The Development of Chemical and Computational Tools to Study Transcriptional Regulation in Cancer

The Development of Chemical and Computational Tools to Study Transcriptional Regulation in Cancer

A dissertation presented

by

Alexander Joel Federation

to

The Committee on Higher Degrees in Chemical Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemical Biology

Harvard University

Cambridge, Massachusetts

April 2015

Dissertation Advisor: Professor James Eliot Bradner **Alexander Joel Federation**

The Development of Chemical and Computational Tools to Study Transcriptional Regulation in Cancer

Abstract

Eukaryotic gene regulation is a complex process requiring the action of many multicomponent complexes in the cell. Specific inhibitors of chromatin-associacted factors allow the functional study of protein domains without genetic removal of the entire protein. Here, two small molecule probes were used to study the role of DOT1L and BET proteins in cancer biology.

DOT1L is a histone methyltransferase with activity correlating with positive regulation of transcription. In MLL-rearranged leukemia, DOT1L is recruited abberently to early developmental transcription factors, leading to their inappropriate expression and leukemia manintenence. The development of an assay platform for DOT1L allowed the investigation of many small molelcule DOT1L inhibitors, leading to compounds with improved potency and pharmacokinetics.

Studying the action of BET bromodomain inhibitors led to the identification of super enhancers, large tiussue-specific regulatory elements driving the expression of genes criticial for the function of the cell. Super enhancers are often found in oncogenic transcloation events, especially in B cell malignancies. This study identified a subset of super enhancers that promote off-target DNA damage from the B cell antibody diversity enzyme AID, leading to double strand break events and translocations.

Super enhancers also regulate the expression of master transcription factors (TFs) in a given cell type. Using the topology of the super enhancer, the sites of master TF bindng can be predicted, allowing the construction of network models for transcriptional regulation. These models were built in a large number of healthy and diseased cell types, including the pedicatric malignancy medulloblastoma. In medulloblastoma, a network motif was identified that matches an expression pattern seen in a transient

cell population in the developing cerebellum, providing evidence for the previously unknown cell of origin for Group 4 medulloblasoma.

Table of Contents

Acknowledgements

"When you're in Disney Land, be sure to go on all the rides."

- James E. Bradner

I still remember hearing that advice from Jay in our first meeting after I had joined the lab. It's a philosophy that I've taken to heart during my time in graduate school, and it would have been impossible if it had not been for how welcoming and collaborative everyone in the lab has been during my time here. Jun Qi mentored me closely during my rotation in the chemistry lab and thankfully continued afterwards when I moved on to other disciplines. He was a close collaborator for Chapters 2 and 3 of this work. Chris Ott and Michael McKeown provided guidance while learning biochemistry, assay development and cell biology. Allowing me to jump into my project with the resource of their combined expertise accelerated my science tremendously.

My course in the Bradner Lab would have been much different if Charles Lin had not joined. Charles taught a computational biology class his first summer here where I first became engrossed in the subject. He continued to share his code and mentor me as I developed as a programmer and has shaped the way I now approach new problems in biology. Rhamy Zeid was my baymate for the majority of our time in the lab. I've learned a great amount from Rhamy's technical expertise, but more importantly he is a person I look up to and a great friend. Thank you also to all my other labmates, past and current, for their support and friendship.

Of course, Jay – I could not have asked for a better mentoring experience during my time in graudate school. From Jay, I've learned the importance of identifying problems in biology that might have a positive translational outcome. I've learned that the most important step in a successful scientific pursuit is to assemble a motivated, talented and goal-oriented team of scientists to work with. I've essentially learned what kind of scientist I'd like to someday be – one that rigorously undertakes impactful projects with openness and enthusiasm, with the goal of making a positive impact on human medicine.

All of the collaborations we've undertaken towards the projects discussed here have been wonderful experiences. The SGC in Chapter 2, the Alt Lab and Liu Lab in Chapter 5, the Young Lab in

Chapter 6 and DKFZ in Chapter 7 have all been tremendous opportunities for me to learn and meet incredible people in other fields of science.

Thank you to the institutions that have provided financial support during time as a student. The National Science Foundation and Ashford family provided funding for my stipend and allowed maximal freedom for me to learn all I possibly could during my time here. The National Institutes of Health R21 program funded the CBX chemical discovery effort, which unfortunately will not be presented in this dissertation.

The Chemical Biology Program – Jason Millburg, KeyAnna Schmiedl, and Samantha Reed have built a supportive graduate program that has quickly grown during my time here. Additionally, the Leder Program gave me the opportunity to see first-hand human translational biology and medicine in the GI cancer clinic, an experience that has framed my approach to all subsequent problems I've worked on.

My dissertation committee, composed of Prof. Timothy Mitchison, Prof. Nathanael Gray, Prof. Len Zon and Prof. Alex Meissner, has been an invaluable resource for me scientifically and personally. I look to all four of them as role models as truly exceptional scientists and I'm grateful for their time and investment in my education.

Teachers throughout my life have helped steer me down the path of scientific pursuit and I'm grateful to all of them. Thank you to Arnold Serotski, Shayne Watterson, Laura Coleman, Marc Fleming, Michael Snyder and Eric Rathfelder. I will always be indebted to Professors Thomas Krugh and Bradley Nilsson of the University of Rochester. They fostered my personal and scientific growth from my first days in college, giving me the resources and opportunities to thrive and the freedom to enable my growth as a young scientist. My graduate school career would have been much more difficult without their guidance, trust and support and during my college years.

I have to thank all of my friends that have supported me at various times during this journey. The University of Rochester cross country and track teams, as well as my other close friends from home, college and Boston have always been there to support me during the most difficult times. A special thank you must go to Megan Fritz, who was my closest supporter during most of this work and will join me as I move on this coming fall.

Lastly, thank you to my family – my Mom and Dad, Lindsay and Kevin.

Chapter 1

The Regulation of Eukaryotic Gene Expression

The development and function of the human organism requires precise temporal and spatial control of the estimated 25,000 genes contained in the human genome. The DNA encoding for these genes occupies only 1.5% of genomic space, allowing much of the remaining space to contribute to complex, combinatorial control of gene expression. Initial large-scale studies suggest that 80% of the genome participates in some measureable biochemical function[1]. The nature of these functions is the subject of current debate[2], however, these studies support a model in which large domains of the human genome are devoted to controlling the binding, initiation and elongation of RNA polymerase II (RNA PolII). This degree of regulation enables a single template of genomic information to produce the hundreds of cell types found throughout the many stages of human development.

The earliest studies of gene regulation in bacteria uncovered short DNA sequences necessary for the stimulus-dependent transcription of metabolic enzymes. Functionally related genes in prokaryotes are found in close proximity to each other, allowing their simultaneous regulation by nearby genetic elements. The promoter is defined as the DNA sequence immediately upstream of the transcription start site (TSS) capable of localizing RNA polymerase at beginning of the gene (-10 and -35 bp in E. coli)[3]. Operator sequences near the TSS can cooperate with the promoter to control the rate of transcription, often recruiting repressive transcription factors (TFs) that bind the operator and block the function of RNA polymerase.

Eukaryotic gene expression shares some of these features, notably the promoter element 5' of the TSS. In the human genome, however, genes are not organized in operons based on function. Instead, each gene is regulated by its own promoter, as well as through inputs from other nearby cis-regulatory elements. Enhancers are DNA sequences that can bind transcription factors and increase the level of transcription of a nearby gene while functioning irrespective of their orientation in the genome[4]. The current model of transcription postulates that enhancer elements loop to the promoter to cooperate in recruiting and regulating the rate of elongating RNA PolII (Figure 1.1). Other cis-elements have been reported that influence gene regulation, including insulating structural elements that block the action of enhancers. These also contribute to loop structures, though these loops typically isolate enhancer-promoter interactions from nearby inactive genes.

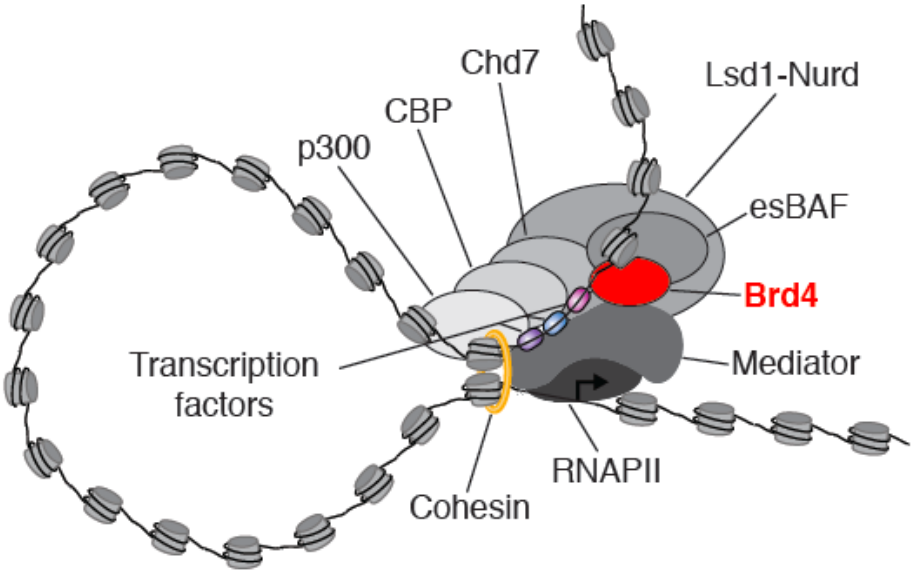


Figure 1.1: The looping model of chromatin. Transcription factors bind to enhancer and promoter elements to recruit various cofactors. These include BRD4, mediator, CBP/p300, as well as chromatin remodeling complexes (BAF, NuRD). Cohesion stabilized enhancer-promoter looping, allowing the recruitment of RNA PolII and it's eventual phosphorylation and productive elongration through the gene body.

As the molecules that directly interact with promoters, enhancers and insulators, many transcription factors (TFs) have been deeply characterized to understand their role in gene regulation. DNA binding is achieved through one of several conserved protein folds that interact directly with the major and minor grooves of DNA in a sequence-specific manner[5]. Recent systematic efforts have aimed to define these operator-binding sequences of the human transcription factors. When bound to the DNA, TFs can interact with and recruit other transcriptional cofactors through their effector domains to influence transcriptional activity, transcriptional dynamics, integrate signaling from upstream pathways and control the architecture and accessibility of chromatin. TF genes are critical for human development, can promote cellular reprogramming by their forced expression and are commonly altered in disease – exemplified by the two most commonly deregulated proteins in cancer – MYC and p53[6,7].

In eukaryotic transcription, an additional layer of regulatory information is encoded independently of DNA sequence through covalent modifications to DNA and associated proteins. The large size of eukaryotic genomes necessitates their compaction, and this is accomplished by the wrapping of DNA around histone particles. Histones are highly conserved basic complexes composed of four core particles arranged in two H2A-H2B dimers and a H3-H4 tetramer. The histones contain unstructured tail regions that do not directly interact with DNA which are heavily modified by nuclear enzymes. When wrapped with DNA, these modifications of chromatin can influence the density of histone packing, recruitment of TFs and cofactors, and consequently, the expression of nearby genes[8].

The first covalent modification to chromatin to be described is DNA cytidine methylation, a truly epigenetic modification in that it can lead to heritable changes in phenotype through generations of cell division[9]. Additional covalent modifications are found on the histone proteins themselves, including methylation as well as acetylation, phosphorylation, ubiquitination and crotonylation, with new modifications continuing to be discovered[10]. These modifications are enzymatically added and removed by histone modifying enzymes and recognized by chromatin “reader” proteins, analogous to the way signaling is propagated in the cytoplasm through kinase cascades[11]. Other mechanisms for gene regulation non encoded in DNA exist, including the action of noncoding RNAs and post-transcriptional regulation of RNA and protein translation, but these mechanisms were not studied in the course of this work.

The complexity of eukaryotic gene expression requires precise tools to interrogate the functions of different components in the system, which is a main goal of this work. The use of chemical tools to manipulate biological systems is referred to as *chemical biology*, and this is the main approach that will be utilized in this dissertation to study chromatin-associated proteins. Many chemicals in the scientific literature claim to specifically alter the function of a protein target, however, a useful research probe for mechanistic biological study must meet certain criteria for activity and specificity[12]. The goal of this thesis is to discover and characterize chemical probes of chromatin-associated proteins and use these probes to gain deeper understanding of the mechanisms of gene regulation. Summarized as the central hypothesis of this dissertation:

> *The perturbation of chromatin regulating complexes with small molecules will result in specific modulation of pathogenic transcriptional programs in models of human cancer. The mechanistic understanding gained by this study of transcription will allow generation of improved models of transcriptional regulation in human biology and disease.*

Testing this hypothesis requires two experimental approaches. First, small molecules that target transcriptional proteins of interest must by synthesized and characterized for their function. This thesis focuses on two small molecules targeting either the DOT1L methyltransferase or BET family of bromodomains. The following chapters will detail their chemical synthesis, structural characterization and functional characterization in various biochemical, cellular and animal models of disease.

In Chapter 2, a medicinal chemistry effort around a cofactor-mimicking scaffold for the DOT1L methyltransferase leads to the development of molecules that allow for crystallographic structure determination of the enzyme binding to small molecule inhibitors. These structures reveal a dramatic remodeling of the enzymatic pocket, which accounts for the high selectivity and long residence time of tis compound class. Chapter 3 builds upon these results, using the structural information to synthesize a chemical affinity probe for DOT1L. The use of this tool allows the development of two biochemical assays for DOT1L binding. These assays, coupled with a cell-based imaging assay were used to evaluate a focused library of compounds in a second round of medicinal chemistry. These efforts led to the

identification of inhibitors highly active in cellular systems as well as inhibitors with long *in vivo* stability for the study of DOT1L biology in animal models of cancer.

The second class of compounds discussed in this thesis will be introduced in Chapter 4. Inhibition of BET bromodomain proteins by JQ1 led to distinct effects in different cell types. Further investigation of this phenomenon uncovered a cis-regulatory element that is highly cell-type specific and exquisitely sensitive to BET bromodomain inhibition; these are termed super enhancers.

It is well established that chromosomal translocation can constitute a driving event during oncogenesis. Often these translocations involve the rearrangement of super enhancer elements, exemplified by the IgH-MYC translocations that characterize multiple myeloma. In B cell malignancies, off-target DNA damage activity of an antibody diversification enzyme called activation-induced cytidine deaminase (AID) leads to double strand breaks that can promote these oncogenic translocations to occur. Chapter 5 explores the role of super enhancers in AID off-target localization and proposes a set of epigenomic and transcriptional criteria to predict sites highly prone to AID off-target activity.

The tissue specific regulatory information encoded in super enhancers provided the rationale for Chapter 6. Super enhancers and their underlying sequences are used to search for relevant TF binding sites, allowing the prediction of TF-TF interactions. These interactions can be assembled into a network model of transcriptional regulation for a given cell type. These network models were then used to study medulloblastoma, a highly heterogeneous pediatric brain tumor (Chapter 7). The transcriptional models identified a small set of transcription factors highly specific for the cryptic Group 4 subset of medulloblastoma, leading the putative cell of origin for this tumor and the potential for model organism development.

Finally, this dissertation concludes with a short discussion on the history of systems pharmacology, the potential uses of cellular network models, how they might contribute to the field of chemical biology and to therapeutic development for cancer and other diseases of transcriptional deregulation.

Chapter 2

Catalytic Site Remodeling of the DOT1L Methyltransferase by Selective Inhibitors

Contributors: Wenyu Yu, Emma J. Chory, Amy K. Wernimont, Alex Scopton, Alexander J. Federation, Jason J. Marineau, Jun Qi, Dalia Barsyte-Lovejoy, Joanna Yi, Richard Marcellus, Roxana E. Iacob, John R. Engen, H Erno Wienholds, Fengling Li, Javier Pineda, Guille Estiu, Tatiana Shatseva, Taraneh Hajian, Rima Al-Awar, John E. Dick, Masoud Vedadi, Peter J. Brown, Cheryl H. Arrowsmith, James E. Bradner, Matthieu Schapira

Corresponding Supplementary Material can be found in Appendix A.

This chapter originally appeared in *Nature Communications*, Vol 3 (2012).

Introduction

Protein methyltransferases (PMTs) play essential roles in epigenetic regulation of gene expression and chromatin dependent signaling via their methylation of histones and other chromatin-associated substrates. Mutation or deregulation of PMTs is linked to many diseases, especially cancer, and there is strong interest in this family of proteins as potential drug targets[1-3]. Targeting the common S-adenosylemethionine (SAM) cofactor binding site of PMTs is an attractive strategy for this family, analogous to targeting the ATP binding site of protein kinases[4,5]. There are sixty PMTs encoded in the human genome including 51 lysine methyltransferases (PKMTs) and nine protein arginine methyltransferases (PRMTs)[1]. Most PKMTs contain a conserved SET-domain, with the exception of DOT1L which has a protein fold that more closely resembles PRMTs[6]. DOT1L is also unique in that it is the only PKMT to methylate histone H3 on Lysine 79 (H3K79)[7,8], a chromatin mark associated with active chromatin and transcriptional elongation[9].

Aberrant mono- and dimethylation of H3K79 by DOT1L is an essential step in the development of MLL-rearranged mixed lineage leukemia (MLLr), an acute form of the disease that, in infants, constitutes 70% of acute lymphoid and over 35% of acute myeloid leukemias[10-13]. MLLr is characterized by chromosomal translocations that result in an oncogenic fusion protein comprising the N-terminal region of MLL and the C-terminus of one of ~70 translocation partners[14], and pharmacological targeting of MLL translocation complexes was recently shown to reverse oncogenic activity of MLL fusion proteins in leukemia[15]. A subset of MLL fusion partners including AF4, AF9 and AF10 have been shown to aberrantly recruit DOT1L to select genomic loci leading to increased H3K79 methylation and transcriptional activation of genes essential for leukemogenesis[16]. Recently, the anti-leukemic activity of DOT1L inhibition by either genetic or pharmacologic approaches resulted in selective killing of MLL-AF4/9/10 translocation carrying cells [10,17]. In particular, EPZ004777 (Figure 2.1c) is a picomolar, specific, SAM competitive inhibitor of DOT1L that selectively kills cells bearing *MLL* chromosomal rearrangements and extends survival in a murine xenograft model of MLL[17]. DOT1L inhibition by EPZ004777 also accelerated reprogramming of somatic cells into induced pluripotent stem cells[18] suggesting that DOT1L inhibition may be useful in regenerative medicine.

To understand the structural mechanism of SAM competitive inhibition of PMTs, and of DOT1L in particular, we solved the structure of DOT1L in complex with EPZ004777, and uncovered novel and unexpected conformational variability of the cofactor binding site that can accommodate compounds significantly larger and more hydrophobic than SAM. We also present chemical analogs with improved solubility or potency including SGC0946 which will serve as a useful chemical probe of DOT1L activity[19]. These results provide important insight into SAM competitive inhibition of HMTs in general, and reconcile the contradictory observations that both very polar and very hydrophobic compounds can bind at the DOT1L cofactor site. Our data will also guide the design of new inhibitors with improved drug-like properties.

Results

The catalytic subunit of DOT1L is comprised by the first 416 residues[6]. Mono- and di-methylation of H3K79 is dependent on SAM (Figure 2.1a), which binds within a cofactor-binding site that is largely enclosed by an activation loop (residues 122 to 140; Figure 2.1b). Structural and biochemical studies by Rui-Ming Xu and colleagues characterize a 4 Å methyl transfer channel partitioning the SAM binding cavity from the putative substrate binding site, which is formed by a substrate binding loop (residues 301 to 311). This spatial organization is consistent with an in-line methyl transfer reaction, whereby a likely deprotonated acceptor lysine executes nucleophilic attack on the SAM methyl group[6].

EPZ004777 (Figure 2.1c) is a near chemical derivative of SAM, which has been reported to inhibit DOT1L with extraordinary potency in a radionucleotide homogeneous assay (IC_{50} = 400 pM), and possesses surprising selectivity for DOT1L compared to other SAM-dependent lysine and arginine methyltransferases[17]. Limited information is available regarding the design of EPZ004777, but most striking is the para-tert-butylphenyl appending group coupled via a urea linkage. Structural modeling studies of EPZ004777 bound to the SAM binding pocket of DOT1L [PDB:1NW3] failed to identify a ligand conformation that could accommodate this bulky substituent within the enclosed amino acid binding pocket. The only pose of EPZ004777 with a suitable fit within the DOT1L active site was accomplished by telescoping the urea capping feature through the methyl transfer channel and into the substrate-binding site (data not shown).

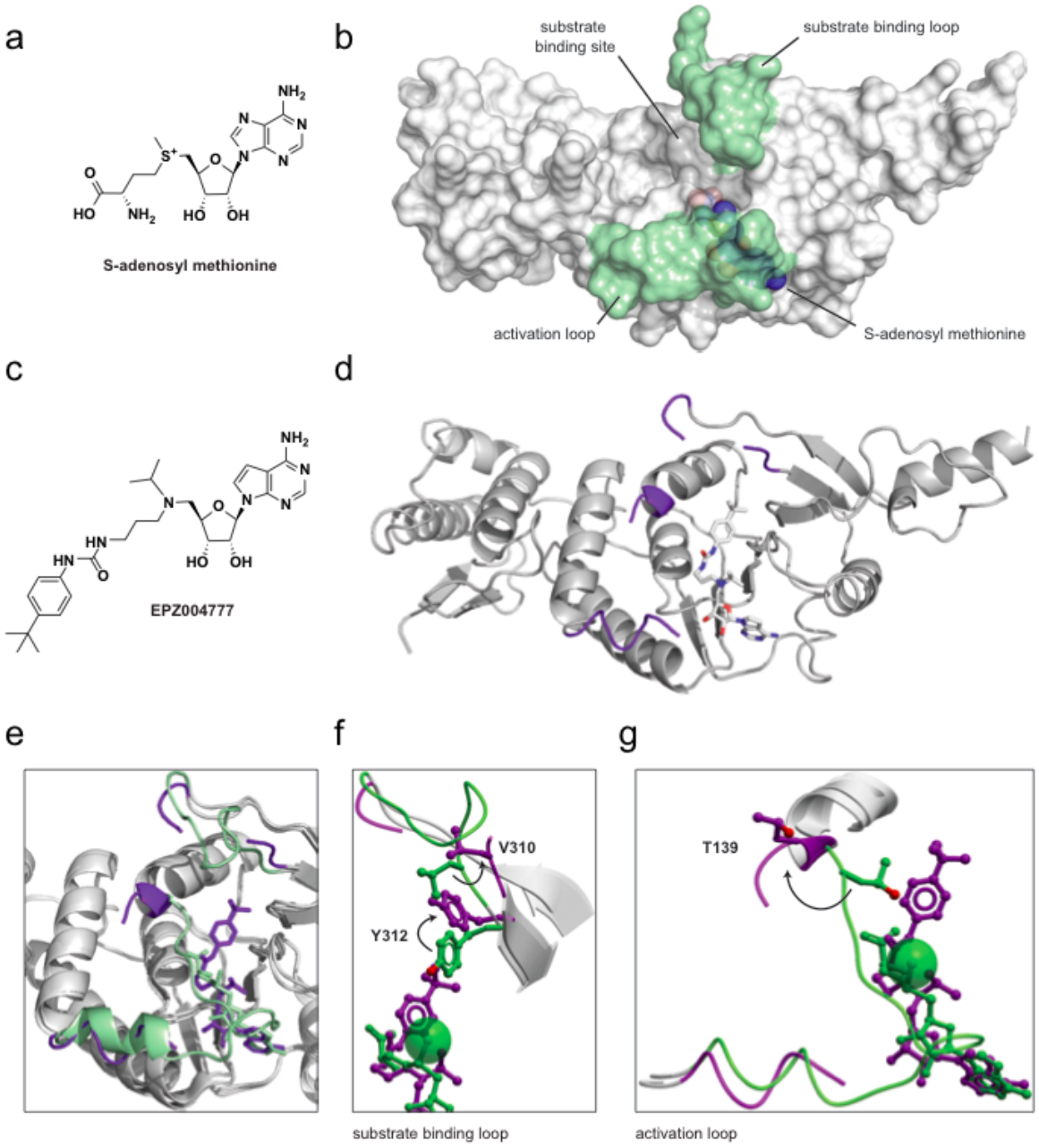


Figure 2.1: Structural mechanism of DOT1L inhibition. Binding of the cofactor SAM (a) stabilizes a catalytically competent conformation of DOT1L where the activation and substrate binding loops (green) guide the substrate towards the site of methyl transfer (b). Binding of EPZ004777 (c) affects the conformation of these loops which become partly disordered (d - magenta). EPZ004777 occupies the same site as SAM (e), but conformational rearrangement of substrate binding and activation loop residues (f, g respectively) are necessary to accommodate the t-butylphenyl group.

To definitively resolve the binding mechanism, we crystallized a complex of DOT1L bound to EPZ004777 by soaking a crystal of DOT1L in complex with the reaction by-product S-adenosylhomocysteine (SAH) in a solution of EPZ004777. The binary crystal structure of DOT1L in complex with EPZ004777 was solved to a resolution of 2.4 Å (PDB code 4ER3). The overall structure is very similar to that of DOT1L bound to SAM (1.1 Å backbone RMSD relative to PDB code 1NW3) with the exception of the activation and substrate-binding loops which are disordered in the EPZ004777-bound structure (Figure 2.1d). As expected, EPZ004777 occupies the cofactor-binding site, and consistent with molecular modeling the t-butylphenyl moiety of EPZ004777 cannot fit in the SAM-bound conformation of the cofactor pocket. Large structural rearrangements at Tyr312 and Thr139, in the substrate-binding and activation loops respectively, are necessary to accommodate this portion of the inhibitor (Figure 2.1f-g). Importantly, Tyr312 and Thr139 are considered key partners in forming the methyl transfer channel, and mutation of Tyr312 to Ala abrogates methyltransferase activity[6]. In the presence of EPZ004777, Tyr312 is rotated away from this channel to accommodate the t-butyl moiety (Figure 2.1f) while Thr139 is repositioned ~6 Å away from its SAM-bound state (Figure 2.1g). These conformational changes therefore disrupt the structural integrity of the catalytic pocket, and must necessarily propagate along the activation and substrate binding loops, possibly further disrupting these catalytically important structural features (Figure 2.1e).

Unexpectedly, EPZ004777 adopts an extended conformation in the complex structure, and multiple interactions scattered along the inhibitor likely contribute to its potent binding. The deazaadenosine moiety of the inhibitor recapitulates critical interactions with Glu186, Asp222 and Phe223 previously observed with SAM, SAH and close cofactor mimetics (Figure 2.2a-b)[5,6,20]. The t-butylphenyl end of the compound is surrounded by a cluster of hydrophobic side-chains (L143, V144, M147, V169, F239, V267, Y312). The urea linker is elegantly exploited via interaction with the carboxylic group of D161, and the protonated nitrogen of the isopropyl ammonium is favorably engaged in a hydrogen-bond with the backbone carbonyl of G163. However, it is not clear whether the hydrophobic isopropyl moiety is interacting favorably with the activation loop, which is disordered in our structure. Finally, a hydrophobic cleft composed of side-chains from F223, F245 and V249 located in close proximity to position 7 of the deazaadenine ring remains unexploited (Figure 2.2b).

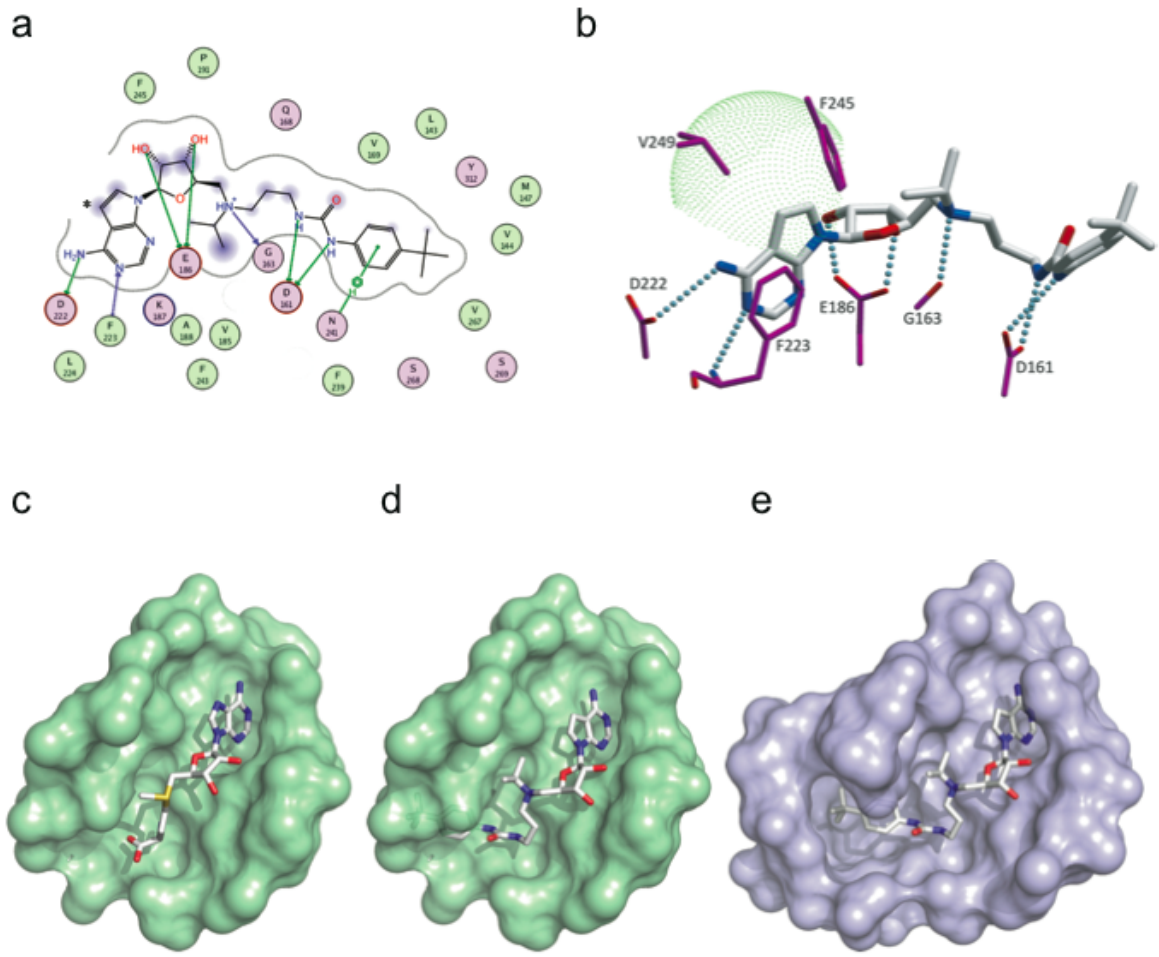
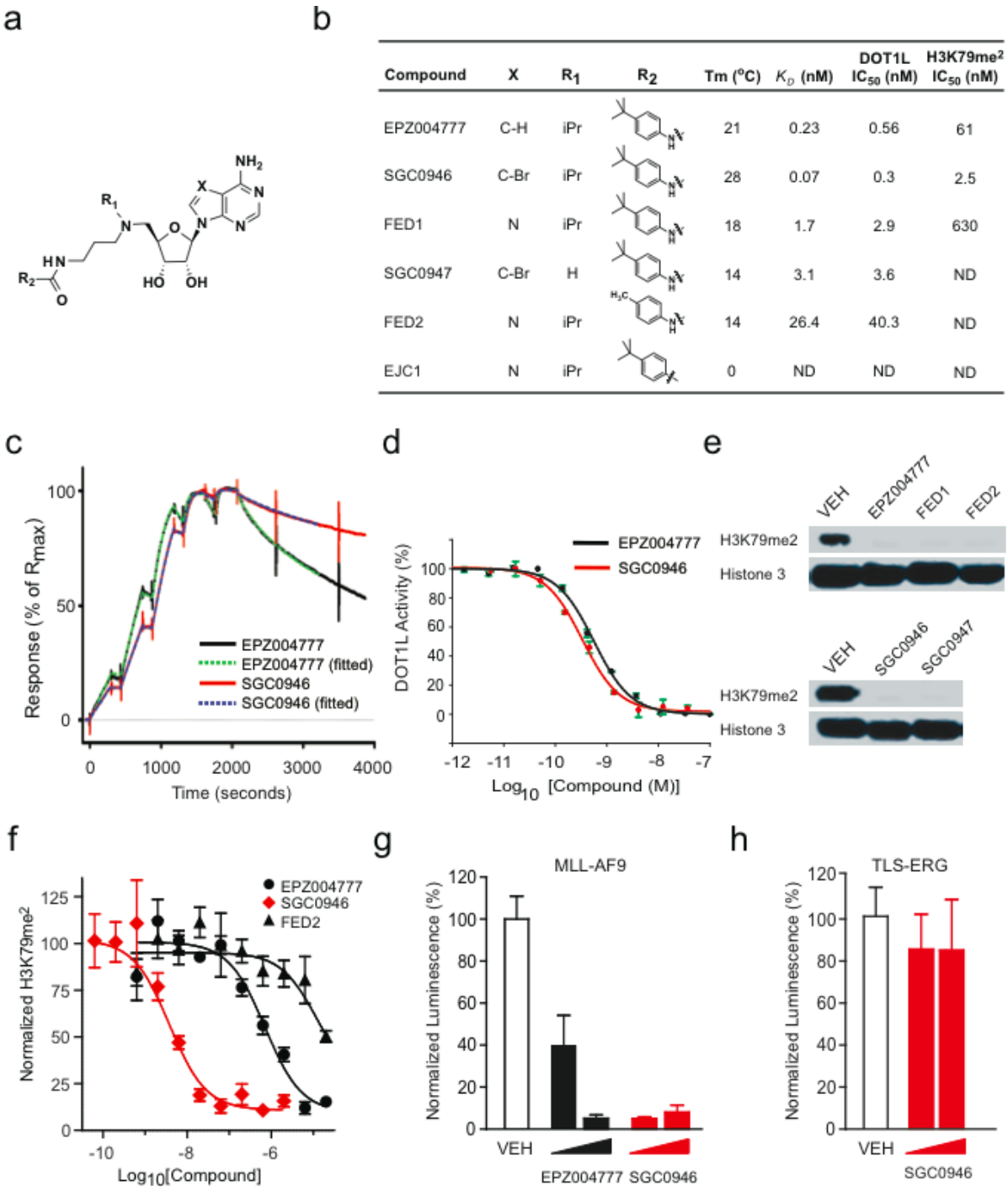


Figure 2.2: Molecular recognition of the DOT1L cofactor pocket by EPZ004777. (a) Two-dimensional projection of DOT1L-EPZ004777 interactions generated with MOE (Chemical Computing Group, Montreal, Canada). Residues within 4.5 Å of the ligand are shown, unless they form no hydrogen-bond and their side-chains are pointing away from the inhibitor. Gray line: DOT1L surface envelope; blue shading: solvent accessibility; hydrogen bonds with DOT1L side-chain/backbone atoms: green/blue arrows; green aromatic rings symbolize arene-cation interactions; amino-acid color code is pink: polar, green: hydrophobic, blue rim: basic, red rim: acidic. Position 7 of EPZ004777 is indicated by a star. (b) Three-dimensional representation highlighting hydrogen-bonds that mediate the interaction (doted lines) and a hydrophobic cluster (green sphere) that is not exploited by EPZ004777. (c) cofactor binding site in complex with SAM (PDB code 1NW3). (d) The t-butylphenyl group of EPZ004777 does not fit in the SAM-bound conformation of DOT1L. (e) Conformational rearrangements of DOT1L open-up a cavity to accommodate the t-butylphenyl group.

Overall, comparison of DOT1L bound to SAM (1NW3) and EPZ004777 (4ER3) reveals structural remodeling of the cofactor-binding site (Figure 2.2c-e). SAM exhibits excellent shape complementarity with the internal face of the binding pocket (Figure 2.2c), whereas EPZ004777 is incompatible with this pocket geometry (Figure 2.2d). Visualization of the internal face of the SAM binding pocket in the EPZ004777 complex reveals a new internal cavity formed around the t-butylphenyl feature (Figure 2.2e). Thus, our structure clearly indicates positions – such as the urea and t-butylphenyl group – that make favorable interactions, and other positions – such as the pseudo adenosine moiety – that can be further explored to improve potency.

To explore putative determinants of molecular recognition, we prepared a focused library of analogous molecules to EPZ004777 for biochemical and biophysical study (Figure 2.3a-b; Supplementary Figure A.2), using an efficient seven step synthesis. Taking advantage of the hydrophobic cleft in DOT1L surrounding position 7 of the adenine ring, we prepared SGC0946, an EPZ004777 analog with a bromine atom at position 7, which is expected to increase potency by occupying this cleft. Indeed, SGC0946 is a more potent DOT1L inhibitor than EPZ004777. Surface plasmon resonance and a homogeneous biochemical assay with recombinant, purified DOT1L acting on a nucleosomal substrate yielded a K_D of 0.23 nM and IC_{50} of 0.5 ± 0.1 nM for EPZ004777 (in agreement with previous work[17]), while a K_D of 0.07 nM and IC_{50} 0.3 ± 0.1 nM were measured for SGC0946 (Figure 2.3b-d). The gain in potency was even more substantial in cells: the brominated compound reduced the level of methylation of H3K79 in A431 cells with an IC_{50} of 2.5 nM, compared with an IC_{50} of 61 nM for EPZ004777, as established by quantitative, automated epifluorescence microscopy (Figure 2.3f). Bromination of the adenine ring did not affect selectivity: like EPZ004777, SGC0946 was inactive against a panel of 12 protein methyltransferases and DNMT1 (Supplementary Figure A.1). We found both compounds inactive against PRMT5, while an IC50 of ~500 nM was originally reported for EPZ004777[17]. This discrepancy may come from the fact that our assay is based on the PRMT5-MEP50 complex, while the previously reported inhibition was for isolated PRMT5.



Compound	X	R_1	R_2	Tm (°C)	K_D (nM)	DOT1L IC_{50} (nM)	H3K79me² IC_{50} (nM)
EPZ004777	C-H	iPr		21	0.23	0.56	61
SGC0946	C-Br	iPr		28	0.07	0.3	2.5
FED1	N	iPr		18	1.7	2.9	630
SGC0947	C-Br	H		14	3.1	3.6	ND
FED2	N	iPr		14	26.4	40.3	ND
EJC1	N	iPr		0	ND	ND	ND

Figure 2.3: Biochemical and cellular characterization of DOT1L inhibitors. (a, b) EPZ004777 and analogs define an emerging structure activity relationship. (c) Biacore SPR sensorgram from single cycle kinetics runs with 5 concentrations, normalized to the calculated Rmax, indicate that EPZ004777 binds DOT1L with 1:1 stoichiometry and a dissociation constant (K_D) of 0.23 nM. (d) EPZ004777 inhibits DOT1L enzymatic activity with an IC_{50} of 0.56 nM. (e) Inhibition of H3K79me2 by a series of DOT1L inhibitors by immunoblot (10 μM except SGC0946, 1 μM; 4 days), and (f) high-content imaging performed in dose-response. SGC0946 (1 and 5 μM) kills human cord blood cells transformed with an MLL-AF9 fusion oncogene, while EPZ004777 was less efficacious at 1 μM (g). SGC0946 (1 and 5 μM) does not affect viability of cord blood cells transformed with an unrelated oncogene, TLS-ERG (h).

Structural advantages to the position 7 substitution of the basic adenine nitrogen with carbon in EPZ004777 were not identified. Indeed, this modification serves to impair solubility and increase the complexity and expense of the synthetic effort. We therefore prepared the adenine derivative, FED1, which demonstrates potent inhibitory activity for DOT1L in homogeneous (IC_{50} = 2.9 +/- 0.2 nM) and cell-based (IC_{50} = 0.63 μM) assays (Figure 2.3b, Supplementary Figure A.2). To confirm that the hydrophobic interactions at the t-butyl group of EPZ004777 contribute significantly to binding, we synthesized FED2, a derivative of FED1 in which the t-butyl was replaced by a methyl group. FED2 exhibited a reduction in target potency in biochemical (IC_{50} = 40.3 +/- 7.3 nM), biophysical (K_D = 26.4 nM) and cellular (IC_{50} > 1 μM) assays. We also investigated the importance of the isopropyl group by measuring the effect of its deletion from SGC0946. The resulting compound (SGC0947) had a ten-fold increase in IC_{50}, which reflects a non-negligible contribution of this group to the interaction. Finally, we confirmed the essentiality of the urea linking feature by synthesizing and testing EJC1, which lacks demonstrable inhibitory or binding activity for DOT1L.

We next tested the active compounds, EPZ004777, FED1, FED2, SGC0946 and SGC0947, in several MLL and non-MLL cell lines in order to assess their effects on DOT1L cellular function. Depletion of H3K79me2 was evident at 48 hours for all compounds in MV4;11 leukemia cells expressing an MLL/AF4 fusion protein (Figure 2.3e). Similarly, EPZ004777 and SGC0946 both showed time and dose dependent reductions in the H3K79me2 mark in the Molm13 MLL cell line that has the MLL/AF9 translocation (Supplementary Figure A.3). Quantitative assessment of H3K79me2 levels as measured by automated epifluorescence microscopy in A431 cells (Figure 2.3b,f) showed a substantially improved DOT1L inhibitory potency of SGC0946 (IC_{50} = 2.5 nM) compared to EPZ004777 (IC_{50} = 61 nM). Similar cellular potencies for SGC0946 (IC_{50} = 6 nM) and EPZ004777 (IC_{50} = 97 nM) were also observed in MCF10A breast epithelial cells (data not shown). We attribute this increase in cellular potency to the lower K_D and extended residence time of SGC0946 (Figure 2.3b; Supplementary Figure A.2)[21]. Consistent with a biologically selective effect on DOT1L, SGC0946 displayed selective reduction of cell viability in an experimental leukemia model derived from human cord blood cells transformed with the MLL-AF9 fusion oncogene[22], without apparent effect on the viability of cells transformed with an MLL-unrelated translocation (TLS-ERG)[23] (Figure 2.3g-h). SGC0946 also showed increased efficacy in this

model relative to EPZ004777. Taken together, these data demonstrate a consistent trend between *in vitro* biochemical and biophysical activity, and selective, on-target cellular activity of DOT1L inhibitors.

To verify that the binding mode of EPZ004777 is conserved with that of its chemical analogs, the DOT1L structure was solved in complex with the four active derivatives. As expected, the compounds recapitulated the EPZ004777 binding pose (Figure 2.4a). Strikingly, however, each of the four structures captured the activation and substrate binding loops in a different conformation. The activation loop was fully ordered only in the DOT1L-FED2 complex while the substrate-binding loop was ordered in all four complex structures (Figure 2.4a). While the original DOT1L-EPZ004777 structure was obtained by soaking a crystal of the DOT1L-SAH complex in a solution of EPZ004777, this new series of structures was produced by soaking crystals of DOT1L in complex with 5-iodotubercidin, a kinase inhibitor that we had previously identified as weakly inhibiting DOT1L (IC_{50} = 18 ± 0.9 μM; Supplementary Figure A.1). Unlike the crystals of the DOT1L-SAH complex, which generally broke during soaking and rarely allowed displacement of SAH by the inhibitors, we found soaking experiments much more efficient with the DOT1L-5-iodotubercidin crystals. Surprisingly, when we soaked the latter in a solution of EPZ004777, we found two molecules of EPZ004777 bound to a single DOT1L molecule. One EPZ004777 ligand was in the SAH binding pocket, in the same pose as previously described, and a second EPZ004777 molecule was bound in a pocket generated by yet a different conformation of the activation loop. Taken together, these results indicate that the activation and substrate binding loops are capable of adopting highly variable conformations from one crystal structure to another. Indeed, within the structured regions of the activation and substrate binding loops, variability in atomic coordinates was evident also by crystallographic B-factors (Figure 2.4c-e).

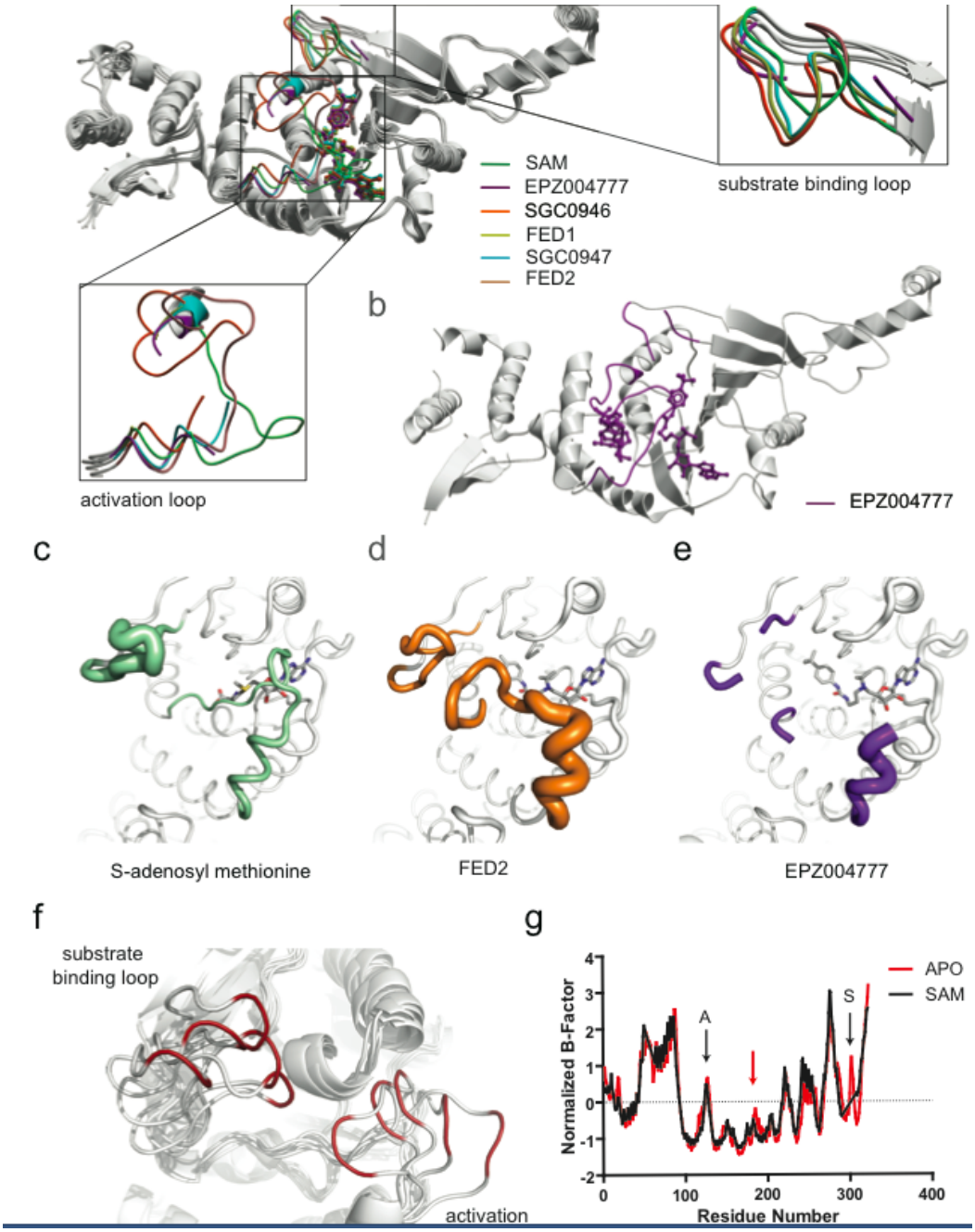


Figure 2.4: Conformational variation of the DOT1L activation and substrate binding loops. (a) The activation and substrate binding loops adopt diverse conformations in complex with SAM (green), EPZ004777 (magenta), SGC0946 (orange), FED1 (yellow), SGC0947 (cyan), and FED2 (beige). (b) Soaking of the 5-iodotubercidin-bound crystals with EPZ004777 results in a complex that contains a second binding pocket occupied by a second EPZ004777 molecule. (c-e) Variability in atomic coordinates is evident in loop regions by crystallographic B-factors as illustrated by ribbon thickness in (c) SAM, (d) FED2 and (e) EPZ004777 complexes. (f) MD simulation of apo DOT1L (1NW3) shows a diverse population of loop conformations. (g) Quantification of protein flexibility through calculation of normalized MD B-factors identifies regions of high mobility in the activation loop (arrow labeled “A”) and substrate binding loop (black arrow labeled “S”). SAM binding reduces MD B-factors near to the open SAM nucleoside binding pocket.

In all structures, the crystal lattice is such that adjacent DOT1L molecules are in contact with both the activation and the substrate binding loops. Thus, our observations of conformational heterogeneity of the loops and potential 2:1 binding stoichiometry derived from crystal structures might be influenced by surface contacts in the crystal lattice. Therefore we performed solution state binding studies in order to better understand these properties. Surface plasmon resonance indicated unambiguously that the DOT1L:inhibitor complex has a binding stoichiometry of 1:1 in solution for EPZ004777 and its four analogs (Supplementary Figure A.2). This result was also confirmed by isothermal titration calorimetry for EPZ004777 (Supplementary Figure A.4). The second inhibitor-binding site observed in one of our DOT1L-EPZ004777 complex structures is therefore most likely an artifact of the high concentration of inhibitor used for soaking, and/or crystal packing contacts. Nevertheless these structures clearly show that the substrate and activation loops of DOT1L are able to adopt a wide variety of conformations, suggesting a significant degree of flexibility and/or conformational mobility.

To explore the dynamic features of DOT1L, we performed 23 ns and 31 ns molecular dynamics (MD) production runs on DOT1L in the presence and absence of SAM, respectively. The initial coordinates were taken from crystallographic data. For the dynamics of apo DOT1L, the coordinates of the cofactor were deleted from the protein-cofactor complex (PDB: 1NW3). The system was stable (average rms of 3.1 Å; Supplementary Figure 5), and revealed a large conformational flexibility of the activation loop and substrate binding loop as shown by the superposition of several snapshots saved along the MD (Figure 2.4f) and quantified by the calculation of the theoretical B-factor per residue (computational approximations of the atomic displacement parameter; Figure 2.4g). The maximum flexibility is centered around residue W301 for the substrate-binding loop, and around residue P126 for the activation loop. A residue-by-residue comparison of molecular dynamical B-factors obtained from apo and SAM-bound simulations identified a putative region of conformational stabilization by engagement of the cofactor binding site, spanning residues 183-193, which defines a portion of the substrate binding pocket that engages the nucleoside features of SAM and analogous small-molecule inhibitors (Figure 2.4g). These data support a model in which ligand binding partially organizes a highly flexible DOT1L catalytic site.

With an interest in experimentally interrogating the conformational flexibility of DOT1L, we performed hydrogen exchange mass spectrometry (HX MS) on apo DOT1L, as well as saturated complexes with selective small-molecule inhibitors and the cofactor product, S-adenosylhomocysteine (SAH). With HX MS, the relative dynamics within a protein are monitored by measuring the exchange of backbone amide hydrogens with the bulk solvent[24]. First, experimental methods were developed to provide adequate DOT1L catalytic domain protein coverage (> 98%) following pepsin digestion (Supplementary Figure A.6). We then sought to investigate if conformational or dynamic changes occur in DOT1L upon inhibitor binding and where the changes were located with respect to the substrate-binding site. Three inhibitors defining a broad range of target-specific biochemical inhibition for DOT1L were selected for study using HX MS: SAH, FED2 and EPZ004777. Hydrogen exchange was measured for each of the three inhibitors and the results were compared with the hydrogen exchange of DOT1L alone. As shown in Figure 2.5a, apo DOT1L is a dynamic protein that exchanges amide hydrogen atoms rather rapidly. Notably, rapid exchange was evident in both the substrate binding and activation loops at the earliest time point following deuterium exchange (1 sec), indicating a relatively higher degree of solvent exposure and/or conformational flexibility consistent with crystallographic findings and computational predictions by MD. Exchange, and therefore dynamics, was comparable between apo DOT1L and DOT1L incubated with SAM-competitive inhibitors for the majority of the protein throughout the time-frame of the experiment (4 hours). However, certain regions demonstrated ligand-associated reduction in hydrogen exchange, consistent with meaningful changes in protein dynamics upon inhibitor binding.

Regional changes in HX MS with ligand binding were observed most frequently near the cofactor binding site, consistent with the mode of binding determined by crystallography (Figure 2.5b, c). For example all compounds contributed marked decreases in exchange in the region of peptide (188-199), consistent with structural and MD data which show improved conformational stability attributable to SAM nucleoside binding (Figure 2.5c, d). Further stabilization of the interior face of the SAM binding pocket was proportionate to inhibitor potency, evidenced by decreasing exchange in peptide (218-224) and peptide (160-172) in the presence of SAM, FED2 and EPZ004777 (Figure 2.5d, Supplementary Figure A.6). Upon binding by the potent DOT1L inhibitor EPZ004777, considerably stronger effects on HX MS were observed in peptide (293-304), which is precisely the structured region of the substrate-binding loop

resolved by co-crystallographic studies as reorganized to accommodate the t-butylphenyl group (Figure 2.5c, d).

Taken together, our crystal structures, biochemical analysis and molecular dynamics simulations indicate that the activation and substrate binding loops of DOT1L are inherently flexible and that the cofactor and substrate binding sites can undergo dramatic conformational remodeling to accommodate a variety of hydrophobic ligands.

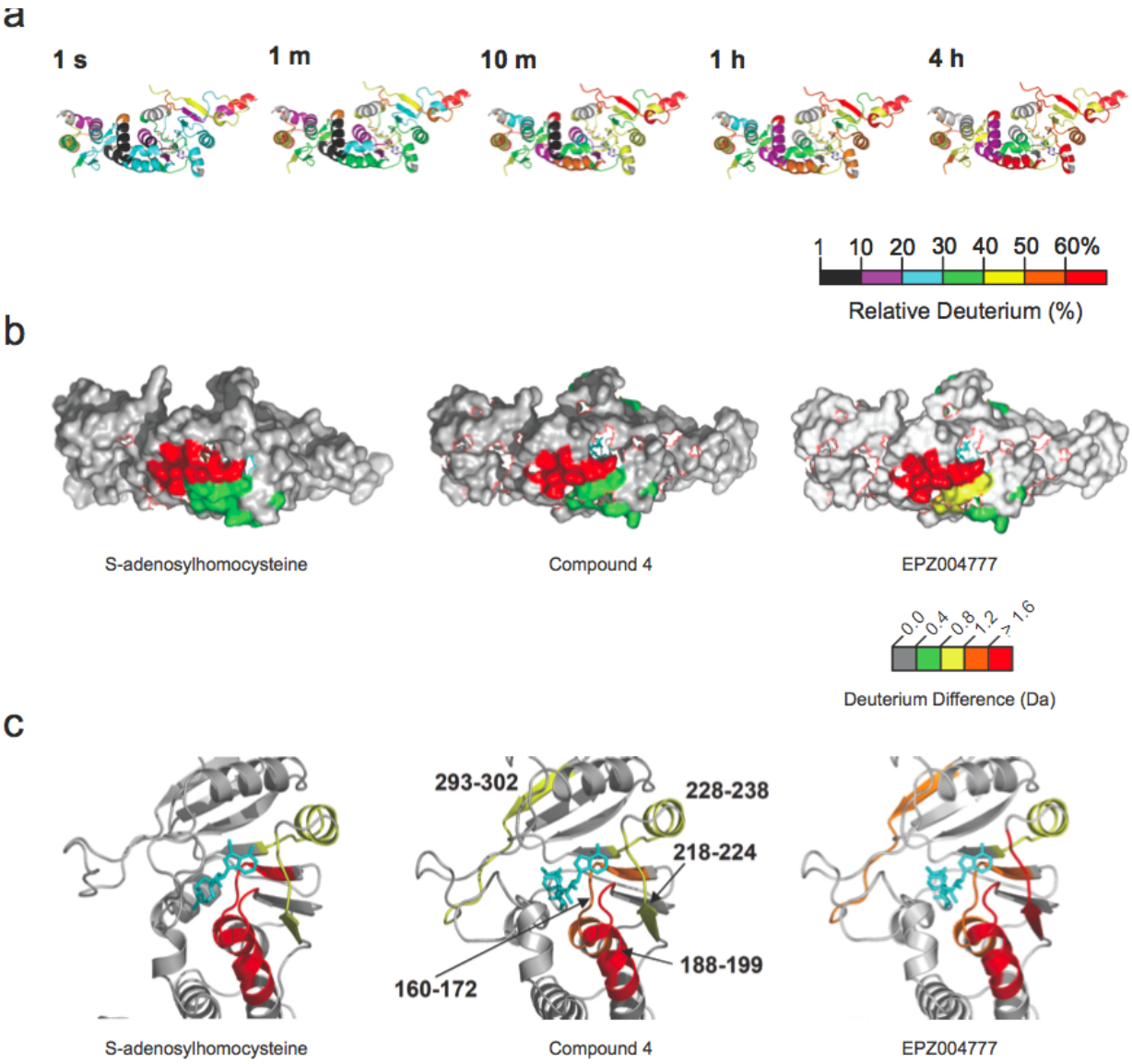


Figure 2.5: Hydrogen-deuterium exchange. (a) Summary of all HX MS data for DOT1L alone mapped onto the crystal structure of DOT1L. (b, c) Comparison of deuterium exchange with DOT1L inhibitor binding, as shown in a (b) space-filling model and (c) a ribbon model at higher magnification. The relative deuterium levels for all residues of DOT1L that were probed with hydrogen exchange are mapped with color for each exchange point; the color code is explained at the bottom of each figure. Regions colored gray after exchange began represent residues where deuterium levels were not determined. (d) The deuterium uptake curves for 3 representative peptides are shown. The maximum of the *y* axis in each graph is the maximum amount of deuterium that could be incorporated. Because these experiments were performed as comparisons under identical experimental conditions no corrections have been made for back-exchange thus the absolute value of each deuterium level is 18–25% higher than plotted based on totally deuterated standards [described in detail in[24]. The cumulative error of measuring deuterium uptake in these assays is approximately ± 0.20 Da. Any differences larger than this value were considered significant for the purposes of comparing the datasets[39]. The location of each peptide, according to the labels, is shown on the crystal structures in (c). All HX MS data are shown mapped to PDB: 1NW3, for ease of comparison.

Discussion

The DOT1L-inhibitor structures presented here are the first structures of a protein methyltransferase in complex with cofactor-competitive inhibitors that are active in cells. Integrated structural, biochemical, computational and biophysical analysis of DOT1L bound to a focused library of informative small-molecule inhibitors has defined a unique mode of inhibition, associated with active site remodeling and local stabilization of a highly dynamical protein target. Based on the previous structural data of SAM-bound DOT1L[6], the high affinity, SAM competitive binding of EPZ004777[17] was puzzling for two reasons. First, EPZ004777 is a very hydrophobic compound, while the SAM pocket is hydrophilic. Second, it is significantly larger than SAM, and therefore would not be expected to fit into the SAM-binding pocket (Figure 2.2d). Thus, binding of EPZ004777 to the cofactor site of DOT1L seemed to defy basic principles of shape and electrostatic complementarity.

While the SAM binding pocket of many protein methyltransferases is expected to be druggable[4], the high polarity of the pocket represents a challenge for the design of small-molecule inhibitors sufficiently hydrophobic to cross cell membranes. EPZ004777 has emerged as an important first-generation selective inhibitor of DOT1L, exhibiting high target potency in cellular assays perhaps attributable to its hydrophobicity. However, this SAM analogue possesses limited solubility (50 μM in aqueous solution). Recently, additional mimetics of SAM were reported as nanomolar inhibitors of DOT1L, but no cellular activity was provided, perhaps owing to the high polarity of the compounds[20] which would likely confer poor cellular permeability. Guided by crystallographic and computational data characterizing a focused library of DOT1L inhibitors, we observe remodeling of the catalytic site to accommodate bulky hydrophobic groups tethered by a urea linkage. As evidenced by the FED2 structure, for example, hydrophobic DOT1L inhibitors induce the formation of a largely enclosed binding pocket which is on average 68% more hydrophobic than the corresponding site in all available structures of DOT1L in complex with SAM, SAH, or close analogs (Supplementary Figure A.7).

The poor pharmacokinetic properties of EPZ004777 limit its utility as a chemical probe in the *in vivo* setting where much of the DOT1L developmental and disease biology is studied[10,11,25]. Indeed, in the article first reporting EPZ004777, a pharmacologic target validation study performed in a murine model of MLL required administration by osmotic pumps implanted subcutaneously to overcome evidently poor

pharmacokinetics[17]. Our structural analysis clearly indicates positions of the compound that could be exploited to improve its pharmacokinetics properties and potency. Here we show that SGC0946 has improved potency in cellular assays measuring the levels of H3K79 methylation. SGC0946 also has improved efficiency at specifically reducing the cell viability of a human MLL cell line derived from human cord blood that is transformed with the MLL-AF9 translocation oncogene[22]. SGC0946 specifically killed these MLL translocated cells but not cells transformed with another leukemogenic fusion protein[23].

Inhibition of histone methyltransferases is a promising new avenue for therapeutic discovery[1,3], and potent, selective and cell permeable reagents are urgently needed to link pharmacologic inhibition of specific PMTs with desirable phenotypes. We show here that remodeling of the DOT1L cofactor pocket allows competition by chemical inhibitors with physico-chemical properties distinct from and more drug-like than those of SAM. We suggest that this phenomenon may be a general principle that could be exploited for other PMTs because the post-SET element of SET-domain methyltransferases, and the a-X helix of PRMTs, both located at the SAM binding site, have been shown to be flexible and dynamic regions that adopt a catalytically competent conformation upon SAM binding[26,27].

We show here that remodeling of the DOT1L cofactor pocket allows competition by chemical inhibitors with physicochemical properties very distinct from those of SAM. Similarly, the structural variability of the post-SET element of SET-domain methyltransferases, and of the a-X helix of PRMTs, both located at the SAM binding site [26,27], may be exploited towards the design of drug-like SAM competitors for these distinct classes of methyltransferases.

The potent DOT1L inhibitor, SGC0946, presented here demonstrates how limited chemical modification can significantly improve cellular activity, probably a direct effect of a lower K_D and extended residence time[21]. Importantly, this compound will serve as a useful chemical probe[19] to further investigate the cellular mechanism of DOT1L in both normal and diseased cells. Furthermore, structural resolution of the mode of molecular recognition of DOT1L by prototype inhibitors using data and methods outlined here will guide development of improved compounds and will accelerate the development of clinical candidates.

Materials and Methods

Protein expression and purification

To produce recombinant DOT1L fragments containing N-terminal 351 and 420 amino acids in *E. coli*, the corresponding cDNA fragments were amplified by PCR and cloned into a modified pET28-MHL vector with an N-terminal 6-His tag. The proteins were overexpressed in *E.coli* BL21 (DE3) V2R-pRARE in Terrific Broth medium in the presence of 50 µg/mL kanamycin and chloramphenicol. Cells were grown at 37°C to an OD_{600} of 1.5 and induced by isopropyl-1-thio-D-galactopyranoside (IPTG, final concentration 1 mM) and incubated overnight at 15°C. The cell pellets were frozen in liquid nitrogen and stored at -80°C. For purification, the cell paste was thawed and resuspended in lysis buffer with 1mM phenylmethyl sulfonyl fluoride (PMSF). Slightly different purification protocols were used for DOT1L (1-420) and DOT1L (1-351). DOT1L (1-420) was purified by Ni-NTA column (Qiagen) and processed by in-house produced TEV protease to remove the His tag. The protein was then incubated in 50 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$ with benzonase nuclease for 2 hours at room temperature to remove DNA which binds to the C-terminal region of DOT1L (1-420), but not DOT1L (1-351). The filtered protein sample was diluted with 50 mM K_iPO_4 pH 7.0, and further purified by HiTrap-SP (GE Healthcare). The protein was finally purified by gel filtration (Superdex 200, GE Healthcare). For purification of DOT1L (1-351), only Ni-NTA affinity chromatography and size exclusion chromatography were used.

DOT1L purified by above-mentioned methods always contains co-purified SAM as reported[6]. This partially occupied form of DOT1L is applicable for use in the enzyme assay, crystallography and SPR experiment. However, for ITC and hydrogen-deuterium exchange experiments, DOT1L(1-420) in apo form is required. To obtain SAM-free DOT1L, 2 mL of 10mg/mL partially occupied form of DOT1L (1-420) was incubated with 1.0 mM 5-iodotubercidin for one hour at room temperature. The sample was then concentrated to 0.5 mL and then diluted with 50 mM K_iPO_4 pH 7. The diluted sample was loaded onto a HiTrap SP FF column (GE Healthcare), and then washed with 4 liters of buffer (50 mM K_iPO_4, pH 7.0) with a flow rate of 5 mL/min to remove SAM before elution with NaCl. The collected protein fraction was further purified by gel filtration (Superdex 200, GE Healthcare). SAM-free DOT1L(1-420) shows an A260/A280 around 0.57, while the partially occupied form of DOT1L shows an A260/A280 around 0.66.

Compound Synthesis

Reactions were run as described in the individual procedures using standard double manifold and syringe techniques; glassware was dried by baking in an oven at 130 °C for 12h prior to use. Solvents for reactions were purchased anhydrous from Sigma-Aldrich and used as received; the only exception being EtOH, which was stored over 4 Å molecular sieves. HPLC grade solvents were used for aqueous work ups and chromatography. Reagents were used as received. Reactions were monitored by thin-layer chromatography using EMD silica gel 60 F_{254} (250-micron) glass-backed plates (visualized by UV fluorescence quenching and staining with $KMnO_4$) and by LC-MS using a Waters Aquity BEH C18 2 x 50 mm 1.7 μm particle column (50 °C) eluting at 1 mL/min with H_2O/acetonitrile [0.2% v/v added formic acid or concentrated $NH_4OH_{(aq)}$ solution; 95:5(0min)→1:99(3.60min)→1:99(4.00min)] using alternating positive/negative electrospray ionization (125-1000 amu) and UV detection (210-350 nm). Flash column chromatography was carried out using Merck grade 9385 silica gel 60 Å pore size (230-400 mesh). Melting points were obtained using a capillary melting point apparatus and are uncorrected. ^{1}H NMR spectra were recorded at 400 MHz on a Bruker spectrometer and are reported in ppm using the residual solvent signal (dimethylsulfoxide-d_6 = 2.50 ppm; chloroform-d = 7.27 ppm; methanol-d_4 = 3.31 ppm; dichloromethane-d_2 = 5.32 ppm) as an internal standard. Data are reported as: {(δ shift), [(s = singlet, d = doublet, dd, doublet of doublets, ddd = doublet of a dd, t = triplet, quin = quintet, sept = septet, br = broad, ap = apparent), (*J* = coupling constant in Hz) and (integration)]}. Proton-decoupled ^{13}C NMR specta were recorded at 100 MHz on a Bruker spectrometer and are reported in ppm using the residual solvent signal (chloroform-d = 77.0 ppm; dimethylsulfoxide-d_6 = 39.51 ppm; methanol-d_4 = 49.15 ppm) as an internal standard. Infrared spectra were recorded using an ATR-FTIR instrument. High resolution mass spectra were acquired by flow injection on a qTOF Premiere Mass Spectrometer operating in ES+ ionization with resolution ~15,000. Detailed scheme and methods can be found in the online version of this manuscript

Crystallization

To obtain crystals of DOT1L(1-351)/EPZ004777, crystals of DOT1L(1-351) with SAH were first prepared for displacement soaking. Crystals of DOT1L (1-351) complexed with SAH were obtained at 18 °C using the vapor diffusion method by mixing a protein solution at a concentration of 20 mg/mL (in 20 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, and 1 mM TCEP) with a 5-fold excess of SAH with an equal volume of reservoir solution (1.6M $(NH4)_2SO4$, 0.01M $MgCl_2$, 0.1M NaCaCo, pH 5.5). Solid

EPZ004777 was dissolved in Milli-Q water to obtain a 10 mM stock solution. For soaking, 1 mL of 10 mM EPZ004777 stock solution was mixed with 19 mL of reservoir buffer to prepare 0.5 mM EPZ004777 in soaking buffer. Crystals of DOT1L(1-351)/SAH were transferred into 1.5 mL soaking buffer and incubated for 12 hours during which time EPZ004777 displaced SAH in the crystals.

To obtain crystals of DOT1L (1-420) with FED1, FED2, SGC0946, SGC0947 and EPZ004777, crystals of DOT1L(1-420)/5-iodotubercidin were first prepared for displacement soaking. Purified protein DOT1L (1-420) was concentrated to 16 mg/mL in a buffer containing 20 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, and 1 mM TCEP. To obtain crystals of DOT1L/5-iodotubercidin, DOT1L (1-420) was mixed with 5-iodotubercidin by directly adding a 5 fold molar excess of compound to the protein solution, and the sitting drop vapor diffusion method was used at 18°C in a buffer containing 3.5 M sodium formate, and 100 mM sodium acetate, pH 4.6.

For displacement soaking, all solid compounds were first dissolved in Milli-Q water to obtain 10 mM aqueous stock solutions. Compound solutions for displacement soaking were prepared by diluting 1 µL stock solution of each compound with 19 µL reservoir buffer to make 0.5 mM compound solutions. Crystals of DOT1L (1-420)/5-iodotubercidin were then transferred into 1.5 µL of each compound solution and incubated for 4-24 hours at 18°C. Prior to being flash-frozen in liquid nitrogen, the crystals were soaked in a cryoprotectant consisting of 80% reservoir solution and 20% glycerol.

Data Collection and Indexing

The following beam lines at the Advanced Photon Source were used for data collection: 31ID (LRL-CAT), 23ID (GMCA-CAT, www.gmca.anl.gov), and 19ID (SBC-CAT, 222.sbc.anl.gov).

Data were indexed using the XDS program[28] or within the HKL suite of programs[29] or using Mosflm[30]. Original index and test set reflections were selected from the 2.1 Å DOT1L structure, PDB accession number 3QOW and implemented using the pointless program followed by scala from the CCP4i suite of programs[31]. A paired down model of DOT1L was then used for rigid body refinement directly into each dataset. Missing loops and ligands were then manually built and real space refined within COOT[32], followed by iterative rounds of refinement using refmac 5[33]. Geometric restraints for compounds were created using the program Elbow from the phenix suite of programs[34] or using the

online program PRODRG[35]. All models were refined with good geometric restraints and excellent clash and Molprobity scores and deposited in the PDB.

Surface Plasmon Resonance (SPR)

SPR studies were performed using a Biacore™ T200 instrument (GE Health Sciences Inc.). DOT1L (1-420 approximately 4000RU) was stably immobilized to CM5 chips through amine coupling at pH 7.4 according to the manufacture's protocol (GE Health Sciences Inc.). Compounds were dissolved in 100% DMSO and 2-fold serial dilutions were performed in 100% DMSO. For testing, the serially diluted compounds were diluted 1:20 into HBS-EP buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.05% P-20) giving a final concentration of 5% DMSO. The flow rate was set at 70 mL/min. For each compound, single-cycle kinetic analysis was performed with an on-time of 300 seconds, and an off-time of 1800 seconds. SAM and SAH were used as positive controls. Curve fitting and K_D determinations were performed with the Biacore T200 Evalutation software (GE Health Sciences Inc). Five concentrations of each compound were tested in two-fold serial dilution experiments within the ranges listed. All compounds were fitted using the on and off rates, which were then used to calculate the K_D. The stoichiometry was determined by reference to the binding levels of the 1:1 binding controls, SAM and SAH. The SPR profile is unusual for the most potent compounds due to their extreme binding affinity.

Isothermal titration calorimetry

SAM-free DOT1L(1-420), purified by above-mentioned method, was used in the ITC experiment with a VP-ITC calorimeter (MicroCal, LLC). The enthalpy of binding of DOT1L (15 μM in cell) and EPZ004777 (0.2 mM in syringe) was measured at 25°C in 20 mM Tris-HCl pH 8.0, 200 mM NaCl, 5% DMSO. A clear 1:1 stoichiometry was observed. We did not attempt to calculate the K_D by fitting the binding isotherms since the binding was too tight to be determined directly by ITC, instead, we used SPR to measure the K_D.

DOT1L methyltransferase assay

The reported assay condition was used with minor modification[17]. Nucleosomes purified from chicken blood cells was used as substrate. 60 nM nucleosome solution was added into 40 μL assay buffer (20 mM Tris-HCl, 2 mM DTT, 10 mM $MgCl_2$, 0.01% Triton X-100), which contained 0.25 nM recombinant DOT1L (1-420) with inhibitors of different concentrations (or DMSO as control). A similar 40

μL mixture without DOT1L was prepared as background. The mixture was incubated at room temperature for 30 minutes before 0.75 μM ^{3}H-SAM (PerkinElmer, catalog number NET155001MC) was added to start the reaction. The reaction mixture was incubated at room temperature for two hours and quenched by the addition of 160 μL 10% trichloroacetic acid (TCA). The mixture was transferred into glass fiber filter plates (Millipore, catalog number MSFBN6B) and washed twice with 10% TCA. An ethanol wash (100 mL) was followed by drying. 100 μL Microscint Zero (PerkinElmer, catalog number 6013611) was finally added into each well of the filter plates and centrifuged to flow through filters. Tritium incorporation was detected on a TopCount (PerkinElmer).

Immunoblot of H3K79me2

For western blot analysis in human cell lines, exponentially growing human leukemia cell line MV4;11 or Molm13 cells were plated in 6-well plates at 2 × 10^5 cells/well in a final volume of 2 mL. Cells were incubated in the presence of DMSO or 10 μM of EPZ004777 or other synthesized derivatives, excepting SGC0946 which was incubated at a concentration of 1μM. For the Molm13 experiments, 1 μM of compounds were used for times indicated. Cells (2-4 × 10^6) were harvested at 4 days (MV4:1) and histones were extracted as detailed in Experimental Procedures. Histones (3 μg) were separated on 10% Bis-Tris gels (Invitrogen NP0315BOX), transferred to 0.2 μM nitrocellulose membranes using the IBLOT and probed with the appropriate primary antibodies (see Supplementary Experimental Procedures for a list of primary antibodies used) diluted 1:1000 in 5% milk in TBS-T. Following primary antibody incubation, membranes were probed with ECL donkey anti Rabbit IgG (Thermo Fisher Scientific 45-000-683) secondary antibody and signal was detected using SuperSignal West Pico Chemiluminescent Substrate Fisher Scientific (Thermo Fisher Scientific PI-34077).

Cellular H3K79 methylation assay

A431 cells were plated at 1,000 cells/well in 50μL in 384-well clear bottom plates (Corning 3712) and incubated overnight. Cells were treated with compound with an automated pin transfer instrument (Janus, Perkin Elmer) and incubated for 3-4 days. Following incubation with compound, media was aspirated (EL406, BioTek), cells were fixed in 50 μL formaldehyde solution (3.7% formaldehyde in PBS) and incubated 10 minutes at room temperature. Fixation solution was aspirated and cells were rinsed in 50μL of blocking solution (1% BSA in PBS) twice. Then 50uL permeabilization solution was added (1%

SDS in blocking solution) for 2 minutes, then aspirated. After rinse with blocking solution, cells were incubated in 50uL of blocking solution for 30 minutes at room temperature. Blocking solution was aspirated, and cells were incubated for 1+ hour at room temperature in 10 µL of primary antibody for dimethylated histone 3 K79 (ab1791) at a 1:5000 dilution in blocking solution. Primary antibody solution was aspirated, and cells were washed twice in 50µL of blocking solution. Following the second wash, cells were incubated in 10 µL for 60 minutes at room temperature in secondary antibody (Invitrogen A-21244) and nuclear staining (Invitrogen H3570) solution at 1:1000 dilution in blocking solution. Secondary antibody and nuclear staining solution was aspirated and cells were washed two times in 50µL of blocking solution. Then 50µL of PBS was added to each well. Image acquisition was performed on a high content screening microscope (ImageXpress Micro, Molecular Devices), and image analysis (MetaXpress3.0, Molecular Devices) was performed to obtain average dimethylated lysine 79 and acetylated-tubulin signal per cell based on treatment. Replicate experimental data from incubations with inhibitor were normalized to DMSO controls. Dose response data was generated (Graphpad Prism) by normalization of maximum and minimum dimethylated lysine 79 compared to EPZ004777.

Cell Viability

Human cord blood cells transformed with MLL-AF9 and MLL-ENL were cultured as described elsewhere[22]. The cells were treated with indicated inhibitor concentration for 14 days. Cell viability assay was performed by incubating cells with resazurin solution (0.1mg/ml) for 4h and fluorescent resorufin measurement at 584nM.

Inhibitor selectivity profiling

Selectivity of DOT1L inhibitors was assessed by screening a panel of SAM-dependent methyltransferases (SUV39H2, G9a, EHMT1, SETDB1, PRMT3, SETD7, MLL, SETD8, SUV420H1, SUV420H2, PRC2, DNMT1 and PRMT5) by a radioactivity based assay. In this assay ^{3}H-SAM (Cat.# NET155V250UC, Perkin Elmer) was used as a methyl donor to methylate peptide substrates. Peptide substrates were biotinylated to be captured in each well through their interaction with streptavidin using a streptavidin-coated flash plate (96-well FlashPlate, Cat#: SMP103, Perkin Elmer, http://www.perkinelmer.com/). Amount of the product (methylated peptide) was quantified by tracing the radioactivity (counts per minute measured by a TopCount reader from Perkin Elmer). Assay conditions

were optimized for each protein separately, and all experiments were performed at linear initial velocity. The enzymatic reactions were conducted in triplicate at room temperature for 1 hour in a 20 µL reaction mixture in 20 mM Tris-HCl, pH 8.0, 5 mM DTT, and 0.01% Triton X-100 containing ^{3}H-SAM at concentrations around Km value of each enzyme. The reaction was quenched with equal volumes of 7.5 M guanidine-HCl. 10 µL of the reaction mix containing guanidine-HCl was mixed with 190 µL of 20 mM Tris Buffer, pH 8 and transferred into a flash plate. The plate was incubated for an hour prior to reading using a TopCount (Perkin Elmer, http://www.perkinelmer.com/) to accumulate maximum signal.

Hydrogen-deuterium exchange

Hydrogen exchange experiments were performed essentially as described in Iacob et al.[36]. 70 pmol of DOT1L protein was incubated with different concentrations of inhibitor for a protein:inhibitor ratio of 1:6. Kds were 242nM (for EPZ), 500nM (for FED2) and 750nM (for SAH). At the above mentioned ratios, 96.71% protein was bound to SAH, 97.78% was bound to FED2 and 98.91% was bound to EPZ. All mixtures were incubated for 30 min at room temperature before deuterium labeling. As a control, DOT1L was incubated in 20 mM Tris, 150 mM NaCl, 3mM DTT (pH 8.0) buffer and treated exactly as the inhibitor bound protein. Deuterium exchange was initiated by dilution of each protein with 15-fold 20 mM Tris, 150 mM NaCl, 3mM DTT (pD 8.0), D_2O buffer at room temperature. At each deuterium exchange time point (from 10 s to 4 hours) an aliquot from the exchange reaction was removed and labeling was quenched by adjusting the pH to 2.5 with an equal volume of quench buffer (0.8M GnHCl, 0.8% Formic Acid, H_2O). Quenched samples were immediately frozen on dry ice and stored at -80°C until analysis.

Each frozen sample was thawed rapidly and injected into a custom Waters nanoACQUITY UPLC HDX Manager™ and analyzed as previously described[37]. The protein samples were digested using a Poroszyme immobilized pepsin cartridge (Applied Biosystems) which was accommodated within the UPLC system. The cooling chamber of the UPLC system, which housed all the chromatographic elements, was held at 0.1°C for the entire time of the measurements. The injected peptides were trapped and desalted for 3 min at 100 µL/min and then separated in 6 min by an 8–40% acetonitrile:water gradient at 40 µL/min. The separation column was a 1.0×100.0 mm ACQUITY UPLC C18 BEH (Waters) containing 1.7 µm particles and the back pressure averaged 8800 psi at 0.1°C. The average amount of back-exchange using this experimental setup was 18% to 25%, based on analysis of highly deuterated

peptide standards. Deuterium levels were not corrected for back-exchange and are therefore reported as relative[24]; however, all comparison experiments were done under identical experimental conditions thus negating the need for back exchange correction[24]. The UPLC step was performed with protonated solvents, thereby allowing deuterium to be replaced with hydrogen from side chains and the amino/carboxyl terminus that exchange much faster than amide linkages[38]. All experiments were performed in duplicate. The error of determining the deuterium levels was ±0.20 Da in this experimental setup consistent with previously obtained values[39]. Mass spectra were obtained with a Waters XEVO G2 TOF equipped with standard ESI source (Waters Corp., Milford, MA, USA). The instrument configuration was the following: capillary was 3.2 kV, trap collision energy at 6 V, sampling cone at 35 V, source temperature of 80°C and desolvation temperature of 175°C. Mass spectra were acquired over an m/z range of 100 to 2000. Mass accuracy was ensured by calibration with 500 fmol/μL GFP, and was less than 10 ppm throughout all experiments. The mass spectra were processed with the software DynamX^{TM40}(Waters Corp., Milford, MA, USA) by centroiding an isotopic distribution corresponding to the +2, +3, or +4 charge state of each peptide. Deuteration levels were calculated by subtracting the centroid of the isotopic distribution for peptide ions of undeuterated protein from the centroid of the isotopic distribution for peptide ions from the deuterium labeled sample. The resulting relative deuterium levels were automatically plotted versus the exchange time. Identification of the peptic fragments was accomplished through a combination of exact mass analysis and MS^{E41} using Identity Software (Waters Corp., Milford, MA, USA). MS^{E} was performed by a series of low-high collision energies ramping from 5–30 V, therefore ensuring proper fragmentation of all the peptic peptides eluting from the LC system. DOT1L peptic peptide map was prepared using MSTools[42].

Molecular dynamics simulations

The DOT1L protein structures were isolated from the relevant crystal structures (PDBs: 1NW3 and 3QOW) and prepared using the Protein Preparation Module from Schrödinger/2011 (Schrodinger, NY). All crystallographic waters and counterions were removed. The Ligprep Module was then used to obtain minimized 3D structures of SAM.

The resulting structures for the proteins and small molecules served as the starting point for the molecular dynamics (MD) simulations, which were carried out using the PMEMD version included in the

AMBER11 Molecular Dynamics Package. These MD simulations were performed after careful relaxation of the systems using minimization and equilibration protocols. The ionizable residues were set to their normal ionization states at neutral pH. The protein atoms were surrounded by a periodic box of TIP3P[43] water molecules that extended 10 Å from the protein. Na^+ counterions were placed by LEaP REF to neutralize the system.

The ff03.r1 version of the all-atom AMBER force field was used to model the protein, and the GAFF force field was used for the SAM structure[44]. After geometry optimization at the B3LYP/6-31G* level, atom-centered partial charges were derived using the AMBER antechamber program (RESP methodology)[45]. In the MD simulation protocol, the time step was chosen to be 2 fs and the SHAKE algorithm[46] was used to constrain all bonds involving hydrogen atoms. A non-bonded cutoff of 10.0 Å was used and the non-bonded pair list was updated every 25 time steps. Langevin dynamics was used to control the system temperature (300K) using a collision frequency of 1.0 ps^{-1}. Isotropic position scaling was used to maintain the system pressure at 1 atm. Periodic boundary conditions were applied to simulate a continuous system. To include the contributions of long-range interactions, the Particle-Mesh-Ewald (PME) REF method was used with a grid spacing of ~1 Å combined with a fourth-order B-spline interpolation. PME also enabled computation of the potential and forces in between grid points. The trajectories were analyzed using the PTRAJ module of AMBER. The same module was used to calculate the theoretical B-factors. They were normalized by taking the difference between the raw B-factor and the average B-factor and dividing this by the standard deviation.

Pocket hydrophobicity

Structures of DOT1L in complex with SAM (PDB codes 1NW3 and 3QOW), SAH (PDB code 3QOX), Bromo-deaza-SAH (PDB code 3SX0), a methylated SAH analog (PDB code 3SR4), and compound **4** were loaded in ICM version 3.7-2b (Molsoft, San Diego, CA), and the adenosine end common to all compounds was deleted from the structures. The molecular surface of the receptor surrounding the non-adenosine portion of the compound was generated with Molsoft's ICM (San Diego, CA), and all DOT1L atoms within 1.5 Å of this surface that were lining the binding pocket were selected. Oxygen, nitrogen atoms, and hydrogen atoms attached to them were considered polar. Other atoms were considered hydrophobic. The hydrophobicity was quantified as the ratio of hydrophobic over polar atoms.

Accession codes

Protein Data Bank: The coordinates and structure factors for human DOT1L complex structures presented in this work have been deposited in the Protein Data Bank (PDB) with the following accession numbers: single EPZ004777 molecule: 4ER3; 5-iodotubercidin: 3UWP; SGC0946: 4ER6; compound 2**:** 4ER0; compound 3: 4ER7; compound 4: 4EQZ; two EPZ004777 molecules: 4ER5

Acknowledgements

We wish to thank Dr. Wolfram Tempel for reviewing crystal structures and Dr. Abdellah Allali-Hassani and Guillermo Senisterra for their expert advice on biochemical assays. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. The computations were performed on Kraken at the National Institute for Computational Sciences (http://www.nics.tennessee.edu/). X-ray diffraction data was collected at the LRL Collaborative Access Team (LRL-CAT) beam line facilities at Sector 31 of the Advanced Photon Source operated by Eli Lilly & Company, GM/CA CAT which is funded in whole or in part with Federal funds from the National Cancer Institute (Y1-CO-1020) and the National Institute of General Medical Sciences (Y1-GM-1104), and the Advanced Photon Source supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. J.E.B., J.Q., J.J.M. are supported by the Leukemia & Lymphoma Society (SCOR), the National Institutes of Health (R01 CA152314 and K08CA128972), the American Society of Hematology, and the William Lawrence & Blanche Hughes Foundation. E.J.C. is supported by a Steamboat Foundation Award. A.J.F is supported by the National Science Foundation and an Ashford Fellowship. We would like to acknowledge Northeastern University and the Barnett Institute of Chemical and Biological Analysis for financial support, partial funding from NIH R01-GM086507 and a research agreement with the Waters Corporation (J.R.E.). The Structural Genomics Consortium is a registered charity (number 1097737) that receives funds from Canadian Institutes of Health Research, Eli Lilly Canada, Genome Canada, GlaxoSmithKline, the Ontario Ministry of Economic Development and Innovation, the Novartis Research Foundation, Pfizer, Abbott, Takeda and the Wellcome Trust. C.H.A. holds a Canada Research Chair in Structural Genomics.

Contributions

W.Y. and T.H. purified the protein; W.Y. and F.L. performed methyltransferase assays; E.J.C., A.S., A.F., J.J.M. and J.Q. synthesized compounds; E.J.C. and A.J.F. developed DOT1L biochemical and cellular assays; W.Y. crystallized DOT1L complexes; A.K.W. and W.T. solved the crystal structures; D.B.-L., J.Y., T.S., E.W. and J.E.D. conducted cell-based experiments; C.G. and A.A. performed Caco-2

experiments, R.E.I. and J.R.E. performed hydrogen–deuterium exchange data, J.P. and G.E. performed MD simulations, R.M. and R.A. generated the SPR data; and M.V., P.J.B., C.H.A., J.E.B. and M.S. designed experiments and analyzed results with the respective co-authors above. All authors participated in preparing the manuscript.

Chapter 3

Structure-Guided DOT1L Probe Optimization by Label-Free Ligand Displacement

Additional Contributors: Joanna Yi*, Alexander J. Federation*, Jun Qi*, Sirano Dhe-Paganon, Michael Hadler, Xiang Xu ,Roodolph St. Pierre, Anthony C. Varca, Lei Wu, Jason J. Marineau, William B. Smith, Amanda Souza, Emma J. Chory, Scott A. Armstrong, James E. Bradner

* Denotes equal contribution

Corresponding Supplementary Material can be found in Appendix B.

This chapter originally appeared in *ACS Chemical Biology*, Online (2015).

Introduction

Mixed-Lineage Leukemia (*MLL*) gene rearrangements occur in 5-10% of all acute leukemia patients and in greater than 70% of infants with acute lymphoblastic leukemia (ALL). The presence of the rearrangement portends a poor prognosis, despite aggressive therapy with significant associated morbidity[1]. Cooperation between specific chromatin-modifying complexes and *MLL*-rearranged gene products defines disease pathogenesis and has prompted efforts to target modulators of chromatin structure and function in this cancer[2-3].

MLL is a member of the Trithorax family of proteins and functions as a histone lysine methyltransferase (KMTase)[1]. During development, MLL catalyzes trimethylation of lysine 4 on histone 3 from the methyl donor S-adenosylmethionine (SAM) at homeobox genes, promoting their expression. In MLL-rearranged leukemia, however, the SET domain responsible for KMTase activity is uniformly lost with translocation and replaced by one of more than 70 known fusion partners[3]. Many of these fusion partners recruit DOT1L, which is the only known methyltransferase responsible for the mono-, di-, and trimethylation of lysine 79 of histone 3 (H3K79). H3K79 methylation is associated with most actively transcribed genes and marks regions of elongating RNA Pol II typically within the first intron of gene bodies[4]. The recruitment of DOT1L by MLL fusion partners to developmental MLL-target genes results in aberrant hypermethylation of H3K79 at these loci, contributing to leukemogenesis by inappropriately sustained gene expression, namely at the HOXA locus[5-8].

The therapeutic significance of DOT1L in established MLL-rearranged leukemia has been validated by genetic and chemical genetic approaches. Conditional inactivation of Dot1L in MLL-AF9 (and AF6, AF10) leukemia models results in diminished H3K79 methylation and prolonged survival[6,9-12]. Recently, SAM-competitive small-molecule inhibitors of DOT1L have been developed, first reported in 2011[13], and further characterized biochemically and structurally by our group in collaboration with Professor Cheryl Arrowsmith[14]. Structurally, these SAM mimetics featured high potency and selectivity for DOT1L. However, the cellular activity of these compounds is rather low in potency relative to the extraordinary sub-nanomolar binding potency in homogeneous assays *in vitro.* Notably, the anti-leukemic effect of DOT1L inhibition requires 10-14 days of continuous dosing at high (1-3 μM) concentrations in cell culture models using current inhibitors[14-19]. In animal studies, this translates to a modest benefit in

survival while requiring high doses through continuous osmotic subcutaneous infusion[13,15]. Further optimization of DOT1L inhibitors is therefore needed.

To date, development of structurally divergent DOT1L inhibitors has been slow in the broader epigenetics community, perhaps relating to the challenges in biochemistry and cell biology platforms that underlie ligand discovery and optimization. Thus far, biochemical assays of DOT1L use radioligands and often require specialized synthetic or highly purified histone particles as substrates. Additionally the ubiquitinylation of nucleosomes strongly influences DOT1L activity and poses difficulties to ligand discovery[20]. The delayed cellular effects of DOT1L inhibition challenge the miniaturization of cell-based measures of compound potency. Simple dose-ranging comparisons have proven time-consuming and low-throughput. We therefore identified an opportunity to create a facile discovery platform enabling the characterization of existing DOT1L inhibitors, and the preparation of new compounds with improved properties. Herein, we report the development of tagged DOT1L ligands used in robust and miniaturized biochemical assays, as well as a high-throughput, high-content assay system that reports on phamacodynamic H3K79 methylation abundance in short incubation windows. Together, these three orthogonal assays have defined a platform capable of discovering and optimizing novel DOT1L inhibitors.

Results

Toward the development of DOT1L chemical probes, we chose a SAM-competitive inhibitor from our laboratory (**FED1**) as a suitable starting point to develop assay ligands for DOT1L (Figure 3.1a). **FED1** is a near chemical derivative of **EPZ004777** that features a more efficient and high-yielding synthesis[14]. Additionally, **FED1** has a modestly reduced binding potency for DOT1L that was postulated to improve utility in competition binding assay development across a broad range of inhibitors. Given the extended residence times of DOT1L inhibitors (**EPZ004777** K_{off} 9.29E-4 s^{-1} and **FED1** K_{off} 2.20E-3 s^{-1})[14], a less potent inhibitor such as **FED1** may provide the opportunity to discover weaker initial assay positives in high-throughput screening. The crystal structure of **FED1** shows a binding mode similar to **EPZ004777**, with the tert-butyl phenyl urea motif further extending the binding pocket compared to SAM (Supplementary Figure B.1a, PDB: 4ER0)[14]. While most of the molecule is deeply obscured in the binding pocket and inaccessible to solvent, the more open position of the nucleotide base suggested a tolerance

for further chemical substitution. We postulated that modification on the N6 position of the **FED1** adenine would not interfere with the activity of the molecule, allowing the installation of features with functional utility (e.g. retrievable chemical linkers and fluorophores).

We synthesized two probes (Supplementary Scheme B.1 and Scheme B.2), **1** with polyethyleneglycol (PEG) linked biotin, and **2** with a thiourea-coupled fluorescein (FITC) (Figure 3.1b). The binding affinities of these two modified inhibitors were confirmed by isothermal titration calorimetry and are comparable to the parent compound (Figure 3.1c and Supplementary Figure B.1b). The apparent potency of **1** was also similar to the parent compound by differential scanning fluorimetry (Supplementary Figure B.1d). The crystal structure of **1** with DOT1L was then obtained. Ligand **1** bound to the SAM pocket as expected, with the structured features of the linker protruding out toward solvent. The lack of atomic density for the remaining PEG and biotin features likely reflects unrestricted mobility in solvent (Figure 3.1d). The ligand-interaction diagram further confirmed that DOT1L binds to **1** and **FED1** in a manner dictated by common determinants of molecular recognition (Figure 3.1e, Supplementary Figure B.1c). The amide bond of the linker unit also formed a hydrogen bond with a structured water molecule, which may contribute to preservation of the ligand binding activity. Cell-permeability was confirmed by immunoblot for the H3K79 dimethyl (H3K79me2) histone mark, which was efficiently depleted by both ligands (Figure 3.1f). As expected, both compounds inhibited cell proliferation comparably to **EPZ004777** after 10-14 days (Figure 3.1g).

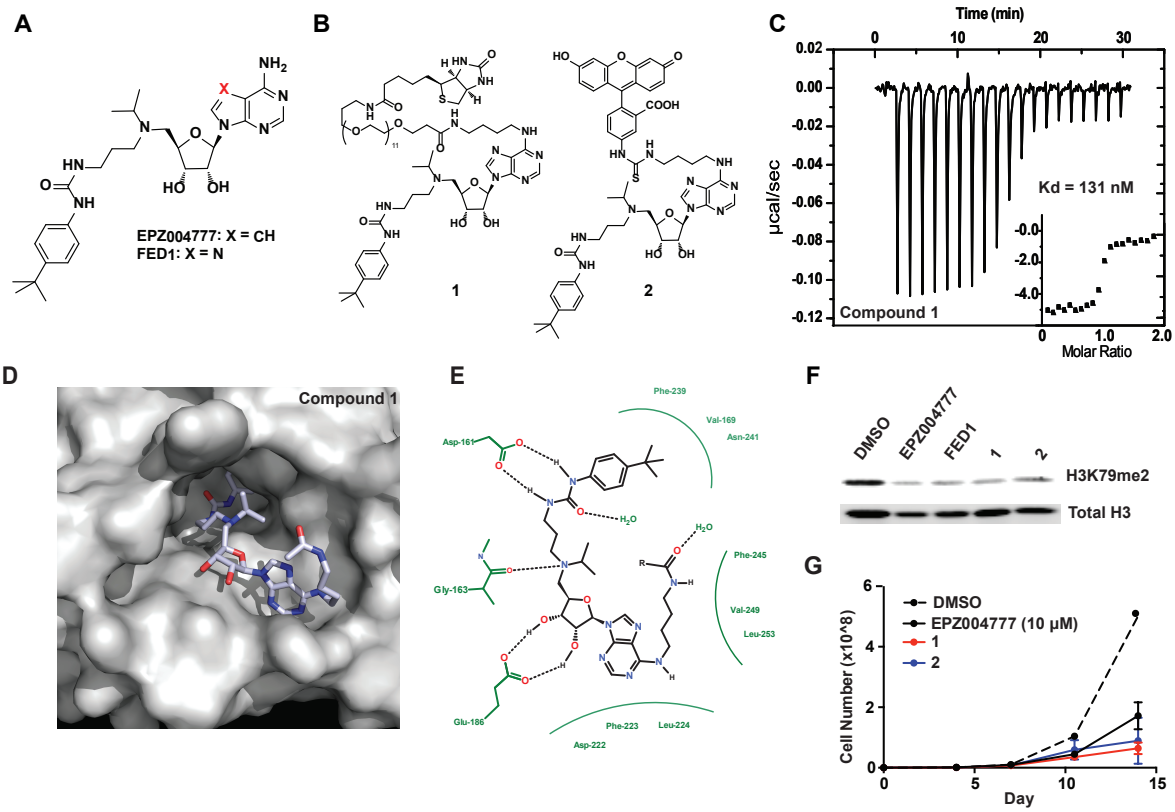


Figure 3.1: Design and characterization of chemical probes of DOT1L. a) Structures of **EPZ004777**[13] and **FED1**.[14] b) Structures of biotinylated (**1**) and FITC-labelled (**2**) **FED1** attached via a linker to the N6 position of the base. c) Isothermal calorimetry analysis of **1** demonstrating strong 1:1 binding with DOT1L. d) Binding of **1** to DOT1L demonstrates linker exposure and a similar binding mode as **FED1**. e) Detailed ligand-interaction diagram of **1** demonstrating new hydrogen bond formation. f) Inhibition of H3K79me2 by indicated DOT1L inhibitors (10 uM) in MV4;11 cells treated for 4 days. g) DOT1L probe treatment results in inhibition of MV4;11 cell growth over time.

With these validated ligands in hand, we next developed orthogonal biochemical assays capable of detecting small molecule binding to purified, recombinant human DOT1L protein. The first assay utilizes **1** and employs a nanomaterial proximity assay (AlphaScreen, Perkin Elmer; Figure 3.2a-c). The biotin on **1** recruits a streptavidin-coated donor bead while the **FED1** portion of the molecule recruits a nickel-coated receptor bead via binding to recombinant HIS_6-DOT1L methyltransferase domain. Illumination of the donor bead releases singlet oxygen, which diffuses to activate *in situ* synthesis of a chemiluminescent lanthanide within the acceptor bead only when the two are in close proximity, here dependent on the DOT1L-ligand interaction. Displacement of **1** from DOT1L disrupts the proximity of the two beads and diminishes chemiluminescence. Finally, we have miniaturized the assay to microtiter plate format (384-well) and improved robustness compatible with high-throughput screening (Z' = 0.78). Using a set of resynthesized DOT1L chemical tools, we confirmed faithful utility in comparative ligand potency determination (IC_{50} for **EPZ004777** 5.3nM, **FED1** 22.4 nM, **SAH** 1299 nM).

Next, a fluorescence polarization (FP) assay was developed to monitor binding of inhibitors to DOT1L using the fluorescent probe **2** (Figure 3.2d-f). After excitation with plane-polarized light, binding of **2** to DOT1L increases anisotropy of the bound state relative to free **2**. Therefore, displacement of non-covalently bound **2** from DOT1L by a competitive ligand leads to a decrease in detectable signal. This FP assay has proven amenable to high-throughput screening (Z' = 0.91), and accurately discriminates compounds in our reference set by relative potencies. In our experience to date, the FP assay is more suitable to HTS than the bead-based proximity assay owing to false-positives that are assay specific (nickel chelation, biotin mimetics). Together, these two assays are highly complementary in screening and lead optimization.

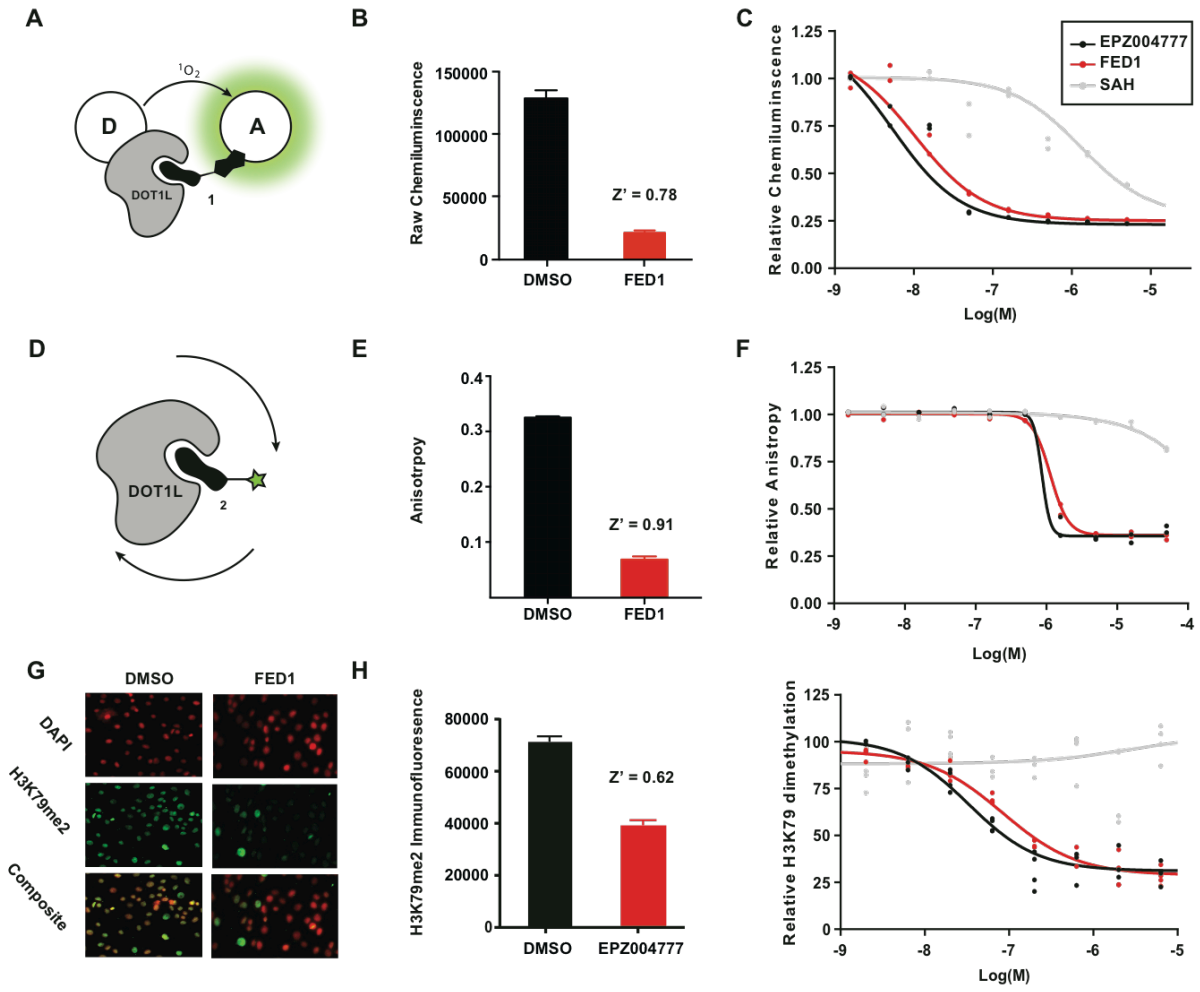


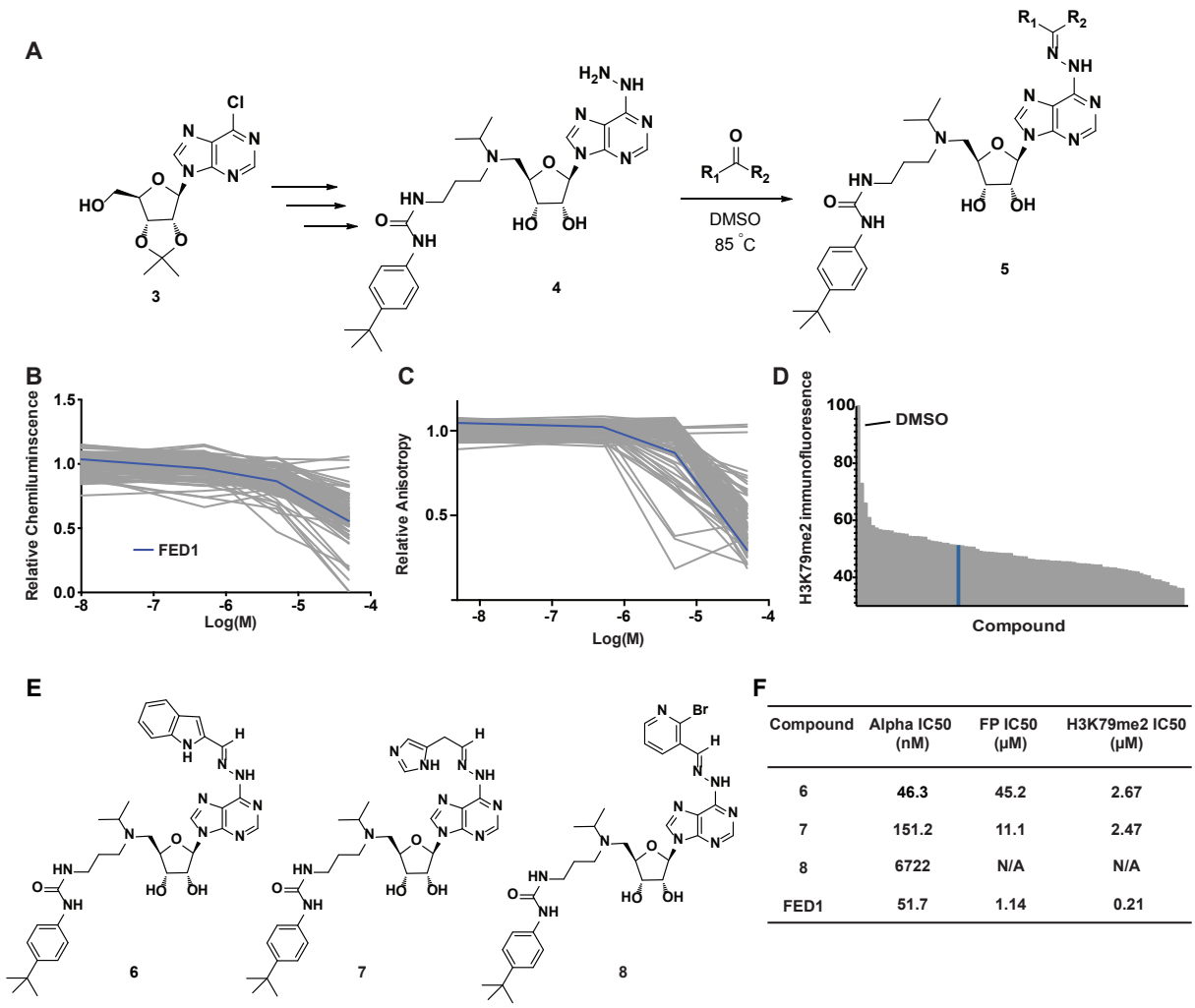
Figure 3.2: Development of non-radiometric biochemical and cellular assays for DOT1L. a-c) AlphaScreen proximity bead-based assay demonstrating adaptability to high-throughput screening (Z' calculated by 1- ((3*σ_{FED1}+σ_{DMSO})/Absolute value(μ_{FED1}-μ_{DMSO})), and expected comparable potency differentiation of known inhibitors. d-f) Fluorescence polarization assay demonstrating significant assay robustness (Z' calculated with above formula), and separation of weak DOT1L inhibitors (**SAH**) from more potent compounds (**FED1** and **EPZ004777**). g-i) High-content imaging assay evaluating H3K79me2 abundance by immunofluorescence in A431 cells after 4 days of indicated DOT1L inhibitors, with diminished H3K79me2 compared to DMSO. Assay is robust (Z' calculated as above) and reports cellular EC_{50}. Biochemical assays were performed in duplicate, and high-content assays were performed as four replicates.

To support iterative development of DOT1L inhibitors using a cellular assay capable of reporting on methyltransferase function, we developed a high-content imaging assay to measure global nuclear H3K79me2 with short-term compound incubation (3-4 days; Figure 3.2g-i). The A431 epidermoid carcinoma cell line was selected based on high basal H3K79me2 and strong adherence, which tolerated the harsh immunofixation conditions required to detect staining of this mark on the histone body. Cells were treated for four days with compound in 384-well plates, and then assessed for H3K79me2 using primary and secondary (fluorescent-conjugated) antibodies. Using our reference compound set, the dynamic range of this assay format proved capable of supporting dose-ranging studies and correctly ranked the cellular potency of all inhibitors: in comparative format **EPZ004777** was more potent than **FED1**, and **SAH** was ineffective at reducing H3K79 methylation levels in cell culture. This cellular assay allows both the comparisons between compounds to measure in-cell efficacy and the evaluation of multiple compounds simultaneously in dose-response. It is also amenable to further evaluate any screening hits obtained from the AlphaScreen and FP assays, and could also be employed as a primary assay for cell-based, high-throughput screening (Z' = 0.62).

To test the DOT1L discovery platform, we generated a small focused library of compounds exploring the tolerability of substitutions of the adenosine base using a highly parallel chemistry we co-developed to explore surface-recognition features of organic ligands (hydrazine cap-scan technology)[21-22]. Informed by the DOT1L-**FED1** and DOT1L-**1** crystal structures[18], we developed a focused library of N6-substituted ligands. A hydrazine functional group was introduced to the adenosine ring via a nine-step synthesis starting from known compound **3** (Supplementary Scheme B.3 and B.4. The hydrazine **4** then condensed with 90 divergent aldehydes/ketones in 96-well format to generate the hydrazone library **5** (Figure 3.3a) without the need for further purification. The facile assembly of the library allowed rapid exploration of structure-activity relationships (SAR) at N6, and the exercise provided firm validation of the assay cascade. All analogues were evaluated in both biochemical assays in four-point dose response, followed by the H3K79me2 high content screening at a single dose. Most members of the hydrazine library showed high levels of activity in all 3 assays, suggesting that this site is highly permissive for a wide range of chemical functionalities (Figure 3.3b-d). Comparison of library activity in the two biochemical assays demonstrated generally good agreement between both assays (Supplementary

Figure B.2). A group of outliers that were highly active in only the bead-based assay were triaged as likely assay-interference agents. Three of the hydrazones that showed the most consistent activity between all assays were resynthesized for re-test with a ten-point, dose-response curve; these three compounds contained nitrogenous heterocyclic rings, including an indole (**6**), an imidazole (**7**), and a pyridine (**8**) (Figure 3.3e). While these three compounds are more potent than the parent hydrazine **4,** they had modestly weaker biochemical activity than **FED1** (Figure 3.3f). Although the modifications explored in the focused library did not provide large gains in potency, the study validated all assays in a high-throughput format.

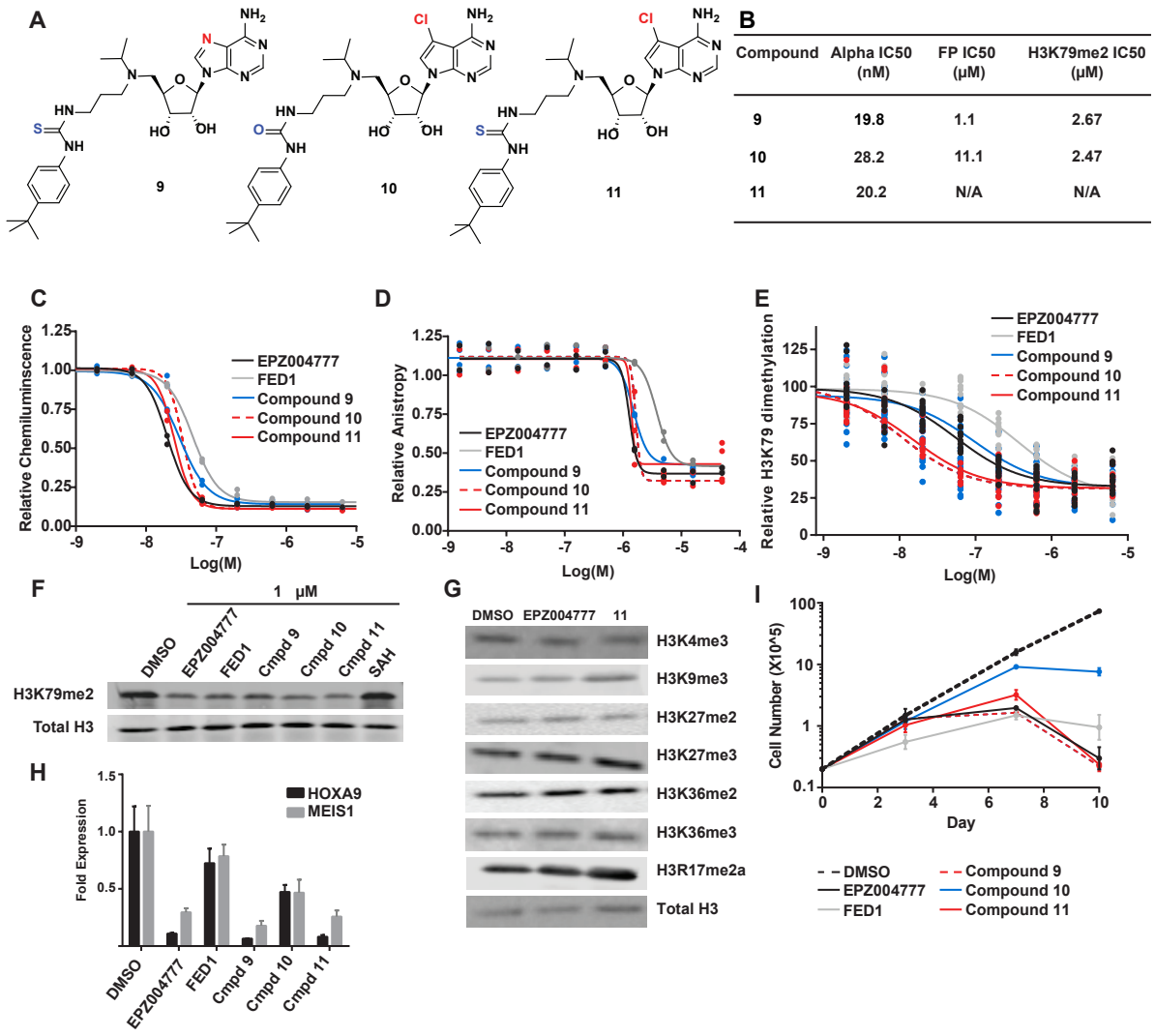
We then explored further features of the SAM-like core structure. In addition to the N6 substitutions, we modified the urea tail along with another location on the adenosine base. The urea motif has also been further explored with generating a small urea library (Supplementary Scheme B.5), and we discovered introduction of the thiourea group to **FED1** produced a 10-fold increase in cellular activity (**9**) without significantly changing biochemical activity. On the adenosine base, we focused on the C7 position of the ring, and discovered that introduction of a chlorine group on the C7 position of deazaadenosine motif could generate a desirable interaction with the nearby hydrophobic pocket (**10**), leading to an increase in potency in all assays. By combining these two beneficial optimizations, we generated **11** with introduction of both C7-chlorine and thiourea (Figure 3.4a). **11** was more potent as we expected in both the biochemical assays, and both **10** and **11** were more potent than **EPZ004777** in the H3K79me2 high-content assay (Figure 3.4b-e).



F

Compound	Alpha IC50 (nM)	FP IC50 (μM)	H3K79me2 IC50 (μM)
6	46.3	45.2	2.67
7	151.2	11.1	2.47
8	6722	N/A	N/A
FED1	51.7	1.14	0.21

Figure 3.3: Purine substitutions of adenosyl DOT1L inhibitors. a) Synthetic illustrating generation of focused library of hydrazine inhibitors, modified off the N6 position. b, c) Biochemical screening results of the library at 4 doses, with d) cellular H3K79me2 screen at 20 uM compared to **FED1** (indicated in blue). e) Structures of resynthesized assay positives. f) Profiling table of validated assay positives compared to **FED1**.



Compound	Alpha IC50 (nM)	FP IC50 (μM)	H3K79me2 IC50 (μM)
9	19.8	1.1	2.67
10	28.2	11.1	2.47
11	20.2	N/A	N/A

Figure 3.4: Development of potent chlorinated DOT1L inhibitors. a) Structures of **9**, **10**, and **11**. b-e) Evaluation of SAM-like derivatives by biochemical assays demonstrating similar potency, and high content assay demonstrating increased potency of chlorinated inhibitors (**10,11**) and improvement of **FED1** potency with substitution of a thiourea tail. Biochemical assays were run in duplicate and high-content assays in quadruplicate. (**9**). f) Immunoblot for H3K79me2 demonstrating improved DOT1L inhibition by **11** in MOLM-13 cells treated for 4 days (all at 1 uM). g) **11** (10 uM) demonstrates specificity for H3K79me2 inhibition by immunoblot for histone methylation marks in MOLM-13 cells. h) HoxA9 and Meis1 mRNA in MV4;11 cells decreases with DOT1L inhibition (10 uM) in proportion to cellular IC_{50} after 7 days of treatment (RT-PCR). i) Inhibition of MV4;11 cell growth over time with DOT1L inhibitor treatment (10 uM) demonstrates correlation with HCS potency.

With the improved potency in the cellular assay, we then confirmed the relative potencies of these compounds by immunoblotting for H3K79me2 in MLL cells (MOLM-13). This was followed by evaluation of the selectivity of the representative **11** to inhibit only the DOT1L methyltransferase by immunoblotting for a number of histone methylation marks mediated by other KMTases and an arginine methyltransferase. The only mark affected by these compounds is H3K79me2 (Figure 3.4f-g, Supplementary Figure B.2). We then further ensured the on-target effect of these compounds by assessing gene-expression changes in MV4;11 cells. Decreased expression of both MLL-target genes *HOXA9* and *MEIS1* was observed after 7 days of incubation. The potency in gene expression correlated to effects on H3K79me2 reported by high-content screening, further validating that the 4-day H3K79me2 measurement accurately predicts on-target biological activity previously observed after 7-10 days of treatment (Figure 3.4h). As expected, these measurements also correlated with an anti-proliferative effect in treated MV4;11 cells (Figure 3.4i). Therefore, utilizing our novel assay cascade and structural information, we developed inhibitors of DOT1L with enhanced cellular activity and maintained selectivity compared to previously reported compounds.

Discussion

Our approach to affinity ligand design for assay development was based on a structural understanding of the binding mode between small molecule and target. Since the addition of the handle on the small molecule does not impact its DOT1L potency, the resultant probes **1** and **2** reported here can be used as chemical tools for assay development and further mechanistic studies of the DOT1L complex and its function in MLL[15]. The hydrazine library demonstrated the accommodation of DOT1L to large substituents off the base but potency was not maintained, perhaps from impurities in the original screen. However, this site appears to be permissible for future medicinal chemistry efforts towards improving pharmacokinetics or compound stability. Further exploration of the base and urea tail moiety, as accurately characterized by our assay cascade, led to the identification of more potent compounds than **EPZ004777** with improved cellular activity.

Together, these chemical biology tools for the study of DOT1L provide a nimble platform for discovery chemistry. The label-free biochemical assays and rapid cellular assay will be useful for

discovering both allosteric and direct SAM-competitive DOT1L inhibitors, although substrate-competitive inhibitors may be silent in these biochemical assays. The high content assay, however, should be agnostic to the mode of inhibition. It also has the potential to detect inhibitors of other proteins that modulate DOT1L activity or the rate of H3K79me2 removal. These tagged and potent inhibitors are openly available for use to probe DOT1L biology. We hope this design principle will be adapted to inhibitor discovery for other critical methyltransferases implicated in disease, including EHZ2 and MMSET.

Materials and Methods

DOT1L AlphaScreen Binding Assay

All reagents were diluted in 50mM HEPES, 150mM NaCl, 0.5% BSA (w/v), 0.05% Tween20 (w/v), pH 8.0 with 1mM DTT added. The final concentrations of His6-DOT1L was 80nM and **1** was 40nM. Addition of 10uL of 2x this solution to the plates (AlphaScreen plates, Perkin Elmer #6005359) was performed with a liquid handler. 100 nL of compounds was added by pin transfer using a Janus Workststation (PerkinElmer, USA). After a brief centrifugation, plates were incubated at room temperature for 30 minutes. A 2x solution of beads was made such the final concentrations of both the acceptor and donor beads were 25 μg mL^{-1}. Ten μL of this solution was added to the plate, and after centrifugation and 20 minute incubation, plates were read on the Envision 2104 plate reader (PerkinElmer, USA). Dose response data normalized to DMSO controls.

DOT1L Fluorescence Polarization Assay

All reagents were diluted in PBS with 1mM DTT freshly added. Five μL of DOT1L solution (final concentration 1μM) was added to 384-well plates (Thermo Scientific 262260) with a Biotek EL406 liquid handler. 100 nL of compounds from stock plates was added by pin transfer using a Janus Workststation. After a brief centrifugation, plates were incubated at room temperature for 30 minutes. Five μL of a 2x solution of **2** (final concentration 10nM) was added, and the plate briefly centrifuged. After 30 minute incubation at room temperature, plates were read on the Envision 2104 plate reader. Flatfield and polarization calculations were performed by manufacturer's protocol, and anisotropy was calculated based on the formula 2*P/3-P in which P = polarization. Results were normalized to DMSO controls.

High content imaging assay

A431 cells were plated at 1,000 cells/well in 50μL in 384-well clear bottom plates (Corning 3712) and incubated for 1 hour at room temperature. Compounds were added using a Janus Workstation and incubated for 4 days. After this, cells were fixed in 3.7% formaldehyde in PBS at room temperature x10 minutes. After two rinses with blocking solution (1% BSA in PBS), 1% SDS in PBS was added for 2 minutes. After 1 wash, cells were incubated with blocking solution at room temperature x30 minutes. Cells were then incubated for either 1 hour at room temperature or overnight at 4°C in 10 μL of primary antibody for H3K79me2 (ab3594) at a 1:500 dilution in blocking solution. After rinsing with blocking solution, cells were then incubated for 1 hour at room temperature in 10 μL of secondary antibody (Invitrogen A-21244) and nuclear staining (Invitrogen H3570) solution at 1:1000 dilution in blocking solution. Cells were washed twice, after which 50μL of PBS was added to each well. Images were acquired on a high content screening microscope (ImageXpress Micro, Molecular Devices), and image analysis (MetaXpress3.0, Molecular Devices) was performed to obtain average H3K79me2 signal per cell. Dose response data (normalized to DMSO) was generated (Graphpad Prism) by normalization of maximum and minimum H3K79me2.

Acknowledgments

We thank R. Paranal for his assistance in developing the high content assay; we thank C. Ott for helpful discussions, and D. Buckley for thoughtful review of the manuscript. This work was funded by the US National Institutes of Health grant R01 CA176745, the William Lawrence & Blanche Hughes Foundation, the Leukemia & Lymphoma Society SCOR grant, and the American Society of Hematology-Scholar Award.

Contributions

JSY, AJF, JQ, JEB, and SAA designed the experiments and analyzed the data. JSY, JQ and AJF developed and optimized assays. SDP, RSP, AJF, AS generated protein and XX solved crystal structures. JSY, MH, WBS, AJF performed the cellular studies. JQ, AJF, JSY, JJM, EJC, and ACV

designed and synthesized all the compounds. SAA provided guidance and advice. JSY, AJF, JQ, and JEB wrote the manuscript and all authors reviewed the manuscript.

Chapter 4

Chemical Inhibition of BET Bromodomains as a Strategy to Target Super Enhancers

The remaining chapters build on research enabled by a small molecules probe that was reported prior to the beginning of this work, namely the BET bromodomain inhibitor JQ1. The BET family of transcriptional cofactors contains bromodomains responsible for binding acetylated chromatin, and this function is blocked by JQ1. JQ1 was first tested in a preclinical model of NUT midline carcinoma (NMC), an aggressive and rare cancer characterized by a rearrangement of the BET protein BRD4 with the nuclear protein of the testes (NUT) that conserves the tandem bromodomains of BRD4 (Figure 4.1a)[1]. The molecule acts by directly binding the bromodomain pocket, preventing BRD4 chromatin localization and cofactor activity, leading to a disruption of transcription. JQ1 is highly effective in models of NMC (Figure 4.1), but since this disease is so rare, subsequent efforts were undertaken to find additional settings without a direct genetic lesion of BET proteins where JQ1 might still provide a therapeutic benefit.

JQ1 efficacy in a non-BET rearranged cancer was showcased in multiple myeloma, an incurable malignancy of antibody-producing plasma cells[2]. The molecule was effective at halting proliferation of myeloma cells in culture, surprisingly through the suppression of a MYC transcriptional signature. Further investigation showed that this effect was due to downregulation of MYC itself, owing to a depletion in *MYC* transcript levels upon treatment with JQ1.

MYC is central to the pathogenesis of many cancer types. However, as a transcription factor, finding direct inhibitors of the MYC protein is notoriously difficult. This strategy to indirectly target MYC activity through inhibition of *MYC* expression was quite promising, and other cancer models were then tested to look for this effect.

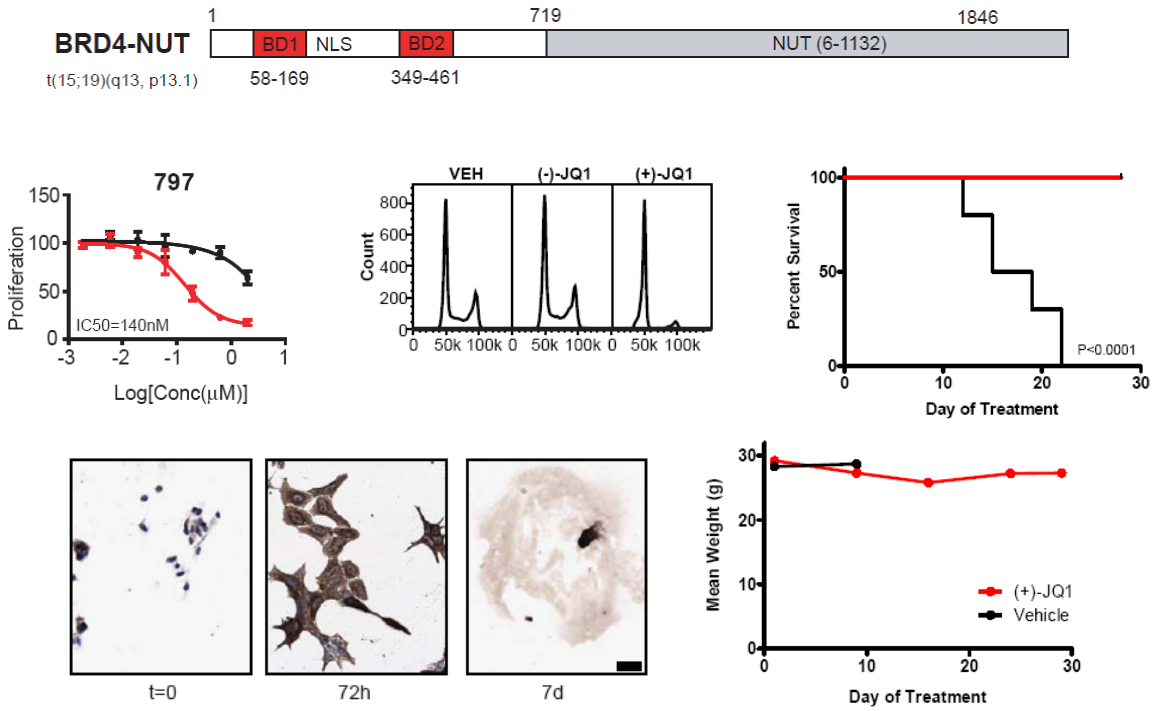


Figure 4.1: The activity of JQ1 in NUT-midline carcinoma. NUT midline carcinoma is characterized by an in-frame translocation of the N-terminus of BRD4 with the NUT protein. 797 patient-derived cultured NMC cells are sensitive to JQ1 treatment *in vitro*. Drug treatment results in a G1 cell cycle arrest, a phenotypic differentiation effect and prolonged survival in a mouse xenograft model of NMC. Mice treated with JQ1 maintain a constant weight during the course of treatment.

Further investigation found multiple models of acute lymphoblastic leukemia that are sensitive to bromodomain inhibition, again leading to the downregulation of the MYC oncogene[3-4]. Pediatric neuroblastoma is also sensitive, with a different MYC isoform (MYCN) driving these cancers and being suppressed by JQ1 treatment[5]. Enabled by the compound's high tolerability in mice, non-cancer indications for BET inhibition were investigated as well, with the compound showing activity in models of heart failure[6] and atherosclerosis[7]. These disease models are not solely MYC driven, and a broad but disease-specific transcriptional response was seen in these models. Notably, the activities of master transcription factors NFKB, NFAT and GATA, which are important for inflammatory pathogenesis, were blunted by JQ1 treatment. A reversible contraceptive effect was also seen in mice due to the inhibition of BRDT[8], a transcriptional response characterized by loss of KLF17 and MSY2 transcription.

The cell-type specific effects of JQ1 remained at odds with observations suggesting that BETs were part of the general transcriptional machinery and found genome-wide at active cis-regulatory promoter and enhancer elements (Figure 4.2a,b). Looking at transcriptional responses to the molecule, it was unclear why disrupting the binding of a constitutive cofactor would exert effects on a relatively small number (100-1,000) of genes. To investigate this further, we mapped the chromatin landscape of cultured multiple myeloma (MM) cells in response to treatment with JQ1. As reported in the initial studies of JQ1, MM cells display a rapid and selective loss of *MYC* oncogene transcription upon treatment. Interestingly, in MM a reciprocal 8;14 chromosomal translocation is places *MYC* under the control of the *IgH* 3' enhancer. At this enhancer, ChIP-seq showed a massively high occupancy of BRD4 that was rapidly lost upon JQ1 treatment. Based on this observation, the lab developed algorithms to quantitatively measure the amount of BRD4 at cis-elements genome-wide and their sensitivity of to BRD4 loss with JQ1 treatment. This analysis revealed that the promoters in the cell have similar levels of BRD4 binding. Most enhancers follow this pattern as well, however, a small subset of enhancers (about 300 in MM) are asymmetrically bound by BRD4 (Figure 4.2 c,d)[9]. These 3% of enhancers contain almost 50% of the BRD4 bound on the genome and were termed "super enhancers". On average, super enhancers spanned 20 kilobases of DNA and contained an order of magnitude higher occupancy of BRD4 compared to typical enhancers. Super-enhancers associated with and drove the expression of key drivers of the MM cell state including MYC, IRF4, XBP1, and BCL-xL. Inhibition of BRD4 led to disproportionate loss of

chromatin and transcriptional regulators from super-enhancers resulting in potent and selective loss of transcription from super-enhancer associated genes.

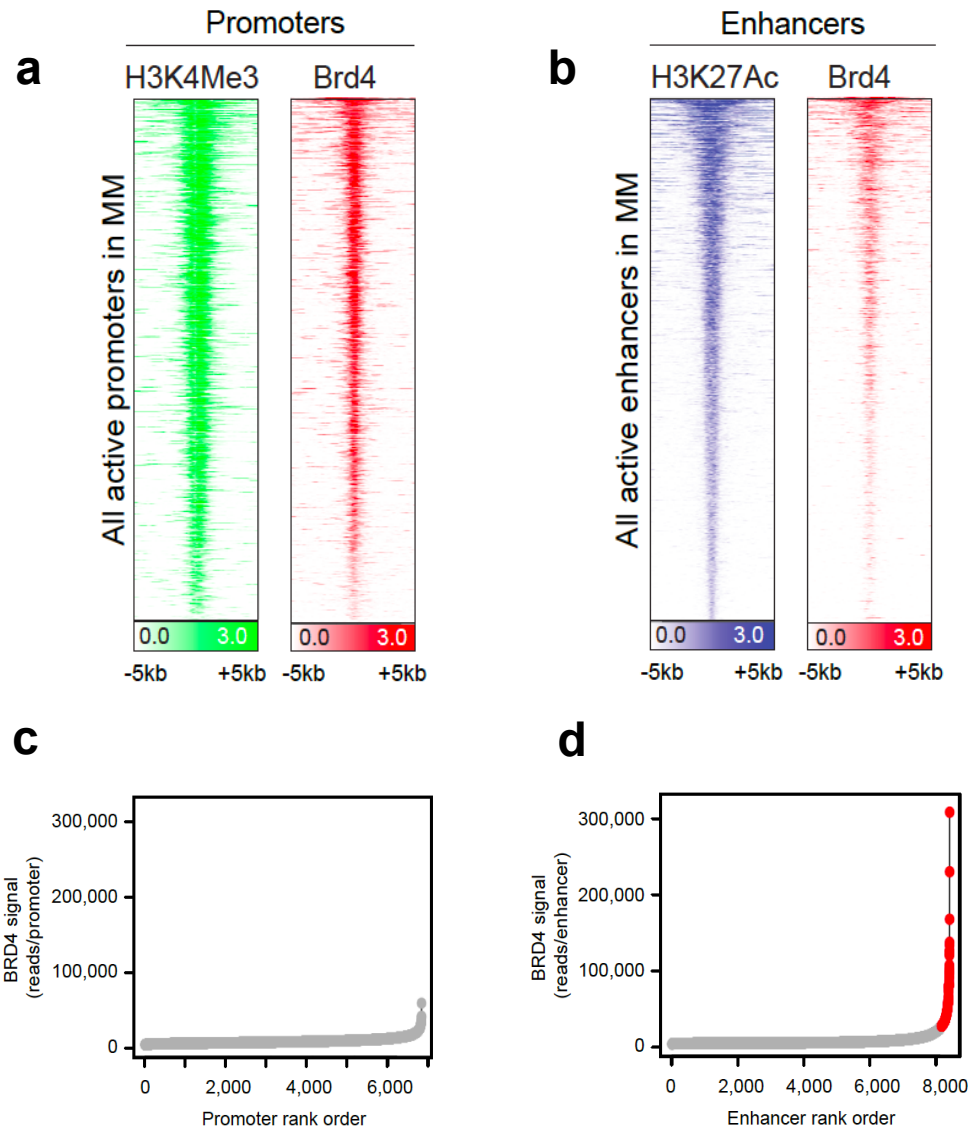


Figure 4.2: BRD4 binds genome wide and defines super enhancers. Genome-wide ChIP-seq profiles for BRD4 at all promoters (a) or enhancers (b) defined by H3K4me3 and non-promoter H3K27ac, respectively. Ranking of BRD4 ChIP-seq read density at all promoters (c) and enhancers (d), with super enhancers highlighted with red points.

Follow-up studies mapping super-enhancers in large numbers of healthy and tumor cells have identified several additional key features they play in transcriptional regulation. Most super enhancers were observed in only one or a small number of cell types[10]. In all profiled cell types, super enhancers are found to regulate genes critical for the specialized biological function of that cell (Figure 4.3). Additionally, super-enhancers are found regulating master transcription factors (TFs) that drive the expression of cell-type specific genes, including positive feedback reinforcement of the master TFs themselves. Lastly super enhancers are highly enriched for genetic risk variants for disease, but only in the relevant cell type for that disease.

The following chapters aim to utilize these properties of super enhancers as a window to investigate biological processes in development and oncogenesis. Chapter 5 investigates the role that super enhancers play in genetic translocation events that underlie the development of B cell malignancies. Chapter 6 describes the use of super enhancers to model master TF regulatory networks in diverse cell types using a single epigenomic measurement, and Chapter 7 applies these techniques in pediatric medulloblastoma. Strikingly, the super enhancer network of a poorly characterized subgroup of medulloblastoma reveals a small sub-network highly specific for a transient cell population in the developing cerebellum, pointing to a putative cell of origin for this malignancy.

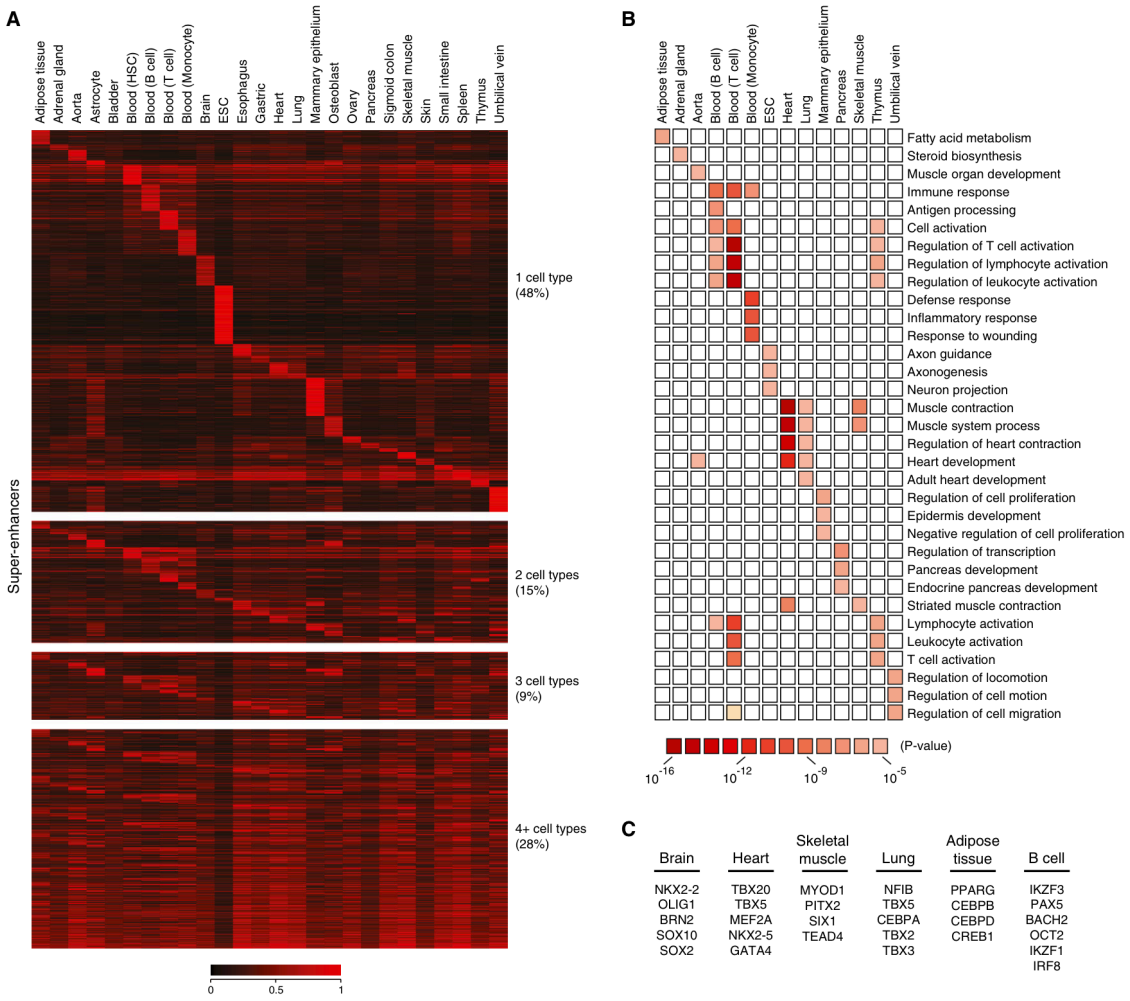


Figure 4.3: Properties of super enhancers. (a) The number of cell types that contain super enhancer loci across a survey of over 80 primary and cell line samples. (b) GO analysis of super enhancer associated genes in various cell types. (c) Master regulator transcription factors regulated by super enhancers in selected tissue types.

Chapter 5

Convergent Sense/Antisense Transcription At Intragenic Super-Enhancers Targets AID-initiated Genomic Instability

Contributors: Fei-Long Meng*, Zhou Du*, Alexander J. Federation*, Jiazhi Hu, Qiao Wang, Kyong-Rim Kieffer-Kwon, Robin M. Meyers, Corina Amor, Caitlyn R. Wasserman, Donna Neuberg, Rafael Casellas, Michel C. Nussenzweig, James E. Bradner, X. Shirley Liu, and Frederick W. Alt

* Denotes equal contribution

Corresponding Supplementary Material can be found in Appendix C.

This chapter originally appeared in *Cell*, Vol 159 (2014).

Introduction

The B cell antigen receptor ("BCR") is comprised of immunoglobulin (Ig) heavy (IgH) and light (IgL) chains. In response to antigen activation, B lymphocytes in peripheral lymphoid organs undergo somatic hypermutation (SHM) and IgH class switch recombination (CSR), and ultimately secrete their BCR as an antibody. SHM diversifies antibody repertoires by introducing high-frequency mutations into IgH and IgL variable region exons[1]. SHM occurs in germinal centers (GCs) of peripheral lymphoid tissues, where B cells are selected for mutations that generate BCRs with increased antigen affinity[2]. IgH CSR involves generation and joining of *IgH* locus DSBs in switch (S) regions that precede various sets of *IgH* C_H exons ("C_Hs") to replace the initially expressed C_H with a downstream C_H, thereby, producing antibodies with different effector functions[3]. Both SHM and CSR are initiated by activation induced cytidine deaminase (AID)[4-5], which deaminates cytosine to uridine on single- stranded DNA (ssDNA)[1]. Mismatches created by these deaminated cytidines are processed into mutations or DSBs during SHM and CSR, respectively, through a process that employs activities of normal base excision or mismatch repair pathways[1].

Within target sequences, AID cytidine deamination focuses on 3-4bp "SHM" motifs that are greatly enriched in S regions and in portions of variable region exons that encode antigen- binding sites[1,6]. Transcription is required for AID targeting during SHM and CSR[7-9]. In this regard, SHM of V(D)J exons in GC B cells begins about 150bp downstream of the transcription start site (TSS) and tapers off 1-2 kb downstream[10]. Likewise, each C_H has a promoter upstream of the S region that upon induction by external signals generates transcription through the S region and, thereby, targets AID[3,11]. Mouse and human S regions also have a highly G-rich non- template strand that upon transcription forms stable R-loops that provide ssDNA to augment AID targeting[12]. RNA polymerase II (Pol II) has been implicated in directing AID to Ig gene SHM and CSR targets through a transcription coupled mechanism[13] that involves AID association with the Spt5 transcription cofactor in the context of Pol II stalling[14]. R loops or other aspects of repetitive S region structure may augment AID access by promoting Pol II stalling[15,16]. Once AID is recruited to Ig targets, replication protein A (RPA) and the RNA exosome RNA degradation complex contribute to generating requisite ssDNA substrates[3,17-19].

Beyond Ig gene targets, AID initiates recurrent mutations or DSBs in a small subset of non-Ig

genes collectively termed AID "off-target" genes[20-23]. Off-target AID activity promotes translocations between Ig loci and cellular oncogenes, as well as SHMs of oncogenes associated with B cell lymphomas[7,24,25]. Identification of AID off-targets has been facilitated by genome-wide translocation cloning methods[21,22] and other large-scale approaches[23,26]. In general, AID activity occurs at much lower levels on off-targets than on Ig genes[10,21,22,26], likely due to specialized AID-targeting features of the latter. AID off-target sequences are not enriched in AID hotspot motifs relative to the genome in general[27]. Consistent with a role for transcription, AID off-target activity is most abundant on transcribed genes downstream of their TSSs[20,21-23]. However, transcription *per se* is not sufficient to target AID, as most transcribed genes are not AID off-targets[7,10]. Next- generation sequencing studies revealed unexpected transcriptional features, including divergent sense and antisense transcription at TSSs[28,29] and frequent promoter proximal Pol II pausing[29]. But, divergent transcription ("DivT") from TSSs occurs in over half of all genes and does not map to sites of AID off-target activity[21] (see below). Likewise, transcriptional pausing alone cannot explain AID off-targeting, since more than 30% of transcribed genes have paused Pol II[29]. Thus, mechanisms that lead to recurrent AID targeting may arise from previously unrecognized transcriptional or epigenetic determinants[7].

Global Run-on Sequencing (GRO-Seq) detects nascent transcripts generated by transcriptionally engaged RNA polymerases[30]. GRO-Seq revealed that a large fraction of intergenic regions are transcribed, with a subset emanating from transcriptional enhancers[31]. Enhancers are sequence-defined, cis-regulatory elements that influence target gene expression irrespective of orientation[32]. Both enhancers within genes (intragenic) and intergenic enhancers may regulate target promoters locally and over long distances[32]. Active enhancer sequences are commonly transcribed by RNA Pol II generating so-called “enhancer RNAs (eRNAs)”; and transcription arising from enhancers is often divergent, with both sense and antisense transcription emanating from enhancer elements[31,33]. Various regulatory functions have been ascribed to eRNAs and other non-coding RNAs[34]; however, much of non-coding RNA biology is not fully understood.

Enhancers are comprised of discrete or clustered transcription factor binding sequences. A common feature of active enhancers is chromatin that is characteristically modified by acetylation (e.g. histone 3 lysine 27; H3K27Ac) and methylation (e.g. histone 3 lysine 4 mono- methylation; H3K4me1)[36].

An unexpected asymmetry in the regional allocation of enhancer factors and enrichment for enhancer marks within and unique to each mammalian cell type studies revealed a subset of so-called super-enhancers (SEs) that feature clusters of highly hyperacetylated and actively transcribed enhancers that, on average, are 10- fold longer than other "typical" enhancers[37,38]. Like locus control regions, SEs regulate genes involved in specialized cellular function[39] and are found within or adjacent to lineage-specifying transcription factor genes[37,40]. In cancer, SEs frequently enforce oncogene expression[38] and, thereby, contribute to tumor pathogenesis. For example, translocations that juxtapose *c-myc* to the IgH 3' regulatory region, a known SE[41,42], promote B cell lymphoma by activating *c-Myc* over long distances[43]. In this context, selectively blocking SE activity with bromodomain and extra-terminal domain (BET) inhibitors is a promising cancer therapeutic strategy[38,41,42].

Here, we report that the majority of detectable AID off-target activity in a variety of mouse and human lymphoid or non-lymphoid cell types occurs within focal regions of overlapping sense/anti-sense transcription within intragenic SEs.

Results

To elucidate transcriptional features that influence AID targeting genome-wide, we applied GRO-Seq to splenic naïve, GC and CSR-activated B cells at much greater depth than done previously. Naïve splenic B cells were purified and then cultured in the presence of αCD40 plus interleukin-4 (IL4) for 60 hours to stimulate AID induction and CSR to IgG1 and IgE (Supplementary Figure C.1A). Splenic GC B cells were purified from sheep red blood cell immunized mice (Supplementary Figure C.1A) and confirmed to be greater than 90% pure (Supplementary Figure C.1B-D). Three independent GRO-Seq biological replicates were performed for each cell type and gave highly reproducible results (Supplementary Figure C.1E). Transcription profiles of over 20,000 genes revealed distinct (but over-lapping) gene expression patterns for each cell types that were further classified by gene ontology terms (Supplementary Figure C.1G). As expected[21,30], GRO-Seq revealed divergent sense and anti-sense transcription at TSSs of over 50% of the genes in each of the three cell types (Figure 5.1). In depth examination of sense transcription profiles of several “signature” genes illustrates the specificity of purified cell populations. For example, *Aicda* sense transcription reflects AID protein expression in the

three cell types, with high levels in GC B cells and activated B cells; but none detectable in naïve B cells (Figure 1). In contrast, several GC B cell-specific genes, including *SLIP-GC*[44] and *Bcl6*[45], had high sense transcription through their gene bodies in GC B cells, but not in naïve or CSR-activated B cells (Figure 1). Finally, *Bcl2*, which is expressed in CSR-activated but not in GC B cells[46], showed corresponding sense transcription patterns (Figure 1).

While *IgH* C_H exons were appropriately transcribed in the three cell populations (Supplementary Figure C.1H), transcription within core S regions could not be mapped due to their abundant repetitive sequence[14]. All analyzed mice had a clonal knock-in $V_H(D)J_H$ exon47 (V_HB1-8), which showed active transcription at its upstream regions in all three cell types. However, detailed analyses of transcription through the body of the V_HB1-8 allele was not possible (Supplementary Figure C.1H); because it uses a member of the V_HJ558 family, which contains many highly related, unexpressed upstream copies[47].

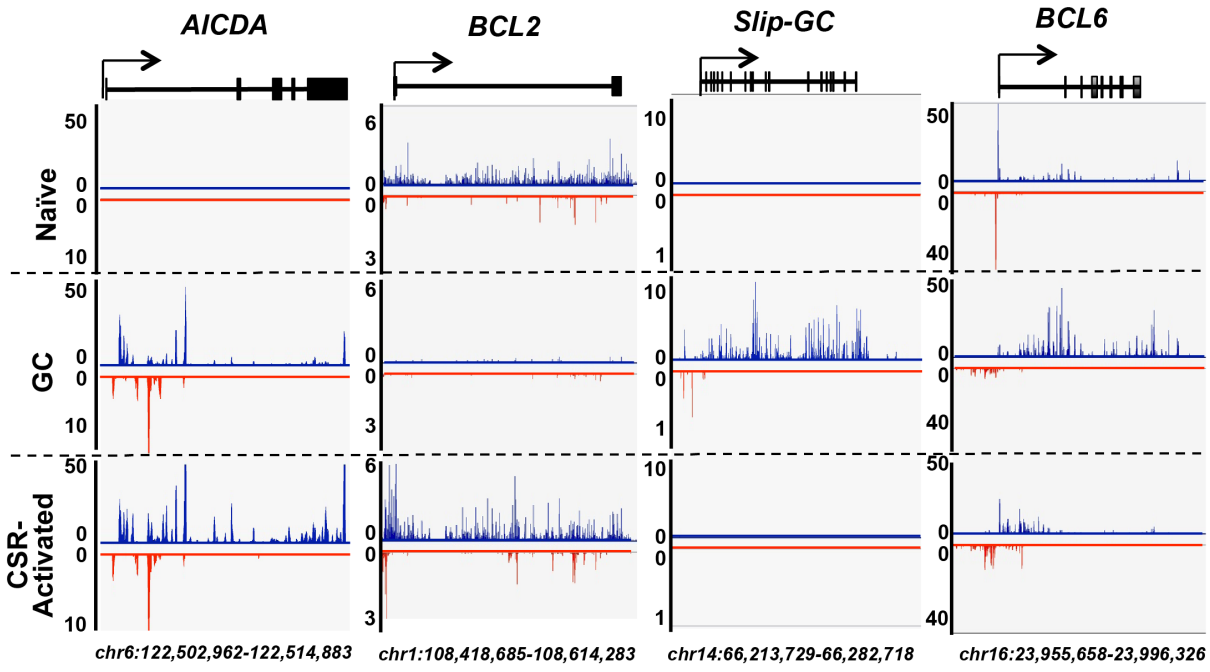


Figure 5.1: GRO-Seq Profiles of Naïve, Germinal Center, and CSR-activated B Cells. GRO-Seq profiles of four representative genes are shown for different B cell types. The Y-axis indicates GRO-Seq counts normalized to number of reads per million. Gene sense and antisense transcription are displayed in blue and red, respectively. Gene exons are illustrated by squares along gene bodies at the top of each panel. Arrows indicate TSSs and direction of sense transcription. Genome coordinates (mm9/NCBI37) are labeled at the bottom. All the profiles were generated from merged data of three independent experiments, which individually showed similar patterns.

We developed High-Throughput Genome-wide Translocation Sequencing ("HTGTS") to map, at the nucleotide level, translocation junctions between bait I-SceI nuclease generated DSBs in *c-myc* and other endogenous DSBs[21]. Identification of DSB hotspots from a fixed chromosomal site is facilitated by ability of recurrent DSBs to dominate genome-wide translocation landscapes due to cellular heterogeneity in three-dimensional genome organization[48]. Beyond expected Ig locus targets, our prior HTGTS studies revealed15 non-Ig genes that are recurrent targets of AID-initiated DSBs and translocations. To increase the depth of HTGTS AID off-target data and allow better comparison with deeper GRO-Seq transcription profiles, we further employed a modified, more sensitive HTGTS approach[49], coupled with Ataxia Telangiectasia Mutated (ATM)-deficient CSR-activated B cells[50]. This combined approach identified highly clustered AID-dependent off-target DSB sites within 36 additional genes (Supplementary Figure C2.A). Overall, we now have identified 51 highly focal AID off-target DSB/translocation sites in αCD40 plus IL4-stimulated B cells. Nearly 90% of the new off-target set was validated via in WT B cells by HTGTS and/or by an independent method (Qian et al, unpublished). As previously found for our more limited set of AID off-target sites[21] many of our new AID off targets occurred within genes that have divergently transcribed TSSs; but their focal sites of HTGTS junctions again were downstream of and distinct from divergently transcribed TSSs. Thus, we were compelled to search for other factors that promote focal AID off-targeting. As we found no enrichment for known AID targeting motifs in these regions, we focused our search on potentially novel transcriptional and/or epigenetic features and, as described below, consistently identified both.

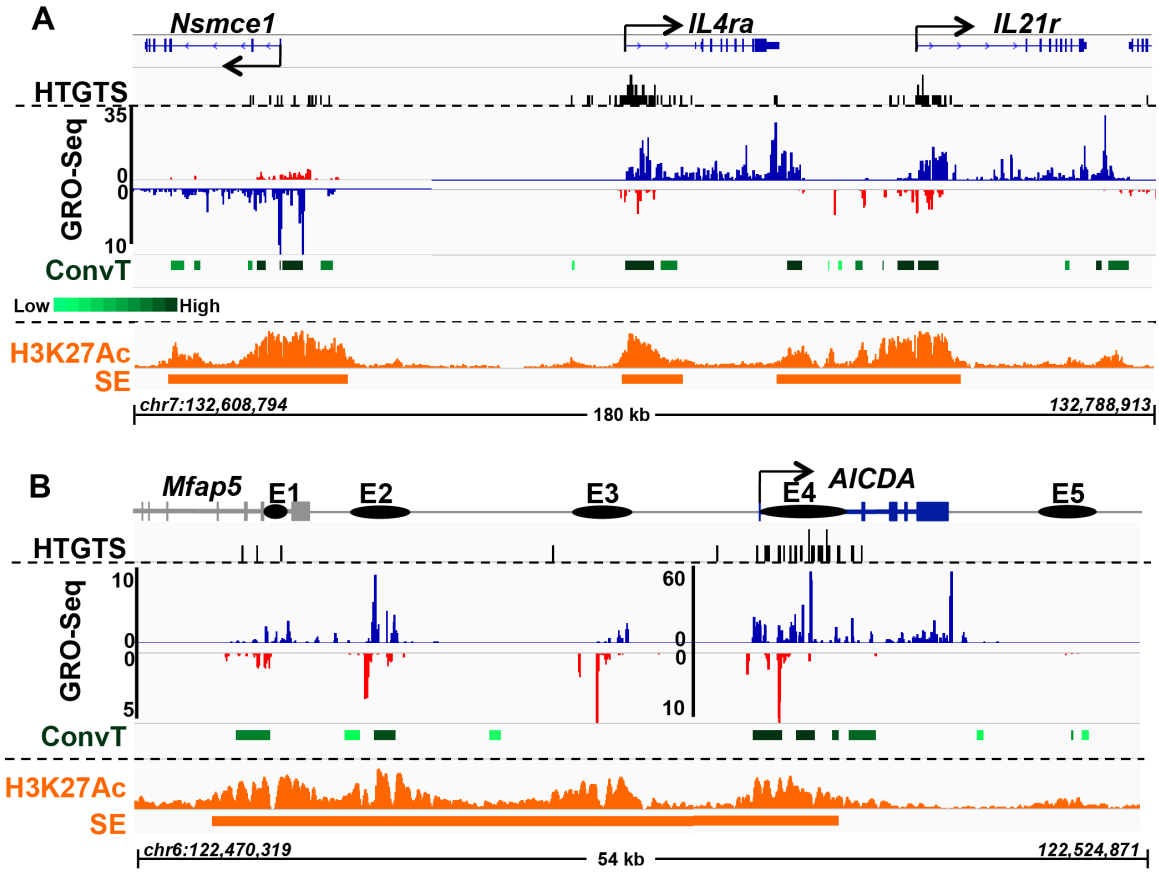


Figure 5.2: AID Off-Target Translocations Cluster Within Regions of ConT and SEs . (A) HTGTS, GRO-Seq, ConT and H3K27Ac profiles in the vicinity of *Nsmce1, IL4ra,* and *IL21r* genes. Top Panel: HTGTS junctions are indicated by black bars. Middle Panel (GRO-Seq): GRO-Seq-determined sense and antisense transcription is displayed in blue and red, respectively. ConvT regions are shown as green bars at the bottom with the darkest shades corresponding to highest levels of ConT as calculated by the geometric means of sense and antisense transcription reads. A scale bar is shown below the ConvT label. Bottom Panel (H3K27Ac and SE): The H3K27Ac ChIP-Seq profile is shown in orange and identified SEs depicted below with orange bars. (B) Profile of *AICDA* gene. Known *AICDA* enhancers are represented as E1-5 with solid circles. To represent lower level transcription of certain enhancers, a smaller scale is used for E1-3. Genome coordinates (mm9/NCBI37) are at the bottom of each panel. Other details are the same as for panel A.

With our present, substantially deeper, GRO-Seq data sets, we further analyzed potential relationships between sense/antisense transcription and AID off-target sites in αCD40 plus IL4 activated B cells. Initially, we visually inspected three linked AID off-target sites, including sites in the previously characterized *IL4r and IL21r* genes and a newly identified site in *Nsmce1*. In each of these linked genes, HTGTS translocation junctions were tightly clustered downstream of the TSS (Figure 5.2A). Moreover, in each, translocation clusters fell within sites that exhibited enriched, overlapping sense and antisense transcription to which we heretofore apply the term "convergent transcription" (ConvT) (Figure 5.2 and 5.3). We also found a robust AID off-target site within the AID gene ("*Aicda*") itself (Figure 5.2B). *Aicda* is associated with five enhancers that lie upstream, within, or downstream of the gene body[18,51] (Figure 5.2B). Four of these enhancers showed both sense and antisense transcription, likely at least in part in the context of generating eRNAs. Notably, the major focal cluster of AID off-target sites in and around *Aicda* fell within a ConvT region associated with enhancer 4 downstream of the TSS (Figure 5.2B).

Visual inspection of AID off-target sites in additional genes revealed similar coincidence of regions of robust sense/antisense ("S/AS") ConvT downstream of the TSS with focal clusters of AID-dependent off-target translocations (Supplementary Figure C.2.), leading us to examine this potentially striking association genome-wide. While metagene profiles of GRO-Seq data from αCD40 plus IL4 activated B cells confirmed expected DivT at many TSSs[28], they did not reveal similarly abundant convergent transcription (Supplementary Figure C.1F). Thus, at least at robust levels, convergent transcription likely occurs in a much smaller fraction of genes. For further analyses, we developed a computational pipeline to specifically identify S/AS ConvT regions genome-wide using deep GRO-Seq data sets (Figure 5.3A). Strikingly, among the 51 AID off-target genes, 48 (94%) had their highly clustered AID off-target translocations within regions associated with S/AS convergent transcription (Figure 5.3B). We randomly sampled convergent transcription of regions of genes, in the top three transcription-level deciles, that were similar in size to those of AID off-target regions and found a much lower association with convergent transcription than for AID off-target regions (Supplementary Figure C.3A). This finding shows that AID off-targets are highly enriched at ConvT sites. Finally, concurrency between S/AS convergent transcription and AID off-target translocations was much higher in αCD40 plus IL4 activated B cells (94%) than in naïve (49%) or GC (63%) B cells, consistent the notion that not all AID off-targets

would be shared among three cell types with over-lapping, but clearly distinct, transcription profiles (Figure 5.3B).

To further examine the relationship between ConvT and AID targeting, we calculated the geometric mean of GRO-Seq sense and antisense transcription reads in regions of interest to quantify degree of convergent transcription , and divided the values into deciles displayed by different shades of green bars below the GRO-Seq profiles. For most AID off-targets, HTGTS junctions clustered in regions with the most abundant ConvT. Furthermore, ConvT associated with AID off-targets was substantially greater than that at other genomic loci (Supplementary Figure C.3C). In addition, within AID off-target ConvT regions, the highest density of translocations occurred at sites with the most robust ConvT (Figure 5.3C). We further evaluated this relationship by determining how variations in sequencing depth influenced identification of ConvT. Even with our current very deep sequencing depth (>306 million mappable reads), we did not reach saturation of the total length of ConvT regions (Supplemental Figure C.3D), consistent with (at least low-level) pervasive transcription of the genome[52]. In contrast, we reached saturation of the concurrency of AID off-targets with ConvT regions at about 40% of our current GRO-Seq depth (120 million mappable reads; Supplementary Figure C.3D), confirming that most AID off-target DSB/translocation regions detectable by HTGTS in αCD40 plus IL4 stimulated B cells are associated with relatively strong convergent transcription (Figure 5.3C).

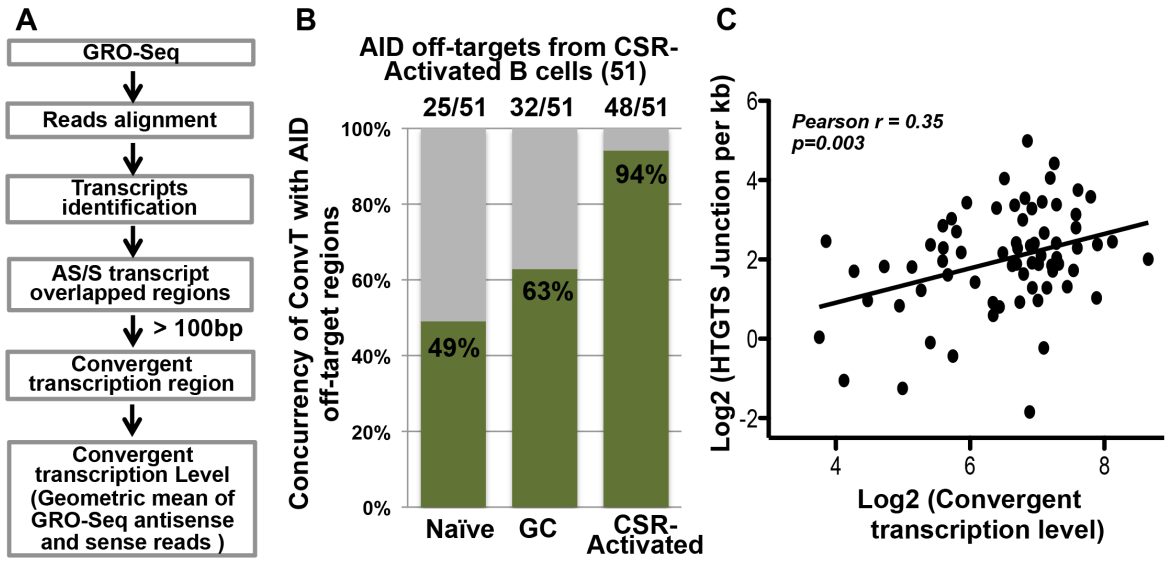


Figure 5.3: AID Off-targets Correlate with ConT in CSR-activated B Cells. (A) Pipeline for identification of ConvT regions. Raw GRO-Seq reads were aligned to the genome and transcripts were identified *de novo*. A "ConvT" region was defined as sense and antisense transcription overlaps that were longer than 100bp. See supplementary methods for details. (B) The percentage of the 51 AID off-target regions identified in CSR-activated B cells that were associated with ConvT regions in the three listed cell populations is the indicated by the green bars. (D) Numbers of translocation junctions per kilobase (kb) (Y axis) plotted against ConvT levels (X axis) of all individual AID off-target regions except *Pvt1* (see Supplementary Methods). Pearson's correlation coefficient and two-tailed *p* value are indicated.

ConvT of overlapping genes was first described in bacteriophage lambda[53], and has been associated with transcriptional gene silencing[54] and RNA Pol II collision[55]. Considering that intragenic antisense transcription associated with AID-off targets sequences may arise from enhancer elements, we explored whether intragenic SEs were enriched for AID off-targets compared to typical enhancers. Enhancer regions were identified by triplicate chromatin immunoprecipitation with massively parallel genome sequencing (ChIP-Seq) using an antibody to the active enhancer histone mark H3K27Ac in chromatin purified from αCD40 plus IL4 stimulated B cells. SEs were called based on outlier analysis for regions of asymmetric, high enrichment for H3K27Ac, as previously described[37]. We found the *Aicda* locus to be largely encompassed within a SE in CSR-activated B cells with robust H3K27Ac signals over E1, E2, E3 and E4 (Figure 5.2B), the active enhancers in CSR-activated B cells[51]. Notably, E4 also corresponds in position to a cluster of HTGTS junctions and robust ConvT (Figure 5.2B). Likewise the *Nsmce1*, *IL4ra*, *Il21r*, and many other AID off-target genes were each associated with SEs and again the peak of HTGTS junctions and regions of robust ConvT occurred within regions of robust H3K27Ac SE signals (Figure 5.2A).

We performed an unbiased association analysis between the 51 AID off-targets identified by HTGTS and the non-Ig 448 SEs that we identified in αCD40 plus IL4 activated B cells. These studies revealed that 50 of the 51 AID off-target genes in these cells are associated with SEs and that the discrete translocation clusters were within SEs (Figure 5.4A). Notably, the single AID off-target region not within SE (under the current cutoff for SE identification) was in a typical enhancer. In addition, 47 (92%) of the AID off-target translocation clusters were within regions of SEs that overlap with annotated gene bodies (Figure 5.4A). The other 3 HTGTS off-target translocation clusters occurred within transcribed regions of SEs that have not yet been assigned to a target gene. As a comparison, random samplings of transcribed genomic regions corresponding in size to those of AID off-targets yielded at most three (6%) that overlapped with SEs. Independent analysis of the relationship between HTGTS hotspots and H3K27Ac ChIP-Seq using an orthogonal computational method identified 41 AID off-targets within SE domains (Supplementary Figure C.4), including additional novel off-targets that correlated with robust ConvT. Finally, within a given AID off-target region, translocation junction frequency highly correlated with H3K27Ac abundance (Supplementary Figure C.4B). In this regard, SEs associated with AID off-target

sequences were more enriched for H3K27Ac, compared to other SEs (Figure 5.4B). Thus, the relative activity of SEs, estimated by regional histone acetylation, correlates with the frequency of AID off-targets within them.

The majority (30 of 51) of the AID off-target genes had a SE that overlapped with the region just downstream of the TSS that was enriched in AID off-targets, as represented by the *CD83* gene (Figure 5.4C). In addition, a number (12 of 51) of the AID targets were relatively small genes, such as *Pim1,* that were located within large SEs and, correspondingly, off-target translocations tended to span the gene body (Figure 5.4D). Several AID off- target genes (3 of 51) were large genes, such as *Pvt1*, the well-known translocation target downstream of *c-myc*, in which translocations clustered in within SEs that occurred inside the gene body (Figure 5.4E). Finally, the remainder (6 of 51) fell into a heterogeneous set in which AID off-target translocations clustered into convergently transcribed SE domains that, for various reasons were not yet assignable to a specific gene (e.g. *Gpr183*).

Nearly all AID off-target clusters identified by HTGTS in αCD40 plus IL4 activated B cells are associated with SEs; yet, only a subset of SEs are AID off-targets. Motivated by the putative contribution of S/AS eRNA transcription to translocation frequency, we compared regions of AID off-target genes where SEs overlap with the gene body (intragenic SEs) to regions where SEs lie outside the gene body (intergenic SEs) and to regions of gene bodies that do not overlap with SEs ("non-overlapping gene region"), for translocation density (translocations per 1kb; Figure 5.5A) and for ConvT levels (geometric means; Figure 5.5A). We observed that translocation junction density and ConvT levels in AID off-target regions are highly enriched among intragenic SEs compared to both intergenic SEs and non-overlapping gene regions (Figure 5.5A; upper). Despite this enrichment, only about 10% of all intragenic SEs in the CSR-activated B cells are AID off-targets (Figure 5.4A) and other SE-gene overlap regions exist that are not enriched in AID off-target activity (Figure 5.5B; upper). Comparison of ConvT levels in each of the three regions outlined above (Figure 5.5A, B; lower panels) revealed that intragenic SEs featuring high levels of ConvT were more frequently AID off-target regions than intragenic SEs lacking high-level S/AS transcription (Figure 5.5A, B; lower panels).

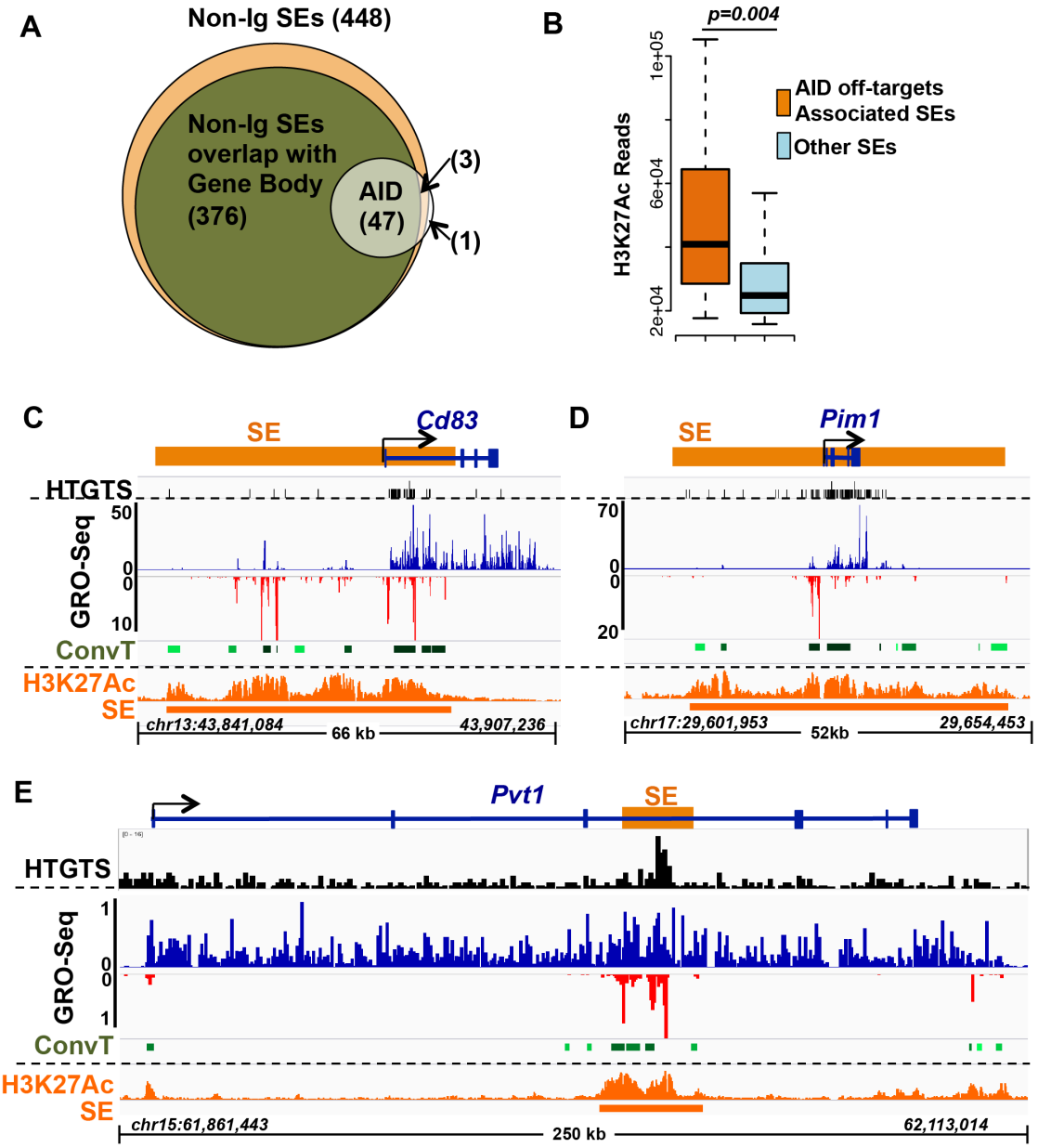


Figure 5.4: AID Off-target ConvT Arising from Intragenic SEs . (A) Venn diagram showing the number of AID off-target regions that overlapped with total non-Ig SEs (448) and with non-Ig SEs overlapping with Gene Bodies (376). (B) H3K27Ac signals of AID off-target-associated SEs (orange) and the other SEs (cyan) are plotted. AID off- target associated SEs had a stronger H3K27Ac signal (Mann-Whitney U-test, *p* value =0.004). Representative AID off-targets are shown based on the SE location indicated in the diagram at the top of each panel. (C) Many AID targets locate downstream of TSSs where SEs and genes overlap. *CD83* is shown as an example. (D) For some relatively small genes located within a larger SE, nearly the whole gene body is an AID off-target, as shown for *Pim1*. (E) SEs inside of very long genes, like *P*vt1 also provide focal AID off-targets. HTGTS, GRO-Seq and H3K27Ac/SE data is illustrated for each panel as described in Figure 2A. The relatively high HTGTS background in *Pvt1* results from long resections downstream of the HTGTS bait DSB in *c-myc*.

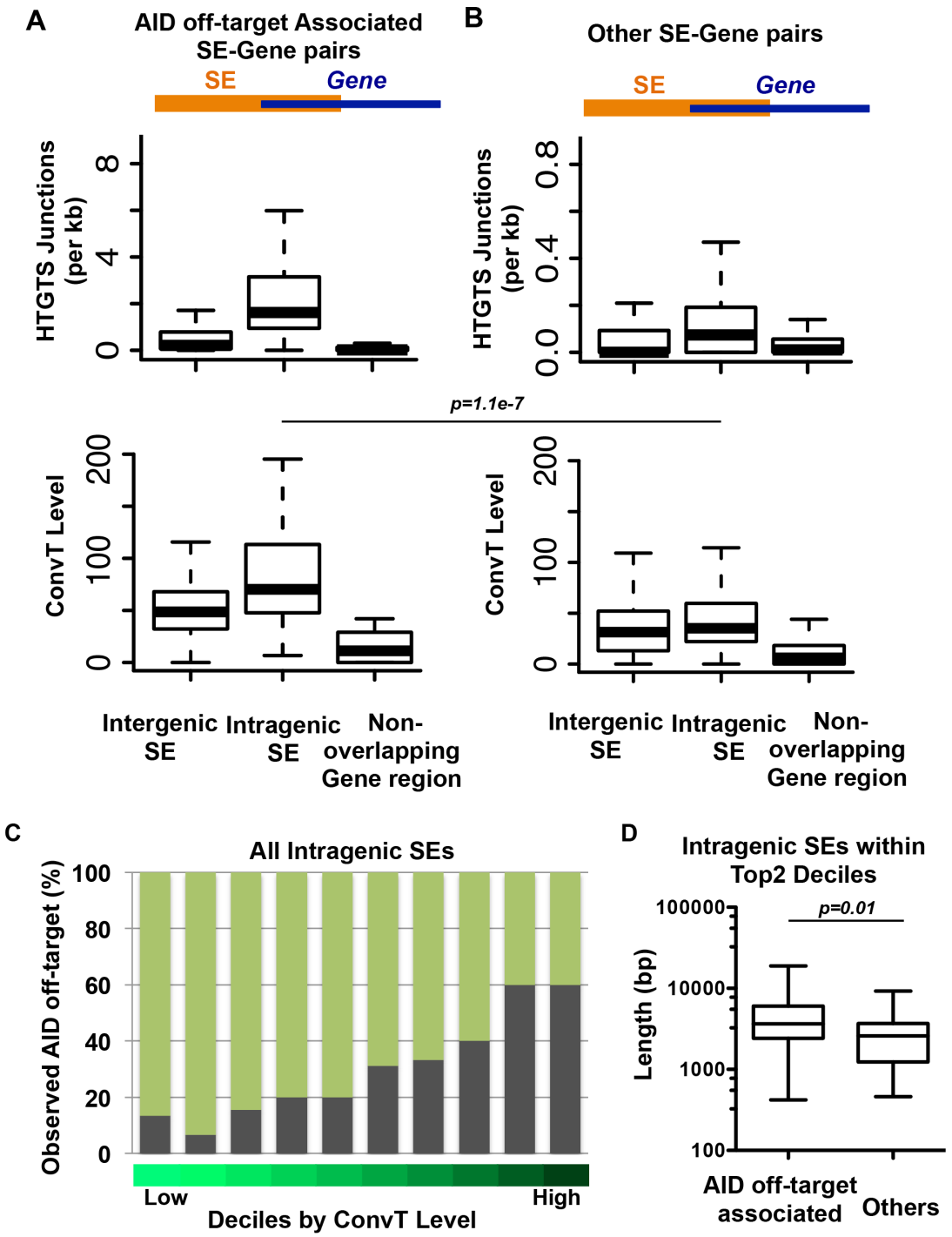


Figure 5.5: Convergently Transcribed Intragenic SEs are Preferred AID off-targets. (A) Upper and Lower: Each SE associated with an AID off-target region and its overlapping gene body were divided into intergenic SEs, intragenic SEs, and non-overlapping gene regions as described in the text and outlined at the top of the Panels. For all AID off- targets, the number of translocation junctions per kb in each of the 3 regions (upper panel) and convergent transcription levels of each region (lower panel) are plotted. (B) Upper and Lower: Each SE that was not associated with an AID off-target region and its overlapping gene body were divided into regions as describe for panel C and translocation junction numbers per kb (upper panel) and convergent transcription levels (lower panel) plotted for each region. A Mann- Whitney U-test was performed to compare two classifications of SEs for convergent transcription

Figure 5.5 (Continued)

ratios within each of the 3 regions; the only significant difference found was that the AID-off-target intragenic SEs has a significantly higher convergent transcription ratio than non- AID off-target intragenic SEs (p value =1.1e-7). (C) All intragenic SEs were grouped into deciles based on the ConvT levels. The fraction of AID off-targets in each decile is indicated by grey bar. (D) Intragenic SEs in the top 2 deciles are divided into those associated with AID off-targets (60%) and those that are not (40%). Length of ConvT regions was plotted and found to be significantly longer in the AID off-target associated intragenic SEs (Mann-Whitney U-test, *p* value =0.01).

Finally, to further address why some SEs are AID targets and others are not, we grouped all intragenic SEs into deciles based on low to high convergent transcription (Figure 5.5C). We then calculated the percentage of the 228 unique AID off-targets revealed by HTGTS (this study) and RPA-ChIP (Qian et al; personal communication) in CSR-activated B cells in each decile. Strikingly, 60% of all SEs within the top 2 deciles (highest convergent transcription) were sites of clustered AID off-target DSBs and translocations. Analysis for SEs in these top two deciles that were AID off-targets versus those that were not did not reveal any obvious sequence differences (e.g. GC content or WRCH and AGCT motifs density). However, ConvT regions associated with SEs in the top two deciles that were AID off-targets were significantly longer than those that were not (Figure 5.5D). These studies strong provide strong evidence that ConvT from intergeneic SEs generates a major class of focal AID off-target regions.

Prior studies of a selected set of AID off-targets divided them into three groups in GC B cells based on mutation frequency in Ung/Msh2 double deficient B cells versus AID-deficient B cells, including 15 Group A genes that had high levels of mutation, 21 group B genes that had substantially lower levels, and 47 group C genes that were infrequently mutated[23]. Our GRO-Seq analyses of GC B cells revealed that nearly 70% of the highly mutated Group A gene off-target regions, including *Pim1*, *Ebf1*, *CD83* and *Ocab*, overlapped with ConvT regions (Figure 5.6A, C) that were well above simulated background levels expected for the most highly transcribed genes (Supplementary Figure C.5A). In contrast, regions reported to have low level mutation frequency (Group B and C genes) showed low correlations with convergent transcription (33% and 32%, respectively; Figure. 5.6A) that were not above simulated background concurrency. Finally, of the five Group A genes that did not associate directly with convergent transcription, SHMs in four occurred quite proximal to ConvT regions (Supplementary Figure C.5C). We identified the SEs by using H3K27Ac ChIP-Seq in GC B cells, and that some were shared between GC and CSR-activated B cells while many others were distinct, consistent with overlapping but distinct GRO-Seq profiles. Of the highly mutated Group A gene regions, nearly half associated with SEs (Figure 5.6B), and all were associated with H3K27Ac peaks (Figure 5.6C). For Group B and C gene regions, concurrencies with SE were 20% and 2%, respectively. Thus, under physiological conditions in the GC, AID also tends to target the convergently transcribed intragenic SEs or, occasionally, typical enhancers.

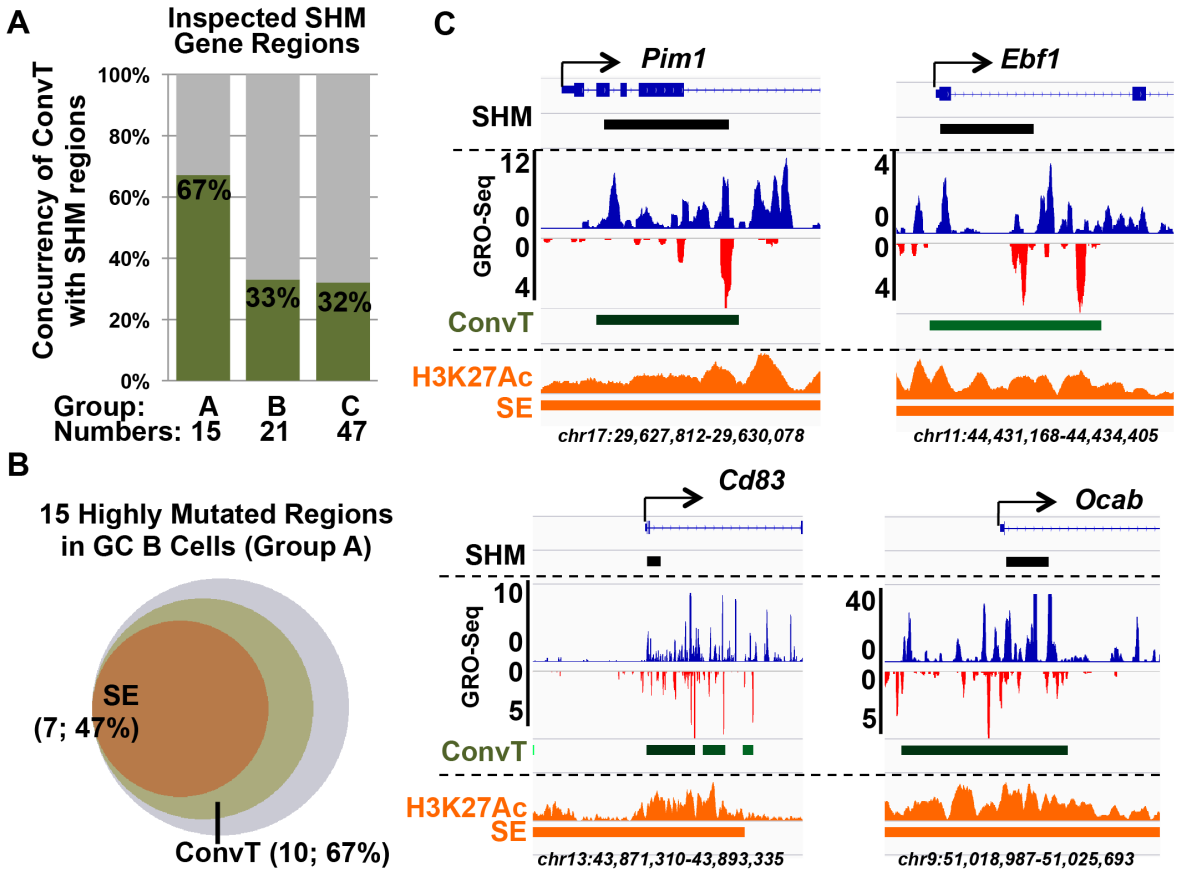


Figure 5.6: Transcription of AID off-targets in GC B Cells. Regions of genes containing SHMs in Ung/Msh2 double deficient GC B cells were analyzed for convergent transcription as determined by GRO-Seq and outlined in Figure 3. GC AID off-target Group A, B and C genes include gene regions with high, intermediate, and low frequencies of AID-dependent mutations, respectively. (A): Concurrency of Group A, B and C gene ConvT regions in GC B cells. (B): Venn diagram showing the number of Group A gene regions that overlapped with SEs and ConvT. (C): Examples of Group A gene regions are shown. Approximately 2-3 kb regions around the TSSs of the indicated genes are shown. The "SHM" diagram at the top of each sub-panel indicates regions of these genes included in the prior SHM analyses with a black bar. GRO-Seq profile, ConvT, H3K27Ac ChIP-Seq profile, and SEs are shown as in Figure 2A.

Ectopic manipulation of endogenous SEs and ConvT regions to assess affects on AID targeting would be problematic since these regions are the actual AID targets. As an alternative approach, we performed GRO-Seq on mouse embryonic fibroblasts (MEFs) in which ectopic AID expression revealed a set of 29 AID off-target sequences, most of which were novel (Wang et al., unpublished). Remarkably, we found that the great majority of these clustered MEF translocations occurred in ConvT regions (Figure 5.7A) that also were mostly also associated with SEs (Qian et al., unpublished data). We also tested the generality of our ConvT findings with respect to AID off-target events observed during SHM in the human Ramos Burkitt's lymphoma cell line. Strikingly, the majority of fifty-four AID off-targets identified in this line were associated with SEs (Qian et al, unpublished data) and we found that most were clustered in regions of strong ConvT (Figure 5.7B). As discussed below, we have also extended our finding to human B cell lymphoma translocations.

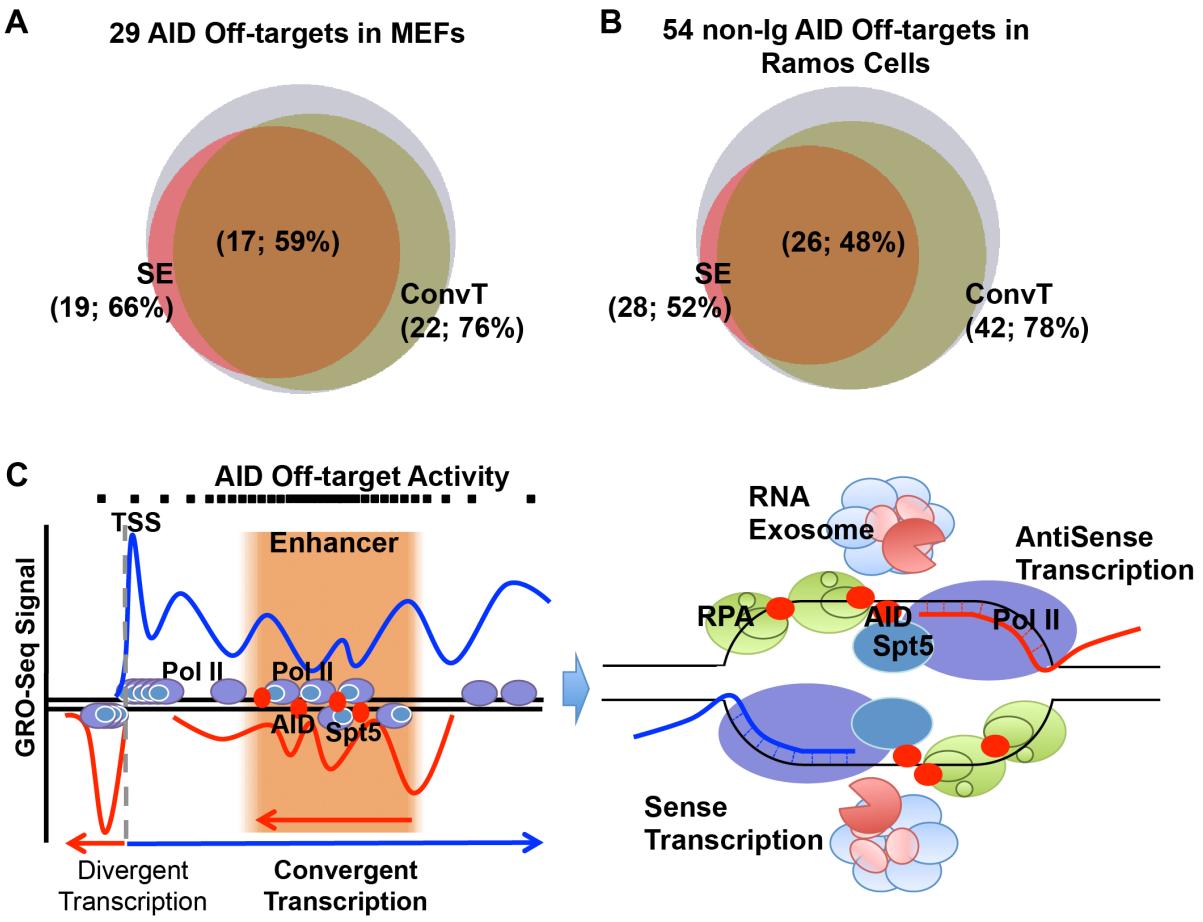


Figure 5.7: Model of AID Targeting at Off-targets. (A) Venn diagram showing the number of AID off-target regions that overlapped with SEs and ConvT in MEFs with ectopic AID overexpression. (B) Venn diagram showing number of AID off-target regions that overlapped with SEs and ConvT in Ramos Human Burkitt's lymphoma cell line. (C) Model of AID "off-targeting". *Left:* At AID off-targets, SEs overlap with gene bodies and this combination generates regions of sense/antisense convergent transcription due to sense gene transcription encountering the enhancer antisense transcription. *Right:* Stalled RNA polymerase with the help of Spt5 recruits AID and generates regions of ssDNA. RNA Exosome or other RNases degrade the aborted sense and antisense transcripts, and works together with RPA to help AID access to the ssDNA substrates. Some aspects adapted from Basu et al., 2011; See Discussion for other details. See also Figure S6 and Table S4.

Discussion

We report that most AID off-target DSBs and translocations in αCD40 plus IL4 CSR- activated B cells occur in and around ConvT regions within genes (Figure 5.3). Furthermore, most of these AID off-target sites in CSR-activated B cells occurred within portions of genes that overlapped with enhancers, the vast majority of which were SEs (Figure 5.4; Casellas et al., unpublished data). Together, these findings implicate a role for SEs within genes in generating robust ConvT and, thereby, in creating susceptibility to AID off-target activity. Notably, we also found that the majority of the regions with highest levels of off-target AID activity in GC B cells or in human Ramos cells undergoing SHM are in focal areas of target genes that contain SEs and undergo robust ConvT. Even in a non-lymphoid cells (MEFs) in which AID was ectopically expressed, we found that the great majority of 29 AID dependent translocation clusters occurred in regions that underwent robust ConvT, confirming our findings for a totally different set of genes in a different cell type. Together, these finding strongly support a mechanistic link between AID off-target sequences and S/AS convergent transcription.

RNA polymerase II (Pol II) transcriptional pausing or stalling contributes to directing AID to Ig gene SHM and CSR targets via a process thought to involve AID association with the Spt5 transcription cofactor[8]. Ig gene V(D)J exons and S regions likely evolved specific features to promote AID targeting[7]. As AID off-target genes lack consistent sequence features of Ig gene AID targets[27], the question of how they attract AID has been long-standing. Our current findings implicate a mechanism that answers this question for the majority of AID off-targets (Figure 5.7C). Thus, most robust AID off-target DSBs, SHMs and translocations occur within intragenic SEs, where we find ConvT that includes sense gene transcription and antisense transcription emanating from the SEs. In such AID off-target regions, antisense eRNA transcription generally occurs at lower levels than sense transcription (Figures 5.2 and 5.4). Thus, most genic sense transcription likely proceeds unimpaired to generate full length mRNAs with only a small fraction encountering antisense transcription, consistent with ability of cells to generate products of these gene[9]. Prior yeast studies showed that, within convergently transcribed sequences, Pol II elongation complexes proceeding in opposite directions cannot bypass each other, and that consequential Pol II collisions lead to stalling or stopping[55]. We propose that such Pol II stalling due to

convergent transcription leads to AID recruitment and further downstream events similar to those implicated in specialized Ig targets (Figure 5.7C) [14,17]. Beyond AID recruitment, convergent transcription could also generate ssDNA substrates for AID. Thus, following Pol II collisions, RNA exosome or other RNase activities could remove nascent transcripts to provide local ssDNA targets (Figure 5.7C).

AID activity generally occurs at much higher levels on specialized Ig gene targets than on off-targets[10,21,22,26]. Whether or not the ConvT mechanism we propose for off-targets can be applied to on-targets remains to be determined. In CSR-activated B cells, we observed ConvT within the very 5' Sγ region (Supplementary Figure C.1H). However, the transcription profile of core S regions cannot be obtained due to poor mappability of repetitive S regions. Clearly, S regions evolved specialized structural features that facilitate AID recruitment and access to the ssDNA substrates[7]. However, mechanisms by which AID specifically targets Ig variable region exons for SHM in GCs may be a more relevant. In this regard, a long- standing paradox involves that fact that SHM of variable region exons occurs only in GC B cells and not in CSR-activated B cells, even though the variable region exons are transcribed in both[10]. Our preliminary analyses reveal potentially higher relative levels of antisense to sense transcription on the downstream edge of the KI V(D)J (VB1-8) exon in GC versus naive or CSR activated B cells (Supplementary Figure C.1H). However, as we cannot map transcription within the main body of the KI VB1-8 due to many highly related unexpressed, upstream V_HJ558 sequences, final testing of this potential mechanism for specific AID targeting of VI(D)J exons will require additional mouse models that eliminate sequence redundancies.

SEs are important for establishment of cell lineage and expression of cell lineage- specific genes[37,40]. Correspondingly, SEs are associated frequently with genes highly expressed in activated B cells. Many of the 51 genes that we have shown to have SEs that are AID off-targets are B cell-specific genes and a notably high proportion (25%) are known oncogenes. In this regard, many human B cell lymphomas contain translocations or mutations of oncogenes that are initiated by off-target AID activity[24,25]. Reminiscent of the AID off-targeting pattern in mouse CSR-activated and GC B cells, human B cell oncogene translocation sites that often occur several kb downstream the TSS[56,57]. Indeed, we have analyzed SEs in human tonsil B cells (enriched in GC B cells) and now found many to be sites of oncogene translocation in human B cell lymphoma, including those in *c-myc*, *Pax5*, *Bcl6*, *Bcl2*, *Pim1*,

Ocab, *Lcp*, and *Bcl7a*, occur in regions where SEs overlap with the gene bodies tonsil B cells (Supplementary Figure C.6C). Thus, beyond contributing to de-regulated oncogene expression[42], our findings suggest that SEs may target oncogenes for translocations in B cell lymphoma. Finally, AID has also been implicated in genomic instability and translocations in cells beyond those of the immune system[58,59]. Our MEF studies suggest ConvT from SEs could play a role in such settings.

Author Contributions

F.L.M., Z.D., A.J.F., J.B., X.S.L. and F.W.A. designed the study. F.L.M, C.A and C.R.W. purified B cells and performed GRO-Seq. Z.D. analyzed GRO-Seq data. F.L.M., A.J.F. and C.A performed H3K27Ac ChIP-Seq and Z.D. and A.J.F. analyzed ChIP-Seq data. J.H. performed HTGTS and J.H. and R.M analyzed HTGTS data. Q.W and M.C.N prepared MEFs and supplied the AID off-target and SE list for MEFs. K.K. and R.C. supplied the AID off-target and SE list for Ramos cells and shared their AID off-target list for CSR-activated B cells. D.N. advised on statistical analysis and various aspects of data analysis. F.L.M, Z.D., and A.J.F. and F.A. designed figures. F.W.A and F.L.M. drafted the manuscript, and F.L. M. Z.D, A.J.F., J.B., X.S.L and F.W.A. polished the manuscript.

Acknowledgements

F.W.A was supported by NIH grants R01AI077595 and P01CA109901 and is an investigator of the Howard Hughes Medical Institute. X.S.L was supported by NIH grant 1R01GM099409. A.J.F and J.E.B. were supported by a Leukemia & Lymphoma Society SCOR, the National Science Foundation, and NIH grants 1R01 CA176745-01 and P01 CA109901. F.L.M. is a Lymphoma Research Foundation postdoctoral fellow and was a Cancer Research Institute postdoctoral fellow. J.Z. is supported by a Robertson Foundation/Cancer Research Institute Irvington Fellowship. ZD was supported by the National Science Foundation of China grant NSFC 31329003. The authors are grateful to Drs. Yi Zhang and Li Shen for assistance with DNA sequencing and Dr. David Schatz (Yale University) for providing primer sequences for GC SHM targets.

Materials and Methods

B Cell Purification.

Splenic naïve B cells were purified from V_HB1-8 heavy chain knock-in mice as described[60]. Naïve B cells were activated with αCD40 plus IL4 for 60 hours to generate CSR-activated B cells. V_HB1-8 knock-in mice were immunized with 5 x 10^8 sheep red blood cells (SRBCs) for 9 days. Splenic GC B cells were purified as described[60].

GRO-Seq and ChIP-Seq.

GRO-Seq[30] and H3K27Ac ChIP-Seq[42] were performed as described. Three biological replicates of each mouse B cell type were performed. Two biological replicates of mouse MEF experiments and one biological replicate of Ramos experiments were performed.

AID Off-targets.

HTGTS was performed with αCD40 plus IL4 or RP105 activated ATM deficient CSR- activated B cells as described (Hu et al., 2014) and also with a new HTGTS method[49]. AID off-target coordinates were retrieved via a new HTGTS pipeline[49].

Data Analysis.

GRO-Seq and ChIP-Seq data sets were aligned using Bowtie[61] to mouse genome build mm9/NCBI37 or human genome build hg19/NCBI37. Uniquely mapped, non-redundant sequence reads were retained. We used Homer[63] to *de novo* identify transcripts from both strands of the genome in the context of the GRO-Seq data, and considered broad sense/antisense overlap regions (>100bp) as ConvT regions. We used the MACS1.4 software[62] to identify regions of ChIP-Seq enrichment over background with a P value threshold of 10^{-5}. We used ROSE software to identify SEs[37].

Chapter 6

Models of Transcriptional Regulatory Circuits in Mammalian Cells

Contributors: Alexander J. Federation*, Violaine St.-Andre*, Brian J. Abraham, Angela Fan, Tong Ihn Lee, Charles Y. Lin, Richard A. Young, James E. Bradner

* Denotes equal contribution

At the time of publication, this chapter has been submitted to *The Proceedings of the National Academy of Sciences*

Introduction

The pathways involved in some complex biological processes, such as these that describe the series of chemical reactions and the accompanying energy flow in metabolism, have been mapped through the efforts of many laboratories over many years and have proven exceptionally valuable for much basic and applied science[1-5]. The control of gene expression programs is a complex biological process that also involves a series of reactions[6-11], but we have limited understanding of the pathways by which key transcription factors (TFs) control the gene expression program of each mammalian cell. These gene control pathways are important to decipher because they have the potential to define cell identity, enhance cellular reprogramming for regenerative medicine and improve our understanding of transcriptional dysregulation in disease.

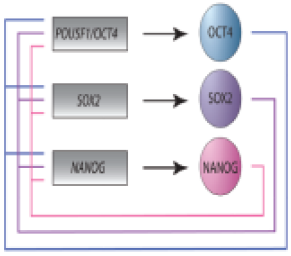
There is considerable evidence that the control of cell-type specific gene expression programs in mammals is dominated by a small number of the many hundreds of TFs that are expressed in each cell type[12-16]. These master, or core, TFs are generally expressed in a cell-type specific or lineage-specific manner and have a powerful ability to reprogram cells from one cell type to another. In embryonic stem cells (ESCs), where transcriptional control has been most extensively studied, the master TFs Oct4/Pou5f1, Sox2 and Nanog have been shown to be essential for establishment or maintenance of embryonic stem cell identity and are among the factors capable of reprogramming most any cell into ESC-like induced pluripotent stem cells (iPSCs)[17-20]. These core TFs bind to their own genes and those of the other core TFs, forming an interconnected autoregulatory loop (Figure 6.1A)[21], a property that is shared by the core TFs of other cell types[22-24]. The core TFs and the interconnected autoregulatory loop they form have been termed “core regulatory circuitry” (CRC)[21-24]. Because the ESC core TFs also bind to a large portion of the cell-type specific genes expressed in these cells, we can posit that regulatory information flows from the CRC to this key portion of the cell’s gene expression program, thus forming a map of information flow from CRC to cell-type specific genes[20].

With limited knowledge of core TFs in most cell types, efforts to map the control of gene expression programs have thus far been dominated by efforts to integrate global information regarding gene-gene, protein-protein, gene-protein and regulatory element interactions nested in these networks (Figure 6.1B)[25-37]. These global studies have provided foundational resources and important insights into

basic principles governing transcriptional regulatory networks, including the presence of recurring motifs of regulatory interactions[38-40] and of gene modules that participate in common biological processes[41-44]. However, these global network maps do not generally capture the notion that key control information flows from a small number of core TFs. Recent studies have revealed that core TFs bind clusters of enhancers called super-enhancers and that the super-enhancer associated genes include those encoding the core TFs themselves[45-46]. The ability to identify super-enhancer associated TF genes, and thus candidate core TFs, should permit modeling of CRCs for all the human cell types for which super-enhancer data is available.

Here we describe a method to reconstruct cell-type specific CRCs based on the two properties of core TFs identified in ESCs and several other cell types: they are encoded by genes whose expression is driven by super-enhancers and they bind to each other's super-enhancers in an interconnected autoregulatory loop. We report CRC models for 75 cell and tissue types throughout the human body. These models recapitulate and expand on previously described CRCs for well-studied cell types and provide core circuitry models for a broad range of human cell types that can serve as a first step to further mapping of cell-type specific gene expression control pathways.

A Core regulatory circuitry



B

Method	System
Genetic perturbation	Yeast
Chromatin conformation	Human
Co-expression	Yeast, Fly, Human
Protein interactions	Human
ChIP-seq + expression	Human
DNAse hypersensitivity	Human
Mass spectrometry	Human
eQTL	Human

Figure. 6.1: Examples of methods commonly used to map transcriptional networks. (A) ESC Core regulatory circuitry model, (B) A summary of approaches used to construct cell regulatory networks and the model systems used for the methodology

Results

To construct maps of core regulatory circuitry (CRC) of human cell types, we used the logic outlined in Figure 2. Detailed studies of the transcriptional control of cell identity in ESCs and a few other cell types have shown that core TFs are 1) encoded by genes associated with super-enhancers (SEs) 2) bind their own SE[45] and 3) form fully interconnected autoregulatory loops with the other core TFs by binding enhancers together with the other core TFs[21-24] (Figure 6.2A). Candidate core TFs were predicted for multiple cell and tissue types using these three criteria. SE-assigned TFs were first selected from the set of TFs expressed in each cell type using H3K27ac ChIP-seq data. Those SE-assigned TFs predicted to bind their own SE due to the presence of DNA sequence motifs at the SE were next identified. From these TFs, the subset predicted to also bind the SE of every other TF in the subset were selected as the core regulatory circuitry (Figure 6.2B).

For 75 human cell and tissue types, we first identified the set of genes that were expressed, encoded TFs and associated with SEs. SEs were defined as previously described[46]. Briefly, genomic regions with exceptionally high levels of signal density from chromatin immunoprecipitation against the enhancer-associated H3K27Ac histone modification were identified as SEs. These regions commonly comprised clusters of enhancers and these individual enhancers were termed SE constituents. SEs were assigned as regulating the closest expressed gene. Recent chromatin conformation data indicates that SEs indeed loop to the promoter of the closest expressed gene in the wide majority of cases, supporting the use of closest expressed gene as the regulatory target of the SE.

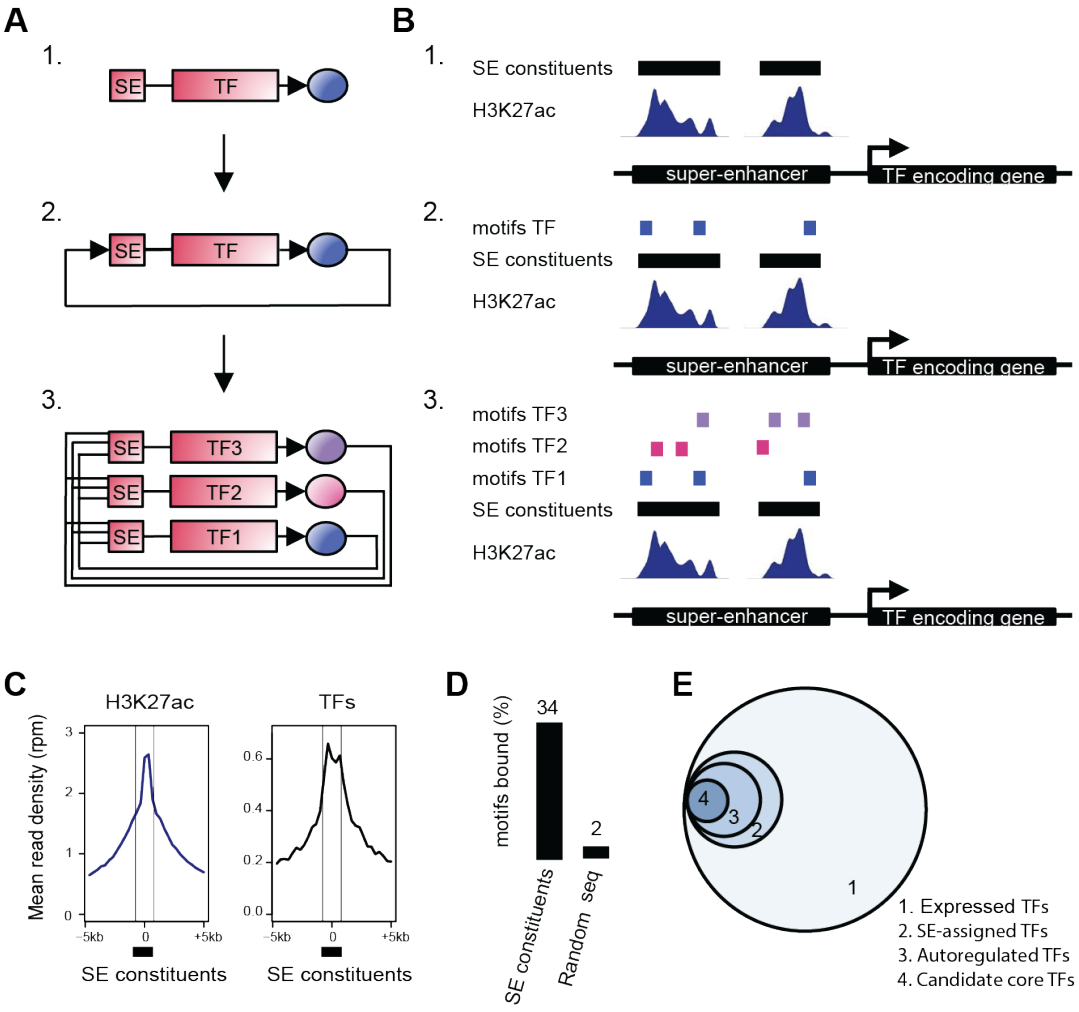


Figure. 6.2. A method to build Core Regulatory Circuitry. (A) Graphical description of the method used to create Core Regulatory Circuitry (CRC) models. 1. Identification of SE-associated TFs. 2. Identification of the TFs that are predicted to bind their own SE and considered as autoregulated. 3. CRCs are assembled as the most representative set of fully inter-connected auto-regulated TFs. (B) Cartoon showing 1. TF associated SE constituents defined by H3K27ac ChIP-seq peak signals. 2. TFs having at least 3 DNA binding sequence motif instances in the sum of their SE constituents are considered autoregulated. 3. SEs having at least 3 DNA binding sequence motif instances for each other predicted autoregulated TFs are associated to genes that together form an interconnected autoregulatory loop. (C) Metagenes of H3K27ac and of the average ChIP-seq signal for Oct4/Pou5f1, Sox2 and Nanog in H1 hESCs on SE constituents +-5kb. (D) Average percentage of DNA binding motifs that are actually bound by the TFs from ChIP-seq data for Oct4/Pou5f1, Sox2 and Nanog in H1 hESCs, in either SE constituents or sets of random genomic sequences of the same size. (E) Venn diagram showing the average numbers, for 84 samples, of: 1. TFs having motifs that are expressed (445), 2. TFs having motifs, being expressed and that are assigned to an SE (61), 3. TFs having motifs, being expressed and assigned to an SE that are predicted to bind their own SE (39), 4. TFs that are part of the CRC (15).

Previous studies have shown that core TFs bind their own enhancers[48-50] so the set of SE-assigned TF genes whose products were predicted to bind their own SEs were next identified. Binding was predicted by searching SE constituents for DNA sequence motifs corresponding to the product of the gene assigned to that SE. DNA-binding sequence motifs for 695 TFs were compiled from multiple published sources[51-55] and SE constituent sequences were scanned for the presence of the TF binding motifs, using the FIMO software package from the MEME suite[56]. SE constituents, as opposed to full SEs, were used as TF binding distributions peak on the SE constituent sequences defined by H3K27ac ChIP-seq peak signal (Figure 6.2C). Furthermore, the presence of multiple DNA sequence motifs at SE constituents is predictive of the binding of a TF, whereas this is not the case on average across the genome (Figure 6.2D). SE-assigned TF genes that were predicted to bind their own SE were considered autoregulated, as prior evidence in ESCs indicates that such genes do regulate their own expression[21,49,57-61].

In ESCs and a few other cell types, the core TFs occupy both the enhancers of their own genes and those of other core TFs, forming an interconnected autoregulatory loop[21,23,62-63]. From the set of TFs considered autoregulated, we identified those that are predicted to bind the other autoregulated TFs based on the presence of motifs in SE constituents sequences, and assembled candidate interconnected autoregulatory loops. For each cell or tissue type, we selected the loop containing the set of TFs most often represented across the possible loops as the representative model of CRC (Supplementary Figure D.1). On average, across 75 cell types, 15% of the genes considered expressed and encoding TFs were assigned to an SE, 9% were predicted to be autoregulated, and 3% were identified as core TF candidates (Figure 6.2E).

The CRC predicted for human H1 ESCs (Figure 6.3A, left panel) indicates that the approach described here captures the previously described TFs of ESC core regulatory circuitry and suggests that additional TFs contribute to this circuitry. The H1 ESC CRC contains three factors - OCT4/POU5F1, SOX2, and NANOG - that are considered the foundation of the core regulatory circuitry in ESCs[66-69]. All three factors are essential for the pluripotent state[70-77], regulate their own genes and those encoding the other two factors[21,49,60,78-79], and can be used to reprogram fibroblasts to an induced pluripotent state[17-19,80].

The results of the algorithm we developed suggest that seven additional TFs contribute to the ESC CRC (Figure 6.3A); most of these factors have previously been implicated in control of stem cell state, and there is ChIP-seq evidence indicating that their super-enhancers are bound by Oct4/Pou5f1, Sox2 and Nanog (Figure 6.3B). FOXO1 and ZIC3 have previously been shown to be essential for the maintenance of pluripotency[42,81-82]. In hESC, Foxo1 regulates OCT4/POU5F1 and SOX2 expression[81]. In mESCs, ZIC3 directly activates Nanog expression and can contribute to reprogramming of human fibroblasts into an induced pluripotent state[82]. NR5A1/SF1 can influence the pluripotent state[83]. NR5A1 and RARG both bind to regulatory regions of the OCT4/POU5F1 gene and regulate its expression[84-87]. The other three TFs - MYB, RORA and SOX21 - are best known for their roles in other stem cells. MYB and RORA have roles in establishing or maintaining self-renewing populations of hematopoietic cells[88-92], while SOX21 is involved in regulating pluripotency in intestinal stem cells, where its expression is influenced by Sox2[93]. Thus, there are multiple lines of evidence that support the inclusion of OCT4/POU5F1, SOX2, NANOG, FOXO1, ZIC3, NR5A1, RARG, MYB, RORA and SOX21 in a model of hESC core regulatory circuitry.

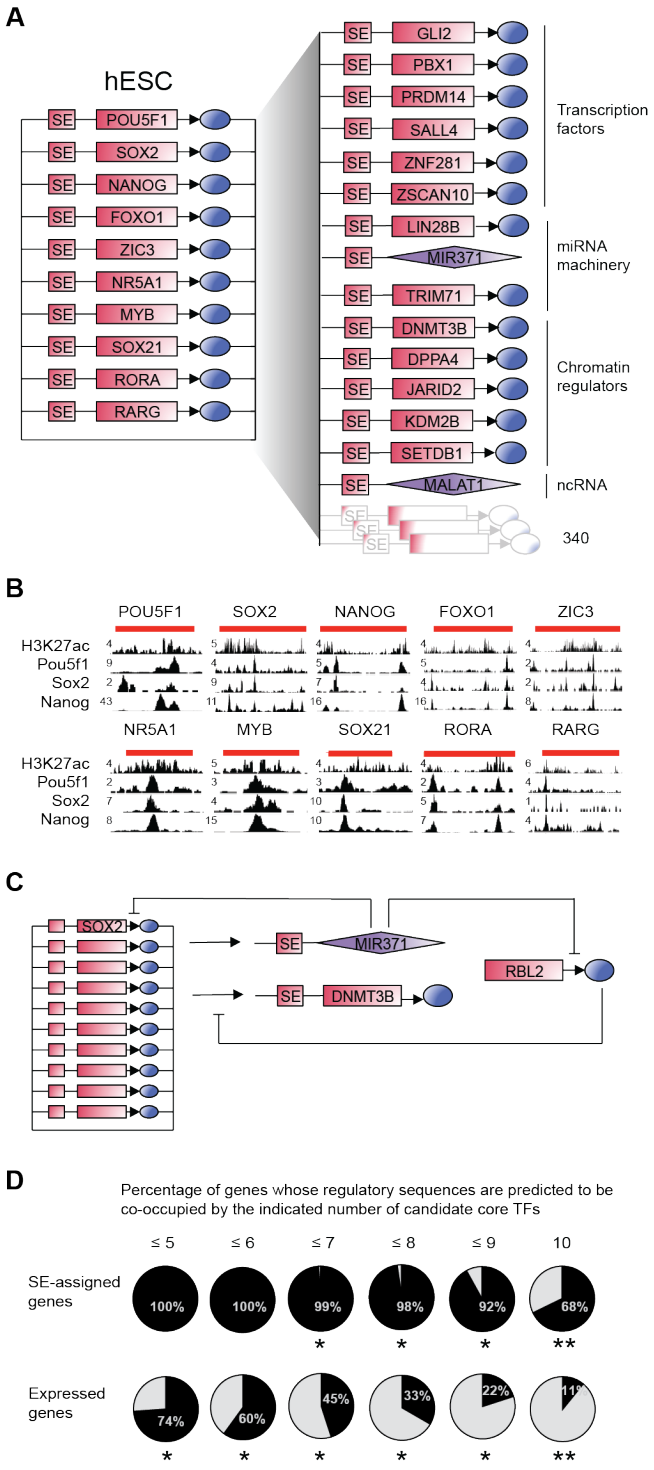


Figure. 6.3: H1 core and extended circuitry. (A) Left: CRC map for human embryonic stem cells (H1). The role of each TF in ESC pluripotency and self-renewal is listed in Table1. Right: H1 hESC extended regulatory circuit. Example of SE-assigned genes that are predicted to be bound by each of the TFs in the

Figure. 6.3 (continued)

CRC. (B) ChIP-seq data for H3K27ac and Oct4/Pou5f1, Sox2 and Nanog showing binding of each TF to each of the SE of H1 CRC. SE genomic locations are depicted by red lines on top of the tracks. (C) Diagram showing transcriptional regulation of mir371-373 on Sox2 core TF expression and on Dnmt3b expression through regulation of Rbl2. (D) Pie-charts showing the percentages of SE-assigned genes (up) or expressed genes (bottom) whose regulatory sequences are predicted to be bound by increasing fractions of candidate core TFs. P-values were calculated using random sampling on the set of TFs that are expressed in H1. P-values < 1.e-2 and < 1.e-3 are represented by * and ** respectively.

OCT4/POU5F1, SOX2 and NANOG contribute to the formation of SEs at hundreds of active ESC genes that play prominent roles in cell identity[45], suggesting that a simple extended model of regulatory information can be constructed to include these additional SE-assigned genes downstream of the core TFs (Figure 6.3A, right panel). This model of extended hESC regulatory circuitry contains many genes that are known to play prominent roles in ESC biology[20]. These include the TFs PRDM14, SALL4 and ZNF281, the chromatin regulators DNMT3B, JARID2 and SETDB1 and the miRNA cluster miR-371-373, all of which have established roles in pluripotency, self-renewal or differentiation. We therefore suggest that the ESC gene expression program is controlled by a CRC consisting of ten key TFs that 1) bind the SEs of their own genes and autoregulate their own expression and 2) co-bind the SEs of many other genes important for ESC identity and regulate their expression.

Among the SE-assigned genes, some transcription regulators may create feedforward or feedback loops of regulation with the genes in the extended CRC to modulate the direct effect of core TFs. This could be the case, for example, of miR371-373, which may fine-tune the expression of Sox2 hESC core TF. Indeed, Sox2 is given as a highly probable target of miR371-375 by the TargetScan software[94]. miR371-375 may also up-regulate Dnmt3b DNA methylase through Rbl2 expression inhibition, as does its murine homolog in mESCs[95-96] (Figure 6.3C).

The regulatory regions of SE-assigned genes are predicted to be co-occupied by most of the candidate core TFs. SE-assigned genes have motifs in their enhancers and promoters predicting the binding of higher fractions of candidate core TFs than the average expressed genes (Figure 6.3D). 68% of the SE-assigned genes are predicted to be bound by each of the core TFs. This is not the case when random sets of expressed TFs are used instead of the set of candidate core TFs (permutation test p-value = 4 e-4), showing that co-occupancy of SE–assigned gene is a feature of candidate core TFs. Experimental evidence shows that Oct4/Pou5f1 contributes to the regulation of at least 30% (proportion test p-value < 2.2 e-16) of these downstream SE-assigned target genes. Thus, in the model of extended ESC regulatory circuitry, the core TFs co-occupy and likely regulate the SEs of a large portion of genes that are key to ESC identity.

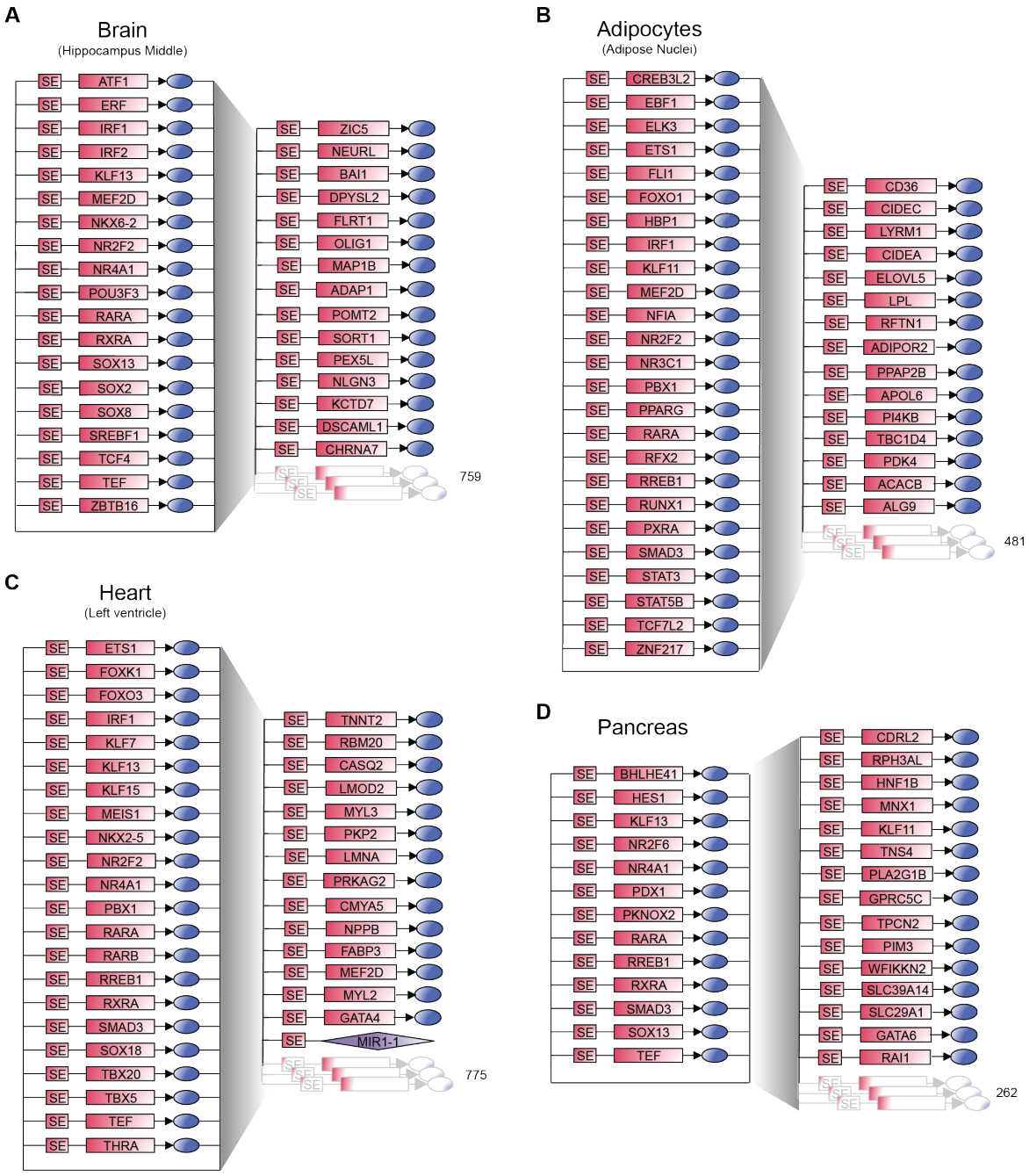


Figure. 6.4: **Core and extended regulatory circuitry maps for multiple cells and tissue types.** Core and extended circuitry maps for A) brain (hippocampus middle), B) adipocytes (adipose nuclei), C) heart (left ventricle), and D) pancreas. The number of SE-assigned genes predicted to be co-occupied by the candidate core TFs and 15 examples of those are displayed on the right part of the maps.

A model of CRC and extended regulatory circuitry was predicted for each of 75 human cell and tissue types (Figure 6.4). The predicted CRCs contain key transcriptional regulators of cell identity that have been previously identified (Supplementary Figure D.2). This includes, for example, TBX5 in the heart CRC (left ventricle)[97], PDX1 in the pancreas CRC[98], and SOX2 in the brain CRC[99] (hippocampus middle). ChIP-seq data for TFs in the CRCs are available for 3 cell types and support the predicted binding interactions (Supplementary Figure D.3). This indicates that the CRC models capture much existing knowledge of TFs that play key roles in control of cell identity across cell and tissue types.

The candidate core TFs identified across a wide range of cell types are cell-type-specific or lineage-specific. Analysis of the TF composition of the CRCs across samples shows that the majority (2/3) of the core TFs are cell-type specific, but a substantial fraction are expressed in multiple cell types, typically within a lineage (Supplementary Figure D.4). This feature of shared core TFs in lineages is evident in hierarchical clustering of candidate core TFs (Figure 6.5A).

Models for extended regulatory circuitry were generated for 75 cell and tissue types using the same process described above for the hESC extended regulatory circuitry (Figure 6.4). The features of these extended circuitries are consistent with those observed for hESCs. On average, across samples, 73% of the SE-assigned genes are predicted to be co-occupied by each of the core TFs (Figure 6.5B) and these SE-assigned genes play prominent roles in specific cell identities. The CRC and extended regulatory circuitry models provide the foundation for further study of the transcriptional regulation of human cell identity.

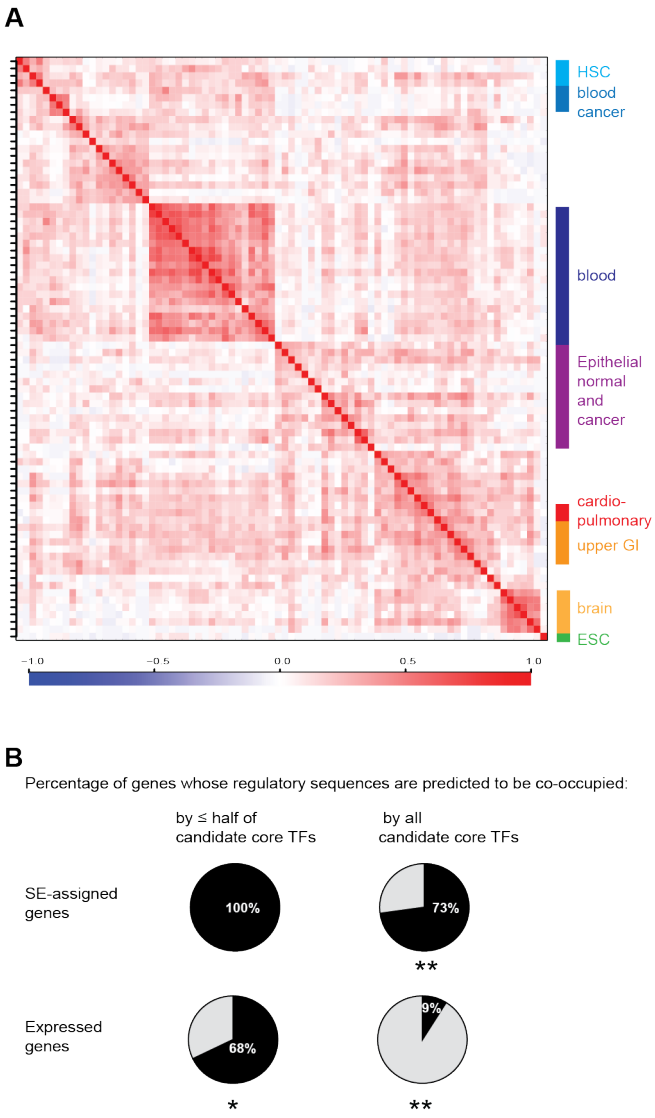


Figure. 6.5: Properties of CRCs of multiple human cell and tissue types. (A) CRCs cluster according to cell type similarities. Hierarchical clustering of candidate core TFs for 80 human samples. The matrix of correlation based on Pearson coefficients identifies specific clusters for Hematopoietic Stem Cells (HSC), blood cancer cells, blood cells, epithelial normal and cancer cells, cardio-pulmonary system cells, upper gastro-intestinal system and brain cells. Correlation values range from -1 to 1 and are colored from blue to red according to the color scale. (B) Pie-charts showing the average percentages, on 84 samples, of SE-associated genes (up) or of all expressed genes (bottom), whose regulatory sequences are predicted to be co-occupied by more than half or by each of the TFs in the CRC. P-values were calculated using random sampling on the set of TFs expressed in each cell or tissue type. P-values < 1.e-2 and < 1.e-3 are represented by * and ** respectively.

Discussion

We describe here the first maps of core regulatory circuitry of cell identity for 75 human cell types and tissues. Identifying the small number of TFs responsible for establishment and maintenance of cell identity is critical to decipher the transcriptional pathways that have the potential to define cell identity in the hundreds of cell types of the human body. The CRCs we predict include known master TFs and reprogramming TFs that have been previously identified in a few cell types through extensive genetic studies. In particular, the CRC predicted for hESCs recapitulates and expands on the previous model of hESC CRC, and may be critical towards realizing their therapeutic potential. Most importantly, these models predict novel putative regulators of cell identity in many cell types for which no core factors have yet been characterized.

Key target genes of the CRC were identified in a first step towards understanding how the information flows from the core TFs to all expressed genes. Across all cell and tissue types, the candidate core TFs were predicted to preferentially co-occupy the regulatory sequences of SE-assigned genes, compared to the ones of the average expressed genes. As SE-assigned genes are typically key for cell identity[45,46,100-101], this shows that the concerted action of candidate core TFs may be preferentially targeted to those key cell identity genes. A small number of core TFs organized into an interconnected autoregulatory loop, and able to bind most of the key cell identity genes, thus appears to be a common theme for the regulation of cell-type-specific gene expression programs. This topology of interactions may provide regulatory mechanisms by which cell identity can be robustly maintained, while allowing cells to respond to developmental cues. This led us to envision a model whereby the core TFs promote hallmarks of cell identities through 1) co-binding the SEs of their own genes and regulating their own expression, and 2) co-binding of the SEs of many other genes important for cell identity and regulating their expression. The maps of CRC were thus extended to include the SE-assigned target genes of the CRC. Those maps of extended regulatory circuitry are founding models for the description of more comprehensive networks. These may include more complex feed-forward and feedback loops of regulation and their relationship to signaling pathways.

The circuitry maps for many cell types, along with the possibility of predicting the circuitry of any cells for which SE data can be generated, should provide guidance for reprogramming studies.

Identification of CRC in ESCs has largely contributed to the success of reprogramming of differentiated cells into undifferentiated cells. Better knowledge of the core TFs of more differentiated cells will help with reprogramming differentiated cells into other cell types, which has high potential for clinical applications. Hierarchical clustering of candidate core TFs shows that CRCs of cell-types from the same lineages share candidate core TFs. This may result from the sequential activation and repression of TFs during the course of differentiation[102] and indicates that specific combinations of TFs may be required to control complementary aspects of cell identity. It also suggests that minor changes in core TFs might enable trans-differentiation between similar cell-types in reprogramming experiments.

The circuitry maps may also prove particularly valuable for better understanding transcriptional dysregulation in disease. Previous studies have shown that mutations in the binding sites of a core TF could mediate the formation of a SE involved in regulating the expression of its associated gene[103-104]. SEs are hotspots of non-coding disease-associated sequence variation[105-106] and some of these variants may modify the binding sites for core TFs and lead to such gene expression dysregulation mechanisms. Because of the cell-type specific usage of SEs, these disease-associated variants should have phenotypic consequences in the tissue in which the SEs are functional only, which may help explain the cell-type specificity of diseases. Extended regulatory circuitry maps integrating candidate core TFs and their SE-assigned target genes for many human cell-types, may thus help better understand disease-associated genetic variation function and the transcriptional pathways that lead to pathologies.

Methods

ChIP-seq data

H3K27ac ChIP-seq sequence reads were either downloaded from GEO or generously shared by the NIH Roadmap Epigenome project and were aligned to the hg19 version of the human genome using Bowtie 0.12.9 with parameters –k1 –m1 --best.

Identification of H3K27ac ChIP-seq peaks and super-enhancers

CRC Mapper

SEs identified with ROSE are assigned to the closest expressed transcript, considering the distance of the transcription start site (TSS) to the center of the SE. H3K27ac read density at TSS+-1kb is

used to rank the transcripts in each sample and the top 2/3 genes ranked by the highest value of any of their transcripts define a threshold value of read densities to select for expressed genes. The ratio of top 2/3 genes was determined as the ratio that allows recovering the highest percentage of expressed genes, minimizing false negatives, based on comparisons with micro-array and RNA-seq data in ESC (data not shown). This ratio of expressed genes is consistent with the ratio of genes considered expressed across cell types. Transcripts that have a signal at their promoter above this threshold were selected for subsequent analysis.

SE-associated transcription factors (TF) are then selected from the lists of SE-associated genes using a list of 1253 TFs consisting in the intersection of AnimalTFDB and TcoF lists of TFs minus CTCF, GTF2I and GTF2IRD1 that are usually not considered as TFs.

Motif analysis

For each TF encoded by a gene assigned to a SE, we identified whether sequence-specific binding motif information has been predicted for these proteins. We compiled a database of DNA sequence motifs for 695 TFs – about 60% of known TFs in vertebrates - from multiple sources. The database of motifs used is composed of the TRANSFAC database of motifs and the vertebrate motifs from the MEME database (January 23rd 2014 update): JASPAR CORE 2014 vertebrates, Jolma 2013, Homeodomains, mouse UniPROBE, mouse and human ETS factors. For TFs with previously identified sequence-specific binding motifs, the motif linked to the TF was searched for in the SE sequences assigned to the gene encoding that TF. For the motif search, the search space in SEs was restricted to SE constituents, as these are the regions that capture most of the TF binding in SEs. SE constituent DNA sequences from all the identified SEs in a given sample were extracted and extended on each side (500 bp by default) to allow for TF binding motif identification using FIMO (Find Individual Motif Occurrences) from the MEME (Multiple Em for Motif Elicitation) suite with p-value threshold of 1e-4, specific background, and our compiled library of motifs.

Identification of fully interconnected auto-regulatory loops

The SE constituents that have motifs for their associated TF are then identified from the FIMO output. From this list of genes, the TFs that have at least 3 DNA sequence motif instances for their own protein products in the sum of their assigned SE constituents are defined as autoregulated TFs. All

possible fully interconnected autoregulatory loops are then identified using an algorithm based on the recursive identification of all possible cliques from a graph. When multiple possibilities of fully interconnected autoregulatory loops were identified for a sample, their composition in TFs highly overlapped between the loop possibilities, there were not multiple independent fully interconnected autoregulatory loops. In those cases, in order to select a representative CRC, we selected the most representative fully interconnected autoregulatory loop as the one containing the TFs that appeared the most often across all possible loops. For each TF in the cliques, its number of occurrences across all possible fully interconnected autoregulatory loops was calculated. The loops were ranked based on the sum of the scores of each TF in the loop, divided by the number TFs in the loop. The best ranked loop was selected as the CRC.

Metagenes

Genome-wide meta-representations of ChIP-seq density were created by mapping aligned reads to SE constituents +- 5kb. Each SE constituent and flanking region was split into equally sized bins and the average ChIP-seq density in each bin was calculated.

Transcription factor binding to motif in SE constituents

H1 human embryonic stem cells Oct4/Pou5f1, Sox2 and Nanog ChIP-seq data were used to quantify the binding of TFs to their cognate motifs +-1kb in SE constituents extended 500 bp on each side or on the same number of random genomic regions of the same size. We quantified the number of sequences containing motifs that overlapped with the ChIP-seq peaks identified by MACS ran with parameter -p 1e-9 keep-dup=auto -w -S -space=50. The true positive rates of TF binding was calculated by dividing the number of motif containing sequences that were bound by the TF from the ChIP-seq data analysis, over the total number of motif containing sequences.

ChIP-seq tracks

H1 hESC ChIP-seq data for Oct4/Pou5f1, Sox2 and Nanog were downloaded from GEO and processed similarly to H3K27ac ChIP-seq data. Chip-seq data for Creb1, Ebf1, Elf1, Ets1, Ikzf1, Pax5 and Pou2f2 in GM12878 lymphoblastoid B cells; for Tcf7L2 in HCT-116 colon cancer cell line; and for Esr1 in T-47D breast cancer cell line; were downloaded from ENCODE and processed similarly to H3K27ac ChIP-seq data.

CRC target gene analysis

For the CRC target gene analysis two groups of target genes were considered: expressed genes and SE-assigned genes. Expressed genes correspond to the top 2/3 genes ranked based on H3K27ac signal at their TSS+-1kb. SE-assigned genes were identified from the list of expressed genes as described above. In each group, genes that have motifs instances predicting the binding of a defined number of candidate core TFs in the sum of their enhancer + promoters sequences were quantified. TSS+-1kb and associated super or typical enhancer constituents extended 500bp on each side were used for the motif search when all expressed genes were considered, and SE constituents extended 500bp on each side + corresponding TSS+-1kb sequence of the SE-assigned gene were used for the motif search when SE assigned genes were considered. The same analysis was done with random sampling selection of the same number of expressed TFs that have motifs in our database to calculate p-values of significance. In H1 hESC, 888 TFs of the list of 1253 vertebrate TFs (71%) are expressed in H1 and 389 of these 888 (44%) have motifs in our database.

Lineage clustering

Hierarchical clustering on candidate core TFs was done in R. A matrix of distances was calculated based on Pearson correlations between the candidate core TF lists and plotted using the R image function. For a better robustness of the clustering, only the 80 samples that had more than 7 TFs in their CRC were used for this analysis.

Chapter 7

Medulloblastoma regulatory circuitries reveal subgroup-specific cellular origins

Contributors: Charles Y. Lin, Serap Erkek, Daisuke Kawauchi, Alexander J. Federation, Rhamy Zeid, Marc Zapatka, Barbara Worst, Hans-Jörg Warnatz, Sebastian Waszak, David T.W. Jones, Marcel Kool, Volker Hovestadt, Ivo Buchhalter, Laura Sieber, Pascal Johann, Thomas Risch, Vyacheslav Amstislavskiy, Marie-Laure Yaspo, Hans Lehrach, Marina Ryzhova, Andrey Korshunov, Roland Eils, Peter Lichter, Jan O. Korbel, Stefan M. Pfister, James E. Bradner and Paul A. Northcott

At the time of submission, this chapter is under review for publication at *Nature*

Introduction

Medulloblastoma is a highly malignant paediatric brain tumour consisting of four biologically and clinically distinct molecular subgroups[1,2]. Transcriptional diversity underlying WNT, SHH, Group 3, and Group 4 subgroup medulloblastoma is partially explained through activation of discriminatory signaling pathways, including the Wingless/WNT and Sonic hedgehog/SHH developmental cascades inherent to WNT and SHH medulloblastomas, respectively. Recurrent, somatically altered driver genes such as *MYC* (Group 3), *KDM6A* (Group 4), the recently implicated *GFI1/GFI1B* (Group 3 and Group 4), and others are likewise suspected to be responsible for diversity between medulloblastoma subgroups[3-5]. Furthermore, distinct cellular origins have been experimentally substantiated for WNT[6] and SHH[7-9] tumours, whereas clues into the origins of Group 3 and Group 4 remain elusive. Understanding the molecular, cellular, and biological diversity underpinning medulloblastoma subgroups is of paramount interest to the paediatric neuro-oncology community as current treatment regimens involving invasive surgery, cranio-spinal irradiation, and cytotoxic chemotherapy too often impose serious detrimental consequences on the developing child. Development of more effective therapies with reduced toxicity will necessitate a more complete understanding of medulloblastoma, with the expectation that in the future it will be treated not as a single disease but more aptly as four distinct diseases according to subgroup.

Recent next-generation sequencing (NGS) studies of medulloblastoma have improved our perspective of recurrently mutated genes and pathways, the proportion of cases affected by such alterations, and their respective subgroup distribution[3,10,11]. The bulk of these efforts have thus far been focused on somatic, DNA-level genomic alterations, especially non-synonymous single nucleotide variants (SNVs), indels, and focal copy-number aberrations. Recurrent targeting of genes involved in chromatin modification has been the most consistent theme to emerge from these studies, strongly suggesting deregulation of the epigenome as a critical step during medulloblastoma pathogenesis. Still this hypothesis has yet to be substantiated and knowledge pertaining to how the medulloblastoma epigenome influences subgroup-specific transcriptional programs remains in its infancy. Recent DNA methylome-based analyses[12] have shed some light in this realm, but mostly without the added information gained from studying histone modifications, especially those that demarcate active regulatory elements such as enhancers that drive transcription.

Enhancers are distal *cis*-acting regulatory elements that serve as sites of recruitment for transcription factors and the associated chromatin machinery, potentiating transcriptional control of target gene(s)[13]. Massive catalogues of genome-wide enhancers have been inferred and published by large consortia such as ENCODE[14,15] and Roadmap[16], dramatically advancing our understanding of enhancer-gene regulation across a comprehensive spectrum of cell lines and tissues from different species. However, since enhancers exhibit extensive diversity between cell types, such enhancer 'encyclopaedias' may have context-specific utility and share limited overlap with regulatory elements specific to cell types and entities that are currently underrepresented in these large consortia-based studies. Herein, we used histone ChIP-sequencing to describe the enhancer landscape of medulloblastoma across a series of 28 primary tumour tissue specimens and 3 additional cell lines representative of the molecular subgroups. Our approach to studying enhancers genome-wide in a large set of primary tissue samples led to the identification of a wealth of previously unknown regulatory elements, including a large proportion that are medulloblastoma subgroup-specific. Moreover, our data provide novel insight into potential targetable oncogenic pathways and medulloblastoma cellular origins, especially for the poorly characterized Group 3 and Group 4 subgroups.

Results

Large-scale efforts aimed at systematically annotating active regulatory elements genome-wide (e.g. through DNase I hypersensitivity, H3K27ac and BRD4 ChIP-seq) have recently been published by large consortia[14,16]. Although incredibly comprehensive in scope, these studies have primarily utilized high passage cancer cell lines, immortalized non-neoplastic cell lines, or bulk normal human tissues for cataloguing active enhancers. Medulloblastoma has been severely under-represented in such reports, with only a single long-term culture cell line (D721; first reported in 1997) included amongst 125 cell types initially studied by ENCODE[15].

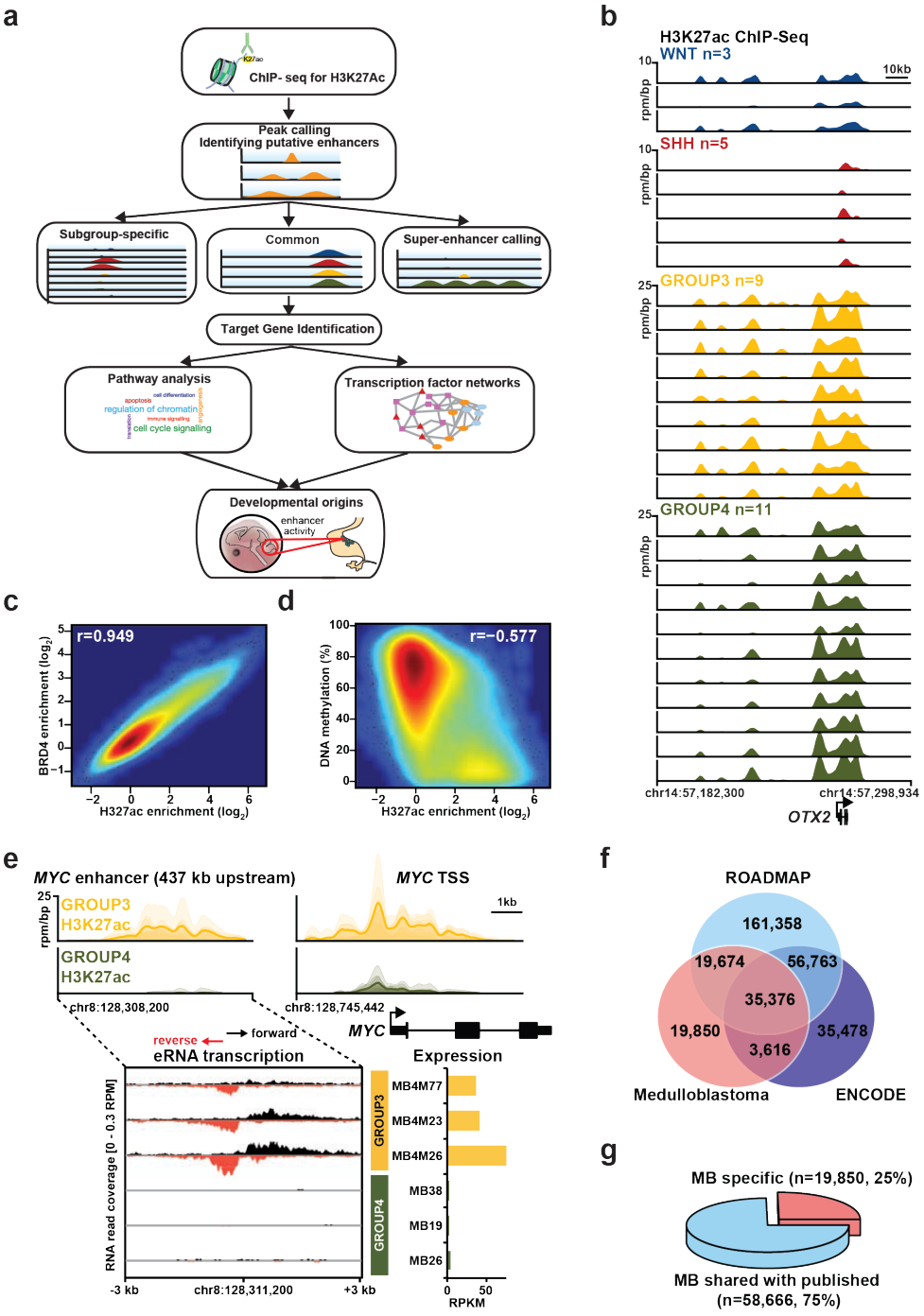


Figure 7.1: The enhancer landscape of primary medulloblastoma. (a) Experimental workflow for studying enhancers and super-enhancers in primary medulloblastomas. (b) H3K27ac ChIP-seq data showing a highly active enhancer at the *OTX2* locus across all 28 primary medulloblastoma samples from our series. (c) Scatter plot comparing the enrichment (log_2) of H3K27ac signal versus BRD4 signal at medulloblastoma enhancers (n=78,516) as determined by ChIP-seq. (d) Scatter plot comparing the enrichment (log_2) of H3K27ac signal versus DNA methylation at medulloblastoma enhancers (n=78,516) as determined by ChIP-seq and WGBS, respectively. (e) RNA-seq data showing Group 3-specific enhancer RNA (eRNA) expression (lower left) overlapping a Group 3-specific *MYC* enhancer (upper left) in a subset of Group 3 and Group 4 medulloblastomas. MYC gene expression (RPKM) is also shown for the same subset of cases (lower right). (f) Venn diagram showing the overlap of medulloblastoma enhancers with those reported by ENCODE and Roadmap.

Cancer cell lines often exhibit drastic genomic and transcriptional divergence from their corresponding primary tumour tissues. This is exemplified in Non-Hodgkin's lymphoma where there were stronger epigenomic similarities between primary tumour samples and normal tissues than between tumours and cell lines[17]. Given the apparent limitations of using cell lines to faithfully study the tumour epigenome, and the recognized subgroup-dependent heterogeneity of medulloblastoma, we collected a series of 28 treatment-naïve, fresh-frozen medulloblastoma tumour specimens for studying the active enhancer landscape by H3K27ac ChIP-Seq (Figure 7.1a,b). The cohort was selected to be inclusive of all four medulloblastoma subgroups (WNT, n=3, SHH, n=5, Group 3, n=9, Group 4, n=11). Three additional Group 3 cell lines (MED8A, D425, and HD-MB03) were also included in our experimental workflow. Parallel ChIP-Seq was performed for *Bromodomain Containing 4* (BRD4), a chromatin reader and transcriptional coactivator required for enhancer activity[18,19], in 27/31 cases (Figure 7.1c). Enrichment of H3K27ac and BRD4 ChIP-Seq signals was highly correlated (Pearson correlation, r=0.949) at putative enhancer loci, suggesting that the regions we have inferred are indeed active enhancers (Figure 7.1c)[18,19]. In contrast, regions enriched for H3K27ac were strongly anti-correlated with DNA methylation (Pearson correlation, r=-0.577; Figure7.1d). Finally, analysing strand-specific RNA-Seq data generated from the same tumour samples subjected to ChIP-Seq, we observed notable short, unspliced, bidirectional RNA transcripts overlapping active enhancers (Figure 7.1e), in accordance with recently described enhancer RNAs (eRNAs) associated with active enhancers[20]. Using MACS[21] to identify significantly enriched H3K27ac peaks, we inferred 78,516 medulloblastoma enhancers, which mainly (~80%) covered introns and intergenic regions. Comparison of predicted medulloblastoma enhancers with those reported using analogous methods by ENCODE and Roadmap revealed 19,850 novel regions, indicative of potentially cerebellar cell type- or medulloblastoma-specific enhancers in our dataset (Figure 7.1f, g). Importantly, primary medulloblastoma enhancer landscapes exhibited poor overlap and correlation with those generated from medulloblastoma cell lines, further emphasizing the importance of profiling regulatory elements in primary tumours.

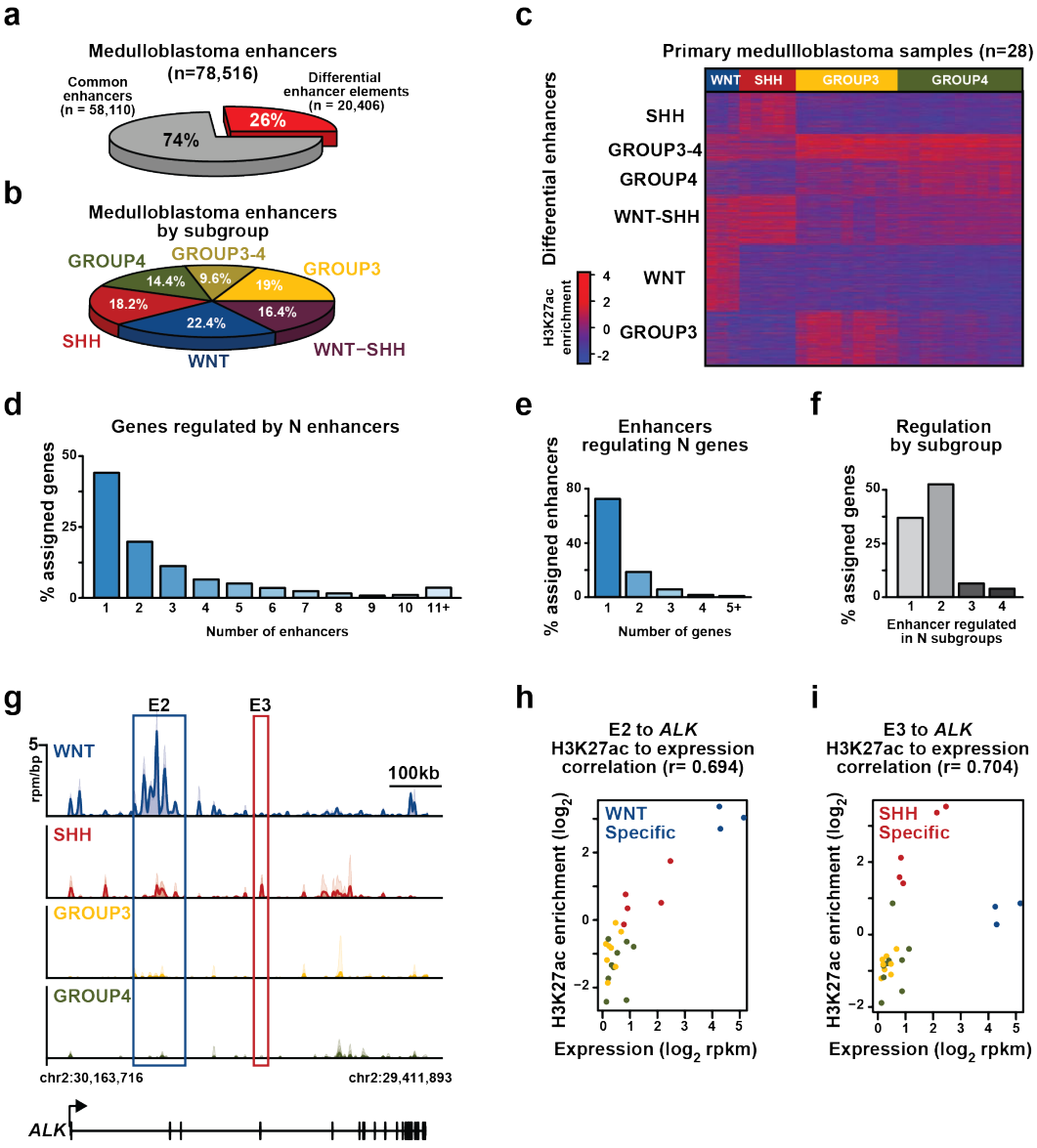


Figure 7.2: Differentially regulated enhancers and enhancer-gene assignments in medulloblastoma subgroups. (a) ANOVA classification of medulloblastoma enhancers displayed as a pie chart. (b) Pie chart showing the distribution of differentially regulated enhancers among medulloblastoma enhancer classes. (c) K-means clustering of differentially regulated medulloblastoma enhancers (n=20,406). (d) Bar plot showing the proportion of enhancer-gene assignments to N enhancers. (e) Bar plot displaying the proportion of enhancer-gene assignments to N genes. (f) Bar plot summarizing the proportion of enhancer-gene assignments to N subgroups. (g) WNT (E2) and SHH (E3) subgroup-specific enhancers inferred to regulate *ALK*. (h,i) Scatter plots correlating sample-matched gene expression (RPKM, x-axis) of *ALK* with H3K27ac enrichment (log_2; y-axis) for the WNT-specific (E2) and SHH-specific (E3) enhancers shown in (g).

Since medulloblastoma subgroups were first described based on their transcriptional diversity[22], we sought to explore subgroup-specific enhancers potentially driving inter-subgroup heterogeneity in anticipation of gaining insight into the regulatory landscape responsible for subgroup identity. ANOVA was used to identify sets of enhancers differing according to known molecular subgroup, revealing 20,406 differentially regulated enhancers (26% of all inferred medulloblastoma enhancers; Figure 7.2a, b). The remaining 74% (n=58,110) displayed activity across all groups suggesting a general role in medulloblastoma or cerebellar identity (Figure 7.2a). K-means clustering of differentially regulated enhancers delineated six distinct medulloblastoma enhancer classes, including one for each subgroup (i.e. WNT, SHH, Group 3 and Group 4) as well as WNT-SHH and Group 3-Group 4 shared classes (Figure 2b, c). Group 3 and Group 4 subgroups are known to exhibit some degree of transcriptional similarity[23-25], consistent with the enhancer clustering results. In contrast, WNT and SHH subgroups tend to be mostly dissimilar from a transcriptional perspective, and thus the WNT-SHH set of shared enhancers identified here is unexpected but intriguing.

We next sought to assign enhancer elements to target genes, a process typically hindered by the fact that enhancers may regulate multiple genes, and that a majority of enhancer/promoter interactions occur at distances > 50kb[26]. To overcome challenges in enhancer/gene assignment, we leveraged sample-matched RNA-Seq gene expression data to identify enhancer/gene pairs contained in the same topologically associated domain (TAD[27]) that exhibit strong positive correlations between enhancer H3K27ac levels and mRNA expression (ρ > 0.6 and FDR <0.05). This approach assigned 8,775 enhancers (including 43% of all medulloblastoma differential enhancers) to at least one protein-coding target gene. The majority (44%) of inferred target genes were assigned to a single enhancer, but in many cases, several enhancers were predicted to converge on the regulation of a single gene (Figure 7.2d). Likewise, 73% of enhancers were assigned to only a single gene target and rarely was a given enhancer assigned to more than two candidate genes (Figure 7.2e). Compelling subgroup-related diversity with respect to inferred enhancer-target gene regulation was prevalent in our dataset (Figure 7.2f-i), with numerous genes exhibiting convergent regulation by distinct subgroup specific differential enhancer loci. For example, we identified alternative subgroup-specific enhancers predicted to regulate known oncogenes, including WNT-specific and SHH-specific enhancers inferred to target *ALK*, and WNT-specific

and Group 3-specific enhancers inferred to target *MYC* (Figure 7.2g-i). These data provide a rational, computationally robust approach for assigning medulloblastoma enhancers to their potential targets and underscore the apparent complexity inherent to enhancer-gene regulation across medulloblastoma subgroups.

Group 3 and Group 4 medulloblastoma remain the least well understood subgroups of the disease, despite collectively accounting for ~60% of all diagnoses[2,10]. Group 3 patients have a dismal prognosis, are frequently metastatic, and are restricted to infancy/childhood, whereas survival rates for Group 4 can be quite heterogeneous and diagnoses occur across all age groups. Molecularly, Group 3 and Group 4 are only crudely defined, with aberrant *MYC* amplification and over-expression characteristic of Group 3 but not Group 4, a signature feature discriminating these subgroups. In contrast to WNT patients (who almost universally survive current treatment regimens), and SHH patients (who represent rational candidates for SHH pathway inhibitors such as SMO inhibitors), novel treatment options remain scarce for Group 3 and Group 4 patients.

We hypothesized that enhancer-driven functional pathways distinguishing Group 3 from Group 4 medulloblastoma might better characterize the more aggressive nature of tumorigenesis in Group 3 and potentially provide novel therapeutic insights. We first validated that Group 3 or Group 4 differential enhancer target genes showed reciprocal patterns of H3K27ac enrichment. We ranked the top 1,000 enhancers (by H3K27ac signal) in either Group 3 or Group 4 by fold change in acetylation and found a strong leading edge enrichment of Group 3 and Group 4 specific target genes (Figure 7.3a). Functional pathway analysis performed on Group 3 and Group 4 enhancer-gene target assignments identified prominent neuronal gene sets enriched in both subgroups (Figure 7.3b, c), consistent with published transcriptional studies[23-25] and validating the computational approach implemented to assign enhancers to target genes. Neuronal development driven by transcriptional regulators dominated the Group 4 functional annotation, whereas Group 3 enhancer target genes prominently included thematic pathways associated with TGFβ signaling (Figure 7.3b,c). Group 3 specific enhancers included the TGFβ family type I and II membrane receptors (*ACVR2A* and *TGFBR1*) as target genes (Figure 7.3c), whereas enhancers regulating *SMAD5*, *TGFB1*, and *TGFB3* showed equivalent acetylation between Group 3 and Group 4. Overall components of the TGFβ signaling pathway showed a strong enrichment for enhancer

regulation in Group 3 (Figure 7.3d). Notably, we uncovered a ~450kb focal amplification at the *ACVR2A* locus in one of the Group 3 samples that encompassed both the gene and the upstream enhancer regions. These data, combined with our prior observations that TGFB receptor genes are recurrently amplified at low frequency in Group 3[4], implicate TGFβ signalling as an important oncogenic driver in Group 3 medulloblastoma.

The Group 3-specific presence of large enhancer clusters at TGFβ signalling pathway components prompted a consideration that super-enhancers (SEs), broad spatially co-localized enhancer domains[28-31], might play an essential role in establishing subgroup-specific identity. SEs are established by cell state-defining transcription factors and transcription factors at the termini of signalling pathways[29,31]. In multiple tumour types, SEs have been shown to drive oncogenes, genes required for maintenance of tumour cell identity, and genes associated with cell type-specific functions. As catalogues of these gene categories in medulloblastoma subgroups are poorly understood, we undertook a systematic mapping of SEs across all 28 medulloblastoma samples. SE maps revealed massive (>50kb) SE domains at the cerebellar-specific transcription factors *ZIC1* and *ZIC4*[32,33] (Figure 7.4b), and at 70% of a queried set of established medulloblastoma driver/signalling pathway genes, including *GLI2*, *MYC*, and *OTX2*[4].

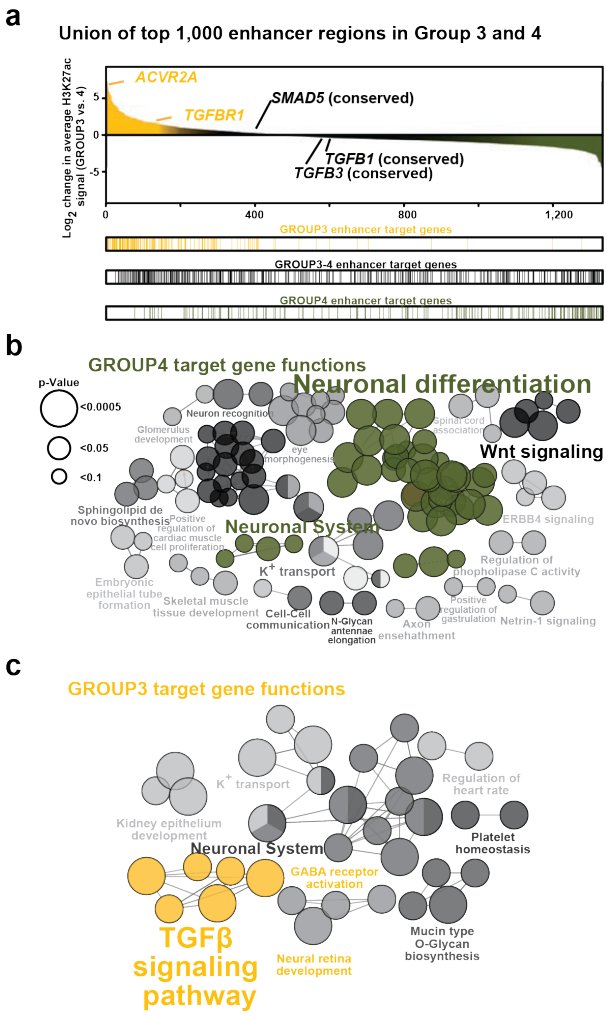


Figure 7.3: Functional characterization of enhancer-gene targets in medulloblastoma subgroups. (a) Waterfall plot discriminating the top 1,000 Group 3 and Group 4 subgroup-specific enhancers as defined by total H3K27ac signal. The distribution of assigned targets in Group 3, Group 4, and shared Group 3-4 targets are shown below the waterfall. (b,c) Functional annotation of target genes assigned to Group 3 and Group 4 subgroup-specific enhancers based on their significant overlap with gene sets annotated in Gene Ontology (GO Biological Process) and pathway databases (KEGG, Reactome). (d) Convergence of Group 3-specific enhancers on TGFβ pathway genes. Subgroup-specific enhancers are summarized as nodes according to their respective medulloblastoma enhancer class – Group 3, Group 4, and shared Group 3-4 – with edges representing individual enhancer-TGFβ pathway gene assignments.

To map SEs in a subgroup specific manner, we determined and ranked the average H3K27ac occupancy across samples of a subgroup at the union of all enhancer regions identified in samples of that subgroup. Subgroup SEs were identified from this meta H3K27ac signal using previously established methods[31], resulting in ~3,000 distinct SE containing loci with ~600-1,100 SEs identified per subgroup (Figure 7.4a). Compared to typical enhancers, subgroup SEs showed higher occupancy of BRD4 and greater enhancer signal dynamic range between subgroups. Targets of differential enhancers contained within SEs (i.e. SE target genes) included numerous exemplar medulloblastoma signature genes as well as novel candidates including *NKD1/NKD2* (WNT subgroup), *PCNT* (SHH subgroup), *HLX* (Group 3), and *SNCAIP* (Group 4) (Figure 7.4d-f). Medulloblastoma SEs were inferred to regulate known cancer genes, including *ALK* in WNT, *SMO* and *NTRK3* in SHH, *LMO1*, *LMO2*, and *MYC* in Group 3, and *ETV4* and *PAX5* in Group 4, among others. Furthermore, several actionable, SE-regulated genes were revealed in our analysis including several kinases (NTRK1, SGK1) and chromatin modifying enzymes (PNMT, HDAC4), which have available small molecule inhibitors.

Rank transformation of subgroup SE elements across all samples enabled a systematic identification of SEs displaying either conserved SE activity across samples, or highly subgroup specific patterns of activity (Figure 7.4b-f). Importantly, subgroup-specific SEs were predicted to regulate a large fraction of established medulloblastoma signature genes (32%), suggesting that SEs might play an important role in driving subgroup-specific identity. Consistent with this hypothesis, patterns of SEs across all primary medulloblastoma samples were sufficient to recapitulate transcriptional subgroupings in an unbiased hierarchical clustering using no prior knowledge of subgroup status (Figure 7.4a). As shown with all enhancer elements, SEs from established Group 3 medulloblastoma cell lines clustered with one another, but failed to show similarity to primary samples from any subgroup.

Among subgroup-specific SE target genes, we observed an enrichment of transcription factors (TFs) involved in neuronal development (p-Value ~ 0.0001, Fisher's exact test; Extended Data Figure 6). XX% of these SEs are previously uncharacterized suggesting a tissue or cell type-specific role. Overall, subgroup-specific TFs showed similar patterns of expression, enhancer motif enrichment, and strong overlap of target gene pathways. TFs were also enriched in subgroup-specific SE targets as compared to subgroup-specific non-SE targets (p-Value ~ 0.002, Fisher's exact test), consistent with prior observations

in other cancers that SEs regulate key TFs required for tumour identity and maintenance[17,19,29]. Given prior evidence in embryonic stem cells that pluripotency master regulator TFs (OCT4, SOX2, NANOG) are driven by SEs and themselves bind to and establish SEs[31], we hypothesized that a *reverse* analysis of SEs in medulloblastoma might enable a *de novo* reconstruction of tumour identity-defining TFs and their associated regulatory circuitry, thereby providing novel insights into medulloblastoma origins.

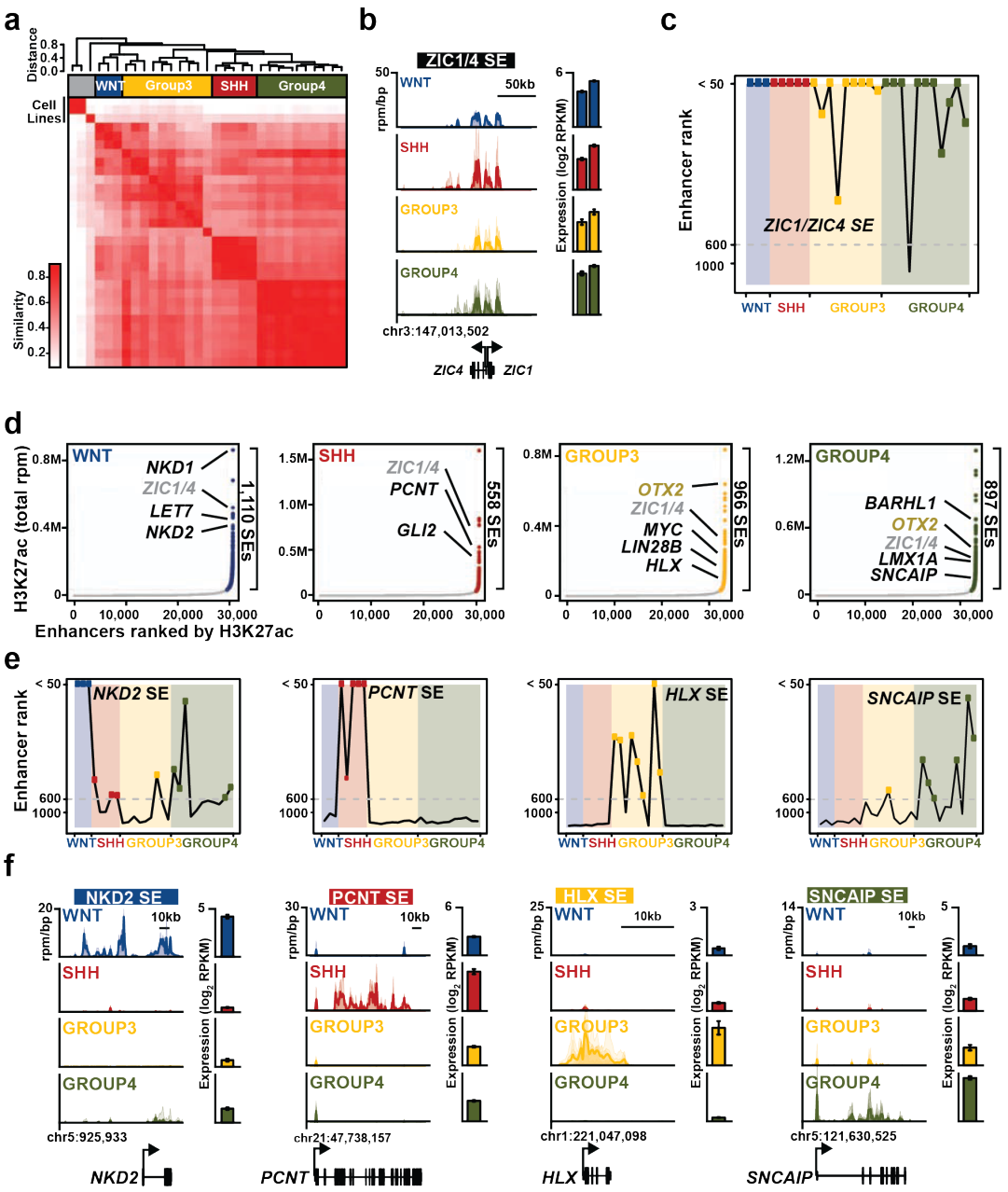


Figure 7.4: Medulloblastoma super-enhancers define subgroup-specific identity. (a) Unsupervised hierarchical clustering of primary medulloblastomas and cell lines using H3K27ac signal calculated at all SEs identified in each individual sample. (b) Meta tracks of H3K27ac ChIP-seq signal for the conserved *ZIC1/ZIC4* SE locus. Expression (mean RPKM) for both *ZIC4* (left) and *ZIC1* (right) is displayed as bar graphs to the right of each H3K27ac track. (c) Line plot showing the enhancer rank for the conserved *ZIC1/ZIC4* SE locus across all samples according to subgroup. (d) Ranked plots of enhancers defined across composite H3K27ac landscapes of WNT, SHH, Group 3, and Group 4 medulloblastomas. Enhancers are ranked by increasing group average H3K27ac signal (rpm). The cut-off discriminating typical enhancers (TEs) from super-enhancers (SEs) is shown as a dashed line. Select genes associated with SEs in each subgroup are highlighted and shaded according to enhancer class specificity. (e) Line plots showing the enhancer rank for candidate SE loci across all samples according to subgroup.

Figure 7.4 (Continued)

Examples of subgroup-specific (WNT=*NKD2*, SHH=*PCNT*, Group 3=*HLX*, Group 4=*SNCAIP*) SEs are shown. (f) Meta tracks of H3K27ac ChIP-Seq signal across medulloblastoma subgroups for the loci described in (e). The y-axis shows ChIP-Seq signal (rpm/bp) for each individual sample (shaded regions) with the average signal across the group shown in a line. The x-axis depicts genomic position with SE boundaries demarcated as rectangles. Bar graphs shown to the right of each H3K27ac track summarize the expression (mean RPKM) of the relevant candidate genes as determined by RNA-seq.

Pursuant to this idea, we proposed a set of criteria for TF inclusion into the core regulatory circuitry of medulloblastoma. Specifically, (1) core regulatory circuitry TFs are SE-regulated and (2) the TFs themselves bind to SEs of one another (Figure 7.5a). For each SE regulated TF, these criteria can be quantified through a measurement of the *in* and *out* degree of regulation, whereby the *in degree* represents the total number of SE regulated TFs that bind to a TF's SE, and the *out degree* represents the total number of other TF SEs bound by a given TF (Figure 7.5a). Using these criteria in the poorly characterized Group 4 medulloblastoma, we observe interconnected binding at the SEs of three neuronal TFs, *LMX1A*, *LHX2*, and *EOMES* (Figure 7.5b). Inspection of their respective gene loci revealed large SEs containing clustered binding sites for these factors present only in Group 3 and Group 4 (Figure 7.5b,c). Additionally, Group 4-specific enhancers for *LMX1A, LHX2,* and *EOMES* binding sites linked those TFs with Group 4-specific target genes. Extending regulatory circuitry reconstruction across all SE associated TFs in medulloblastoma, we identified regulatory cliques of TFs with similar patterns of *in/out* degree, strong interconnectivity via motif binding, and higher likelihoods of pairwise protein/protein interaction and motif co-occurrence at enhancers (Figure 7.5d). This reconstruction creates for the first time a candidate core regulatory circuitry in each subgroup, and implicates specific sets of TFs in establishing medulloblastoma subgroup identity (Figure 7.5d).

Cellular origins for WNT and SHH medulloblastomas have been experimentally established using mouse models genetically engineered to aberrantly activate the WNT and SHH signaling pathways, respectively, in distinct cerebellar stem/progenitor cells during development[6-9]. The origins of Group 3 and Group 4 medulloblastoma, however, are unknown and yet essential to define as these tumours account for ~60% of all diagnoses, lack targeted therapies, and are frequently associated with a poor clinical outcome secondary to current standard of care[2].

Cell identity is most essentially defined by the activity of master regulator TFs. In reprograming and trans-differentiation studies, the activity of these TFs (e.g. OCT4, SOX2, and NANOG in embryonic stem cells, MYOD in myoblasts, Pu.1 in B cells) is sufficient to induce transitions between cell states[34,35]. As such, we hypothesized that the regulatory SE regions governing endogenous expression of candidate master TFs and embedded in the core regulatory circuitry of medulloblastoma subgroups might inform the cellular origins of the disease via their cell type-specific activity. Examination of the expression of Lmx1a,

Eomes, and Lhx2 in the developing mouse cerebellum (e13.5) using the Mouse Allen Brain Atlas database showed spatiotemporal patterns of restricted expression in the nuclear transitory zone (NTZ) located below the pial surface at the rostral end of the cerebellar plate and serving as an assembly point for immature deep cerebellar nuclei (DCN). DCN residing in the NTZ at this time point are predominantly glutamatergic projection neurons that originate from earlier progenitors of the rhombic lip, a transient structure producing progenitors with distinct cellular fates, including DCN and cerebellar granule neurons[36] (Figure 7.5e,g). Immunohistochemical staining for Lmx1a and Eomes at the same time point (e13.5) recapitulated these findings (Figure 7.5f). To validate the spatiotemporal expression pattern observed for these predominantly Group 4-specific TFs, we cloned constituent regions of the Lmx1a SE into a *LacZ* reporter construct, introduced the reporter into the e11.5 developing mouse hindbrain by *ex utero* electroporation, and assayed enhancer activity via x-gal staining at 48 hours post-transfection (Figure 7.5h). X-gal staining revealed spatially restricted activity of the Lmx1a SE reporter in the developing cerebellum (Figure 7.5i). These findings validate the specific activity of the Lmx1a SE observed in Group 4 medulloblastoma and implicate precursors of glutamatergic DCN as potential cells-of-origin for this subgroup. Finally, these findings establish SE core regulatory circuitry as a novel method to infer cell of origin for poorly classified primary tumours.

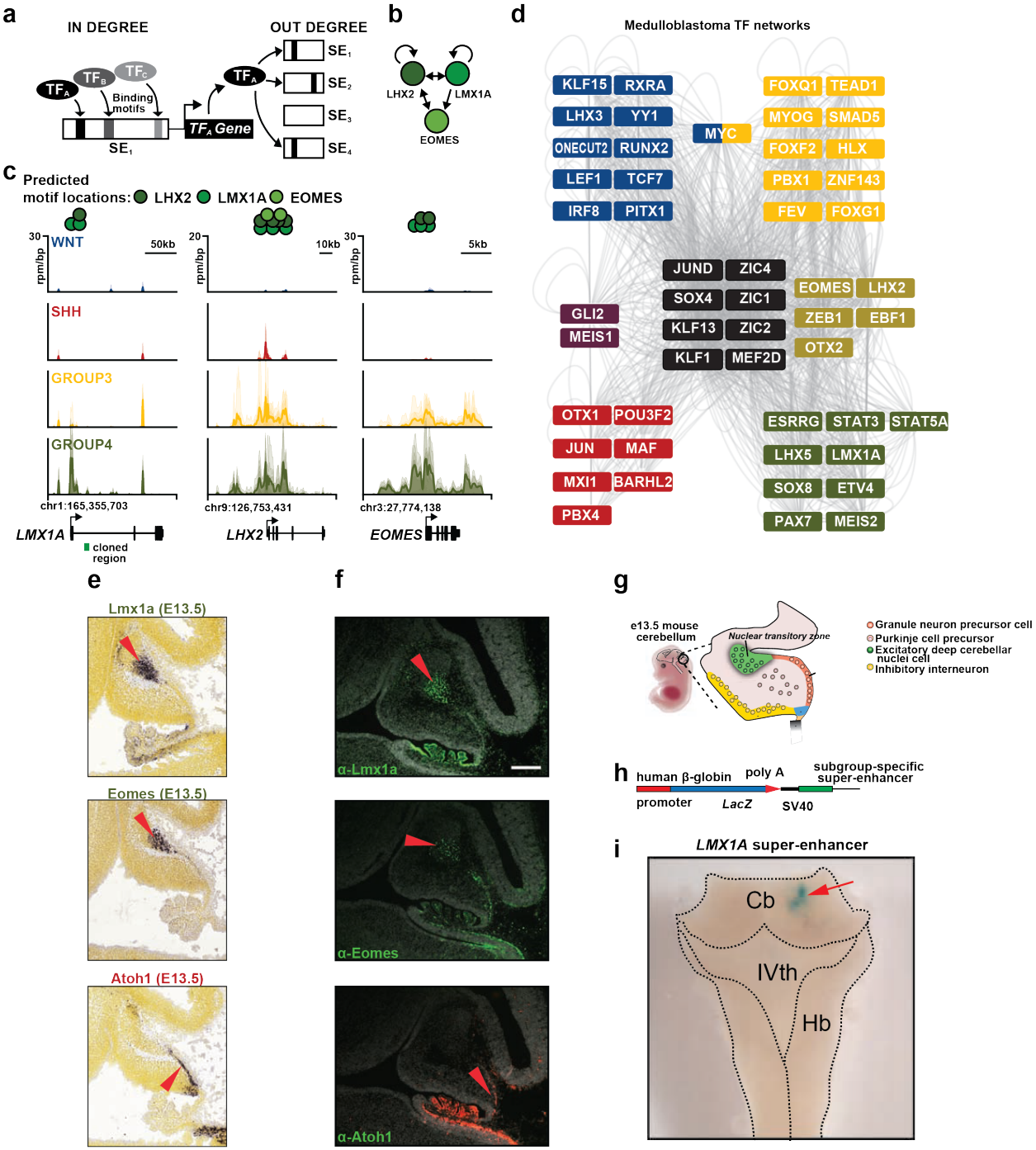


Figure 7.5: Super-enhancers define medulloblastoma regulatory circuitry and identify putative cellular origins. (a) Methodology for inferring transcriptional regulatory circuitry driven by medulloblastoma SEs. (b) Subset of the Group 4-specific transcriptional network predicted to be driven by *LMX1A*, *LHX2*, and *EOMES*.

Figure 7.5 (Continued)

(c) H3K27ac ChIP-seq meta tracks for the SE-regulated transcription factors *LMX1A*, *LHX2*, and *EOMES*. Locations of enriched motifs for each of the respective transcription factors are highlighted at the top of the panel. (d) Subgroup-specific TF circuitry. Nodes are TFs associated with a SE in a subgroup-specific context. Edges indicate co-regulating TFs as defined by enrichment of TF binding motifs in respective regulatory regions. (e) In situ hybridization data showing highly localized expression of Lmx1a (upper panel) and Eomes (middle panel) in the embryonic cerebellum at e13.5. Red arrows indicate highly localized expression of both TFs in deep cerebellar nuclei (DCN) of the nuclear transitory zone (NTZ). Atoh1 expression (lower panel) is shown at the same developmental time-point to serve as a marker of the external granule layer (EGL). (f) Immunofluorescence microscopy for Lmx1a (upper panel) and Eomes (middle panel) performed on sagittal sections of the e13.5 murine cerebellum. Red arrows indicate highly localized expression of both TFs in DCN of the NTZ. Atoh1 staining (lower panel) is shown at the same developmental time-point to serve as a marker of the EGL. (g) Atlas (sagittal) of the e13.5 murine cerebellum, highlighting the main cell types and compartments contributing to the cerebellar anlage at this time point in development. (h) Strategy for validating medulloblastoma subgroup-specific enhancers *in vivo*. (g) X-gal staining of an embryonic cerebellum (e13.5) transfected with the reporter construct shown in (h) containing constituents of the Group 4-specific *Lmx1a* SE. Red arrow indicates cells positive for *in vivo* enhancer activity driving LacZ expression.

Discussion

Recent, large-scale epigenome classification studies have effectively integrated ChIP-Seq data for core histone modifications, DNA methylation, transcriptome, and other complementary data to comprehensively ascribe function to the mammalian epigenome across cell types[37,38]. Vast catalogs of active enhancers, including SE's, have been generated for 100's of different cell types and tissues, including both non-neoplastic and disease-derived entities[29]. In most cases, largely due to logistical and technical limitations, these data have been generated using immortalized, cultured cell line material. Herein, we describe the medulloblastoma enhancer landscape across a series of 28 fresh-frozen, treatment-naïve tissue samples and 3 cultured cell lines, demonstrating dramatic divergence between primary tumour and tumour cell line material. Moreover, by studying an unprecedented cohort from a single cancer entity, we uncover considerable regulatory element heterogeneity between subgroups of the disease that would be overlooked and unsubstantiated in series limited to just a few cases.

Clinically relevant medulloblastoma subgroups are principally defined based on their underlying transcriptional profiles. Differentially regulated medulloblastoma enhancers and SEs are likewise capable of recapitulating these subgroups. Biological themes and signalling networks extracted from transcriptional data have served as the primary source of annotation for medulloblastoma subgroups, with WNT and SHH subgroups characterized by activation of their respective signalling pathways, and Group 3 and Group 4 recognized for their GABAergic and glutamatergic expression phenotypes, respectively. Although these data provide a functional and phenotypic annotation of medulloblastoma, they fail to articulate the cell of origin and developmental identity of individual subgroups. Using a reverse analysis of the medulloblastoma chromatin landscape starting at the level of differentially regulated enhancers and SEs, we have reconstructed the core regulatory circuitry inherent to medulloblastoma subgroups, and inferred master transcriptional regulators responsible for subgroup-specific transcriptional programs. The majority of these master regulator TFs were not previously implicated in medulloblastoma developmental biology, nor were they visible amongst transcriptionally-derived gene sets dominated by aberrant signalling and overwhelming phenotypic signatures. Through tracing the spatiotemporal activity of a subset of Group 4 master TFs, these studies identified DCN of the cerebellar NTZ, or plausibly their earlier precursors originating from the rhombic lip, as putative cells-of-origin for this large subgroup of

patients. Together these approaches establish a framework for the inference of tumour cell of origin through enhancer core regulatory circuitry mapping.

In medulloblastoma, knowledge of tumour cell of origin has broad implications for the understanding and treatment of the disease. Numerous cancers, especially those of the immune compartment are treated through targeting of the lineage (e.g. anti-B cell therapies). As medulloblastoma arises from cell populations that normally exist ephemerally during development, targeting the aberrant persistence of tumour cells from these lineages may represent a novel therapeutic strategy. Consistent with this approach, we note that many of the subgroup specific master TFs identified in our core regulatory circuitry show minimal enhancer activity in bulk adult cerebellum. Additionally, we demonstrate that core circuitry TF-regulating SE elements can drive spatiotemporal patterns of expression when inserted into reporter constructs. As such, use of these elements in *Cre*-inducible knockout/over-expression systems may accelerate the development of more faithful mouse models recapitulating medulloblastoma subgroups. Finally, elucidation of core regulatory circuitry implicates upstream signalling dependent regulators of these TFs, their co-activators, and their downstream effectors as potential subgroup-specific therapeutic targets. These insights demonstrate the critical importance of epigenetic analyses of primary tumours as opposed to cell line model systems and highlight the broad utility of core regulatory circuitry inference especially in poorly characterized and clinically diverse tumours.

Methods

All patient material included in this study was collected after receiving informed consent from the patients and their families. Medulloblastoma samples were collected at first resection, before adjuvant chemotherapy or radiotherapy. Subgroup assignments were made using the Illumina 450K DNA methylation array as described[39]. Chromatin extraction and library preparation for ChIP-seq of H3K27ac and BRD4 were performed at ActiveMotif (Carlsbad, CA) using proprietary protocols. Alignment and filtering of ChIP-seq data was performed as described[5]. H3K27ac enhancer peaks were called using MACS[21]. H3K27ac peaks were classified as being subgroup-specific or as common enhancers by first calculating H3K27ac enrichment on the merged peaks followed by ANOVA and k-means clustering. Target gene identification of enhancers was performed by correlating H3K27ac enrichment at the

enhancers with expression levels of genes located in the same topologically associated domain[27] as the enhancers. Candidate gene(s) showing the highest correlation were selected as the putative target(s) of the enhancer. Gene Ontology/Pathway analysis of enhancer-gene targets was performed using the ClueGO plugin for cytoscape[40]. Super-enhancers (SEs) were called using ROSE2[31] and subgroup specificity of super-enhancers were assigned via ranking average H3K27ac signal across the subgroups. Medulloblastoma core-transcriptional circuitry analysis was performed by calculating inward and outward degree regulation of SE-regulated transcription factors. Reporter assays for validating enhancers *in vivo* were performed by *ex utero* electroporation of reporter constructs into the hindbrain of murine CD1 embryos (e11.5). Enhancer activity was measured by X-gal staining of transfected cerebella. Endogenous expression of candidate transcription factors was determined by querying the Allen Brain Atlas Data Portal (http://developingmouse.brain-map.org) or by immunohistochemistry performed on embryonic cerebella.

Chapter 8

Modeling of Cellular Networks to Enable Systems Pharmacology

Conclusion

In this dissertation, two high-quality small molecule probes were used to interrogate the ole of their protein targets (DOT1L and BETs) in mammalian gene regulation. Both targets also have critical roles in cancer biology, and their respective chemical inhibitors will serve as templates for further therapeutic development. Moreover, insights gained through the use of these probes uncovered a rationale to consider enhancer-linked systems-level models of transcription factor networks. These networks are being built with enhancer landscapes of many healthy and diseased tissues, and have already uncovered developmental transcriptional signatures that define a putative cell-of-origin for Group 4 medulloblastoma.

In Chapter 2, a structural biology effort was initiated concurrently with medicinal chemistry for the DOT1L methyltransferase. A previously reported small molecule inhibitor proved difficult to crystalize, so the newly synthesized near chemical derivatives were critical for obtaining a first useful structure. The molecular binding pose revealed regions of the molecule amenable to modification towards improving potency and pharmacokinetics. Moreover, a dynamic loop in the enzyme was identified that may explain the compound's extraordinary potency and selectivity for DOT1L. Alternative biophysical methods like HX-MS were utilized to confirm this observation.

These efforts were critical for Chapter 3, where structural information of the DOT1L inhibitor-binding pose allowed the design of ligand-affinity probes for use in biochemical assays. The suite of assays presented in this chapter will be critical for continued DOT1L therapeutic development, both for supporting iterative medicinal chemistry on the Epizyme scaffold, as well as for high-throughput screening efforts to find scaffolds with better pharmacokinetic properties and *in vivo* stability.

The remaining chapters in the thesis built off of observations made about JQ1, an inhibitor of the BET family of bromodomains. JQ1 has exquisite cell-type specificity in its phenotype resulting from it's ability to disrupt the function of super enhancers, large cis-domains driving lineage gene expression programs.

Chapter 5 considered a long-standing problem in immunology and oncogenesis – defining genomic features that cause off-target activity of the AID antibody diversification enzyme. Using an engineered system to generate in high throughput large numbers of AID-mediated break sites, AID hotspots were

identified genome wide. These hotspots almost all occurred within super enhancers, more specifically super enhancers that were intragenic. Measurements of nascent RNA generation allowed even more precise definition of these hotspots, as they tend to occur in intragenic super enhancer regions also displaying a pattern of convergent transcription. The mechanism by which these genomic features contribute to AID-dependent DNA cutting is currently under investigation.

Super enhancers were also used as a lens in Chapter 6, here to better predict the transcription factor interactions that maintain the gene expression program of the cell. Enhancer topology was used to predict sites of TF binding. The underlying DNA sequence of these regions was then used as a template to search for known TF operator sequences, allowing the prediction of which TFs can regulate each super enhancer. From this information, the enhancer-connected transcriptional network can be constructed for a given cell type. These algorithms were used in medulloblastoma in Chapter 7. This led to the identification of medulloblastoma subgroup-specific transcriptional regulators. A novel set of regulators was predicted for Group 4 medulloblastoma, pointing to a potential cell of origin for this cancer. Current efforts are underway to further validate this prediction and use these regulatory elements to breed the first mouse model of group 4 medulloblastoma.

Systems pharmacology

As outlined in Chapter 6, there is longstanding interest in understanding the network architecture of complex biological processes. *Systems pharmacology* is an emerging field that attempts to unite complex, quantitative biological models with tools from chemical biology and pharmacology to innovate therapeutic discovery. Many of the ideas presented here mirror the goals of systems pharmacology: using small molecule probes to understand the underlying biological network of a disease and to predict how perturbations to the system may provide a therapeutic benefit.

Despite the new name, scientists and physicians have been using systems approaches to study the biological activity of chemical substances for many years. In the late 19th century, pharmacologists including John Langley and Paul Ehrlich isolated intact organs to understand the nature of pharmacologically active substances. These experiments eventually led to the proposition of receptor

theory, further elaborated by Raymond Ahlquist in 1948 describing the distinguishing characteristics between α- and β- adrenoceptors[1,2]. These studies allowed definition of the effector molecules in these systems (receptors) and led to the development of clinical-grade drugs, possibly because the complexity of the downstream receptor biology remained intact, allowing the study of interesting phenotypes arising from these molecules. Additionally, the molecules were tested in a setting that selected for compounds that had acceptable pharmacokinetic properties to show any effect in the experimentation, removing the need to optimize these features later in the development process.

With the advent of methodologies to isolate receptors for study outside the context of the tissue, pharmacology shifted its focus towards target-based methods. This began with radioligand displacement assays, and as molecular biology and automation technology matured through the 1980s and 1990s, the main approach to finding new pharmacological agents focused on screening collections of molecules against isolated, recombinant protein targets. This continued after the sequencing of the human genome when the entire collection of potential drug targets was revealed. These targeted approaches brought much success, with drugs like Gleevec and Erlotinib providing extraordinary benefit cancer. However, there are drawbacks to these methods. Resistance to targeted therapy in cancer occurs rapidly and nearly universally. Additionally, rates of success for currently pharmacological development are staggeringly low, often due to poor target selection or unexpected toxicity[3].

While drug discovery programs focused on targeted therapy, systems biologists continued building sophisticated descriptions of various biological networks and their properties. Great progress has been made in understand metabolic networks, transcriptional networks and functional networks, among others. In simple organisms, complete network descriptions have been solved, leading to the identification of patterns or “network motifs” that are often preferred in natural biological networks[4]. Studying the motif logic embedded in these networks reveals phenotypic consequences of motifs, with different motifs being preferred in different biological systems. While much of this work is done in simple gene regulatory network model systems, network motifs in the context of human biology can be identified that allow some understanding of the logic of human cellular networks.

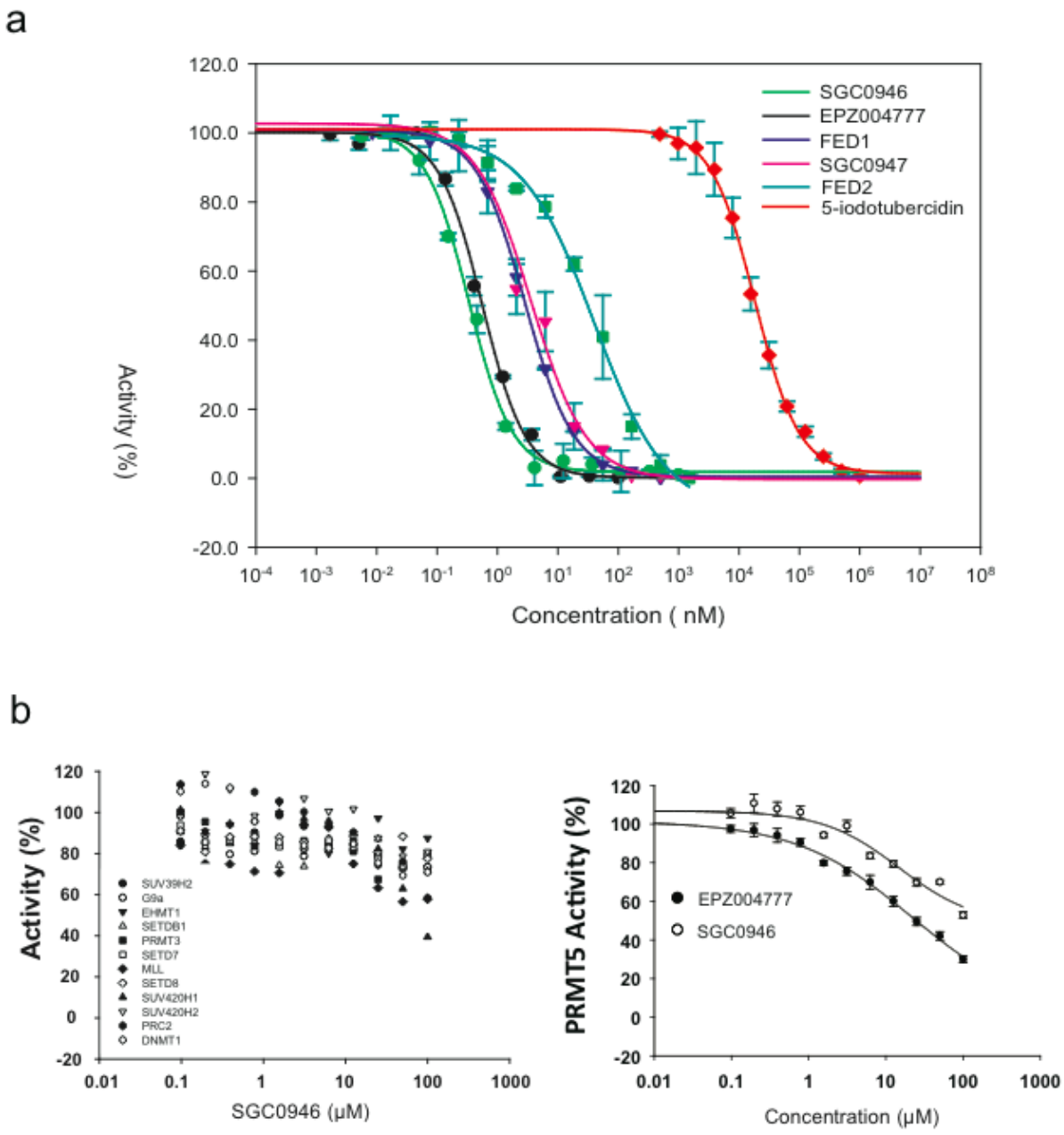
Beyond these approaches, many other avenues are being taken in modern systems biology to gain greater understanding of biological networks. Truly remarkable work describing the first computational

simulation of a whole cell in 2012 provided proof of concept for the integration of different regulatory and biochemical modules to model behavior in a predictive way[5]. Open source projects like Biopax are attempting to formalize and standardize the way scientists communicative about networks and hosting databases of published pathways and interaction data[6]. Algorithms that integrate large datasets in epigenomics or gene expression are providing tissue-specific and disease-specific networks with orthogonal measurements by which the research community can generate hypotheses for therapeutic approaches[7,8].

The emergence of facile genomic, proteomic and other high-dimensional biological measurements provides a unique opportunity in the coming years to build upon classic pharmacology by incorporating systems approaches during therapeutic development. As a field, there is opportunity for sophisticated adaption of these technologies (both experimental and computational) with chemical approaches to study the action of bioactive molecules in complex, relevant biological models of disease. Measuring drug action in these models might allow improvements in diagnostic capabilities, modeling of disease in cellular, organ and patient-level networks, and prediction therapeutic efficacy while understanding toxicity in different organs thorough an organism. Beyond these goals, the field will aim to incorporate tools traditionally found in industrial drug discovery including pharmacokinetic and pharmacodynamics modeling in an attempt to understand drug action at the level of the organism. Should systems pharmacology realize some of these goals, they will certainly have a positive impact on human health.

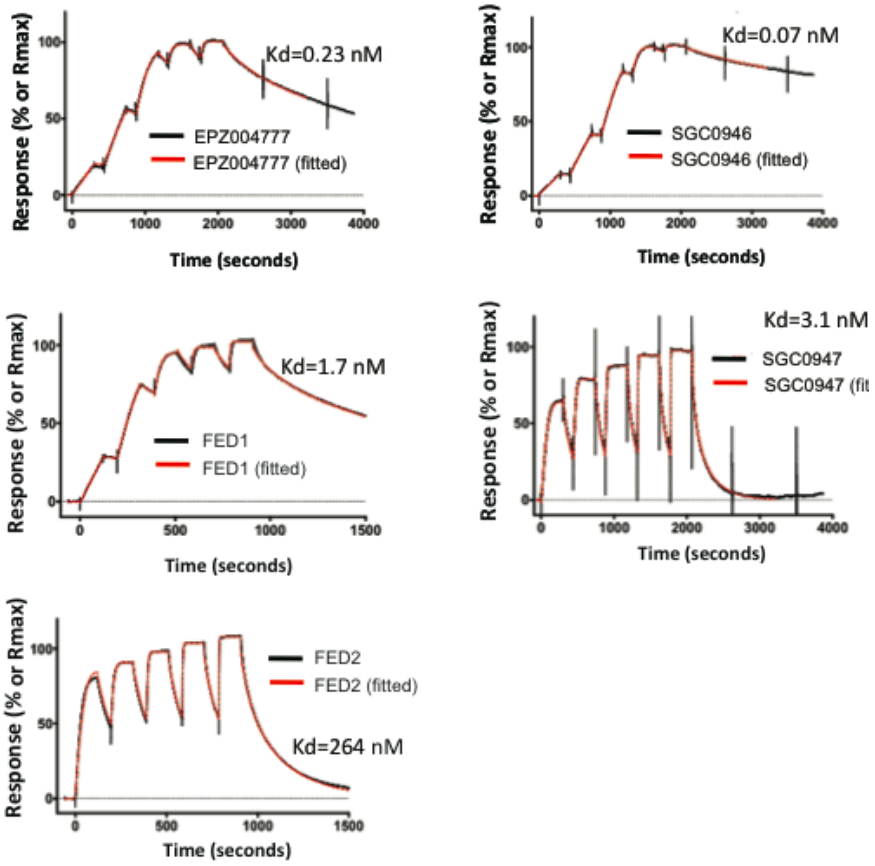
Appendix A

Supplementary Materials for Chapter 2



Supplementary Figure A.1: **Effect of inhibitors on DOT1L catalytic activity, and selectivity profiles of EPZ004777 and SGC0946**. (a) A radioactivity-based assay is used to follow the dose response effect of chemical inhibitors on DOT1L activity. (b) EPZ004777 and SGC0946 are inactive against a panel of 12 protein methyltransferases and DNMT1.

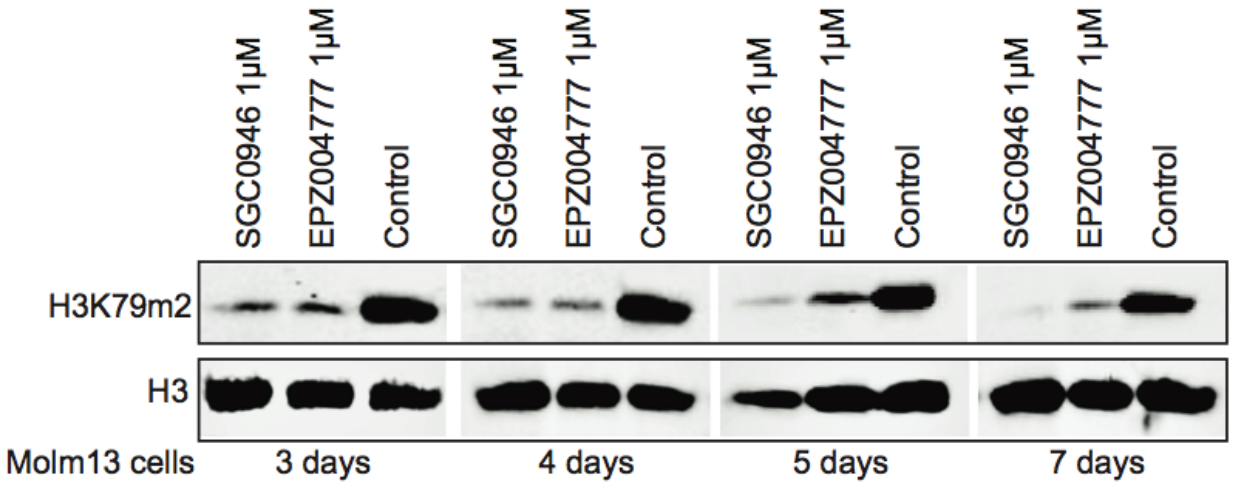
a



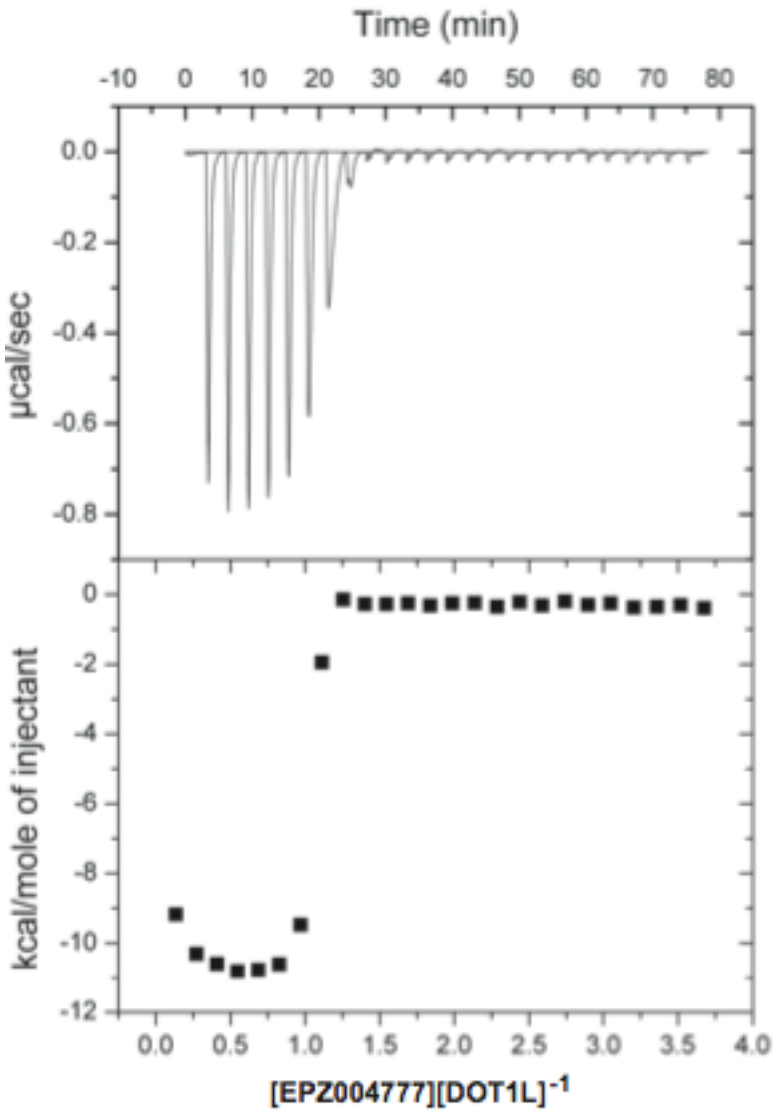
b

	EPZ004777	SGC0946	FED1	SGC0947	FED2
KD (M)	2.31E-10	7.31E-11	1.76E-09	3.19E-09	2.64E-08
ka (1/Ms)	3.25E+06	2.40E+06	1.65E+06	1.31E+07	4.53E+05
SE (ka)	7.30E+03	4.40E+03	1.20E+04	2.20E+05	2.20E+03
kd (1/s)	7.52E-04	1.76E-04	0.002901	4.17E-02	0.01194
SE (kd)	1.30E-06	2.00E-07	1.90E-05	7.20E-04	5.60E-05
Conc. Range Tested	20nM - 1.25nM	20nM - 1.25nM	200nM - 12.5nM	100nM - 6.25nM	2uM - 125nM
Fitting Model	1:1 kinetics	1:1 kinetics	1:1 kinetics	1:1 kinetics	1:1 kinetics
Stoichiometry	1	1	1	1	1

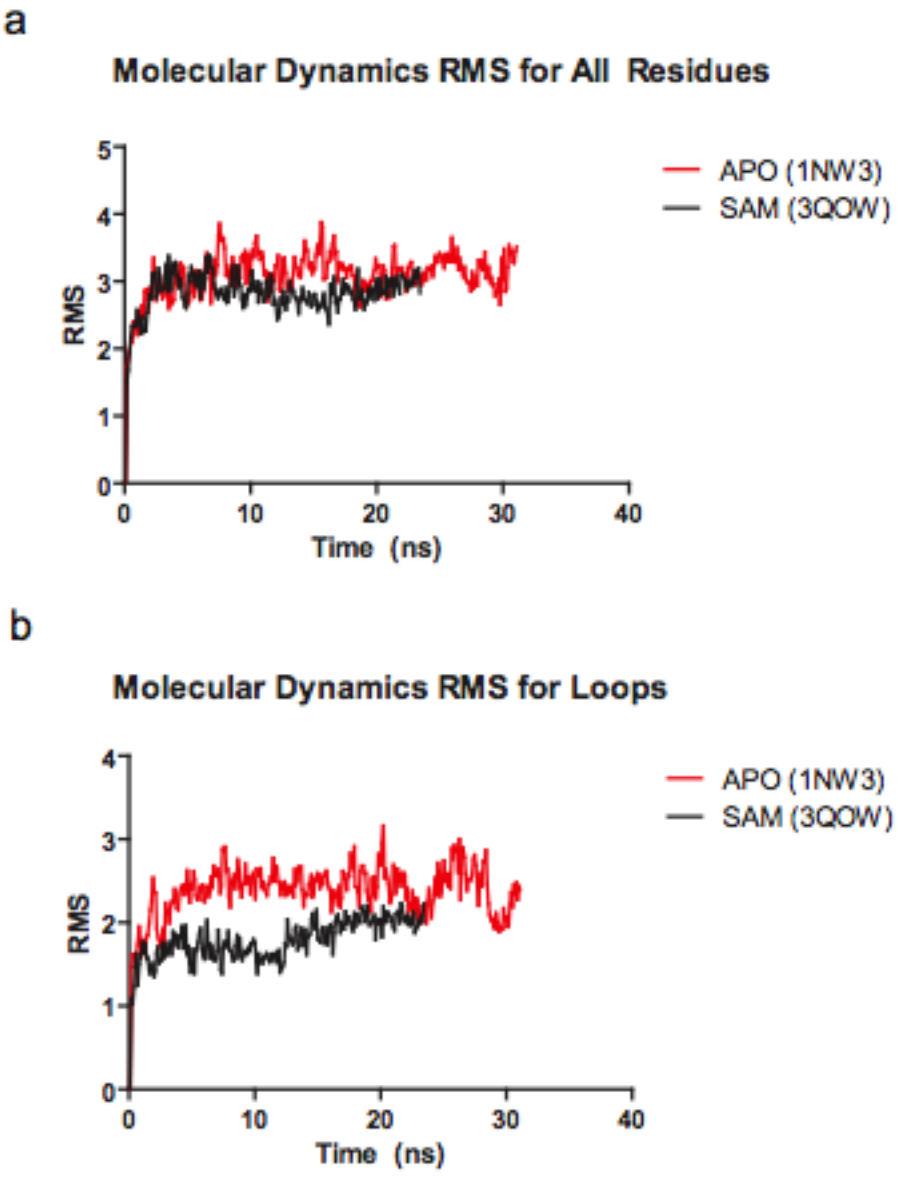
Supplementary Figure A.2: SPR on EPZ004777. (a) Biacore SPR sensorgrams of the tested compounds from single cycle kinetics runs with 5 concentrations. The fitted curves (in red) are overlaid on the experimental data (in black). The Raw experimental data was normalized to the calculated RMax for each compound. SPR indicates dissociation constants (K_D) of 0.23 nM, 0.07 nM, 1.7 nM, 3.1 nM and 264 nM for EPZ004777, SGC0946, FED1, SGC0947 and FED2 respectively. The DOT1L binding stoichiometry was 1:1 for all inhibitors. Five concentrations of each compound were tested in two-fold serial dilution series within the range listed. All compounds were fitted using the on and off rates, which were then used to calculate the K_D. The stoichiometry was determined by reference to the binding levels of the 1:1 binding controls, SAM and SAH. (b) Summary of DOT1L SPR data.



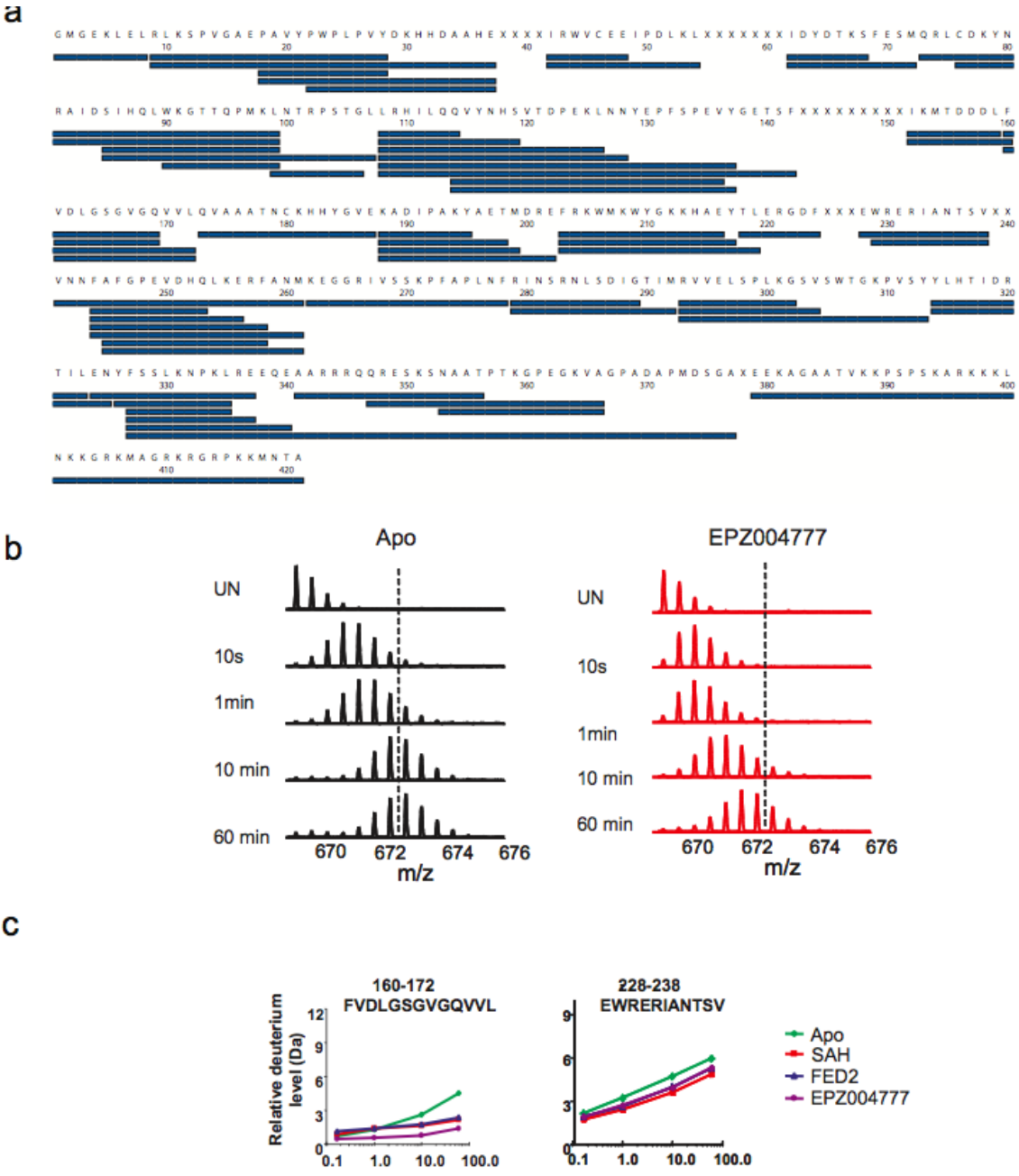
Supplementary Figure A.3: DOT1L inhibitors decrease H3K79me2 in cultured cells. Comparative immunoblot of EPZ004777 and SGC0946 in Molm13 MLL cells, demonstrating reduced H3K79me2.



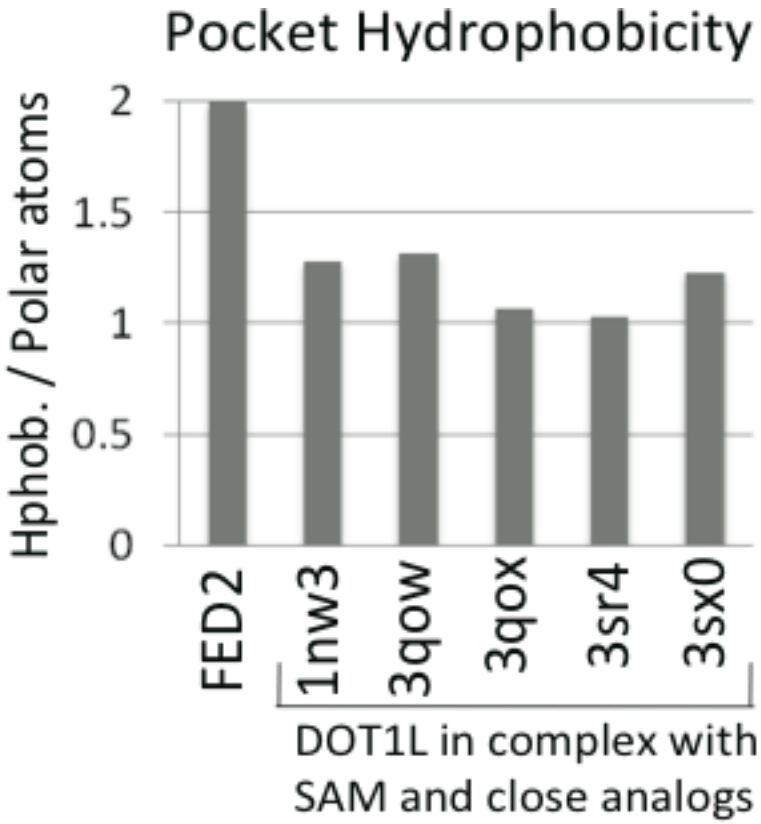
Supplementary Figure A.4: Isothermal titration calorimetry on DOT1L inhibitors. Isothermal titration calorimetry indicates a 1:1 binding stoichiometry between EPZ004777 and DOT1L in solution.



Supplementary Figure A.5: **Molecular dynamics.** RMS values for (a) all residues and (b) loop regions of apo DOT1L and SAM-bound DOT1L.



Supplementary Figure A.6: **Conformational flexibility of DOT1L as determined by HX MS**. (**a**) Coverage map of all the peptic peptides in DOT1L. 98% total protein coverage was obtained under the experimental conditions described. Not all these peptides were followed with HX MS. Amino acids that were not covered by digestion are represented with an X. (**b**) Representative mass spectra of a peptide (188-199) showing that upon drug binding the isotopic distribution changes showing less deuterium incorporation due to protection by the drug. (**c**) HX MS Relative Difference Plots for two peptides (as labeled) from DOT1L in the presence of EPZ004777, FED2 and SAH. The differences in deuterium incorporation between the bound protein and the native protein are plotted for deuterium exchange time points of 10 s (blue bars), 1 min (red bars), 10 min (green bars), and 1 hr (purple bars). The data shown are for peptides common to all three experiments, ordered from N to C terminus (top to bottom).



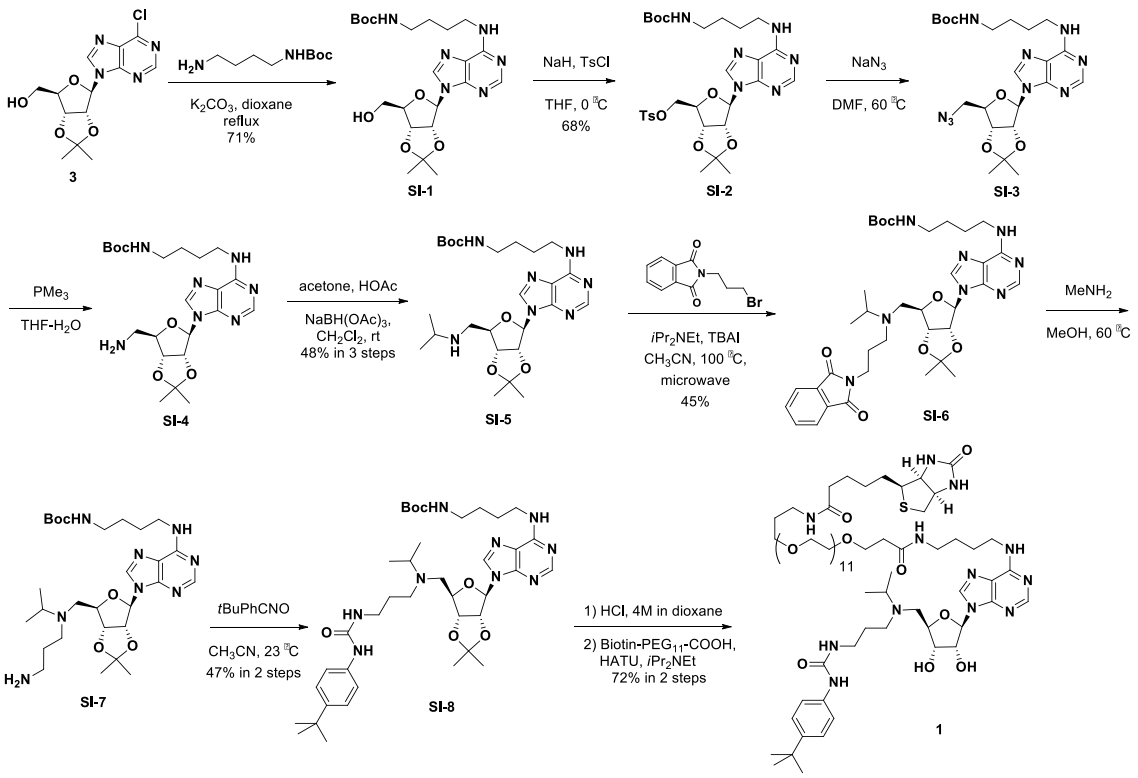
Supplementary Figure A.7: **DOT1L hydrophobicity**. The enclosed portion of the cofactor site is more hydrophobic in complex with FED2 than with SAM and close analogs: SAM (PDB codes 1NW3 and 3QOW), SAH (PDB code 3QOX), a methylated SAH analog (PDB codes 3SR4), Bromo-deaza-SAH (PDB code 3SX0).

Appendix B

Supplementary Materials for Chapter 3

Synthesis of probes and inhibitors:

Reactions were run as described in the individual procedures using standard double manifold and syringe techniques; glassware was dried by baking in an oven at 130 °C for 12h prior to use. Solvents for reactions were purchased anhydrous from Sigma-Aldrich and used as received; the only exception being EtOH, which was stored over 4 Å molecular sieves. HPLC grade solvents were used for aqueous work ups and chromatography. Reagents were used as received. Reactions were monitored by thin-layer chromatography using EMD silica gel 60 F254 (250-micron) glass-backed plates (visualized by UV fluorescence quenching and staining with KMnO4) and by LC-MS using a Waters Aquity BEH C18 2 x 50 mm 1.7 μm particle column (50 °C) eluting at 1 mL/min with H2O/acetonitrile [0.2% v/v added formic acid or concentrated NH_4OH(aq) solution; 95:5(0min)→1:99(3.60min)→ 1:99(4.00min)] using alternating positive/negative electrospray ionization (125-1000 amu) and UV detection (210-350 nm). Flash column chromatography was carried out using Merck grade 9385 silica gel 60 Å pore size (230-400 mesh). Melting points were obtained using a capillary melting point apparatus and are uncorrected. 1H NMR spectra were recorded at 400 MHz on a Bruker spectrometer and are reported in ppm using the residual solvent signal (dimethylsulfoxide-d6 = 2.50 ppm; chloroform-d = 7.27 ppm; methanol-d4 = 3.31 ppm; dichloromethane-d2 = 5.32 ppm) as an internal standard. Data are reported as: {(δ shift), [(s = singlet, d = doublet, dd, doublet of doublets, ddd = doublet of a dd, t = triplet, quin = quintet, sept = septet, br = broad, ap = apparent), (J = coupling constant in Hz) and (integration)]}. Proton-decoupled 13C NMR specta were recorded at 100 MHz on a Bruker spectrometer and are reported in ppm using the residual solvent signal (chloroform-d = 77.0 ppm; dimethylsulfoxide-d6 = 39.51 ppm; methanol-d4 = 49.15 ppm) as an internal standard. Infrared spectra were recorded using an ATR-FTIR instrument. High resolution mass spectra were acquired by flow injection on a qTOF Premiere Mass Spectrometer operating in ES+ ionization with resolution ~15,000.



Supplementary Scheme B.1: Synthesis of compound 1.

A mixture of compound **3** (6 g, 18.4 mmol), *tert*-butyl 4-aminobutylcarbamate (6.9 g, 36.8 mmol) and K_2CO_3 (7.6 g, 55.2 mmol) in 1,4-dioxane (80 mL) was refluxed for 3 h. The mixture was concentrated *in vacuo,* and the residue was diluted with ethyl acetate (50 mL). The solution was washed with water (20 mL) and brine (20 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by silica gel column (petroleum ether/ethyl actete, v/v = 1/3) to give 5.5 g of **SI-1** (71% yield) as a yellow oil. MS: m/z 479.1 $(M+H)^+$.

NaH (0.5 g, 21.0 mmol, 60% in mineral oil) was added to a solution of **SI-1** (5 g, 10.5 mmol) and TsCl (2.38 g, 12.6 mmol) in THF (50 mL) at 0 ^oC, and the mixture was stirred at 0 ^oC for 2 h. Water (10 mL) was added to the reaction mixture; and the resulting mixture was further stirred at room temperature for 10 min. The mixture was diluted with ethyl acetate (50 mL) and was washed with water (20 mL) and brine (20 mL). The organic layer was dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by short silica gel path (petroleum ether/ethyl acetate: 1/2) to give **SI-2** as white solid (4.5 g, 68%). MS: m/z 633.1 $(M+H)^+$.

A mixture of **SI-2** (6 g, 10.3 mmol) and NaN_3 (1.34 g, 20.6 mmol) in DMF (30 mL) was stirred at 60 ^oC for 2 h. The reaction mixture was diluted with ethyl acetate (150 mL) and washed with water (3 x 30 mL). The organic phase was dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo* to give the crude product **SI-3**, which was used for the next step without further purifications. MS: m/z 504.1 $(M+H)^+$.

PMe_3 (1.0 M in THF, 30 mL, 30 mmol) was added dropwise to a solution of **SI-3** (5 g, 10 mmol) in THF (40 mL); and the resulting mixture was stirred at room temperature overnight. The reaction mixture was quenched with water (2.1 mL) and stirred at room temperature for 2 h. The mixture was concentrated *in vacuo* and the residue was diluted with DCM (90 mL). The organic layer was washed with water (30 mL) and brine (15 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo* to give the crude **SI-4** as yellow oil. MS: m/z 478.1 $(M+H)^+$. The compound **SI-4** was used directly for next step.

A mixture of acetone (0.52 mL, 5.76 mmol) and acetic acid (0.32 mL, 5.5 mmol) was added dropwise to a solution of **SI-4** (2.5 g, 5.24 mmol) in DCM (40 mL), followed by addition of $NaBH(OAc)_3$ (3.8 g, 18 mmol) in portions. The mixture was stirred at room temperature for 1 h, was then diluted with DCM (50 mL) and washed with sat $NaHCO_3$ (50 mL). The aqueous phase was extracted with DCM (30 mL) and the combined organic layers were washed with brine (50 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by flash chromatography column (DCM/MeOH, v/v = 20, with addition of 1% NH_4OH) to give 1.3 g of **SI-5** (48% yield in 3 steps) as a colorless oil. MS: m/z 520.1 $(M+H)^+$.

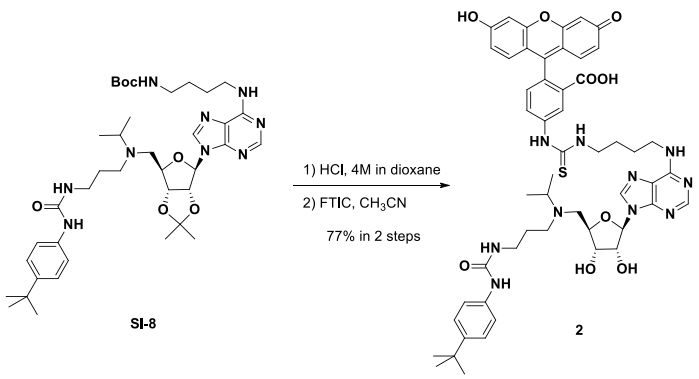
A mixture of **SI-5** (1.3 g, 2.5 mmol), 2-(3-bromopropyl) isoindoline-1,3-dione (1 g, 3.75 mmol), TBAI (95 mg, 0.25 mmol) and DIPEA (3.6 mL, 20.6 mmol) in acetonitrile (15 mL) was heated in microwave at 100 °C for 20 h. The mixture was cooled to room temperature, diluted with ethyl acetate (50 mL). The organic layer was washed with water (3 x 20 mL) and brine (30 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by flash chromatography column (DCM/MeOH, v/v = 50, with addition of 1% NH_4OH) to give 800 mg of **SI-6** (45% yield) as yellow oil. MS: m/z 707.1 $(M+H)^+$.

SI-6 (800 mg, 1.13 mmol) was dissolved in 2M methylamine in MeOH (20 mL). The mixture was stirred at room temperature for 5 minutes and was then heated at 55-60 °C for 16 h. The reaction mixture was concentrated *in vacuo*. The resulting oil was azeotroped with MeOH (20 mL) three times to afford the crude product **SI-7**, which was used for the next step without further purification. MS: m/z 577.4 $(M+H)^+$.

A solution of 1-tert-butyl-4-isocyanatobenzene (0.22 mL, 1.24 mmol) in DCM (2 mL) was added dropwise to a suspension of **SI-7 (**650 mg, 1.13 mmol, crude) in DCM (15 mL), and the resulted mixture was stirred at room temperature for 16 h. The reaction mixture was concentrated *in vavuo*, and the residue was purified by prep-HPLC to give 400 mg of **SI-8** (47% yield) as a white solid. MS: m/z 752.5 $(M+H)^+$; ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.36 (s, 1H), 7.77 (d, *J* = 10.0 Hz, 1H), 7.29 (m, 4H), 6.77 (m, 2H), 6.06 (s, 1H), 5.56 (s, 1H), 5.17 (m, 1H), 4.41 (br s, 1H), 3.65 (m, 2H), 3.17 (br m, 4H), 2.99 (br s, 1H), 2.73 (br s, 1H), 2.62 (m, 3H), 1.81 (m, 5H), 1.74 (s, 6H), 1.56 (s, 9H), 1.40 (s, 9H), 0.99 (s, 3H), 0.80 (m, 3H) ppm.

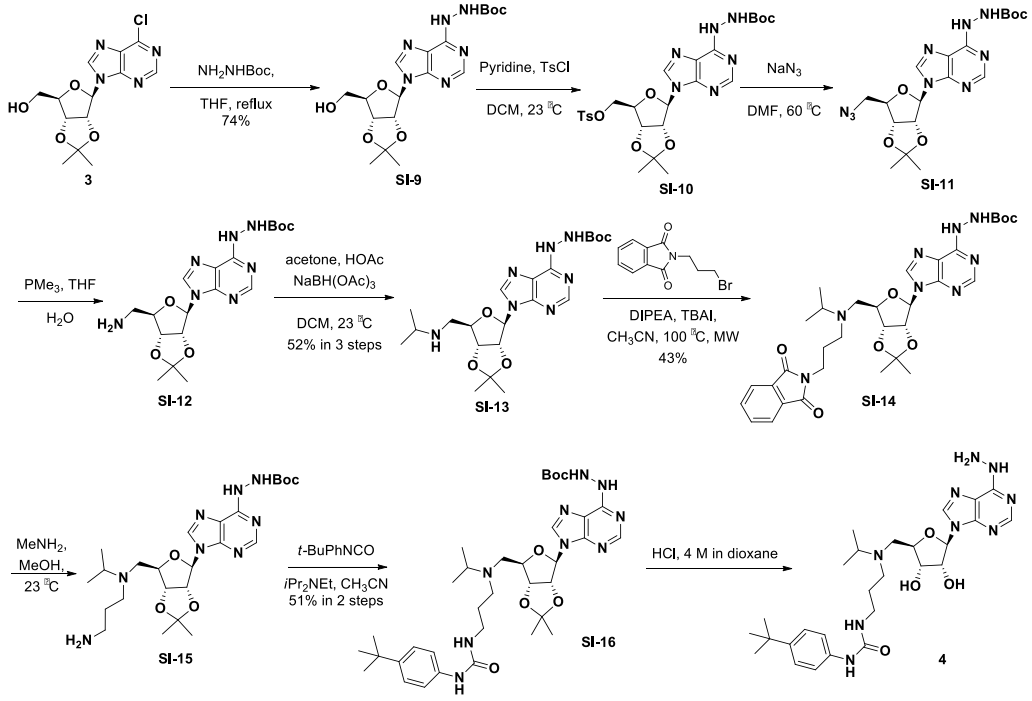
To a solution of compound **SI-8** (17.5 mg, 0.024 mmol) in THF (1 mL, 0.024 M) was added HCl solution (4 M in dioxane, 2 mL). The resulted mixture was stirred for 16 h and all solvent was removed under vacuum to afford crude free amine that was used for next step without further purification.

To a solution of the crude amine in DMF (1.2 mL, 0.2 M), DIPEA (16 mg, 0.12 mmol), Biotin-PEG_{11}-COOH (22 mg, 0.026 mmol), and HATU (11 mg, 0.026 mmol) were added sequentially. The resulted mixture was stirred at room temperature for 4 h, and was diluted with 10 mL of methanol. The crude reaction was then directly purified on HPLC to afford compound **1** as a white solid. MS: m/z 1395.8 $(M+H)^+$.



Supplementary Scheme B.2.: Synthesis of FITC probe 2.

To a solution of free amine obtained directly from deprotection of **SI-8 (**17.5 mg, 0.024 mmol, crude) in THF-MeOH mixture (1:1, total 1 mL), DIPEA (15 mg, 0.24 mmol) was added followed by the dropwise addition of a solution of FITC (7.3 mg, 0.188 mmol) in THF (0.5 mL). The resulted mixture was stirred at room temperature for 1h. The reaction mixture was quenched with 5 mL MeOH, and was purified by prep-HPLC to give 400 mg of **2** (47% yield) as a yellow solid. MS: m/z 1001.4 $(M+H)^+$.



Supplementary Scheme B.3: Synthesis of hydrazine 4.

A mixture of compound **3** (5 g, 15.3 mmol), tert-butyl 4-aminobutylcarbamate (5.7 g, 30.6 mmol) and DIPEA (7.9 mL, 45.9 mmol) in THF (70 mL) was refluxed overnight. The mixture was concentrated *in vacuo*. The residue was purified by flash chromatography column (petroleum ether/ethyl aceate, v/v = 2) to give 4.8 g of **SI-9** (74% yield) as a white solid. MS: m/z 423.1 $(M+H)^+$.

Pyridine (35.5 mmol) was added to a solution of compound **SI-9** in DCM, followed by addition of TsCl (2.0 g, 10.6 mmol) in portions. The mixture was allowed to stir at room temperature overnight. The reaction mixture was washed with water (3 x 20 mL) and brine (30 mL), dried over anhydrous sodium sulfate and concentrated *in vacuo*. The residue was purified by flash chromatography column (PE/EA: 1/2) to give **SI-10** as a white solid (2.3 g, 55% yield). MS: m/z 577.1 $(M+H)^+$.

A mixture of **SI-10** (2.3 g, 4.0 mmol) and NaN_3 (0.52 g, 8.0 mmol) in DMF (20 mL) was stirred at 60 °C for 2 h. The reaction mixture was diluted with ethyl acetate (100 mL) and washed with water (3 x 30 mL). The organic phase was dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo* to

give the crude product **SI-11**, which was used for the next step without further purifications. MS: m/z 448.1 $(M+H)^+$.

PMe_3 (1.0M in THF, 16 mL, 16.0 mmol) was added dropwise to a solution of **SI-11** (1.8 g, 4.0 mmol) in THF (40 mL), and the mixture was stirred at room temperature overnight. The reaction mixture was treated with water (1.5 mL) and stirred at room temperature for 2 h. The mixture was concentrated *in vacuo* and the residue was redissolved in DCM (90 mL). The resulting mixture was washed with water (30 mL) and brine (15 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo* to give the crude **SI-12** as yellow oil. MS: m/z 422.1 $(M+H)^+$. Compound **SI-12** was used for next step without further purification.

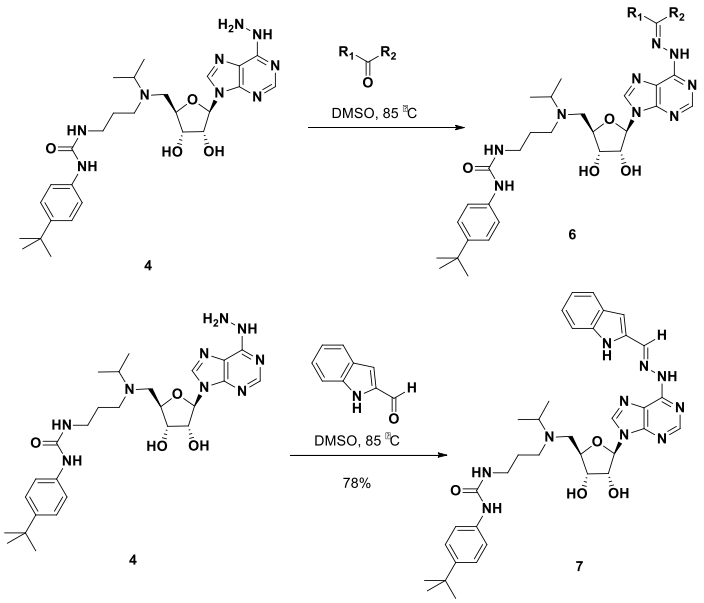
A solution of acetone (0.255 g, 4.4 mmol) and acetic acid (0.24 g, 4.0 mmol) was added dropwise to a solution of **SI-12** (1.6 g, 4.0 mmol, crude) in DCM (20 mL), followed by addition of $NaBH(OAc)_3$ (1.02 g, 4.8 mmol) in portions, and the mixture was stirred at room temperature for 1 h. The mixture was diluted with DCM (50 mL) and washed with sat $NaHCO_3$ (50 mL). The aqueous phase was extracted with DCM (2 x 30 mL), and the combined organic layers were washed with brine (30 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by silica gel column (DCM/MeOH, v/v = 20, with addition of $NH_3{\cdot}H_2O$ 1%) to give 0.93 g of **SI-13** (52% yield in 3 steps). MS: m/z 464.1 $(M+H)^+$.

A mixture of **SI-13** (0.5 g, 1.08 mmol), 2-(3-bromopropyl) isoindoline-1,3-dione (430 mg, 1.62 mmol), TBAI (42 mg, 0.11 mmol) and DIPEA (1.5 mL, 8.64 mmol) in acetonitrile (15 mL) was heated in a microwave at 100^{o}C for 20 h. The mixture was diluted with ethyl acetate (50 mL), washed with water (20 mL) and brine (20 mL), dried over anhydrous sodium sulfate and concentrated *in vacuo*. The residue was purified by flash chromatography column (DCM/MeOH, v/v = 50, with addition of $NH_3{\cdot}H_2O$ 1%) give 300 mg of **SI-14** (43% yield). MS: m/z 651.1 $(M+H)^+$.

SI-14 (200 mg, 0.31 mmol) was dissolved in 2M methylamine in MeOH (5 mL) and the mixture was stirred at room temperature for 5 min, then heated at 55-60 °C for 2 h. The reaction mixture was concentrated *in vacuo*. The resulting oil was azeotroped with MeOH (10 mL) twice to afford the crude product **SI-15**, which was used for the next step without further purifications. MS: m/z 521.4 $(M+H)^+$.

A solution of 1-tert-butyl-4-isocyanatobenzene (0.03 mL, 0.19 mmol) in DCM (1 mL) was added dropwise to a suspension of **SI-15 (**120 mg, 0.17 mmol, crude) in DCM (5 mL), and the resulted mixture was stirred at room temperature for 1 h. The reaction mixture was concentrated *in vacuo* and the crude product was purified by prep-HPLC to give 60 mg of **SI-16** (51% yield) as a white solid. MS: m/z 696.5 $(M+H)^+$; ^1H NMR (500 MHz, $DMSO_{d6}$) δ 9.27 (m, 2H), 8.32 (m, 3H), 7.28 (d, *J* = 9.0 Hz, 2H), 7.21 (d, *J* = 9.0 Hz, 2H), 6.19 (s, 1H), 6.01 (t, *J* = 5.5 Hz, 1H), 5.55 (s, 1H), 5.00 (s, 1H), 4.17 (s, 1H), 3.08 (m, 2H), 2.87 (m, 1H), 2.64 (m, 1H), 2.35 (m, 3H), 2.50 (m, 16H), 1.35 (m, 10H), 0.95 (d, *J* = 6.0 Hz, 3H), 0.77 (d, *J* = 6.0 Hz, 3H) ppm.

HCl solution (4.0 M in dioxane, 4 mL) was added to a solution of **SI-16** (32 mg, 0.045 mmol) in THF (2 mL), and the resulting mixture was stirred at room temperature overnight. The solvent was removed *in vacuo*, and the resulting product was used directly without purification.

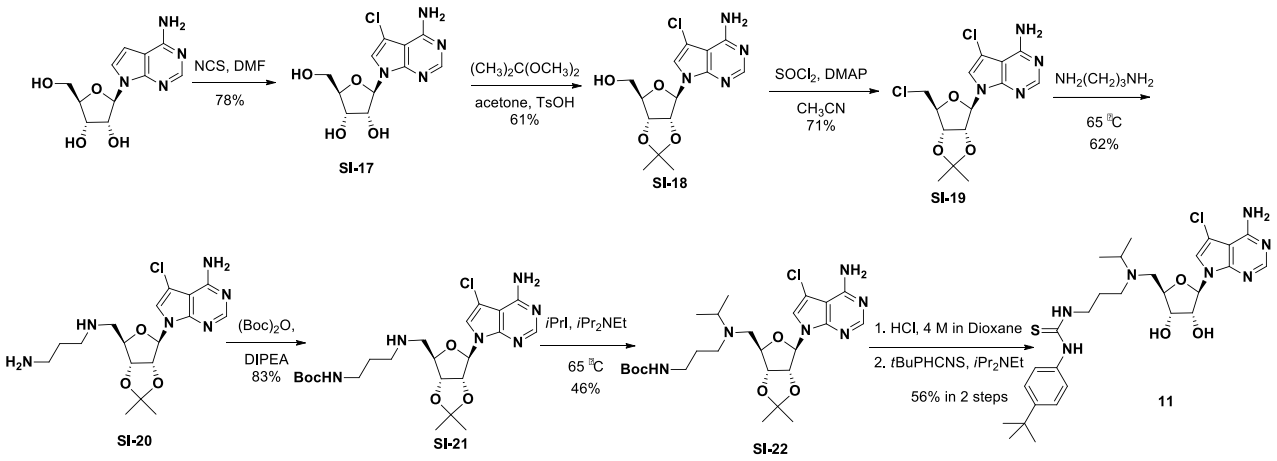


Supplementary Scheme B.4: Library synthesis and the resynthesis of library hits.

The crude compound **4** was dissolved into DMSO to generate 11mM stock solution. The solution was then transferred to 96-well deep well plate (90 µL in each well), followed by addition of aldehyde solution (100 mM, 10 µL) in each well. The plate was sealed and heated in oven at 85 °C for 24 h to

afford 96 well hydrazone library plate. All the hydrazones on the plate were checked by LCMS and were observed to have 85%-90% purity. The compound plate was then directly used for screening without further purification.

Based on the screening data, three compounds, **7, 8, 9** were synthesized in the same manner as the library synthesis. The products were then purified by prep-HPLC to give the hydrazones in 78-83% yield. These three compounds were then tested to generate a dose response curve.



Supplementary Scheme B.5: Synthesis of compound 11.

A solution of N-chlorosuccinimide (251.7 mg, 1.88 mmol) in DMF (5 mL) was added to a solution of 7-deazaadenosine (250.0 mg, 0.94 mmol) in DMF (5 mL). The reaction was stirred at room temperature for 2.5 h. The reaction was diluted with DCM (100 mL) and hexanes (900 mL) was added slowly with stirring to precipitate out the desired product. The reaction was then filtered to give the desired product **SI-17** (221.5 mg, 78%) as a light pink solid. MS *m/z* 300.93 $(M+H)^+$.

2,2-Dimethyoxypropane (7.2 mL, 58 mmol) was added to a suspension of **SI-17** (1 g, 3.3 mmol) in acetone (33 mL) followed by addition of TsOH-H_2O (620 mg, 3.3 mmol). The reaction was stirred at room temperature for 5 h. $NaHCO_3$ (840 mg, 10 mmol) was added to the reaction mixture and was further stirred at room temperature for 0.5 h. The solvent was then removed *in vacuo*, and the residue was resuspended in DCM (150 mL) and washed with sat. $NaHCO_3$ (10 mL) and H_2O (10 mL). The organic layer was collected, dried over anhydrous sodium sulfate, was filtered and concentrated *in vacuo*.

The residue was purified by flash chromatography column (0-15% MeOH:DCM) to afford the desired product **SI-18** (685 mg, 61%). MS *m/z* 340.95 $(M+H)^+$.

DMAP (61.5 mg, 0.50 mmol) was added to a solution of **SI-18** (535 mg, 1.6 mmol) in acetonitrile (16 mL) under nitrogen. A solution of $SOCl_2$ (89.8 uL, 1.23 mmol) in acetonitrile (0.3 mL) was added to reaction mixture dropwise at 0 °C. The reaction mixture was warmed to room temperature overnight, and was quenched with MeOH (20 mL), water (2 mL) and aq. NH_4OH (3 mL). The solvent was removed *in vacuo* and the residue was diluted with DCM (150 mL). The organic layer was washed with sat. $NaHCO_3$ (5 mL) and the aqueous layer was extracted with DCM (100 mL). The combined organic layers were washed with 5% HCl (5 mL), were dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by flash chromatography column (0-5% MeOH:DCM) to afford **SI-19** (400 mg, 71%). ^{1}H NMR (400 MHz, CD_3OD) δ ppm 1.37 (s, 3 H) 1.59 (s, 3 H) 3.66 - 3.72 (m, 1 H) 3.73 - 3.79 (m, 1 H) 4.31 - 4.40 (m, 1 H) 5.03 (dd, *J*=6.26, 3.13 Hz, 1 H) 5.25 (dd, *J*=6.26, 3.13 Hz, 1 H) 6.23 (d, *J*=2.74 Hz, 1 H) 7.35 (s, 1 H) 8.13 (s, 1 H). ^{13}C NMR (400 MHz, CD_3OD) δ ppm 25.7, 27.6, 45.2, 83.8, 85.6, 86.6, 91.5, 106.2, 115.9, 121.3, 150.3, 153.9, 158.7. MS *m/z* 358.70 $(M+H)^+$.

A solution of SI-19 (13.6 mg, 0.038 mmol) in 1,3-diaminopropane (2 mL) was stirred at room temperature for 10 min and then heated to 60 °C overnight. The diamine was removed in vacuo, and the residue was then purified by flash chromatography column (0-15% MeOH:DCM+1%NH4OH) to afford SI-20 (205 mg, 62%). 1H NMR (400 MHz, CD3OD) δ ppm 1.38 (s, 3 H) 1.56 - 1.65 (m, 3 H) 1.98 - 2.11 (m, 2 H) 3.01 (t, J=7.63 Hz, 2 H) 3.11 (t, J=8.02 Hz, 2 H) 3.44 - 3.57 (m, 2 H) 4.45 - 4.52 (m, 1 H) 5.12 (dd, J=6.46, 4.11 Hz, 1 H) 5.30 (dd, J=6.26, 2.74 Hz, 1 H) 6.19 - 6.23 (m, 1 H) 7.44 - 7.48 (m, 1 H) 8.17 (s, 1 H). 13C NMR (400 MHz, CD3OD) δ ppm 25.3, 25.7, 27.6, 38.1, 46.5, 51.0, 83.0, 83.4, 85.6, 92.3, 106.0, 116.5, 122.3, 150.0, 152.9, 153.9, 158.7. MS m/z 397.33 $(M+H)^+$.

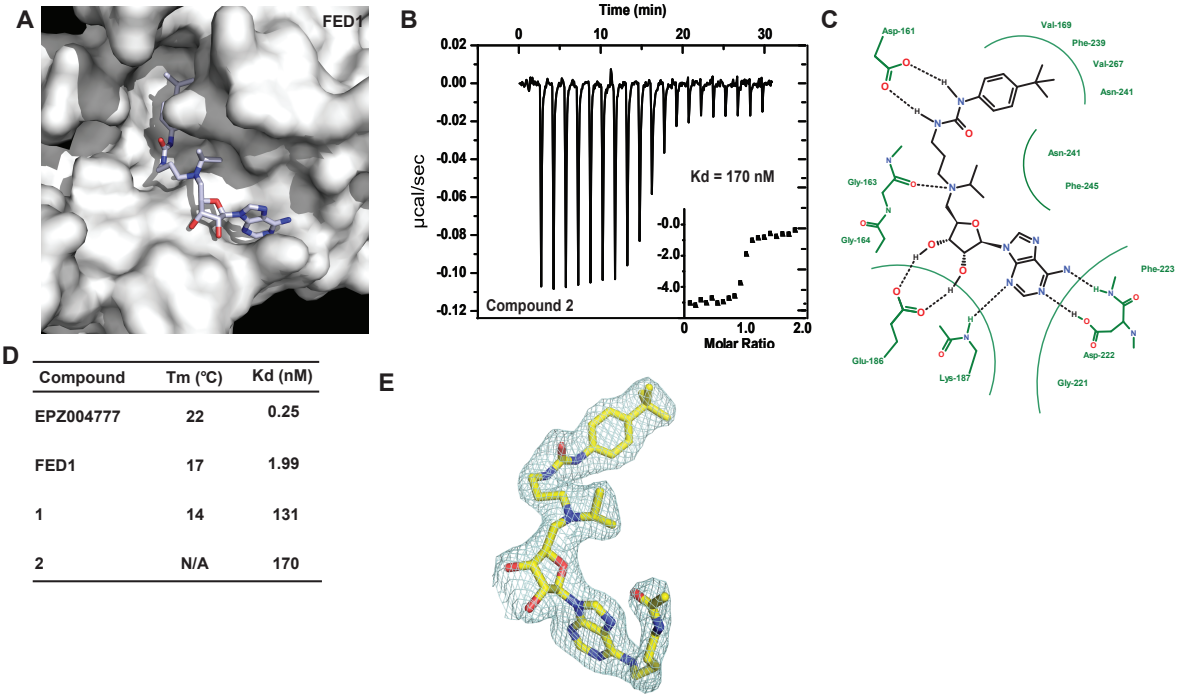
A solution of Boc_2O (153 mg, 0.7 mmol) in acetonitrile (0.5 mL) was added slowly to a solution of **SI-20** (279 mg, 0.7 mmol) in acetonitrile (7 mL) with DIPEA (611 uL, 3.5 mmol) at 0 °C. The reaction was stirred at 0°C for 10 min. The reaction was diluted with DCM (20 mL), washed with sat. $NaHCO_3$ (3 mL) and then the aqueous layer was extracted with DCM (2 x 20 mL). The combined organic layers were then washed with H_2O (3 mL), dried over anhydrous sodium sulfate, filtered and concentrated *in vacuo*. The residue was purified by flash chromatography column to afford **SI-21** (288 mg, 83%). ^{1}H NMR (400 MHz,

CD_3OD) δ ppm 1.37 (s, 3 H) 1.40 (s, 9 H) 1.59 (s, 3 H) 1.64 (t, *J*=7.04 Hz, 2 H) 2.70 (br. s., 2 H) 3.01 - 3.07 (m, 3 H) 3.47 - 3.62 (m, 1 H) 4.30 (d, *J*=3.91 Hz, 1 H) 4.96 (d, *J*=1.96 Hz, 1 H) 5.28 (dd, *J*=6.46, 2.93 Hz, 1 H) 6.19 (d, *J*=2.74 Hz, 1 H) 7.39 (s, 1 H) 8.13 (s, 1 H). ^{13}C NMR (400 MHz, CD_3OD) δ ppm 25.7, 27.6, 28.9, 29.9, 38.8, 52.2, 64.6, 65.7, 83.7, 85.2, 85.5, 91.6, 102.4, 106.0, 116.1, 121.8, 150.2, 153.9, 158.7. MS *m/z* 496.58 $(M+H)^+$.

SI-21 (288 mg, 0.58 mmol) was dissolved in acetonitrile (4 mL), 2-iodopropane (1.8 mL) and DIPEA (1 mL). The reaction mixture was heated to 65°C overnight, and was concentrated *in vacuo*. The residue was then purified by flash chromatography column (0-15% MeOH:DCM) to afford **SI-22** (144 mg, 46 % with recovering 140 mg of starting material **SI-21**). ^1H NMR (400 MHz, CD_3OD) δ ppm 0.85 - 0.92 (m, 5 H) 1.02 (d, *J*=7.04 Hz, 3 H) 1.27 - 1.34 (m, 9 H) 1.38 (s, 3 H) 1.41 (s, 9 H) 1.59 (s, 4 H) 1.91 (s, 1 H) 2.49 (br. s., 1 H) 3.05 (d, *J*=6.65 Hz, 1 H) 4.19 (d, *J*=3.52 Hz, 1 H) 4.56 (s, 4 H) 5.25 - 5.35 (m, 1 H) 6.18 (d, *J*=2.74 Hz, 1 H) 7.37 (s, 1 H) 8.13 (s, 1 H). MS *m/z* 538.58 $(M+H)^+$.

To a suspension of **SI-22** (8.0 mg, 0.015 mmol) in tetrahydrofuran (1 mL) was added 4M HCl in dioxane (1 mL). The reaction was stirred at room temperature overnight and then concentrated *in vacuo* to give the desired free amine. MS *m/z* 399.31 $(M+H)^+$.

To a solution of free amine (5.55 mg, 0.014 mmol, crude) in a 1:1 mixture of acetonitrile:tetrahydrofuran (0.4 mL) was added DIPEA (18.0 uL, 0.139 mmol) followed by a solution of 4-tertbutylphenyl isothiocyanate (2.7 mg, 0.014 mmol) in the 1:1 mixture of acetonitrile:tetrahydrofuran (0.1 mL). The reaction was stirred at room temperature for 30 min. The reaction was diluted with MeOH (1.5 mL) and purified by prep-HPLC to afford the desired product **11** (4.6 mg, 56% in 2 steps). ^1H NMR (600 MHz, CD_3OD) δ ppm 1.31 (s, 9 H) 1.36 - 1.40 (m, 6 H) 2.02 - 2.11 (m, 1 H) 3.23 (q, *J*=7.24 Hz, 2 H) 3.49 - 3.54 (m, 1 H) 3.55 - 3.60 (m, 1 H) 3.62 - 3.69 (m, 1 H) 3.73 (dt, *J*=13.06, 6.68 Hz, 2 H) 4.32 (br. s., 4 H) 6.14 (br. s., 1 H) 6.99 - 7.24 (m, 2 H) 7.38 (s, 3 H) 8.11 (s, 1 H). ^{13}C NMR (400 MHz, CD_3OD) δ ppm 17.6, 19.0, 32.0, 35.7, 40.7, 56.2, 73.8, 126.2, 127.8, 154.0

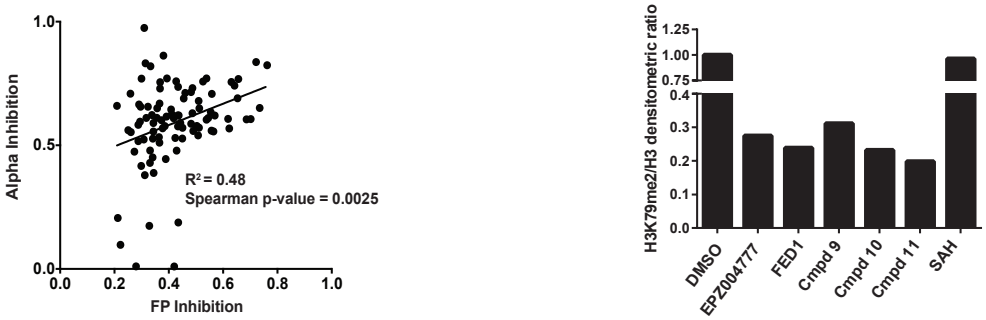


D

Compound	Tm (°C)	Kd (nM)
EPZ004777	22	0.25
FED1	17	1.99
1	14	131
2	N/A	170

E

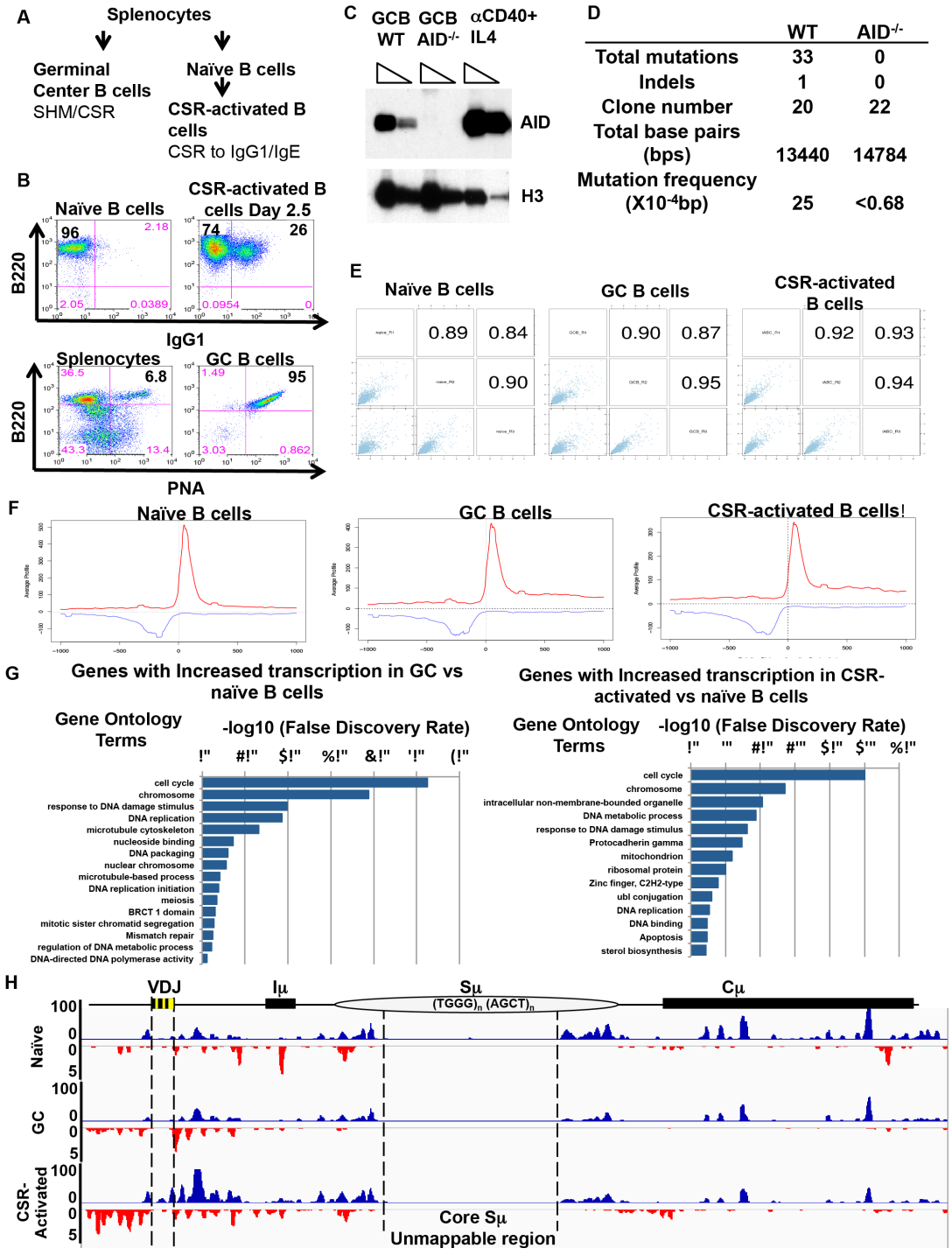
Supplementary Figure B.1: Structural characterization of probe compounds (a) FED1 crystallized with DOT1L. (b) ITC analysis of 2 demonstrating strong 1:1 binding with DOT1L. (c) FED1 ligand-interaction diagram with DOT1L. (d) DOT1L protein thermal melt shifts (°C) of indicated compounds. Note 2's fluorophore interfered with interpretation of SYPRO orange fluorescence.



Supplementary Figure B.2: Assay measurements on synthesized compounds (a) Scatter plot comparing the correlation between degree of inhibition of the hydrazine library in the AlphaScreen vs. the fluorescence polarization screening assay at 50uM. Generally acceptable correlation was seen between the 2 assays (b) Quantification of blot against H3K79me2 upon treatment with various inhibitors

Appendix C

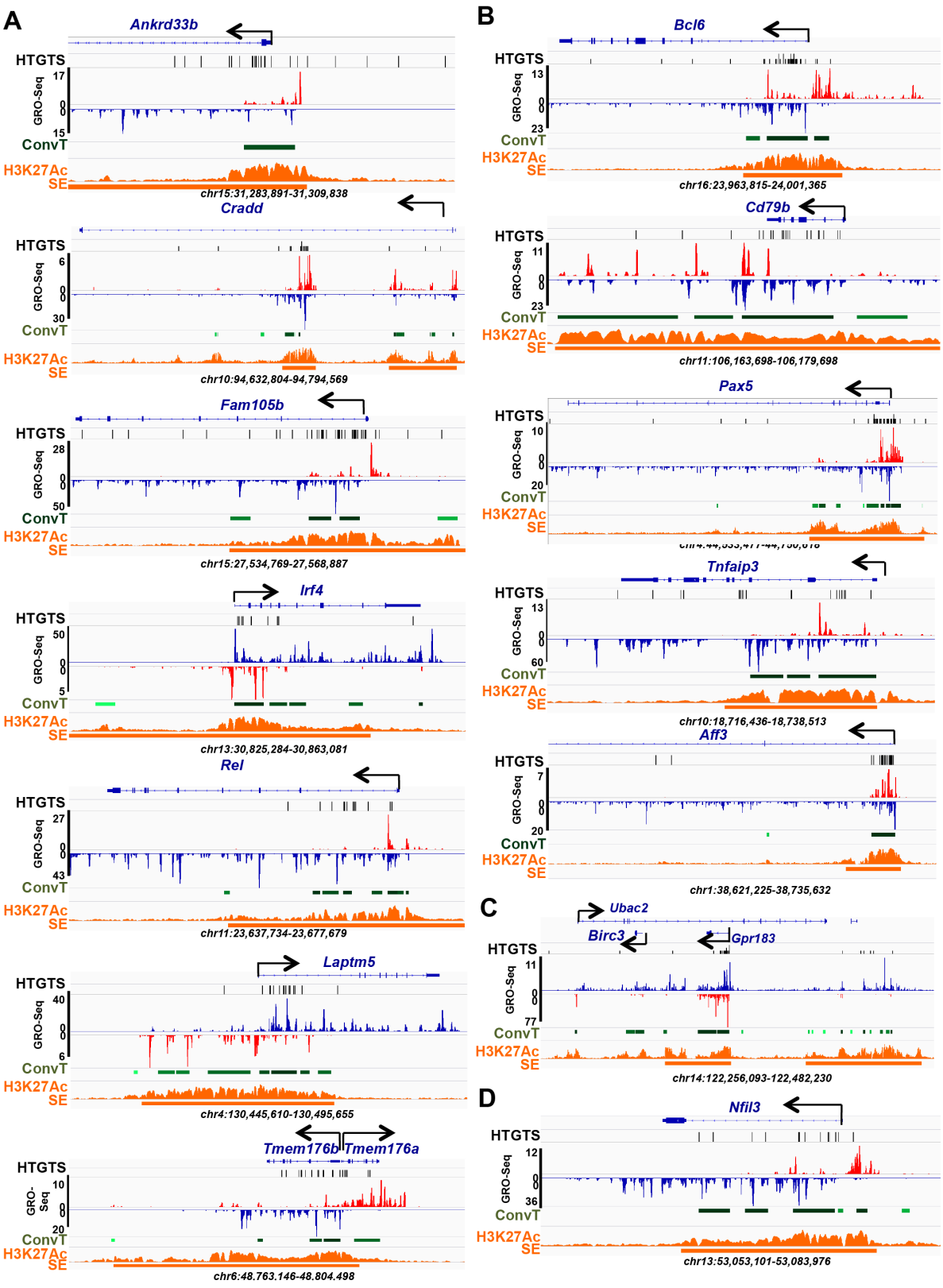
Supplementary Materials for Chapter 5



Supplementary Figure C.1: Transcriptional profiles of three different types of B cells

Supplementary Figure C.1 (Continued)

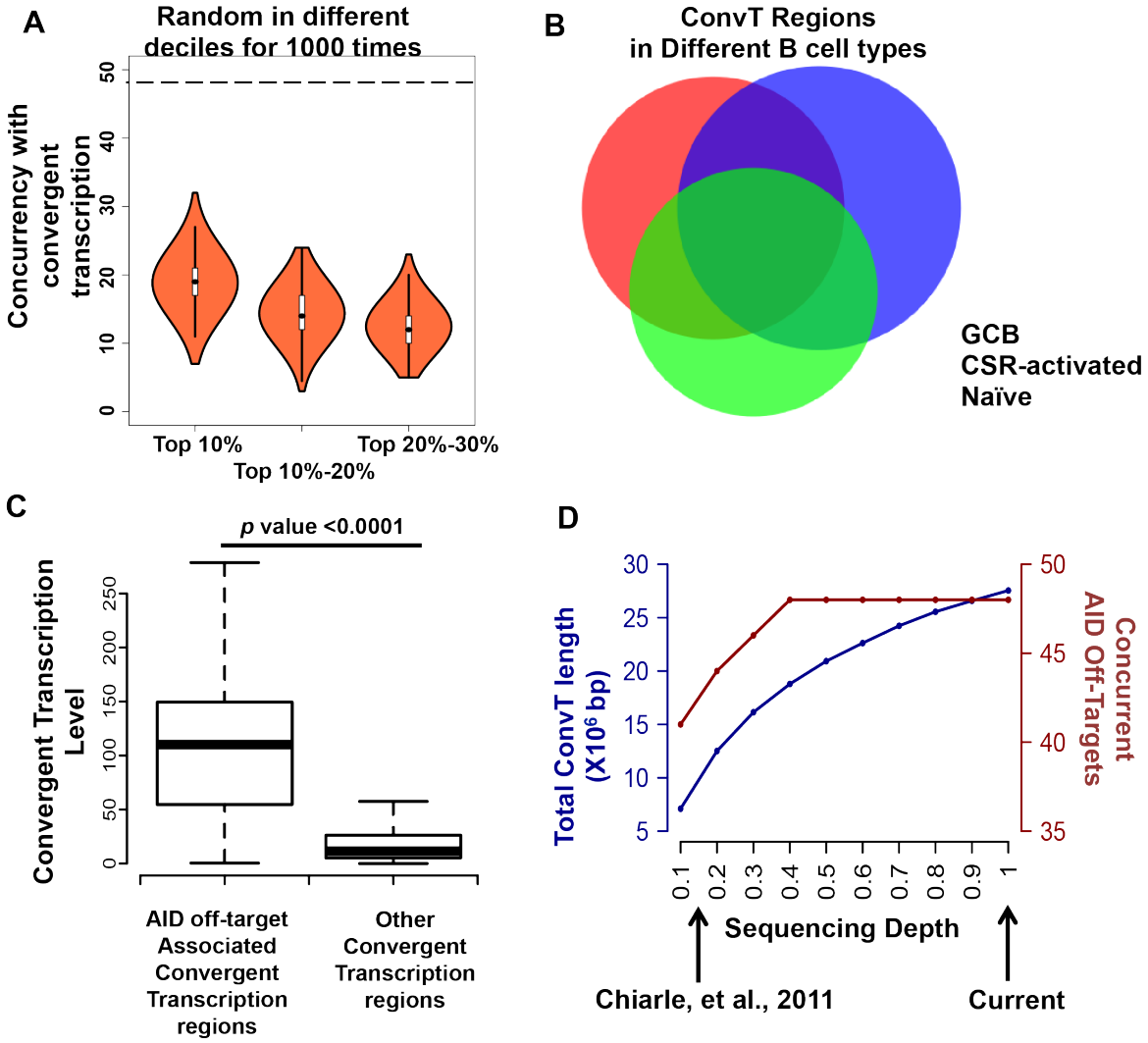
(A) General B cell purification Approach. See text for details. (B) Representative FACS plots of naïve, CSR-activated and GC B cells. Upper panels: At Day 2.5 after activation, about 25% V_HB1-8 B cells switched to IgG1. Lower Panels: FACS staining of total splenocytes and purified GC B cells for B220 versus PNA. GC B cells are the $B220^+PNA^{hi}$ population. (C). Western blot of AID protein in purified WT, $AID^{-/-}$ GC B cells and *ex vivo* αCD40 plus IL4 activated B cells. Histone H3 was blotted as a loading control. The anti-AID antibody used has been described[3] and anti-histone H3 antibody was from Abcam (ab1791). (D). SHM of the J_H4 intron in the purified WT and $AID^{-/-}$ purified GC B cells performed as described previously (Jolly et al., 1997). (E) Correlation among different GRO-Seq replicates done on the same cell type. GRO-Seq signals per 10kb bin along the mouse genome were plotted for the three independent replicates of each cell type. The Pearson Correlation Co- efficient (*r*) among the replicates was calculated and indicated on the figures. (F) Aggregate patterns of gene transcription profiles (metagene profile) at 1kb region on either side of TSSs. Antisense transcription is indicated in blue (bottom) and sense transcription is indicated in red (top). (G) Gene Ontology (GO) enrichment analysis of genes showing increased expression in GC B cells compared to naïve B cells (left), and GO enrichment analysis of genes showing increased expression in *ex vivo* CSR-activated B cells compared to in naïve B cells (right). False Discovery Rate (FDR) is -log10 converted and represented by a bar to show the significance of enrichment in different GO concepts. (H) GRO-Seq profiles of V_HB1-8 preassembled V(D)J exon and the downstream C_H region in the three cell types analyzed. The V(D)J exon, Iμ and Cμ are indicated by solid bars. Complementarity determining regions (CDRs) are shown by yellow bars inside of V(D)J exon, which contains WRCH (W=A/T, R=A/G and H=A/C/T) "SHM" hotspots. Predominant motifs (TGGG and AGCT) are highlighted for S region DNA. The unmappable core Sμ region is indicated by two vertical dashed lines. The Y-axis indicates the GRO-Seq read counts normalized to reads per million reads. Reads aligned to annotated gene sense and antisense strands are displayed in blue and red. The profiles of core Sμ region and V(D)J exon were incomplete because of the low mappability.



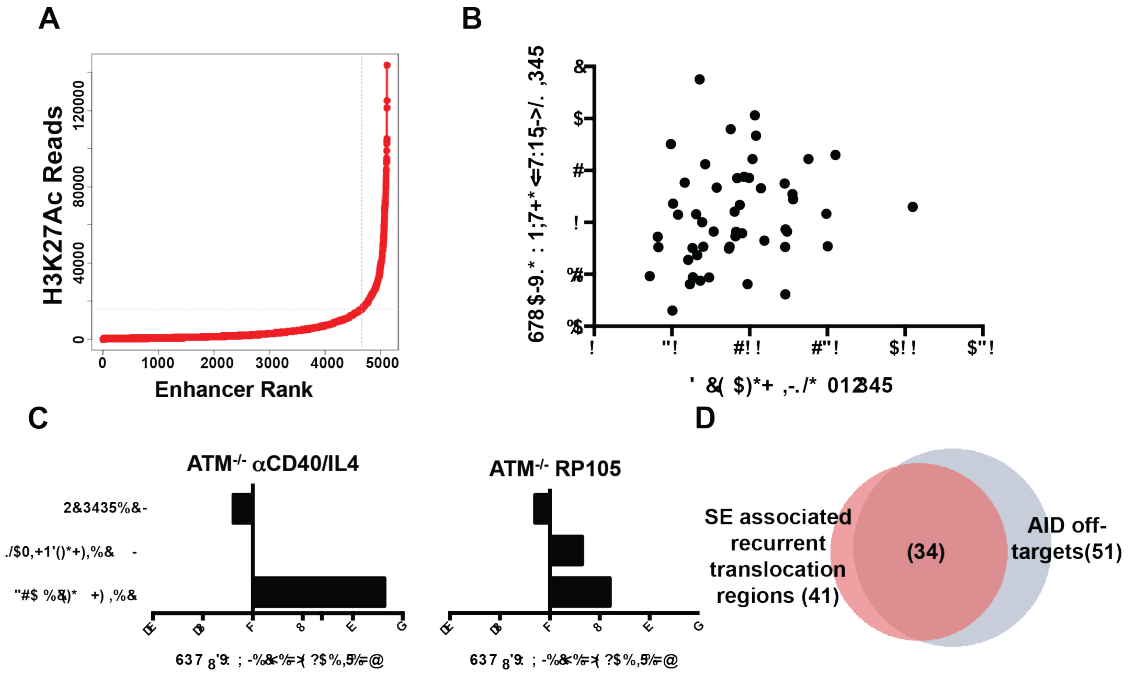
Supplementary Figure C.2: HTGTS, GRO-seq and H3K27Ac profiles of several AID off-target sites

Supplementary Figure C.2 (Continued):

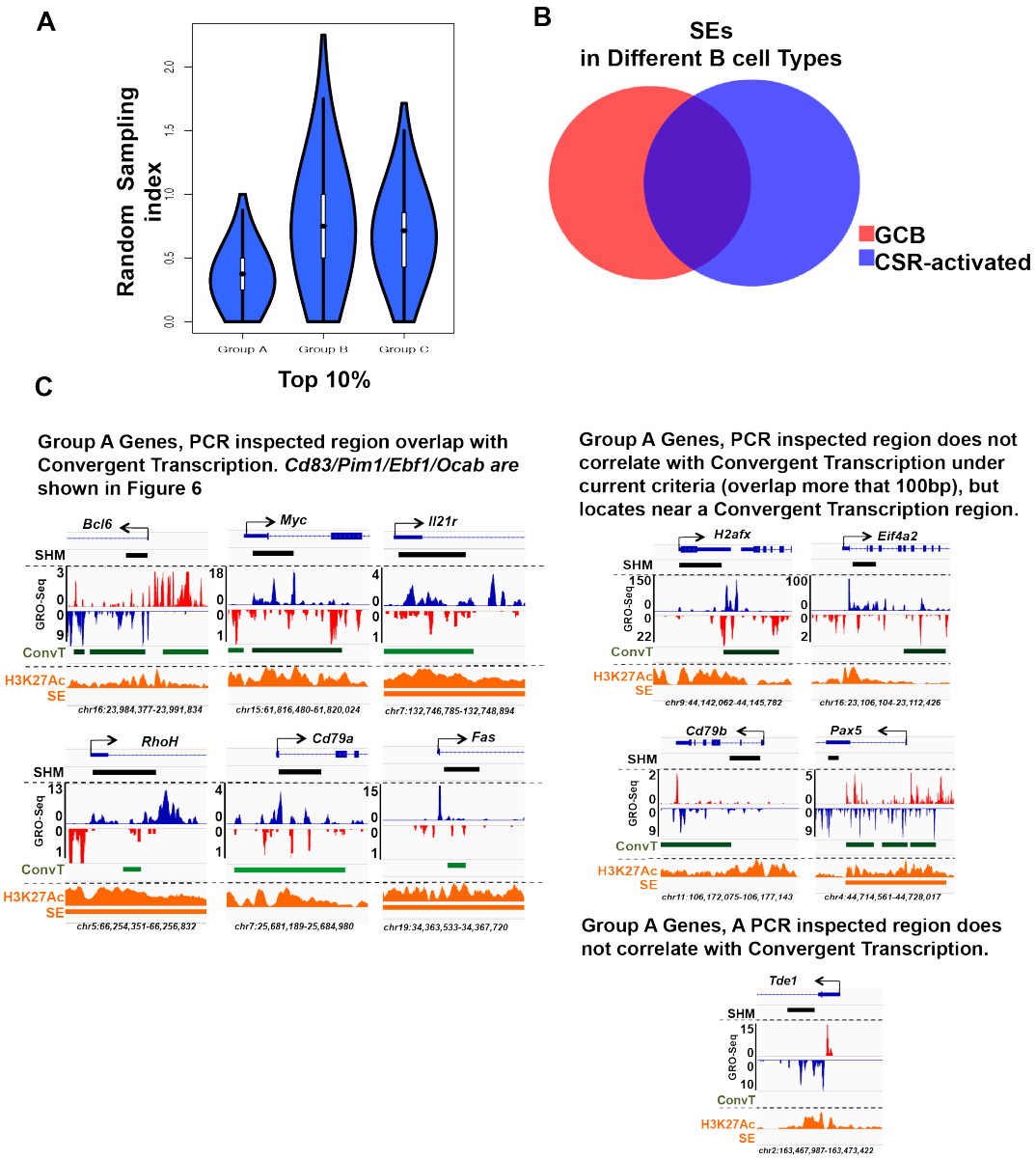
Translocation junctions from ATM$^{-/-}$ CSR-activated B cell HTGTS data are indicated in the HTGTS row (top), except for *Myc* which was the DSB bait site for HTGTS cloning. GRO- Seq determined gene sense and antisense transcription is displayed in blue and red, respectively in the middle panel. Convergent transcription ("ConvT") is shown as green bars at the bottom of the GRO-Seq panel with the darkest shades corresponding to highest levels of convergent transcription as calculated by the geometric means of antisense and sense transcription reads (see Fig. 5.2A). The H3K27Ac ChIP-Seq profile is shown in orange, and Super-Enhancers (SEs) are shown below it in the bottom panel. (A). This set of panels shows 7 newly identified AID off-targets. (B). This panel shows AID off-targets whose human orthologs are oncogenes (see text for details). Panel C shows an example of a Class 4 gene (see text for details). Panel D is an example of a novel AID hotspot gene identified by the independent pipeline for SE-associated recurrent AID dependent HTGTS hotspots.



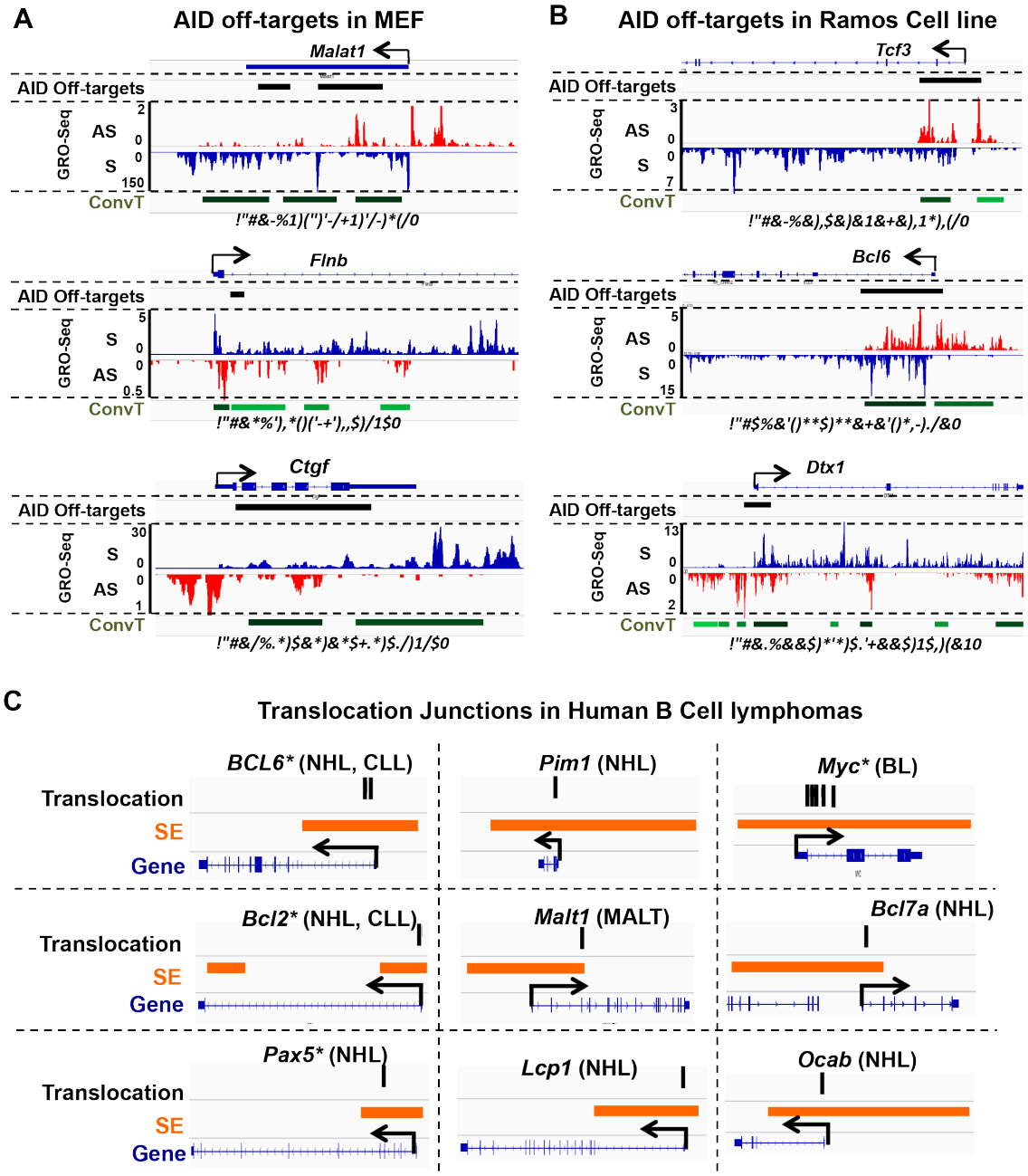
Supplementary Figure C.3: AID off-targets associated with convergent transcription. (A) Random sampling in transcription regions. Random sampling of regions corresponding in size to those of AID off-target regions in three highest deciles (with respect to transcription levels) of transcribed genes revealed that the numbers of regions associated with convergent transcription in each sampling was substantially lower than that of regions containing AID off-targets. Random-sampling results are displayed in violin plots. The dashed line indicates the observed number of AID off-target regions associated with convergent transcription. (B) Venn diagram showing the overlap of convergent transcription regions among the three B cell types analyzed. (C) Convergent transcription levels of AID off-target associated convergent transcription regions and other non-AID off target associated convergent transcription regions are plotted. AID off-target associated convergent transcription regions had a significantly higher level of convergent transcription (Mann-Whitney U-test, *p* value <0.0001). (D) Sequencing Depth affects convergent transcription identification. The 306 million total mappable-reads from CSR-activated B cells were pooled and then randomly sampled. Random fractions of sequences at different sequencing depth were subjected to convergent transcription identification and AID off-target association analysis. The total convergent transcription region length continued increased with deeper sequencing depth (blue line). The numbers of AID off- targets associated convergent transcription reached saturation at about 120 millions mappable reads. The sequencing depth of our previously published GRO-Seq dataset[21] is indicated in the figure that as shown was not sufficient to identify the convergent transcription correlation with AID off-targets.



Supplementary Figure C.4: AID off-targets located at SE-gene overlap. *(A) Identification SEs with ROSE.* All enhancers were ranked by their level of H3K27Ac ChIP-Seq reads. Enhancers with values above 15755 were considered SEs. (B) HTGTS translocation junction numbers and H3K27Ac read density are plotted for individual AID off-targets. Spearman's rank correlation coefficient and two-tailed *p* value are indicated. (C) Observed versus Expected HTGTS translocation frequency in αCD40 plus IL4 activated and RP105 activated ATM deficient B cells. The filtered HTGTS junctions were grouped into three different genomic regions, typical enhancer, super-enhancer and promoter. Typical and super enhancers were defined by using H3K27Ac ChIP-Seq peaks and promoters were defined at +/- 1kb from the annotated TSSs. Areas that overlap multiple regions (e.g. super-enhancers that cover promoter regions) were assigned to both categories. Expected values were estimated based on relative sizes of the three regions. The ratio of observed event versus the expected event was calculated for each category. (D) The overlap between SE-associated recurrent translocation regions identified by an independent pipeline and AID off-target sites identified with previously described pipeline[21]. Among the 41 SE-associated recurrent translocation regions identified, 34 were identified as AID off-targets by the other pipeline and 7 were novel and all correlated with convergent transcription.



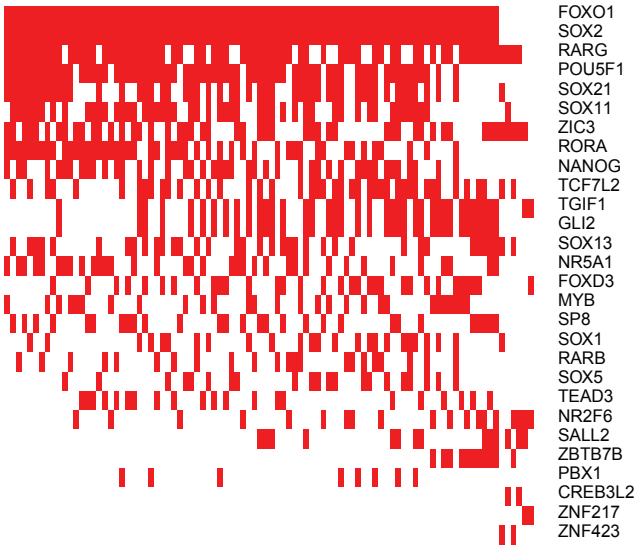
Supplementary Figure C.5: AID off-targets in GC B cells. (A) Venn diagram showing the overlap of SEs between CSR-activated and GC B cells. (B) Random sampling in highly transcribed gene regions revealed by GRO-Seq in GC B cells. The randomly sampling was done as described in the legend to Supplementary Figure C.3A. The numbers of gene regions with convergent transcription in each random sampling was normalized to number of genes in Group A, B and C that convergently transcribed region (Figure. 5.6A). The normalized values (termed “random sampling index”) are displayed in violin plots. The closer a random sampling index is to 1, the more likely it is random. (C) GRO-Seq, ConvT, H3K27Ac and SE profiles of gene regions in highly mutated Group A genes. Regions that are around the annotated TSS are shown. Regions included in prior mutation analyses are marked as black bar in the “SHM” row, and regions of AS/S convergent transcription (ConvT) are labeled below as green bars.



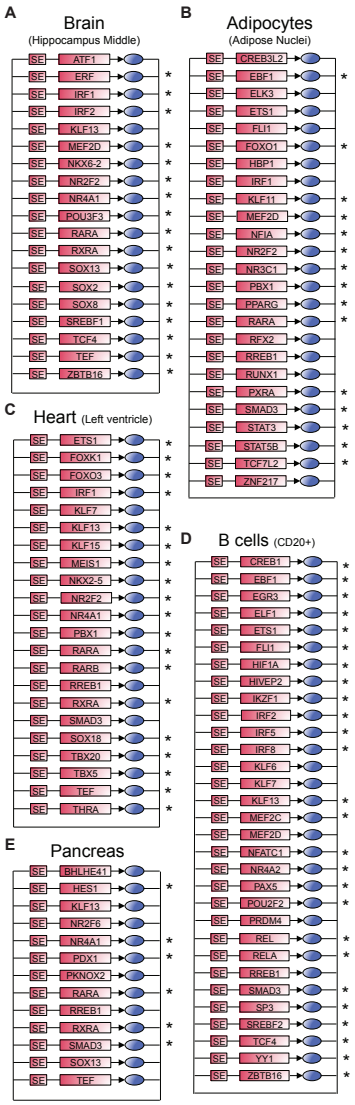
Supplementary Figure C.6: Profiles of AID targets in MEFs, a Ramos Human B cell lymphoma line, Human B cell lymphoma translocations. (A) GRO-Seq and ConvT profiles of exampled AID off-targets in MEF cells. AID off- target Regions are shown. AID off-target information was retrieved from Wang et al., unpublished. (B) GRO-Seq and ConvT profiles of exampled AID off-targets in Ramos cell lines. AID off-target regions are shown. AID off-target information was retrieved from Qian et al., unpublished. (C) Human B cell lymphoma translocation junctions often occur in Human Tonsil Cell SE-Gene overlap regions. Translocation junction information was retrieved form TICdb database. SE location information of human tonsil was retrieved from. NHL: non-Hodgkin lymphoma; CLL: chronic lymphocytic leukaemia; BL: Burkitt lymphoma; MALT: mucosa-associated lymphoid tissue lymphoma. Asterisks indicate oncogenes that were implicated as AID off-targets in human lymphomas.

Appendix D

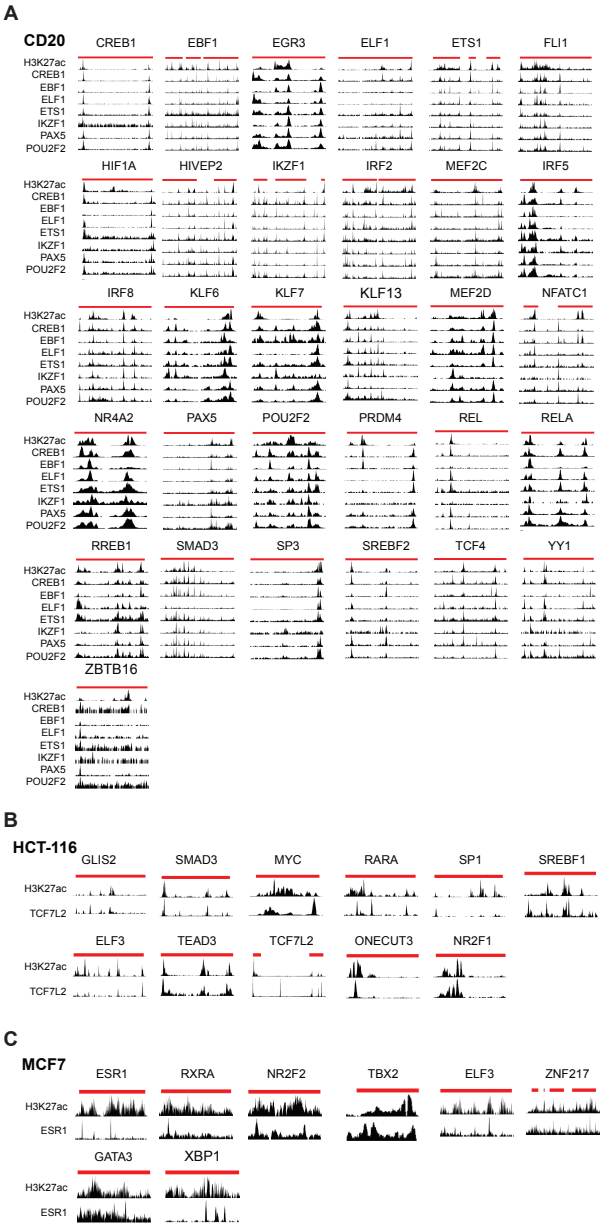
Supplementary Materials for Chapter 6



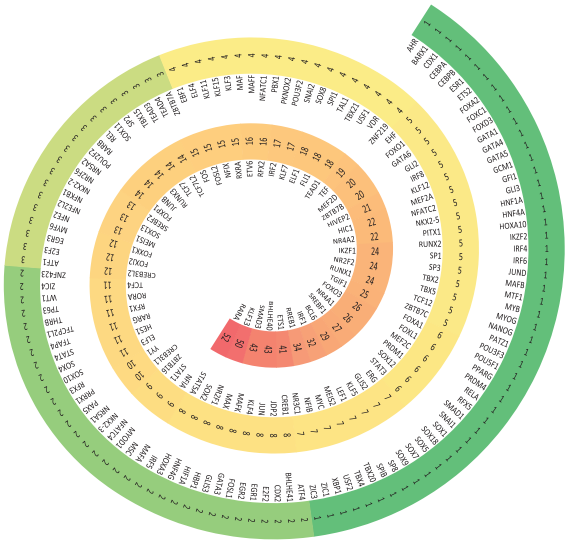
Supplementary Figure D.1: **Example of CRC selection out of multiple fully interconnected autoregulatory loop (IAL) possibilities.** The TF content of each possible IAL of H1 hESC is plotted on each column. The TFs were ranked vertically by decreasing fraction of their occurrences across all the possible IALs. The loops are ranked from left to right by average fraction of TF occurrence in the loops. The CRC corresponds to the leftmost loop.



Supplementary Figure D.2: **Examples of CRC models for 5 well studied cell-types.** A) brain (hippocampus middle), B) adipocytes (adipose nuclei), C) heart (left ventricle), D) B-cells (CD20+) and E) pancreas. The TFs which role have been described as critical for the control of cell identity of those cell or tissue types are marked with stars.



Supplementary Figure D.3: ChIP-seq data support candidate core TF predictions for 3 cell types. ChIP-seq data showing binding of the TF to the SEs of the candidate core TF for (A) CD20+ B-cell (ChIP-seq data from GM12878 lymphoblastoid B-cells), (B) HCT-116 colon cancer cell line (ChIP-seq data from HCT-116) and (C) MCF-7 breast cancer cell line (ChIP-seq data from T-47D breast cancer cell line). SE genomic locations are depicted by red lines on top of the tracks.



Supplementary Figure D.4: Candidate core TFs are cell-type-specific or lineage-specific. Number of samples, among 84 samples, in which a TF is considered a candidate core TF.

References

Chapter 1

1. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012). doi:10.1038/nature11247
2. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* (2014). doi:10.1073/pnas.1318948111
3. Ippen, K., Miller, J., Scaife, J. & Beckwith, J. New Controlling Element in the Lac Operon of E. coli. *Nature* (1968). doi:10.1038/217825a0
4. Ong, C.-T. & Corces, V. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* (2011).doi:10.1038/nrg2957
5. Spitz, F. & Furlong, E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* (2012).doi:10.1038/nrg3207
6. Dang, C. MYC on the Path to Cancer. *Cell*(2012).doi:10.1016/j.cell.2012.03.003
7. Zilfou, J. & Lowe, S. Tumor Suppressive Functions of p53. *Cold Spring Harbor Perspectives in Biology*(2009).doi:10.1101/cshperspect.a001883
8. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–95 (2011).
9. Holliday, R., Pugh, J.E. DNA modification mechanisms and gene activity during development. *Science.* 187, 4174, 336-32 (1975).
10. Kouzarides, T. Chromatin modifications and their function. *Cell* 128, 693–705 (2007).
11. Schreiber, S. & Bernstein, B. Signaling Network Model of Chromatin. *Cell* **111**, 771–778 (2002).
12. Frye, S. The art of the chemical probe. *Nat Chem Biol* (2010).doi:10.1038/nchembio.296

Chapter 2

13. Arrowsmith, C.H., Bountra, C., Fish, P.V., Lee, K. & Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nat Rev Drug Discov* 11, 384-400 (2012).
14. Chi, P., Allis, C.D. & Wang, G.G. Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer* 10, 457-69 (2010).
15. Copeland, R.A., Solomon, M.E. & Richon, V.M. Protein methyltransferases as a target class for drug discovery. *Nat Rev Drug Discov* 8, 724-32 (2009).
16. Campagna-Slater, V. et al. Structural chemistry of the histone methyltransferases cofactor binding site. *J Chem Inf Model* 51, 612-23 (2011).
17. Richon, V.M. et al. Chemogenetic analysis of human protein methyltransferases. *Chem Biol Drug Des* 78, 199-210 (2011).
18. Min, J., Feng, Q., Li, Z., Zhang, Y. & Xu, R.M. Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase. *Cell* 112, 711-23 (2003).
19. Feng, Q. et al. Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr Biol* 12, 1052-8 (2002).
20. McGinty, R.K., Kim, J., Chatterjee, C., Roeder, R.G. & Muir, T.W. Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature* 453, 812-6 (2008).
21. Steger, D.J. et al. DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol Cell Biol* 28, 2825-39 (2008).
22. Bernt, K.M. et al. MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer Cell* 20, 66-78 (2011).
23. Muntean, A.G. & Hess, J.L. The Pathogenesis of Mixed-Lineage Leukemia. *Annu Rev Pathol* (2011).
24. Nguyen, A.T., Taranova, O., He, J. & Zhang, Y. DOT1L, the H3K79 methyltransferase, is required for MLL-AF9-mediated leukemogenesis. *Blood* 117, 6912-22 (2011).
25. Krivtsov, A.V. et al. H3K79 methylation profiles define murine and human MLL-AF4 leukemias. *Cancer Cell* 14, 355-68 (2008).
26. Krivtsov, A.V. & Armstrong, S.A. MLL translocations, histone modifications and leukaemia stem-cell development. *Nat Rev Cancer* 7, 823-33 (2007).

27. Grembecka, J. et al. Menin-MLL inhibitors reverse oncogenic activity of MLL fusion proteins in leukemia. *Nat Chem Biol* 8, 277-84 (2012).
28. Nguyen, A.T. & Zhang, Y. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev* 25, 1345-58 (2011).
29. Daigle, S.R. et al. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell* 20, 53-65 (2011).
30. Onder, T.T. et al. Chromatin-modifying enzymes as modulators of reprogramming. *Nature* 483, 598-602 (2012).
31. Frye, S.V. The art of the chemical probe. *Nat Chem Biol* 6, 159-161 (2010).
32. Yao, Y. et al. Selective inhibitors of histone methyltransferase DOT1L: design, synthesis, and crystallographic studies. *J Am Chem Soc* 133, 16746-9 (2011).
33. Copeland, R.A. Conformational adaptation in drug-target interactions and residence time. *Future Med Chem* 3, 1491-501 (2012).
34. Barabe, F., Kennedy, J.A., Hope, K.J. & Dick, J.E. Modeling the initiation and progression of human acute leukemia in mice. *Science* 316, 600-4 (2007).
35. Warner, J.K. et al. Direct evidence for cooperating genetic events in the leukemic transformation of normal human hematopoietic cells. *Leukemia* 19, 1794-805 (2005).
36. Wales, T.E. & Engen, J.R. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom Rev* 25, 158-70 (2006).
37. Mohan, M. et al. Linking H3K79 trimethylation to Wnt signaling through a novel Dot1-containing complex (DotCom). *Genes Dev* 24, 574-89 (2010).
38. Schapira, M. Structural Chemistry of Human SET Domain Protein Methyltransferases. *Curr Chem Genomics* 5, 85-94 (2011).
39. Troffer-Charlier, N., Cura, V., Hassenboehler, P., Moras, D. & Cavarelli, J. Functional insights from structures of coactivator-associated arginine methyltransferase 1 domains. *EMBO* J 26, 4391-401 (2007).
40. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr* 66, 133-44 (2010).

41. Otwinowski, z. & Minor, w. Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology* 276, 307-326 (1997).
42. Battye, T.G., Kontogiannis, L., Johnson, O., Powell, H.R. & Leslie, A.G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* 67, 271-81 (2011).
43. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50, 760-3 (1994).
44. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66, 486-501 (2010).
45. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53, 240-55 (1997).
46. Moriarty, N.W., Grosse-Kunstleve, R.W. & Adams, P.D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr D Biol Crystallogr* 65, 1074-80 (2009).
47. Schuttelkopf, A.W. & van Aalten, D.M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* 60, 1355-63 (2004).
48. Iacob, R.E., Zhang, J., Gray, N.S. & Engen, J.R. Allosteric interactions between the myristate- and ATP-site of the Abl kinase. *PLoS One* 6, e15929 (2011).
49. Wales, T.E., Fadgen, K.E., Gerhardt, G.C. & Engen, J.R. High-speed and high-resolution UPLC separation at zero degrees Celsius. *Anal Chem* 80, 6815-20 (2008).
50. Englander, S.W. & Kallenbach, N.R. Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q Rev Biophys* 16, 521-655 (1983).
51. Burkitt, W. & O'Connor, G. Assessment of the repeatability and reproducibility of hydrogen/deuterium exchange mass spectrometry measurements. *Rapid Commun Mass Spectrom* 22, 3893-901 (2008).
52. Wei, H. et al. Using hydrogen/deuterium exchange mass spectrometry to study conformational changes in granulocyte colony stimulating factor upon PEGylation. *J Am Soc Mass Spectrom* 23, 498-504 (2012).

53. Plumb, R.S. et al. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom* 20, 1989-94 (2006).
54. Kavan, D. & Man, P. MSTools-Web based application for visualization and presentation of HXMS data. *Internat J Mass Spectrom* 302, 52-58 (2011).
55. Jorgensen, W.L.C., Madura, J., Impey, R.W. & Klein, M.L. Comparison of simple potential functions for the simulation of liquid water. *J Chem Phys*. 79, 926-935 (1983).
56. Wang, J., Wang, W., Kollman, P.A. & Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25, 247-60 (2006).
57. Bayly, C.A., Cieplak, P., Cornell, W.D. & Kollman, P.A. A well behaved elestrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem* 97, 10269-80 (1993).
58. Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23, 327-41 (1977).

Chapter 3

1. Krivtsov, A., and Armstrong, S. MLL translocations, histone modifications and leukaemia stem-cell development. *Nature Reviews Cancer* 7, 823-833 (2007).
2. Goodell, M. Epigenetics in hematology: introducing a collection of reviews. *Blood* 121, 3059-3060 (2013).
3. Neff, T., and Armstrong, S. A. Recent progress toward epigenetic therapies: the example of mixed lineage leukemia. *Blood* 121, 4847-4853 (2013).
4. Steger, D., Lefterova, M., Ying, L., Stonestrom, A., Schupp, M., Zhuo, D., Vakoc, A., Kim, J.-E., Chen, J., Lazar, M., Blobel, G., and Vakoc, C. DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol. Cell. Biol*. 28, 2825-2839 (2008).
5. Guenther, M., Lawton, L., Rozovskaia, T., Frampton, G., Levine, S., Volkert, T., Croce, C., Nakamura, T., Canaani, E., and Young, R. Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. *Genes Dev*. 22, 3403-3408 (2008).
6. Bernt, K., Zhu, N., Sinha, A., Vempati, S., Faber, J., Krivtsov, A., Feng, Z., Punt, N., Daigle, A., Bullinger, L., Pollock, R., Richon, V., Kung, A., and Armstrong, S. MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer Cell* 20, 66-78 (2011).
7. Bitoun, E., Oliver, P., and Davies, K. The mixed-lineage leukemia fusion partner AF4 stimulates RNA polymerase II transcriptional elongation and mediates coordinated chromatin remodeling. *Hum. Mol. Genet.* 16, 92-106 (2007).
8. Krivtsov, A., Feng, Z., Lemieux, M., Faber, J., Vempati, S., Sinha, A., Xia, X., Jesneck, J., Bracken, A., Silverman, L., Kutok, J., Kung, A., and Armstrong, S. H3K79 methylation profiles define murine and human MLL-AF4 leukemias. *Cancer Cell* 14, 355-368 (2008).
9. Okada, Y., Feng, Q., Lin, Y., Jiang, Q., Li, Y., Coffield, V., Su, L., Xu, G., and Zhang, Y. hDOT1L links histone methylation to leukemogenesis. *Cell* 121, 167-178 (2005).
10. Deshpande, A., Chen, L., Fazio, M., Sinha, A., Bernt, K., Banka, D., Dias, S., Chang, J., Olhava, E., Daigle, S., Richon, V., Pollock, R., and Armstrong, S. Leukemic transformation by the MLL-AF6 fusion oncogene requires the H3K79 methyltransferase Dot1l. *Blood* 121, 2533-2541 (2013).

11. Jo, S. Y., Granowicz, E. M., Maillard, I., Thomas, D., and Hess, J. L. Requirement for Dot1l in murine postnatal hematopoiesis and leukemogenesis by MLL translocation. *Blood* 117, 4759-4768 (2011).
12. Nguyen, A. T., Taranova, O., He, J., and Zhang, Y. DOT1L, the H3K79 methyltransferase, is required for MLL-AF9-mediated leukemogenesis. *Blood* 117, 6912-6922 (2011).
13. Daigle, S., Olhava, E., Therkelsen, C., Majer, C., Sneeringer, C., Song, J., Johnston, L., Scott, M., Smith, J., Xiao, Y., Jin, L., Kuntz, K., Chesworth, R., Moyer, M., Bernt, K., Tseng, J.-C., Kung, A., Armstrong, S., Copeland, R., Richon, V., and Pollock, R. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell* 20, 53-65 (2011).
14. Yu, W., Chory, E., Wernimont, A., Tempel, W., Scopton, A., Federation, A., Marineau, J., Qi, J., Barsyte-Lovejoy, D., Yi, J., Marcellus, R., Iacob, R., Engen, J., Griffin, C., Aman, A., Wienholds, E., Li, F., Pineda, J., Estiu, G., Shatseva, T., Hajian, T., Al-Awar, R., Dick, J., Vedadi, M., Brown, P., Arrowsmith, C., Bradner, J., and Schapira, M. Catalytic site remodelling of the DOT1L methyltransferase by selective inhibitors. *Nature Comm* 3, 1-11 (2012).
15. Daigle, S., Olhava, E., Therkelsen, C., Basavapathruni, A., Jin, L., Boriack-Sjodin, P., Allain, C., Klaus, C., Raimondi, A., Scott, M., Waters, N., Chesworth, R., Moyer, M., Copeland, R., Richon, V., and Pollock, R. Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. *Blood* 122, 1017-1025 (2013).
16. Deng, L., Zhang, L., Yao, Y., Wang, C., Redell, M., Dong, S., and Song, Y. Synthesis, Activity and Metabolic Stability of Non-Ribose Containing Inhibitors of Histone Methyltransferase DOT1L. *MedChemComm* 4, 822-826 (2013).
17. Yao, Y., Chen, P., Diao, J., Cheng, G., Deng, L., Anglin, J., Prasad, B. V., and Song, Y. Selective inhibitors of histone methyltransferase DOT1L: design, synthesis, and crystallographic studies. *J. Am. Chem. Soc.* 133, 16746-16749 (2011).
18. Anglin, J., Deng, L., Yao, Y., Cai, G., Liu, Z., Jiang, H., Cheng, G., Chen, P., Dong, S., and Song, Y. Synthesis and structure-activity relationship investigation of adenosine-containing inhibitors of histone methyltransferase DOT1L. *J. Med. Chem.* 55, 8066-8074 (2012).

19. Yu, W., Smil, D., Li, F., Tempel, W., Fedorov, O., Nguyen, K., Bolshan, Y., Al-Awar, R., Knapp, S., Arrowsmith, C., Vedadi, M., Brown, P., and Schapira, M. Bromo-deaza-SAH: a potent and selective DOT1L inhibitor. *Biorg. Med. Chem*. 21, 1787-1794 (2013).
20. McGinty, R. K., Köhn, M., Chatterjee, C., Chiang, K. P., Pratt, M. R., and Muir, T. W. Structure-activity analysis of semisynthetic nucleosomes: mechanistic insights into the stimulation of Dot1L by ubiquitylated histone H2B. *ACS Chem. Biol*. 4, 958-968 (2009).
21. Vegas, A., Bradner, J., Tang, W., McPherson, O., Greenberg, E., Koehler, A., and Schreiber, S. Fluorous-based small-molecule microarrays for the discovery of histone deacetylase inhibitors. *Angewandte Chemie* (International ed. in English) 46, 7960-7964 (2007).
22. Bradner, J., West, N., Grachan, M., Greenberg, E., Haggarty, S., Warnow, T., and Mazitschek, R. Chemical phylogenetics of histone deacetylases. *Nat. Chem. Biol.* 6, 238-243 (2010).

Chapter 4

1. Filippakopoulos, P., Qi, J., et al. Selective inhibition of BET bromodomains. *Nature* 468, 1067–73 (2010).
2. Delmore, J. E. et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* 146**,** 904–17 (2011).
3. Ott, C. J. *et al.* BET bromodomain inhibition targets both c-Myc and IL7R in high-risk acute lymphoblastic leukemia. *Blood* **120,** 2843–52 (2012).
4. Costa, D. *et al.* BET inhibition as a single or combined therapeutic approach in primary paediatric B-precursor acute lymphoblastic leukaemia. *Blood Cancer J*(2013).doi:10.1038/bcj.2013.24
5. Puissant, A. *et al.* Targeting MYCN in Neuroblastoma by BET Bromodomain Inhibition. *Cancer Discovery* (2013).doi:10.1158/2159-8290.CD-12-0418
6. Anand, P. *et al.* BET Bromodomains Mediate Transcriptional Pause Release in Heart Failure. *Cell* (2013). doi:10.1016/j.cell.2013.07.013
7. Brown, J. D. *et al.* NF-κB Directs Dynamic Super Enhancer Formation in Inflammation and Atherogenesis. *Molecular Cell* (2014).doi:10.1016/j.molcel.2014.08.024
8. Matzuk, M. *et al.* Small-Molecule Inhibition of BRDT for Male Contraception. *Cell* (2012). doi:10.1016/j.cell.2012.06.045
9. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–34 (2013).
10. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,**934–47 (2013).

Chapter 5

1. Di Noia, J.M., Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. *Annual review of biochemistry 76*, 1-22 (2007).
2. Victora, G.D., Nussenzweig, M.C. Germinal centers. *Annual review of immunology* 30, 429-457 (2012).
3. Chaudhuri, J., Basu, U., Zarrin, A., Yan, C., Franco, S., Perlot, T., Vuong, B., Wang, J., Phan, R.T., Datta, A.*, et al*. Evolution of the immunoglobulin heavy chain class switch recombination mechanism. *Advances in immunology 94*, 157-214 (2007).
4. Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102, 553-563 (2000).
5. Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N., Forveille, M., Dufourcq-Labelouse, R., Gennery, A.*, et al*. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* 102, 565-575 (2000).
6. Hackney, J.A., Misaghi, S., Senger, K., Garris, C., Sun, Y., Lorenzo, M.N., and Zarrin, A.A. DNA targets of AID evolutionary link between antibody somatic hypermutation and class switch recombination. *Advances in immunology* 101, 163-189 (2009).
7. Alt, F.W., Zhang, Y., Meng, F.L., Guo, C., and Schwer, B. Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* 152, 417-429 (2013).
8. Pavri, R., Nussenzweig, M.C. AID targeting in antibody diversity. *Advances in immunology* 110, 1-26 (2011).
9. Storb, U. Why does somatic hypermutation by AID require transcription of its target genes? *Advances in immunology* 122, 253-277 (2014).
10. Liu, M., and Schatz, D.G. Balancing AID and DNA repair during somatic hypermutation. *Trends in immunology* 30, 173-181 (2009).
11. Stavnezer, J., Guikema, J.E., and Schrader, C.E. Mechanism and regulation of class switch recombination. *Annual review of immunology* 26, 261-292 (2008).

12. Yu, K., Chedin, F., Hsieh, C.L., Wilson, T.E., and Lieber, M.R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nature immunology* 4, 442-451 (2003).
13. Peters, A., Storb, U. Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. *Immunity* 4, 57-65 (1996).
14. Pavri, R., Gazumyan, A., Jankovic, M., Di Virgilio, M., Klein, I., Ansarah-Sobrinho, C., Resch, W., Yamane, A., Reina San-Martin, B., Barreto, V., et al. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* 143, 122-133 (2010).
15. Rajagopal, D., Maul, R.W., Ghosh, A., Chakraborty, T., Khamlichi, A.A., Sen, R., and Gearhart, P.J. Immunoglobulin switch mu sequence causes RNA polymerase II accumulation and reduces dA hypermutation. *The Journal of experimental medicine* 206, 1237- 1244 (2009).
16. Wang, L., Wuerffel, R., Feldman, S., Khamlichi, A.A., and Kenter, A.L. S region sequence, RNA polymerase II, and histone modifications create chromatin accessibility during class switch recombination. *The Journal of experimental medicine* 206, 1817-1830 (2009).
17. Basu, U., Meng, F.L., Keim, C., Grinstein, V., Pefanis, E., Eccleston, J., Zhang, T., Myers, D., Wasserman, C.R., Wesemann, D.R., et al. (2011). The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell* 144, 353- 363 (2011).
18. Matthews, A.J., Zheng, S., DiMenna, L.J., and Chaudhuri, J. Regulation of immunoglobulin class-switch recombination: choreography of noncoding transcription, targeted DNA deamination, and long-range DNA repair. *Advances in immunology* 122, 1-57 (2014).
19. Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. *Nature* (2014).
20. Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R.S., Kuppers, R., and Dalla-Favera, R. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* 412, 341-346 (2001).
21. Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* 147, 107-119 (2011).
22. Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M.,

Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* 147, 95-106 (2011).

23. Liu, M., Duke, J.L., Richter, D.J., Vinuesa, C.G., Goodnow, C.C., Kleinstein, S.H., and Schatz, D.G. Two levels of protection for the B cell genome during somatic hypermutation. *Nature* 451, 841-845 (2008).
24. Robbiani, D.F., and Nussenzweig, M.C. Chromosome translocation, B cell lymphoma, and activation-induced cytidine deaminase. *Annual review of pathology* 8, 79-103 (2013).
25. Kuppers, R., and Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* 20, 5580-5594 (2001).
26. Yamane, A., Resch, W., Kuo, N., Kuchen, S., Li, Z., Sun, H.W., Robbiani, D.F., McBride, K., Nussenzweig, M.C., and Casellas, R. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nature immunology* 12, 62-69 (2011).
27. Duke, J.L., Liu, M., Yaari, G., Khalil, A.M., Tomayko, M.M., Shlomchik, M.J., Schatz, D.G., and Kleinstein, S.H. Multiple transcription factor binding sites predict AID targeting in non-Ig genes. *Journal of immunology* 190, 3878-3888 (2013).
28. Wu, X., Sharp, P.A. Divergent transcription: a driving force for new gene origination? *Cell* 155, 990-996 (2013)
29. Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews Genetics* 13, 720-731.
30. Core, L.J., Waterfall, J.J., and Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845-1848 (2008).
31. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474, 390-394 (2011).
32. Levine, M., Cattoglio, C., and Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* 157, 13-25 (2014).
33. Natoli, G., Andrau, J.C. Noncoding transcription at enhancers: general principles and functional

models. *Annual review of genetics* 46, 1-19 (2012).

34. Lam, M.T., Li, W., Rosenfeld, M.G., and Glass, C.K. Enhancer RNAs and regulated transcriptional programs. *Trends in biochemical sciences* 39, 170-182 (2014).
35. Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* 110, 2876-2881 (2013).
36. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27Ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America 107, 21931-21936.
37. 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. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307-319 (2013).
38. 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. Selective inhibition of tumor oncogenes by disruption of super- enhancers. *Cell* 153, 320-334 (2013).
39. Parker, S.C., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Program, N.C.S., et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* 110, 17921- 17926 (2013).
40. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. Super-enhancers in the control of cell identity and disease. *Cell* 155, 934-947 (2013).
41. Delmore, J.E., Issa, G.C., Lemieux, M.E., Rahl, P.B., Shi, J., Jacobs, H.M., Kastritis, E., Gilpatrick, T., Paranal, R.M., Qi, J., et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* 146, 904-917 (2011).
42. Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. Discovery and characterization of super-enhancer- associated

dependencies in diffuse large B cell lymphoma. *Cancer cell* 24, 777-790 (2013).

43. Gostissa, M., Yan, C.T., Bianco, J.M., Cogne, M., Pinaud, E., and Alt, F.W. Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3' regulatory region. *Nature* 462, 803-807 (2009).
44. Richter, K., Brar, S., Ray, M., Pisitkun, P., Bolland, S., Verkoczy, L., and Diaz, M. Speckled-like pattern in the germinal center (SLIP-GC), a nuclear GTPase expressed in activation-induced deaminase-expressing lymphomas and germinal center B cells. *The Journal of biological chemistry* 284, 30652-30661 (2009).
45. Basso, K., Dalla-Favera, R. BCL6: master regulator of the germinal center reaction and key oncogene in B cell lymphomagenesis. *Advances in immunology* 105, 193-210 (2010).
46. Liu, Y.J., Mason, D.Y., Johnson, G.D., Abbot, S., Gregory, C.D., Hardie, D.L., Gordon, J., and MacLennan, I.C. Germinal center cells express bcl-2 protein after activation by signals which prevent their entry into apoptosis. *European journal of immunology* 21, 1905- 1910 (1991).
47. Brodeur, P.H., and Riblet, R. The immunoglobulin heavy chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes. *European journal of immunology* 14, 922-930 (1984).
48. Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148, 908-921 (2012).
49. Frock R.L., Hu J., Meyers R.M., Ho Y., Kii E., Alt F.W. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nature biotechnology* 33 179-186 (2015).
50. Hu, J., Tepsuporn, S., Meyers, R.M., Gostissa, M., and Alt, F.W. Developmental propagation of V(D)J recombination-associated DNA breaks and translocations in mature B cells via dicentric chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* Advance online (2014).
51. Kieffer-Kwon, K.R., Tang, Z., Mathe, E., Qian, J., Sung, M.H., Li, G., Resch, W., Baek, S., Pruett, N., Grontved, L., et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* 155, 1507-1520 (2013).
52. Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel

small RNAs. *Nature reviews Genetics* 10, 833-844 (2009).

53. Ward, D.F., and Murray, N.E. Convergent transcription in bacteriophage lambda: interference with gene expression. *Journal of molecular biology* 133, 249-266 (1979).
54. Gullerova, M., Proudfoot, N.J. Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. *Nature structural & molecular biology* 19, 1193-1201 (2012).
55. Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. RNA polymerase II collision interrupts convergent transcription. *Molecular cell* 48, 365-374 (2012).
56. Migliazza, A., Martinotti, S., Chen, W., Fusco, C., Ye, B.H., Knowles, D.M., Offit, K., Chaganti, R.S., and Dalla-Favera, R. Frequent somatic hypermutation of the 5' noncoding region of the BCL6 gene in B-cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America* 92, 12520-12524 (1995).
57. Shen, H.M., Peters, A., Baron, B., Zhu, X., and Storb, U. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science* 280, 1750-1752 (1998).
58. Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.G., Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K., et al. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* 139, 1069-1083 (2009).
59. Marusawa, H., Takai, A., and Chiba, T. Role of activation-induced cytidine deaminase in inflammation-associated cancer development. *Advances in immunology* 111, 109- 141 (2011).
60. Cato, M.H., Yau, I.W., and Rickert, R.C. (2011). Magnetic-based purification of untouched mouse germinal center B cells for ex vivo manipulation and biochemical analysis. Nature protocols 6, 953-960.
61. Langmead, B., and Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 357-359 (2012)3.
62. Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* May 28;38(4):576-589 (2010).
63. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9, R137

(2008).

Chapter 6

1. Krebs, H A. 1940. The Citric Acid Cycle and the Szent-Györgyi Cycle in Pigeon Breast Muscle. *The Biochemical Journal* 34 (5): 775–79.
2. Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, et al. 2008. KEGG for Linking Genomes to Life and the Environment. *Nucleic Acids Research* 36 (Database issue): D480–84. doi:10.1093/nar/gkm882.
3. Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2012. KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets. *Nucleic Acids Research* 40 (Database issue): D109–14. doi:10.1093/nar/gkr988.
4. Papin, Jason A, Tony Hunter, Bernhard O Palsson, and Shankar Subramaniam. 2005. Reconstruction of Cellular Signalling Networks and Analysis of Their Properties. *Nature Reviews. Molecular Cell Biology* 6 (2): 99–111. doi:10.1038/nrm1570.
5. Thiele, Ines, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, et al. 2013. A Community-Driven Global Reconstruction of Human Metabolism. *Nature Biotechnology* 31 (5): 419–25. doi:10.1038/nbt.2488.
6. Adelman, Karen, and John T Lis. 2012. Promoter-Proximal Pausing of RNA Polymerase II: Emerging Roles in Metazoans. *Nature Reviews: Genetics* 13 (10): 720–31. doi:10.1038/nrg3293.
7. Bonasio, Roberto, Shengjiang Tu, and Danny Reinberg. 2010. Molecular Signals of Epigenetic States. *Science* (New York, N.Y.) 330 (6004): 612–16. doi:10.1126/science.1191078.
8. Conaway, Ronald C, and Joan Weliky Conaway. 2011. Origins and Activity of the Mediator Complex. *Seminars in Cell & Developmental Biology* 22 (7): 729–34. doi:10.1016/j.semcdb.2011.07.021.
9. Roeder, Robert G. 2005. Transcriptional Regulation and the Role of Diverse Coactivators in Animal Cells. *FEBS Letters* 579 (4): 909–15. doi:10.1016/j.febslet.2004.12.007.
10. Spitz, François, and Eileen E M Furlong. 2012. Transcription Factors: From Enhancer Binding to Developmental Control. *Nature Reviews. Genetics* 13 (9): 613–26. doi:10.1038/nrg3207.

11. Zhou, Qiang, Tiandao Li, and David H Price. 2012. RNA Polymerase II Elongation Control. *Annual Review of Biochemistry* 81 (January): 119–43. doi:10.1146/annurev-biochem-052610-095910.
12. Graf, Thomas, and Tariq Enver. 2009. Forcing Cells to Change Lineages. *Nature* 462 (7273): 587–94. doi:10.1038/nature08533.
13. Lee, Tong Ihn, and Richard A Young. 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell* 152 (6): 1237–51. doi:10.1016/j.cell.2013.02.014.
14. Vierbuchen, Thomas, Austin Ostermeier, Zhiping P Pang, Yuko Kokubu, Thomas C Südhof, and Marius Wernig. 2010. Direct Conversion of Fibroblasts to Functional Neurons by Defined Factors. *Nature* 463 (7284): 1035–41. doi:10.1038/nature08797.
15. Buganim, Yosef, Dina A Faddah, and Rudolf Jaenisch. 2013. Mechanisms and Models of Somatic Cell Reprogramming. *Nature Reviews. Genetics* 14 (6): 427–39. doi:10.1038/nrg3473.
16. Morris, Samantha A, and George Q Daley. 2013. A Blueprint for Engineering Cell Fate: Current Technologies to Reprogram Cell Identity. Cell Research 23 (1): 33–48. doi:10.1038/cr.2013.1.
17. Yamanaka, Shinya, and Helen M Blau. 2010. Nuclear Reprogramming to a Pluripotent State by Three Approaches. Nature 465 (7299): 704–12. doi:10.1038/nature09229.
18. Hanna, Jacob H, Krishanu Saha, and Rudolf Jaenisch. 2010. Pluripotency and Cellular Reprogramming: Facts, Hypotheses, Unresolved Issues. Cell 143 (4): 508–25. doi:10.1016/j.cell.2010.10.008.
19. Stadtfeld, Matthias, and Konrad Hochedlinger. 2010. Induced Pluripotency: History, Mechanisms, and Applications. Genes & Development 24 (20): 2239–63. doi:10.1101/gad.1963910.
20. Young, Richard A. 2011. Control of the Embryonic Stem Cell State. Cell 144 (6): 940–54. doi:10.1016/j.cell.2011.01.032.
21. Boyer, Laurie A, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, et al. 2005. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell 122 (6): 947–56. doi:10.1016/j.cell.2005.08.020.
22. Odom, Duncan T, Robin D Dowell, Elizabeth S Jacobsen, Lena Nekludova, P Alexander Rolfe, Timothy W Danford, David K Gifford, Ernest Fraenkel, Graeme I Bell, and Richard A Young. 2006.

Core Transcriptional Regulatory Circuitry in Human Hepatocytes. Molecular Systems Biology 2 (January): 2006.0017. doi:10.1038/msb4100059.

23. Odom, Duncan T, Nora Zizlsperger, D Benjamin Gordon, George W Bell, Nicola J Rinaldi, Heather L Murray, Tom L Volkert, et al. 2004. Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science (New York, N.Y.) 303 (5662): 1378–81. doi:10.1126/science.1089769.
24. Sanda, Takaomi, Lee N Lawton, M Inmaculada Barrasa, Zi Peng Fan, Holger Kohlhammer, Alejandro Gutierrez, Wenxue Ma, et al. 2012. Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T Cell Acute Lymphoblastic Leukemia. Cancer Cell 22 (2): 209–21. doi:10.1016/j.ccr.2012.06.007.
25. Ito, T, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. 2001. A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome. Proceedings of the National Academy of Sciences of the United States of America 98 (8): 4569–74. doi:10.1073/pnas.061034498.
26. Stuart, Joshua M, Eran Segal, Daphne Koller, and Stuart K Kim. 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (New York, N.Y.) 302 (5643): 249–55. doi:10.1126/science.1087447.
27. Stelzl, Ulrich, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, et al. 2005. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell 122 (6): 957–68. doi:10.1016/j.cell.2005.08.029.
28. Carro, Maria Stella, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, et al. 2010. The Transcriptional Network for Mesenchymal Transformation of Brain Tumours. Nature 463 (7279): 318–25. doi:10.1038/nature08712.
29. Gerstein, Mark B, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, et al. 2012. Architecture of the Human Regulatory Network Derived from ENCODE Data. Nature 489 (7414): 91–100. doi:10.1038/nature11245.
30. Neph, Shane, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. 2012. Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. Cell 150 (6): 1274–86. doi:10.1016/j.cell.2012.04.040.

31. Sandhu, Kuljeet Singh, Guoliang Li, Huay Mei Poh, Yu Ling Kelly Quek, Yee Yen Sia, Su Qin Peh, Fabianus Hendriyan Mulawadi, et al. 2012. Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. Cell Reports 2 (5): 1207–19. doi:10.1016/j.celrep.2012.09.022.

32. Yosef, Nir, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, et al. 2013. Dynamic Regulatory Network Controlling TH17 Cell Differentiation. Nature 496 (7446): 461–68. doi:10.1038/nature11981.

33. Galagan, James E, Kyle Minch, Matthew Peterson, Anna Lyubetskaya, Elham Azizi, Linsday Sweet, Antonio Gomes, et al. 2013. The Mycobacterium Tuberculosis Regulatory Network and Hypoxia. Nature 499 (7457): 178–83. doi:10.1038/nature12337.

34. Fairfax, Benjamin P, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, et al. 2014. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science (New York, N.Y.) 343 (6175): 1246949. doi:10.1126/science.1246949.

35. Kemmeren, Patrick, Katrin Sameith, Loes A L van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, et al. 2014. Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. Cell 157 (3): 740–52. doi:10.1016/j.cell.2014.02.054.

36. Morris, John H, Giselle M Knudsen, Erik Verschueren, Jeffrey R Johnson, Peter Cimermancic, Alexander L Greninger, and Alexander R Pico. 2014. Affinity Purification-Mass Spectrometry and Network Analysis to Understand Protein-Protein Interactions. Nature Protocols 9 (11): 2539–54. doi:10.1038/nprot.2014.164.

37. Suvà, Mario L, Esther Rheinbay, Shawn M Gillespie, Anoop P Patel, Hiroaki Wakimoto, Samuel D Rabkin, Nicolo Riggi, et al. 2014. Reconstructing and Reprogramming the Tumor-Propagating Potential of Glioblastoma Stem-like Cells. Cell 157 (3): 580–94. doi:10.1016/j.cell.2014.02.030.

38. Lee, Tong Ihn, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, et al. 2002. Transcriptional Regulatory Networks in Saccharomyces Cerevisiae. Science (New York, N.Y.) 298 (5594): 799–804. doi:10.1126/science.1075090.

39. Alon, U. 2003. Biological Networks: The Tinkerer as an Engineer. Science (New York, N.Y.) 301 (5641): 1866–67. doi:10.1126/science.1089072.

40. Milo, R, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. 2002. Network Motifs: Simple Building Blocks of Complex Networks. Science (New York, N.Y.) 298 (5594): 824–27. doi:10.1126/science.298.5594.824.

41. Mitra, Koyel, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. 2013. Integrative Approaches for Finding Modular Structure in Biological Networks. Nature Reviews. Genetics 14 (10): 719–32. doi:10.1038/nrg3552.

42. Lim, Linda Shushan, Yuin-Han Loh, Weiwei Zhang, Yixun Li, Xi Chen, Yinan Wang, Manjiri Bakre, Huck-Hui Ng, and Lawrence W Stanton. 2007. Zic3 Is Required for Maintenance of Pluripotency in Embryonic Stem Cells. Molecular Biology of the Cell 18 (4): 1348–58. doi:10.1091/mbc.E06-07-0624.

43. Bar-Joseph, Ziv, Georg K Gerber, Tong Ihn Lee, Nicola J Rinaldi, Jane Y Yoo, François Robert, D Benjamin Gordon, et al. 2003. Computational Discovery of Gene Modules and Regulatory Networks. Nature Biotechnology 21 (11): 1337–42. doi:10.1038/nbt890.

44. Dutkowski, Janusz, Michael Kramer, Michal A Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan, and Trey Ideker. 2013. A Gene Ontology Inferred from Molecular Networks. Nature Biotechnology 31 (1): 38–45. doi:10.1038/nbt.2463.

45. Whyte, Warren A, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong Ihn Lee, and Richard A Young. 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell 153 (2): 307–19. doi:10.1016/j.cell.2013.03.035.

46. Hnisz, Denes, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla a Sigova, Heather a Hoke, and Richard a Young. 2013. Super-Enhancers in the Control of Cell Identity and Disease. Cell 155 (4): 934–47. doi:10.1016/j.cell.2013.09.053.

47. Dowen, Jill M., Zi Peng Fan, Denes Hnisz, Gang Ren, Brian J. Abraham, Lyndon N. Zhang, Abraham S. Weintraub, et al. 2014. Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell 159 (2): 374–87. doi:10.1016/j.cell.2014.09.030.

48. Chen, X, V B Vega, and H-H Ng. 2008. Transcriptional Regulatory Networks in Embryonic Stem Cells. Cold Spring Harbor Symposia on Quantitative Biology 73 (January): 203–9. doi:10.1101/sqb.2008.73.026.

49. Chew, Joon-Lin, Yuin-Han Loh, Wensheng Zhang, Xi Chen, Wai-Leong Tam, Leng-Siew Yeap, Pin Li, et al. 2005. Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells. Molecular and Cellular Biology 25 (14): 6031–46. doi:10.1128/MCB.25.14.6031-6046.2005.

50. Matoba, Ryo, Hitoshi Niwa, Shinji Masui, Satoshi Ohtsuka, Mark G Carter, Alexei A Sharov, and Minoru S H Ko. 2006. Dissecting Oct3/4-Regulated Gene Networks in Embryonic Stem Cells by Expression Profiling. PloS One 1 (January): e26. doi:10.1371/journal.pone.0000026.

51. Mathelier, Anthony, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, et al. 2014. JASPAR 2014: An Extensively Expanded and Updated Open-Access Database of Transcription Factor Binding Profiles. Nucleic Acids Research 42 (Database issue): D142–47. doi:10.1093/nar/gkt997.

52. Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. DNA-Binding Specificities of Human Transcription Factors. Cell 152 (1-2): 327–39. doi:10.1016/j.cell.2012.12.009.

53. Berger, Michael F, Gwenael Badis, Andrew R Gehrke, Shaheynoor Talukder, Anthony A Philippakis, Lourdes Peña-Castillo, Trevis M Alleyne, et al. 2008. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell 133 (7): 1266–76. doi:10.1016/j.cell.2008.05.024.

54. Robasky, Kimberly, and Martha L Bulyk. 2011. UniPROBE, Update 2011: Expanded Content and Search Tools in the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions. Nucleic Acids Research 39 (Database issue): D124–28. doi:10.1093/nar/gkq992.

55. Wei, Gong-Hong, Gwenael Badis, Michael F Berger, Teemu Kivioja, Kimmo Palin, Martin Enge, Martin Bonke, et al. 2010. Genome-Wide Analysis of ETS-Family DNA-Binding in Vitro and in Vivo. The EMBO Journal 29 (13): 2147–60. doi:10.1038/emboj.2010.106.

56. Matys, V, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, et al. 2006. TRANSFAC and Its Module TRANSCompel: Transcriptional Gene Regulation in Eukaryotes. Nucleic Acids Research 34 (Database issue): D108–10. doi:10.1093/nar/gkj143.

57. Tomioka, Mizuho, Masazumi Nishimoto, Satoru Miyagi, Tomoko Katayanagi, Nobutaka Fukui, Hitoshi Niwa, Masami Muramatsu, and Akihiko Okuda. 2002. Identification of Sox-2 Regulatory Region Which Is under the Control of Oct-3/4-Sox-2 Complex. Nucleic Acids Research 30 (14): 3202–13.

58. Rodda, David J, Joon-Lin Chew, Leng-Hiong Lim, Yuin-Han Loh, Bei Wang, Huck-Hui Ng, and Paul Robson. 2005. Transcriptional Regulation of Nanog by OCT4 and SOX2. The Journal of Biological Chemistry 280 (26): 24731–37. doi:10.1074/jbc.M502573200.

59. Okumura-Nakanishi, Sayaka, Motoki Saito, Hitoshi Niwa, and Fuyuki Ishikawa. 2005. Oct-3/4 and Sox2 Regulate Oct-3/4 Gene in Embryonic Stem Cells. The Journal of Biological Chemistry 280 (7): 5307–17. doi:10.1074/jbc.M410015200.

60. Loh, Yuin-Han, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, et al. 2006. The Oct4 and Nanog Transcription Network Regulates Pluripotency in Mouse Embryonic Stem Cells. Nature Genetics 38 (4): 431–40. doi:10.1038/ng1760.

61. Navarro, Pablo, Nicola Festuccia, Douglas Colby, Alessia Gagliardi, Nicholas P Mullin, Wensheng Zhang, Violetta Karwacki-Neisius, et al. 2012. OCT4/SOX2-Independent Nanog Autorepression Modulates Heterogeneous Nanog Gene Expression in Mouse ES Cells. The EMBO Journal 31 (24): 4547–62. doi:10.1038/emboj.2012.321.

62. Lien, Ching-Ling, John McAnally, James A Richardson, and Eric N Olson. 2002. Cardiac-Specific Activity of an Nkx2-5 Enhancer Requires an Evolutionarily Conserved Smad Binding Site. Developmental Biology 244 (2): 257–66. doi:10.1006/dbio.2002.0603.

63. Novershtern, Noa, Aravind Subramanian, Lee N Lawton, Raymond H Mak, W Nicholas Haining, Marie E McConkey, Naomi Habib, et al. 2011. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell 144 (2): 296–309. doi:10.1016/j.cell.2011.01.004.

64. Orkin, S H, J Wang, J Kim, J Chu, S Rao, T W Theunissen, X Shen, and D N Levasseur. 2008. The Transcriptional Network Controlling Pluripotency in ES Cells. Cold Spring Harbor Symposia on Quantitative Biology 73 (January): 195–202. doi:10.1101/sqb.2008.72.001.

65. Silva, Jose, and Austin Smith. 2008. Capturing Pluripotency. Cell 132 (4): 532–36. doi:10.1016/j.cell.2008.02.006.

66. Jaenisch, Rudolf, and Richard Young. 2008. Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. Cell 132 (4): 567–82. doi:10.1016/j.cell.2008.01.015.

67. Chambers, Ian, and Simon R Tomlinson. 2009. The Transcriptional Foundation of Pluripotency. Development (Cambridge, England) 136 (14): 2311–22. doi:10.1242/dev.024398.

68. Ng, Huck-Hui, and M Azim Surani. 2011. The Transcriptional and Signalling Networks of Pluripotency. Nature Cell Biology 13 (5): 490–96. doi:10.1038/ncb0511-490.

69. Huang, Xin, and Jianlong Wang. 2014. The Extended Pluripotency Protein Interactome and Its Links to Reprogramming. Current Opinion in Genetics & Development 28C (August): 16–24. doi:10.1016/j.gde.2014.08.003.

70. Nichols, J, B Zevnik, K Anastassiadis, H Niwa, D Klewe-Nebenius, I Chambers, H Schöler, and A Smith. 1998. Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. Cell 95 (3): 379–91. http://www.ncbi.nlm.nih.gov/pubmed/9814708.

71. Niwa, H, J Miyazaki, and A G Smith. 2000. Quantitative Expression of Oct-3/4 Defines Differentiation, Dedifferentiation or Self-Renewal of ES Cells. Nature Genetics 24 (4): 372–76. doi:10.1038/74199.

72. Avilion, Ariel A, Silvia K Nicolis, Larysa H Pevny, Lidia Perez, Nigel Vivian, and Robin Lovell-Badge. 2003. Multipotent Cell Lineages in Early Mouse Development Depend on SOX2 Function. Genes & Development 17 (1): 126–40. doi:10.1101/gad.224503.

73. Masui, Shinji, Yuhki Nakatake, Yayoi Toyooka, Daisuke Shimosato, Rika Yagi, Kazue Takahashi, Hitoshi Okochi, et al. 2007. Pluripotency Governed by Sox2 via Regulation of Oct3/4 Expression in Mouse Embryonic Stem Cells. Nature Cell Biology 9 (6): 625–35. doi:10.1038/ncb1589.

74. Chambers, Ian, Douglas Colby, Morag Robertson, Jennifer Nichols, Sonia Lee, Susan Tweedie, and Austin Smith. 2003. Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. Cell 113 (5): 643–55. http://www.ncbi.nlm.nih.gov/pubmed/12787505.

75. Mitsui, Kaoru, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. 2003. The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. Cell 113 (5): 631–42. http://www.ncbi.nlm.nih.gov/pubmed/12787504.

76. Silva, Jose, Jennifer Nichols, Thorold W Theunissen, Ge Guo, Anouk L van Oosten, Ornella Barrandon, Jason Wray, Shinya Yamanaka, Ian Chambers, and Austin Smith. 2009. Nanog Is the Gateway to the Pluripotent Ground State. Cell 138 (4): 722–37. doi:10.1016/j.cell.2009.07.039.

77. Theunissen, Thorold W, Anouk L van Oosten, Gonçalo Castelo-Branco, John Hall, Austin Smith, and José C R Silva. 2011. Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. Current Biology : CB 21 (1): 65–71. doi:10.1016/j.cub.2010.11.074.

78. Catena, Raffaella, Cecilia Tiveron, Antonella Ronchi, Silvia Porta, Anna Ferri, Laura Tatangelo, Maurizio Cavallaro, et al. 2004. Conserved POU Binding DNA Sites in the Sox2 Upstream Enhancer Regulate Gene Expression in Embryonic and Neural Stem Cells. The Journal of Biological Chemistry 279 (40): 41846–57. doi:10.1074/jbc.M405514200.

79. Kuroda, Takao, Masako Tada, Hiroshi Kubota, Hironobu Kimura, Shin-ya Hatano, Hirofumi Suemori, Norio Nakatsuji, and Takashi Tada. 2005. Octamer and Sox Elements Are Required for Transcriptional Cis Regulation of Nanog Gene Expression. Molecular and Cellular Biology 25 (6): 2475–85. doi:10.1128/MCB.25.6.2475-2485.2005.

80. Ichida, Justin K, Joel Blanchard, Kelvin Lam, Esther Y Son, Julia E Chung, Dieter Egli, Kyle M Loh, et al. 2009. A Small-Molecule Inhibitor of Tgf-Beta Signaling Replaces sox2 in Reprogramming by Inducing Nanog. Cell Stem Cell 5 (5): 491–503. doi:10.1016/j.stem.2009.09.012.

81. Zhang X, Yalcin S, Lee D-FF, Yeh T-YJY, Lee S-MM, Su J, Mungamuri SK, Rimmelé P, Kennedy M, Sellers R, Landthaler M, Tuschl T, Chi N-WW, Lemischka I, Keller G & Ghaffari S (2011) FOXO1 is an essential regulator of pluripotency in human embryonic stem cells. *Nat. Cell Biol.* 13, 1092–9.

82. Declercq, Jeroen, Preethi Sheshadri, Catherine M Verfaillie, and Anujith Kumar. 2013. Zic3 Enhances the Generation of Mouse Induced Pluripotent Stem Cells. Stem Cells and Development 22 (14): 2017–25. doi:10.1089/scd.2012.0651.

83. Guo, Ge, and Austin Smith. 2010. A Genome-Wide Screen in EpiSCs Identifies Nr5a Nuclear Receptors as Potent Inducers of Ground State Pluripotency. Development (Cambridge, England) 137 (19): 3185–92. doi:10.1242/dev.052753.

84. Barnea, E, and Y Bergman. 2000. Synergy of SF1 and RAR in Activation of Oct-3/4 Promoter. The Journal of Biological Chemistry 275 (9): 6608–19. http://www.ncbi.nlm.nih.gov/pubmed/10692469.

85. Schoorlemmer, J, L Jonk, S Sanbing, A van Puijenbroek, A Feijen, and W Kruijer. 1995. Regulation of Oct-4 Gene Expression during Differentiation of EC Cells. Molecular Biology Reports 21 (3): 129–40. http://www.ncbi.nlm.nih.gov/pubmed/8832901.

86. Sylvester, I, and H R Schöler. 1994. Regulation of the Oct-4 Gene by Nuclear Receptors. Nucleic Acids Research 22 (6): 901–11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=307908&tool=pmcentrez&rendertype=abstract.

87. Yang, Heung-Mo, Hyun-Jin Do, Dong-Ku Kim, Jin-Ki Park, Won-Kyong Chang, Hyung-Min Chung, Sang-Yun Choi, and Jae-Hwan Kim. 2007. Transcriptional Regulation of Human Oct4 by Steroidogenic Factor-1. Journal of Cellular Biochemistry 101 (5): 1198–1209. doi:10.1002/jcb.21244.

88. Cheasley, Dane, Lloyd Pereira, Sally Lightowler, Elizabeth Vincan, Jordane Malaterre, and Robert G Ramsay. 2011. Myb Controls Intestinal Stem Cell Genes and Self-Renewal. Stem Cells (Dayton, Ohio) 29 (12): 2042–50. doi:10.1002/stem.761.

89. Zuber, Johannes, Amy R Rappaport, Weijun Luo, Eric Wang, Chong Chen, Angelina V Vaseva, Junwei Shi, et al. 2011. An Integrated Approach to Dissecting Oncogene Addiction Implicates a Myb-Coordinated Self-Renewal Program as Essential for Leukemia Maintenance. Genes & Development 25 (15): 1628–40. doi:10.1101/gad.17269211.

90. White, J R, and K Weston. 2000. Myb Is Required for Self-Renewal in a Model System of Early Hematopoiesis. Oncogene 19 (9): 1196–1205. doi:10.1038/sj.onc.1203394.

91. Lieu, Yen K, and E Premkumar Reddy. 2009. Conditional c-Myb Knockout in Adult Hematopoietic Stem Cells Leads to Loss of Self-Renewal due to Impaired Proliferation and Accelerated Differentiation.

92. Doulatov, Sergei, Linda T Vo, Stephanie S Chou, Peter G Kim, Natasha Arora, Hu Li, Brandon K Hadland, et al. 2013. Induction of Multipotential Hematopoietic Progenitors from Human Pluripotent Stem Cells via Respecification of Lineage-Restricted Precursors. Cell Stem Cell 13 (4): 459–70. doi:10.1016/j.stem.2013.09.002.

93. Kuzmichev, Andrey N, Suel-Kee Kim, Ana C D'Alessio, Josh G Chenoweth, Ina M Wittko, Loraine Campanati, and Ronald D McKay. 2012. Sox2 Acts through Sox21 to Regulate Transcription in Pluripotent and Differentiated Cells. Current Biology : CB 22 (18): 1705–10. doi:10.1016/j.cub.2012.07.013.

94. Lewis, Benjamin P, Christopher B Burge, and David P Bartel. 2005. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are microRNA Targets. Cell 120 (1): 15–20. doi:10.1016/j.cell.2004.12.035.

95. Sinkkonen, Lasse, Tabea Hugenschmidt, Philipp Berninger, Dimos Gaidatzis, Fabio Mohn, Caroline G Artus-Revel, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. 2008. MicroRNAs Control de Novo DNA Methylation through Regulation of Transcriptional Repressors in Mouse Embryonic Stem Cells.

96. Marson, Alexander, Stuart S Levine, Megan F Cole, Garrett M Frampton, Tobias Brambrink, Sarah Johnstone, Matthew G Guenther, et al. 2008. Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells. Cell 134 (3): 521–33. doi:10.1016/j.cell.2008.07.020.

97. Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B., and Srivastava, D. (2010). Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. Cell.

98. Hale, M., Kagami, H., Shi, L., Holland, A., Elsässer, H.-P., Hammer, R., and MacDonald, R. (2005). The homeodomain protein PDX1 is required at mid-pancreatic development for the formation of the exocrine pancreas. Developmental Biology

99. Miyagi, S., Masui, S., Niwa, H., Saito, T., Shimazaki, T., Okano, H., Nishimoto, M., Muramatsu, M., Iwama, A., and Okuda, A. (2008). Consequence of the loss of Sox2 in the developing brain of the mouse. FEBS Letters.

100. Parker, Stephen C J, Michael L Stitzel, D Leland Taylor, Jose Miguel Orozco, Michael R Erdos, Jennifer A Akiyama, Kelly Lammerts van Bueren, et al. 2013. Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants. Proceedings of the National Academy of Sciences of the United States of America 110 (44): 17921–26. doi:10.1073/pnas.13170

101. Lovén, Jakob, Heather A Hoke, Charles Y Lin, Ashley Lau, David A Orlando, Christopher R Vakoc, James E Bradner, Tong Ihn Lee, and Richard A Young. 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell 153 (2): 320–34. doi:10.1016/j.cell.2013.03.036.

102. Loh, Kyle M, and Bing Lim. 2011. A Precarious Balance: Pluripotency Factors as Lineage Specifiers. Cell Stem Cell 8 (4): 363–69. doi:10.1016/j.stem.2011.03.013.

103. Herranz, Daniel, Alberto Ambesi-Impiombato, Teresa Palomero, Stephanie A Schnell, Laura Belver, Agnieszka A Wendorff, Luyao Xu, et al. 2014. A NOTCH1-Driven MYC Enhancer Promotes T Cell Development, Transformation and Acute Lymphoblastic Leukemia. Nature Medicine 20 (10): 1130–37. doi:10.1038/nm.3665.

104. Mansour, M. R., B. J. Abraham, L. Anders, A. Berezovskaya, A. Gutierrez, A. D. Durbin, J. Etchin, et al. 2014. An Oncogenic Super-Enhancer Formed through Somatic Mutation of a Noncoding Intergenic Element. Science, November. doi:10.1126/science.1259037.

105. Maurano, Matthew T, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (New York, N.Y.) 337 (6099): 1190–95. doi:10.1126/science.1222794.

106. Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2014. Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants. Nature, October. doi:10.1038/nature13835.

Chapter 7

1. Northcott, P. A., Dubuc, A. M., Pfister, S. & Taylor, M. D. Molecular subgroups of medulloblastoma. *Expert review of neurotherapeutics* 12, 871-884, doi:10.1586/ern.12.66 (2012).
2. Northcott, P. A., Korshunov, A., Pfister, S. M. & Taylor, M. D. The clinical implications of medulloblastoma subgroups. *Nature reviews*. Neurology 8, 340-351, doi:10.1038/nrneurol.2012.78 (2012).
3. Jones, D. T. et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature* 488, 100-105, doi:10.1038/nature11284 (2012).
4. Northcott, P. A. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* 488, 49-56, doi:10.1038/nature11327 (2012).
5. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* 511, 428-434, doi:10.1038/nature13379 (2014).
6. Gibson, P. et al. Subtypes of medulloblastoma have distinct developmental origins. *Nature* 468, 1095-1099, doi:10.1038/nature09587 (2010).
7. Grammel, D. et al. Sonic hedgehog-associated medulloblastoma arising from the cochlear nuclei of the brainstem. *Acta neuropathologica* 123, 601-614, doi:10.1007/s00401-012-0961-0 (2012).
8. Yang, Z. J. et al. Medulloblastoma can be initiated by deletion of Patched in lineage-restricted progenitors or stem cells. *Cancer cell* 14, 135-145, doi:10.1016/j.ccr.2008.07.003 (2008).
9. Schuller, U. et al. Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. *Cancer cell* 14, 123-134, doi:10.1016/j.ccr.2008.07.005 (2008).
10. Northcott, P. A. et al. Medulloblastomics: the end of the beginning. *Nature reviews*. Cancer 12, 818-834, doi:10.1038/nrc3410 (2012).
11. Kool, M. et al. Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. *Cancer cell* 25, 393-405, doi:10.1016/j.ccr.2014.02.004 (2014).
12. Hovestadt, V. et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* 510, 537-541, doi:10.1038/nature13268 (2014).

13. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews*. Genetics 15, 272-286, doi:10.1038/nrg3682 (2014).
14. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74, doi:10.1038/nature11247 (2012).
15. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75-82, doi:10.1038/nature11232 (2012).
16. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317-330, doi:10.1038/nature14248 (2015).
17. Chapuy, B. et al. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer cell* 24, 777-790, doi:10.1016/j.ccr.2013.11.003 (2013).
18. Anand, P. et al. BET bromodomains mediate transcriptional pause release in heart failure. *Cell* 154, 569-582, doi:10.1016/j.cell.2013.07.013 (2013).
19. Loven, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* 153, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).
20. Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182-187, doi:10.1038/nature09033 (2010).
21. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
22. Thompson, M. C. et al. Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 24, 1924-1931, doi:10.1200/JCO.2005.04.4974 (2006).
23. Cho, Y. J. et al. Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 29, 1424-1430, doi:10.1200/JCO.2010.28.5148 (2011).
24. Northcott, P. A. et al. Medulloblastoma comprises four distinct molecular variants. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 29, 1408-1414, doi:10.1200/JCO.2009.27.4324 (2011).

25. Kool, M. et al. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PloS one* 3, e3088, doi:10.1371/journal.pone.0003088 (2008).
26. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290-294, doi:10.1038/nature12644 (2013).
27. Pope, B. D. et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* 515, 402-405, doi:10.1038/nature13986 (2014).
28. Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159, 374-387, doi:10.1016/j.cell.2014.09.030 (2014).
29. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* 155, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).
30. Parker, S. C. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* 110, 17921-17926, doi:10.1073/pnas.1317023110 (2013).
31. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
32. Aruga, J. et al. Mouse Zic1 is involved in cerebellar development. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 18, 284-293 (1998).
33. Grinberg, I. et al. Heterozygous deletion of the linked genes ZIC1 and ZIC4 is involved in Dandy-Walker malformation. *Nature genetics* 36, 1053-1055, doi:10.1038/ng1420 (2004).
34. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* 462, 587-594, doi:10.1038/nature08533 (2009).
35. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237-1251, doi:10.1016/j.cell.2013.02.014 (2013).
36. Fink, A. J. et al. Development of the deep cerebellar nuclei: transcription factors and cell migration from the rhombic lip. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26, 3066-3076, doi:10.1523/JNEUROSCI.5203-05.2006 (2006).

37. Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* 518, 314-316, doi:10.1038/518314a (2015).
38. Skipper, M. et al. Presenting the epigenome roadmap. *Nature* 518, 313, doi:10.1038/518313a (2015).
39. Hovestadt, V. et al. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta neuropathologica* 125, 913-916, doi:10.1007/s00401-013-1126-5 (2013).
40. Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091-1093, doi:10.1093/bioinformatics/btp101 (2009).

Chapter 8

1. Ahlquist, R.P., A study of the adrenotropic receptors. *Am J Physiol*, 1948. 153(3): p. 586-600.
2. Lands, A.M., et al., Differentiation of receptor systems activated by sympathomimetic amines. *Nature*, 1967. 214(5088): p. 597-8.
3. Kola, I. and J. Landis, Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*, 2004. 3(8): p. 711-5.
4. Alon, U. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8,** 450–461 (2007).
5. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150,** 389–401 (2012).
6. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nat Biotechnol* (2012). doi:10.1038/nbt.1666
7. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**, 382–90 (2005).

8. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*(2012).doi:10.1038/nature11212