3'Eκ (B), Eβ (C), Eα (D), and Eγ2 (E). Regions identified as chromatin state 4 are highlighted in gray.



promoter	-	pV1	pV1	pV1	pV1	pV1	pV1	pV1	pV1	pSV40
enhancer 1	-	-	Eλ31	-	Eλ31	Eλ24	-	Eλ24	-	ESV40
enhancer 2	-	-	-	λRE1	λRE1	-	λRE3	λRE3	λRE2	-

Figure 3.4: Identification and functional validation of novel *Ig* L chain enhancers
(A, B) Tracks for the indicated chromatin features as visualized in the IGV browser (see Fig. 3.2). Regions identified as chromatin state 4 are highlighted in gray. Locations of candidate enhancer elements κRE1 in *Igk* (A), λRE1, λRE2 and λRE3 (B) are shown as red blocks. The known Eλ enhancer elements are indicated as green boxes. (C) Luciferase data for κRE1 enhancer activity in lymphocytes. Reporter plasmids containing combinations of the SV40 promoter (pSV40), SV40 enhancer (ESV40), κRE1, or lacking control elements (P-E-) were tested in the following cell lines: 63-12 pro-B cells (red bars), J558L plasmacytoma (blue bars), and P5424 pro-T cells (gray bars). All data are normalized for transfection efficiency and presented relative to pSV40 activity, which is set to 1. Representative data from at least two biological replicates are shown for all luciferase data. (D) Luciferase data for the indicated combinations of regulatory elements as described in (C).

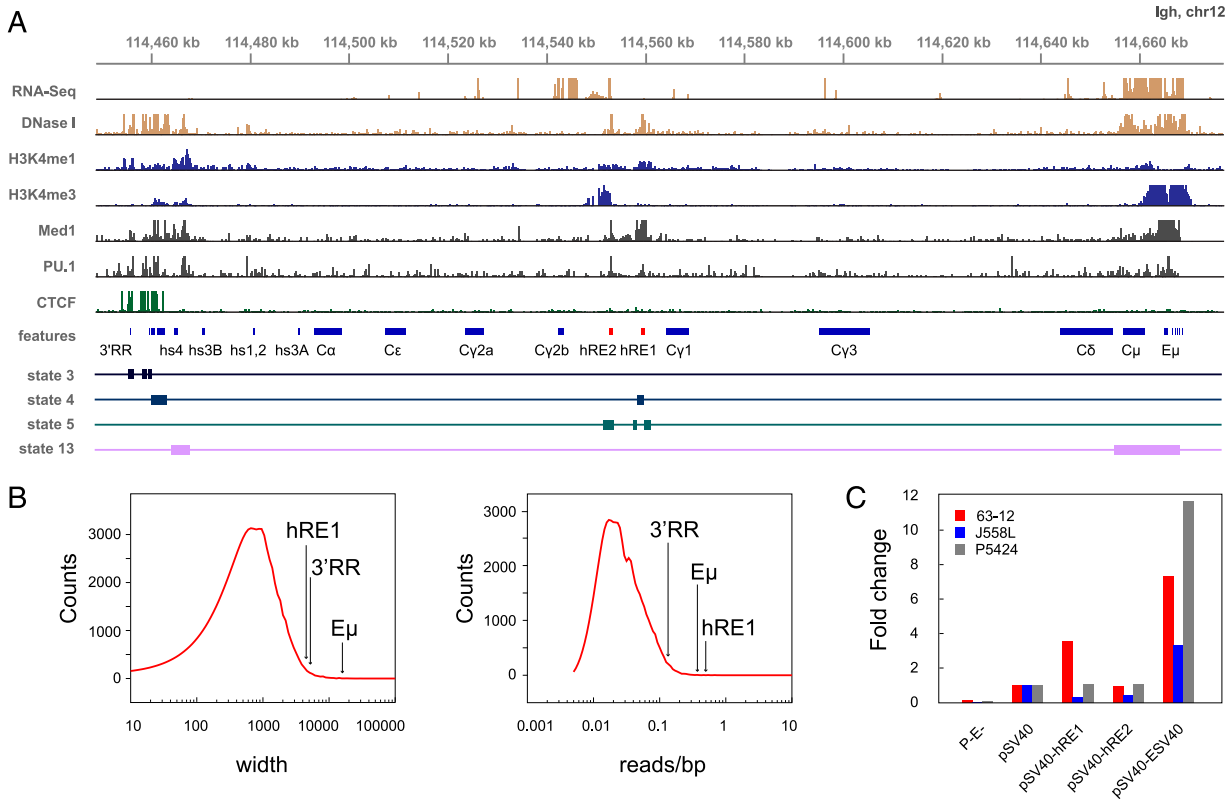


Figure 3.5: Functional definition of a novel *Igh* super-enhancer
(A) Tracks for the indicated chromatin features as visualized in the IGV browser (see Fig. 3.2). Chromatin states 3 (black), 4 (blue), 5 (green), and 13 (lavender) are shown in the bottom four tracks. The location of candidate enhancer elements hRE1 and hRE2 are highlighted as red boxes. (B) The distribution of Med1 peaks (identified using SICER) in pro-B cells by width (left panel) and read count-to-width ratio (right panel). Arrows indicate positions of Med1 peaks overlapping the three super-enhancer regions within *Igh*, highlighting their extreme breadth (left panel) and read densities (right panel). (C) Luciferase data for hRE1 and hRE2 as described in Fig. 3.3C.

3.7 Acknowledgements

We are grateful to Dr. Stephanie Kolar for helpful discussions. This research was supported by NIH Grants AI 079732, AI 081224 and CA 156690 (to E.M.O.) and AI 082918 (to A.J.F.).

3.8 Supplemental Tables

Supplemental Table 3.1: All datasets used in the analysis
All samples were processed de-novo from SRA files as described in Methods section.

Feature	PI	Exp. type	Cell source	Series (GSE)	Accession number (GSM)	Additional notes
H3K4me2	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987804	
H3K36me3	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987807	
p300	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987808	
c-Myc	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987810	
CTCF	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987805	
Rad21	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987806	
Input	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE40173	GSM987812	This input was used for all experiments in GSE40173 superset
E2A	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE21978	GSM546523	
EBF	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE21978	GSM546524	
Input2	C. Murre	ChIP-seq	Rag1(-/-) C57BL/6 mouse	GSE21978	GSM546540	This input was used for E2A and EBF experiments
H3K27ac	E. Oltz	ChIP-chip	Rag1(-/-) C57BL/6 mouse	Chip-chip	Available online	http://artyomovlab.wustl.edu/publications/supp_materials/AgR_2013/
Pol II	E. Oltz	ChIP-chip	Rag1(-/-) C57BL/6 mouse	Chip-chip	Available online	http://artyomovlab.wustl.edu/publications/supp_materials/AgR_2013/
H3K4me1	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932934	exp. 8666
H3K4me3	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932939 - 42	exp. 8110, 8115
H3K27me3	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932947 - 51	exp. 8111, 8116
H3K9ac/K14ac	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932943 - 46	exp. 8108, 8113
DNase I hypersensitivity	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932968 - 69	exp. 8439
RNA-Seq	M. Busslinger	RNA-seq	Rag2(-/-) C57BL/6 mouse	GSE38046	GSM932910 - 13	exp. 8275
Input	M. Busslinger	ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE35857	GSM876635 -42	exp. 8091.5, 8091.6, 8112.1, 8112.6, 8123.2, 8123.3, 8149.8.30222AAXX, 8149.8.301DTAAXX
Pax5	M. Busslinger	Bio-ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE35857	GSM932921 - 23	exp. 8093
Bio-Chip-Seq Input	M. Busslinger	Bio-ChIP-seq	Rag2(-/-) C57BL/6 mouse	GSE35857	GSM932932 - 33	exp. 8095
Med1	R. Young	ChIP-seq	38B9 pro-B cell line, C57BL/6-129 mouse	GSE44288	GSM1038263	
Med1 Input	R. Young	ChIP-seq	38B9 pro-B cell line, C57BL/6-129 mouse	GSE44288	GSM1038264	
PU.1	R. Young	ChIP-seq	38B9 pro-B cell line, C57BL/6-129 mouse	GSE21614	GSM539538	
PU.1 Input	R. Young	ChIP-seq	38B9 pro-B cell line, C57BL/6-129 mouse	GSE44288	GSM1038265	

Supplemental Table 3.2: List of all primers used for cloning

Note that all primers include a restriction site. Coordinates are based on mm9 assembly of mouse genome. Bases selected in bold are added for technical purposes; the rest correspond to genomic regions of interest.

Primers	Sequences (5'-3')	Chr	Strand	Begin	End
pVλ1 F	**GACACTCGAG**GAGCTCTGTTCTTAGTAACA	16	-	19085581	19085749
pVλ1 R	**GATAAGCTT**AATATTGGTCAGCAGCAGGC	16	-	19085581	19085749
Eλ24 F	**TAAGGATCC**TCTCCTGAGATATTGCATAGGCCTGCCC	16	-	19179217	19179887
Eλ24 R	**CTTGGATCC**ACTCCTTTGTGCTCTGATAGCA	16	-	19179217	19179887
Eλ31 F	**TAAGGATCC**TCTCCTGAGATGTTACATAGGCCTGCCA	16	-	19152593	19154022
Eλ31 R	**GACGGATCC**ACTCCTTTGTGCTCTGATAACC	16	-	19152593	19154022
λRE1 F	**TAAGTCGAC**GACCTGGATTCAATTCACAGTACCT	16	-	19007160	19008717
λRE1 R	**GATGTCGAC**CTATTATAATCACCTAGGACACTGC	16	-	19007160	19008717
λRE2 F	**TAAGTCGAC**CCAATGGAGGCCAGAATGGGATAAC	16	-	19010593	19011993
λRE2 R	**GACGTCGAC**TAATGGTGACAGTGACATCCAAACT	16	-	19010593	19011993
λRE3 F	**GCTGTCGAC**CAATTCACACTATCCACGTAATAAG	16	-	19152593	19154022
λRE3 R	**TAAGTCGAC**GATGTCGGATTCAGGCCTGGTAA	16	-	19152593	19154022
hRE2 F	**CTAGGTCGAC**CTCTCCACCTGATTTGCTGCAC	12	-	114552827	114553398
hRE2 R	**TAGCGGATCC**TTCTCAGCTGACCACACTCACA	12	-	114552827	114553398
hRE1 F	**CTAGGTCGAC**TATGCTATACAGGATAAACTC	12	-	114559154	114559750
hRE1 R	**TAGCGGATCC**CATGGAAGATAAAAACTGTACA	12	-	114559154	114559750
κRE1 F	**TAGCAGATCT**TATTTAACTAGGTATGATTAT	6	+	68834022	68834718
κRE1 R	**CTAGGTCGAC**AACTATAAAGTCAGTGGATTTC	6	+	68834022	68834718

Supplemental Table 3.3: List of all states
Generated in 38-state model with 19 features, classified by number of marks that have emission probability of 0.5 and higher. States that are enriched in 5 or more marks are highlighted. States that are also enriched in Med1 and key transcription factors are shown in bold.

State	No. of intervals	Coverage, bp	Average conservation score	No. of enriched marks	Marks
10	182	112000	0.56202	1	H3K4me1
12	45	20000	0.40003	1	DNASE
15	52	120800	0.49497	1	H3K36me3
19	172	135600	0.37406	1	PU1
23	72	44400	0.48466	1	H3K4me2
25	256	212600	0.26356	1	Med1
29	76	284200	0.88749	1	H3K27me3
31	155	189200	0.32503	1	Pol2
33	93	108400	0.43606	1	H3K27ac
34	88	39600	0.21294	1	Pax5
37	15	8000	0.54671	1	E2A
38	10	13000	0.397	1	p300
1	95	62600	0.52599	2	Med1, H3K4me1
9	72	37000	0.64537	2	H3K4me2, H3K4me1
11	79	47200	0.76623	2	DNASE, H3K4me1
16	41	51200	0.60315	2	H3K36me3, H3K4me1
17	42	30400	0.53316	2	H3ac, H3K4me1
18	38	18400	0.47392	2	PU.1, H3K4me1
20	162	160800	0.29633	2	PU.1, Med1
22	28	21600	0.66138	2	PU.1, H3K4me2
24	31	21600	0.40012	2	H3K4me2, Med1
26	66	32800	0.56474	2	PU.1, Med1

30	7	11400	1.05072	2	H3K27me3, Pol2
32	24	29200	0.3418	2	Pol2, H3K27ac
2	128	91000	0.43376	3	H3K4me1, PU.1, Med1
6	58	31600	0.6256	3	H3ac, H3K4me2, H3K4me1
14	16	22400	0.73986	3	H3K36me3, H3K4me3, H3K4me2
21	34	25400	0.48801	3	H3K4me2, PU.1, Med1
3	**61**	**38800**	**0.60471**	**5+**	**DNASE, E2A, H3K4me1, PU.1, Med1**
4	**41**	**42400**	**0.68189**	**5+**	**DNASE, H3K4me3,H3K4me2, H3K4me1, H3ac, PU.1, Med1, E2A, Pax5, p300**
5	**79**	**71400**	**0.66883**	**5+**	**DNASE, H3K4me2, H3K4me1, PU.1, Med1**
7	30	29800	0.82031	5+	DNASE, H3K4me3, H3K4me2, H3K4me1, H3ac
8	33	22000	0.74818	5+	DNASE, H3K4me2, H3K4me1,H3ac, Pax5
13	**9**	**22800**	**1.21196**	**5+**	**DNASE, H3K36me3, H3K4me3,H3K4me2, H3K4me1, H3ac, E2A, Med1**
27	103	56800	0.54211	0	Not quite empty state, CTCF emission is just under 0.5
28	29	2.646E +09	1.8151	0	Global empty state neighbored by other empty state
35	114	128600	0.34177	0	Empty state that follows states 15,17 and 22
36	926	638040 0	0.54239	0	Empty state that follows states 10,12,23,27 and 34

Supplemental Table 3.4: All regions identified as states 4, 5, and 13
The conservation score for each interval was calculated as described in Methods section. Regions that were tested in luciferase assays and found to be active enhancers are highlighted in green. The region that represent known enhancers are highlighted in red.

Chr	Begin	End	State	Locus	Cons. Score	Annotation
chr12	114460200	114463400	4	*Igh*	0.605	hs7, hs6, hs5
chr12	114558600	114559800	4	*Igh*	0.723	**hRE1**

chr12	115420600	115422800	4	*Igh*	0.96	-
chr12	115511600	115512400	4	*Igh*	0	-
chr12	115578000	115578800	4	*Igh*	0	-
chr12	115732000	115732400	4	*Igh*	0	-
chr12	115783600	115784000	4	*Igh*	0	-
chr12	115847000	115849200	4	*Igh*	0.524	-
chr12	115853400	115854400	4	*Igh*	0.262	Igh-V-J558.9.99
chr12	116219800	116220800	4	*Igh*	0	-
chr12	116235400	116235800	4	*Igh*	0.678	-
chr12	116261600	116263600	4	*Igh*	0.831	Igh-V-3609.2pg.138
chr12	116403400	116404200	4	*Igh*	0.262	Igh-V-3609.5.147
chr12	116461000	116461600	4	*Igh*	0.262	-
chr12	116495800	116496600	4	*Igh*	0.597	Igh-V-3609.6pg.161
chr12	116500400	116501600	4	*Igh*	0.678	-
chr12	116642000	116642800	4	*Igh*	0.765803	Igh-V8-8-1*01
chr12	116645400	116645800	4	*Igh*	0.262	-
chr12	116757200	116757600	4	*Igh*	0.674	Igh-V-3609.10pg.167
chr13	19273800	19275200	4	*Tcrg*	1.227	-
chr13	19311200	19312400	4	*Tcrg*	1.153	**Eγ2**
chr14	53070400	53071400	4	*Tcra/d*	2.001	Olfr1909
chr14	54846200	54848200	4	*Tcra/d*	1.809	**Eα**
chr14	54852600	54853600	4	*Tcra/d*	1.68	-
chr16	19003000	19003600	4	*Igl*	0.516063	-
chr16	19007400	19008200	4	*Igl*	1.054	**λRE1**
chr16	19025800	19026000	4	*Igl*	0	-
chr16	19027000	19027800	4	*Igl*	0	**Eλ31**
chr16	19052800	19053400	4	*Igl*	0	-

chr16	19152600	19153000	4	*Igl*	0.706	**λRE3**
chr16	19178800	19180600	4	*Igl*	0.588	**Eλ24**
chr6	41503800	41505200	4	*Tcrb*	1.708	**Eβ**
chr6	41508400	41508800	4	*Tcrb*	1.81477	Trb-V31 promoter
chr6	68387200	68388200	4	*Igk*	1.097	Igk-V15-103
chr6	68431000	68432000	4	*Igk*	0.944	Igk-V15-101
chr6	68548000	68548600	4	*Igk*	0	-
chr6	68700000	68701200	4	*Igk*	0	-
chr6	68833400	68835200	4	*Igk*	0.597	**κRE1**
chr6	69990400	69991400	4	*Igk*	0.606	-
chr6	70659600	70660200	4	*Igk*	0.579	-
chr6	70685000	70686000	4	*Igk*	1.794	**3'Eκ**
chr12	114551600	114553600	5	*Igh*	0.7	**hRE2**
chr12	114557800	114558600	5	*Igh*	0.262	-
chr12	114559800	114561200	5	*Igh*	0.582	-
chr12	115232000	115233400	5	*Igh*	0.955	Igh-V-SM7.2.49
chr12	115297800	115298600	5	*Igh*	0.808	Igh-V-SM7.3.54
chr12	115331800	115332400	5	*Igh*	0.732	Igh-V-GAM3.8-1-57
chr12	115414600	115415400	5	*Igh*	0.582	Igh-V-SM7.4.63
chr12	115419000	115420600	5	*Igh*	0.831	-
chr12	115422800	115423200	5	*Igh*	0	-
chr12	115511400	115511600	5	*Igh*	0	-
chr12	115512400	115512600	5	*Igh*	0	-
chr12	115578800	115579400	5	*Igh*	1.012	-
chr12	115699600	115701000	5	*Igh*	0.597	Igh-V-3609.1.84
chr12	115732400	115733200	5	*Igh*	0	-
chr12	115784000	115785200	5	*Igh*	0	Igh-V10-4*01 promoter

chr12	115802400	115803200	5	*Igh*	0.835	Igh-V15.1.95
chr12	115835800	115836600	5	*Igh*	0.674	Igh-V-J558.7pg.97
chr12	115849200	115851000	5	*Igh*	0.554803	Igh-V-J558.8.98
chr12	115853200	115853400	5	*Igh*	0	-
chr12	115854400	115855000	5	*Igh*	0.606	proximal to Igh-V-J558.9.99
chr12	116235800	116236400	5	*Igh*	0.262	-
chr12	116241000	116242400	5	*Igh*	0.606	Igh-V-3609.2pg.138
chr12	116263600	116265200	5	*Igh*	0.262	proximal to Igh-V-3609.3.139
chr12	116402800	116403400	5	*Igh*	0	-
chr12	116404200	116404800	5	*Igh*	0.262	Igh-V-3609.5.147
chr12	116410400	116410800	5	*Igh*	0.249992	-
chr12	116431400	116432600	5	*Igh*	0.584	Igh-V-J558.54.148
chr12	116461600	116462400	5	*Igh*	0.262	-
chr12	116495200	116495800	5	*Igh*	0.262	proximal to Igh-V-3609.6pg.161
chr12	116499800	116500400	5	*Igh*	0.248268	-
chr12	116550000	116550800	5	*Igh*	0.734	Igh-V-J558.58.154
chr12	116642800	116644000	5	*Igh*	0.867	Igh-V8-8-1*01
chr12	116645000	116645400	5	*Igh*	0.262	-
chr12	116712800	116713000	5	*Igh*	0.262	-
chr12	116756800	116757200	5	*Igh*	0	-
chr12	116757600	116758000	5	*Igh*	0.674	Igh-V-3609.10pg.167
chr12	116765200	116766400	5	*Igh*	0	-
chr12	116805200	116805800	5	*Igh*	0.734	Igh-V-3609.11.169
chr12	116885400	116887000	5	*Igh*	1.232	Igh-V-3609.12.174
chr12	116930200	116930400	5	*Igh*	0.579	Igh-V-J558.73pg.175
chr12	117003600	117004000	5	*Igh*	0.867	Igh-V-3609.13pg.178
chr12	117008800	117010200	5	*Igh*	0.262	-

chr12	117028600	117029400	5	*Igh*	0.832	Igh-V-J558.76pg.179
chr12	117046200	117047200	5	*Igh*	0.868	Igh-V-3609.14pg.181
chr12	117100000	117100800	5	*Igh*	1.152	Igh-V-J558.80.186
chr12	117126400	117126600	5	*Igh*	1.037	Igh-V-J558.82.188
chr13	19273200	19273800	5	*Tcrg*	0.848465	-
chr13	19276400	19277600	5	*Tcrg*	1.625	Trg-V4
chr13	19281800	19282800	5	*Tcrg*	1.395	Trg-V6
chr13	19300200	19305200	5	*Tcrg*	1.618	Trg-J1
chr13	19310000	19311200	5	*Tcrg*	0.262	proximal to Eγ2
chr13	19312400	19314400	5	*Tcrg*	0.816622	proximal to Eγ2
chr13	19363600	19365400	5	*Tcrg*	1.051	-
chr13	19392600	19393400	5	*Tcrg*	0.951	-
chr14	53070200	53070400	5	*Tcra/d*	2.001	Olfr1909
chr14	53108600	53109000	5	*Tcra/d*	0.720724	-
chr14	54845600	54846200	5	*Tcra/d*	1.663	-
chr14	54850000	54852600	5	*Tcra/d*	1.422	-
chr16	19003600	19004000	5	*Igl*	0.795	-
chr16	19006800	19007400	5	*Igl*	1.054	-
chr16	19008200	19008400	5	*Igl*	0	-
chr16	19024000	19025800	5	*Igl*	0	-
chr16	19026000	19027000	5	*Igl*	0	-
chr16	19027800	19028800	5	*Igl*	0.641	proximal to Eλ31
chr16	19052000	19052800	5	*Igl*	0	-
chr16	19152200	19152600	5	*Igl*	0	-
chr16	19172200	19173000	5	*Igl*	0	-
chr6	40840400	40842200	5	*Tcrb*	1.717	Trb-V1
chr6	41502800	41503800	5	*Tcrb*	1.442	proximal to Eβ

chr6	41505200	41506000	5	*Tcrb*	1.708	proximal to Eβ
chr6	41507400	41507800	5	*Tcrb*	1.63294	Trb-V31
chr6	41508200	41508400	5	*Tcrb*	1.77354	Trb-V31
chr6	68547800	68548000	5	*Igk*	0.262	-
chr6	68630200	68630800	5	*Igk*	0.882	Igk-V10-95
chr6	68686400	68687400	5	*Igk*	0.79	Igk-V19-93
chr6	68836400	68837000	5	*Igk*	0	-
chr6	69989800	69990400	5	*Igk*	0.606	-
chr6	69991400	69992600	5	*Igk*	0.582	-
chr6	70684800	70685000	5	*Igk*	1.487	proximal to 3'Eκ
chr12	114464000	114467800	13	*Igh*	0.664	**hs4**
chr12	114655000	114668200	13	*Igh*	1.809	**Eμ + Cμ**
chr12	116497000	116497800	13	*Igh*	0.582	proximal to Igh-V-3609.6pg.161
chr12	117106800	117107400	13	*Igh*	0.808	Igh-V-J558.81.187
chr13	19447800	19449000	13	*Tcrg*	2.058	artifact - Stard3nl gene overflow
chr14	54179200	54180200	13	*Tcra/d*	1.418	Tra-V15-1/Trd-V6-1
chr14	54181400	54182000	13	*Tcra/d*	0.262	-
chr6	40996800	40997600	13	*Tcrb*	1.51262	Trb-V2 promoter
chr6	41554200	41555000	13	*Tcrb*	1.794	artifact - Ephb6 gene overflow

3.9 References

Abeel, T., Van de Peer, Y., and Saeys, Y. (2009). Toward a gold standard for promoter prediction evaluation. Bioinformatics *25*, i313-320.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Choi, N.M., Loguercio, S., Verma-Gaur, J., Degner, S.C., Torkamani, A., Su, A.I., Oltz, E.M., Artyomov, M., and Feeney, A.J. (2013). Deep sequencing of the murine igh repertoire reveals complex regulation of nonrandom v gene rearrangement frequencies. J Immunol *191*, 2393-2402.

Degner-Leisso, S.C., and Feeney, A.J. (2010). Epigenetic and 3-dimensional regulation of V(D)J rearrangement of immunoglobulin genes. Seminars in immunology *22*, 346-352.

Ebert, A., McManus, S., Tagoh, H., Medvedovic, J., Salvagiotto, G., Novatchkova, M., Tamir, I., Sommer, A., Jaritz, M., and Busslinger, M. (2011). The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. Immunity *34*, 175-187.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology *28*, 817-825.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature methods *9*, 215-216.

Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O.I., Verma-Gaur, J., Loguercio, S., Su, A.I., Feeney, A.J., Artyomov, M.N.*, et al.* (2013). Unifying model for molecular determinants of the preselection Vbeta repertoire. Proceedings of the National Academy of Sciences of the United States of America *110*, E3206-3215.

Hagman, J., Rudin, C.M., Haasch, D., Chaplin, D., and Storb, U. (1990). A novel enhancer in the immunoglobulin lambda locus is duplicated and functionally independent of NF kappa B. Genes & development *4*, 978-992.

Han, J.H., Akira, S., Calame, K., Beutler, B., Selsing, E., and Imanishi-Kari, T. (2007). Class switch recombination and somatic hypermutation in early mouse B cells are mediated by B cell and Toll-like receptors. Immunity *27*, 64-75.

Hobert, O. (2010). Gene regulation: enhancers stepping out of the shadow. Current biology : CB *20*, R697-699.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods *9*, 473-476.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.

Matthews, A.G., Kuo, A.J., Ramon-Maiques, S., Han, S., Champagne, K.S., Ivanov, D., Gallardo, M., Carney, D., Cheung, P., Ciccone, D.N.*, et al.* (2007). RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature *450*, 1106-1110.

Medvedovic, J., Ebert, A., Tagoh, H., Tamir, I.M., Schwickert, T.A., Novatchkova, M., Sun, Q., Huis In 't Veld, P.J., Guo, C., Yoon, H.S.*, et al.* (2013). Flexible Long-Range Loops in the VH Gene Region of the Igh Locus Facilitate the Generation of a Diverse Antibody Repertoire. Immunity *39*, 229-244.

Mercer, E.M., Lin, Y.C., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C.K., and Murre, C. (2011). Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. Immunity *35*, 413-425.

Osipovich, O., and Oltz, E.M. (2010). Regulation of antigen receptor gene assembly by genetic-epigenetic crosstalk. Seminars in immunology *22*, 313-322.

Perlot, T., and Alt, F.W. (2008). Cis-regulatory elements and epigenetic changes control genomic rearrangements of the IgH locus. Advances in immunology *99*, 1-32.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Shiina, T., Inoko, H., and Kulski, J.K. (2004). An update of the HLA genomic region, locus information and disease associations: 2004. Tissue antigens *64*, 631-649.

Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R., and Liu, X.S. (2007). Model-based analysis of two-color arrays (MA2C). Genome biology *8*, R178.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Verma-Gaur, J., Torkamani, A., Schaffer, L., Head, S.R., Schork, N.J., and Feeney, A.J. (2012). Noncoding transcription within the Igh distal V(H) region at PAIR elements affects the 3D structure of the Igh locus in pro-B cells. Proceedings of the National Academy of Sciences of the United States of America *109*, 17004-17009.

Vincent-Fabert, C., Fiancette, R., Pinaud, E., Truffinet, V., Cogne, N., Cogne, M., and Denizot, Y. (2010). Genomic deletion of the whole IgH 3' regulatory region (hs3a, hs1,2, hs3b, and hs4) dramatically affects class switch recombination and Ig secretion to all isotypes. Blood *116*, 1895-1898.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M.*, et al.* (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome biology *13*, R48.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zhang, S., Li, Q., Liu, J., and Zhou, X.J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics *27*, i401-409.

Chapter 4 : *cis*-Regulatory Circuits Regulating *NEK6* Kinase Overexpression in Transformed B Cells Are Super-Enhancer Independent

This paper has been published in Cell Reports:

Huang, Y., Koues, O. I., Zhao, J., Liu, R., Pyfrom, S. C., Payton, J. E., Oltz, E. M. (2017). *cis*-Regulatory Circuits Regulating NEK6 Kinase Overexpression in Transformed B Cells Are Super-Enhancer Independent. Cell Reports 18, 2918-2931.

https://www.ncbi.nlm.nih.gov/pubmed/28329684

Author contributions: E.M.O., J.E.P., Y.H. and O.I.K. conceptualized the study and designed experiments. E.M.O. supervised all aspects of the project. Experiments and data analyses were performed by all authors. The manuscript was written by E.M.O. and Y.H. with input from J.E.P.

4.1 Abstract

Alterations in distal regulatory elements that control gene expression underlie many diseases, including cancer. Epigenomic analyses of normal and diseased cells have produced correlative predictions for connections between dysregulated enhancers and target genes involved in pathogenesis. However, with few exceptions, these predicted *cis*-regulatory circuits

remain untested. Here, we dissect *cis*-regulatory circuits that lead to overexpression of *NEK6,* a mitosis-associated kinase, in human B cell lymphoma. We find that only a minor subset of predicted enhancers is required for *NEK6* expression. Indeed, an annotated super-enhancer is dispensable for NEK6 overexpression and for maintaining the architecture of a B cell-specific regulatory hub. A CTCF cluster serves as a chromatin and architectural boundary to block communication of the *NEK6* regulatory hub with neighboring genes. Our findings emphasize that validation of predicted *cis*-regulatory circuits and super-enhancers is needed to prioritize transcriptional control elements as therapeutic targets.

4.2 Introduction

Cell identity and function rely on stringently controlled programs of gene expression, perturbations of which underlie diseases, including autoimmunity and cancer. Genome-wide association studies have revealed that most pathogenic changes in gene expression are linked to variants in regulatory elements rather than coding sequences (Maurano *et al.*, 2012). A dissection of *cis*-regulatory circuits controlling transcriptomes in normal and diseased cells remains an important objective. Most *cis*-regulatory circuits are composed of gene-proximal promoters and distal enhancers, which serve as conduits for transcription factors (TFs) and communicate with each other via physical contact, forming a series of loops in nuclear chromatin (Bulger and Groudine, 2011).

Conventional enhancers (CEs), both active and poised, can be identified in the genome as nucleosome-free regions. The activity level of each CE is revealed by the density of certain histone modifications, prototypically histone H3 acetylated at lysine 27 (H3K27ac) (Bulger and Groudine, 2011). Recent epigenome analyses have revealed a new class of regulatory regions,

coined super-enhancers (SEs) (Whyte et al., 2013), which are characterized by broad stretches of H3K27ac. Most SEs are dense clusters of highly active CEs, which bind lineage-restricted TFs. Indeed, SEs normally co-localize with a limited set of genes that are most essential for cell identity and function. The acquisition or amplification of SEs near oncogenes contributes to several classes of cancer (Hnisz *et al.*, 2013; Mansour *et al.*, 2014). SEs are also enriched for disease-associated sequence variants, some of which presumably disrupt TF binding sites to alter SE function and expression of its associated gene(s) (Hnisz et al., 2013; Koues et al., 2016). However, contributions of SEs to gene expression programs have been mostly assumed from correlative chromatin profiling, rather than by direct testing (Proudhon et al., 2016). Furthermore, it remains controversial whether SEs represent a new paradigm in transcriptional regulation, or merely clusters of CEs that additively promote transcription (Dukler et al., 2016; Hay et al., 2016).

In addition to *cis*-regulatory elements, gene expression programs are significantly influenced by chromosome architecture, which facilitates or impairs promoter-enhancer contacts. The architecture of mammalian genomes is compartmentalized into topologically associated domains (TADs), which are highly conserved among cell types and species (Dixon *et al.*, 2012). Loci within each TAD interact with one another, but are largely cordoned off from neighboring TADs. Each of these architectural building blocks is subdivided into structures called sub-TADs or contact domains, which are composed of loops between CTCF binding elements (structural loops) or between promoters and enhancers (regulatory loops). At a biochemical level, structural loops form via dimeric interactions between CTCF proteins bound in a convergent orientation at two distinct sites and are stabilized by association with the ring-like cohesin complex (Ghirlando and Felsenfeld, 2016; Rao et al., 2014). The bases of many structural loops serve as boundary

elements that partition active and inactive chromatin domains within TADs and limit inappropriate interactions of regulatory elements with neighboring genes (Hnisz et al., 2016a; Ong and Corces, 2014). In keeping with their structural determinants, contact domains, unlike TADs, may vary significantly between cell types, developmental stages, or activation status (Dixon et al., 2016). Indeed, key questions remain about how intra-TAD architectures form and change during cellular differentiation and transformation. Answers to these fundamental questions will not only impact our understanding of basic gene regulatory mechanisms, but also the etiology of many diseases. A substantial subset of disease-associated SNPs and genomic alterations disrupt CTCF sites, breaking architectural borders, allowing inappropriate communication between enhancers and alternative genes (Lupiáñez *et al.*, 2015; Hnisz *et al.*, 2016).

Similarly, a deeper understanding of the regulatory determinants that underlie oncogenic gene expression programs remains a basic mission of cancer research (Sur and Taipale, 2016). Pathogenic expression programs have been characterized for many cancers, including various types of B cell lymphoma (BCL) (Jiang et al., 2016; Morin et al., 2010). A common class of BCL, termed follicular lymphoma (FL), is incurable. Most FLs exhibit an indolent clinical course, but often transform to a more aggressive cancer, termed diffuse large BCL (DLBCL) (Lenz and Staudt, 2010). Recently, we showed that pathogenic gene expression programs in FL are coordinated by a common set of TFs that, in turn, augment or attenuate activities of their target enhancers when compared with normal B cell counterparts, termed centrocytes (CCs) (Koues et al., 2015). Integrative transcriptome and epigenome analyses revealed a blueprint of pathogenic *cis*-regulatory circuits associated with FL, which predicted connections between distal enhancers and promoters of dysregulated genes. Similar correlation-based circuitries

governing gene expression have been constructed for many normal and transformed cell types (Thurman et al., 2012), revealing a new collection of potential targets for epigenetic therapeutics. However, the validity of predicted circuits remains largely untested at the functional level. This gap is particularly important given that a majority of predicted *cis*-regulatory circuits consist of multiple enhancers connected to a single gene or, conversely, multiple genes connected to a single enhancer (Thurman et al., 2012).

Here, we functionally dissect a predicted *cis*-regulatory circuit for the mitosis-associated kinase, *NEK6*, which is commonly overexpressed in BCL (Mareschal et al., 2015). We find that only a subset of CEs, predicted by correlative algorithms to regulate *NEK6* in BCL, is required to maintain its elevated expression. Strikingly, a B cell-specific super-enhancer is completely dispensable for *NEK6* expression and maintenance of a regulatory hub that co-localizes its promoter with many distal CEs. A cluster of CTCF sites at one border of the *NEK6* contact domain serves as a chromatin and architectural boundary to minimize the functional impact of its regulatory hub with neighboring genes. Our study not only provides insights into how *NEK6* expression is regulated in normal and pathogenic B cells, but also emphasizes the need to rigorously test predictions, based solely on chromatin landscapes, regarding *cis*-regulatory circuits and super-enhancer function.

4.3 Results

The NEK6 Cis-Regulatory Circuit Distinguishes FL Subsets

Very few correlation-based predictions for *cis*-regulatory circuits in normal or transformed cells have been validated functionally by targeted engineering of control elements

within their native chromosomal context (Sur and Taipale, 2016). To rigorously test a manageable set of predictions, we prioritized pathogenic *cis*-regulatory circuits associated with CC transformation into FL (Koues et al., 2015). Prioritization of differentially expressed genes and their corresponding regulomes was tiered for recurrence of pathogenic enhancers in FL samples, altered levels of gene expression, relevant TF binding, and gene function (Fig. S4.1A, Materials and Methods, online Table S1). The scheme yielded seven regulatory clusters and accompanying genes, which we considered to be of high priority for functional dissection (online Table S2). Each of the seven regions consists of multiple enhancers and potential target genes, which renders comprehensive analysis of all prioritized circuits unwieldy. From the seven, we selected a region spanning *NEK6* and several neighboring genes for in depth functional studies, based on multiple criteria. We first tested enhancer activities using luciferase reporters for a series of regulatory elements from the seven surviving regions, each of which displays augmented H3K27ac in FL compared with CC. A regulatory element in the *NEK6* region (CE1) displays the most robust enhancer activity in both an EBV-transformed B cell line (GM12878) and a human BCL line (Farage, Fig. S4.1B). Moreover, *NEK6*, a central gene in the identified circuit, encodes a serine/threonine kinase that mediates mitotic progression, is overexpressed in many cancers, and is essential for sustained growth of tumors derived from numerous tissues (Fry *et al.*, 2012).

With regard to B cell oncogenesis, *NEK6* expression distinguishes the two known subtypes of DLBCL, exhibiting elevated expression in germinal center (GC-) compared with the activated B cell (ABC-) subtype (Mareschal et al., 2015). Epigenome analyses revealed that FL also segregates into two analogous classes (Koues et al., 2015), called subtype 1 (GC-like) and 2 (ABC-like). Strikingly, *NEK6* expression is significantly elevated in subtype 1 FL, further

highlighting its similarity to GC-DLBCL (Fig. 4.1A). One final criterion in selecting the *NEK6* region for further study is its rich regulatory landscape, which seemingly consists of multiple enhancers augmented in BCL and a series of potential architectural elements (see below). Thus, we suspected that analysis of *NEK6 cis*-regulatory circuits would provide insights into enhancer and architectural elements important for cell type-, lymphoma-, or FL subtype-specific expression of this mitosis-associated kinase.

The NEK6 Regulatory Landscape

To identify the collection of distal architectural and regulatory elements that contribute to elevated *NEK6* expression in BCL, we leveraged data from public databases (ENCODE Project Consortium, 2012; Koues et al., 2015). Nucleosome-depleted regions demarcate more than a dozen active or poised elements spread over a 500 kb region encompassing *NEK6* and its neighboring genes (Fig. 4.1B, FAIRE/DNase-seq). Several of these regions are bound by architectural factors, CTCF and RAD21, in GM12878, suggesting they may serve as structural or boundary elements (CTCF sites, CS1-7). *NEK6* has two annotated transcription start sites (TSSs), which are both active in human B cells and GM12878 (Fig. S4.1C). H3K27ac peaks coincide with 14 nucleosome-depleted regions in FL samples, indicating positions of active conventional enhancers (CE1-14). Importantly, many of these enhancers exhibit a higher density of H3K27ac in FL compared with normal CC counterparts, suggesting they are hyperactive in transformed B cells. A subset of active enhancers (CE3-9) is clustered in a region -63 to -40 kb upstream of *NEK6*, which is designated as a super-enhancer (SE1) in both FL and CC samples using the ROSE algorithm (Lovén et al., 2013; Whyte et al., 2013) to analyze H3K27ac ChIP-seq data (Fig. 4.1C). When compared with other cell types, the activities of CE1, CE10, and SE1

are primarily restricted to B cells (Fig. S4.1D). Another conventional enhancer region, CE13-14, is also active in a subset of other cell types that express *NEK6*. These epigenome analyses suggest that CE1, CE10, SE1 and, perhaps, CE13-14 are critical enhancers for driving high levels of *NEK6* expression in activated or transformed B cells.

Extensive genetic manipulations are required to dissect the *NEK6* regulome; however, this approach is currently infeasible using primary human B cells. As such, we identified a tractable cell model that mirrors the *NEK6* chromatin landscape in primary FL. As shown in Fig. 4.1B, the transformed human B cell line, GM12878, meets this criterion, while the human T lymphocyte cell line, Jurkat, exhibits a chromatin landscape largely devoid of active regulatory elements near *NEK6*, thus providing a negative control. In addition to recapitulating patterns of active enhancers in primary B and FL cells, the CE3-9 region is classified as an SE in GM12878 (Fig. 4.1C). *NEK6* expression in GM12878 is comparable to levels observed in tonsillar B cells, the majority of which are activated, whereas *NEK6* transcripts are nearly undetectable in Jurkat (Fig. 4.1D).

In addition to *NEK6*, two neighboring genes, *LHX2* and *PSMB7*, are predicted to connect with many of the B cell-restricted enhancers in FL using a gene circuitry algorithm (Koues et al., 2015). *LHX2* is a TF involved in the differentiation of developing lymphoid and neural cell precursors and is a putative oncogene for pancreatic tumors (Zhou et al., 2014). *PSMB7* is a proteasome subunit that was identified as a biomarker for breast and colon cancers (Munkácsy *et al.*, 2010). As shown in Fig. 4.1E, expression of these two genes, but not the more distal *DENND1A*, are modestly elevated in FL and/or tonsillar B cells compared with human CCs. All of these genes are expressed at varying levels in GM12878 (Fig. S4.1E). As such, functional

dissection of the *NEK6* *cis*-regulatory circuit can be achieved using GM12878, which recapitulates prominent features of the FL regulome.

Spatial Convergence of NEK6 Distal Regulatory Elements

Proper control of gene expression requires direct contact of distal regulatory elements with their target promoters. Many cell type-specific contacts between enhancers and promoters are confined within TADs and further restricted by boundary elements to minimize inappropriate enhancer-promoter communication. To elucidate the *NEK6* interactome within its chromosomal neighborhood, we analyzed publicly available Hi-C data for a ~2 Mb region in GM12878 (Fig. 4.2A) (Rao et al., 2014). Based on interactomes conserved among cell types, the TAD containing *NEK6* spans ~1 Mb encompassing *DENND1A*, *LHX2*, *NEK6* and *PSMB7*. In GM12878, this region also contains several sub-TADs, one of which includes *NEK6*, spanning from the *DENND1A* promoter to *PSMB7* (~500 kb). Within the sub-TAD, there is a robust contact domain spanning from the cluster of upstream CTCF sites (CS2-4) to the downstream *NEK6* promoters (TSS1-2). More focal contacts are observed between both *NEK6* promoters and pockets of upstream regulatory elements, especially with CE1 and SE1. Hi-C data revealed associations of the *NEK6* locus with *PSMB7* and, to a lesser extent, with *LHX2*, suggesting a potential mechanism for their elevated expression in FL. Finally, *NEK6* is flanked by two sets of CTCF sites pointing in convergent orientations, a trio located approximately 130 kb upstream of TSS1 (CS2-4) and a pair located in a *NEK6* intron (CS5) and near the *PSMB7* promoter (CS6). The convergent orientation favors loop formation between CTCF regions (Ghirlando and Felsenfeld, 2016; Rao et al., 2014), perhaps spatially sequestering the *NEK6* regulome.

To determine whether this regulatory architecture is cell type-specific, we performed 3C assays in GM12878 (*NEK6*$^{+}$) and Jurkat (*NEK6*$^{-}$), which directly probes interactions between a given viewpoint and selected regions of the *NEK6* chromosomal neighborhood. As shown in Fig. 4.2B, a viewpoint spanning TSS1 interacts with upstream regulatory regions and with TSS2 at significantly higher frequencies in GM12878 compared with Jurkat. Peak TSS1 associations are with the CTCF cluster (CS2-4), CE1, CE2, and sites within SE1. To further validate the *NEK6* interactome, we assayed a number of complementary viewpoints. Interactions with the distal CE1 element are significantly higher throughout the *NEK6* sub-TAD in GM12878 compared with Jurkat. The enhanced CE1-*PSMB7* contacts were confirmed using a *PSMB7* promoter viewpoint (Fig. S4.2A). Coupled with 3C assays using viewpoints in SE1 (Fig. S4.2B, C), TSS2 (Fig. S4.2D), and the CTCF cluster (Fig. S4.2E), we conclude that the upstream region of *NEK6* folds into a cell type-specific regulatory conformation, forming a hub for enhancers, promoters, and CTCF sites, which likely drives higher levels of *NEK6* expression in activated B cells.

Conventional Enhancers Augment NEK6 Expression in Transformed B Cells

Our ultimate goal is to test predictions for key components of the *cis*-regulatory circuit associated with elevated *NEK6* expression in transformed B cells. Chromatin profiling and interactome analyses revealed over a dozen enhancer elements that could potentially augment *NEK6* expression in FL. To prioritize functional analyses, we first measured enhancer activities for each candidate regulatory element in GM12878 and Jurkat (Fig. 4.3A, Fig. S4.3A). In addition to the robust, GM12878-specific enhancer activity of CE1, four other elements augment luciferase expression from SV40 promoter-driven reporters. These include two regions in SE1 (CE5 and 9), the CE10 region upstream of TSS1 and the CE13 region upstream of TSS2. Despite

its significant levels of interaction with *NEK6* promoters (Fig. 4.2B), CE2 lacks enhancer activity in GM12878, which is consistent with minimal deposition of H3K27ac over this region (Fig. 4.1B). The activity status of CEs was bolstered by ChIP-seq data from GM12878 (ENCODE Project Consortium, 2012), which reveals significant peaks for EP300 and TFs important in B cell biology, including EBF1, OCT2, PU.1, PAX5, RELA and TCF3 (Fig. S4.3B). In contrast, CE2 lacks significant binding by any of these factors. These functional data led us to first focus on the role of three CEs located outside of SE1, which had the most robust activities in GM12878 (CE1, CE10 and CE13).

To test the contributions of selected CEs to *NEK6* expression, we individually deleted each enhancer from its endogenous site in GM12878 using CRISPR/Cas9 technology (online Table S3). Deletion of CE13, which is proximal to TSS2, produces a modest, but significant decrease in *NEK6* expression when compared with subclones retaining the enhancer on both alleles (Fig. 4.3B). Ablation of CE10 has no significant impact on *NEK6* expression, despite its enhancer activity in luciferase assays. Importantly, *NEK6* expression is attenuated substantially in subclones lacking the most distal enhancer, CE1, located 120 kb from TSS1. Consistently, NEK6 protein levels are dramatically reduced in $CE1^{-/-}$ subclones as measured by western blotting (Fig. S4.3C). The effects of each enhancer deletion are indistinguishable for transcripts derived from either TSS1 or TSS2 (Fig. S4.3D). Moreover, neither the CE1 nor the CE13 enhancer deletion affects expression of neighboring *LHX2* and *PSMB7* genes (Fig. S4.3E). These data suggest that CE1 and CE13 both contribute to augmented *NEK6* expression in transformed B cells. Indeed, compound deletion of both elements further diminishes NEK6 mRNA and protein expression (Fig. 4.3C, Fig. S4.3C). We conclude that two conventional enhancers,

positioned outside of the large super-enhancer, additively potentiate *NEK6* expression in GM12878.

To probe the effects of enhancer deletions on *NEK6* chromatin and interaction landscapes, we analyzed subclones using ChIP and 3C, respectively. Deletion of CE13 reduces H3K27ac to near background levels at an adjacent region, verifying removal of the core enhancer (Fig. 4.3D). H3K27ac levels in CE13$^{-/-}$ mutants are unaffected at all other *NEK6* enhancers tested. In sharp contrast, deletion of CE1 leads to significant reductions in H3K27ac not only at an adjacent region, but also at many locations within SE1 and other enhancers that associate with CE1. These data suggest that CE1 is a dominant element in sculpting the active epigenetic landscape near *NEK6*, perhaps through spatial interactions that form its regulatory hub. However, the TSS1 interactome is unaffected by deletion of either CE13 or CE1 (Fig. 4.3E). Likewise, CE1 deletion does not alter long-range interactions between this region and downstream regulatory elements, including the TSSs (Fig. S4.3F). However, deletion of CE13 slightly boosts associations of CE1 with downstream enhancers, as well as *NEK6* TSSs (Fig. S4.3G). This finding suggests that CE13 may partially compete with CE1 for association with TSSs and other elements of the regulatory hub. When CE13 is deleted, there may be a compensatory increase in CE1 interactions.

To further test whether the dominant CE1 element is dispensable for maintaining the *NEK6* interactome, we performed 4C-seq on GM12878, as well as three independent CE1$^{-/-}$ and two wild-type subclones. Genome-wide interactome data probed from CE1 and TSS1 viewpoints show that CE1 deletion subclones have no significant differences for interactions with regions between CS2 and downstream of TSS2 (Fig. 4.3F and Fig. S4.3H), validating our 3C findings. These data indicate that maintenance of the *NEK6* regulatory hub, which includes the distal

CTCF cluster, CE1, SE1, CE13, and TSSs, is independent of the dominant conventional enhancer, CE1. However, this element contributes significantly to the maintenance of active chromatin marks at other CEs in the regulatory hub, boosting *NEK6* expression in GM12878.

The NEK6 Super-Enhancer Is a Bystander

Super-enhancers are thought to be dominant regulatory elements for genes controlling cell identity, major cellular functions and, in some cases, oncogenesis (Hnisz et al., 2013). Our chromatin analysis identified SE1, a 23 kb region located between CE1 and the TSSs, as a B cell-specific *NEK6* super-enhancer. Although two conventional enhancers (CE1 and CE13) contribute to *NEK6* expression, a substantial level of transcripts remains following their deletion, further implicating SE1 as an important regulatory element. To test this directly, we deleted the entire SE1 region from both alleles of GM12878 using CRISPR/Cas9. Surprisingly, multiple independent clones lacking SE1 consistently express *NEK6* mRNA at modestly higher levels when compared with subclones retaining an $SE1^{+/+}$ configuration (Fig. 4.4A). Removal of SE1 also enhances or has minimal impact on NEK6 protein expression (Fig. S4.3C). ChIP analysis revealed a depletion of H3K27ac neighboring the deleted SE1, confirming removal of the super-enhancer (Fig. 4.4B). However, SE1 deletion does not impact H3K27ac levels at other tested CEs. Moreover, compound deletion of SE1 on one allele of $CE1^{-/-}$ clones has no significant impact on *NEK6* expression (Fig. 4.4A).

One potential explanation for enhanced *NEK6* expression following SE1 removal is that CE1 resides 23 kb closer to its promoters. However, this would imply that SE1 itself does not contribute fundamentally to *NEK6* expression. To explore the impact of SE1 on the *NEK6* regulatory hub, we performed 3C. As shown in Fig. 4.4C, SE1 deletion potentiates interactions

between TSS1 and more distal elements (CE1 and CE2). The $SE1^{-/-}$ clones also show enhanced associations between TSS1 and more proximal regulatory regions (CE10 and TSS2), whose linear distances are unaffected by SE1 deletion. These data suggest that SE1 has a modest inhibitory impact on the frequency of enhancer associations in the *NEK6* regulatory hub, as well as overall expression of this gene in GM12878.

An alternative explanation for the lack of SE1 regulatory function is that removal of critical enhancer elements drop NEK6 levels below a threshold required for GM12878 proliferation or survival. To test this possibility, we depleted NEK6 using several independent shRNAs. Reduced levels of NEK6 protein (20-30% normal, Fig. S4.4A, B) have no detectable impact on either proliferation or survival of GM12878 (Fig. S4.4C, D). The lack of a biological phenotype may also stem from expression of NEK7 in these cells, a closely related kinase with significant functional overlap (Fry et al., 2012). These data indicate that selective pressure from reduced NEK6 levels cannot reasonably explain the lack of a significant expression phenotype in SE1-deficient cells.

Although SE1 is dispensable for *NEK6* expression in GM12878, it remains possible that this broad regulatory region may target another gene in its chromosomal neighborhood. Focused RT-qPCR analysis of $SE1^{-/-}$ clones revealed no significant change in *PSMB7* expression (Fig. 4.4D). Similar to its effect on *NEK6*, SE1 deletion modestly enhances levels of *LHX2* transcripts. To explore potential SE1 roles on a more global level, we analyzed three independent GM12878 subclones with $SE1^{+/+}$ or $SE1^{-/-}$ genotypes using RNA-seq. SE1 deletion does not significantly change steady-state expression of any gene located within 5 Mb (Fig. 4.4E). On the transcriptome level, six genes are significantly increased or decreased in $SE1^{-/-}$ clones compared with their wild-type counterparts (Fig. S4.4E). The six genes are located on five different

chromosomes; however, published promoter-capture Hi-C data reveal no significant inter-chromosomal interactions between any of the gene promoters and SE1 in GM12878 (Mifsud et al., 2015). We conclude that SE1, although clearly assigned as a super-enhancer using current algorithms, has no identifiable regulatory impact for maintaining expression of its nearest neighbors or any gene in a large chromosomal swath centered on *NEK6*.

A CTCF Cluster Establishes the NEK6 Contact Domain but Not the Regulatory Hub

Our functional data clearly demonstrate that two conventional enhancers, CE1 and CE13, additively increase *NEK6* expression in transformed B cells. The more distal of these two elements, CE1, requires long-range looping (>120 kb) to communicate with *NEK6* promoters. Architectural elements, largely consisting of CTCF sites, are common mediators of long-range looping that facilitate enhancer contact with gene promoters. Moreover, some CTCF sites serve as boundary elements to compartmentalize chromatin domains and inhibit inappropriate communication between enhancers and other neighboring genes (Ghirlando and Felsenfeld, 2016). CE1 is flanked by a cluster of CTCF sites positioned at one border of a robust contact domain containing *NEK6*. All three sites in this cluster are oriented convergently with a pair of downstream CTCF sites, located in a *NEK6* intron (CS5) and near the *PSMB7* promoter (CS6). The convergent orientation favors intermolecular CTCF interactions, which could form loops to cordon off *NEK6*-associated enhancers from other genes in the TAD. To explore architectural logic in the *NEK6 cis*-regulatory circuit, we deleted a region spanning all three sites in the upstream CTCF cluster (CS2-4). Minimal CTCF binding is detected at sites flanking CS2-4 following its deletion, compared with wild-type loci (Fig. 4.5A), whereas CTCF ChIP signals are

unaffected at CS5 and CS6. *NEK6* expression is reduced ~20% in subclones harboring the CS2-4 deletion on both alleles (Fig. 4.5B). In contrast, *LHX2* expression is enhanced ~60% in knock-out subclones, while expression of the two other genes in this TAD, *DENND1A* and *PSMB7*, remains unchanged.

These data suggest that CS2-4 serves as a boundary element to prevent the spread of active chromatin from *NEK6* to *LHX2*, or to minimize long-range interactions between *NEK6* enhancers and *LHX2*, or both (Ghirlando and Felsenfeld, 2016; Ong and Corces, 2014). To test the first possibility, we measured H3K27ac densities at sites in the *NEK6* contact domain and adjacent *LHX2* regions (Fig. 4.5C). Consistent with a role for CS2-4 as a chromatin boundary, its deletion permits H3K27ac spreading upstream of CE1 into the *LHX2* locus. The CS2-4 deletion had an opposite effect on H3K27ac densities within the *NEK6* contact domain, which are significantly reduced, and accompanied by an increase in the H3K27me3 modification (Fig. S4.5A). Thus, perturbed patterns of chromatin modifications correlate well with altered gene expression upon deletion of the 5' CTCF cluster, supporting its functional assignment as a boundary element.

To determine whether CS2-4 also serves as a spatial boundary, precluding communication between *NEK6* enhancers and other promoters, we performed 3C on subclones with wild-type and CS2-4$^{-/-}$ genotypes. As expected, mutant subclones generate no 3C signal for interactions between TSS1 and the deleted CS3 region (Fig. 4.5D). All other interactions between TSS1 and *NEK6* regulatory elements are unaffected by the CS2-4 deletion. In contrast, TSS1 interactions with the *LHX2* and *DENND1A* promoters, located further upstream in the sub-TAD, are significantly increased in mutant subclones. A similar enhancement of upstream interactions is observed for the CE1 element with *LHX2* but not *DENND1A*, which correlates

with the differential impacts of CS2-4 deletion on expression levels. Conversely, CE1 associations are decreased with downstream regions, including CS5 and the *PSMB7* promoter. The enhanced interactions with *LHX2* were confirmed using a complementary viewpoint corresponding to its promoter (Fig. 4.5E).

To support these findings, we performed 4C-seq on GM12878, as well as independent CS2-4$^{-/-}$ and wild-type subclones (Fig. 4.5F, Fig. S4.5B and C). Genome-wide interactome data probed from TSS1 and CE1 viewpoints reveal that, in general, CS2-4$^{-/-}$ subclones have more robust associations with upstream regions in the sub-TAD, reaching to the *DENND1A* promoter, as reflected in percent total normalized reads (Fig. 4.5F) (Guo et al., 2015). In contrast, interactions within the *NEK6* contact domain itself are slightly attenuated following CS2-4 deletion (diminished percent normalized reads in Fig. 4.5F, Fig. S4.5B). In addition, 4C-seq data identify several interactions that differ significantly between CS2-4$^{-/-}$ and control clones. Deletion of the CTCF cluster significantly augments interactions between CE1 and several regions upstream (Fig. 4.5F, green asterisks), as well as with the *LHX2* promoter, although the latter does not attain statistical significance in 4C data. Conversely, multiple interactions of CE1 with downstream regions in the *NEK6* gene body and *PSMB7* promoter region are significantly diminished following CS2-4 removal (Fig. 4.5F, red asterisks), consistent with our 3C data (Fig. 4.5D). Similarly, upon CS2-4 deletion, TSS1 has significantly elevated associations with the *DENND1A* and *LHX2* promoters (Fig. 4.5F).

A potential explanation for the latter finding is that new contact loops may be formed between *NEK6*-proximal CTCF sites (e.g., CS5) and the properly oriented CTCF site upstream of the deleted CS2-4 region. A CTCF site located between the *DENND1A* promoter and *LHX2*, designated as CS0, has the same orientation as those deleted from the CS2-4 cluster (Fig. 4.5F).

Indeed, 3C analyses indicate that the CS2-4 deletion enhances CS0-CS5 interactions, whereas CS0-CS6 crosslinking remains unaffected (Fig. 4.5G). The architectural remodeling of CTCF interactions, which may place the *NEK6* gene in closer proximity to *LHX2* and *DENND1A*, was confirmed using the complementary CS5 viewpoint (Fig. 4.5G). Together, these data indicate that CS2-4 contributes modestly to establishing the regulatory hub between *NEK6* promoters and enhancers. Instead, this CTCF cluster predominantly functions as a chromatin and architectural boundary, minimizing the impact of the *NEK6* regulatory hub on neighboring genes in its TAD.

4.4 Discussion

Developmental and cell type-specific regulation of genes is orchestrated by changes in TF expression, enhancer activation, and alterations in chromatin landscapes, including architecture. Deciphering the contributions of each process to gene regulation is especially important given that a vast majority of disease-associated changes in the genome affect expression levels rather than coding potentials (Maurano *et al.*, 2012). A prerequisite for understanding *cis*-regulatory circuits that govern normal or pathogenic gene expression is the profiling of enhancers and their contacts in distinct cell types. This milestone has largely been achieved in several hematologic malignancies and normal cellular counterparts (Chapuy et al., 2013; Koues et al., 2015). Based on chromatin and architectural profiles, pattern-based algorithms have been used to predict key regulatory connections between enhancers and their target genes. However, there is a critical need to test predicted circuits using reductionist, genetic approaches.

In this study, we dissected *cis*-regulatory circuits within a chromosomal neighborhood spanning at least three genes overexpressed in human BCL. Importantly, many predictions from

pattern-based algorithms for *NEK6* were not substantiated when tested directly. The predicted circuitry for pathogenic *NEK6* expression involved at least a dozen enhancers with augmented H3K27ac loads in FL versus normal B cells. All of the CEs, including those comprising a super-enhancer, directly contact the *NEK6* promoter in transformed B cells, further strengthening their predicted contributions to its elevated expression in BCL. Instead, we find that the *NEK6* regulome is dominated by two conventional enhancers – one located near the TSSs (CE13), and a second, more powerful enhancer (CE1), located ~100 kb upstream. Although some of the predicted enhancers for *NEK6* bind an overlapping set of factors, CE1 exhibits higher loads of TF binding than other enhancers (Fig. S4.3B), potentially explaining its dominant regulatory function. CE13 has lower levels of bound TFs and enhancer activity in luciferase assays, yet its proximity to TSSs may elevate its role in *NEK6* regulation. The remaining CEs and, surprisingly, the super-enhancer, are all dispensable for *NEK6* expression in transformed B cells, despite correlative changes in epigenetic and architectural landscapes. Thus, our study underscores the pressing need to hone predicted circuitry through rigorous testing. Although tedious, the emergence of high throughput methods for genetic dissection of TFs, enhancers, and chromosome architecture will speed achievement of this goal.

We suspect several potential reasons for disconnects between predictive algorithms and direct validation of *cis*-regulatory circuits. First, as shown here for *NEK6*, a dominant enhancer can affect the chromatin profile of other regulatory elements in its interactome. Deletion of CE1 attenuated H3K27ac loads on other CEs spread throughout the *NEK6* region. Thus, increased CE1 activity in BCL likely augments H3K27ac on other elements in the regulatory hub, even if they do not contribute substantially to enhanced gene expression. Second, we cannot rule out that some CEs function as "back-up" elements to partially sustain *NEK6* expression if CE1 activity is

destroyed. This may be true for CE13, which contributes modestly to *NEK6* expression in the absence of CE1. However, SE1 does not appear to have such a back-up role since deletion of the entire region or its composite CEs have no significant effect on *NEK6* expression, whether CE1 is present or not.

The most surprising and significant finding from our study is that a clearly established SE has no discernable impact on the expression of *NEK6* or any other gene on its chromosome. This finding is especially notable given the building dogma that SEs are a collection of key elements controlling high-level expression of genes critical for cell identity and function, as well as oncogenesis (Lovén et al., 2013). Not only does this finding underscore the need for functional evaluation of SEs in many cell types, but it also brings to light a third potential explanation for discrepancies between predicted and validated *cis*-regulatory circuits. Although the SE and a subset of other CEs are dispensable for *NEK6* expression, these elements may be required earlier in B cell development or transformation to initially activate or augment transcription of this kinase gene. After these key activation events, SE1 or other CEs may become dispensable, with CE1 primarily maintaining elevated levels of *NEK6* expression. These issues are currently intractable in primary human B cells, but may be approached in future studies by deletion of analogous regulatory regions for mouse *NEK6*. Notwithstanding, our findings indicate that at least a subset of SEs associated with oncogenesis would not be priority targets for current epigenetic-based therapeutic strategies to squelch expression of associated genes (Lovén et al., 2013).

A second surprise to emerge from our studies concerned determinants for regulatory architecture of the *NEK6* chromosomal neighborhood. We found that most enhancers in this region converge spatially to form a regulatory hub with *NEK6* promoters and flanking CTCF

clusters. Although CE1 is the dominant *NEK6* enhancer, its deletion does not significantly affect maintenance of the regulatory hub. Likewise, deletion of CS2-4 has only a modest impact on spatial interactions within this hub. These findings suggest several intriguing possibilities for architectural determinants of regulatory hubs, which await future dissection, including: (1) direct CE1-promoter interactions are redundant, structurally, with CS2-4 looping to downstream CTCF sites, (2) another element, excluding SE1 and CE1, is the key determinant for initiating regulatory hub formation, or (3) once the *NEK6* sub-TAD is decorated with active histone modifications, homotypic chromatin interactions drive close association of the promoter with regional enhancers (Lieberman-Aiden *et al.*, 2009). Nevertheless, our study identifies important dual roles for CS2-4 as a chromatin and architectural boundary, impairing the spread of active chromatin and enhancer interactions upstream of *NEK6* into *LHX2*. Thus, many CTCF sites or clusters predicted to be important for formation of architectural loops may be more critical in establishing or maintaining borders of regulatory domains.

Our findings will also inform future studies to determine how *NEK6* contributes to B lymphomagenesis. Despite consistent overexpression of the mitosis-associated kinase in BCL, NEK6 depletion had no detectable impact on viability or proliferation of transformed human B cells, including complete *NEK6* knockout in two BCL lines (data not shown). In contrast, NEK6 knockdown in other cancer models significantly attenuated cell growth (Fry et al., 2012). We suspect that, in BCL, partial functional overlap with the closely related kinase, NEK7, may explain the lack of cellular phenotype. Indeed, *NEK7* is overexpressed in primary cells derived from BCL biopsies compared with their normal counterparts (Koues et al., 2015). Human *NEK6* and *NEK7* loci appear to be partial duplicates of one another since both are flanked upstream by additional *LHX* and *DENND* genes. However, unlike *NEK6*, the *NEK7* locus is devoid of

chromatin hallmarks for active distal enhancers in B lymphocytes, FL, or other cell types (ENCODE Project Consortium, 2012; Koues et al., 2015). These correlative data suggest that NEK family kinases are essential components of the program for lymphomagenesis, requiring transformed B cells to augment *NEK6* as a complement, or a back-up, to *NEK7* overexpression, or *vice versa*. Thus, our dissection of the *NEK6* regulome will be an important starting point to test such requirements in the germinal center program and oncogenic conversion to BCL.

4.5 Materials and Methods

Patient Samples. All human samples were obtained under IRB-approved protocols as previously described (Koues et al., 2015).

Prioritization Scheme for *cis*-Regulatory Circuits in FL. We prioritized FL circuits as follows. First, we selected circuits with at least one gene-enhancer pair that was recurrently augmented in more than 6/15 FL samples (Koues et al., 2015): gene expression FL/average CC>1 (quantified by microarray analysis), enhancer histone marks FL/average CC>1.5 (quantified by H3ac, H3K27ac or FAIRE-seq). Second, predicted target gene(s) and relevant enhancers were required to exhibit robust levels of RNA expression (normalized microarray signal > 120) and histone marks (H3K27ac, H3ac and H3K4me1 ChIP-seq RPM > 100) in FL. Third, the surviving list of enhancer-gene combinations was intersected with a manually curated list of ~7000 genes that have been implicated in general oncogenesis, immune modulation, or chromatin modification (online Table S1). Finally, the remaining genetic loci were examined for binding of TFs known to be important for B cell function (EBF1, PU.1, IRF4, IKZF1, POU2F2, PAX5, MEF2A, MEF2C, RUNX3, RELA, TCF3, TCF12, YY1, MAX, STAT1, STAT3, STAT5A, SP1), or the enhancer-associated acetyltransferase, EP300, using public ChIP-seq data

for the transformed B cell line, GM12878, in UCSC Genome Browser (ENCODE Project Consortium, 2012). The remaining list of ~2000 gene-enhancer pairs were ranked based on levels of RNA expression and histone marks, recurrence in FL samples, as well as concordance between expression and histone modifications at putative enhancers. Manual inspection of the top ~200 highest ranked enhancer-promoter pairs yielded seven genetic loci that we considered to be of highest priority. See also Figure S4.1A.

RT-PCR and RNA-Seq. For RT-PCR, total RNA was extracted using TRIzol (Invitrogen), reverse-transcribed (M-MuLV reverse transcriptase, New England Biolabs). SYBR qPCR was carried out using primers in online Table S4. Statistical analysis was performed using Prism. For RNA-seq, total RNA was extracted (RNeasy, Qiagen). Poly (A) mRNA was purified (Dynabeads mRNA Direct, Thermo Fisher Scientific), reverse-transcribed, and used for preparation of indexed libraries. All six libraries were pooled in one lane for 50 bp single-end deep sequencing (Illumina HiSeq2500). RNA-seq reads were aligned to the reference human genome (Ensembl 76) with STAR 2.0.4b (Dobin et al., 2013). Gene counts were derived by Subread:featureCounts 1.4.5 (Liao et al., 2014). Statistical analysis was performed using edgeR 3.14.0 (Robinson et al., 2010).

Luciferase Assay. Candidate enhancers (~800bp) were PCR amplified (online Table S4) and cloned into SV40 promoter-driven pGL3 plasmid (Promega). Reporters were transfected into GM12878 and Farage (Roche 06366236001), or electroporated into Jurkat.

SE Calling. H3K27ac ChIP-seq data for primary B cells (Koues et al., 2015) and GM12878 (ENCODE Project Consortium, 2012) were aligned to the reference human genome (hg19) with

Bowtie2 (Langmead et al. 2012). Peaks were called using MACS, and SEs were called using ROSE under default settings (Lovén et al., 2013; Whyte et al., 2013).

3C and 4C-seq. 3C and 4C-seq assays were performed as described previously (Hagège et al., 2007; Majumder et al., 2015; Splinter et al., 2012) using strategies detailed below. Primers and probes are shown in online Table S4. 4C-seq statistics are shown in Table S4.1. **3C.** In brief, 10^7 cells were crosslinked with 1% formaldehyde, quenched with glycine, lysed, digested with HindIII, religated, and purified with phenol-chloroform followed by Qiagen PCR purification columns. Interactions were measured using a Taqman qPCR assay for ligation products between each anchor HindIII fragment and each target HindIII fragment. Interaction frequencies were normalized for signals obtained from nearest neighbor fragments in the *EEF1G* gene. Standard curves were generated using HindIII digested and religated bacterial artificial chromosomes (RP11-1123P20, RP11-15B9, RP11-902D21 and RP11-259I15 for *NEK6*, RP11-993C15 for *EEF1G*). Amplicons with extreme Ct values in standard curves were either discarded or analyzed using delta Ct values. Statistical analysis was performed using Prism. **4C-Seq.** In brief, 3C DNAs were digested with a second restriction enzyme, DpnII, religated, and purified using Qiagen PCR purification columns. The circularized DNA was amplified using inverse PCR and nested inverse PCR reactions with primers in the anchor HindIII-DpnII fragment. PCR products were used to prepare indexed sequencing libraries. All twelve libraries were pooled in one lane for 50 bp single-end deep sequencing (Illumina HiSeq2500). Reads were aligned to the reference human genome (build hg19) with Bowtie2 2.2.9 (Langmead and Salzberg, 2012). Reads for each HindIII fragment were calculated using r3Cseq 1.18.0 (Thongjuea et al., 2013) and normalized using DESeq2 1.14.1 (Love et al., 2014). Statistical analysis for differential interactions between

genotypes were performed using DESeq2. Spearman correlation of each genotype was performed using R.

Chromatin Immunoprecipitation. ChIP assays were performed as described previously (Koues et al., 2015) using the following antibodies: 1 μg anti-H3K27ac (ab4729), 1 μg anti-H3K27me3 (ab6002), 8 μl anti-CTCF (Cell Signaling 2899) and anti-rabbit IgG (sc2027). ChIP DNA was analyzed with SYBR qPCR assays using primers listed in online Table S4. Statistical analysis was performed using Prism.

CRISPR-Mediated Deletion. 10^7 GM12878 cells or engineered subclones were electroporated with hCas9 plasmid (Addgene 41815), expression plasmids for two gRNAs targeting sequences that flank the region to be deleted, and a plasmid encoding hCD4. $hCD4^+$ cells were purified 24 h post-transfection using magnetic beads (StemCell Technologies 18052), passaged for ~7 days, subcloned by limiting dilution, and screened for deletions using multiple independent primer pairs outside and inside of the gRNA target sites. gRNA sequences are shown in online Table S3. Most gRNAs were cloned into the Addgene vector 41824, while gRNAs for CE13 were cloned into pKLV-U6gRNA(BbsI)-PGKpuro2ABFP (Addgene 50946). PCR primers for screening deletions are provided in online Table S3. PCR products spanning deletion sites were purified and Sanger sequenced (online Table S3). All molecular analyses were performed on sibling subclones corresponding to parental and mutant genotypes in the same experiment to avoid complications that might arise from drifts in bulk GM12878 cultures and experimental variations.

Western Blotting. Western blotting was performed using standard protocols with the following antibodies: NEK6 (ab133494), GAPDH (sc365062).

NEK6 Knockdown. GM12878 cells were transduced with retroviral vectors containing shRNAs specific for either *GFP* (target sequence: AGCACAAGCTGGAGTACAACTA) or *NEK6* (target sequences 1, 2, and 3: CGGCCAGAGTGTCACAGGCAAA, AGGAGAGGACAGTATGGAAGTA, AGCAGATGATCAAGTACTTTAA) and an hCD2 marker as previously described (Bednarski et al., 2012). Transduced cells were subjected to the following assays. Cell death was quantified by Annexin V (BD Biosciences 556422) and hCD2 (BD Biosciences 560642) double staining 72 h post-transduction. Cell proliferation was measured by CFSE dilution (Life technologies C34554), staining cells with CFSE 48 h post-transduction, then with anti-hCD2 at 72, 96 and 120 h post-transduction. Knockdown efficiencies were assessed for $hCD2^+$ cells purified 72 h post transfection using magnetic beads (Miltenyi Biotec 130-091-114) by western blotting.

Accession Numbers. The accession number for raw reads and processed files for RNA-seq and 4C-seq datasets is GEO: GSE87323.

4.6 Figures

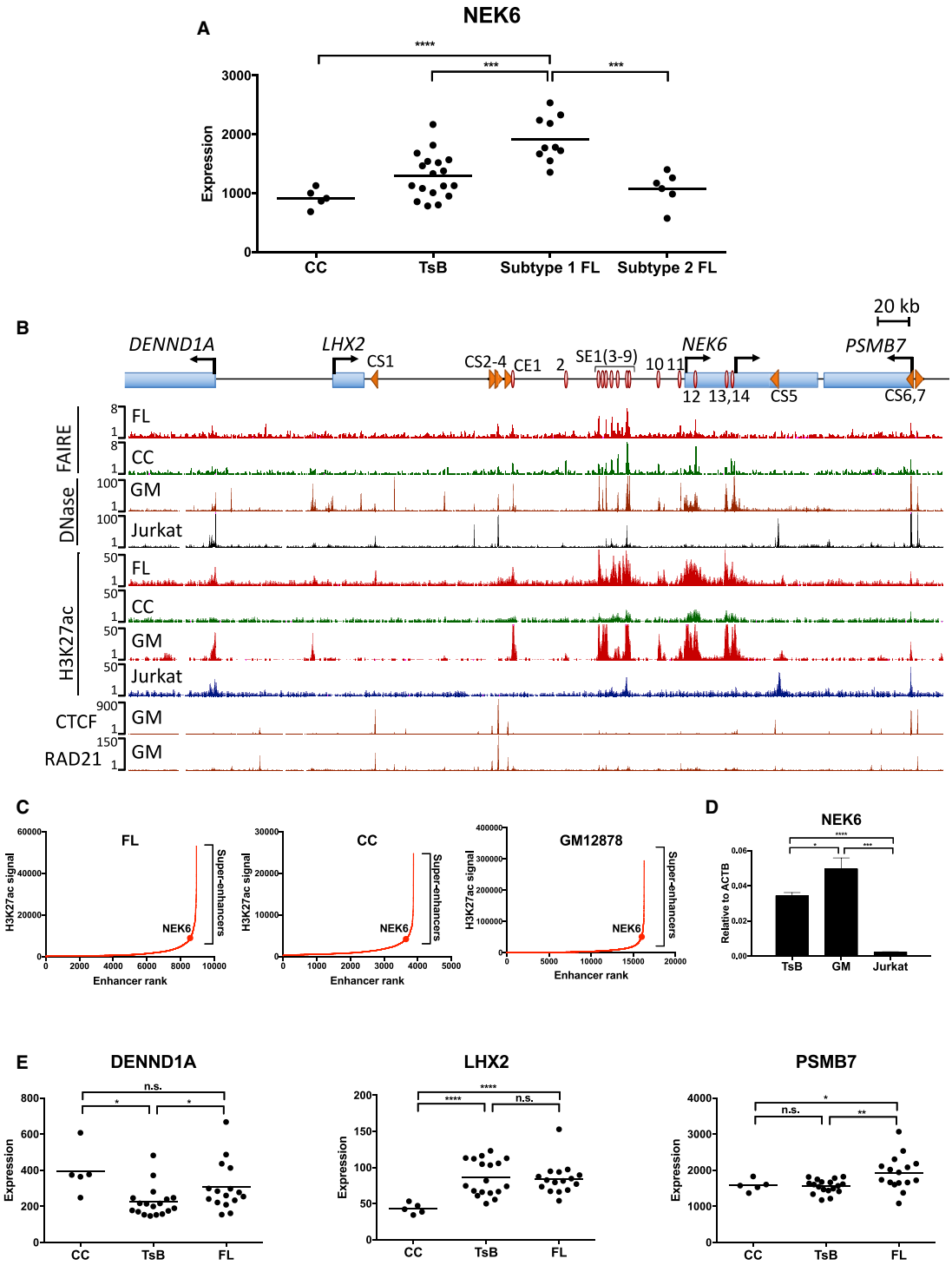


Figure 4.1: The *NEK6* regulatory landscape in normal and transformed cells

(A) Expression levels of *NEK6* in primary human cells. Each dot represents normalized microarray signals for a purified B cell sample from independent healthy volunteers or FL biopsies (CC: tonsillar centrocytes, TsB: unfractionated tonsillar B cells). Statistical tests were performed for subtype 1 or 2 FL versus other cell types. Only significant differences are shown for clarity (unpaired t-test with Welch's correction): ***p<0.005, and ****p<0.001. (B) Scheme depicting genes and regulatory elements in the *NEK6* neighborhood. Red circles represent CEs that are FAIRE- and H3K27ac-positive in at least two FL samples from previously published data (Koues et al. 2015). Orange arrowheads depict CSs, as well as their orientations, as identified by chromatin profiling. UCSC Genome Browser views are shown for FAIRE- and H3K27ac ChIP-seq data from FL and CC samples (Koues et al. 2015), as well as DNase-seq, H3K27ac, CTCF and RAD21 ChIP-seq data in GM12878 (GM) and Jurkat cell lines (ENCODE). All sequencing data are presented as reads per million mapped reads. (C) Rank order of increasing H3K27ac enrichment at enhancers in the indicated cell types. SEs were called using ROSE, with the *NEK6*-associated SE highlighted. (D) *NEK6* transcripts in the indicated cell types measured by RT-qPCR. Results represent the mean ± SEM of three independent experiments. Statistical significance (unpaired t-test with Welch's correction): *p<0.05, ***p<0.005, and ****p<0.001. (E) Expression levels of *NEK6* neighboring genes in primary B cell samples, as measured by microarray. Each dot represents an independent sample. Statistical significance (unpaired t-test with Welch's correction): *p<0.05, and ***p<0.005.

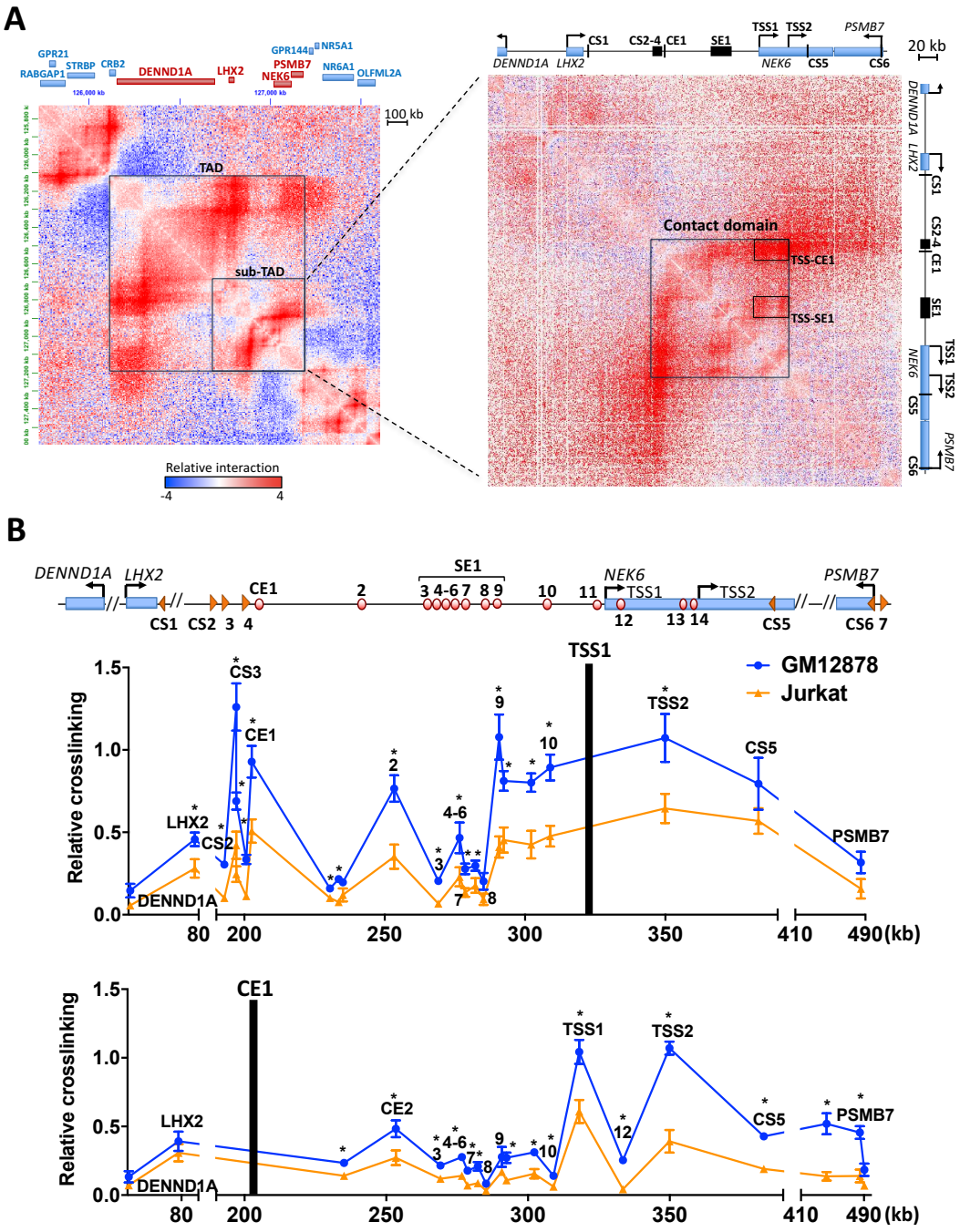


Figure 4.2: The *NEK6* regulatory hub

(A) Hi-C data for the *NEK6* region in GM12878, as visualized in Juicebox (Rao et al. 2014). The intensity of each pixel represents relative normalized numbers of contact between corresponding regions, for which red and blue represent enriched or depleted interaction frequencies, respectively. Knight and Ruiz normalization (balanced) is applied to remove locus-specific biases. The observed over expected (O/E) signal is displayed to account for a higher number of interactions with closer regions due to one-dimensional proximity (Rao et al. 2014). Several chromatin structures and contact points are highlighted with black boxes. In the left panel, genes within the *NEK6*-TAD are colored red and remaining genes are colored blue. (B) Interaction frequencies, as measured by 3C-qPCR,

for *NEK6* TSS1 (top) and CE1 (bottom) viewpoints in GM12878 (*NEK6* expressed) and Jurkat (*NEK6* silent). Results represent the mean ± SEM of three independent experiments. Statistical significance (unpaired t-test with Welch's correction): *p<0.05.

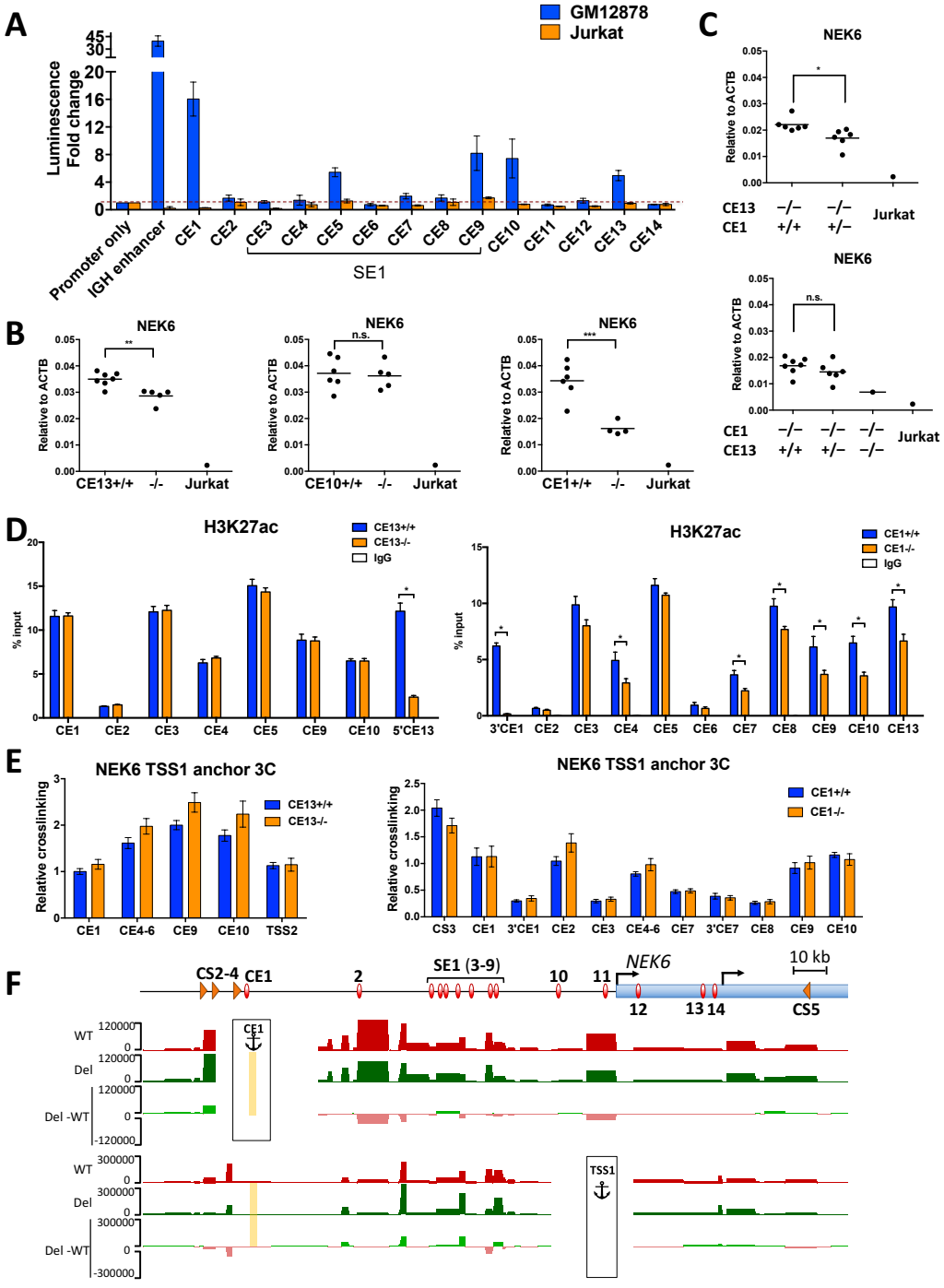


Figure 4.3: CEs potentiate *NEK6* in transformed B cells

(A) Luciferase reporter assays for 14 putative CEs near *NEK6*. Enhancer activities were measured transiently in GM12878 or Jurkat cells and calculated relative to an SV40 promoter-only reporter construct. Human *IGH* enhancer was included as a positive control. Results show the mean ± SEM of at least four independent experiments in GM12878, and at least two in Jurkat. (B and C) *NEK6* transcripts, as measured by RT-qPCR, in different GM12878-derived CRISPR deletion subclones with the indicated genotypes or Jurkat cells, as a negative control. Each dot represents the Jurkat cell line or a unique subclone of GM12878, reported as the average of two independent RNA preparations, reverse transcription, and qPCR assays, the latter performed in triplicate. Statistical significance (unpaired t-test with Welch's correction): *$p<0.05$, **$p<0.01$, and ***$p<0.005$. (D) H3K27ac ChIP assays in GM12878-derived subclones harboring deletions of CE13 (left) or CE1 (right). ChIP-DNA was analyzed by qPCR with primers in or adjacent to indicated CEs. ChIP assays with a non-specific IgG antibody are shown as controls. For panels D and E, each bar represents the mean ± SEM of two subclones, each of which includes two independent experiments. Statistical significance (unpaired t-test with Welch's correction): *$p<0.05$. (E) Interaction frequencies, as measured by 3C-qPCR, in deletion subclones of CE13 (left) and CE1 (right) for the *NEK6* TSS1 viewpoint. (F) UCSC Genome Browser views of interaction profiles, as measured by 4C-seq, for CE1 wild-type and deletion subclones using CE1 and *NEK6*-TSS1 as anchors. For each viewpoint, the average counts per HindIII fragment normalized by DESeq2 are shown for three wild-type (red), and three CS2-4 deletion lines (green). A plot for differential signal between deletion and wild-type samples (Del-WT) is displayed below. None of the differences are statistically significant (DESeq2). The deleted CE1 region is shown as a yellow rectangle.

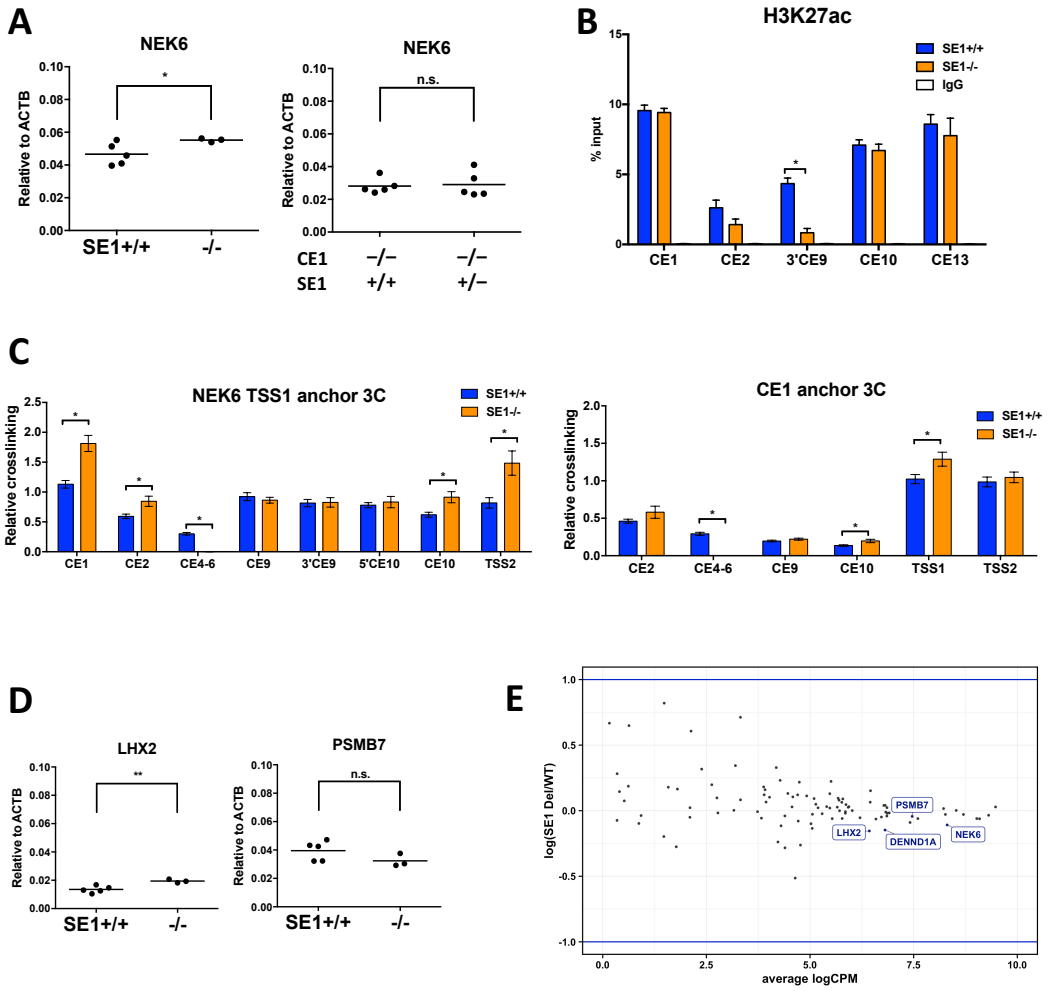


Figure 4.4: SE1 is a dispensable element in the *NEK6* regulome

(A) *NEK6* transcripts measured by RT-qPCR of SE1 deletion subclones. Each dot represents a unique subclone, which is reported as the average of two independent experiments. See Fig. 4.3B and C for details. For panels A-D, statistical significance (unpaired t-test with Welch's correction): *p<0.05. (B) H3K27ac ChIP assays in SE1 deletion subclones. See Fig. 4.3D for details. For panels B and C, each bar represents the mean ± SEM of two subclones, each of which includes two independent experiments. (C) Interaction frequencies, as measured by 3C-qPCR, in SE1 deletion subclones for *NEK6* TSS1 (left) and CE1 (right) viewpoints. (D) *LHX2* and *PSMB7* transcripts measured by RT-qPCR in SE1 deletion subclones. Each dot represents a unique subclone, which is reported as the average of two independent experiments. (E) Expression profile for all genes located within 5 Mb of SE1, as measured by RNA-seq, in SE1 wild-type and deletion subclones of GM12878. Average logCPM indicates the average expression level of each gene among three wild-type and three deletion subclones, reported as log2 read counts per million mapped reads. Log(SE1 Del/WT) represents the log2 fold-change of each gene between the average CPM of deletion versus wild-type subclones. Blue lines denote two-fold differences.

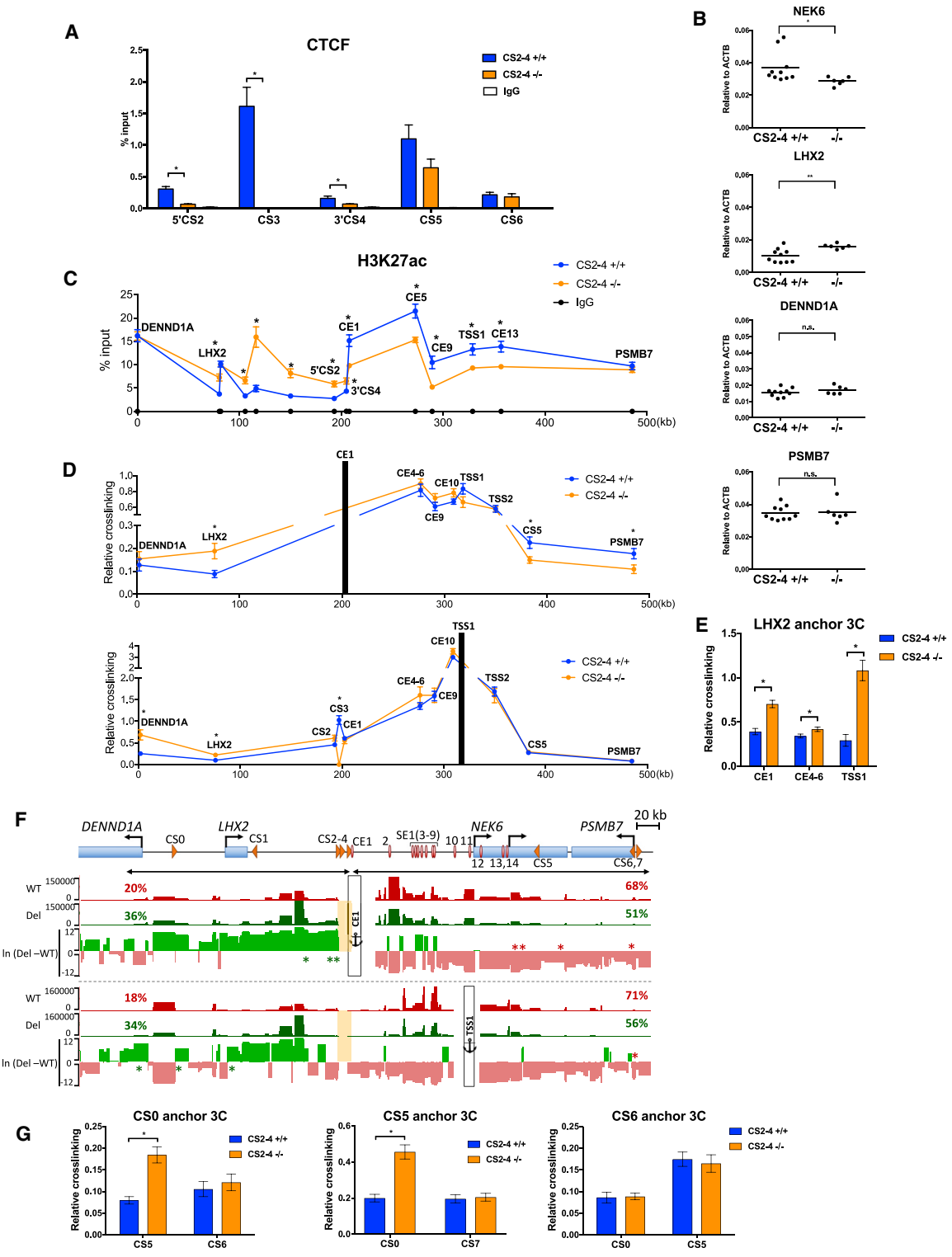


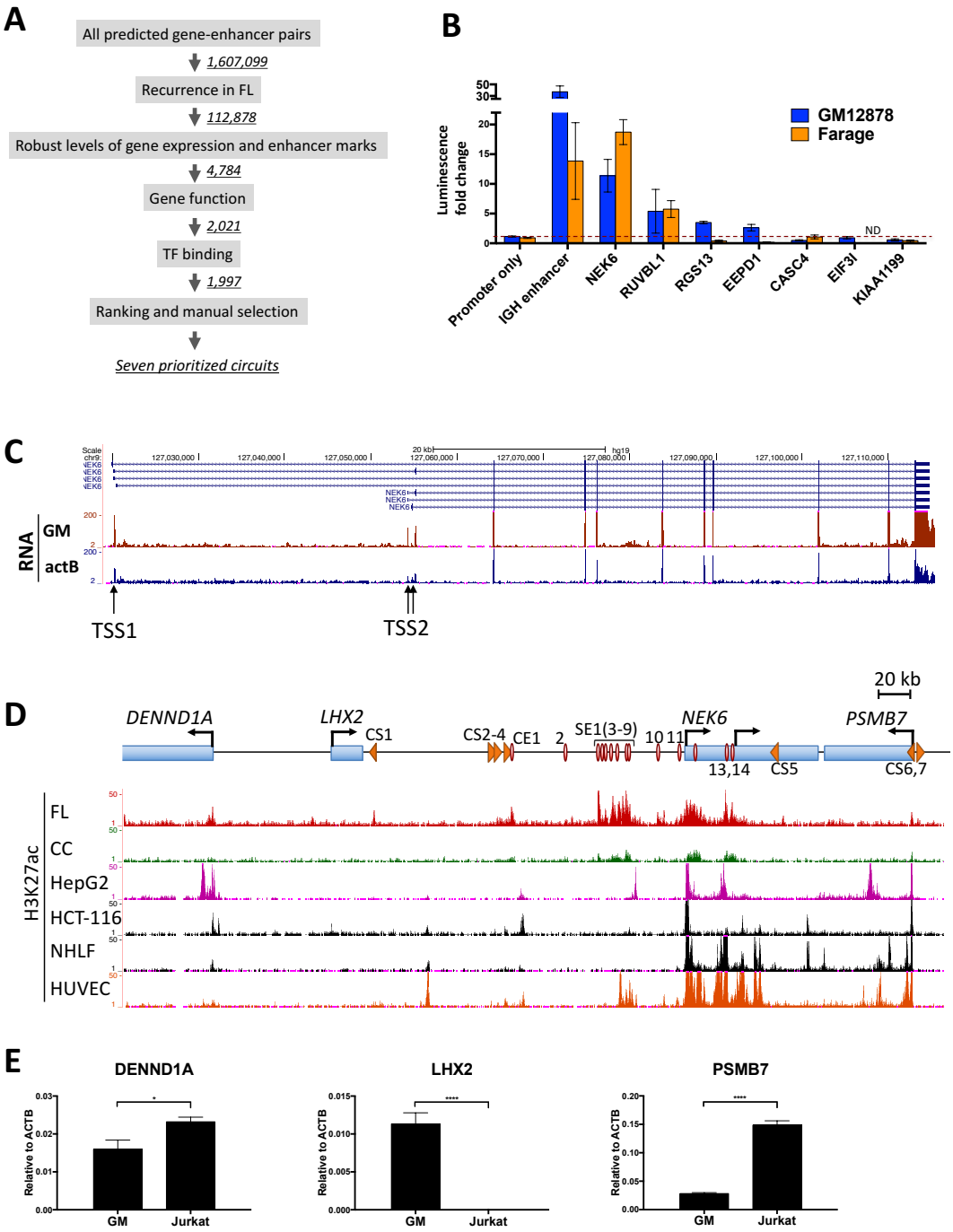
Figure 4.5: CS2-4 serves as a chromatin and architectural boundary for the *NEK6* regulatory hub
(A) CTCF ChIP assays in CS2-4 deletion subclones. ChIP-DNA was analyzed by qPCR using primers within or adjacent to indicated CSs. Each bar represents the mean ± SEM of two subclones, each of which includes two independent experiments. ChIP assays with a non-specific IgG antibody were performed as specificity controls. For

panels A-E and G, statistical significance (unpaired t-test with Welch's correction): *p<0.05, and **p<0.01. (B) Transcript abundance of genes in the *NEK6*-TAD, as measured by RT-qPCR, for CS2-4 deletion subclones. Each dot represents a unique subclone, which is reported as an average of two independent experiments. See Fig. 4.3B and C for details. (C) H3K27ac ChIP assays in C2-4 deletion subclones. See Fig. 4.3D for details. Each bar represents the mean ± SEM of two subclones, each of which includes two independent experiments. (D and E) Interaction frequencies, as measured by 3C-qPCR, in CS2-4 deletion subclones for CE1, *NEK6* TSS1 (D), and the *LHX2* promoter (E) viewpoints. Each dot in (D) or bar in (E) represents the mean ± SEM of two subclones, each of which includes two independent experiments. (F) UCSC Genome Browser views of interaction profiles, as measured by 4C-seq, for CS2-4 wild-type and deletion subclones using CE1 and *NEK6*-TSS1 as anchors. For each viewpoint, the average reads per HindIII fragment normalized by DESeq2 are shown for three wild-type (red), and three CS2-4 deletion lines (green). Reads located within the deleted CS2-4 region (yellow rectangle) are removed from all samples. Percentages of total normalized reads are displayed above each sample for regions upstream and downstream of CS2-4 deletion, as marked by double-headed arrow lines. For each viewpoint, a plot for differential signal between deletion and wild-type samples in natural log scale, ln (Del-WT), is displayed below. Statistical significance (generalized linear model adjusted by Benjamini-Hochberg procedure): p<0.05, are denoted by green or red asterisks for interactions that are increased or decreased in CS2-4 mutants, respectively. (G) Interaction frequencies, as measured by 3C-qPCR, in CS2-4 deletion subclones for CS0 (left), CS5 (middle), and CS6 (right) viewpoints. Each bar represents the mean ± SEM of two subclones, each of which includes two independent experiments.

4.7 Acknowledgements

This work was supported by NIH grants CA156690, CA188286 (J.E.P. and E.M.O.), AI122726, AI115734 (E.M.O.), as well as WU-ICTS (TR000448) and Siteman Cancer Center (CA91842) grants. We thank Patrick Collins for helpful comments and the Genome Technology Access Center for assistance with –omics analyses.

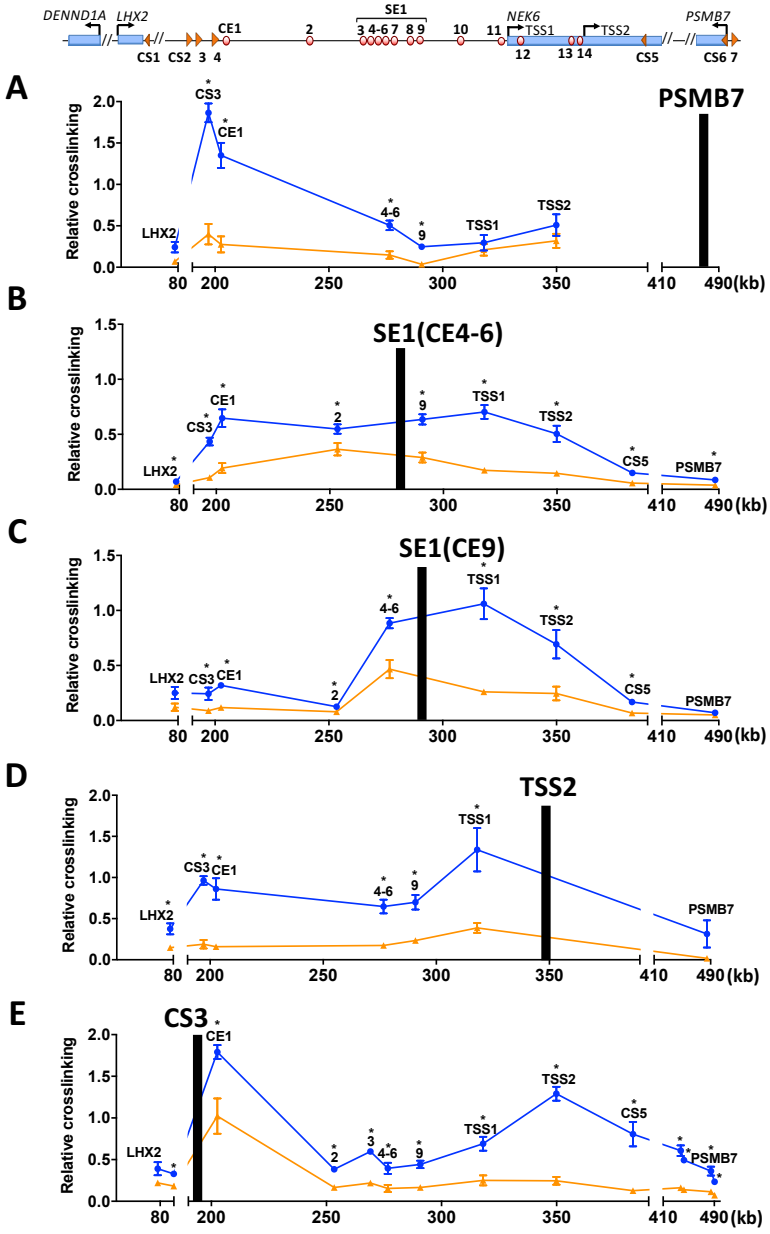
4.8 Supplemental Figures and Tables



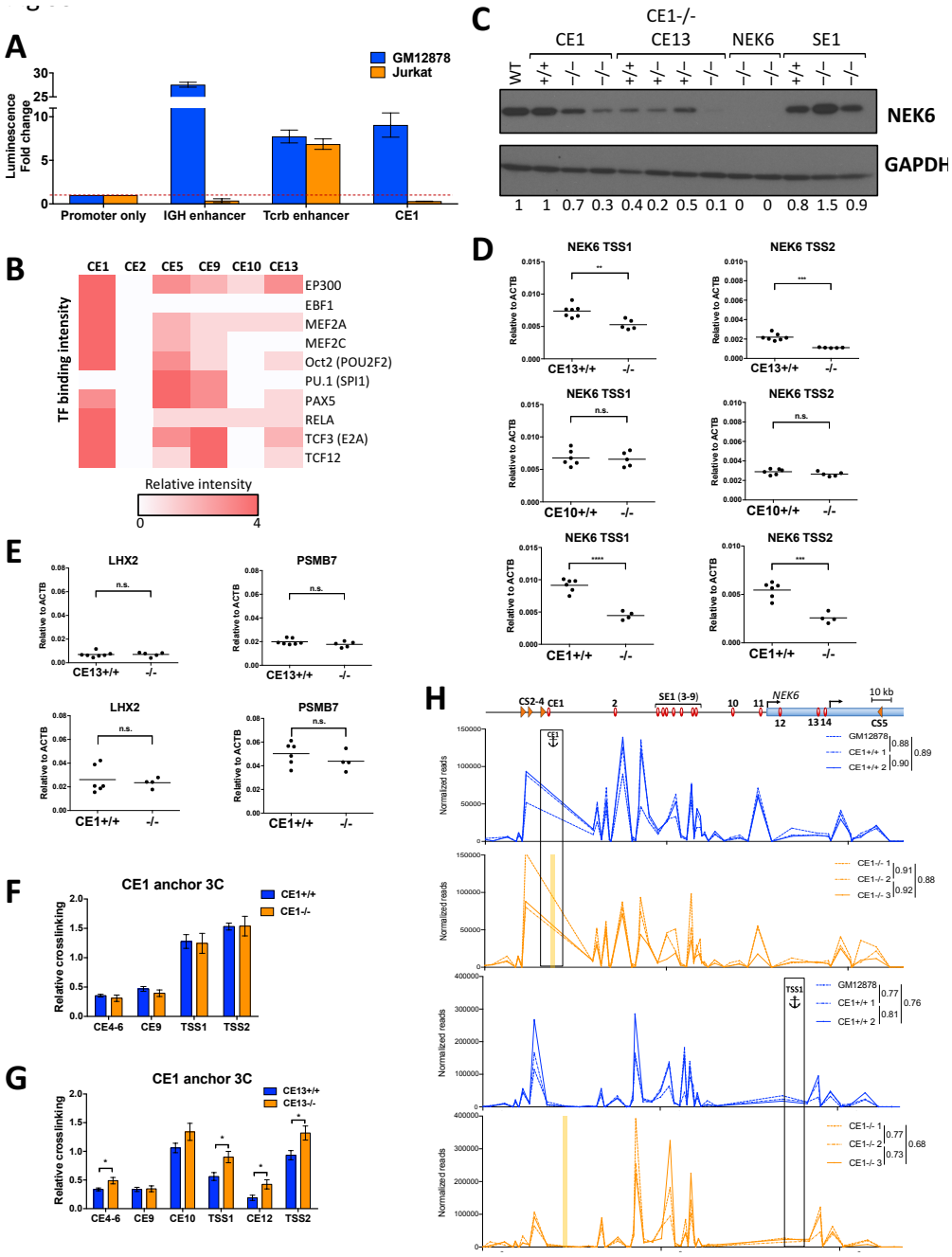
Supplemental Figure 4.1: Prioritization scheme, luciferase assays of putative enhancers, and regulatory landscape of *NEK6* in distinct cell types

(A) A flowchart showing how seven regulatory circuits were selected from genome-wide predicted gene-enhancer pairs. See Supplemental Experimental Procedures for details. (B) Luciferase reporter assays for DREs connected to seven high priority genes with elevated expression in FL. Enhancer activities were measured transiently in

GM12878 and Farage B cell lines and are reported relative to a control reporter construct containing only the SV40 promoter. The potent B cell enhancer associated with the human *IGH* locus is included as a positive control. Results represent the mean ± SEM of at least two independent experiments. ND: Not done. (C) UCSC Genome Browser views of annotated *NEK6* transcript isoforms and RNA-seq data from GM12878 (ENCODE) or in vitro activated B cells (Koues et al. 2015). The two active TSSs for *NEK6* in B cells are indicated. RNA data are presented as the number of aligned, in silico extended reads per 10 bp. (D) UCSC Genome Browser views of H3K27ac ChIP-seq data from FL, CC and other distinct ENCODE cell types. (E) Transcript abundance of *NEK6* neighboring genes measured by RT-qPCR in GM12878 and Jurkat cells. Results represent the mean ± SEM of three independent experiments. Statistical significance (unpaired t-test with Welch's correction): *$p<0.05$, and ****$p<0.001$.

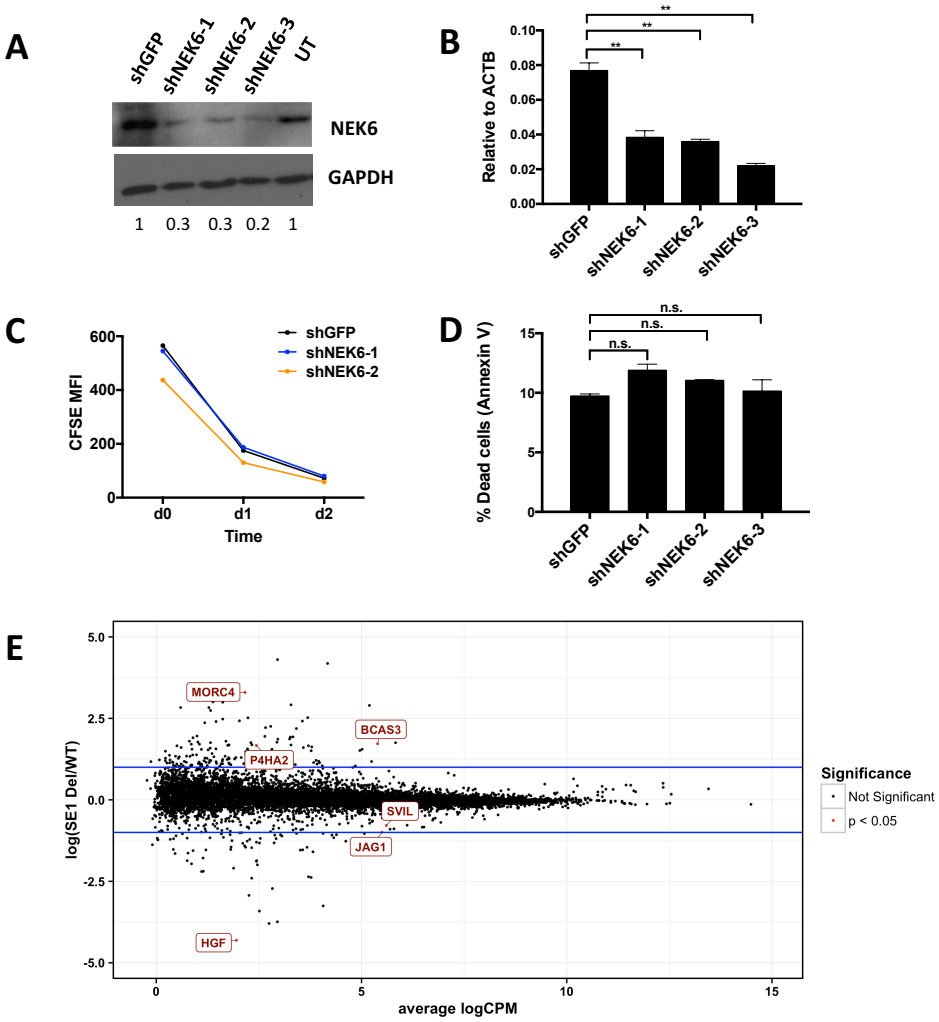


Supplemental Figure 4.2: Interaction frequencies of five additional viewpoints within the *NEK6* sub-TAD
(A-E) Interaction frequencies measured by 3C-qPCR in GM12878 (*NEK6* expressing) and Jurkat (*NEK6* silent) for the indicated viewpoints: *PSMB7* promoter (A), SE1 (CE4-6) (B), SE1 (CE9) (C), *NEK6* TSS2 (D), and CS3 (E). Results represent the mean ± SEM of two independent experiments. Statistical significance (unpaired t-test with Welch's correction): *p<0.05.



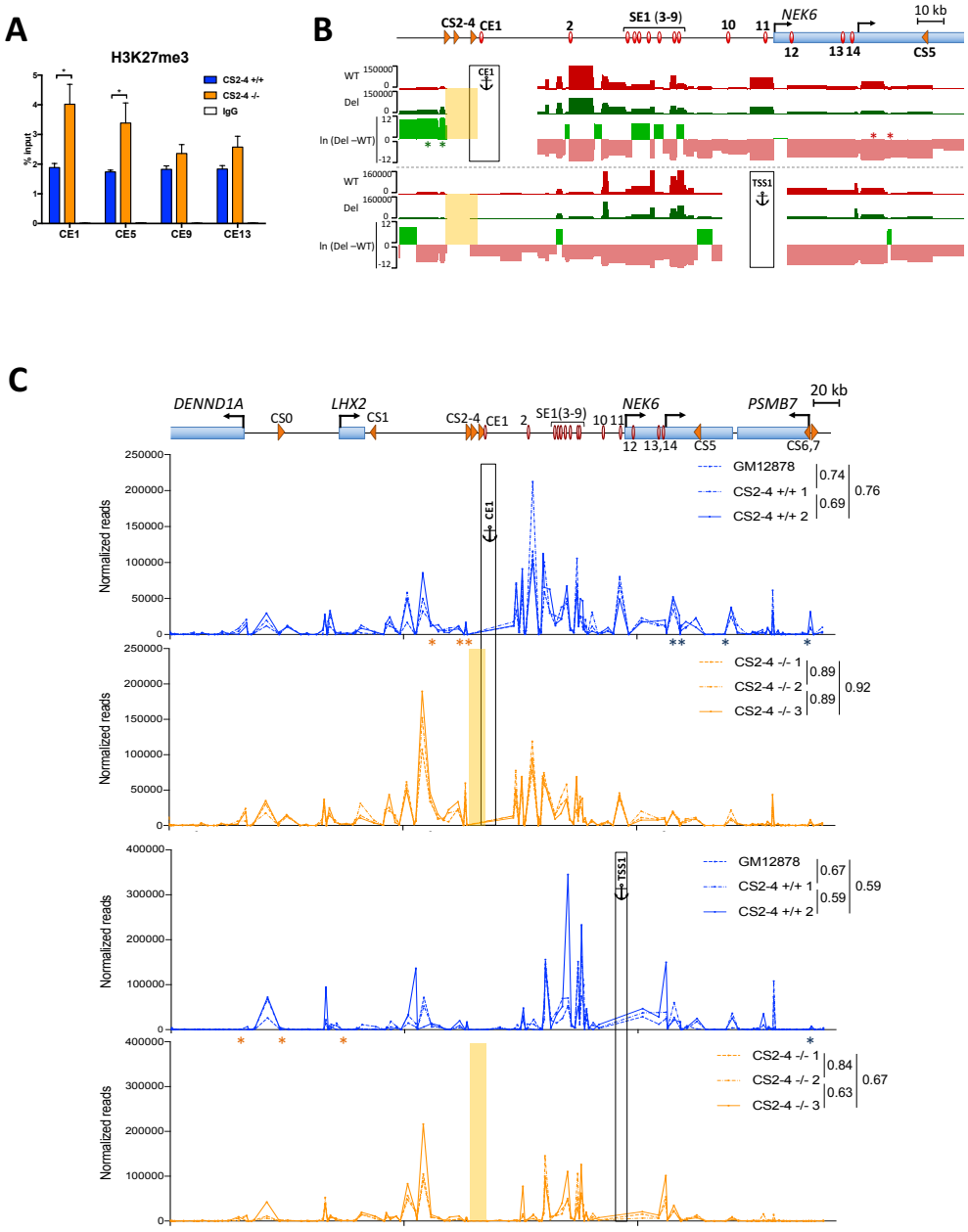
Supplemental Figure 4.3: Luciferase assays, TF binding, expression and interaction analyses of CEs
(A) Luciferase reporter assays for control constructs. Enhancer activities were measured transiently in GM12878 or Jurkat cells and calculated relative to an SV40 promoter-only reporter construct. The human *IGH* enhancer and mouse *Tcrb* enhancer were included as positive controls. Results show the mean ± SEM of two independent experiments. (B) Approximate relative TF binding intensities for six *NEK6* CEs, derived from ChIP-seq data for TFs important in B cell biology in GM12878 (ENCODE). (C) Immunoblots probed with antibodies specific for NEK6 or GAPDH, in different GM12878-derived CRISPR deletion subclones with the indicated genotypes, including parental wild-type cells (WT) and *NEK6* knockout subclones (*NEK6*$^{-/-}$). Normalized NEK6 protein levels relative to WT, as measured by ImageJ, are indicated at the bottom. (D) *NEK6* transcripts derived from the two TSSs, as measured by RT-qPCR, in deletion subclones of CE13 (top), CE10 (middle) and CE1 (bottom). For panels D and E, each dot represents an independent subclone, which is reported as the average of two independent experiments. See

Fig. 4.3B and C for details. Statistical significance (unpaired t-test with Welch's correction): **$p<0.01$, and ***$p<0.005$, ****$p<0.001$. (E) *LHX2* and *PSMB7* transcripts measured by RT-qPCR in CE13 (top) and CE1 (bottom) deletion subclones. (F and G) Interaction frequencies, as measured by 3C-qPCR, in deletion subclones of CE1 (F) and CE13 (G) for CE1 viewpoint. Each bar represents the mean ± SEM of two independent subclones, each of which includes two independent experiments. Statistical significance (unpaired t-test with Welch's correction): *$p<0.05$. (H) Interaction profiles, as measured by 4C-seq, for CE1 wild-type and deletion samples using CE1 and *NEK6*-TSS1 as anchors. For each viewpoint, reads per HindIII fragment normalized by DESeq2 are shown for three wild-type (blue), and three CE1 deletion lines (orange). The deleted CE1 region is shown as a yellow rectangle. Spearman's rank correlation coefficients, as shown on the right of sample names, are calculated for each pair of samples of the same genotype, and are all significant (asymptotic t approximation, $p<2.2 \times 10^{-16}$).



Supplemental Figure 4.4: NEK6 knockdowns in GM12878 and global transcription profiles in SE1 deletion subclones

(A) Immunoblots probed with antibodies specific for NEK6 or GAPDH, in GM12878 cells transduced with shRNAs targeting *GFP* (control) or different regions of *NEK6* transcripts and purified at 72h. Normalized NEK6 protein levels relative to sh*GFP* are indicated at the bottom. (B) *NEK6* transcripts, as measured by RT-qPCR, in GM12878 cells transduced with either *GFP*- or *NEK6*-specific shRNAs and purified at 72h. Statistical significance (unpaired t-test with Welch's correction): **p<0.01. (C) Proliferation rates, as measured by CSFE dilution (flow cytometry), in GM12878 cells transduced with either *GFP*- or *NEK6*-specific shRNAs and analyzed from 72h, which is labeled as d0. Median fluorescence intensities of CFSE are shown on the Y-axis. (D) Cell death, as measured by Annexin V staining (flow cytometry), in GM12878 cells transduced with either *GFP*- or *NEK6*-specific shRNAs and analyzed at 72h. Statistical significance (unpaired t-test with Welch's correction): p<0.05. (E) Global transcription profiles, as measured by RNA-seq, in SE1 wild-type and deletion subclones. Average logCPM indicates the average expression level of each gene among three wild-type and three deletion subclones, reported as log2 read counts per million mapped reads. Log(SE1 Del/WT) represents the log2 fold change of each gene between the average CPM of deletion subclones versus wild-type subclones. Statistical significance is generated using generalized linear model with p-values adjusted by Benjamini-Hochberg procedure. Six genes with p<0.05 are labeled with red color. Blue lines denote two-fold differences.



Supplemental Figure 4.5: H3K27me3 ChIP assays and interaction profiles in C2-4 deletion subclones
(A) ChIP-DNAs were analyzed by qPCR using primers near the indicated CEs. Each bar represents the mean ± SEM of two independent subclones, each of which includes two independent experiments. Statistical significance (unpaired t-test with Welch's correction): *p<0.05. ChIP assays with a non-specific IgG antibody are shown as controls. (B) Zoomed-in UCSC Genome Browser views of interaction profiles, as measured by 4C-seq, for CS2-4 wild-type and deletion subclones using CE1 and *NEK6*-TSS1 as anchors. For each viewpoint, the average reads per HindIII fragment normalized by DESeq2 are shown for three wild-type (red), and three CS2-4 deletion lines (green). Reads located within the deleted CS2-4 region are removed from all samples. Also shown is a plot for differential signal between deletion and wild-type samples in natural log scale, ln (Del-WT). Statistical significance (generalized linear model adjusted by Benjamini-Hochberg procedure): p<0.05, are denoted by green or red asterisks for interactions that are increased or decreased in CS2-4 mutants, respectively. (C) Interaction profiles, as measured by 4C-seq, for CS2-4 wild-type and deletion samples using CE1 and *NEK6*-TSS1 as anchors. For each viewpoint, reads per HindIII fragment normalized by DESeq2 are shown for three wild-type (blue), and three CE1 deletion lines

(orange). The deleted CS2-4 region is shown as a yellow rectangle. Spearman's rank correlation coefficients, as shown on the right of sample names, are calculated for each pair of samples of the same genotype, and are all significant (asymptotic t approximation, $p<2.2 \times 10^{-16}$).

Supplemental Table 4.1: 4C-seq statistics

Sample name	**# of total mapped reads** (reads located within two restriction fragments of the viewpoint are removed)	**Fraction of total mapped reads located in the cis chromosome** (chr9)	**Fraction of total mapped reads located in the NEK6 sub-TAD** (chr9:126,130,000-127,200,000)
CE1viewpoint_GM12878	3,299,896	61%	41%
CE1viewpoint_CE1_WT_clone1	4,712,155	58%	38%
CE1viewpoint_CE1_WT_clone2	4,708,887	57%	39%
CE1viewpoint_CE1_Del_clone1	3,923,007	52%	31%
CE1viewpoint_CE1_Del_clone2	4,000,761	57%	37%
CE1viewpoint_CE1_Del_clone3	4,121,564	64%	44%
CE1viewpoint_GM12878	3,235,282	60%	40%
CE1viewpoint_CS2-4_WT_clone1	3,228,695	61%	44%
CE1viewpoint_CS2-4_WT_clone2	5,606,882	59%	35%
CE1viewpoint_CS2-4_Del_clone1	4,088,825	64%	47%
CE1viewpoint_CS2-4_Del_clone2	5,208,257	63%	45%
CE1viewpoint_CS2-4_Del_clone3	5,819,853	67%	48%
TSS1viewpoint_GM12878	3,650,318	61%	40%
TSS1viewpoint_CE1_WT_clone1	5,801,414	60%	39%
TSS1viewpoint_CE1_WT_clone2	6,220,468	61%	45%
TSS1viewpoint_CE1_Del_clone1	5,300,021	55%	37%
TSS1viewpoint_CE1_Del_clone2	5,587,267	63%	43%
TSS1viewpoint_CE1_Del_clone3	5,307,205	71%	55%
TSS1viewpoint_GM12878	3,824,313	63%	43%
TSS1viewpoint_CS2-4_WT_clone1	2,733,325	64%	39%
TSS1viewpoint_CS2-4_WT_clone2	4,785,052	60%	38%
TSS1viewpoint_CS2-4_Del_clone1	3,281,914	68%	52%
TSS1viewpoint_CS2-4_Del_clone2	2,793,105	66%	49%
TSS1viewpoint_CS2-4_Del_clone3	4,328,501	74%	57%

4.9 References

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.

Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G.M., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. (2013). Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell *24*, 777–790.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell *62*, 668–680.

Dukler, N., Gulko, B., Huang, Y.-F., and Siepel, A. (2016). Is a super-enhancer greater than the sum of its parts? Nat. Genet. *49*, 2–3.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fry, A.M., O'Regan, L., Sabir, S.R., and Bayliss, R. (2012). Cell cycle regulation by the NEK family of protein kinases. J. Cell Sci. *125*, 4423–4433.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Hagège, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forné, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat. Protoc. *2*, 1722–1733.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hnisz, D., Day, D.S., and Young, R.A. (2016a). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016b). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Jiang, Y., Dominguez, P.M., and Melnick, A.M. (2016). The many layers of epigenetic dysfunction in B-cell lymphomas. Curr. Opin. Hematol. *23*, 377–384.

Koues, O.I., Kowalewski, R.A., Chang, L.W., Pyfrom, S.C., Schmidt, J.A., Luo, H., Sandoval, L.E., Hughes, T.B., Bednarski, J.J., Cashen, A.F., et al. (2015). Enhancer Sequence Variants and Transcription-Factor Deregulation Synergize to Construct Pathogenic Regulatory Circuits in B-Cell Lymphoma. Immunity *42*, 186–198.

Koues, O.I., Collins, P.L., Cella, M., Robinette, M.L., Porter, S.I., Pyfrom, S.C., Payton, J.E., Colonna, M., and Oltz, E.M. (2016). Distinct Gene Regulatory Pathways for Human Innate versus Adaptive Lymphoid Cells. Cell *165*, 1134–1146.

Lenz, G., and Staudt, L.M. (2010). Aggressive lymphomas. N. Engl. J. Med. *362*, 1417–1429.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Majumder, K., Rupp, L.J., Yang-Iott, K.S., Koues, O.I., Kyle, K.E., Bassing, C.H., and Oltz, E.M. (2015). Domain-Specific and Stage-Intrinsic Changes in Tcrb Conformation during Thymocyte Development. J. Immunol. *195*, 1262–1272.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., et al. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science *346*, 1373–1377.

Mareschal, S., Ruminy, P., Bagacean, C., Marchand, V., Cornic, M., Jais, J.-P., Figeac, M., Picquenot, J.-M., Molina, T.J., Fest, T., et al. (2015). Accurate Classification of Germinal Center B-Cell-Like/Activated B-Cell-Like Diffuse Large B-Cell Lymphoma Using a Simple and Rapid

Reverse Transcriptase-Multiplex Ligation-Dependent Probe Amplification Assay: A CALYM Study. J. Mol. Diagn.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Morin, R.D., Johnson, N.A., Severson, T.M., Mungall, A.J., An, J., Goya, R., Paul, J.E., Boyle, M., Woolcock, B.W., Kuchenbauer, F., et al. (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat. Genet. *42*, 181–185.

Munkácsy, G., Abdul-Ghani, R., Mihály, Z., Tegze, B., Tchernitsa, O., Surowiak, P., Schäfer, R., and Györffy, B. (2010). PSMB7 is associated with anthracycline resistance and is a prognostic biomarker in breast cancer. Br. J. Cancer *102*, 361–368.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Proudhon, C., Snetkova, V., Raviram, R., Lobry, C., Badri, S., Jiang, T., Hao, B., Trimarchi, T., Kluger, Y., Aifantis, I., et al. (2016). Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. Cell Rep. *15*, 2159–2169.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Splinter, E., de Wit, E., van de Werken, H.J.G., Klous, P., and de Laat, W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods *58*, 221–230.

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nat. Rev. Cancer *16*, 483–493.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Zhou, F., Gou, S., Xiong, J., Wu, H., Wang, C., and Liu, T. (2014). Oncogenicity of LHX2 in pancreatic ductal adenocarcinoma. Mol. Biol. Rep. *41*, 8163–8167.

Chapter 5 : Conclusions and Future Directions

Gene regulation is controlled fundamentally by epigenetic mechanisms, including *cis*-regulatory elements, chromatin states and chromosomal architecture. However, it remains a challenging issue to understand how these features collectively work in normal and pathogenic cellular processes. To start addressing these questions, we studied gene regulation in normal development using antigen receptor loci as model systems, and in disease by focusing on human B cell lymphoma, using chromatin profiling and computational algorithms. In Chapter 2, using the mouse T cell receptor beta locus as a model, we quantified features using chromatin profiling and other assays, and utilized a subset of these features to computationally predict recombination frequencies of Vβ gene segments. In Chapter 3, we applied chromatin profiling data of other mouse antigen receptor loci to bioinformatically classify and identify novel regulatory elements. In Chapter 4, we dissected a pathogenic *cis*-regulatory circuit composed of dysregulated enhancers and target genes in lymphoma, which was predicted *in silico* from chromatin profiling data, in order to validate and provide insights into prediction methods.

Multivariate model to predict pre-selection Vβ repertoires

In Chapter 2, we quantified pre-selection Vβ repertoires for the mouse *Tcrb* locus and reported 13 distinct genetic and epigenetic features, including chromatin states and chromosomal architecture, to evaluate their relative contributions in sculpting the pre-selection Vβ repertoire. We constructed a computational model to predict Vβ usage and discovered dominating predictors, including chromatin modifications and transcription. Our findings reinforce the

importance of chromatin states in regulating *Tcrb* recombination. Next, comprehensive mechanisms establishing chromatin states at Vβs remain to be identified. For example, we found that most pseudogene V segments flanked by functional RSSs are located in a repressive chromatin state. The suppression of pseudogenes can be partially explained by dysfunctional germline promoters and localization to lamina in the nuclear periphery (Reddy et al., 2008). These possibilities can be evaluated using genetic studies and imaging technologies. In addition, our study provides a framework to predict pre-selection repertoires for all AgR loci. For example, improvement of multiplex PCR approaches quantifying pre-selection repertoires will accelerate generating predictive models of other AgR genes. Ultimately, these models will be valuable to predict the effects of designed or natural variations of AgR loci on pre-selection V repertoires in mouse and human, since paralleled mechanisms exist between the usage of human Vβ orthologs and those of mouse (Livák, 2003).

Targeted analyses to identify regulatory elements in antigen receptor loci

In Chapter 3, we classified and defined regulatory elements in mouse antigen receptor loci using focused computational analysis of chromatin profiles. 38 distinct epigenetic states were identified, one of which corresponds to known enhancers and defines novel enhancer candidates in *immunoglobulin* loci. Several selected elements were verified in functional assays, validating our computational approach. In the future, these validated enhancers should be studied *in vivo* for their functions in Ig gene expression, assembly, class switching, and somatic hypermutation. This is especially true for hRE1, a super-enhancer, which potentially contributes to class switching and V(D)J recombination (Han et al., 2007; Medvedovic et al., 2013; Vincent-Fabert et al., 2010). In addition, other regulatory elements can be pursued with functional

validations, including state 4 corresponding to enhancers and state 5 representing promoters. Furthermore, our success indicates that functional assignment and novel element identification in complex loci benefit from focused analysis of chromatin profiles tailored to these regions. In future, our unbiased computational approach can be utilized to define the regulomes of *T cell receptor* and other complex loci, including NK cell receptors and *MHC* loci (Shiina et al., 2004). Finally, this study demonstrates that we can use chromatin profiling and computational analyses to identify enhancers in our hands. Therefore, we pursued and translated these approaches to human B cell lymphoma in Chapter 4 to identify pathogenic enhancers and their associated genes.

Functional dissection of predicted cis-regulatory circuits in B cell lymphoma

In Chapter 4, we selected and systematically dissected a *cis*-regulatory circuit in B cell lymphoma that was predicted by computational methods. This circuit is composed of many augmented enhancers and multiple overexpressed genes, including *NEK6*, a mitotic kinase gene. We find that only a subset of predicted enhancers, excluding a super-enhancer, is required to maintain elevated *NEK6* expression. A CTCF cluster serves as a boundary element for cordoning off the *NEK6* regulatory hub from neighboring genes. Our work provides the framework for dissecting predictions of *cis*-regulatory circuits. In addition, our findings highlight the importance to rigorously test predictions regarding enhancers, super-enhancers and circuits generated by current computational algorithms, because a large portion of these predictions were not substantiated in our study.

Two novel features revealed from our work should be considered in order to improve the predictive power of algorithms defining *cis*-regulatory circuits, beyond the standard

correlative chromatin and expression patterns, coupled with low-resolution Hi-C data. Future studies should consider to measure enhancer activities in parallel using high-throughput assays (Kwasnieski et al., 2012; Melnikov et al., 2012). Using this approach, thousands of enhancer candidates or enhancer variants, as well as unique sequence tags, are precisely synthesized by microarray, cloned into reporter constructs, and transfected into cells. Expression levels of tags are subsequently measured by high-throughput sequencing, corresponding to the functionality and activities of tested enhancers. This information will not only screen out false positive enhancer candidates, improving the accuracy of circuit-predicting methods, but also will aid prioritization of circuits based on the robustness of associated enhancer activities. Besides a deeper understanding of enhancer activities, a better knowledge of chromatin structure will help in correctly connecting enhancers and target genes. Currently, chromatin domains are defined to be of different sizes, depending on the resolution of input data and the computational method applied (Hnisz et al., 2016). Because chromatin domains are considered to limit enhancer-gene interactions, wrongly assigned chromatin domains may lead to enhancers linked with false targets. In addition, low-resolution and noisy Hi-C data make it difficult to robustly identify chromatin loops with high-confidence. Because direct promoter-enhancer looping suggests functional interaction, errors in predicting chromatin loops will cause mistakes in connecting promoters and enhancers. These problems need to be addressed with upcoming higher-resolution chromatin interaction data and novel prediction algorithms (Whalen et al., 2016).

The most surprising and significant finding in our study is that a super-enhancer, SE1, is not required for expression of genes within its chromosomal neighborhood or within 5 Mb of its genomic space. This result challenges the current dogma that SEs are clusters of key regulatory elements required for high-level expression of genes essential for cell identity, cell function, and

oncogenesis. Our finding for SE1 highlights the necessity to functionally evaluate SEs predicted solely by computational algorithms. Although SE1 and a subset of other CEs are not required for expression of *NEK6* and neighboring genes, these regulatory elements may be required for expression earlier in B cell development or in initial stages of transformation. This possibility is currently intractable in primary human B cells, but can be evaluated by exploring whether chromatin pattern at SE1 correlates with gene expression in earlier human B cell stages.

A second surprise emerging from our studies concerns structural determinants of the *NEK6* regulatory hub. Deletion of either CE1, the dominant enhancer, or CS2-4, a CTCF cluster, does not affect chromosomal interactions within the *NEK6* locus. This suggests several possibilities regarding architectural determinants. First, CE1-promoter interaction is redundant structurally with CS2-4 looping to downstream CTCF sites, which can be tested by deleting both CE1 and CS2-4. Second, another element is the key determinant awaiting to be identified. Third, once the NEK6 region is loaded with active chromatin, homotypic chromatin interactions drive *NEK6* promoters and enhancers in spatial proximity (Lieberman-Aiden et al., 2009). The third possibility can potentially be studied by reverting active chromatin states at *NEK6* using chemicals or engineered sequence-specific TFs.

In this work, we focused on a pathogenic *cis*-regulatory circuit in lymphoma predicted *in silico*. In this circuit, we identified a potential oncogene, *NEK6*, overexpressed in human B cell lymphoma, and several regulatory elements that maintain the *NEK6* regulatory hub and expression. So how is NEK6 upregulated and does this upregulation directly contribute to lymphomagenesis? To answer this question, the biological relevance of NEK6 to B lymphomagenesis needs to be established. Our preliminary data suggest that depletion of NEK6 does not affect proliferation or survival in GM12878 and two B lymphoma cell lines. In contrast,

knocking down NEK6 in other cancer models attenuates cell growth (Fry et al., 2012). This lack of phenotype in B cells may stem from the compensatory role of its homolog, NEK7, which is also overexpressed in FL. Therefore, depletion of NEK7 may sensitize NEK6-deficient cells and reveal novel mechanisms for B cell transformation. In addition, mechanisms to augment *NEK6* enhancer activities remain to be defined. Our preliminary finding suggests that several TFs overexpressed in FL may contribute to higher enhancer activities, including MAX, MYC and STAT family members, which directly bind to *NEK6* enhancers in GM12878 cells (ENCODE Project Consortium, 2012).

The overarching theme of my dissertation is to identify new regulatory elements and assess their impacts on regulating expression of target genes using chromatin profiling coupled with computational algorithms. We predicted Vβ repertoires in mouse using a multivariate computational model, including features of chromatin states and chromosomal interactions. Then we classified and identified distinct regulatory potentials of new *cis*-elements in other mouse antigen receptor loci using tailored chromatin profiling analyses. Finally, we systematically dissected a pathogenic *cis*-regulatory circuit for *NEK6* in human B cell lymphoma, predicted by correlative chromatin states and chromosomal interactions. My dissertation presents a framework to predict and validate gene regulatory mechanisms, which could be extrapolated to nearly any locus or cell type. This work provides useful metrics to improve the power of computational algorithms predicting gene regulatory networks. Discrepancies between predicted and validated circuits underscore the need to functionally validate predictions as well as to generate features with higher-resolution and higher-throughput. Our findings provide biological insights into the functions of elements with different regulatory potentials, and yield a list of important candidate elements for in vivo studies. These discoveries in developing lymphocytes and transformed B

cells can be translated into studies on gene regulation in human B cell lymphoma, as well as many other malignancies. In conclusion, my dissertation paves the way for future investigations on the roles of *cis*-regulatory elements and chromatin architecture in normal development and disease.

References

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fry, A.M., O'Regan, L., Sabir, S.R., and Bayliss, R. (2012). Cell cycle regulation by the NEK family of protein kinases. J. Cell Sci. *125*, 4423–4433.

Han, J.-H., Akira, S., Calame, K., Beutler, B., Selsing, E., and Imanishi-Kari, T. (2007). Class switch recombination and somatic hypermutation in early mouse B cells are mediated by B cell and Toll-like receptors. Immunity *27*, 64–75.

Hnisz, D., Day, D.S., and Young, R.A. (2016). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. U. S. A. *109*, 19498–19503.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Livák, F. (2003). Evolutionarily conserved pattern of gene segment usage within the mammalian TCRbeta locus. Immunogenetics *55*, 307–314.

Medvedovic, J., Ebert, A., Tagoh, H., Tamir, I.M., Schwickert, T.A., Novatchkova, M., Sun, Q., Huis In 't Veld, P.J., Guo, C., Yoon, H.S., et al. (2013). Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. Immunity *39*, 229–244.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Reddy, K.L., Zullo, J.M., Bertolino, E., and Singh, H. (2008). Transcriptional repression

mediated by repositioning of genes to the nuclear lamina. Nature *452*, 243–247.

Shiina, T., Inoko, H., and Kulski, J.K. (2004). An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens *64*, 631–649.

Vincent-Fabert, C., Fiancette, R., Pinaud, E., Truffinet, V., Cogné, N., Cogné, M., and Denizot, Y. (2010). Genomic deletion of the whole IgH 3' regulatory region (hs3a, hs1,2, hs3b, and hs4) dramatically affects class switch recombination and Ig secretion to all isotypes. Blood *116*, 1895–1898.

Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496.