UNIVERSITY OF CALIFORNIA

IRVINE

Genome-wide analysis of NIPBL/cohesin binding in mouse and human cells: Implications for gene regulation and human disease

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Daniel Newkirk

Dissertation Committee:
Professor Kyoko Yokomori, Chair
Professor Xing Dai
Professor Bogi Anderson
Assistant Professor Ali Mortazavi
Associate Professor Xiaohui Xie

2015

Dedication

This dissertation is dedicated

to my family and many friends

who have encouraged me to pursue this dream

Table of Contents

Abbreviations

CdLS: Cornelia de Lange Syndrome

ChIP-seq: Chromatin immunoprecipitation couple to high-throughput sequencing

DEGs: differentially expressed genes

DMRs: differentially methylated regions

EM: expectation maximization

FSHD: Facioscapulohumeral Muscular Dystrophy

MEFs: mouse embryonic fibroblasts

MES: mouse embryonic stem cells

RNAPII: RNA polymerase II

RNA-seq: high-throughput sequencing of messenger RNA

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor Dr. Kyoko Yokomori for the opportunity to work on fascinating projects within the lab. She has worked tirelessly to help all of us in the lab to think critically, ask good questions, and find the best route to those answers; I am grateful for her mentorship! Dr. Xiaohui Xie, my co-advisor, has also been incredibly helpful in teaching me how to formulate problems, and design good approaches to answer those questions computationally. Whether in teaching me how to write out an algorithm mathematically, or in carefully critiquing the assumptions made in my statistical testing, he has been a wonderful teacher and mentor. The rest of my thesis committee, Dr. Xing Dai, Dr. Bogi Anderson, and Dr. Ali Mortazavi, have been incredibly helpful with good advice throughout the course of my research.

I am thankful for the rest of the Yokomori Lab as well! Dr. Yen-Yun Chen, and Dr. Richard Chien made the work done on the Cornelia de Lange Syndrome (CdLS) project possible, with Dr. Chien performing the ChIP-sequencing, and both aiding in the manual testing of specific regions. Dr. Chen performed the ChIP-sequencing for the FSHD muscular dystrophy project, and has fought through many of the difficulties in culturing the myoblasts in the process. Alex Ball and Dr. Xiangduo Kong have also been vital members of the lab, with critical instruction on cell culture, cloning techniques, and much beside. Dr. Weihua Zeng pioneered much of the work on the FSHD project that I've been able to help extend, and his insight has been instrumental. Everyone has been a joy to work with.

I would like to thank our collaborators at UCI, Dr. Arthur Lander and Dr. Anne Calof for developing and giving us access to the mouse model for CdLS. Dr. Shimako Kawauchi and Rosie Santos in the Calof lab provided us with the mouse tissue and helped us prepare the MEFs necessary for our study. Dr. Jacob Biesinger and Aniello Infante both helped to perform some of the bioinformatics analysis in on the CdLS project.

For the FSHD project, our collaborators Dr. Silvère van der Maarel and members of his lab including Jessica de Greef. Dr. Morena Mora and Dr. Rabi Tawil provided many of the primary myoblast lines used in the FSHD project. Dr. Ali Mortazavi has provided crucial help both in advice and sequencing for the project.

Finally, I would like to thank my family and friends, who have chosen to support me and care for me at every stage, and without whom none of this would be possible. My parents and sister are amazing! I'm grateful for my father, who has helped me more than words can describe. My girlfriend and best friend, Lily Kehoe, has also been an immense blessing to me, and encouraged me greatly.

My work on the projects in the following s was under support of the Bioinformatics Training Program grant from the Institute for Genomics and Bioinformatics T15LM07443, and by the Developmental Systems Biology training grant from the Center for Complex Biological Systems.

Chapter 2 of this dissertation was coauthored by Jake Biesinger, Alvin Chon, Kyoko Yokomori, and Xiaohui Xie. It is reprinted with permission from JOURNAL OF COMPUTATIONAL BIOLOGY, 21011, Volume 18, Issue 11, pp. 1495-1505, published by Mary Ann Liebert, Inc., New Rochelle, NY.

Curriculum Vitae

Daniel Newkirk

Education:

Bachelors of Science in Biochemistry: Biola University, 2003

Masters of Arts in Science and Religion: Biola University, 2009

Doctor of Philosophy in Biomedical Sciences: University of California, Irvine, 2014

Publications:

Zeng, W., Chien, R., **Newkirk, D**., Chen, Y.-Y., Xie, X., and Yokomori, K. *Identification and characterization of Scc2/NIPBL binding sites in mammalian cells*. (Manuscript in Preparation)

Newkirk, D. A.*, Chien, R.*, Chen, Y-Y.*, Zeng, W., Biesinger, J., Flowers, E., Infante, A., Kawauchi, S., Santos, R., Calof, A. L., Lander, A. D., Xie, X., and Yokomori, K., *The effect of Nipbl haploinsufficiency on genome-wide cohesin binding and target gene expression: modeling Cornelia de Lange Syndrome* (Manuscript Submitted) *equal contribution

Zeng W, Chen YY, **Newkirk DA**, Wu B, Balog J, Kong X, Ball AR Jr, Zanotti S, Tawil R, Hashimoto N, Mortazavi A, van der Maarel SM, Yokomori K. Genetic and epigenetic characteristics of FSHD-associated 4q and 10q D4Z4 that are distinct from non-4q/10q D4Z4 homologs. Hum Mutat. 2014 Aug;35(8):998-1010. doi: 10.1002/humu.22593. Epub 2014 Jun 24. PubMed PMID: 24838473.

Newkirk D, Biesinger J, Chon A, Yokomori K, Xie X. *AREM: aligning short reads from ChIP-sequencing by expectation maximization*. J Comput Biol. 2011 Nov; 18(11): 1495-505.

Kong X, Stephens J, Ball AR Jr, Heale JT, **Newkirk DA**, Berns MW, Yokomori K. *Condensin I recruitment to base damage-enriched DNA lesions is modulated by PARP1*. PLoS One. 2011; 6(8): e23548.

Chien R, Zeng W, Kawauchi S, Bender MA, Santos R, Gregson HC, Schmiesing JA, **Newkirk DA**, Kong X, Ball AR Jr, Calof AL, Lander AD, Groudine MT, Yokomori K. *Cohesin mediates chromatin interactions that regulate mammalian β-globin expression*. J Biol Chem. 2011 May 20; 286(20): 17870-8.

Nishimoto KP, **Newkirk D**, Hou S, Fruehauf J, Nelson EL. *Fluorescence activated cell sorting (FACS) using RNAlater to minimize RNA degradation and perturbation of mRNA expression from cells involved in initial host microbe interactions.* J Microbiol Methods, 2007. **70**(1): p. 205-8.

Awards and Honors:

CCBS Opportunity Award, 2011

Best Poster Award Day 1, NLM Bioinformatics Training Program Conference, 2011

CCBS Opportunity Award, 2013

Fellowships:

Bioinformatics Training Program Fellowship, Institute for Genomics and Bioinformatics, UCI 2010-2012

Trainee President, Bioinformatics Training Program, 2011-2012

Developmental Systems Biology Training Grant, Center for Complex Biological Systems, UCI, 2013

Meeting Presentations:

Posters:

- "Global analysis of cohesin-mediated gene regulation in a Cornelia de Lange Syndrome Mouse Model," Informatics Training Conference, National Library of Medicine; Bethesda, Maryland, 2011.

- "AREM: Aligning short reads using expectation maximization," High-throughput Sequencing to P4 Medicine at UCI, Institute for Genomics and Bioinformatics; UCI, 2011

Presentations:

- "AREM: Aligning short reads from ChIP-sequencing using expectation maximization," Artificial Intelligence/Machine Learning Seminar, Dept. of Computer Science; UCI, 2011

Papers presented at conferences:

- "AREM: Aligning short reads from ChIP-sequencing using expectation maximization," RECOMB; Vancouver, British Columbia, Canada, 2011

Teaching Experience:

Staff professor, Introduction to Public Speaking Laboratory (Fall 2004, 2 sections)

- Lectured on material in complement to what was presented in class
- Developed activities to give the students enjoyable ways to practice many types of speeches
- Aided students in recording their speeches to see themselves speak and understand ways they can improve while helping to ease the fear of public speaking

Assistant Parliamentary Debate Coach (2004-2009)

- Worked with novice students to learn the basics of debate
- Guided teams in how to coordinate and perform research on debate topics
- Coached and judged debate rounds at tournaments throughout the year
- Worked with the Head Coach and students to develop arguments and debate theory for the benefit of the team and the debate community

Teaching Assistant, Genetics (Fall 2010)

- Developed worksheets, activities, and materials to help students understand critical concepts in genetics
- Lead discussions to further develop material covered in lecture
- Provided feedback and suggestions to help the professors continue to develop the course
- Graded exams and course materials

Teaching Assistant, Genetics (Fall 2011)
Teaching Assistant, Genetics (Fall 2012)
Teaching Assistant, Genetics (Summer 2012)
Teaching Assistant, Developmental Biology (Summer 2012)

- Lead weekly discussions to further develop material covered in the course
- Developed problem sets to help students identify what they know and where they needed further study
- Graded exams and course materials

Teaching Assistant, Genetics (Summer 2013)
Teaching Assistant, Nutrition (Summer 2013)

- Graded exams and other class assignments
- Answered student questions and developed review sessions and activities

Teaching Assistant, Genetics (Summer 2014)
Teaching Assistant, Nutrition (Summer 2014)
Teaching Assistant (Head TA), Genetics (Fall 2014)

Abstract of the Dissertation

Genome-wide analysis of NIPBL/cohesin binding in mouse and human cells:

Implications for gene regulation and human disease

by

Daniel Newkirk

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2014

Professor Kyoko Yokomori, Chair

One of the most powerful tools to arrive in biology in the past decade is high-throughput sequencing, such as Illumina sequencing. These platforms have allowed an unparalleled look at protein-chromatin interactions, mRNA expression, chromatin topology, and a great deal more. Our lab has successfully used these tools to better understand the distribution of cohesin and Nipbl binding in mouse and human cells, with the aim of further clarifying how cohesin and Nipbl regulate gene expression and what genes they regulate. Moreover, we have been able to identify how this binding can change in disease, and correlate these changes with the corresponding gene expression changes taking place in vivo. Careful analysis of ChIP-seq data (chromatin immunoprecipitation coupled with sequencing) has indicated that cohesin binding decreases genome-wide in mouse embryonic fibroblasts (MEFs) derived from a mouse model for Cornelia de Lange Syndrome (CdLS). In fact, cohesin's role in gene activation is most susceptible to Nipbl haploinsufficiency. Moreover, we find that decreased cohesin binding is correlated with the gene expression changes taking place between

wildtype and mutant MEFs. Enhancer-promoter interactions, one mechanism by which cohesin can regulate gene expression, are decreased in the mutant MEFs. Our studies have helped characterize how Nipbl haploinsufficiency affects cohesin binding, and suggest how this effect on cohesin binding can affect gene expression in the context of CdLS.

Based on the increased severity of the disease phenotype of CdLS patients with mutations in *NIPBL* versus mutations in the cohesin subunits, it was postulated that NIPBL might have cohesin-independent functions in the cell. To examine this, we have used ChIP-seq in human cells to identify the global distribution of NIPBL-chromatin interactions. We found that, similar to cohesin, NIPBL is enriched at the promoter region. In contrast to other studies however, we found that NIPBL is present at sites also bound by cohesin and CTCF. While most NIPBL –bound regions are shared with cohesin, about 10% of these sites are free of cohesin and CTCF. Further examination of the cohesin-free sites show 273 genes where NIPBL is bound at the promoter, and could be direct genes targets. Of these, 73 are differentially expressed upon siRNA depletion of NIPBL, two of which were examined in detail to show that NIPBL normally represses the expression of these genes independently of cohesin, which is the first time any ability of NIPBL to repress gene expression has been shown. Taken together, our data indicate that mutations in NIPBL may indeed effect expression of NIPBL target genes, suggesting that this may explain in part the differences in disease severity in CdLS patients.

Over the past decade, our lab has examined the importance of intact heterochromatin on chromosome 4q in FSHD muscular dystrophy. While we have shown that the loss of cohesin and H3K9me3 present at the D4Z4 repeat array on chromosome 4q is characteristic of the disease, and may underlie expression changes seen in patient myoblasts in vivo, the global differences in heterochromatin in FSHD have not been previously studied. Therefore, we have used several sequencing techniques to identify the genome-wide changes to heterochromatin and gene expression in FSHD, with the intent of being able to identify disease-specific signatures and illuminate the interdependence of the two in disease.

Chapter 1

Introduction

1.1 The cohesin complex

The cohesin complex is an evolutionarily conserved, essential protein complex composed of the Structural Maintenance of Chromosomes proteins (SMCs) SMC1 and SMC3. It also contains the non-SMC subunits RAD21 and SA1 or SA2 (in the case of mammals; *Saccharomyces* only contain one homolog, Scc3). The cohesin complex is highly conserved in eukaryotes, and has homologs in bacteria [1]. Structurally, the cohesin complex forms a ring-shape, allowing it to encircle chromatin. Each SMC subunit is characterized by a long coiled-coil domain on either side of a central hinge domain, and an ATPase head domain; the protein folds in half at the hinge region, allowing the globular domains containing Walker A (N-terminal) and Walker B (C-terminal) motifs to form the ATPase head domain [2-4]. The SMC1 and SMC3 proteins then form a heterodimer through interactions between their hinge domains. The protein RAD21, which interacts with the head domains of SMC1 and SMC3, closes the ring while the SA subunits are bound to RAD21 (Figure 1.1).

Cohesin's canonical role is to mediate sister chromatid cohesion. It binds both sister chromatids along their entire length and holds the pair together. Separation of the sister chromatids takes place in two steps. In the first, more than 90% of the cohesin bound to chromosomes is released during prophase from the arms of the chromosomes. While it isn't entirely clear how cohesin is released from the chromosome arms, data shows that the phosphorylation of the RAD21 and SA1/SA2 subunits by Plk1 is required for separation, and is also dependent on Aurora B kinase and WAPL. Centromeric cohesin is protected from this phosphorylation by the Sgo1-PP2A complex. For the

second step, cohesin's RAD21 subunit is cleaved by the protease separase, and the remaining cohesin located at the centromere dissociates during anaphase. Dissociation of cohesin is also dependent on the acetylation of SMC3. Moreover, the dissociation of the cleaved form of RAD21 after cleavage by separase requires the deacetylation activity of HDAC8 [5].

Inactivation of cohesin or regulators of cohesin results in precocious sister chromatid cohesion, while depletion of WAPL, SGO1, PLK1, HDAC8, and Aurora B kinase results in unresolved chromatids (Figure 1.3). Sister chromatid cohesion defects can increase genome instability, and have been associated with some forms of disease, including cancer and Robert's Syndrome. As genome instability has been thought to be a hallmark of cancer, mutation of cohesin and other factors may be one of the critical components to cancer development and progression [6, 7].

1.1.1 Cohesin regulates gene expression

While the canonical role for cohesin is sister chromatid cohesion, other roles for cohesin have been identified as well. These include roles in DNA replication, DNA repair, and gene regulation (reviewed in [8]). Many recent studies have focused on exploring cohesin's role in gene regulation, and this work will also focus on this role primarily.

1.1.2 Cohesin as an activator and repressor

Cohesin has been shown to affect gene expression in three different contexts: gene activation, gene silencing, and insulation. Evidence for cohesin's role in gene regulation stems from multiple organisms, from yeast to flies to humans, though there are differences in how cohesin functions from organism to organism. The role of cohesin in each of these expression contexts can be cell-type specific and is critical for development.

1.1.3 Cohesin regulates expression of genes with RNAPII pausing

In the 1970s, studies began to show that elongation could be a rate-limiting step in gene expression. They found that initiation of transcription did not always lead to elongation, suggesting that there was some sort of barrier present [9, 10]. In more recent years, this phenomenon has been further characterized as pausing of RNA Polymerase II (RNAPII), where initiation occurs normally, but the polymerase pauses downstream of the transcription start site (often within the first 100 nucleotides) [11-13]. Recent publications of genome-wide ChIP-seq data show uneven distributions of RNAPII along the gene body, with larger peaks near the transcription start site being common. Factors such as the NELF complex (negative elongation factor) and Spt5 have been shown to be sufficient to induce pausing of RNAPII in purified systems. Surprisingly, Nipbl has been found at—and predictive of—NELF bound genes in Drosophila [14]. Further study revealed that although cohesin was not required for RNAPII pausing, it regulated the expression of these genes; depletion of both NELF and cohesin showed increased transcript levels as compared to depletion of either alone, suggesting the two worked in concert [14]. Since cohesin depletion had no effect on either initiation or transcription of

these genes, it was thought that cohesin regulates expression after the pausing occurs. While there is good evidence for the interaction of cohesin and NELF in Drosophila, it is unclear yet how similar a role cohesin may play in mammals.

1.1.4 Cohesin cooperates with CTCF at insulator boundaries

Insulator boundaries serve to divide the genome into active and inactive domains in higher eukaryotes, can limit or regulate promoter and enhancer interactions, and can limit the spread of heterochromatin [15-18]. One protein that is central to insulator function is the zinc finger protein CTCF, a sequence-specific transcription factor that is ubiquitously expressed and essential [19] (reviewed in [18]). CTCF has been shown to block the communication between enhancers and promoters and prevent transcriptional activation [15, 17, 20], and is capable of mediating chromatin looping between its binding sites [16, 21]. Intriguingly, cohesin was shown to overlap with CTCF at many of its binding sites [22-25], with CTCF able to directly interact with the SA1 subunit of cohesin and recruit cohesin to many of its binding sites genome wide. Even though CTCF and cohesin may not always co-localize directly, the sequence-binding motif for CTCF can be found in at most cohesin binding sites [22].

Cohesin has been shown to be important for the insulator function of CTCF. More than merely co-localizing, CTCF may require cohesin's ability to encircle chromatin and maintain chromatin interactions to establish its insulator function [22]. Conversely, because cohesin does not appear to bind specific sequences on its own, it requires CTCF to recruit it to insulator boundaries for their establishment. While CTCF

and cohesin are capable of establishing and maintaining insulator boundaries, the process of distant regions of chromatin coming into close three-dimensional proximity for this to occur remains an area of active investigation.

Figure 1.1. Cohesin structure and function

A. Cohesin is composed of SMC1, SMC3, RAD21, and SA proteins.

B. Two models for how cohesin interacts with chromatin, either the embrace model (left), or the handcuff model (right)

C. The domain layout of the SMC proteins.

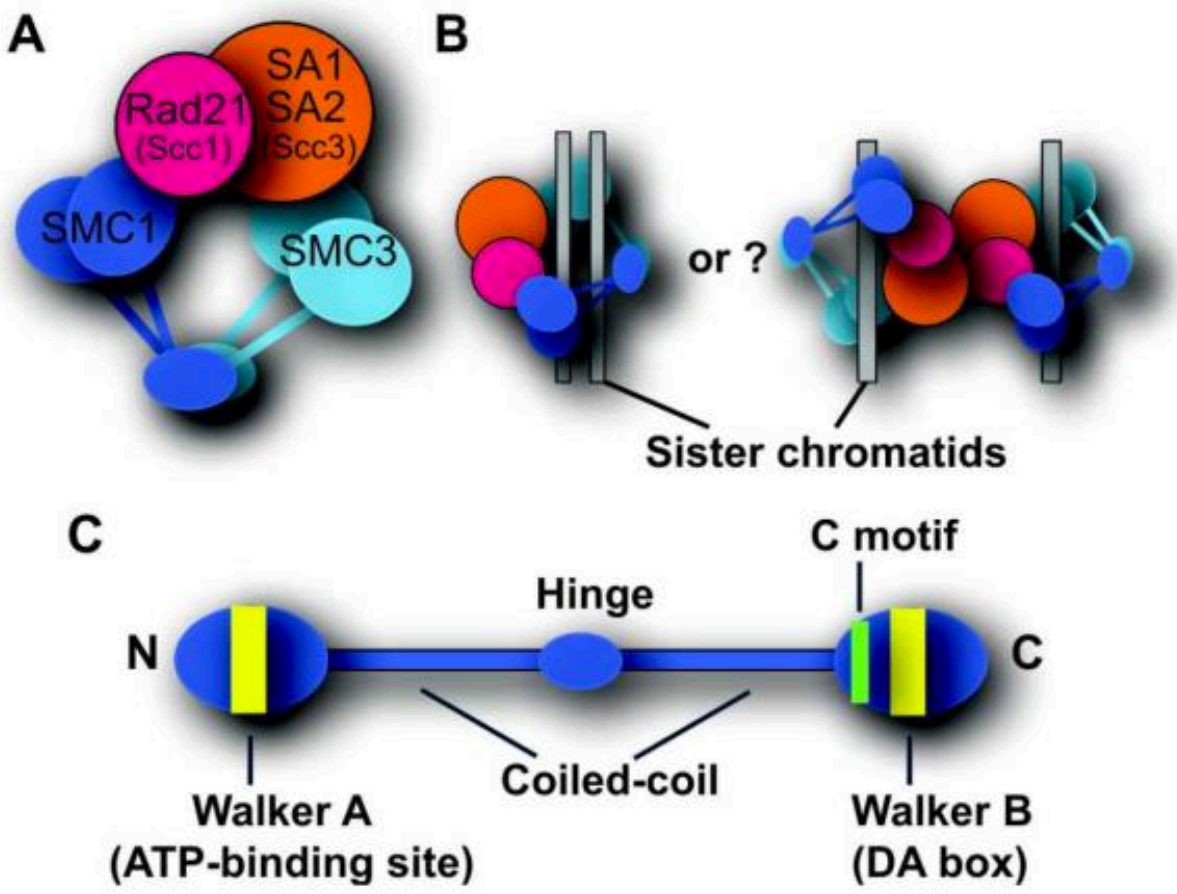


Figure adapted from Chien et. al. [8]

Figure 2.2. Cohesin can mediate long-range chromatin interactions.

Cohesin has the ability to regulate gene expression through mediating long-distance chromatin interactions. Two forms of these interactions, namely insulator loop formation (left) and enhancer and promoter interactions (right) are illustrated here.

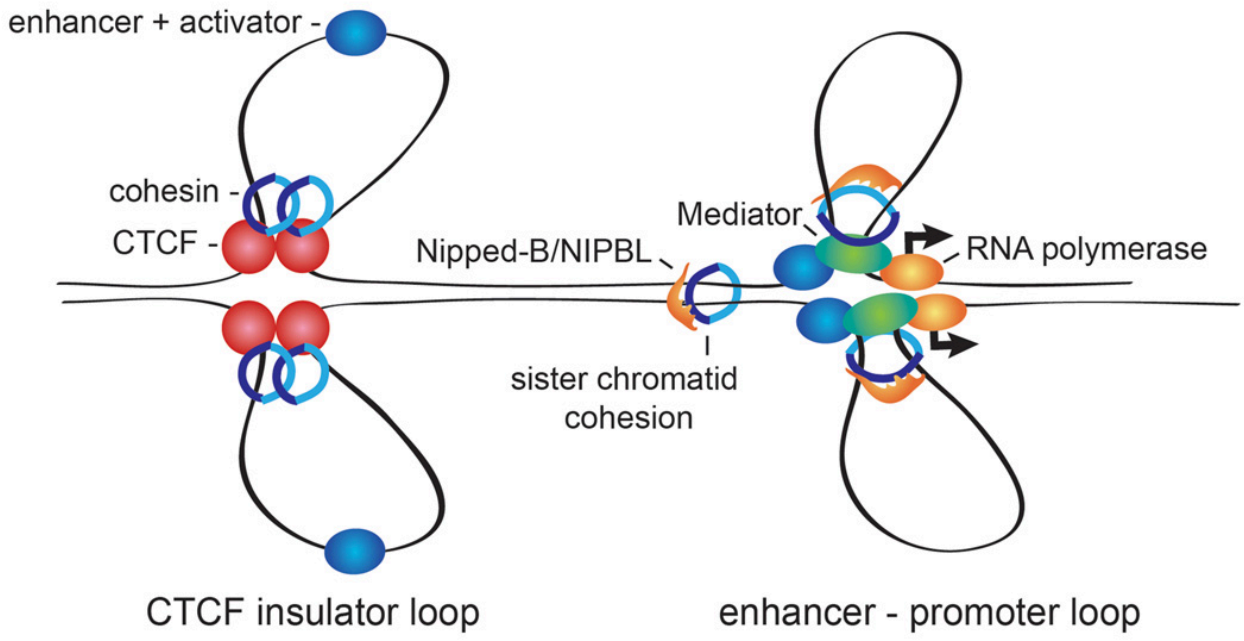


Figure adapted from Dorsett et. al. [26]

1.2 NIPBL

1.2.1 The Nipbl and Mau2 heterodimer

NIPBL is a large, 315 kd protein whose canonical function is in loading cohesin onto chromatin. The *NIPBL* gene contains 47 exons, with two primary isoforms (A and B), though others isoforms are likely to exist [27, 28]. MAU2 forms a heterodimer with NIPBL, and like NIPBL is required for loading of cohesin onto chromatin [29]. Both NIPBL and MAU2 are conserved across species, with orthologs of NIPBL in fission yeast (Mis4), drosophila (Nipped-B), and zebrafish (nipbl) [30-33]; some orthologs for Mau2 have also been more recently characterized [34, 35]. Mau2 interacts with the N-terminal portion of NIPBL, and study of mutations in the NIPBL interaction region in disease—such as in the case of Cornelia de Lange Syndrome (CdLS)—suggest a detrimental effect of disrupting Nipbl-Mau2 interactions on cohesin loading [36].

NIPBL contains other important domains aside from where it interacts with Mau2. NIPBL contains a protein binding motif for HP1 [37], HDAC1/3 [38], and a series of HEAT repeats (cohesin and a related complex condensin also contain heat repeats) [28, 39]. These motifs are thought to be able to wrap around substrates and serve as scaffolding [40]. The varieties of proteins that contain these HEAT repeats suggest that they may be important for chromosome dynamics and function.

1.2.2 Nipbl and Mau2 load the cohesin complex onto chromatin

The cohesin complex is loaded at the end of telophase in mammals, and requires the heterodimer of Nipbl and Mau2 (Scc2-Scc4 in yeast) [29, 41]. Just how Nipbl and Mau2 facilitate the loading of cohesin is unclear. Nipbl and Mau2 were shown to interact with cohesin in yeast [42, 43], and cohesin is unable to stably associate with chromatin when either of them is impaired—thereby preventing sister chromatid cohesion. The heterodimer interacts with the cohesin complex at all four subunits, at places thought to be important for cohesin's function [44]; interactions between Nipbl-Mau2 and the SMC subunits were less critical to cohesin loading than were those with the SA subunit. Also necessary to cohesin's loading was ATP hydrolysis, with either mutation of the Walker A/B motif or lack of ATP preventing interaction with DNA in vitro [42, 44].

1.3 Cornelia de Lange Syndrome

Cornelia de Lange Syndrome (CdLS) (OMIM #122470, #300590 and #610759) is a multi-phenotypic developmental disorder that affects many different organ systems. It occurs at a frequency of 1:10,000 to 1:30,000 individuals, and is characterized in part by craniofacial, neurological, gastrointestinal, and heart defects, as well as limb deformity [26, 28]. Initial descriptions of patients with CdLS occurred in the mid 19th century and early 20th [45, 46], with subsequent work categorizing the clinical phenotypes common to these patients. Individuals with CdLS can show a wide range of severities (Figure 1.2), with some probands showing severe craniofacial defects, limb deformity, and behavioral issues, and others showing only minor facial and limb effects [47]. More than 65% of the studied cases of CdLS have mutations in *NIPBL*, *SMC1*, or *SMC3* [47], with more recent research also implicating mutations in *HDAC8* [5]. Mutations in the cohesin subunit *RAD21* also result in CdLS-like phenotypes, particularly limb deformity, while having little to no cognitive defects [48]. Even after identification of these causative mutations, a substantial percentage of CdLS cases still have undefined causes.

CdLS is caused by mutations in *NIPBL* in more than 65% of cases studied [28, 39]. There are many different mutations that can occur in *NIPBL* (reviewed in [49]), with 278 different heterozygous mutations having been discovered so far. Since NIPBL is an essential protein, mutations are typically heterozygous, with mutations in both alleles believed to be embryonic lethal in humans and mice [50]. Typically, individuals with CdLS have less than a 30% reduction of NIPBL mRNA transcripts, but even a 15% reduction produces clinical phenotypes [28, 39]. The resulting effect is

haploinsufficiency, with the intact allele for *NIPBL* being unable to supply enough transcript/protein even after increased transcription to compensate [50].

Mutations in proteins other than NIPBL are less frequent in CdLS patients than those in *NIPBL* itself. The SMC proteins 1 and 3 have been found to have mutations (only one proband in the case of SMC3) in roughly 5% of cases. Unlike mutations in *NIPBL*, patients with mutations in *SMC1A* show no difference in *SMC1A* transcript or protein levels [47, 51]. Moreover, inclusion of the mutated proteins into the holo-complex is unaffected [52]. It has been suggested that the *SMC1A* mutations function as a dominant negative, with the complexes containing the mutant *SMC1A* subunits having higher affinity for chromatin, and the potential to interfere with biological processes [51]. The *SMC3* mutation is a small, in-frame deletion, and its pathogenic mechanism is unclear [49]. Lastly, patients with mutations in *HDAC8* have been identified, with the mutated versions of HDAC8 incapable of deacetylating SMC3 during S-phase and thereby inhibiting cohesin availability after cell division [5].

1.3.1 Cornelia de Lange Syndrome is linked to gene dysregulation

Although NIPBL is important for the loading of cohesin onto chromatin, it was discovered early on that there was no significant effect of the NIPBL haploinsufficiency on sister chromatid cohesion [30, 50, 53, 54]. This led to the hypothesis that the developmental defects present in CdLS are a result of gene dysregulation. This hypothesis was further supported by work done in a mouse model and in zebrafish [33, 50]. Our own data suggests that it is primarily cohesin's ability to affect gene activation

that is most susceptible to the NIPBL haploinsufficiency (Chapter 3). Since cohesin is important for establishing long-range chromatin interactions, many of these genes may be disregulated by disruption of enhancer-promoter interactions, evidence of which has been shown in mice (Chapter 3, [55, 56]). Other groups have studied the gene expression profiles of patients, indicating a large of number of genes are misregulated in patients in a manner consistent with that seen in the mouse model [54]. Intriguingly, the gene expression changes in either patients or in the mouse model are small, with less than ~4 fold differences [50, 54]. This has led to the hypothesis that the phenotypic severity is not due to the dramatic differences in the expression of the cohesin target genes, but is a result of collective effects of subtle misregulation of many genes [50]. More on this discussion can be found in Chapter 3.

Figure 1.3. Patients with CdLS show a range of severity

The figure below indicates the range of phenotypic severity in patients with CdLS. Some patients show less severe craniofacial defects and limb deformity (E, F, G, H), whereas others show more severe effects (A, B, C, D).

Figure adapted from Mannini et. al. [49]

1.4 FacioScapuloHumeral Muscular Dystrophy

FacioScapuloHumeral Muscular Dystrophy (FSHD) is one of the most common forms of muscular dystrophy in the United States, with a prevalence of between 1:14,000 and 1:20,000 [57-60]. An autosomal dominant disease, it is characterized by the progressive atrophy of the facial and shoulder muscles, with severe cases showing atrophy of the trunk and leg muscles [61] (see figure 1.1). The age of onset for FSHD ranges from infancy to middle age, with most patients presenting symptoms in their twenties to thirties [62, 63]. While the primary disease phenotype of FSHD is atrophy of skeletal muscle, secondary phenotypes including hearing loss and retinal vascular disease have been identified in rare cases [64-66].

There are two forms of FSHD. The first is referred to as 4q-linked, or FSHD1 (OMIM: 158900), and is characterized by a mono-allelic contraction of the D4Z4 subtelomeric macrosatellite repeat array on chromosome 4q35 [67]. A majority of FSHD patients (~95%) have the 4q-linked form of the disease [68]. A minority of patients has an alternate form of FSHD termed phenotypic FSHD or FSHD2 (OMIM: 158901). FSHD2 is phenotypically indistinguishable from FSHD1, but does not have the D4Z4 repeat contraction on chromosome 4q35. Instead, 80% of FSHD2 cases in a particular study showed mutations in an SMC homolog *SMCHD1* [68] (see section 1.2.1 for further discussion). Under normal conditions, individuals have between 11-150 D4Z4 repeats, while in FSHD1 fewer than 10 repeats remain [67]; at least one repeat is required however. Interestingly, there appears to be some correlation between the severity of the disease in FSHD1 patents and the number of remaining repeats; patients with 3 or fewer

repeats show more severe phenotypes, while those with more than 3 tend to have milder phenotypes [69].

1.4.1 DUX4 expression

The D4Z4 repeat array (hereafter just D4Z4) is composed of tandem 3.3 kb macrosatellite sequences, each of which contains a copy of the DUX4 retrogene (figure 1.2.2). DUX4 was shown to encode a transcription factor able to activate genes upstream of D4Z4 and elsewhere in the genome, and may be linked to differentiation defects, atrophy, and muscle defects when expressed in FSHD [67, 70-72]. While the presence of the *DUX*4 gene in D4Z4 has been known since the late 90s [73], its role in FSHD has only recently become clear. Early work suggested that DUX4 might be important to the disease mechanism, but the inability to identify the low-abundance transcripts for *DUX4* in FSHD caused the field to look to other nearby genes [74, 75]. More recent work has gone on to show that *DUX4* is often expressed in FSHD muscle [67], and is toxic and promotes apoptosis.

Having a contraction of D4Z4 repeats is not enough to result in FSHD1. Individuals with the contraction must also have a specific genetic background, with the contraction co-occurring with a polymorphism after the last D4Z4 repeat on chromosome 4q35 [67]. The polymorphism on this allele (referred to as A161) has been shown to encode a polyadenylation signal after the last repeat, allowing for polyadenylation of the *DUX4* transcript —stabilizing its transcripts in vivo [67]. Importantly, the A161 allele is required in FSHD2 as well, with mutations in *SMCHD1* resulting in epigenetic changes

that allow for expression of the *DUX4* transcript [68]. Expression of *DUX4* appears to be important to the disease mechanism in both FSHD1 and FSHD2.

1.4.2: Mutations in SMCHD1 are associated with hypomethylation of D4Z4.

SMCHD1, or Structural Maintenance of Chromosomes hinge domain containing 1, encodes a protein required for gene silencing and DNA methylation on the inactive X chromosome [76]. While *SMCHD1* mutations were identified in connection with approximately 80 % of FSHD2 cases, some patients with FSHD1 also have mutations in *SMCHD1*; these patients show increased severity of the disease even though they have nine copies of D4Z4—typically leading to a less severe form of the disease [77]. SMCHD1 therefore can serve as a modifier of FSHD1 disease severity. Mutations in SMCHD1 correlated exactly with hypomethylation of D4Z4 DNA in patients, suggesting that SMCHD1 is also required for the hypermethylation at D4Z4 typical in normal individuals as well as that on the inactive X chromosome [68]. Depletion of SMCHD1 results in upregulation of *DUX4*, connecting the mutation of *SMCHD1* with the disease mechanism [68]. Interestingly however, SMCHD1 has recently been shown to be important for regulation of some gene clusters, including the protocadherin β cluster, and imprinted genes such as the H19/IGF2 imprinting locus [78]. Therefore, the effect of SMCHD1 mutations could extend beyond D4Z4 to impact the disease etiology.

1.4.3: Loss of H3K9me3, HP1, and cohesin is common to both FSHD1 and FSHD2

FSHD has been characterized as an epigenetic abnormality disease [79], and D4Z4 as a metastable epiallele [68]. The term metastable epiallele refers to genes whose

variable expression is dependent on the probability of repression by other factors [80], in this case referring to the DNA methylation at D4Z4 and it's requirement of SMCHD1 (discussed shortly). Both of these statements are derived from the supporting data that epigenetics undergirds the disease mechanism in FSHD. Our lab previous showed in 2009 that H3K9me3 is lost at D4Z4 in both FSHD1 and FSHD2 patient cells [79]. Upon loss of H3K9me3, both cohesin and HP1γ are lost at D4Z4 [79]. HP1γ is a transcriptional repressor that directly binds H3K9me3 through its chromodomain (reviewed in [81]). In *S. pombe*, its homolog Swi6 has been known to recruit cohesin to the pericentromere [82, 83]. However, Weihua Zeng in our lab showed that cohesin and HP1γ were co-recruited to D4Z4 [79]. The discovery that H3K9me3 was lost in both FSHD1 and FSHD2 was very important, as it was the first time that a common molecular mechanism was shown for the two genetically distinct forms of the disease, suggesting that FSHD is an "epigenetic abnormality disease".

Figure 1.4. Heterochromatin at D4Z4

A. HP1γ, SMCHD1, and cohesin are recruited to D4Z4 by H3K9me3
B. Heterochromatin formed by H3K9me3, HP1γ, SMCHD1, and cohesin represses expression of the DUX4 gene in normal individuals
C. Loss of these heterochromatin components leads to upregulation of DUX4 expression

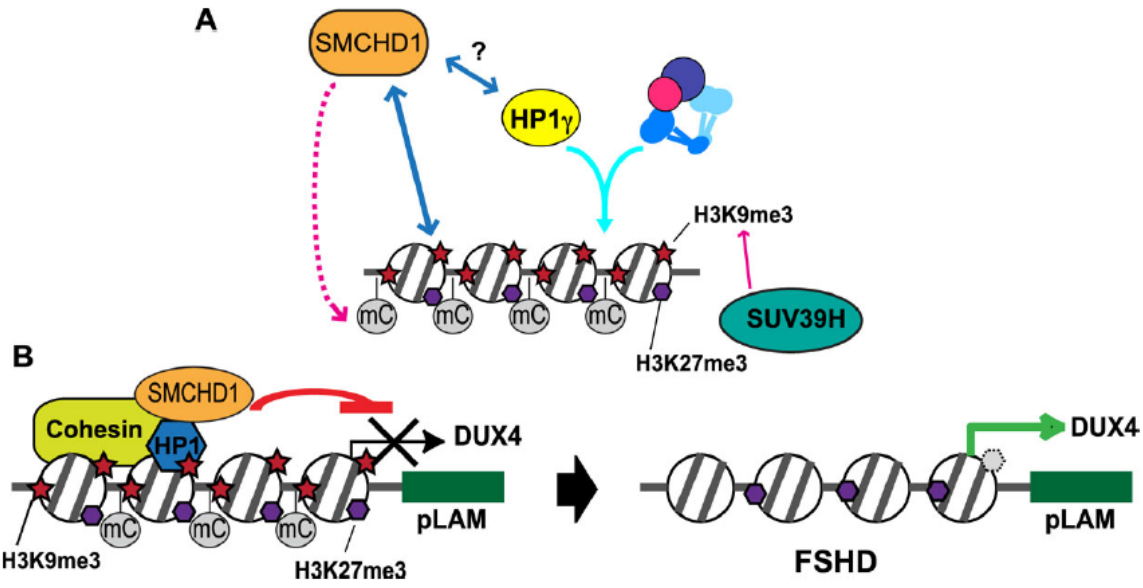


Figure adapted from Zeng et. al. [84]

1.4.4: Loss of H3K9me3 impacts SMCHD1 binding at D4Z4

Our group has recently published an analysis of the D4Z4 homologs present on many chromosomes. In particular, we found that the 4q/10q D4Z4 chromatin showed the characteristic loss of H3K9me3 in FSHD while the homologous regions on other chromosomes did not [84]. Moreover, we found that the open reading frames of the *DUX4* transcripts in these homologs was disrupted, with either missing start codons, early stop codons, and frameshifts [84]. Even more strikingly, we found that the loss of H3K9me3 at D4Z4 results in SMCHD1 displacement and *DUX4* upregulation in immortalized myoblasts, further solidifying the importance of H3K9me3 at D4Z4 to the disease mechanism [84]. The dependency of SMCHD1 on H3K9me3 for localization to D4Z4 suggests that SMCHD1 localization at D4Z4 is already affected in both FSHD1 and FSHD2, and that the mutations in *SMCHD1* further increase the severity of the disease. This will require further validation in patient cells however.

1.5 References

1. Dervyn, E., et al., *The bacterial condensin/cohesin-like protein complex acts in DNA repair and regulation of gene expression.* Mol Microbiol, 2004. **51**(6): p. 1629-40.
2. Hirano, M. and T. Hirano, *Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA.* EMBO J, 2002. **21**(21): p. 5733-44.
3. Haering, C.H., et al., *Molecular architecture of SMC proteins and the yeast cohesin complex.* Mol Cell, 2002. **9**(4): p. 773-88.
4. Melby, T.E., et al., *The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge.* J Cell Biol, 1998. **142**(6): p. 1595-604.
5. Deardorff, M.A., et al., *HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle.* Nature, 2012. **489**(7415): p. 313-7.
6. Xu, H., et al., *Cohesin Rad21 Mediates Loss of Heterozygosity and Is Upregulated via Wnt Promoting Transcriptional Dysregulation in Gastrointestinal Tumors.* Cell Rep, 2014.
7. Matynia, A.P., et al., *Molecular Genetic Biomarkers in Myeloid Malignancies.* Arch Pathol Lab Med, 2014.
8. Chien, R., et al., *Cohesin: a critical chromatin organizer in mammalian gene regulation.* Biochem Cell Biol, 2011. **89**(5): p. 445-58.
9. Gariglio, P., M. Bellard, and P. Chambon, *Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes.* Nucleic Acids Res, 1981. **9**(11): p. 2589-98.
10. Fraser, N.W., P.B. Sehgal, and J.E. Darnell, *DRB-induced premature termination of late adenovirus transcription.* Nature, 1978. **272**(5654): p. 590-3.
11. Wada, T., et al., *DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs.* Genes Dev, 1998. **12**(3): p. 343-56.
12. Kephart, D.D., N.F. Marshall, and D.H. Price, *Stability of Drosophila RNA polymerase II elongation complexes in vitro.* Mol Cell Biol, 1992. **12**(5): p. 2067-77.
13. Marshall, N.F. and D.H. Price, *Control of formation of two distinct classes of RNA polymerase II elongation complexes.* Mol Cell Biol, 1992. **12**(5): p. 2078-90.
14. Fay, A., et al., *Cohesin selectively binds and regulates genes with paused RNA polymerase.* Curr Biol, 2011. **21**(19): p. 1624-34.
15. Bell, A.C., A.G. West, and G. Felsenfeld, *The protein CTCF is required for the enhancer blocking activity of vertebrate insulators.* Cell, 1999. **98**(3): p. 387-96.
16. Splinter, E., et al., *CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.* Genes Dev, 2006. **20**(17): p. 2349-54.

17. Hark, A.T., et al., *CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus.* Nature, 2000. **405**(6785): p. 486-9.
18. Holwerda, S.J. and W. de Laat, *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120369.
19. Heath, H., et al., *CTCF regulates cell cycle progression of alphabeta T cells in the thymus.* EMBO J, 2008. **27**(21): p. 2839-50.
20. Recillas-Targa, F., et al., *Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6883-8.
21. Handoko, L., et al., *CTCF-mediated functional chromatin interactome in pluripotent cells.* Nat Genet, 2011. **43**(7): p. 630-8.
22. Wendt, K.S., et al., *Cohesin mediates transcriptional insulation by CCCTC-binding factor.* Nature, 2008. **451**(7180): p. 796-801.
23. Rubio, E.D., et al., *CTCF physically links cohesin to chromatin.* Proc Natl Acad Sci U S A, 2008. **105**(24): p. 8309-14.
24. Parelho, V., et al., *Cohesins functionally associate with CTCF on mammalian chromosome arms.* Cell, 2008. **132**(3): p. 422-33.
25. Stedman, W., et al., *Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators.* EMBO J, 2008. **27**(4): p. 654-66.
26. Dorsett, D. and I.D. Krantz, *On the molecular etiology of Cornelia de Lange syndrome.* Ann N Y Acad Sci, 2009. **1151**: p. 22-37.
27. Tonkin, E.T., et al., *A giant novel gene undergoing extensive alternative splicing is severed by a Cornelia de Lange-associated translocation breakpoint at 3q26.3.* Hum Genet, 2004. **115**(2): p. 139-48.
28. Krantz, I.D., et al., *Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B.* Nat Genet, 2004. **36**(6): p. 631-5.
29. Ciosk, R., et al., *Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins.* Mol Cell, 2000. **5**(2): p. 243-54.
30. Rollins, R.A., P. Morcillo, and D. Dorsett, *Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes.* Genetics, 1999. **152**(2): p. 577-93.
31. Furuya, K., K. Takahashi, and M. Yanagida, *Faithful anaphase is ensured by Mis4, a sister chromatid cohesion molecule required in S phase and not destroyed in G1 phase.* Genes Dev, 1998. **12**(21): p. 3408-18.
32. Gillespie, P.J. and T. Hirano, *Scc2 couples replication licensing to sister chromatid cohesion in Xenopus egg extracts.* Curr Biol, 2004. **14**(17): p. 1598-603.
33. Muto, A., et al., *Multifactorial origins of heart and gut defects in nipbl-deficient zebrafish, a model of Cornelia de Lange Syndrome.* PLoS Biol, 2011. **9**(10): p. e1001181.
34. Bernard, P., et al., *A screen for cohesion mutants uncovers Ssl3, the fission yeast counterpart of the cohesin loading factor Scc4.* Curr Biol, 2006. **16**(9): p. 875-81.

35. Seitan, V.C., et al., *Metazoan Scc4 homologs link sister chromatid cohesion to cell and axon migration guidance.* PLoS Biol, 2006. **4**(8): p. e242.
36. Braunholz, D., et al., *Isolated NIBPL missense mutations that cause Cornelia de Lange syndrome alter MAU2 interaction.* Eur J Hum Genet, 2012. **20**(3): p. 271-6.
37. Lechner, M.S., et al., *The mammalian heterochromatin protein 1 binds diverse nuclear proteins through a common motif that targets the chromoshadow domain.* Biochem Biophys Res Commun, 2005. **331**(4): p. 929-37.
38. Jahnke, P., et al., *The Cohesin loading factor NIPBL recruits histone deacetylases to mediate local chromatin modifications.* Nucleic Acids Res, 2008. **36**(20): p. 6450-8.
39. Tonkin, E.T., et al., *NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome.* Nat Genet, 2004. **36**(6): p. 636-41.
40. Neuwald, A.F. and T. Hirano, *HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions.* Genome Res, 2000. **10**(10): p. 1445-52.
41. Uhlmann, F. and K. Nasmyth, *Cohesion between sister chromatids must be established during DNA replication.* Curr Biol, 1998. **8**(20): p. 1095-101.
42. Arumugam, P., et al., *ATP hydrolysis is required for cohesin's association with chromosomes.* Curr Biol, 2003. **13**(22): p. 1941-53.
43. Toth, A., et al., *Yeast cohesin complex requires a conserved protein, Eco1p(Ctf7), to establish cohesion between sister chromatids during DNA replication.* Genes Dev, 1999. **13**(3): p. 320-33.
44. Murayama, Y. and F. Uhlmann, *Biochemical reconstitution of topological DNA binding by the cohesin ring.* Nature, 2014. **505**(7483): p. 367-71.
45. Brachmann, W., *Ein fall von symmetrischer monodaktylie durch Ulnadefekt, mit symmetrischer flughautbildung in den ellenbeugen, sowie anderen abnormitaten (zwerghaftogkeit, halsrippen, behaarung).* Jarb Kinder Phys Erzie, 1916. **84**: p. 225-235.
46. Vrolik, W., *Tabulae ad illustrandam embryogenesin hominis et mammalium tam naturalem quam abnormem*. 1849: London, Amsterdam.
47. Liu, J. and I.D. Krantz, *Cornelia de Lange syndrome, cohesin, and beyond.* Clin Genet, 2009. **76**(4): p. 303-14.
48. Deardorff, M.A., et al., *RAD21 mutations cause a human cohesinopathy.* Am J Hum Genet, 2012. **90**(6): p. 1014-27.
49. Mannini, L., et al., *Mutation spectrum and genotype-phenotype correlation in Cornelia de Lange syndrome.* Hum Mutat, 2013. **34**(12): p. 1589-96.
50. Kawauchi, S., et al., *Multiple organ system defects and transcriptional dysregulation in the Nipbl(+/-) mouse, a model of Cornelia de Lange Syndrome.* PLoS Genet, 2009. **5**(9): p. e1000650.
51. Revenkova, E., et al., *Cornelia de Lange syndrome mutations in SMC1A or SMC3 affect binding to DNA.* Hum Mol Genet, 2009. **18**(3): p. 418-27.
52. Gimigliano, A., et al., *Proteomic profile identifies dysregulated pathways in Cornelia de Lange syndrome cells with distinct mutations in SMC1A and SMC3 genes.* J Proteome Res, 2012. **11**(12): p. 6111-23.

53. Castronovo, P., et al., *Premature chromatid separation is not a useful diagnostic marker for Cornelia de Lange syndrome.* Chromosome Res, 2009. **17**(6): p. 763-71.
54. Liu, J., et al., *Transcriptional dysregulation in NIPBL and cohesin mutant human cells.* PLoS Biol, 2009. **7**(5): p. e1000119.
55. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.
56. Chien, R., et al., *Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression.* J Biol Chem, 2011. **286**(20): p. 17870-8.
57. Flanigan, K.M., et al., *Genetic characterization of a large, historically significant Utah kindred with facioscapulohumeral dystrophy.* Neuromuscul Disord, 2001. **11**(6-7): p. 525-9.
58. Mostacciuolo, M.L., et al., *Facioscapulohumeral muscular dystrophy: epidemiological and molecular study in a north-east Italian population sample.* Clin Genet, 2009. **75**(6): p. 550-5.
59. Padberg, G.W., et al., *Facioscapulohumeral muscular dystrophy in the Dutch population.* Muscle Nerve Suppl, 1995(2): p. S81-4.
60. Norwood, F.L., et al., *Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population.* Brain, 2009. **132**(Pt 11): p. 3175-86.
61. Tawil, R. and S.M. Van Der Maarel, *Facioscapulohumeral muscular dystrophy.* Muscle Nerve, 2006. **34**(1): p. 1-15.
62. Scionti, I., et al., *Large-scale population analysis challenges the current criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy.* Am J Hum Genet, 2012. **90**(4): p. 628-35.
63. Lunt, P.W., D.A. Compston, and P.S. Harper, *Estimation of age dependent penetrance in facioscapulohumeral muscular dystrophy by minimising ascertainment bias.* J Med Genet, 1989. **26**(12): p. 755-60.
64. Brouwer, O.F., et al., *Hearing loss in facioscapulohumeral muscular dystrophy.* Neurology, 1991. **41**(12): p. 1878-81.
65. Lutz, K.L., et al., *Clinical and genetic features of hearing loss in facioscapulohumeral muscular dystrophy.* Neurology, 2013. **81**(16): p. 1374-7.
66. Fitzsimons, R.B., E.B. Gurwin, and A.C. Bird, *Retinal vascular abnormalities in facioscapulohumeral muscular dystrophy. A general association with genetic and therapeutic implications.* Brain, 1987. **110 (Pt 3)**: p. 631-48.
67. Lemmers, R.J., et al., *A unifying genetic model for facioscapulohumeral muscular dystrophy.* Science, 2010. **329**(5999): p. 1650-3.
68. Lemmers, R.J., et al., *Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2.* Nat Genet, 2012. **44**(12): p. 1370-4.
69. Larsen, M., et al., *Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1.* Eur J Hum Genet, 2014.
70. Young, J.M., et al., *DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis.* PLoS Genet, 2013. **9**(11): p. e1003947.

71. Yao, Z., et al., *DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle.* Hum Mol Genet, 2014. **23**(20): p. 5342-52.
72. Tassin, A., et al., *DUX4 expression in FSHD muscle cells: how could such a rare protein cause a myopathy?* J Cell Mol Med, 2013. **17**(1): p. 76-89.
73. Gabriels, J., et al., *Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element.* Gene, 1999. **236**(1): p. 25-32.
74. Gabellini, D., M.R. Green, and R. Tupler, *Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle.* Cell, 2002. **110**(3): p. 339-48.
75. Klooster, R., et al., *Comprehensive expression analysis of FSHD candidate genes at the mRNA and protein level.* Eur J Hum Genet, 2009. **17**(12): p. 1615-24.
76. Blewitt, M.E., et al., *SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation.* Nat Genet, 2008. **40**(5): p. 663-9.
77. Sacconi, S., et al., *The FSHD2 Gene SMCHD1 Is a Modifier of Disease Severity in Families Affected by FSHD1.* Am J Hum Genet, 2013.
78. Massah, S., et al., *Epigenetic characterization of the growth hormone gene identifies SmcHD1 as a regulator of autosomal gene clusters.* PLoS One, 2014. **9**(5): p. e97535.
79. Zeng, W., et al., *Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD).* PLoS Genet, 2009. **5**(7): p. e1000559.
80. Rakyan, V.K., et al., *Metastable epialleles in mammals.* Trends Genet, 2002. **18**(7): p. 348-51.
81. Zeng, W., A.R. Ball, Jr., and K. Yokomori, *HP1: heterochromatin binding proteins working the genome.* Epigenetics, 2010. **5**(4): p. 287-92.
82. Bernard, P., et al., *Requirement of heterochromatin for cohesion at centromeres.* Science, 2001. **294**(5551): p. 2539-42.
83. Nonaka, N., et al., *Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast.* Nat Cell Biol, 2002. **4**(1): p. 89-93.
84. Zeng, W., et al., *Genetic and epigenetic characteristics of FSHD-associated 4q and 10q D4Z4 that are distinct from non-4q/10q D4Z4 homologs.* Hum Mutat, 2014. **35**(8): p. 998-1010.

Chapter 2

AREM: aligning short reads from ChIP-Sequencing by expectation maximization

Coauthored with: Jake Biesinger, Alvin Chon, Kyoko Yokomori, and Xiaohui Xie

Reprinted with permission from JOURNAL OF COMPUTATIONAL BIOLOGY, 21011, Volume 18, Issue 11, pp. 1495-1505, published by Mary Ann Liebert, Inc., New Rochelle, NY

2.1 Abstract

High-throughput sequencing coupled to chromatin immunoprecipitation (ChIP-Seq) is widely used in characterizing genome-wide binding patterns of transcription factors, cofactors, chromatin modifiers, and other DNA binding proteins. A key step in ChIP-Seq data analysis is to map short reads from high-throughput sequencing to a reference genome and identify peak regions enriched with short reads. Although several methods have been proposed for ChIP-Seq analysis, most existing methods only consider reads that can be uniquely placed in the reference genome, and therefore have low power for detecting peaks located within repeat sequences. Here we introduce a probabilistic approach for ChIP-Seq data analysis, which utilizes all, reads, providing a truly genome-wide view of binding patterns. Reads are modeled using a mixture model corresponding to K enriched regions and a null genomic background. We use maximum likelihood to estimate the locations of the enriched regions, and implement an expectation-maximization (E-M) algorithm, called AREM, to update the alignment probabilities of each read to different genomic locations. We apply the algorithm to identify genome-wide binding events of two proteins: Rad21, a component of cohesin and a key factor involved in chromatid cohesion, and Srebp-1, a transcription factor important for lipid/cholesterol homeostasis. Using AREM, we were able to identify 19,935 Rad21 peaks and 1,748 Srebp-1 peaks in the mouse genome with high confidence, including 1,517 (7.6%) Rad21 peaks and 227 (13%) Srebp-1 peaks that were missed using only uniquely mapped reads. The open source implementation of our algorithm is available at http://sourceforge.net/projects/arem.

2.2 Introduction

In recent years, high-throughput sequencing coupled to chromatin immunoprecipitation (ChIP-Seq) has become one of the premier methods of analyzing protein-DNA interactions [1]. The ability to capture a vast array of protein binding locations genome-wide in a single experiment has led to important insights in a number of biological processes, including transcriptional regulation, epigenetic modification and signal transduction [2–5]. Numerous methods have been developed to analyze ChIP-Seq data and typically work well for identifying protein-DNA interactions located within non-repeat sequences. However, identifying interactions in repeat regions remains a challenging problem since sequencing reads from these regions usually cannot be uniquely mapped to a reference genome. We present novel methodology for identifying protein-DNA interactions in repeat sequences.

ChIP-Seq computational analysis typically consists of two tasks: one is to identify the genomic locations of the short reads by aligning them to a reference genome, and the second is to find genomic regions enriched with the aligned reads, which is often termed "peak finding". Eland, MAQ, Bowtie, and SOAP are among the most popular for mapping short reads to a reference genome [6–9] and provide many or all of the potential mappings for a given sequence read. Once potential mappings have been identified, significantly enriched genomic regions are identified using one of several available tools [10–18]. Some peak finders are better suited for histone modification studies, others for transcription factor binding site identification. These peak finders have been surveyed on several occasions [19–21].

Many short reads cannot be uniquely mapped to the reference genome. Most peak finding workflows throw away these non-uniquely mapped reads, and as a consequence have low power for detecting peaks located within repeat regions. While each experiment varies, only about 60% [in house data] of the sequence reads from a ChIP-Seq experiment can be uniquely mapped to a reference genome. Therefore, a significant portion of the raw data is not utilized by the current methods. There have been proposals to address the non-uniquely mapped reads in the literature by either randomly choosing a location from a set of potential ones [22, 23] or by taking all potential alignments [12], but most peak callers are not equipped to deal with ambiguous reads.

We propose a novel peak caller designed to handle ambiguous reads directly by performing read alignment and peak-calling jointly rather than in two separate steps. In the context of ChIP-Seq studies, regions enriched during immunoprecipitation are more likely the true genomic source of sequence reads than other regions of the genome. We leverage this idea to iteratively identify the true genomic source of ambiguous reads. Under our model, the true locations of reads and binding peaks are treated as hidden variables, and we implement an algorithm, AREM, to estimate both iteratively by alternating between mapping reads and finding peaks.

Two ChIP-Seq datasets were used in this study: 1) cohesin, a new dataset generated in house, and 2) Srebp-1, a previously published dataset [5]. To generate the cohesin dataset, ChIP-Seq was performed using mouse embryonic fibroblasts and an antibody targeting Rad21 [24], a subunit of cohesin. Cohesin is an essential protein

complex required for sister chromatid cohesion. In mammalian cells, cohesin binding sites are present in intergenic, promoter and 3' regions-especially in connection with CTCF binding sites [25, 26]. It was found that cohesin is recruited by CTCF to many of its binding sites, and plays a role in CTCF-dependent gene regulation [27, 28]. Cohesin has been shown to bind to repeat sequences in a disease-specific manner [24], making it a particularly interesting candidate for our study.

The second dataset is Srebp-1, a transcription factor important in allostatic regulation of sterol biosynthesis and membrane lipid composition [29]. This particular dataset [5] examines the genomic binding locations for Srebp-1 in mouse liver. Regulation of expression by Srebp-1 is important for regulation of cholesterol and repeat-binding for this TF has not been previously shown [30, 29]. We choose these datasets because both proteins have well characterized regulatory motifs, allowing us to directly test the validity of our peak finding method directly.

On a 2.8Ghz CPU, AREM takes about 20 minutes and 1.6GB RAM to call peaks from over 12 million alignments and about 30 minutes and 6GB RAM to call peaks from nearly 120 million alignments. Each dataset takes less than 40 iterations to converge. AREM is written in Python, is open-source, and is available at http://sourceforge.net/projects/arem.

2.3 Results

Building on the methodology of the popular peak-caller MACS [13], we implement AREM, a novel peak caller designed to handle multiple possible alignments for each sequence read. AREM's peak caller combines an initial sliding window approach with a greedy refinement step and iteratively aligns ambiguous reads. We use two ChIP-Seq datasets in this study: Rad21, a subunit of the structural protein cohesin, contained 7.2 million treatment reads and 7.4 million control reads (manuscript in preparation). Srebp-1, a regulator of cholesterol metabolism, had 7.7 million treatment reads and 6.4 million control reads [5].

Using AREM, we identify 19,935 Rad21 peaks covering more than 10 million base pairs at a low FDR of 3.7% and 1,474 Srebp-1 peaks covering nearly 1 million bases at a moderate FDR of 8%. For comparison, we also called peaks using MACS and SICER [15], another popular peak finding program. To compare our results, we use FDR and motif presence as indicators of *bona fide* binding sites.

2.3.1 AREM identifies additional binding sites

We seek to benchmark both AREM's peak-calling and its multiread methodology. To benchmark peak-calling, we limit all reads to their best alignment and run AREM, MACS and SICER. In the Rad21 dataset, AREM identifies 456 more peaks than MACS and 1,920 more peaks than SICER but retains a similar motif presence (81.6% MACS, 82.5% SICER, 81.3% AREM) and has a lower FDR (2.8% MACS, 12.7% SICER, 1.9% AREM) (see Table 2.1). For Srebp-1, AREM identifies more than double the number of

peaks compared to MACS and 816 more than SICER, though the FDR is slightly higher (4.85% MACS, 9% SICER, 8% AREM) and motif presence slightly lower (46.6% MACS, 59% SICER, 39% AREM). In both datasets, AREM appears to be more sensitive to true binding sites, picking up more total sites with motif instances, although it trades off some specificity in Srebp-1.

To see if AREM can identify true sites that are not significant without multireads, we performed peak-calling with multireads, removing peaks that overlapped with those identified using AREM without multireads. Up to 1,546 (8.1%) and 272 (18.9%) previously unidentified peaks were called from Rad21 and Srebp-1, respectively. These new peaks have a similar motif presence compared to previous peaks but overlap with annotated repeat regions more often.

2.3.2 AREM's sensitivity is increased with ambiguous reads

Several methods for dealing with ambiguous reads have been proposed, including retaining all possible mappings, retaining one of the mappings chosen at random, and distributing weight equally among the mappings. The first option will clearly lead to false positives, particularly in repeat regions as the number of retained mappings increases. We compare the latter two methods to our E-M implementation, varying the number of retained reads and summarize the results in Table 2.1. Although both random selection and fractionating reads increases the number of peaks called, our E-M method outperforms them, yielding 1546 more peaks for Rad21, and 272 for Srebp-1 with comparable quality. As the number of retained alignments increases, the disparity gets

smaller. AREM shows fairly consistent results across datasets with a large increase in total number of alignments (nearly 40-fold for Rad21, over 10-fold for Srebp-1).

For a given sample, the iterations show a continued shift of the max alignment probabilities to either 1 or 0. This shift is consistent across datasets with larger numbers of max alignments (data not shown), but does depend on other parameters. What is apparent is that AREM's E-M heuristic performs well, allowing for significant shift toward a "definitive" alignment; at the same time, it does not force a shift on reads with too little information, preventing misalignment and resulting spurious peak-calling.

2.3.3 AREM is sensitive to repeat regions

An important parameter in our model is the minimum enrichment score for all K regions. Since repeat regions have such similar sequence content, many reads will share the same repetitive elements. If one of the shared repeat elements has a slightly higher enrichment score by chance, the E-M method will iteratively shift probability into that repeat region, snowballing the region into what appears to be a full-fledged sequence peak. To distinguish repetitive peaks arising by small enrichment fluctuations from true binding sites within or adjacent to repetitive elements, we impose a minimum enrichment score on all regions. Lower threshold scores will be sensitive to these random fluctuations but true binding peaks may be missed if the score is too high.

To explore the effect of varying the minimum enrichment score, we varied the minimum score from 0.1 to 2, keeping the maximum number of alignments fixed at 20.

For Rad21, we see a declining number of discovered peaks ranging from 28,305 to 19,634 peaks respectively. In addition to a decline in discovered peaks as minimum enrichment score increases, we also see a decrease in the reported FDR and the percent of peaks in repeat regions from 11.28% to 2.95% FDR and 71.56% to 59.02%. Lastly, the percent of peaks with motif increases from 63.64% to 81.12%. These additional peaks appear to be of lower quality: motifs are largely absent from them and the FDR is much higher, see Figure 2.2.

For our method, detecting peaks near repeat regions is a tradeoff between sensitivity and specificity. As the minimum score increases, the method approaches the uniform or ”fraction” distribution, in which only the initial mapping quality scores (and not the enrichment) affect alignment probabilities. The fraction method is explored explicitly, showing increased power compared to unique reads only, but decreased sensitivity to true binding sites compared to other AREM runs.

2.4 Discussion

Repetitive elements in the genome have traditionally been problematic in sequence analysis. Since sequenced reads are short and repetitive sequences are similar, many equally likely mappings may exist for a given read. Our method uses the low-coverage unique reads near repeat regions to evaluate which potential alignments for each read are the most likely. Sensitivity to repeat regions is adjustable, however there is a tradeoff: increasing sensitivity may introduce false positives. Further refinement of our methodology may lead to increased specificity.

Our results imply that functional CTCF binding sites exist within repeat regions, revealing an interesting relationship between repetitive sequence and chromatin structure. Another application of our method would be to explore the relationship between repetitive sequence and epigenetic modifications such as histone modifications. Regulation of and by transposable elements has been linked to methylation marks [31], and transposable elements have a major role in cancers [32]. Better identification of histone modifications in regions of repetitive DNA increases our understanding of key regulators of genome stability and diseases sparked by translocations and mutations.

Table 2.1

Three peak callers (MACS, SICER, and AREM) were run on both datasets. For AREM, the maximum number of retained alignments per read is varied (from 1 to 80). The total number of peaks and bases covered by peaks is reported as well as the FDR by swapping treatment and control. For both datasets, AREM's minimum enrichment score was fixed at 1.5 with 20 maximum alignments per read. For comparison, the motif background rate of occurrence was 4.5% (CTCF) and 27% (Srebp-1) in 100,000 genomic samples, sized similarly to Rad21 MACS peaks and Srebp-1 MACS peaks, respectively.

TABLE 1. COMPARISON OF PEAK-CALLING METHODS FOR COHESIN AND SREBP-1.

Method	*No. of alignments*	*No. of peaks*	*Peak bases*	*FDR*	*New peaks*	*Motif*	*Repeat*
Cohesin							
MACS	2,368,229	18,556	9,546,641	2.8%	—	81.67%	56.55%
SICER	2,368,229	17,092	17,374,108	12.71%	—	82.55%	70.42%
AREM 1	2,368,229	19,012	9,353,567	1.9%	—	81.32%	55.30%
AREM 10	7,616,647	19,881	10,225,479	3.8%	1,404	81.04%	58.88%
AREM 20	12,312,878	19,935	10,531,465	3.7%	1,517	80.88%	59.66%
AREM 40	20,527,010	19,863	10,744,836	3.2%	1,546	80.93%	60.34%
AREM 80	34,537,311	19,820	10,972,796	2.9%	1,538	80.73%	60.91%
Srebp-1							
MACS	10,482,005	721	495,968	4.85%	—	46.60%	53.95%
SICER	10,482,005	622	963,778	9.0%	—	59.00%	77.33%
AREM 1	10,482,005	1,438	880,284	8.0%	—	39.08%	53.47%
AREM 10	28,347,869	1,815	996,346	10.5%	262	39.22%	56.04%
AREM 20	44,493,532	1,748	959,646	8.0%	227	39.95%	55.97%
AREM 40	72,453,642	1,685	983,459	8.2%	248	40.34%	56.46%
AREM 80	118,744,757	1,695	987,746	7.3%	272	40.66%	56.73%

Table 2.1

Figure 2.1

(A) AREM workflow diagram.
(B–E) *de novo* discovery of motifs. From top to bottom:
(B) CTCF in MACS peaks from uniquely mapping reads,
(C) CTCF in AREM's peaks with multireads,
(D) Srebp-1 in MACS peaks from uniquely mapping reads and
(E) Srebp-1 in AREM peaks with multireads.

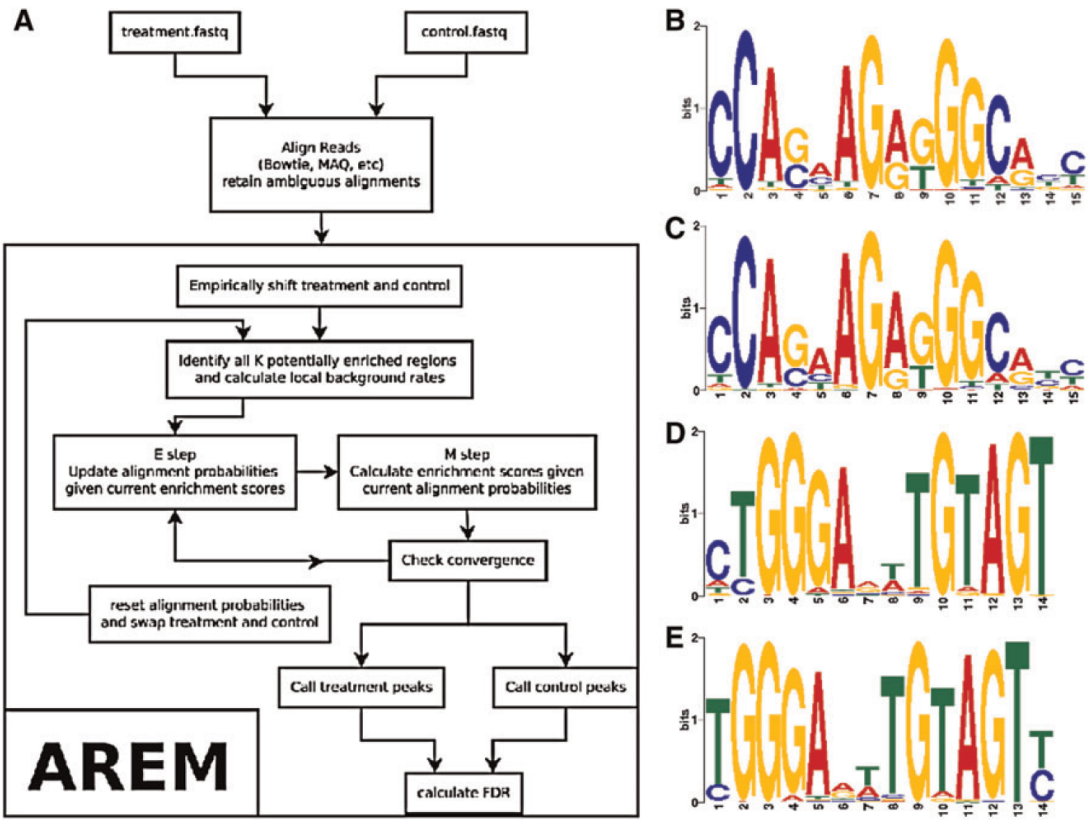

A
treatment.fastq
control.fastq
Align Reads
(Bowtie, MAQ, etc)
retain ambiguous alignments
Empirically shift treatment and control
Identify all K potentially enriched regions
and calculate local background rates
E step
Update alignment probabilities
given current enrichment scores
M step
Calculate enrichment scores given
current alignment probabilities
Check convergence
reset alignment probabilities
and swap treatment and control
Call treatment peaks
Call control peaks
AREM
calculate FDR
B
C
D
E
bits

Figure 2.1

Figure 2.2

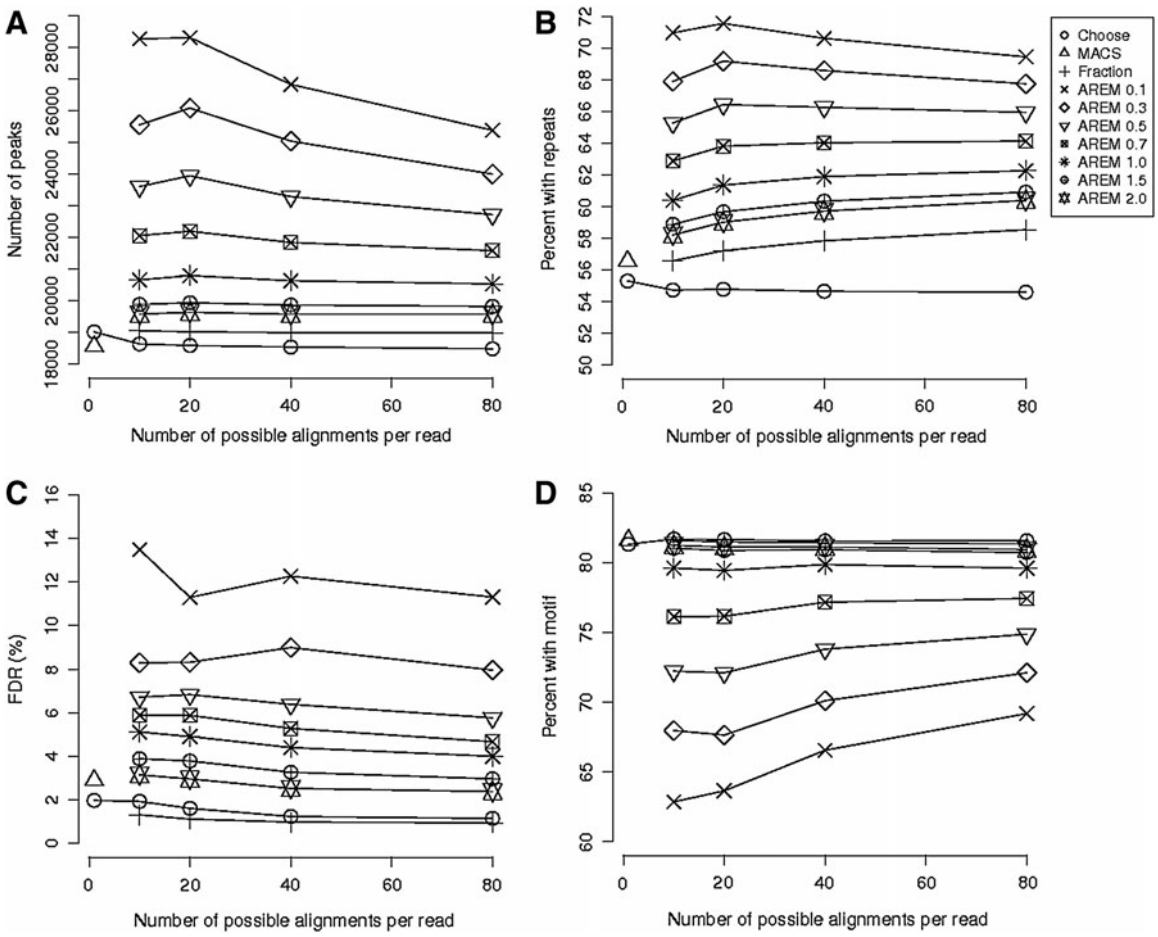
Graphs displaying varying parameters and number of possible alignments per read.

(A) Total number of peaks discovered.

(B) Percentage of peaks with repetitive sequences.

(C) False discovery rate.

(D) Percentage of peaks with motif.


A
Number of peaks
18000
20000
22000
24000
26000
28000
0
20
40
60
80
Number of possible alignments per read
B
Percent with repeats
50
52
54
56
58
60
62
64
66
68
70
72
Number of possible alignments per read
Choose
MACS
Fraction
AREM 0.1
AREM 0.3
AREM 0.5
AREM 0.7
AREM 1.0
AREM 1.5
AREM 2.0
C
FDR (%)
0
2
4
6
8
10
12
14
16
Number of possible alignments per read
D
Percent with motif
60
65
70
75
80
85
Number of possible alignments per read

2.5 Methods

2.5.1 Notations

Let $R = \{r_1,...,r_N\}$ denote a set of reads from a ChIP-Seq experiment with read, $r_i \in \Sigma^l$, where $\Sigma = \{A,C,G,T\}$, l is the length of each read, and N denotes the number of reads. Let $S \in \Sigma^L$ denote the reference sequence to which the reads will be mapped. In real applications, the reference sequence usually consists of multiple chromosomes. For notational simplicity, we assume the chromosomes have been concatenated to form one reference sequence.

We assume that for each read we are provided with a set of potential alignments to the reference sequence. Denote the set of potential alignments of read r_i to S by $A_i = \{(l_{ij},q_{ij}) : j = 1,...,n_i\}$, where l_{ij} and q_{ij} denote the starting location and the confidence score of the j-th alignment, and n_i is the total number of potential alignments. We assume $q_{ij} \in [0,1]$ for all j, and use it to account for both sequencing quality scores and mismatches between the read and the reference sequence. There are several programs available to generate the initial potential alignments and confidence scores.

2.5.2 Mixture model

We use a generative model to describe the likelihood of observing the given set of short reads from a ChIP-Seq experiment. Suppose the ChIP procedure results in the enrichment of K non-overlapping regions in the reference sequence S. Denote the K

enriched regions (also called peak regions) by $\{(s_k, w_k) : k = 1,...,K\}$, where s_k and w_k represent the start and the width, respectively, of the i–th enriched region in S. Let $E_k = \{s_k,...,s_k + w_k - l\}$ denote the set of locations in the enriched region k that can potentially generate a read of length l. Let E_k^s, E_k^w denote the start and width of region k. We will use E_0 to denote all locations in S that are not covered by $\bigcup_{k=1}^{K} E_k$.

We use variable $z_i \in \{1,...,n_i\}$ to denote the true location of read r_i, with $z_i = j$ representing that r_i originates from location l_{ij} of S. In addition, we use variable $u_i \in \{0,1,...,K\}$ to label the type of region that read r_i belongs to. $u_i = k$ represents that read r_i is from the non-enriched regions of S if $k = 0$, and is from k-th enriched region otherwise. Both z_i and u_i are not directly observable, and are often referred to as the hidden variables of the generative model.

Let $P(r_i \mid z_i = j, u_i = k)$ denote the conditional probability of observing read r_i given that r_i is from location l_{ij} and belongs to region k. Assuming different reads are generated independently, the log likelihood of observing R given the mixture model is then

$$\ell = \sum_{i=1}^{N} \log\left[\sum_{j=0}^{n_i}\sum_{k=0}^{K} P(r_i \mid z_i = j, u_i = k) P(z_i = j) P(u_i = k)\right],$$

where $P(z_i)$ and $P(u_i)$ represent the prior probabilities of the location and the region type, respectively, of read r_i. $P(z_i)$ is set according to the confidence scores of different alignments

$$P(z_i = j) = \frac{q_{ij}}{\sum_{k=1}^{n_i} q_{ik}}. \quad (1)$$

$P(u_i)$ depends on both the width and the enrichment ratio of each enriched region. Denote the enrichment ratio of the ChIP regions vs. non-ChIP regions by α, which is often significantly impacted by the quality of antibodies used in ChIP experiments. We parameterize the prior distribution on region types as follows

$$P(u_i = k) = \frac{1}{(\alpha - 1)\sum_j w_j + L} \times \begin{cases} L - \sum_j w_j & \text{if } k = 0 \\ \alpha w_k & \text{o.w.} \end{cases} \quad (2)$$

2.5.3 Parameter estimation

The conditional probability $P(r_i \mid z_i = j, u_i = k)$ can be modeled in a number of different ways. For example, bell-shaped distributions are commonly used to model the enriched regions. However, for computational simplicity, we will use a simple uniform distribution to model the enriched regions. If read r_i comes from one of the enriched regions, i.e., $k \neq 0$, we assume the read is equally likely to originate from any of the potential positions within the enriched region, that is,

$$P(r_i \mid z_i = j, u_i = k) = \frac{1}{w_k - l + 1} I_{E_k}(l_{ij}), \quad (3)$$

where $I_A(x)$ is the indicator function, returning 1 if $x \in A$ and 0 otherwise.

If the read is from non-enriched regions, i.e., $k = 0$, we use p_i^b to model the background probability of an arbitrary read originating from location i of the reference sequence. (We assume p_i^b has been properly normalized such that $\sum_{i=1}^{L} p_i^b = 1$.) Then the conditional probability $P(r_i \mid z_i = j, u_i = k)$ for the case of $k = 0$ is modeled by

$$P\left(r_i \mid z_i = j, u_i = 0\right) = I_{E_0}\left(l_{ij}\right) p_{l_{ij}}^b \tag{4}$$

Numerous ChIP-Seq studies have demonstrated that the locations of ChIP-Seq reads are typically non-uniform, significantly biased toward promoter or open chromatin regions [1]. The p_i^b's takes this ChIP and sequencing bias into account, and can be inferred from control experiments typically employed in ChIP-Seq studies.

Next we integrate out the u_i variable to obtain the conditional probability of observing r_i given only z_i

$$P\left(r_i \mid z_i = j\right) = P\left(u_i = 0\right) I_{E_0}\left(l_{ij}\right) p_{l_{ij}}^b + \sum_{k=1}^{K} \frac{P\left(u_i = k\right)}{w_k = l+1} I_{E_k}\left(l_{ij}\right). \tag{5}$$

Note that because $E_0, E_1, ..., E_K$ are disjoint, only one term in the above summation can be non-zero. This property significantly reduces the computation for parameter estimation since we do not need to infer the values of u_i variables any more.

The log likelihood of observing R given the mixture model can now be written as

$$\ell\left(r_1,...,r_n;\Theta\right)=\sum_{i=1}^{N}\log\left[\sum_{j=0}^{n_i}P\left(r_i\mid z_i=j\right)P\left(z_i=j\right)\right], \tag{6}$$

where $\Theta=(s_1,w_1,...,s_K,w_K,\alpha)$ denotes the parameters of the mixture model. We estimate the values of these unknown parameters using maximum likelihood estimation

$$\hat{\Theta}=\arg\max_{\Theta}\ell(r_1,...,r_n;\Theta)\,. \tag{7}$$

2.5.4 Expectation-maximization algorithm

We solve the maximum likelihood estimation problem in Eq. (7) through an expectation-maximization (E-M) algorithm. The algorithm iteratively applies the following two steps until convergence:

Expectation step: Estimate the posterior probability of alignments under the current estimate of parameters $\Theta^{(t)}$:

$$Q^{(t)}(z_i=j\mid R)=\frac{1}{C}P(r_i\mid z_i=j,\Theta^{(t)})P(z_i=j)\,, \tag{8}$$

where C is a normalization constant.

Maximization step: Find the parameters $\Theta^{(t+1)}$ that maximize the following quantity,

$$\Theta^{(t+1)} = \arg\max_{\Theta} \sum_{i=1}^{N} \sum_{j=0}^{n_i} Q^{(t)}\left(z_i = j \mid R\right) \log P\left(r_i \mid z_i = j, \Theta\right). \qquad (9)$$

2.5.5 Implementation of E-M updates

The mixture model described above contains $2K+1$ parameters. Since K, the number of peak regions, is typically large, ranging from hundreds to hundreds of thousands, exactly solving Eq. (9) in the maximization step is nontrivial. Instead of seeking an exact solution, we identify the K regions from the data by considering all regions where the number of possible alignments is significantly enriched above the background.

For a given window of size w starting at s of the reference genome, we first calculate the number of reads located within the window, weighted by the current estimation of posterior alignment probabilities,

$$f(s,w) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Q^{(t)}(z_i = j \mid R) I_{[s,s+w-l]}(l_{ij}) \qquad (10)$$

We term this quantity the foreground read density. As a comparison, we also calculate a background read density $b(s,w)$, which is estimated using either reads from the control experiment or reads from a much larger extended region covering the window. Different ways of calculating background read density are discussed in [13].

Provided with both background and foreground read densities, we then define an

enrichment score $\phi(s,w)$ to measure the significance of read enrichment within the window starting at position s with width w. For this purpose, we assume the number of reads is distributed according to a Poisson model with mean rate $b(s,w)$. If $f(s,w)$ is an integer, the enrichment score is defined to be $\phi(s,w) = -\log_{10}(1-g(f,b))$, where

$$g(x,\lambda) = e^{-\lambda}\sum_{k=0}^{x}\frac{\lambda^k}{k!} \tag{11}$$

denotes the chance of observing at least x Poisson events given the mean rate of λ. However, if $f(s,w)$ is not an integer, the enrichment score cannot be defined this way. Instead, we use a linear extrapolation to define the enrichment score $\phi(s,w) = -\log_{10}(1-\tilde{g}(f,b))$, where function $\tilde{g}$ is defined as

$$\tilde{g}(x,\lambda) = g(\lfloor x \rfloor,\lambda) + \left[g(\lceil x \rceil,\lambda) - g(\lfloor x \rfloor,\lambda)\right](x - \lfloor x \rfloor). \tag{12}$$

If two potential alignments of a read have the same confidence score and are located in two peak regions with equal enrichment, the update of posterior alignment probabilities in Eq. (8) will assign equal weight to these two alignments. This is so because we have assumed that peak regions have the same enrichment ratio as described in Eq. (2), which is not true as some peak regions are more enriched than others in real ChIP experiments. To address this issue, we have also implemented an update of the posterior probabilities that takes the calculated enrichment scores into account as

$$Q^t(z_i = j \mid R) \leftarrow \sum_{k=1}^{K} [\phi(E_k^s, E_k^w) P(z_i = j) I_{E_k}(z_i)] \tag{13}$$

which is then normalized. In practice, we found this implementation usually behaves better than the one without using enrichment scores.

We use entropy to quantify the uncertainty of alignments associated with each read. For read i, the entropy at iteration t is defined to be

$$H_i^t = -\sum_{j=1}^{n_i} Q^t(z_i = j \mid R) \log Q^t(z_i = j \mid R) . \tag{14}$$

We stop the E-M iteration when the relative square difference between two consecutive entropies is small, that is, when

$$\frac{\sum_{i=0}^{N} (H_i^t - H_i^{t-1})^2}{\sum_{i=0}^{N} (H_i^{t-1})^2} < \varepsilon , \tag{15}$$

where $\varepsilon = 10^{-5}$ for results reported in this paper.

AREM seeks to identify the true genomic source of multiply-aligning reads (also called multireads). Many of the multireads will map to repeat regions of the genome, and we expect repeats to be included in the K potentially enriched regions. To prevent repeat regions from garnering multiread mass without sufficient evidence of their enrichment, we impose a minimum enrichment score. Effectively, unique or less ambiguous

multireads need to raise enrichment above noise levels for repeat regions to be called as peaks. The minimum enrichment score is a parameter of our model and its effect on called peaks is explored in the results.

2.5.6 Alignment

Alignment was performed using Bowtie [7]. We used the Burrows-Wheeler index provided by the Bowtie website to align reads; the index is based on the unmasked MM9 reference genome from the UCSC Genome Browser [33]. The first base of all raw reads was clipped to remove sequencing artifacts and a maximum of two mismatches were allowed in the first 28 bases of the remaining sequence. We generated several alignment collections for both Srebp-1 and Rad21 by varying k, the maximum number of reported alignments. We restricted our study to search the 1, 10, 20, 40, and 80 best alignments. Table 2.1 shows that the total number of alignments was only starting to plateau at k=80, indicating that many sequences have more than 80 possible alignments, for practicality we restricted our search as above. Map confidence scores were calculated from Bowtie output as in [8]. We also provide an option for using the aligner's confidence scores directly rather than recalculating them from mismatches and sequence qualities. During preparation of the sequencing library, unequal amplification can result in biased counts for reads. To eliminate this bias, we limit the number of alignments to one for each start position on each strand. In particular, we choose the best alignment (based on quality score) for each position; in the event that all alignments have the same quality score, we choose a random read to represent that particular position.

2.5.7 Peak Finding

Our peak finding method is an adapted version of the MACS [13] peak finder. Like MACS, we empirically model the spatial separation between +/- strand tags and shift both treatment and control tags. We also continue MACS' conservative approach to background modeling, using the highest of three rates as the background (in this study, genome-wide or within 1,000 or 10,000 bases). As a divergence from MACS, we use a sliding window approach to identify large potentially enriched regions then use a smoothened greedy approach to refine called peaks. We call peaks within this large region by greedily adding reads to improve enrichment, but avoid local optima by always looking up to the full sliding window width away. The initial large regions correspond to the K regions used for the E-M steps of Section 2.5.5. During the E-M steps, local background rates are used as during final peak-calling. Peaks reported in this study are above a p -value of 10e-5. All enrichment scores and p -values are calculated using the Poisson linear interpolation described in equation 12. Once E-M is complete on the treatment data and peaks are called, we reset the treatment alignment probabilities, swap treatment and control and rerun the algorithm, including E-M steps, to determine the False Discovery Rate (FDR). For all algorithms tested in this study, we define the FDR as the ratio of peaks called using control data to peaks called using treatment data. This method of FDR calculation is common in ChIP-Seq studies (e.g., [13, 15]).

2.5.8 Motif finding

Motif presence helps determine peak quality, as shown in [34]. To determine if our new peaks were of the same quality as the other peaks, we performed *de novo* motif

discovery using MEME [35] version 4.4. Input sequence was limited to 150 bp (Rad21) and 200 bp (Srebp-1) around the summit of the peaks called by MACS from uniquely mapping reads. All sequences were used for Srebp-1, while 1,000 sequences were randomly sampled a total of 5 times for Rad21. The motif signal was strong in both datasets and the discovered motif position weight matrix (PWM) was extracted for further use. We also used performed the motif search using Srebp-1 and CTCF motifs catalogued in Transfac 11.3, and found similar results. For the CTCF motif, we did genomic sampling (100,000 samples) to identify a threshold score corresponding to a z-score of 4.29. For Srebp-1, we used the threshold score reported by MEME. See Figure 2.1.

2.6 References

1. Park, P.: ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10 (2009) 669–680
2. Mikkelsen, T., Ku, M., Jaffe, D., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R., et al.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448 (2007) 553–560
3. Ouyang, Z., Zhou, Q., Wong, W.: ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences 106 (2009) 21521
4. Blow, M., McCulley, D., Li, Z., Zhang, T., Akiyama, J., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al.: ChIP-Seq identification of weakly conserved heart enhancers. Nature genetics 42 (2010) 806–810
5. Seo, Y., Chong, H., Infante, A., Im, S., Xie, X., Osborne, T.: Genome-wide analysis of SREBP-1 binding in mouse liver chromatin reveals a preference for promoter proximal binding to a new motif. Proceedings of the National Academy of Sciences 106 (2009) 13765
6. Cox, A.J.: Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. http://bioinfo.cgrb.oregonstate.edu/docs/solexa (2007)
7. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10 (2009) R25
8. Li, H., Ruan, J., Durbin, R.: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 18 (2008) 1851
9. Li, R., Li, Y., Kristiansen, K., Wang, J.: SOAP: short oligonucleotide alignment program. Bioinformatics 24 (2008) 71310.
10. Fejes, A., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., Jones, S.: FindPeaks 3.1: a

tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24 (2008) 1729

11. Ji, H., Jiang, H., Ma, W., Johnson, D., Myers, R., Wong, W.: An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature biotechnology 26 (2008) 1293–1300
12. Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5 (2008) 621–628
13. Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., et al.: Model-based analysis of ChIP-Seq (MACS). Genome biology 9 (2008) R137
14. Spyrou, C., Stark, R., Lynch, A., Tavar´e, S.: BayesPeak: Bayesian analysis of ChIP-seq data. BMC bioinformatics 10 (2009) 299
15. Zang, C., Schones, D., Zeng, C., Cui, K., Zhao, K., Peng,W.: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25 (2009) 1952
16. Blahnik, K., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A., Iyengar, S., Nicolet, C., Ludascher, B., Korf, I., et al.: Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Research 38 (2010) e13
17. Qin, Z., Yu, J., Shen, J., Maher, C., Hu, M., Kalyana-Sundaram, S., Yu, J., Chin-naiyan, A.: HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC bioinformatics 11 (2010) 369
18. Salmon-Divon, M., Dvinge, H., Tammoja, K., Bertone, P.: PeakAnalyzer: Genome- wide annotation of chromatin binding and modification loci. BMC bioinformatics 11 (2010) 415
19. Kharchenko, P., Tolstorukov, M., Park, P.: Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology 26 (2008) 1351–1359
20. Pepke, S., Wold, B., Mortazavi, A.: Computation for ChIP-seq and RNA-seq studies. Nature Methods 6 (2009) S22–S32

21. Wilbanks, E., Facciotti, M.: Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PloS one 5 (2010) e11471
22. Kagey, M., Newman, J., Bilodeau, S., Zhan, Y., Orlando, D., van Berkum, N., Ebmeier, C., Goossens, J., Rahl, P., Levine, S., et al.: Mediator and cohesin connect gene expression and chromatin architecture. Nature (2010)
23. Schmid, C., Bucher, P.: MER41 Repeat Sequences Contain Inducible STAT1 Binding Sites. PloS one 5 (2010) e11425
24. Zeng, W., De Greef, J., Chen, Y., Chien, R., Kong, X., Gregson, H., Winokur, S., Pyle, A., Robertson, K., Schmiesing, J., et al.: Specific loss of histone H3 lysine 9 trimethylation and HP1/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). (2009)
25. Rubio, E., Reiss, D., Welcsh, P., Disteche, C., Filippova, G., Baliga, N., Aeber- sold, R., Ranish, J., Krumm, A.: CTCF physically links cohesin to chromatin. Proceedings of the National Academy of Sciences 105 (2008) 8309
26. Liu, J., Zhang, Z., Bando, M., Itoh, T., Deardorff, M., Clark, D., Kaur, M., Tandy, S., Kondoh, T., Rappaport, E., et al.: Transcriptional dysregulation in NIPBL and cohesin mutant human cells. PLoS Biol 7 (2009) e1000119
27. Wendt, K., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al.: Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451 (2008) 796–801
28. Nativio, R., Wendt, K., Ito, Y., Huddleston, J., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J., Murrell, A.: Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. (2009)
29. Hagen, R., Rodriguez-Cuenca, S., Vidal-Puig, A.: An allostatic control of membrane lipid composition by SREBP1. FEBS letters (2010)
30. Yokoyama, C., Wang, X., Briggs, M., Admon, A., Wu, J., Hua, X., Goldstein, J., Brown, M.:

SREBP-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low density lipoprotein receptor gene. Cell 75 (1993) 187–197

31. Huda, A., Jordan, I.: Epigenetic regulation of Mammalian genomes by transposable elements. Annals of the New York Academy of Sciences 1178 (2009) 276–284
32. Chuzhanova, N., Abeysinghe, S., Krawczak, M., Cooper, D.: Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Human mutation 22 (2003) 245–251
33. Rhead, B., Karolchik, D., Kuhn, R., Hinrichs, A., Zweig, A., Fujita, P., Diekhans, M., Smith, K., Rosenbloom, K., Raney, B., et al.: The UCSC genome browser database: update 2010. Nucleic acids research (2009)
34. Boeva, V., Surdez, D., Guillon, N., Tirode, F., Fejes, A., Delattre, O., Barillot, E.: De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic Acids Research (2010)
35. Bailey, T., Elkan, C.: The value of prior knowledge in discovering motifs with MEME. In: Proc Int Conf Intell Syst Mol Biol. Volume 3. (1995) 21–9

Chapter 3

The effect of Nipbl haploinsufficiency on genome-wide cohesin binding and target gene expression: modeling Cornelia de Lange Syndrome

3.1 Abstract

Cornelia de Lange Syndrome (CdLS) is a multisystem developmental disorder frequently associated with heterozygous loss-of-function mutations of *Nipped-B-like* (*NIPBL*), the human homolog of Drosophila *Nipped-B*. NIPBL loads cohesin onto chromatin. Cohesin mediates sister chromatid cohesion important for mitosis, but is also increasingly recognized as a regulator of gene expression. In CdLS patient cells and animal models, the presence of multiple gene expression changes with little or no sister chromatid cohesion defect suggests that disruption of gene regulation underlies this disorder. However, the effect of *NIPBL* haploinsufficiency on cohesin binding, and how this relates to the clinical presentation of CdLS, has not been fully investigated. We examined genome-wide cohesin binding and its relationship to gene expression using mouse embryonic fibroblasts (MEFs) from *Nipbl* +/- mice that recapitulate the CdLS phenotype. We found a global decrease in cohesin binding, including at CCCTC-binding factor (CTCF) binding sites and repeat regions. Cohesin-bound genes were found to be enriched for histone H3 lysine 4 trimethylation (H3K4me3) at their promoters; were disproportionately downregulated in *Nipbl* mutant MEFs; and displayed evidence of reduced promoter-enhancer interaction. The results suggest that gene activation is the primary cohesin function sensitive to Nipbl reduction. Over 50% of significantly dysregulated transcripts in mutant MEFs come from cohesin target genes, including genes involved in adipogenesis that have been implicated in contributing to the CdLS phenotype.

3.2 Introduction

CdLS (OMIM 122470, 300590, 610759) is a dominant genetic disorder estimated to occur in 1 in 10,000 individuals, characterized by facial dysmorphism, hirsutism, upper limb abnormalities, cognitive retardation, and growth abnormalities [1, 2]. Mutations in the *NIPBL* gene are linked to more than 55% of CdLS cases [3, 4]. NIPBL is an evolutionarily conserved, essential protein that is required for chromatin loading of cohesin [5]. Cohesin is a multiprotein complex, also conserved and essential, which functions in chromosome structural organization important for genome maintenance and gene expression [6-8]. Mutations in the cohesin subunits SMC1 (human SMC1 (hSMC1), SMC1A) and hSMC3 were also found in a minor subset of clinically milder CdLS cases (~5% and <1%, respectively) [9-11]. More recently, mutation of HDAC8, which regulates cohesin dissociation from chromatin in mitosis, was found in a subset of CdLS patients (OMIM 300882) [12]. Mutations in the non-SMC cohesin component *Rad21* gene have also been found in patients with a CdLS-like phenotype (OMIM 606462), with much milder cognitive impairment [13]. Thus, mutations of cohesin subunits and regulators of cohesin's chromatin association cause related phenotypes, suggesting that impairment of the cohesin pathway makes significant contributions to the disease [2, 14].

Because the most common cause of CdLS is *NIPBL* haploinsufficiency [2, 15, 16], *Nipbl* heterozygous mutant (*Nipbl* +/-) mice have been developed as a CdLS disease model. These mice exhibit wide-ranging defects characteristic of the disease, including small size, craniofacial anomalies, microbrachycephaly, heart defects, hearing abnormalities, low body fat, and delayed bone maturation [17]. The mutant mice display

only a 25-30% decrease in *Nipbl* transcripts, presumably due to compensatory upregulation of the intact allele [17]. A similar partial decrease is found in *NIPBL*-haploinsufficient CdLS patients, and even a 15% decrease in expression can cause mild CdLS [18, 19]. These results indicate a high sensitivity of mammalian development to *Nipbl* gene dosage. Although a canonical function of cohesin is sister chromatid cohesion critical for mitosis [8], a role for cohesin in gene regulation has been argued for based on work in multiple organisms [20, 21]. The partial decrease of *Nipbl* expression in CdLS patients and *Nipbl* +/- mice was not sufficient to cause a significant sister chromatid cohesion defect or abnormal mitosis [17, 22-24]. Instead, a distinctive profile of gene expression changes was observed, strongly suggesting that transcriptional dysregulation underlies the disease phenotype [17, 19]. In Nipbl +/- mutant mice, gene expression changes are pervasive, though mostly minor, raising the possibility that small expression perturbations of multiple genes collectively contribute to the disease phenotype [17]. This hypothesis was further tested by combinatorial gene depletion in zebrafish, successfully recapitulating some aspects of the CdLS-like phenotype [25]. However, to what extent Nipbl and cohesin directly regulate affected genes in this CdLS mouse model has not yet been determined.

Cohesin is recruited to different genomic regions and affects gene expression in different ways in mammalian cells [6, 7]. In mammalian cells, one major mechanism of cohesin-mediated gene regulation is through CTCF [26-29]. CTCF is a zinc finger DNA-binding protein and was shown to act as a transcriptional activator/repressor as well as an insulator [30]. Genome-wide chromatin immunoprecipitation (ChIP) analyses revealed

that a significant number of cohesin binding sites overlap with those of CTCF in human and mouse somatic cells [26, 27]. Cohesin is recruited to these sites by CTCF and mediates CTCF's insulator function by bridging distant CTCF sites at, for example, the *H19/IGF2*, *IFNg*, apolipoprotein, and *b-globin* loci [26, 27, 29, 31-34]. While CTCF recruits cohesin, it is cohesin that plays a primary role in long-distance chromatin interaction [32]. A more recent genome-wide Chromosome Conformation Capture Carbon Copy (5C) study revealed that CTCF/cohesin tends to mediate long-range chromatin interactions defining megabase-sized topologically associating domains (TADs) [35], indicating that CTCF and cohesin together play a fundamental role in chromatin organization in the nucleus. Cohesin also binds to other genomic regions and functions in a CTCF-independent manner in gene activation by facilitating promoter-enhancer interactions together with Mediator [31, 35-37]. Significant overlap between cohesin at non-CTCF sites and cell type-specific transcription factor binding sites was found, suggesting a role for cohesin at non-CTCF sites in cell type-specific gene regulation [38, 39]. In addition, cohesin is recruited to heterochromatic repeat regions [40, 41]. To what extent these different modes of cohesin recruitment and function are affected by *NIPBL* haploinsufficiency in CdLS has not been examined.

Here, using MEFs derived from *Nipbl*+/- mice, we analyzed the effect of Nipbl haploinsufficiency on cohesin-mediated gene regulation and identified cohesin target genes that are particularly sensitive to partial reduction of Nipbl. Our results indicate that Nipbl is required for cohesin binding to both CTCF and non-CTCF sites, as well as repeat regions. Significant correlation was found between gene expression changes in Nipbl

mutant cells and cohesin binding to the gene regions, in particular promoter regions, suggesting that even modest Nipbl reduction directly and significantly affects expression of cohesin-bound genes. Target genes are enriched for developmental genes, including multiple genes that regulate adipogenesis, which is impaired in Nipbl +/- mice [17]. The results indicate that Nipbl regulates a significant number of genes through cohesin. While their expression levels vary in wild type cells, the Nipbl/cohesin target genes tend on the whole to be downregulated in Nipbl mutant cells, indicating that Nipbl and cohesin are important for activation of these genes. Consistent with this, these genes are enriched for H3 lysine 4 trimethylation (H3K4me3) at the promoter regions. The long-distance interaction of the cohesin-bound promoter and a putative enhancer region is decreased by Nipbl reduction, indicating that reduced cohesin binding by Nipbl haploinsufficiency affects chromatin interactions. Collectively, the results reveal that Nipbl haploinsufficiency globally reduces cohesin binding, and its major transcriptional consequence is downregulation of cohesin target genes.

3.3 Results

3.3.1 Nipbl haploinsufficiency leads to a global reduction of cohesin binding to its binding sites

In order to investigate how Nipbl haploinsufficiency leads to CdLS, cohesin binding was examined genome-wide by ChIP-seq analyses using antibody specific for the cohesin subunit Rad21, in wild type and *Nipbl* +/- mutant MEFs derived from E15.5 embryos [17] (Figure 3.1A). MEFs derived from five wild type and five mutant pups from two litters were combined to obtain sufficient chromatin samples for ChIP-seq analysis. *Nipbl* +/- mutant MEFs express approximately 30-40% less Nipbl compared to wild type MEFs [17] (Table 3.2). MEFs from this embryonic stage were chosen in order to match with a previous expression microarray study, because they are relatively free of secondary effects caused by *Nipbl* mutation-induced developmental abnormalities compared to embryonic tissue [17]. Consistent with this, there is no noticeable difference in growth rate and cell morphology between normal and mutant MEFs [17]. This particular anti-Rad21 antibody was used previously for ChIP analysis and was shown to identify holo-cohesin complex binding sites [26, 31, 41, 42]. This is consistent with the close correlation of the presence of other cohesin subunits at identified Rad21 binding sites [43] (Figure 3.1B).

Cohesin binding sites were identified using AREM [44], with a significance cut-off based on a *p*-value less than $1x10^{-4}$, resulting in a FDR below 3.0% (Figure 3.1A). Cohesin binding peaks ranged from ~200bp to ~6kb in size with the majority less than

1kb in both wild type and mutant cells (median value of 499 bp in wild type and 481 bp in mutant cells) (Figure 3.1C). Approximately 35% fewer cohesin binding sites were found in *Nipbl* +/- mutant MEFs compared to the wild type MEFs (Figure 3.1A). This is not due to variability in sample preparation since no significant difference in the histone H3 ChIP-seq was observed between the wild type and mutant cell samples (R-value=0.96) (Figure 3.1D). Since the total read number for mutant ChIP-seq was ~15% less than for wild type ChIP-seq (Figure 3.1A), we examined whether the difference was in part due to a difference in the number of total read sequences between the two Rad21 ChIP samples. To address this, we randomly removed reads from the wild type sample to match the number of reads in the mutant sample, and ran the peak discovery algorithm again on the reduced wild type read set. This was repeated 1,000 times. We found that the wild type sample still yielded ~39% more peaks than the mutant, indicating that identification of more peaks in the wild type sample is not due to a difference in the numbers of total read sequences (Figure 3.1E). Thus, cohesin appears to bind to fewer binding sites in *Nipbl* haploinsufficient cells.

The above results might suggest that a significant number of binding sites are unique to the wild type cells (Figure 3.1A). When we compared the raw number of reads located within wild type peaks and the corresponding regions in mutant MEFs, however, we noted a reduced, rather than a complete absence of, cohesin binding in mutant cells (Figure 3.1F). Those regions in mutant cells corresponding to the "WT only" regions consistently contain one to three tags in a given window, which are below the peak cut-off. However, the signals are significant compared to the negative control of preimmune

IgG (Figure 3.1F). Furthermore, even for those sites that are apparently common between the control and mutant MEFs, the binding signals appear to be weaker in mutant cells (Figure 3.1F). To validate this observation, we segmented the genome into nonoverlapping 100bp bins, and plotted a histogram of the log ratios of read counts between the wild type and mutant samples in each bin, with read counts normalized using reads per kb per million total reads (RPKM) [45]. The plot indicates that the read counts for the mutant bins are generally less than those for the wild type bins, even for the binding sites common to both wild type and mutant cells (Figure 3.1G). Signal intensity profiles of the Rad21 ChIP-seq in the selected gene regions also show a general decrease of Rad21 binding at its binding sites in Nipbl+/- MEFs compared to the control MEFs (see below, Figure 3.6B). Decreased cohesin binding was further confirmed by manual ChIP-qPCR analysis of individual cohesin binding sites using at least three independent control and mutant MEF samples supporting the reproducibility of the results (see below, Figure 3.3). Decreased cohesin binding was also observed at additional specific genomic regions in *Nipbl*+/- MEFs [46]. Taken together, the results indicate that cohesin binding is generally decreased at its binding sites found in wild type MEFs, rather than re-distributed, in mutant MEFs.

Figure 3.1. Global decrease of cohesin binding to chromatin in *Nipbl* heterozygous mutant MEFs.

A. Cohesin binding sites identified by ChIP-sequencing using antibody specific for Rad21 in control wild type and Nipbl +/- MEFs.

Peak calling was done using AREM [44]. The *p*-value and false discovery rate (FDR) are shown.

B. Heatmap comparison of Rad21 ChIP-seq data with those of SMC1, SMC3, SA1 and SA2. Rad21 peaks in the wild type MEFs are ranked by strongest to weakest, and compared to the ChIP-seq data of SMC1, SMC3, SA1 and SA2 in MEFs (GSE32320) [43] in the corresponding regions. The normalized (reads per million) tag densities in a 4 kb window around each Rad21 peak are plotted, with peaks sorted from the highest number of tags in the wild type MEFs to the lowest.

C. Histogram of cohesin peak widths in wild type and mutant MEFs, indicating the number of peaks in a given size range. The segmentation of the histogram is at 100bp intervals. The median value is indicated with a vertical black line and labeled.

D. Scatter plot of H3 ChIP-Seq data in both wild type and *Nipbl* +/- MEFs for 500 bp bins over the genome. The values are plotted in log reads per million (RPM).

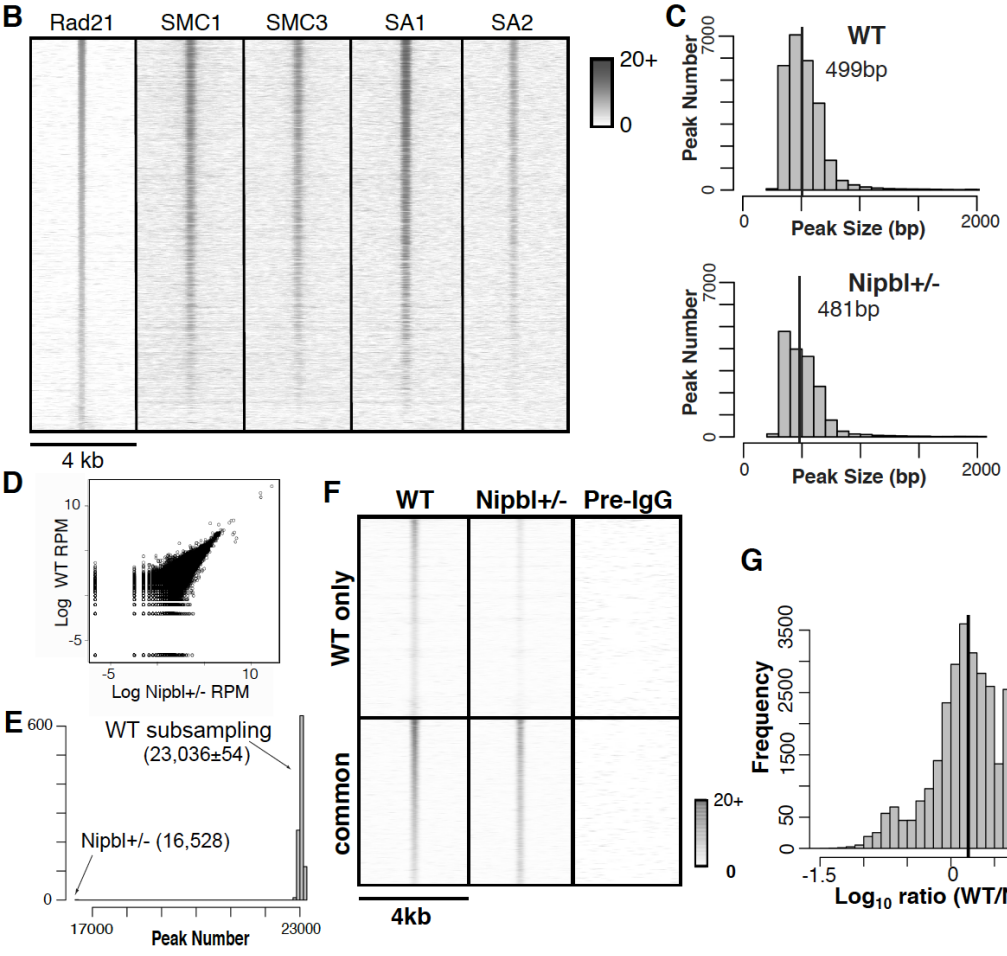
E. Histogram showing the distribution of total peaks called. A comparable number of reads to the *Nipbl*+/- mutant dataset (i.e. 4,740,463) were sub-sampled from the wild type dataset, and peaks called using only the sub-sampled reads. This process was performed 1000 times to produce the histogram above. Mean values with standard deviations are shown.

F. Heatmap analysis of cohesin binding in wild type (WT) MEFs and corresponding peak signals in *Nipbl* +/- MEFs. The normalized (reads per million) tag densities in a 4 kb window around each peak are plotted, with peaks sorted from the highest number of tags in the wild type to the lowest. Peaks are separated into two categories, those that are found only in wild type ("WT only") and those that overlap between wild type and *Nipbl* +/- ("common"). The color scale indicates the number of tags in a given region.

G. Histogram of the ratio between normalized (reads per million total reads) wild type and mutant reads in peaks common to both. Positive values indicate more wild type tags. The black line indicates the mean ratio between wild type and mutant tag counts.

A

Rad21 ChIP-seq	# Tags	% aligned	# Peaks	p-value	FDR
Control MEF	5,501,724	64%	25,407	< $1X10^{-4}$	< 3%
Nipbl (+/-) MEF	4,740,463	75%	16,528	< $1X10^{-4}$	< 3%



3.3.2 The relationship of cohesin binding sites with CTCF binding sites and CTCF motifs

It has been reported that cohesin binding significantly overlaps with CTCF sites and depends on CTCF [26, 27]. A study in mouse embryonic stem cells (mESCs) showed, however, that there is only a limited overlap between CTCF- and Nipbl-bound cohesin sites, suggesting that there are two categories of cohesin binding sites, and the latter may be particularly important for gene activation [36]. Other studies also revealed that ~20-30% of cohesin sites in different human cancer cell lines and up to ~50% of cohesin sites in mouse liver appear to be CTCF-free [38, 39]. Some of these non-CTCF sites overlap with sequence-specific transcription factor binding sites in a cell type-specific manner, highlighting the apparent significance of CTCF-free cohesin sites in cell type-specific gene expression [38, 39]. *De novo* motif discovery by MEME identified the CTCF motif to be the only significant motif associated with cohesin binding sites in our MEFs (Figure 3.2A). Comparing our cohesin peaks with experimentally determined CTCF binding peaks in MEFs [36], we found that approximately two-thirds of cohesin binding sites detected by Rad21 ChIP overlapped CTCF binding sites (Figure 3.2B). This is comparable with what was initially observed in mouse lymphocytes [26] and HeLa cells [27] using antibodies against multiple cohesin subunits. In contrast to recent studies reporting that almost all the CTCF binding sites overlap with cohesin [39], our results show that less than 60% of CTCF binding sites are co-occupied with cohesin (Figure 3.2B). This is consistent with the fact that CTCF binds and functions independently of cohesin at certain genomic regions [30, 37, 47, 48].

Presence of a CTCF motif closely correlates with CTCF binding: over 90% of cohesin binding sites overlapping with CTCF peaks contain CTCF motifs (Figure 3.2C). In contrast, less than half of cohesin binding sites harbor CTCF motifs in the absence of CTCF binding. Cohesin binding sites without CTCF binding tend to be highly deviated from a CTCF motif, reflecting a CTCF-independent mechanism of recruitment (Figure 3.2D). Interestingly, a small population of cohesin-CTCF overlapping sites that also lack any CTCF motif, suggesting an alternative way by which cohesin and CTCF bind to these regions (Figure 3.2C and D).

3.3.3 Nipbl reduction affects cohesin binding at CTCF-bound sites and repeat regions

In mESCs, it was proposed that Nipbl and CTCF recruit cohesin to different genomic regions, implying that cohesin binding to CTCF sites may be Nipbl-independent [36]. We noticed that when we ranked cohesin binding sites based on the read number in wild type peaks, they matched closely with the ranking of cohesin binding sites in mutant MEFs, indicating that the decrease of cohesin binding is roughly proportional to the strength of the original binding signals (Figure 3.2E). This suggests that most cohesin binding sites have similar sensitivity to Nipbl reduction. Importantly, CTCF binding signals also correlate with the ranking of cohesin binding, indicating that CTCF-bound sites are in general better binding sites for cohesin (Figure 3.2E). Because of this, they satisfy the peak definition despite the decrease of cohesin binding in mutant cells (Figure 3.1F and G, and Figure 3.6B). This explains why CTCF-bound cohesin sites are apparently enriched in the sites that are common to both wild type and mutant cells (Figure 3.2F).

Based on the above data, Yen-Yun Chen in our lab further clarified the role of Nipbl in cohesin binding to CTCF sites. She compared the effect of Nipbl reduction on cohesin binding to representative sites, which have either CTCF binding or a CTCF motif or both (Figure 3.3A). Decreased cohesin binding was observed at all sites tested by manual ChIP-qPCR in Nipbl mutant MEFs, correlating with the decreased Nipbl binding (Figure 3.3A). Consistent with the genome-wide ChIP-seq analysis (Figure 3.1D), control histone H3 ChIP-qPCR revealed no significant differences at the corresponding regions, indicating that the decreased cohesin binding is not due to generally decreased ChIP efficiency in mutant MEFs compared to the wild type MEFs (Figure 3.3C). Similar results were obtained using a small interfering RNA (siRNA) specific for *Nipbl* (Figure 3.3A, bottom), which reduced Nipbl to a comparable level as in mutant cells (western blot in Figure 3.3D and RT-qPCR results in Table 3.2). This demonstrates the specificity of the Nipbl antibody and confirms that the decreased cohesin binding seen in Nipbl mutant MEFs is the direct consequence of reduced Nipbl (Figure 3.3A). Thus, Nipbl also functions in cohesin loading at CTCF sites.

Figure 3.2. Most of cohesin binding sites contain CTCF motifs.

A. De novo motif search of cohesin binding sites using MEME. The CTCF motifs identified at the cohesin binding sites in WT and mutant MEFs are compared to the CTCF motif obtained from CTCF ChIP-seq data in MEFs (GSE22562) [36]. E-values are 5.5e-1528 (cohesin binding sites in WT MEFs), 6.6e-1493 (cohesin binding sites in Nipbl MEFs), and 2.6e-1946 (CTCF binding sites in MEFs), respectively.

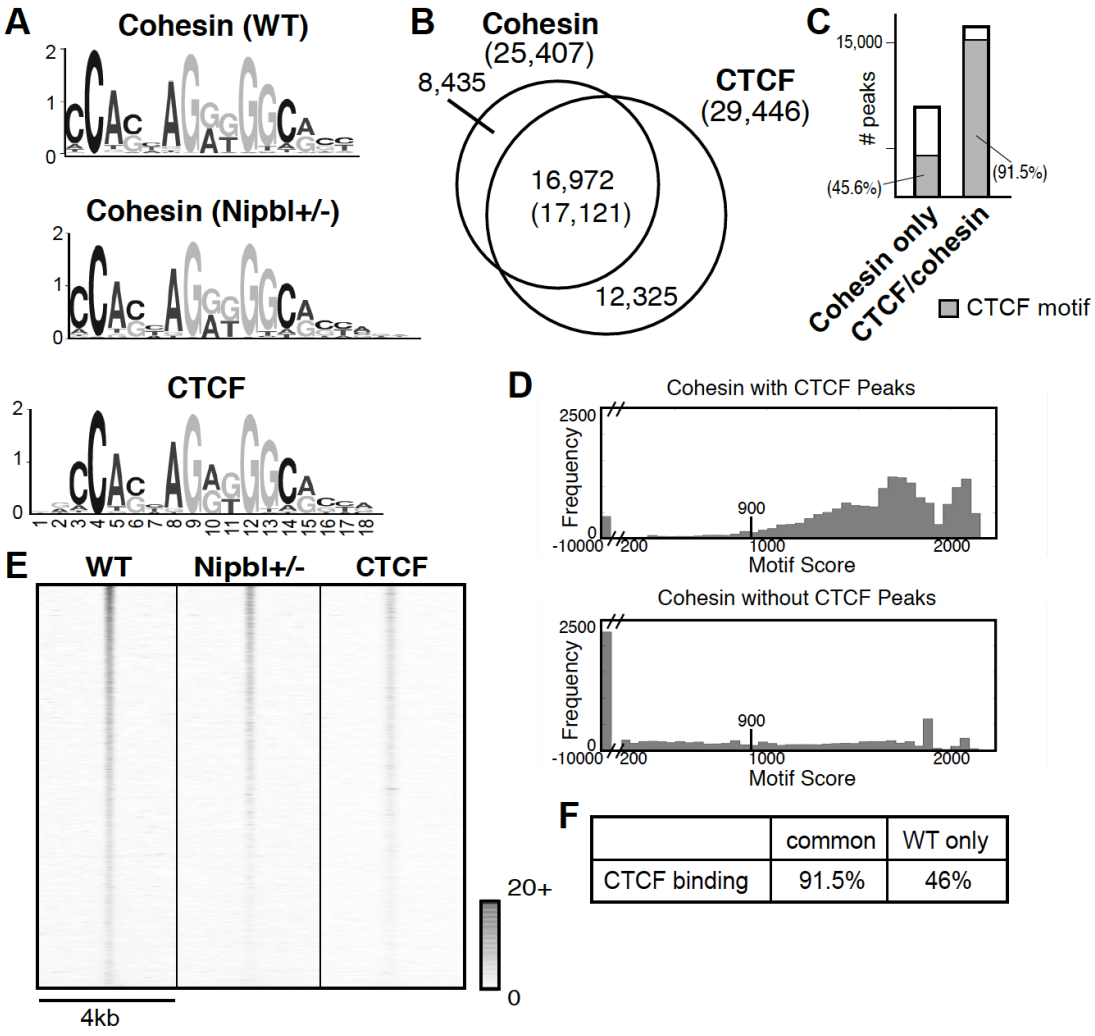
B. Overlap of cohesin binding sites with CTCF binding sites. The number in the parenthesis in overlapping regions between cohesin and CTCF binding represents the number of CTCF binding peaks.

C. Presence of CTCF motifs in cohesin only and cohesin/CTCF binding sites. Shaded area represents binding sites containing CTCF motifs defined in (A) (FDR 4.7%).

D. The CTCF motif score distribution for all cohesin peaks that overlap with a CTCF peak (top) and that don't overlap with a CTCF peak (bottom). Note that the X axis is discontinuous and scores less than 200 are placed in the single bin in each figure. For peaks that contained multiple CTCF motifs, we report the maximum score for the peak. The score threshold (900 with FDR 4.7%) is marked in each figure.

E. Heatmap comparison of cohesin ChIP-seq tags in WT MEFs and Nipbl mutant MEFs with CTCF ChIP-seq tags at the corresponding regions in wild type MEFs [36] as indicated at the top. The normalized (reads per million total reads) tag densities in a 4 kb window (±2kb around the center of all the cohesin peaks) are plotted, with peaks sorted by the number of cohesin tags (highest at the top) in WT MEFs. Tag density scale from 0 to 20 is shown.

F. Percentages of CTCF binding in cohesin binding sites common or unique to WT MEFs

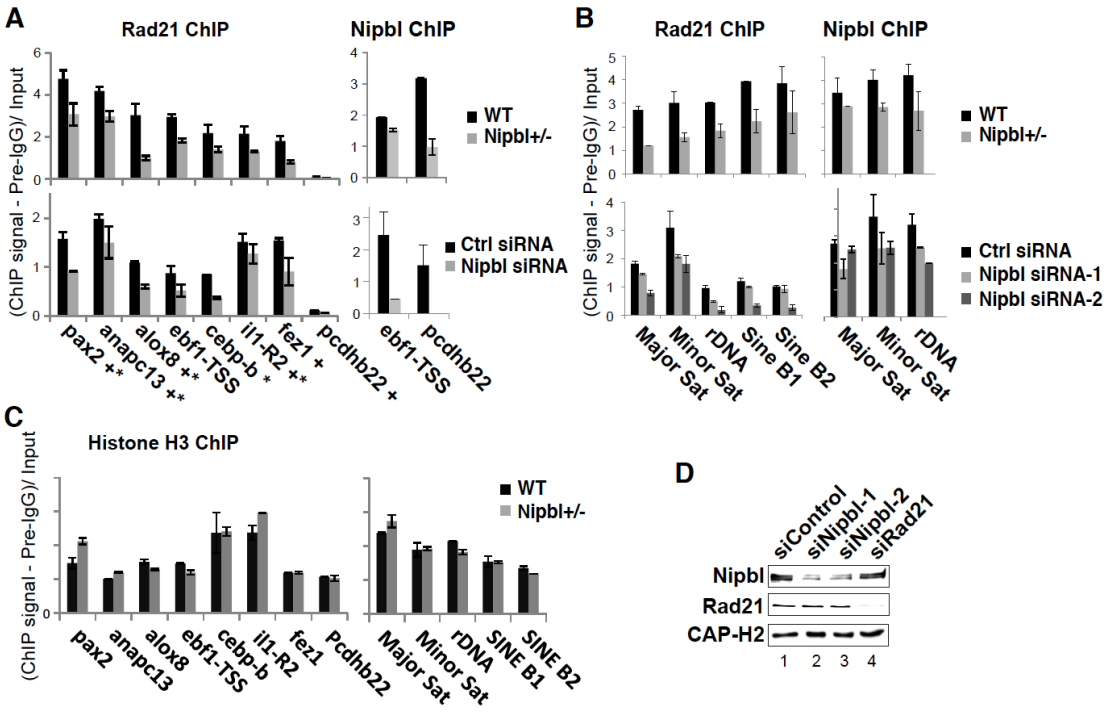


	common	WT only
CTCF binding	91.5%	46%

Figure 3.3. Nipbl reduction decreases cohesin binding.

A. Manual ChIP-q-PCR of cohesin binding sites using anti-Rad21 antibody in mutant and wild type MEFs (top panel) and Nipbl or control siRNA-treated MEFs (bottom panel) as indicated. Representative examples of Nipbl ChIP are also shown. "+" indicates CTCF binding and "*" indicates the presence of motif. Depletion efficiency and specificity of Nipbl siRNA was examined by RT-q-PCR (Table 3.1).

B. Similar manual ChIP-q-PCR analysis as in (A) of repeat regions in wild type and Nipbl mutant MEFs (top) and control and Nipbl siRNA-treated MEFs (bottom).



Provided by Yen-Yun Chen

3.3.4 Cohesin distribution patterns in the genome and enrichment in promoter regions

In order to gain insight into how the weakening of cohesin binding may affect gene expression in mutant cells, the distribution of cohesin binding sites in the genomes of both wild type and mutant MEFs were examined. Approximately 50% of all cohesin binding sites are located in intergenic regions away from any known genes (Figure 3.4A). However, there is a significant enrichment of cohesin binding in promoter regions, and to a lesser extent in the 3' downstream regions, relative to the random genomic distribution generated by sampling from pre-immune ChIP-seq reads (Figure 3.4B). Similar promoter and downstream enrichment has been observed in mouse and human cells [26, 27, 36, 38, 43] as well as in Drosophila [49]. Promoter enrichment is comparable in both wild type and *Nipbl* mutant MEFs, constituting ~10% of all the cohesin binding sites (Figure 3.4A). Thus, there is no significant redistribution or genomic region-biased loss of cohesin binding sites in Nipbl mutant cells.

3.3.5 Cohesin-bound genes are sensitive to Nipbl haploinsufficiency

Based on the significant enrichment of cohesin binding in the promoter regions, we next examined the correlation between cohesin binding to the gene regions and the change of gene expression in mutant MEFs using a KS test. This is a nonparametric test for comparing peak binding sites with gene expression changes in the mutant MEFs (Figure 3.5). Genes that displayed the greatest expression change in mutant MEFs compared to the wild type MEFs showed a strong correlation with cohesin binding to the gene region, indicating that direct binding to the target genes is the major mechanism by

which cohesin mediates gene regulation in a Nipbl dosage-sensitive fashion (Figure 3.5A, left). Random sampling of a comparable number of simulated peaks in the gene regions yielded no correlation (Figure 3.5D, left). Interestingly, cohesin binding to the gene region correlates better with decreased gene expression than increased expression in mutant cells, indicating that gene activation, rather than repression, is the major mode of cohesin function at the gene regions (Figure 3.5A, middle).

When analyzed separately, cohesin binding to the promoter regions (+2.5kb to -0.5kb of transcription start sites (TSS) (Figure 3.5A, right)) showed the highest correlation (*p*-value=3.3e-09) compared to the gene body and downstream (Figure 3.5B). Thus, cohesin binding to the promoter regions is most critical for gene regulation. Similar to the entire gene region, cohesin binding correlates more significantly with a decrease in gene expression in mutant cells, which is particularly prominent at the promoter regions compared to gene bodies or downstream, indicating the significance of cohesin binding to the promoter regions in gene activation (Figure 3.5C). Although cohesin and CTCF binding closely overlapped at promoter regions in HeLa cells [27], the overlap of CTCF binding with cohesin in MEFs is lower in the promoter regions (54%) than that in the intergenic regions (67%) [36]. Consistent with this, there is no significant correlation between CTCF binding in the promoter regions and gene expression changes in *Nipbl* mutant MEFs (*p*-value=0.28) by KS test (Figure 3.5C, right), further indicating the cohesin-independent and Nipbl-insensitive function of CTCF in gene regulation. Taken together, the results suggest that cohesin binding to gene regions (in particular, to

promoters) is significantly associated with gene activation that is sensitive to Nipbl haploinsufficiency.

3.3.6 Identification of cohesin target genes sensitive to Nipbl haploinsufficiency

The results above indicate that cohesin-bound genes sensitive to a partial loss of Nipbl can be considered to be Nipbl/cohesin target genes. Among 218 genes that changed expression significantly in mutant cells compared to the wild type (>1.2-fold change, *p*-value < 0.05) [17], we found that more than half (115 genes) were bound by cohesin, and thus can be considered Nipbl/cohesin target genes (Table 3.3). This is a conservative estimate of the number of direct target genes since cohesin binding sites beyond the upstream and downstream cut-offs (2.5 kb) were not considered for the analysis. Consistent with the KS test analysis (Figure 3.5), ~74% of these cohesin target genes were downregulated in mutant cells, indicating that the positive effect of cohesin on gene expression is particularly sensitive to partial reduction of Nipbl (Table 3.3).

Figure 3.4. Cohesin binding site distribution in the genome in MEFs.

A. Percentage distribution of cohesin peaks in genomic regions. "Promoter" and "Downstream" is defined as 2500bp upstream of the transcription start site (TSS) and 500 bp downstream of the TSS, and "Downstream" represents 500 bp upstream of transcription termination site (TTS) and 2500 bp downstream of TTS. The 3' and 5' untranslated regions (UTRs) are defined as those annotated by the UCSC genome browser minus the 500 bp interior at either the TSS or TTS. When a peak overlaps with multiple regions, it is assigned to one region with the order of precedence of promoter, 5' UTR, Intron, Exon, 3'UTR, downstream, and intergenic.

B. Enrichment of cohesin peaks across genomic regions as compared to randomly sampled genomic sequence. A comparable number of peaks (25,407 and 16,528 peaks in wild type and mutant MEFs, respectively), with the same length as the input set, were randomly chosen 1000 times and the average used as a baseline to determine enrichment in each genomic region category.

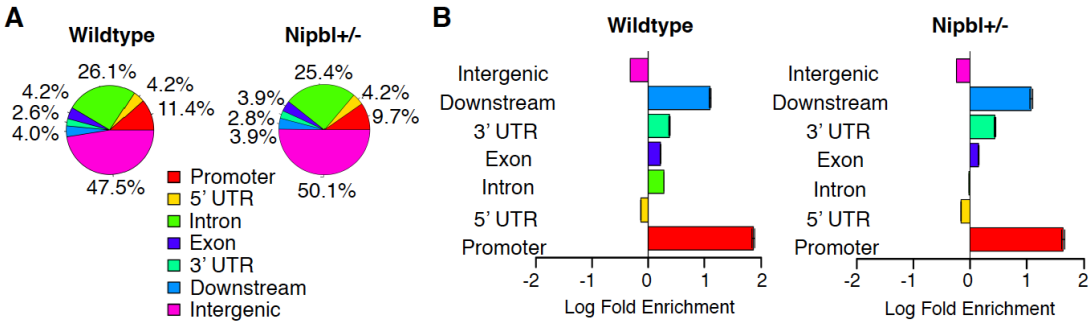

A
Wildtype
Nipbl+/-
26.1%
4.2%
11.4%
4.2%
2.6%
4.0%
47.5%
25.4%
4.2%
9.7%
3.9%
2.8%
3.9%
50.1%
Promoter
5' UTR
Intron
Exon
3' UTR
Downstream
Intergenic
B
Wildtype
Nipbl+/-
Intergenic
Downstream
3' UTR
Exon
Intron
5' UTR
Promoter
-2
-1
0
1
2
Log Fold Enrichment

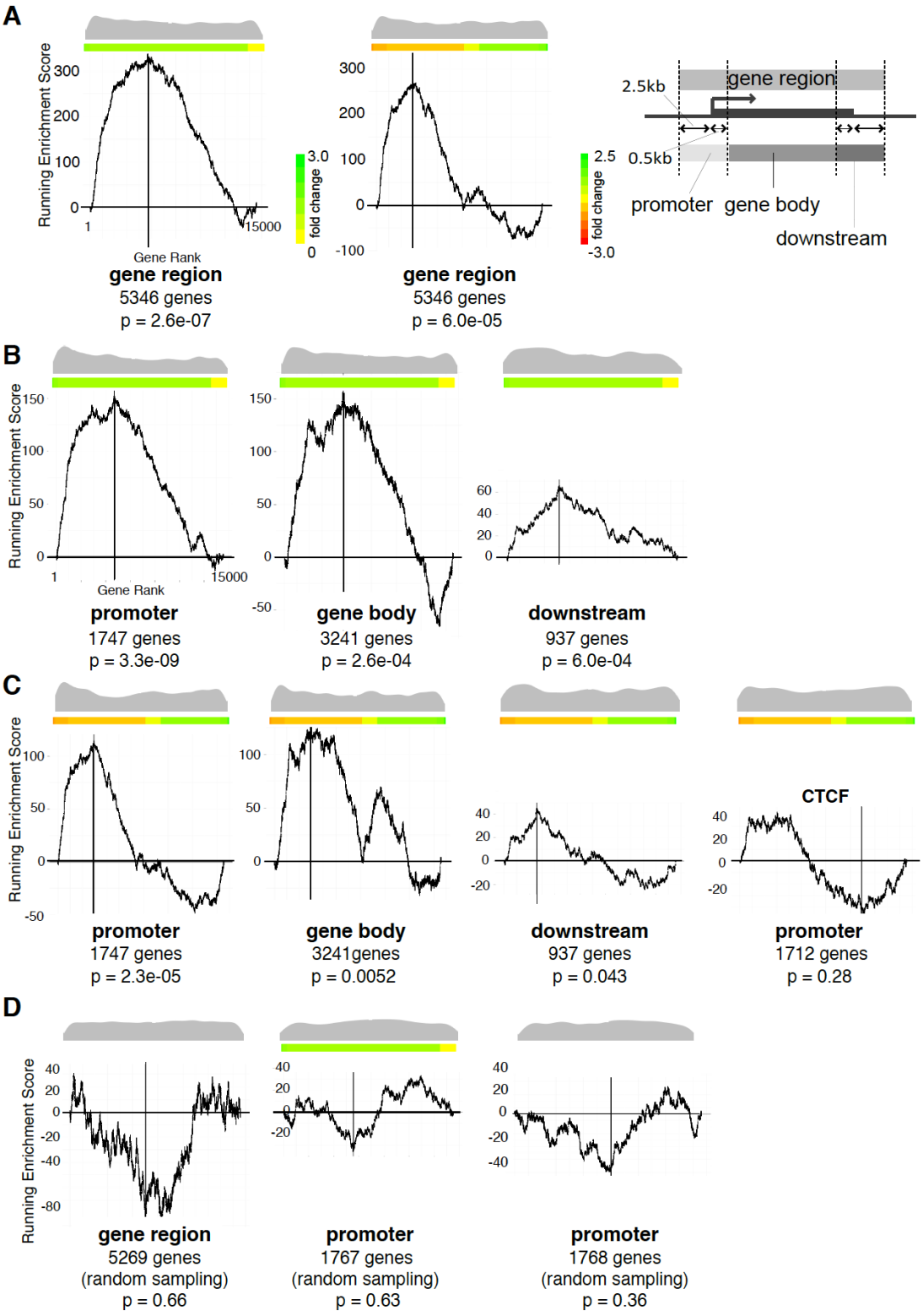
Figure 3.5. Correlation of cohesin binding and gene expression changes in mutant MEFs.

A. KS test indicating the degree of cohesin binding to genes changing expression in *Nipbl* +/- MEFs. X-axis represents all 13,587 genes from the microarray data [17] ranked by absolute fold expression changes from biggest on the left to the smallest on the right in the left panel. Fold changes are shown in different colors as indicated on the side. In the middle panel, gene expression changes were ranked from negative to positive with the color scale shown on the side. Both color scales apply to the rest of the Figure. The Y-axis is the running enrichment score for cohesin binding (see EXPERIMENTAL PROCEDURES for details). Distribution of cohesin-bound genes among 13,587 genes examined is shown as a beanplot [79] at the top, and the number of cohesin-bound genes and *p*-values are shown underneath. The schematic diagram showing the definition of the gene regions, promoter (2.5kb upstream and 0.5kb downstream of TSS), gene body, and downstream (2.5kb downstream and 0.5kb upstream of TTS) regions is shown on the right.

B. Similar KS test analysis as in (A), in which cohesin binding to the promoter, gene body, and downstream regions are analyzed separately.

C. Genes are ranked by expression changes from positive on the left to negative on the right. Fold changes are shown by different colors as indicated on the right. CTCF binding to promoter regions (GSE22562) [36] was analyzed for a comparison.

D. Lack of correlation between the mutant expression changes and randomly chosen genes are shown on the right as a negative control.


A
Running Enrichment Score
300
200
100
0
1
15000
Gene Rank
3.0
fold change
0
gene region
5346 genes
p = 2.6e-07
300
200
100
0
-100
3.0
fold change
0
-3.0
gene region
5346 genes
p = 6.0e-05
2.5kb
gene region
0.5kb
promoter
gene body
downstream
B
Running Enrichment Score
150
100
50
0
1
15000
Gene Rank
promoter
1747 genes
p = 3.3e-09
150
100
50
0
-50
gene body
3241 genes
p = 2.6e-04
60
40
20
0
downstream
937 genes
p = 6.0e-04
C
Running Enrichment Score
100
50
0
-50
promoter
1747 genes
p = 2.3e-05
100
50
0
gene body
3241genes
p = 0.0052
40
20
0
-20
downstream
937 genes
p = 0.043
CTCF
40
20
0
-20
promoter
1712 genes
p = 0.28
D
Running Enrichment Score
40
20
0
-20
-40
-80
gene region
5269 genes
(random sampling)
p = 0.66
40
20
0
-20
promoter
1767 genes
(random sampling)
p = 0.63
40
20
0
-20
-40
promoter
1768 genes
(random sampling)
p = 0.36

Many of these Nipbl/cohesin-target genes contain cohesin binding sites in more than one region (promoter, gene body and/or downstream), suggesting their collaborative effects (Figure 3.6A). In particular, the promoter binding of cohesin is often accompanied by its binding to the gene body. However, binding pattern analysis revealed no significant correlation between a particular pattern and/or number of cohesin binding sites and gene activation or repression (Figure 3.6A). Rad21 ChIP-seq signal intensity profiles of several cohesin target genes (as defined above) reveal decreased cohesin binding in mutant cells at the binding sites originally observed in the wild type cells, supporting the notion that gene expression changes are the direct consequence of the reduced cohesin binding (Figure 3.3A; Figure 3.6B, top). There are other genes, however, that did not change expression significantly in mutant MEFs, but nevertheless also have reduced cohesin peaks nearby (Figure 3.6B, bottom), suggesting that cohesin binding is not the sole determinant of the gene's expression status and that its effect is context-dependent.

Gene ontology analysis revealed that the target genes bound by cohesin at the promoter regions and affected by Nipbl deficiency are most significantly enriched for those involved in development (Table 3.4). The results suggest a direct link between diminished Nipbl/cohesin and the dysregulation of developmental genes, which contributes to the CdLS phenotype.

3.3.8 Cohesin binding correlates significantly with H3K4me3 at the promoter

To investigate the genomic features associated with cohesin target genes, we examined the chromatin status of the target gene promoters. We found that cohesin peaks closely overlap with the peaks of H3K4me3, a hallmark of an active promoter, in a promoter-specific manner (Figure 3.8A). In contrast, there are only minor peaks of H3K27me3 and even less H3K9me3 signal at cohesin-bound promoters, consistent with the results of the KS-test revealing the significant association of cohesin binding to the promoter regions with gene activation rather than repression (Figure 3.5C). Interestingly, however, promoter binding of cohesin was found in genes with different expression levels in wild type MEFs, revealing no particular correlation with high gene expression (Figure 3.8B). Cohesin target genes defined above (Table 3.3) also exhibit variable expression levels in wild type MEFs (Figure 3.8B). Thus, their expression is altered in Nipbl mutant cells regardless of the original expression level in wild type cells, indicating that cohesin binding contributes to gene expression but does not determine the level of transcription *per se*.

When cohesin-bound genes were categorized in five different groups based on the gene expression status in wild type MEFs, significant H3K4me3 enrichment was observed even in the cohesin-bound promoters of genes with low expression, compared to cohesin-free promoters of genes with a similar expression level (Figure 3.8C). Bivalent (H3K4me3 and H3K27me3) modifications are also enriched in the lowest gene expression category (Figure 3.8C). Taken together, the results reveal that there is a close

correlation between cohesin binding and H3K4me3 in the promoter regions regardless of the expression levels of the corresponding genes.

Table 3.1. The list of PCR primers

Unique regions ChIP primers	
pax2-F	CTGGCACTGACATCTTGTGG
pax2-R	TGGGACCTGTAGTCCTGACC
anapc13-F	TCCTAAGCCGTCCTGTAGTCC
anapc13-R	GGGTGTCCATCATCTGAGTCC
alox8-F	GTATGAGGTGGGCCTGAGTG
alox8-R	AAGCCCTGCCTAAATGTGTG
ebf1-F	AACTGAGCCTTAGGGGAAGC
ebf1-R	TCAGGGTTCAATCTCCAAGG
cebpb-F	AGAGTTCTGCTTCCCAGGAGT
cebpb-R	GGAAACAGATCGTTCCTCCA
il1R2-F	TGGAGGCAGTGGAAGAATCA
il1R2-R	ATCCTTGGCAGTGAACCAGA
fez1-F	GAGGGTGGGACGTATTTCAGT
fez1-R	CAGCCTTCTTTCCCTCACAA
pcdhb22-F	GCAGTAATGCCAGCAATGG
pcdhb22-R	TCCAGTTGGTTGGGTTTCAT
RT-qPCR primers	
Rnh1-F (Housing keeping gene)	TCCAGTGTGAGCAGCTGAG
Rnh1-R (Housing keeping gene)	TGCAGGCACTGAAGCACCA
Nipbl-F	AGTCCATATGCCCCACAGAG
Nipbl-R	ACCGGCAACAATAGGACTTG
Rad21-F	AGCCAAGAGGAAGAGGAAGC
Rad21-R	AGCCAGGTCCAGAGTCGTAA
Cebpb-F	GCGGGGTTGTTGATGTTT
Cebpb-R	ATGCTCGAAACGGAAAAGG
Cebpd-F	ACAGGTGGGCAGTGGAGTAA
Cebpd-R	GTGGCACTGTCACCCATACA
Ebf1-F	GCGAGAATCTCCTTCAAGACTTC
Ebf1-R	ACCTACTTGCCTTTGTGGGTT
Il6-F	TAGTCCTTCCTACCCCAATTTCC
Il6-R	TTGGTCCTTAGCCACTCCTTC
Avpr1a-F	TGGTGGCCGTGCTGGGTAATAG
Avpr1a-R	GCGGAAGCGGTAGGTGATGTC
Lpar1-F	ATTTCACAGCCCCAGTTCAC
Lpar1-R	CACCAGCTTGCTCACTGTGT
Adm-F	TATCAGAGCATCGCCACAGA
Adm-R	TTAGCGCCCACTTATTCCAC
Cebpb 3C primers	
cebpb-promoter	ACTCCGAATCCTCCATCCTT
cebpb-region-b	CCTGCCCTGTATCAAAGCAT
cebpb-region-a	CTGCCCAAATCAGTGAGGTT
cebpb-region-c	CCTCTGTGAGGTCTGGTCGT

cebpb-promoter-R	GGTGGCTGCGTTAGACAGTA
cebpb-region-a-R	GTTGTATCCCAAGCCAGCTC
cebpb-region-b-R	CTCCCCACTCTGTTCAGGAC
cebpb-region-c-R	TAACAGCAGGGATGGGTTCT

Table 3.2. Nipbl and Rad21 depletion levels in mutant and siRNA-treated MEFs

Gene	Nipbl+/- mutant	Nipbl siRNA	Rad21 siRNA
Nipbl	0.68±0.003	0.68±0.001	1.04±0.051
Rad21	0.94±0.021	0.99±0.021	0.26±0.018
CTCF	0.95±0.050	0.96±0.066	0.84±0.074

Table 3.3. Gene expression changes and cohesin binding status

	Total	Cohesin binding				
		Gene region	Promoter	Gene body	Downstream	None
Total	218	115	61	83	20	103
Up-regulated	62	30	14	22	6	32
Down-regulated	156	85	47	61	14	71

(Fold change>1.2, *p*-value<0.05)

Table 3.4. Ontology analysis of cohesin target genes.
Biological processes enriched in cohesin target genes with cohesin binding at either promoters or gene regions. "Gene number" is the number of cohesin target genes that belong to a specific category; "Expected number" is the expected gene numbers that belong to a specific category at random.

Altered gene expression in Nipbl+/- MEFs associated with cohesin binding to the promoters

Biological process	Gene number	Expected number	Enrichment	P value	Genes
development	18	7.55	2.38	2.96E-04	Avpr1a, Dner, Fgf7, Thbd, Hoxa5,Hoxb5, Cebpa, Cebpb, Rcan2, Lama2, Ebf1, Klf4, Hunk, Tgfb3, Irx5, Odz4, Ptpre, Lpp
metabolism	33	22	1.50	2.90E-03	Dner, Acvr2a, Hoxa5, Hoxb5, Trib2, Satb1, Cebpa, Cebpb, Gstm2, Amacr, Cd55, Dhrs3, Grk5, Ell2, Serpinb1a, Cyp1b1, Chst1, Hsd3b7, Aldh1a7, Npr3, Man2a1, Klf4, Hunk, Prkd1,Prdx5, Ercc1, Irx5, Odz4, Sox11, Ptpre, Ccrn4l, Rgnef, Bcl11b
cell communication	21	11.53	1.82	2.96E-03	Dner, Acvr2a, Trib2, Cd55, Grk5, Hunk, Odz4, Ptpre, Rgnef, Avpr1a, Fgf7, Thbd, Fam43a, Rcan2, Socs3, Lama2, Cxcr7, Tpcn1, Rerg, Tgfb3, Lpp
immune system	14	6.81	2.06	6.44E-03	Dner, Cd55, Hunk, Ptpre, Thbd, Lama2, Cxcr7, Cebpa, Cebpb, Gstm2, Klf4, Prdx5, Fcgrt, Cd302

Altered gene expression in Nipbl+/- MEFs associated with cohesin binding to the gene regions

Biological process	Gene number	Expected number	Enrichment	P value	Genes
immune system	30	12.83	2.34	6.60E-06	Klf4, Dner, Thbd, Cd55, Lama2, Cd302, Cxcr7, Hunk, Cebpa, Cebpb, Gstm2, Fcgrt, Prdx5, Fmod, Crlf1, Prelp, Svep1, Plac8, Heph, Swap70, Mxra8, Sdc2, Colec12, Pcolce2, Flt4, Gbp1, Hck, Dusp14, Cd109, Ptpre
cell adhesion	19	6.22	3.05	1.33E-05	Dner, Cd55, Lama2, Fmod, Prelp, Svep1, Plac8, Heph, Mxra8, Sdc2, Colec12, Pcolce2, Flt4, Hck, Ptpre, Rerg, Vcan, Odz4, Rgnef
cell communication	41	21.72	1.89	1.65E-05	Dner, Cd55, Lama2, Fmod, Prelp, Svep1, Heph, Sdc2, Colec12, Pcolce2, Flt4, Hck, Ptpre, Rerg, Vcan, Odz4, Rgnef, Thbd, Cxcr7, Hunk, Crlf1, Dusp14, Cd109, Rcan2, Socs3, Fam43a, Trib2, Grk5, Tpcn1, Avpr1a, Fgf7, Acvr2a, Figf, Myh3, Tob1, Acvrl1, Moxd1, Tgfb3, Lpp, Wnt4
development	30	14.22	2.11	4.81E-05	Dner, Lama2, Fmod, Prelp, Heph, Sdc2, Colec12, Pcolce2, Flt4, Ebf1, Hck, Ptpre, Vcan, Odz4, Thbd, Hunk, Crlf1, Rcan2, Socs3, Avpr1a, Fgf7, Figf, Myh3, Tgfb3, Lpp, Klf4, Cebpa, Cebpb, Hoxa5, Hoxb5, Irx5

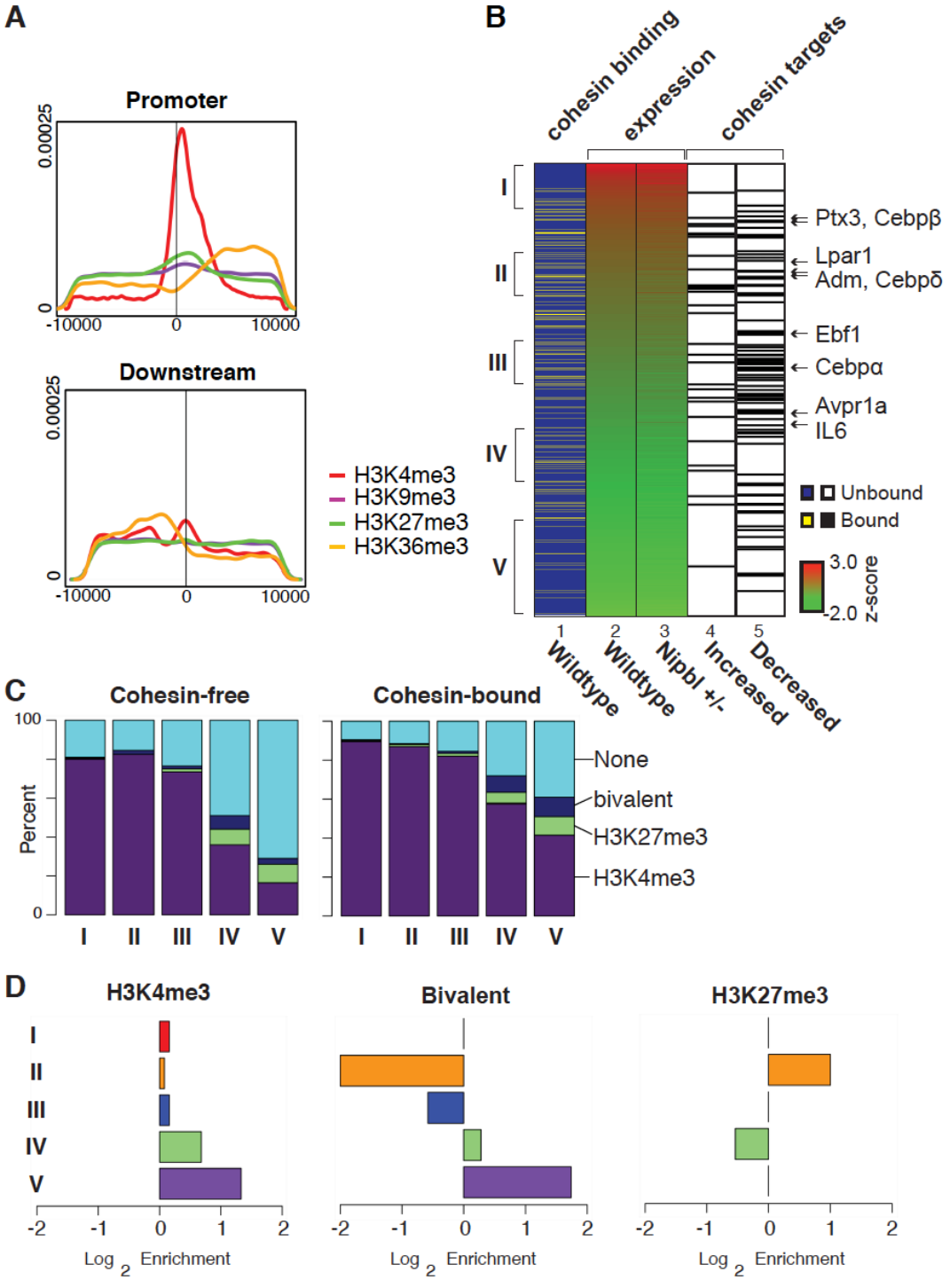
metabolism	57	41.44	1.38	1.91E-03	Dner, Heph, Pcolce2, Flt4, Hck, Ptpre, Odz4, Hunk, Klf4, Cebpa, Cebpb, Hoxa5, Hoxb5, Irx5, Cd55, Svep1, Rgnef, Dusp14, Cd109, Trib2, Grk5, Acvr2a, Acvrl1, Moxd1, Prdx5, Swap70, Satb1, Amacr, Dhrs3, Ell2, Npr3, Man2a1, Prkd1, Cyp1b1, Serpinb1a, Chst1, Hsd3b7, Aldh1a7, H6pd, Serpine2, Cyp7b1, P4ha2, Larp6, Mrps11, Aox1, Hdac5, Cpxm1, Eno2, Sox11, Prkcdbp, Ccrn4l, Ercc1, Pqlc3, Bcl11b

Figure 3.6. Enrichment of H3K4me3 at the promoters of cohesin-bound genes.
A. Density of histone modifications within 10kb of cohesin peaks found in the promoter or downstream regions. Histone methylation data was downloaded from NCBI (GEO: GSE26657). Tags within a 10 kb window around cohesin peaks located in a promoter region were counted and normalized to the total number of tags (reads per million) and used to generate a density plot.

B. Expression status of cohesin target genes. Genes are ranked by their expression status (shown as a z-score) in wild type MEFs (lane 2), and those genes with cohesin binding at the promoter regions are indicated by yellow lines (lane 1). The expression status of the corresponding genes in Nipbl mutant cells is also shown (lane 3), and the cohesin target genes (Table 3.2) (either upregulated (lane 4) or downregulated (lane 5) in mutant cells) are indicated by black lines. Genes in the adipogenesis pathway are indicated with arrows on the right. Five clusters (I through V) of two hundred cohesin-bound genes each in wild type MEFs according to the expression levels are indicated on the left, which were used for the analysis in (C) and (D).

C. The numbers of cohesin target genes containing histone marks in the promoter were tallied for the categories I through V from (B). As a control, the cohesin-free gene directly below each cohesin target gene was also tallied and plotted. H3K4me3, H3K9me3, H3K27me3, bivalent (H3K4me3 and H3K27me3), and the promoters with none of these marks ("None") are indicated. There is almost no signal of H3K9me3 in these categories.

D. Enrichment plot of H3K4me3, H3K27me3, and bivalent (H3K4me3 and K27me3) in promoters of cohesin-bound genes versus cohesin-free genes in the five expression categories as in (C) is shown.


A
Promoter
0.00025
0
-10000
0
10000
Downstream
0.00025
0
-10000
0
10000
H3K4me3
H3K9me3
H3K27me3
H3K36me3
B
cohesin binding
expression
cohesin targets
I
II
III
IV
V
Ptx3, Cebpβ
Lpar1
Adm, Cebpδ
Ebf1
Cebpα
Avpr1a
IL6
Unbound
Bound
3.0
-2.0
z-score
1
2
3
4
5
Wildtype
Wildtype
Nipbl +/-
Increased
Decreased
C
Cohesin-free
Cohesin-bound
100
Percent
0
I
II
III
IV
V
None
bivalent
H3K27me3
H3K4me3
D
H3K4me3
Bivalent
H3K27me3
-2
-1
0
1
2
Log 2 Enrichment

3.4 Discussion

In this study, we used MEFs derived from *Nipbl* heterozygous mutant mice to analyze the effect of *Nipbl* haploinsufficiency (the primary cause of CdLS) on cohesin binding and its relationship to gene expression. We found a genome-wide decrease in cohesin binding even at CTCF sites and repeat regions, indicating the high sensitivity of cohesin binding to even a partial reduction of the Nipbl protein. Importantly, the expression of genes bound by cohesin, particularly at the promoter regions, is preferentially altered in response to Nipbl reduction. While some genes are activated, the majority of cohesin-bound genes are repressed by decreased cohesin binding, indicating the positive role of cohesin in this context. This is consistent with the significant enrichment of H3K4me3 at the promoters of cohesin-bound genes. Our results indicate that more than 50% of genes whose expression is altered significantly in *Nipbl* haploinsufficient cells are cohesin target genes directly influenced by decreased cohesin binding at the individual gene regions. One consequence of reduced cohesin binding at the promoter region is a decrease of a specific long-distance chromatin interaction, raising the possibility that cohesin-dependent higher-order chromatin organization in the nucleus may be globally altered in CdLS patient cells.

3.4.1 Nipbl functions in cohesin loading at both CTCF and non-CTCF sites

In mESCs, it was suggested that Nipbl is involved in cohesin binding to only a subset of cohesin binding sites, which are largely distinct from CTCF-bound sites [36]. However, we found that Nipbl binds to, and its haploinsufficiency decreased cohesin binding to, CTCF sites in MEFs. A similar decrease of cohesin binding was observed at

both CTCF insulators and non-CTCF sites in the *b-globin* locus in *Nipbl* +/- fetal mouse liver [31]. Furthermore, during differentiation in mouse erythroleukemia cells, both Nipbl and cohesin binding is concomitantly increased at these sites [31]. Therefore, while cohesin was suggested to slide from the Scc2 (Nipbl homolog)-dependent loading sites in yeast [50, 51], Nipbl is present and appears to directly affect cohesin loading at CTCF sites in mammalian cells. Nipbl, rather than cohesin, interacts with Mediator and HP1, and appears to recruit and load cohesin onto genomic regions enriched for Mediator and HP1 for gene activation and heterochromatin assembly, respectively [36, 41]. In contrast, cohesin, and not Nipbl, primarily interacts with CTCF [41, 52]. Thus, for cohesin binding to CTCF sites, we envision that cohesin initially recruits Nipbl that in turn stably loads cohesin onto CTCF sites.

A recent study indicated that almost all CTCF sites are bound by cohesin in primary mouse liver [39]. In MEFs, however, we found that ~42% of CTCF-bound sites appear to be cohesin-free. Furthermore, there is less overlap of cohesin and CTCF in the promoter regions compared to the intergenic regions, and little correlation between CTCF binding to the promoter and gene expression changes in *Nipbl* mutant cells was observed. Thus, in contrast to the cooperative function of cohesin and CTCF at distantly located insulator sites [32], cohesin and CTCF appear to have distinct functions at gene promoters. Distinct gene regulatory functions of CTCF and cohesin have also been reported in human cells [37]. Further study is needed to understand the recruitment specificity and functional relationship of cohesin and CTCF in gene regulation.

3.4.2 How does Nipbl haploinsufficiency affect cohesin target gene expression?

Cohesin binding to the gene body regions is found at many of the cohesin target genes. This may represent the cohesin binding at intragenic enhancer elements or may be related to Pol II pausing [53]. While cohesin was shown to facilitate Pol II elongation in Drosophila [54-56], cohesin together with CTCF in the intragenic region was found to cause Pol II pausing at the *PUMA* gene in human cells [57], suggesting that cohesin can have both positive and negative effects on transcriptional elongation in a context-dependent manner. Furthermore, not all the cohesin-bound genes changed expression in *Nipbl*+/- MEFs, echoing this notion that the effect of cohesin binding on gene expression is context-dependent. What determines the effects of cohesin binding at individual binding sites on gene expression requires further investigation.

3.4.3 The role of cohesin in the maintenance of gene expression

While there is now strong evidence for cohesin's role in chromatin organization and gene activation, whether cohesin is involved in initiation or maintenance of gene activation is less clear. Enrichment of cohesin binding at the transcription start sites and termination sites was observed previously in mouse immune cells with no significant correlation to gene expression [26]. Our genome-wide analysis also revealed that cohesin binding to the gene regions has no obvious relationship to the level of gene expression in wild type MEFs. And yet, a decrease in cohesin binding is associated with a tendency to downregulate these genes, indicative of the positive role of cohesin on gene expression, consistent with the enriched presence of H3K4me3 in promoter regions. We speculate that cohesin may not be the primary determinant of gene activation, but rather cohesin

binding may be important for maintaining gene expression status initially determined by sequence- and cell type-specific transcription factors. Similarly, enrichment of bivalent histone modifications in the promoters of cohesin-bound genes with very low expression suggests that cohesin also contributes to the maintenance of the poised state of these genes.

3.4.4 Nipbl haploinsufficiency vs. cohesin mutation

There are two different cohesin complexes in mammalian somatic cells that differ by one non-SMC subunit (i.e., SA1 (STAG1) or SA2 (STAG2)) [58, 59]. A recent report on SA1 knockout mice revealed some phenotypic similarity to what is seen in mice with *Nipbl* haploinsufficiency [43]. Interestingly, the *SA1* gene is one of the cohesin target genes that is slightly upregulated in *Nipbl* mutant cells [17]. Thus, together with the compensatory increase of *Nipbl* expression from the intact allele, there appears to be a feedback mechanism that attempts to balance the expression of *Nipbl* and cohesin in response to *Nipbl* mutation. The fact that upregulation was observed with the *SA1*, but not *SA2*, gene may reflect the unique transcriptional role of SA1 [43]. Interestingly, however, only 10% of 215 genes altered in *Nipbl* mutant MEFs are changed significantly in *SA1* KO MEFs [43]. This discrepancy may, as observed in Drosophila [60], reflect the different effects of decreased binding versus complete knockout of a cohesin subunit on target gene expression. It could also be a result of the decreased binding of the second cohesin complex, cohesin-SA2.

Cohesin binding was relatively uniformly decreased genome-wide in *Nipbl* haploinsufficient cells with no significant redistribution of cohesin binding sites. Point mutations of different subunits of cohesin cause CdLS and CdLS-like disorders with both overlapping and distinct phenotypes compared to CdLS cases caused by NIPBL mutations [9, 10, 13]. Non-overlapping effects of downregulation of different cohesin subunits have been reported in zebrafish [20, 25]. This may reflect an unequal role of each cohesin subunit in gene regulation and it is possible that some of the cohesin target genes may be particularly sensitive to a specific cohesin subunit mutation. For example, similar to the TBP-associating factors (TAFs) in TFIID [61], cohesin subunits may provide different interaction surfaces for distinct transcription factors, which would dictate their differential recruitment and/or transcriptional activities. Furthermore, recent studies provide evidence for cohesin-independent roles of NIPBL in chromatin compaction and gene regulation [62, 63]. Thus, disturbance of cohesin functions as well as impairment of cohesin-independent roles of NIPBL may collectively contribute to CdLS caused by NIPBL mutations.

Our results demonstrate that cohesin binding to chromatin is highly sensitive genome-wide (both at unique and repeat regions) to partial Nipbl reduction, resulting in a general decrease in cohesin binding even at strong CTCF sites. Many genes whose expression is changed by Nipbl reduction are actual cohesin target genes. Our results suggest that decreased cohesin binding due to partial reduction of NIPBL at the gene regions directly contributes to disorder-specific gene expression changes and the CdLS phenotype. This work provides important insight into the function of cohesin in gene regulation with

direct implications for the mechanism underlying *NIPBL* haploinsufficiency-induced CdLS pathogenesis.

3.5 Methods

3.5.1 Cells and antibodies

Mouse embryonic fibroblasts (MEFs) derived from E15.5 wild type and *Nipbl* mutant embryos were used as described previously [17]. In brief, mice heterozygous for *Nipbl* mutation were generated (*Nipbl* +/-) from gene-trap-inserted ES cells. This mutation resulted in a net 30-50% decrease in *Nipbl* transcripts in the mice, along with many phenotypes characteristic of human CdLS patients [17]. Wild type and mutant MEF cell lines derived from the siblings were cultured at 37°C and 5% CO_2 in DMEM (Gibco) supplemented with 10% fetal bovine serum and penicillin-streptomycin (50U/mL). Antibodies specific for hSMC1 and Rad21 were previously described [64]. Rabbit polyclonal antibody against the NIPBL protein was raised against a bacterially-expressed recombinant polypeptide corresponding to the C-terminal fragment of NIPBL isoform A (NP_597677.2) (amino acids 2429–2804) and antigen affinity-purified. CTCF antibody was from Millipore (07-729) and histone H3 from Abcam (ab1791).

3.5.2 ChIP-sequencing (ChIP-seq) and ChIP-PCR

ChIP was carried out as described previously [31]. Approximately 50 mg DNA was used per IP. Cells were crosslinked 10 mins with 1% formaldehyde, lysed, and sonicated using the Bioruptor from Diagenode to obtain ~200bp fragments using a 30 sec on/off cycle for 1 hr. Samples were diluted and pre-cleared for 1 hr with BSA and Protein A beads. Pre-cleared extracts were incubated with Rad21, Nipbl, and preimmune antibodies overnight. IP was performed with Protein A beads with subsequent washes. DNA was eluted off beads, reversed crosslinked for 8 hrs, and purified with the Qiagen PCR

Purification Kit. Samples were submitted to Ambry Genetics (Aliso Viejo, CA) for library preparation and sequencing using the Illumina protocol and the Illumina Genome Analyzer (GA) system. The total number of reads before alignment were: preimmune IgG, 7,428,656; Rad21 in control WT, 7,200,450; Rad21 in Nipbl+/-, 4,668,622; histone H3 in WT, 26,630,000; and histone H3 in Nipbl+/-, 24,952,439. Sequences were aligned to the mouse mm9 reference genome using Bowtie (with parameters–n2, -k20, —best, —strata, —chunkmbs 384) [65]. ChIP-seq data is being submitted to GEO. PCR primers used for manual ChIP confirmation are listed in Table 3.1. Primers corresponding to repeat sequences (major and minor satellite, rDNA, SINEB1 and B2 repeats) were from Matens et al. [66]. For manual ChIP-PCR analysis of selected genomic locations, ChIP signals were normalized with preimmune IgG and input DNA from each cell sample as previously described [31, 41, 67]. The experiments were repeated at least three times using MEF samples from different litters, which yielded consistent results. PCR reactions were done in duplicates or triplicates.

3.5.3 Peak Finding

Peaks were called using AREM (Aligning ChIP-seq Reads using Expectation Maximization) as previously described [44]. AREM incorporates sequences with one or many mappings to call peaks as opposed to using only uniquely mapping reads, allowing one to call peaks normally missed due to repetitive sequence. Since many peaks for Rad21 as well as CTCF can be found in repetitive sequence [44, 68], we used a mixture model to describe the data, assuming K + 1 clusters of sequences (K peaks and background). Maximum likelihood is used to estimate the locations of enrichment, with

the read alignment probabilities iteratively updated using EM. Final peaks are called for each window assuming a Poisson distribution, calculating a *p*-value for each sequence cluster. The false discovery rate for all peaks was determined relative to the pre-immune sample, with EM performed independently for the pre-immune sample as well. Full algorithm details are available, including a systematic comparison to other common peak callers such as SICER and MACS [44]. Overlap between peaks and genomic regions of interest were generated using Perl and Python scripts as well as pybedtools [69, 70]. Figures were generated using the R statistical package [71]. Visualization of sequence pileup utilized the UCSC Genome Browser [72, 73].

3.5.4 Motif Analysis

De Novo motif discovery was performed using Multiple Expectation maximization for Motif Elicitation (MEME) version 6.1 [74]. Input sequences were limited to 200 bp in length surrounding the summit of any given peak, and the number reduced to 1000 randomly sampled sequences from the set of all peak sequences. Motif searches for known motifs were performed by calculation of a log-odds ratio contrasting the position weight matrix with the background nucleotide frequency. Baseline values were determined from calculations across randomly selected regions of the genome. Randomly selected 200bp genomic regions were used to calculate a false discovery rate (FDR) at several position weight matrix (PWM) score thresholds. We chose the motif-calling score threshold corresponding to a 4.7% FDR. The *p*-values were derived for the number of matches above the z-score threshold relative to the background using a hypergeometric test.

3.5.5 Expression data analysis

Affymetrix MOE430A 2.0 array data for mouse embryonic fibroblasts (10 data sets for the wild type and nine for Nipbl+/- mutant MEFS) were previously published [17]. Expression data were filtered for probe sets with values below 300 and above 20,000, with the remainder used for downstream analysis. Differential expression and associated *p*-values were determined using Cyber-t, which uses a modified t-test statistic [75]. Probe sets were collapsed into genes by taking the median value across all probe sets representing a particular gene. Raw expression values for each gene are represented as a z-score, which denotes the number of standard deviations that value is away from the mean value across all genes. Gene ontology analysis was performed using PANTHER [76, 77] with a cutoff of $p < 0.05$.

3.5.6 KS test

Genes were sorted by their fold-change and any adjacent ChIP binding sites were identified. We performed a Kolmogorov-Smirnov (KS) test comparing the expression-sorted ChIP binding presence vs. a uniform distribution of binding sites, similar to Gene Set Enrichment Analysis [78]. If ChIP binding significantly correlates with the gene expression fold-change, the KS statistic, *d*, will also have significant, non-zero magnitude. To better visualize the KS test, we plotted the difference between the presence of cohesin binding at (expression-sorted) genes in Figure 3.5. The x axis of this Figure 3.is the (fold-change-based) gene rank, and the y axis is the KS statistic *d*, which behaves like a running enrichment score and is higher (lower) when binding sites co-occur more (less) often than expected if there were no correlation between ChIP binding

and expression fold-change. The KS test uses only the *d* with the highest magnitude, which is indicated in the plots by a vertical red line. To better visualize ChIP binding presence, we further plot an x-mirrored density of peak presence at the top of each plot; the gray "beanplot" [79] at the top of the plots are larger when many of the genes have adjacent ChIP binding sites.

3.5.7 siRNA depletion

Wild type MEFs were transfected using HiPerFect (Qiagen) following the manufacturer's protocol with 10mM siRNA. A mixture of 30µl HiPerFect, 3µl of 20µM siRNA, and 150µl DMEM was incubated for 10 mins and added to 2 x 10^6 cells in 4 ml DMEM. After 6 hrs, 2ml fresh DMEM with 10% FBS was added. Transfection was repeated the next day. Cells were harvested 48 hrs after the first transfection. SiRNAs against *Nipbl* (Nipbl-1: 5'-GTGGTCGTTACCGAAACCGAA-3'; Nipbl-2: 5'-AAGGCAGTACTTAGACTTTAA-3') and *Rad21* (5'-CTCGAGAATGGTAATTGTATA-3') were made by Qiagen. AllStars Negative Control siRNA was obtained from Qiagen.

3.5.8 RT-q-PCR

Total RNA was extracted using the Qiagen RNeasy Plus kit. First-strand cDNA synthesis was performed with SuperScript II (Invitrogen). Q-PCR was performed using the iCycler iQ Real-time PCR detection system (Bio-Rad) with iQ SYBR Green Supermix (Bio-Rad). Values were generated based on Ct and normalized to control gene RNH1. PCR primers specific for major satellite, minor satellite, rDNA, SINE B1 and

SINE B2 were previously described [66]. Other unique primers are listed in Table 3.1. The RT-qPCR analyses of the wild type and mutant cells were done with two biological replicates with consistent results. The gene expression changes after siRNA treatment were evaluated with two to three biological replicates with similar results.

3.6 References

1. DeScipio, C., et al., *Chromosome rearrangements in cornelia de Lange syndrome (CdLS): report of a der(3 t(3;12)(p25.3;p13.3) in two half sibs with features of CdLS and review of reported CdLS cases with chromosome rearrangements.* Am. J. Med. Genet., 2005. **137**: p. 276-82.
2. Liu, J. and I.D. Krantz, *Cornelia de Lange syndrome, cohesin, and beyond.* Clin. Genet., 2009. **76**: p. 303-14.
3. Krantz, I.D., et al., *Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B.* Nat. genet., 2004. **36**: p. 631-5.
4. Tonkin, E.T., et al., *NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome.* Nat. Genet., 2004. **36**: p. 636-41.
5. Ciosk, R., et al., *Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins.* Mol. Cell, 2000. **5**: p. 243-54.
6. Chien, R., et al., *Cohesin: a critical chromatin organizer in mammalian gene regulation.* Biochem. Cell Biol., 2011. **89**: p. 445-58.
7. Dorsett, D. and L. Ström, *The ancient and evolving roles of cohesin in gene expression and DNA repair.* Curr. Biol., 2012(22).
8. Nasmyth, K. and C.H. Haering, *Cohesin: its roles and mechanisms.* Annu. Rev. Genet., 2009. **43**: p. 525-8.
9. Musio, A., et al., *X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations.* Nat. Genet., 2006. **38**: p. 528-30.
10. Deardorff, M.A., et al., *Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation.* Am. J. Hum. Genet., 2007. **80**: p. 485-94.
11. Mannini, L., et al., *SMC1A codon 496 mutations affect the cellular response to genotoxic treatments.* Am. J. Med. Genet., 2012. **158A**: p. 224-8.
12. Deardorff, M.A., et al., *HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle.* Nature, 2012. **489**: p. 313-7.
13. Deardorff, M.A., et al., *RAD21 mutations cause a human cohesinopathy.* Am. J. Hum. Genet., 2012. **90**: p. 1014-27.
14. Castronovo, P., et al., *Somatic mosaicism in Cornelia de Lange syndrome: a further contributor to the wide clinical expressivity?* Clin. Genet., 2010. **78**: p. 560-4.
15. Dorsett, D. and I.D. Krantz, *On the molecular etiology of Cornelia de Lange syndrome.* Ann. N. Y. Acad. Sci., 2009. **1151**: p. 22-37.
16. Selicorni, A., et al., *Clinical score of 62 Italian patients with Cornelia de Lange syndrome and correlations with the presence and type of NIPBL mutation.* Clin. Genet., 2007. **72**: p. 98-108.
17. Kawauchi, S., et al., *Multiple organ system defects and transcriptional dysregulation in the nipbl mouse, a model of cornelia de lange syndrome.* PLoS Genet., 2009. **5**: p. e1000650.

18. Borck, G., et al., *Father-to-daughter transmission of Cornelia de Lange syndrome caused by a mutation in the 5' untranslated region of the NIPBL Gene.* Hum Mutat, 2006. **27**(8): p. 731-5.
19. Liu, J., et al., *Transcriptional dysregulation in NIPBL and cohesin mutant human cells.* PLoS Biol., 2009. **7**: p. e1000119.
20. Horsfield, J.A., C.G. Print, and M. Mönnich, *Diverse developmental disorders from the one ring: distinct molecular pathways underlie the cohesinopathies.* Front. Genet., 2012. **3**: p. 171.
21. Dorsett, D., *Cohesin: genomic insights into controlling gene transcription and development.* Curr. Opin. Genet. Dev., 2011. **21**: p. 199-206.
22. Kaur, M., et al., *Precocious sister chromatid separation (PSCS) in Cornelia de Lange syndrome.* Am. J. Med. Genet., 2005. **138**: p. 27-31.
23. Castronovo, P., et al., *Premature chromatid separation is not a useful diagnostic marker for Cornelia de Lange syndrome.* Chromosome Res, 2009. **17**(6): p. 763-71.
24. Vrouwe, M.G., et al., *Increased DNA damage sensitivity of Cornelia de Lange syndrome cells: evidence for impaired recombinational repair.* Hum. Mol. Genet., 2007. **16**: p. 1478-87.
25. Muto, A., et al., *Multifactorial origins of heart and gut defects in nipbl-deficient zebrafish, a model of Cornelia de Lange Syndrome.* PLoS Biol., 2011. **9**: p. e1001181.
26. Parelho, V., et al., *Cohesins functionally associate with CTCF on mammalian chromosome arms.* Cell, 2008. **132**: p. 422-33.
27. Wendt, K.S., et al., *Cohesin mediates transcriptional insulation by CCCTC-binding factor.* Nature, 2008. **451**: p. 796-801.
28. Rubio, E.D., et al., *CTCF physically links cohesin to chromatin.* Proc. Natl. Acad. Sci., 2008. **105**: p. 8309-14.
29. Stedman, W., et al., *Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators.* EMBO J., 2008. **27**: p. 654-66.
30. Zlatanova, J. and P. Caiafa, *CTCF and its protein partners: divide and rule?* J. Cell Sci., 2009. **122**: p. 1275-84.
31. Chien, R., et al., *Cohesin mediates chromatin interactions that regulate mammalian β-globin expression.* J. Biol. Chem., 2011. **286**: p. 17870-8.
32. Hadjur, S., et al., *Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus.* Nature, 2009. **460**: p. 410-3.
33. Mishiro, T., et al., *Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster.* EMBO J., 2009. **28**: p. 1234-45.
34. Nativio, R., et al., *Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus.* PLoS Genet., 2009. **5**: p. e1000739.
35. Phillips-Cremins, J.E., et al., *Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment.* Cell, 2013. **doi:pii: S0092-8674(13)00529-1. 10.1016/j.cell.2013.04.053.**
36. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**: p. 430-5.

37. Zuin, J., et al., *Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.* Proc. Natl. Acad. Sci., 2014. **111**: p. 996-1001.
38. Schmidt, D., et al., *A CTCF-independent role for cohesin in tissue-specific transcription.* Genome Res., 2010. **20**: p. 578-88.
39. Faure, A.J., et al., *Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules.* Genome Res., 2012. **22**: p. 2163-75.
40. Shimura, M., et al., *Epigenetic displacement of HP1 from heterochromatin by HIV-1 Vpr causes premature sister chromatid separation.* J. Cell Biol., 2011. **194**: p. 721-35.
41. Zeng, W., et al., *Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD).* PLoS Genet., 2009. **5**: p. e1000559.
42. Hakimi, M.A., et al., *A chromatin remodeling complex that loads cohesin onto human chromosomes.* Nature, 2002. **418**: p. 994-998.
43. Remeseiro, S., et al., *A unique role of cohesin-SA1 in gene regulation and development.* EMBO J., 2012. **31**: p. 2090-102.
44. Newkirk, D., et al., *AREM: aligning short reads from ChIP-sequencing by expectation maximization.* J. Comput. Biol., 2011. **18**: p. 495-505.
45. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat. Methods, 2008. **5**: p. 621-8.
46. Remeseiro, S., et al., *Reduction of Nipbl impairs cohesin loading locally and affects transcription but not cohesion-dependent functions in a mouse model of Cornelia de Lange Syndrome.* Biochim. Biophys. Acta., 2013. **1832**: p. 2097-102.
47. Dunn, K.L. and J.R. Davie, *The many roles of the transcriptional regulator CTCF.* Biochem. Cell Biol., 2003. **81**: p. 161-7.
48. Millau, J.F. and L. Gaudreau, *CTCF, cohesin, and histone variants: connecting the genome.* Biochem. Cell Biol., 2011. **89**: p. 505-13.
49. Misulovin, Z., et al., *Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome.* Chromosoma, 2008. **117**(1): p. 89-102.
50. Lengronne, A., et al., *Cohesin relocation from sites of chromosomal loading to places of convergent transcription.* Nature, 2004. **430**: p. 573-8.
51. Ocampo-Hafalla, M.T. and F. Uhlmann, *Cohesin loading and sliding.* J. Cell Sci., 2011. **124**: p. 685-91.
52. Xiao, T., J. Wallace, and G. Felsenfeld, *Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity.* Mol. Cell Biol., 2011. **31**: p. 2174-83.
53. Ball, A.R., Jr., Y.Y. Chen, and K. Yokomori, *Mechanisms of cohesin-mediated gene regulation and lessons learned from cohesinopathies.* BBA Gene Regul. Mech., 2014. **1839**: p. 191-202.
54. Fay, A., et al., *Cohesin selectively binds and regulates genes with paused RNA polymerase.* Curr. Biol., 2011. **21**: p. 1624-34.

55. Misulovin, Z., et al., *Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome.* Chromosoma, 2008. **117**: p. 89-102.
56. Schaaf, C.A., et al., *Genome-wide control of RNA polymerase II activity by cohesin.* PLoS Genet., 2013. **9**: p. e1003382.
57. Gomes, N.P. and J.M. Espinosa, *Gene-specific repression of the p53 target gene PUMA via intragenic CTCF-Cohesin binding.* Genes Dev., 2010. **24**(10): p. 1022-34.
58. Losada, A., et al., *Identification and characterization of SA/Scc3p subunits in the Xenopus and human cohesin complexes.* J. Cell Biol., 2000. **150**: p. 405-416.
59. Sumara, I., et al., *Characterization of vertebrate cohesin complexes and their regulation in prophase.* J. Cell Biol., 2000. **151**: p. 749-761.
60. Schaaf, C.A., et al., *Regulation of the Drosophila Enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins.* PLoS One, 2009. **4**: p. e6202.
61. Papai, G., P.A. Weil, and P. Schultz, *New insights into the function of transcription factor TFIID from recent structural studies.* Curr. Opin. Genet. Dev., 2011. **21**: p. 219-24.
62. Nolen, L.D., et al., *Regional chromatin decompaction in Cornelia de Lange syndrome associated with NIPBL disruption can be uncoupled from cohesin and CTCF.* Hum. Mol. Genet., 2013. **22**: p. 4180-93.
63. Zuin, J., et al., *A cohesin-independent role for NIPBL at promoters provides insights in CdLS.* PLoS Genet., 2014. **10**: p. e1004153.
64. Gregson, H.C., et al., *A potential role for human cohesin in mitotic spindle aster assembly.* J. Biol. Chem., 2001. **276**: p. 47575-47582.
65. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol., 2009. **10**: p. R25.
66. Martens, J.H., et al., *The profile of repeat-associated histone lysine methylation states in the mouse epigenome.* EMBO J., 2005. **24**: p. 800-12.
67. Zeng, W., et al., *Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs.* Hum. Mutat., 2014. **[Epub].**
68. Schmidt, D., et al., *Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.* Cell, 2012. **148**: p. 335-48.
69. Dale, R.K., B.S. Pedersen, and A.R. Quinlan, *Pybedtools: a flexible Python library for manipulating genomic datasets and annotations.* Bioinformatics, 2011. **27**: p. 3423-4.
70. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**: p. 841-2.
71. Dean, C.B. and J.D. Nielsen, *Generalized linear mixed models: a review and some extensions.* Lifetime Data Anal, 2007. **13**(4): p. 497-512.
72. Rhead, B., et al., *The UCSC Genome Browser database: update 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D613-9.
73. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res., 2002. **12**: p. 996-1006.

74. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc. Int. Conf. Intell. Syst. Mol. Biol., 1994. **2**: p. 28-36.
75. Long, A.D., et al., *Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12.* J. Biol. Chem., 2001. **276**: p. 19937-44.
76. Thomas, P.D., et al., *PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification.* Nuc. Acids Res., 2003. **31**: p. 334-41.
77. Thomas, P.D., et al., *Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.* Nuc. Acids Res., 2006. **34**: p. W645-50.
78. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc. Natl. Acad. Sci., 2005. **102**: p. 15545-50.
79. Kampstra, P., *Beanplot: A Boxplot Alternative for Visual Comparison of Distributions.* J. Statistical Software, 2008. **28**: p. http://www.jstatsoft.org/v28/c01.

Chapter 4

Cohesin-independent gene regulation by NIPBL in HeLa cells

4.1 Abstract

The cohesin complex is an evolutionarily conserved, essential protein complex with critical functions in sister chromatid cohesion, DNA repair, and gene regulation. In Cornelia de Lange Syndrome (CdLS), a developmental disorder affecting many different organ systems, patients fail to show significant sister chromatid cohesion defects. Instead, recent work in a mouse model for CdLS implicates widespread gene dysregulation as a probable cause for the disease. While mutations in cohesin subunits can result in mild forms of CdLS, a majority of cases are due to mutations in NIPBL, the cohesin's loading factor. In order to better understand the chromatin binding patterns of NIPBL in mammalian cells, with implications for CdLS, we performed chromatin immunoprecipitation with sequencing (ChIP-sequencing) of NIPBL. We found that a majority of NIPBL binding sites overlap with cohesin and CTCF. Moreover, there exists a subset of NIPBL binding sites that are free of cohesin and that are enriched at the promoter region of genes. A subset of the genes bound by NIPBL is regulated independently of cohesin, being upregulated upon depletion of NIPBL. Using microarray analysis, we found that 76 genes differentially expressed upon NIPBL depletion are bound by NIPBL at the promoter region. Finally, NIPBL binding sites in HeLa show significant enrichment for the YY1 and HCFC1 transcriptional regulators. Our data suggests that part of the phenotypic diversity present in CdLS patients is due to dysregulation of genes upon loss of NIPBL binding in a cohesin-independent manner.

4.2 Introduction

The cohesin complex has been studied heavily in recent years as researchers have discovered cohesin's importance in DNA repair, replication, and gene regulation (reviewed in [1]). One of the more recent findings, that of gene regulation, has important implications for disease in particular; there are, in fact, a surprising number of genes that cohesin may regulate across the genome (Chapter 3, [2-6]). In conjunction with gene regulation, cohesin has been shown to play an important role in establishing and maintaining long-range chromatin interactions, which were found to be important for IgH diversity [7], regulated gene expression at the IGF2-H19 imprinted locus [5], the β-globin locus [2], and stem cell maintenance in mouse embryonic stem cells (MESs) [3]. Since cohesin may target so many genes, pathways, and systems, proper localization to its binding sites is critical.

Cornelia de Lange Syndrome (CdLS) is a developmental disorder characterized by an array of phenotypes, including limb deformity, cranial-facial defects, heart defects, and neurological delay [8]. Interestingly, these phenotypes can have a wide range of severity in patients. CdLS was shown to be caused primarily by mutations in *NIPBL* [8], with fewer cases due to mutations in either *SMC1* [9, 10] or *SMC3* [10](cohesin), or *HDAC8* [11](which regulates cohesin re-loading after mitosis). Mutations in the RAD21 subunit of cohesin show CdLS –like phenotypes, with similar skeletal and craniofacial defects, but milder cognitive impairment [12]. Various mutations in patients with CdLS had been recorded and their subsequent impact was analyzed in a recent review [13]. These data indicated that there are phenotypic differences in the patients with *NIPBL* mutations in comparison to those with mutations in cohesin, with NIPBL mutations producing a wider variety and more severe phenotypes [13]. The larger array of

phenotypes in particular suggests that NIPBL may be important for the disease mechanism beyond its ability to load cohesin. It was also posited that there might be cohesin-independent functions for NIPBL [13].

Previous work in our lab (see Chapter 3) has better characterized the relationship between *Nipbl* haploinsufficiency and the widespread gene dysregulation in the CdLS mouse model seen by our collaborators [14]. We found that *Nipbl* mutations resulted in a decreased level of cohesin binding across the genome, with many sites occurring near genes that were differentially expressed in the mutant mice. While these data suggest one mechanism by which Nipbl can affect gene expression through cohesin, it does not answer the question whether or not Nipbl can affect gene expression independently of cohesin within these mice (or by extension, within patients with CdLS). To answer this question, we have performed NIPBL ChIP-sequencing in a human cervical cancer cell line HeLa to identify where NIPBL is localized across the genome. Furthermore, we generated global expression data from HeLa cells depleted of either NIPBL or RAD21 (cohesin) to identify genes that are impacted only NIPBL and cohesin levels in the cell. By correlating these data with one another, we have been able to identify a set of genes bound by NIPBL and whose expression has been altered by NIPBL depletion but remain unaffected by RAD21 depletion. Lastly, we also identified transcriptional regulators that may cooperate with NIPBL to regulate expression of target genes. Our data provide evidence that NIPBL is indeed able to regulate expression of genes independently of cohesin, suggesting that the phenotypic diversity of CdLS patients with *NIPBL* mutations may be partly due to the impairment of cohesin-independent gene regulatory function of NIPBL.

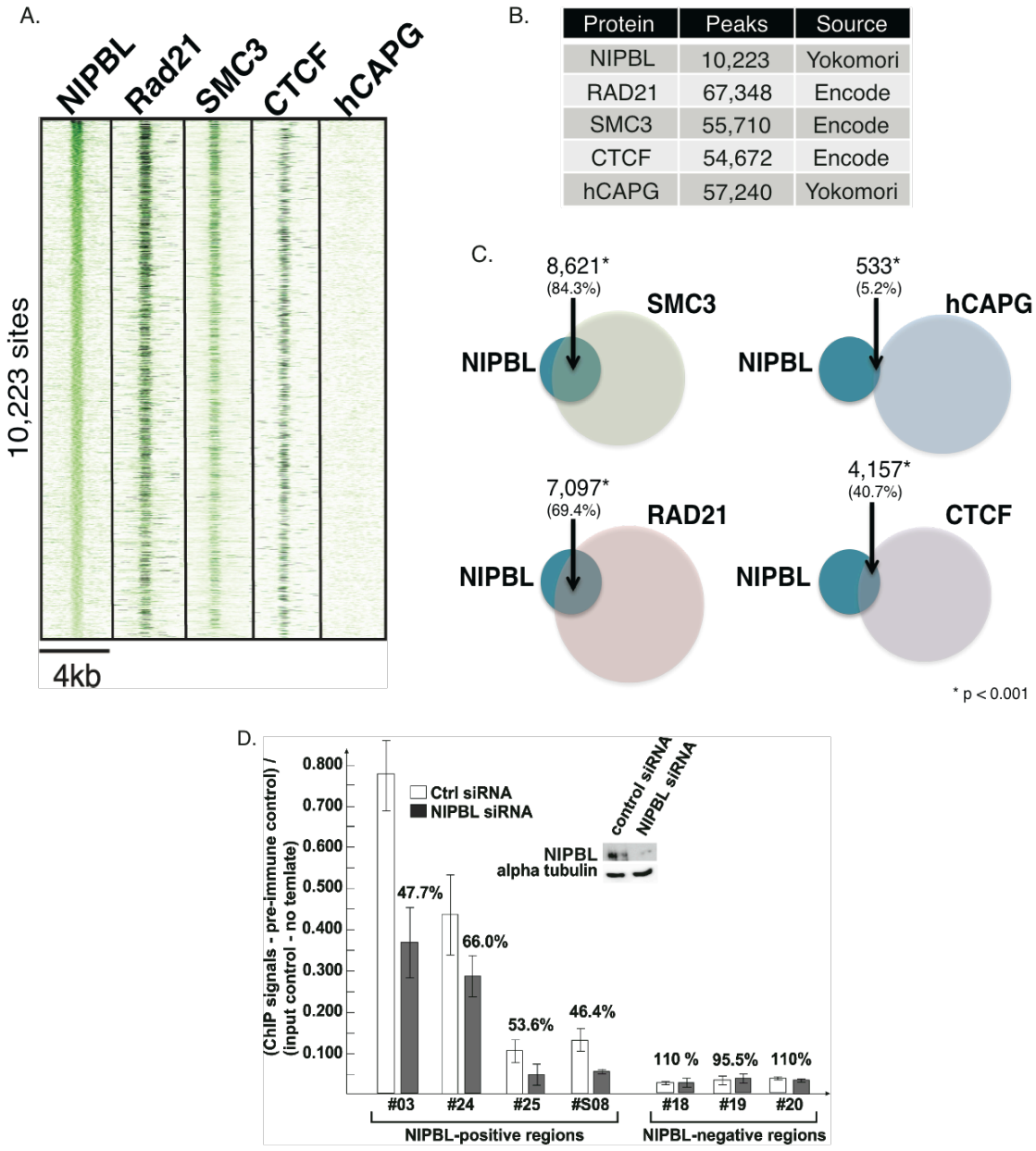
4.3 Results

4.3.1 NIPBL peaks overlap with both cohesin and CTCF

A former graduate student in our lab, Weihua Zeng, performed chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-sequencing) using an antibody directed against the C-terminal portion of NIPBL in asynchronous HeLa. In order to verify the specificity of our antibody, Weiwha Zeng depleted NIPBL in HeLa and performed ChIP at positive and negative sites based on the ChIP-sequencing results (Figure 4.1D), with positive sites showing a decrease in NIPBL binding upon NIPBL depletion. After read mapping and peak calling (see methods), a total of 10,223 peaks were found for NIPBL, with the average peak width for these peaks being 536 bp (Figure 4.3). Using ChIP-sequencing data from the ENCODE project, I mapped reads and called peaks for RAD21, SMC3 (subunits of the cohesin complex), and CTCF (see methods). We found 67,000 peaks for RAD21, 54,000 peaks for SMC3, and 58,000 peaks for CTCF (Figure 4.1). This is in contrast to what has been seen in mouse embryonic stem cells (ESCs) [3], where it was shown that few NIPBL binding sites overlap with CTCF. Recent studies in HB2 cells (human mammary epithelial cells) [15] also indicate that NIPBL does not overlap with either cohesin or CTCF in cell populations enriched for the G1 phase of the cell cycle.

Figure 4.1. NIPBL overlaps with cohesin and CTCF.

A. Heatmap of tag densities for all NIPBL peaks, sorted from highest density to lowest density. Tag densities for cohesin, CTCF, and hCAPG are also shown.
B. Table showing the number of peaks and sample source for each of NIPBL, cohesin, CTCF, and hCAPG.
C. Overlap between NIPBL and cohesin, CTCF and hCAPG. P-value is determined from a hypergeometric test after measuring overlap of randomly sampled "peaks."
D. ChIP confirmation of peaks in the NIPBL ChIP-sequencing data before and after depletion of NIPBL.



Protein	Peaks	Source
NIPBL	10,223	Yokomori
RAD21	67,348	Encode
SMC3	55,710	Encode
CTCF	54,672	Encode
hCAPG	57,240	Yokomori

Figure 4.1

4.3.2 A subset of NIPBL peaks is free of cohesin and CTCF

After examining the degree of overlap between NIPBL and other proteins, we clustered the NIPBL peaks based on whether or not they overlap with either cohesin or CTCF. When examining the set of NIPBL peaks that contain 5 or fewer tags in the NIPBL peak region for RAD21, we found them to be free of SMC3 and CTCF as well, suggesting these peaks are indeed cohesin-free (Figure 4.2). A total of 1,224 NIPBL peaks meet these criteria, about 10% of the total number of NIPBL binding sites. We also examined the number of cohesin-free NIPBL peaks in the promoter region, with about 10% of NIPBL peaks found at the promoter being cohesin-free.

4.3.3 Cohesin-free NIPBL sites are enriched at the promoter

Since we found that a subset of peaks is free of cohesin, we wanted to determine whether these peaks were enriched in specific genomic regions. Using gene locations from the UCSC Genome Browser, we found that 25-27% of NIPBL peaks were found in the promoter region (defined as 2,500 bp upstream and 500 bp downstream of the transcription start site) (Figure 4.3). The proportion of NIPBL peaks in the promoter was the same for NIPBL peaks with or without cohesin (Figure 4.3). When compared to random sampling of the genome, sampling of the same number of genomic regions as NIPBL peaks with or without NIPBL indicates that these peaks are only enriched at the promoter region (Figure 4.3). This apparent enrichment of NIPBL binding peaks in the promoter regions raise the possibility that NIPBL may affect the expression of the corresponding genes.

NIPBL peaks are not only found at the promoter region, but they sit close to the transcription start site (Figure 4.3). A higher proportion of NIPBL peaks appear to be present near the TSS of genes when bound with cohesin (Figure 4.3). NIPBL was previously shown to be present at enhancers and core promoters of genes bound by both cohesin and Mediator [3], and similar levels of promoter enrichment were seen in HB2 cells [15].

4.3.4 NIPBL affects expression of genes

Since NIPBL is enriched near the promoter region of many genes, we chose to examine the expression of cohesin-free, NIPBL-bound genes upon depletion of NIPBL and RAD21 (cohesin). Using siRNA, we depleted either RAD21 or NIPBL (Figure 4.4A) in HeLa. Two genes NSFP1 and FBXL16 showed upregulated expression upon depletion of NIPBL, but not RAD21 (Figure 4.4C), indicating that NIPBL, but not RAD21, regulates their expression. ChIP-PCR for NIPBL and RAD21 at the promoter of these genes confirmed the presence of NIPBL, but not cohesin (Figure 4.4B). Concomitant with the lack of cohesin, there is a lack of CTCF at these locations, but the presence of RNA Polymerase II (RNAPII) and H3K4me3, a histone modification associated with active genes.

Figure 4.2. Some NIPBL binding sites are free of cohesin.

A. Heatmap of tag densities at NIPBL peaks with or without RAD21 (see results). Tag densities of SMC3 and CTCF are also plotted.
B. Heatmap of tag densities at NIPBL peaks in a promoter region of a gene, with or without RAD21. Tag densities of SMC3 and CTCF in these peaks are also plotted.

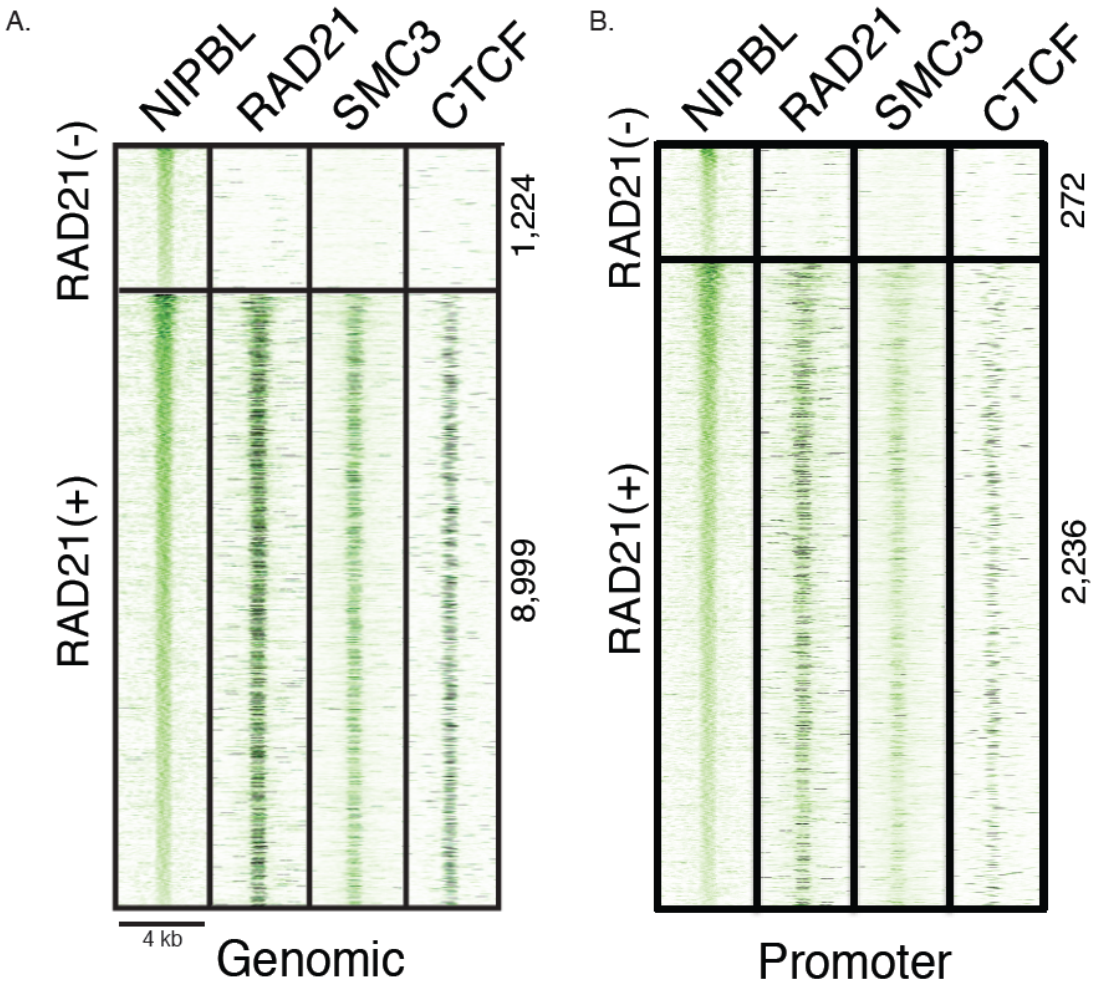

A.
NIPBL
RAD21
SMC3
CTCF
RAD21(-)
RAD21(+)
1,224
8,999
4 kb
Genomic
B.
NIPBL
RAD21
SMC3
CTCF
RAD21(-)
RAD21(+)
272
2,236
Promoter

Figure 4.2

Figure 4.3. NIPBL is enriched near the transcription start site.

A. Pie charts indicating the genomic distribution of NIPBL peaks; either all peaks, peaks with cohesin, or peaks without cohesin are shown.
B. Histogram of NIPBL peak widths.
C. Bar graph comparing the number of NIPBL binding sites in a given genomic region and randomly sampled peaks.
D. Density plot showing the distance of peaks to the TSS of the nearest gene.

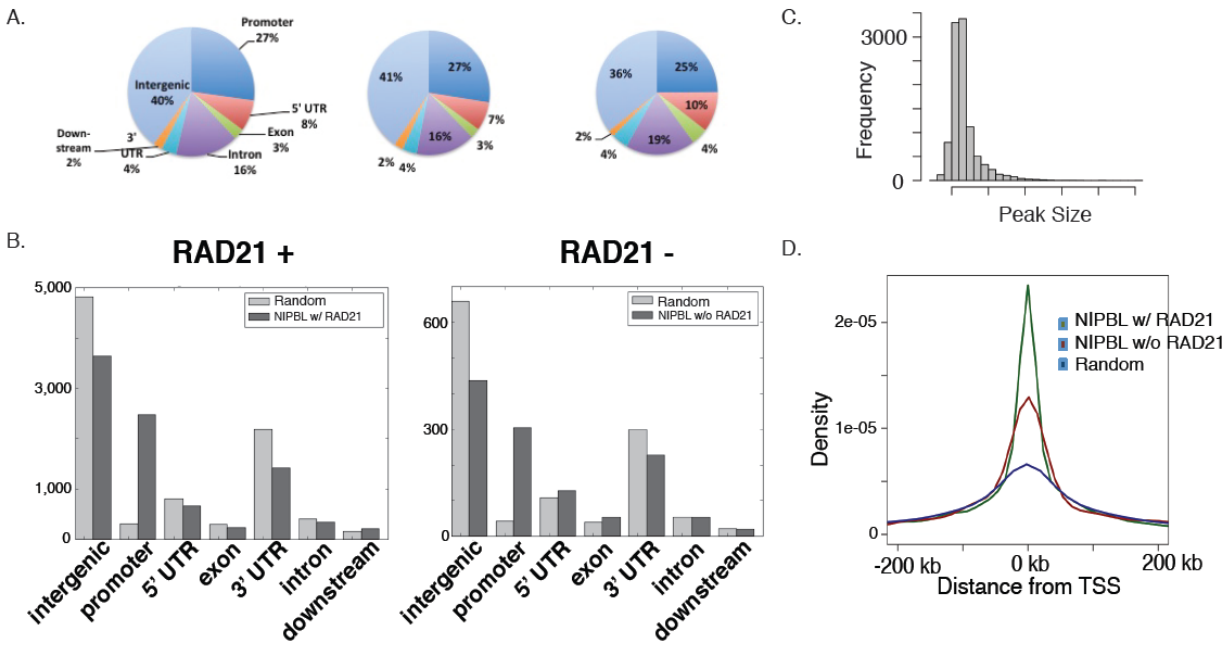

A.
Promoter 27%
Intergenic 40%
5' UTR 8%
Exon 3%
Intron 16%
3' UTR 4%
Down-stream 2%
41%
27%
7%
3%
16%
4%
2%
36%
25%
10%
4%
19%
4%
2%
B.
RAD21 +
Random
NIPBL w/ RAD21
5,000
3,000
1,000
0
intergenic
promoter
5' UTR
exon
3' UTR
intron
downstream
RAD21 -
Random
NIPBL w/o RAD21
600
300
0
intergenic
promoter
5' UTR
exon
3' UTR
intron
downstream
C.
3000
Frequency
0
Peak Size
D.
NIPBL w/ RAD21
NIPBL w/o RAD21
Random
2e-05
1e-05
0
Density
-200 kb
0 kb
200 kb
Distance from TSS

Figure 4.3

4.3.5: NIPBL is bound to the promoter of many genes with altered expression after NIPBL depletion

Using Affymetrix Human Gene ST arrays, we interrogated the global gene expression in HeLa after depletion of either NIPBL or RAD21. Since NIPBL is enriched at the promoter region, we looked at the expression of genes that are bound by NIPBL at the promoter. In order to compare differences between the effect of RAD21 and NIPBL depletion, we examined the expression pattern of the 273 genes bound by NIPBL at the promoter; while many genes have altered expression upon depletion of NIPBL, few genes are affected by RAD21 depletion (Figure 5A), further suggesting that NIPBL can regulate gene expression in a manner independent from cohesin. After filtering genes by absolute fold change greater than 1.2 and p-values of less than 0.05, 76 genes of the 273 NIPBL bound genes are differentially expressed (Tables 4.1).

4.3.5 NIPBL is enriched with YY1 motifs

Previous studies have shown that NIPBL is enriched with motifs of transcription factors such as such as NFYA, SP-1, IRF3, and PBX3 in HB2 cells [15]. We chose to undertake a similar analysis in two ways. First, we searched for motifs *de novo* using MEME [16], followed by scanning the NIPBL peak region with a collection of position weight matrices at a threshold z-score of 4.0. 400 bp around the peak summit for NIPBL were used to detect enriched motifs around NIPBL binding sites. In both cases we find an enrichment of the YY1 motif in NIPBL peaks (Figure 6A), but no enrichment of NFYA, SP-1, IRF3, or PBX3. These results suggest that NIPBL is able to interact with other transcriptional regulators in a cell-type dependent manner. The enrichment of YY1 motifs is present in NIPBL sites with and without cohesin (Figure 6A). A recent study profiled YY1 binding in HeLa, showing variants of the binding

sequence associated with either THAP11 or ZNF143 in conjunction with HCF-1 and YY1 [17]. Using the ChIP-sequencing data from this study, we found there to be significant overlap of NIPBL binding sites with YY1 binding in HeLa (Figure 6B). Combined, these data suggest that NIPBL may act in concert with YY1 to regulate gene expression.

Figure 4.4. NIPBL regulates gene expression independently of cohesin.

A. Western blot showing depletion efficiency of NIPBL and RAD21 in HeLa.
B. ChIP of NIPBL and cohesin at NSFP1 and FBXL16
C. RT-qPCR results showing expression of NSFP1 and FBXL16 before and after depletion of both NIPBL and RAD21.
D. UCSC genome browser image depicting the presence of NIPBL but not cohesin, CTCF at the promoter of both NSFP1 and FBXL16.

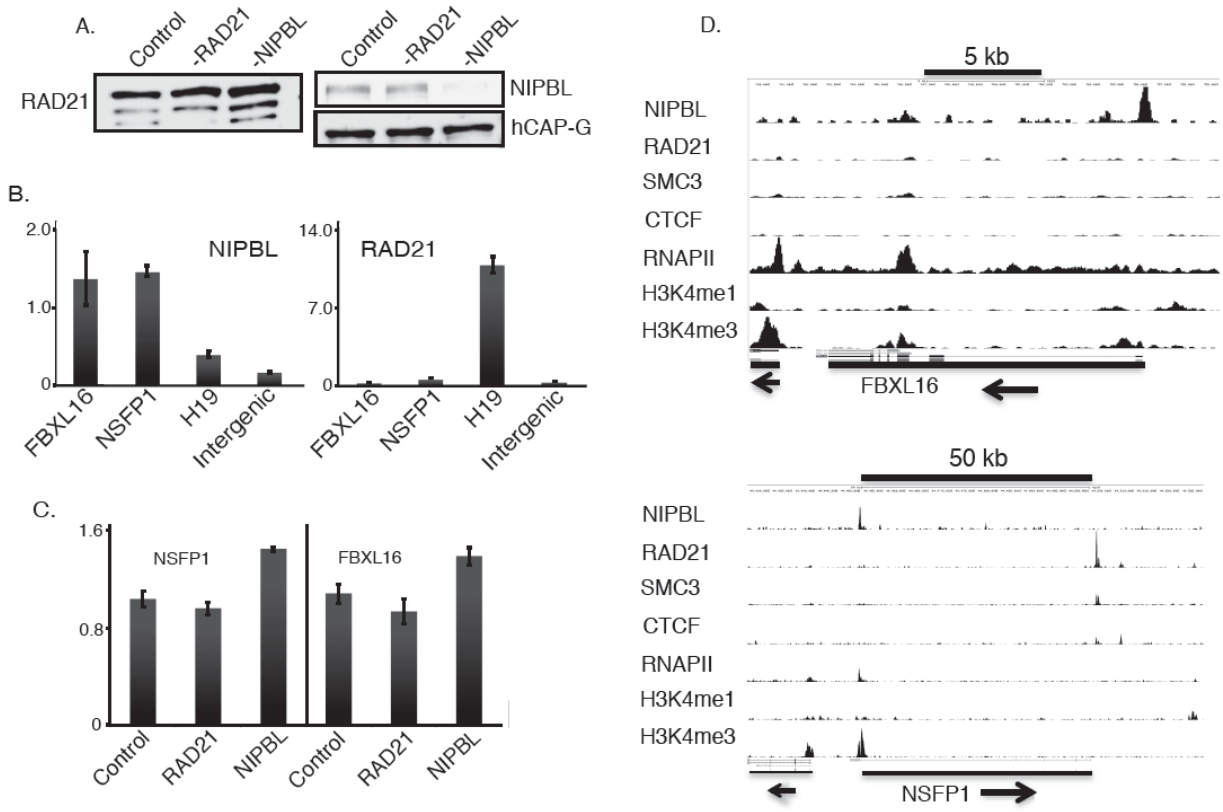

A.
Control
-RAD21
-NIPBL
Control
-RAD21
-NIPBL
RAD21
NIPBL
hCAP-G
B.
NIPBL
RAD21
2.0
1.0
0
14.0
7.0
0
FBXL16
NSFP1
H19
Intergenic
FBXL16
NSFP1
H19
Intergenic
C.
NSFP1
FBXL16
1.6
0.8
0
Control
RAD21
NIPBL
Control
RAD21
NIPBL
D.
5 kb
NIPBL
RAD21
SMC3
CTCF
RNAPII
H3K4me1
H3K4me3
FBXL16
50 kb
NIPBL
RAD21
SMC3
CTCF
RNAPII
H3K4me1
H3K4me3
NSFP1

Figure 4.4

Figure 4.5. NIPBL regulates many genes in HeLa.

A. Expression heatmap showing fold change of NIPBL-bound genes after depletion of NIPBL and RAD21.
B. Volcano plot showing the fold change and associated p-value of expression for genes bound by NIPBL at the promoter region. Points in cyan have an absolute fold change greater than 1.2 and p-value less than 0.05.

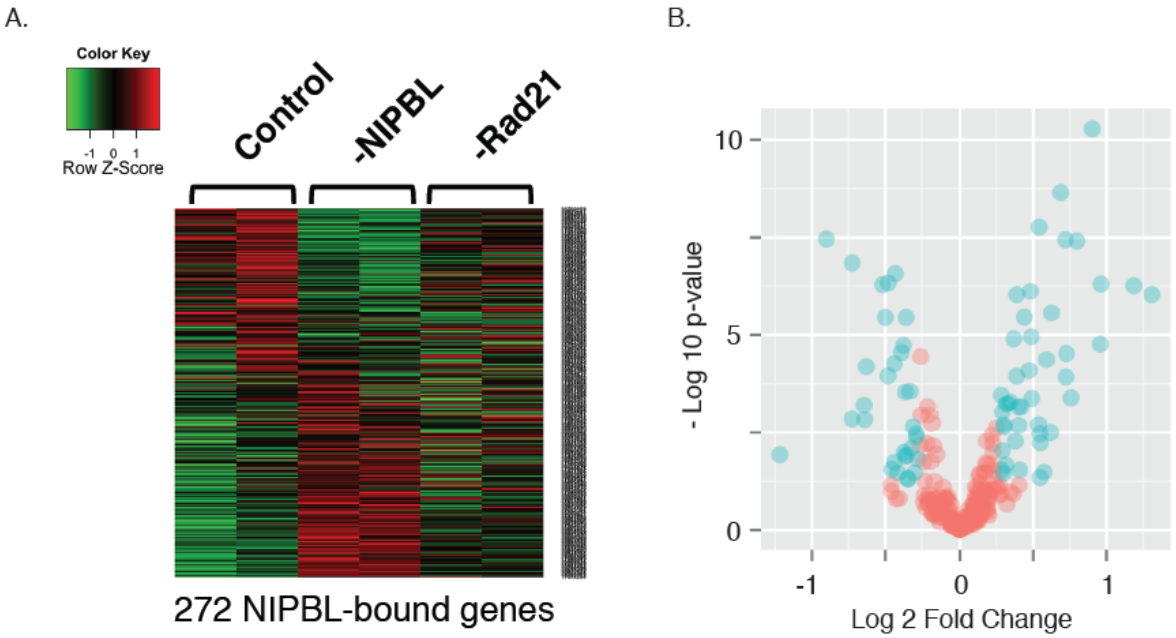

A.
Color Key
-1 0 1
Row Z-Score
Control
-NIPBL
-Rad21
272 NIPBL-bound genes
B.
10
5
0
- Log 10 p-value
-1
0
1
Log 2 Fold Change

Figure 4.5

Table 4.1. Differentially expressed genes bound by NIPBL

Gene	Probe Set ID	Control 1	Control 2	NIPBL 1	NIPBL 2	Avg Control	Avg NIPBL	Fold Change	P-value	Bonferroni	BH	Info
PIWIL1	16759159	25.03323	11.47137	7.875371	7.819393	18.2523	7.847382	-2.325909456	0.011775646	1	0.080203949	piwi-like RNA-mediated gene silencing 1
SGTB	16996771	732.1907	781.0696	452.0379	358.2201	756.63015	405.129	-1.867627719	3.48E-08	0.001305666	3.87E-06	small glutamine-rich tetratricopeptide repeat (TPR)-containing, beta
GNAO1	16819161	80.51002	89.00272	50.15492	52.31038	84.75637	51.23265	-1.654342885	0.001439187	1	0.016793494	guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O
PTPRF	16663621	574.6777	583.4938	328.3806	372.1807	579.08575	350.28065	-1.65320508	1.43E-07	0.005380564	1.14E-05	protein tyrosine phosphatase, receptor type, F
PIP5K1B	17085635	236.6646	306.4132	186.6286	161.0728	271.5389	173.8507	-1.56190858	0.000628409	1	0.008750846	phosphatidylinositol-4-phosphate 5-kinase, type I, beta
AMIGO1	16690612	145.4942	184.6344	105.4206	106.5997	165.0643	106.01015	-1.557061281	0.001495003	1	0.017235715	adhesion molecule with Ig-like domain 1
RASA3	16781315	339.1124	369.4402	243.5203	214.8066	354.2763	229.16345	-1.545954645	6.46E-05	1	0.001417062	RAS p21 protein activator 3
CXCR4	16903140	1846.895	2015.263	1432.086	1264.763	1931.079	1348.4245	-1.432100203	5.17E-07	0.019391697	3.13E-05	chemokine (C-X-C motif) receptor 4
MPV17L	16816178	849.8179	954.0938	648.9741	627.9117	901.95585	638.4429	-1.412743176	3.56E-06	0.133669111	0.000141899	MPV17 mitochondrial membrane protein-like
PDCD2L	16860737	724.7538	616.5132	507.7648	453.6078	670.6335	480.6863	-1.395158339	0.000111647	1	0.002202126	programmed cell death 2-like
LYPLA1	17077244	1858.702	1993.282	1373.244	1391.046	1925.992	1382.145	-1.393480424	0.000000468	0.01755815	2.92E-05	lysophospholipase I
CBFA2T3	16829271	96.93195	96.5461	69.2299	71.92571	96.739025	70.577805	-1.370672055	0.027815674	1	0.14680712	core-binding factor, runt domain, alpha subunit 2; translocated to, 3
CHCHD6	16945034	127.3949	140.6316	93.81182	104.02	134.01325	98.91591	-1.354819968	0.018670204	1	0.112037885	coiled-coil-helix-coiled-coil-helix domain containing 6
AGTPBP1	17095375	711.3773	813.4151	560.6307	565.6161	762.3962	563.1234	-1.353870573	5.50E-05	1	0.001262264	ATP/GTP binding protein 1
HDDC2	17023308	3080.64	3331.69	2502.098	2256.79	3206.165	2379.444	-1.347442932	2.68E-07	0.010047211	1.87E-05	HD domain containing 2
TRIM7	17004066	72.74541	86.11391	56.69193	62.66771	79.42966	59.67982	-1.330929953	0.034411119	1	0.169502776	tripartite motif containing 7
GTF3A	16773507	1379.861	1528.991	1178.426	1041.132	1454.426	1109.779	-1.310554624	2.81E-05	1	0.000735825	general transcription factor IIIA
B4GALT2	16663812	1178.298	1121.344	857.4346	915.6108	1149.821	886.5227	-1.297001194	1.84E-05	0.692052199	0.000526246	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 2
PROSER2	16702443	356.3891	308.1389	270.3387	243.2493	332.264	256.794	-1.29389316	0.010364061	1	0.07322559	proline and serine-rich protein 2
TRMT2A	16932420	279.4653	285.866	226.0925	212.7537	282.66565	219.4231	-1.288221933	0.013056742	1	0.08655658	tRNA methyltransferase 2 homolog A (S. cerevisiae)
LAMA5	16920939	680.239	791.7041	561.7859	583.4025	735.97155	572.5942	-1.285328336	0.00028966	1	0.004705022	laminin, alpha 5
HSBP1L1	16853325	2326.482	2507.791	1906.898	1869.212	2417.1365	1888.055	-1.280225682	3.57E-06	0.1338355	0.000141925	heat shock factor binding protein 1-like 1
PPP2R3B	17116706	265.5316	257.2674	218.2144	191.1674	261.3995	204.6909	-1.277045047	0.012197654	1	0.082371866	protein phosphatase 2, regulatory subunit B", beta
AGO4	16662430	109.7561	136.7666	97.85017	95.99487	123.26135	96.92252	-1.271751395	0.047970202	1	0.211211409	argonaute RISC catalytic component 4
LONRF1	17074815	113.2045	130.5638	97.18026	94.9792	121.88415	96.07973	-1.268572986	0.049070778	1	0.214110737	LON peptidase N-terminal domain and ring finger 1
STRN4	16873608	704.8325	712.2325	593.1768	533.5257	708.5325	563.35125	-1.257709999	0.000278611	1	0.004556276	striatin, calmodulin binding protein 4
BOP1	17082523	466.9799	495.5446	397.4373	372.8322	481.26225	385.13475	-1.24959446	0.007920908	1	0.060569388	block of proliferation 1
BICC1	16705089	581.3826	583.8812	456.1052	484.6763	582.6319	470.39075	-1.238612579	0.002256447	1	0.023665248	bicaudal C homolog 1 (Drosophila)
DDHD2	17068014	306.1788	378.5789	278.4009	277.7708	342.37885	278.08585	-1.231198387	0.034372674	1	0.169446894	DDHD domain containing 2
SNCG	16706896	782.868	729.7493	631.5897	607.1884	756.30865	619.38905	-1.221055894	0.004127906	1	0.037500361	synuclein, gamma (breast cancer-specific protein 1)
EFHD2	16659605	731.383	912.5016	684.6594	661.6499	821.9423	673.15465	-1.221030413	0.003449865	1	0.032773278	EF-hand domain family, member D2
TTC28	16933470	356.101	382.1027	311.8897	297.9011	369.10185	304.8954	-1.210585171	0.015437085	1	0.09784123	tetratricopeptide repeat domain 28
1	16733104	943.5918	951.0653	1114.73	1188.801	947.32855	1151.7655	1.215803641	0.000353312	1	0.005545807	checkpoint kinase 1

NFRKB	16746104	303.4661	360.9683	418.7153	390.3749	332.2172	404.5451	1.217712689	0.028705179	1	0.149961675	nuclear factor related to kappaB binding protein
QSOX2	17099847	819.8497	812.375	944.5494	1049.986	816.11235	997.2677	1.221973543	0.000948032	1	0.012101191	quiescin Q6 sulfhydryl oxidase 2
BTG3	16924305	369.8494	445.5558	504.9923	491.5349	407.7026	498.2636	1.222125147	0.009282824	1	0.067778347	BTG family, member 3
FKRP	16863515	185.7379	197.8354	225.1484	246.191	191.78665	235.6697	1.228811807	0.035146911	1	0.172087302	fukutin related protein
GNB4	16961806	641.9664	609.7824	747.3229	792.2534	625.8744	769.78815	1.229940304	0.001992438	1	0.021515921	guanine nucleotide binding protein (G protein), beta polypeptide 4
MTMR3	16928867	349.7929	365.3895	446.1975	435.4837	357.5912	440.8406	1.232806065	0.002298105	1	0.024001505	myotubularin related protein 3
FBXL17	16998700	240.1966	303.9684	350.6601	325.653	272.0825	338.15655	1.242845644	0.021707038	1	0.124383615	F-box and leucine-rich repeat protein 17
CLEC16A	16815855	723.8721	780.2604	900.1685	972.2794	752.06625	936.22395	1.244868986	0.000592237	1	0.008330931	C-type lectin domain family 16, member A
MECP2	17115453	516.0976	512.5769	652.2802	636.8668	514.33725	644.5735	1.253211779	0.000573665	1	0.008122729	methyl CpG binding protein 2 (Rett syndrome)
PAXIP1	17064724	487.0342	545.5275	645.4814	664.4818	516.28085	654.9816	1.26865368	0.000530663	1	0.00763923	PAX interacting (with transcription-activation domain) protein 1
CD276	16802854	1092.442	1084.423	1392.081	1416.346	1088.4325	1404.2135	1.29012456	1.27E-05	0.478072023	0.000387731	CD276 molecule
ZNRF3	16928699	288.4338	301.082	390.0138	376.0997	294.7579	383.05675	1.299563981	0.005309959	1	0.045355902	zinc and ring finger 3
UHRF1	16857258	711.004	784.791	975.7795	980.7085	747.8975	978.244	1.307992071	0.000113657	1	0.002237056	ubiquitin-like with PHD and ring finger domains 1
TNFRSF12A	16815310	2607.414	2455.11	3279.434	3346.656	2531.262	3313.045	1.308851079	0.000000931	0.034925459	5.03E-05	tumor necrosis factor receptor superfamily, member 12A
BHLHE41	16762470	388.5545	421.4107	558.0768	510.284	404.9826	534.1804	1.31902062	0.000703035	1	0.009552464	basic helix-loop-helix family, member e41
YJEFN3	16860062	356.2509	405.9655	565.9954	440.3185	381.1082	503.15695	1.32024698	0.002091092	1	0.022349666	YjeF N-terminal domain containing 3
CIC	16862670	82.84743	88.42471	112.1526	115.0577	85.63607	113.60515	1.326603965	0.028752305	1	0.150042783	capicua transcriptional repressor
RALB	16885118	492.5674	475.9787	709.4417	585.4435	484.27305	647.4426	1.336937085	0.000693768	1	0.009464262	v-ral simian leukemia viral oncogene homolog B (ras related; GTP binding protein)
NSFP1	16850337	1169.648	1138.911	1485.745	1644.898	1154.2795	1565.3215	1.35610266	3.50E-06	0.131389964	0.000140224	N-ethylmaleimide-sensitive factor pseudogene 1
RASSF5	16676619	441.5291	439.1997	633.3819	587.7014	440.3644	610.54165	1.386446429	8.20E-05	1	0.001709414	Ras association (RalGDS/AF-6) domain family member 5
NFIB	17092490	1376.057	1411.748	1848.37	2042.087	1393.9025	1945.2285	1.395526947	7.60E-07	0.028529994	4.32E-05	nuclear factor I/B
PKMYT1	16823229	413.8439	380.6186	562.5267	550.8839	397.23125	556.7053	1.401464009	1.12E-05	0.419913027	0.000350805	protein kinase, membrane associated tyrosine/threonine 1
CCNF	16815090	249.1364	279.0182	377.0955	364.6722	264.0773	370.88385	1.40445184	0.000425961	1	0.006412494	cyclin F
SOCS3	16849400	129.7057	160.2533	211.9469	208.7209	144.9795	210.3339	1.450783732	0.001991811	1	0.021515343	suppressor of cytokine signaling 3
NSF	16835087	1160.318	1280.248	1747.347	1801.047	1220.283	1774.197	1.453922574	1.74E-08	0.000652185	2.27E-06	N-ethylmaleimide-sensitive factor
NT5M	16831598	164.5242	239.6902	337.4569	252.3383	202.1072	294.8976	1.459114767	0.003394264	1	0.032400968	5',3'-nucleotidase, mitochondrial
PTCH1	17068720	17.64865	26.33855	27.09421	37.12899	21.9936	32.1116	1.460042922	0.045457141	1	0.203620852	patched 1
IQCJ	16947494	70.70164	70.39954	105.2319	101.042	70.55059	103.13695	1.461886428	0.00584529	1	0.048784438	IQ motif containing J
GIMAP6	17064269	24.80815	30.42369	47.50586	34.72089	27.61592	41.113375	1.488756304	0.032983083	1	0.164808528	GTPase, IMAP family member 6
INSL4	17083352	296.9586	353.0164	525.3778	453.7767	324.9875	489.57725	1.506449479	4.29E-05	1	0.001025252	insulin-like 4 (placenta)
MMRN2	16716213	55.18764	43.76784	78.20341	73.57629	49.47774	75.88985	1.533818036	0.003197054	1	0.030879496	multimerin 2
DCK	16967614	331.8959	355.9004	532.0863	528.1923	343.89815	530.1393	1.541559034	2.72E-06	0.102111252	0.000114218	deoxycytidine kinase
STEAP4	17059567	1700.143	2011.991	2942.984	3031.379	1856.067	2987.1815	1.609414693	2.25E-09	8.46E-05	4.65E-07	STEAP family member 4
LRP8	16687352	821.5807	935.2148	1378.033	1515.61	878.39775	1446.8215	1.647114306	3.64E-08	0.001363762	4.02E-06	low density lipoprotein receptor-related protein 8, apolipoprotein e receptor
PLCL2	16938182	129.1956	161.127	248.2515	231.2365	145.1613	239.744	1.651569668	0.000119732	1	0.002326131	phospholipase C-like 2
CADM1	16744616	145.9423	157.1414	237.3808	264.185	151.54185	250.7829	1.654875534	0.00002992	1	0.000769025	cell adhesion molecule 1
HHAT	16677082	46.84926	65.95985	98.07938	92.6376	56.404555	95.35849	1.690616831	0.000409921	1	0.006246213	hedgehog acyltransferase
NLGN2	16830432	388.7027	442.5772	750.8823	694.7847	415.63995	722.8335	1.739085716	3.91E-08	0.001465957	4.16E-06	neuroligin 2
MAPK4	16852322	836.7049	906.1202	1574.012	1682.425	871.41255	1628.2185	1.868481812	5.15E-11	0.000001931	3.11E-08	mitogen-activated protein kinase 4
TUBB3	16822161	73.71715	95.5063	167.0695	161.6743	84.611725	164.3719	1.94266102	1.71E-05	0.641759577	0.000493661	tubulin, beta 3 class III
EDN1	17004903	91.37114	100.9915	195.9044	179.0076	96.18132	187.456	1.94898552	4.99E-07	0.018715013	3.06E-05	endothelin 1
ATOH8	16882352	59.28231	63.40969	137.1734	141.9395	61.346	139.55645	2.274907084	0.000000547	0.020519189	3.28E-05	atonal homolog 8 (Drosophila)
SBSPON	17078254	62.87589	75.73029	182.9942	160.5635	69.30309	171.77885	2.478660764	9.35E-07	0.035085324	5.05E-05	somatomedin B and thrombospondin, type 1 domain containing

Figure 4.6. NIPBL peaks overlap with YY1, HCF-1.

A. MEME motifs for YY1 and HCF-1 found near NIPBL summits with and without cohesin binding.
B. All NIPBL peaks clustered based on the presence of cohesin, CTCF, YY1, and HCF-1.

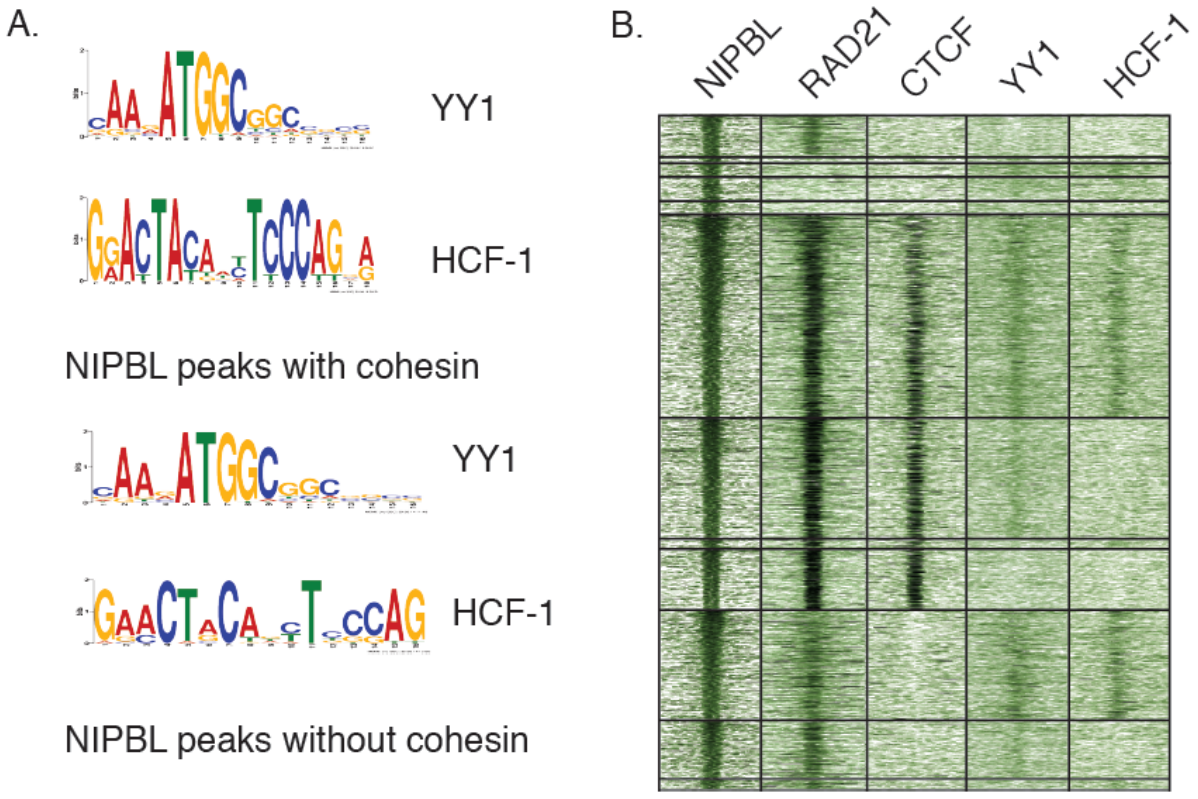

A.
YY1
HCF-1
NIPBL peaks with cohesin
YY1
HCF-1
NIPBL peaks without cohesin
B.
NIPBL
RAD21
CTCF
YY1
HCF-1

Figure 4.6

4.4 Discussion

4.4.1 Characterizing NIPBL binding using ChIP-seq

In the past, characterizing NIPBL binding sites has been difficult. Previous groups have identified NIPBL binding sites, but there is disagreement as to how well the antibodies used were able to capture all potential binding sites for NIPBL across the genome [3, 15]. Our study adds to the available data and helps to further clarify the nature of NIPBL binding using our in-house affinity-purified antibody, which was previously used successfully for ChIP-PCR analyses in human and mouse cells [2, 18]. Specific findings, such as the significant overlap between CTCF and NIPBL, and the overlap of NIPBL with YY1 but not NFYA/B, serve to augment our understanding and may not be in direct conflict to previous studies; instead, our data ought to be considered in addition to work from previous studies. This would suggest that usage of multiple antibodies or a tagged form of NIPBL would be necessary to identify the complete range of NIPBL binding sites and all potential target genes across the genome.

4.4.2 NIPBL can bind to chromatin in a cohesin-independent manner

Relatively little has been known about NIPBL's potential for functions apart from cohesin loading. However, phenotypic differences in patients with CdLS having mutations in NIPBL as opposed to mutations in the cohesin subunits has suggested that NIPBL might regulate gene expression apart from cohesin [13]. In brief, patients with mutations in NIPBL show a wider range and more severe set of phenotypes, especially neurological defects and limb deformity compared to those with mutations in SMC1, SMC3 or Rad21 [13]. While our previous work in Nipbl +/- mice (see Chapter 3) has suggested that a wide range of genes were

affected by Nipbl haploinsufficiency, many are not considered cohesin target genes. While some of these genes are likely affected indirectly by the alteration of upstream cohesin target genes, others suggested that NIPBL may play a direct role in their expression changes (ref). Our data, along with data from the Wendt laboratory [15], supports this hypothesis—that NIPBL can regulate a variety of genes in concert with other sequence-specific transcription factors. Further investigation is necessary to identify how many genes are directly targeted by NIPBL in order to better understand the impact of NIPBL haploinsufficiency in CdLS.

NIPBL contains a variety of protein binding domains, such as the HEAT repeats, an HP1 binding domain, and an interaction domain for Mau2 [19]. NIPBL's recruitment to many genes may be cell-type specific and depends on interactions with a variety of transcription factors. It has been previously established that NIPBL interacts with Mediator, to recruit RNA Polymerase II to the promoter of many genes, although those studies were considered in a cohesin-dependent context [3, 20]. More research will be required to understand how NIPBL interacts with different transcription factors to regulate gene expression, and to elucidate the function of cohesin-independent NIPBL in non-promoter regions.

4.4.3 Can NIPBL and YY1/HCF-1 collectively regulate gene expression?

In this study, we found that there is significant overlap between NIPBL and YY1 and HCF-1 binding sites in HeLa cells. Many of these NIPBL binding sites also contain CTCF, which has been known to interact directly with YY1 in the context of X inactivation [21]. Cohesin has also been shown to interact with YY1 based on mass spec data, with both working in concert with other factors to regulate IgH rearrangement [22]. It is interesting then to find that

YY1 can co-localize with NIPBL independently from either cohesin or CTCF. While YY1's interactions with cohesin (and condensin) at the IgH locus have been important for establishing chromatin interactions necessary for rearrangement ([22], and reviewed in [23]), NIPBL's co-localization with YY1 and HCF-1 may be important for regulation of gene expression through interactions at promoter regions or through regulation of miRNA expression. In particular, YY1 may be necessary for recruitment of NIPBL to specific genomic regions in the absence of CTCF or other sequence-specific transcription factors. Further study will be necessary to identify how NIPBL and YY1/HCF-1 interact at specific loci, and whether their co-localization is required for the proper expression of their target genes.

4.5 Methods

Cell Culture

HeLa were grown at 37° C and 5% CO2 in DMEM (Life Technologies). DMEM was supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin at 50 units/ml.

Antibodies

Antibodies directed against NIPBL, RAD21, hCAPG, and preimmune were previously described [2, 24].

ChIP-sequencing from other groups

The GEO accession numbers for ChIP-sequencing data used were:

Protein	GEO Accession Number
RAD21, SMC3, CTCF, RNAPII	GSE31477
YY1, HCF-1, THAP11, ZNF143	GSE31417

Chromatin Immunoprecipitation

Chromatin Immunoprecipitation was performed as previously described [2]. For each ChIP experiment, about 50 mg of DNA was used. Cells were crosslinked for 10 minutes using 1% formaldehyde diluted in cell culture media, lysed, and sonicated using a Bioruptor sonicator (Diagenode) to a fragment size around 200 bps. Samples were pre-cleared for 1 hr using BSA and Protein A sepharose beads (GE Healthcare). Pre-cleared extracts were incubated with NIPBL and Preimmune antibodies overnight. IP was performed with Protein A beads and then

washed. DNA was eluted off of the beads, reverse-crosslinked overnight, and then purified using a PCR purification kit (Qiagen). Samples were submitted to Ambry Genetics (Aliso Viejo, CA) for library preparation and sequencing using the Illumina protocol and the Illumina Genome Analyzer (GA) system.

siRNA depletion

HeLa were depleted for 24 (RAD21) or 48 (NIPBL, Control) hours with one allotment of siRNA (RAD21, day 2), or two allotments of siRNA (NIPBL, Control, day 1 and day 2). Cells were transfected using 40 µl Hyperfect (Qiagen) and 2 µl of 20 µM siRNA. AllStars control siRNA (Qiagen) was used for the control. Media containing the transfection reagent and siRNA was left on the cells until being split on day 2 of transfection to improve depletion. Cells were harvested on Day 4 after initial depletion.

Microarray

Affymetrix Human Gene ST 2.0 microarrays (Affymetrix) were used to interrogate the global gene expression in control, NIPBL depleted, and RAD21 depleted HeLa. Two biological replicates were used for each experiment, for a total of 6 samples/arrays. Samples were processed according to manufacturer instructions by the University of California, Irvine Genomics and High-Throughput Core Facility. Data were normalized using Plier, and differential gene expression analysis performed using cyber-t [25].

Sequence processing

Pre-processing of the sequenced reads was performed to remove low quality sequence and remove adapter sequence. In order to remove adapter sequence, the adapter sequence used for each library was removed from each read using CutAdapt (Python script, https://code.google.com/p/cutadapt/). After adapter removal, sequences are trimmed for low quality sequence (Phred score <= 20) using FastX. If trimmed sequences are less than 20 bp in length, they are removed entirely. Sequences were mapped back to the human genome draft 19 (hg19) from the UCSC genome browser [26]. Mapping was performed using Bowtie version 1.0 [27], with parameters –n 2 –m 3.

Peak calling

Peak calling was performed using AREM [28] using settings -no-EM, accepting only reads mapping to fewer than 3 locations, and –pval X, where X is adjusted to produce a FDR lower than 5%. For histone modifications, SICER [29] was used to call peaks with width 300, gap size of 600, and FDR less than 0.1%. All other settings were as recommended in the user manual.

Data comparison

All other data processing occurred through the R statistical package [30] and Perl and Python scripts coded by the author. Peak overlap was determined using pybedtools [31].

4.6 References

1. Chien, R., et al., *Cohesin: a critical chromatin organizer in mammalian gene regulation.* Biochem Cell Biol, 2011. **89**(5): p. 445-58.
2. Chien, R., et al., *Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression.* J Biol Chem, 2011. **286**(20): p. 17870-8.
3. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.
4. Liu, J., et al., *Transcriptional dysregulation in NIPBL and cohesin mutant human cells.* PLoS Biol, 2009. **7**(5): p. e1000119.
5. Nativio, R., et al., *Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus.* PLoS Genet, 2009. **5**(11): p. e1000739.
6. Remeseiro, S., et al., *A unique role of cohesin-SA1 in gene regulation and development.* EMBO J, 2012. **31**(9): p. 2090-102.
7. Thomas-Claudepierre, A.S., et al., *The cohesin complex regulates immunoglobulin class switch recombination.* J Exp Med, 2013. **210**(12): p. 2495-502.
8. Krantz, I.D., et al., *Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B.* Nat Genet, 2004. **36**(6): p. 631-5.
9. Musio, A., et al., *X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations.* Nat Genet, 2006. **38**(5): p. 528-30.
10. Deardorff, M.A., et al., *Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation.* Am J Hum Genet, 2007. **80**(3): p. 485-94.
11. Deardorff, M.A., et al., *HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle.* Nature, 2012. **489**(7415): p. 313-7.
12. Deardorff, M.A., et al., *RAD21 mutations cause a human cohesinopathy.* Am J Hum Genet, 2012. **90**(6): p. 1014-27.
13. Mannini, L., et al., *Mutation spectrum and genotype-phenotype correlation in Cornelia de Lange syndrome.* Hum Mutat, 2013. **34**(12): p. 1589-96.
14. Kawauchi, S., et al., *Multiple organ system defects and transcriptional dysregulation in the Nipbl(+/-) mouse, a model of Cornelia de Lange Syndrome.* PLoS Genet, 2009. **5**(9): p. e1000650.
15. Zuin, J., et al., *A cohesin-independent role for NIPBL at promoters provides insights in CdLS.* PLoS Genet, 2014. **10**(2): p. e1004153.
16. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
17. Michaud, J., et al., *HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy.* Genome Res, 2013. **23**(6): p. 907-16.
18. Zeng, W., et al., *Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD).* PLoS Genet, 2009. **5**(7): p. e1000559.
19. Jahnke, P., et al., *The Cohesin loading factor NIPBL recruits histone deacetylases to mediate local chromatin modifications.* Nucleic Acids Res, 2008. **36**(20): p. 6450-8.

20. Muto, A., et al., *Nipbl and mediator cooperatively regulate gene expression to control limb development.* PLoS Genet, 2014. **10**(9): p. e1004671.
21. Donohoe, M.E., et al., *Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch.* Mol Cell, 2007. **25**(1): p. 43-56.
22. Pan, X., et al., *YY1 controls Igkappa repertoire and B-cell development, and localizes with condensin on the Igkappa locus.* EMBO J, 2013. **32**(8): p. 1168-82.
23. Atchison, M.L., *Function of YY1 in Long-Distance DNA Interactions.* Front Immunol, 2014. **5**: p. 45.
24. Heale, J.T., et al., *Condensin I interacts with the PARP-1-XRCC1 complex and functions in DNA single-strand break repair.* Mol Cell, 2006. **21**(6): p. 837-48.
25. Kayala, M.A. and P. Baldi, *Cyber-T web server: differential analysis of high-throughput data.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W553-9.
26. Karolchik, D., et al., *The UCSC Genome Browser database: 2014 update.* Nucleic Acids Res, 2014. **42**(Database issue): p. D764-70.
27. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
28. Newkirk, D., et al., *AREM: aligning short reads from ChIP-sequencing by expectation maximization.* J Comput Biol, 2011. **18**(11): p. 1495-505.
29. Xu, S., et al., *Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells.* Methods Mol Biol, 2014. **1150**: p. 97-111.
30. Team, R.C., *R: A language and environment for statistical computing*. 2013, Vienna, Austria: R Foundation for Statistical Computing.
31. Dale, R.K., B.S. Pedersen, and A.R. Quinlan, *Pybedtools: a flexible Python library for manipulating genomic datasets and annotations.* Bioinformatics, 2011. **27**(24): p. 3423-4.

Chapter 5

Epigenomic characterization of Facioscapulohumeral Muscular Dystrophy:

A resource for future research

5.1 Abstract

Facioscapulohumeral muscular dystrophy (FSHD) is a common type of muscular dystrophy characterized by progressive atrophy of muscle in the upper body. Recent studies by several groups have identified disease-specific epigenetic differences to heterochromatin at chromosome 4q35 near the D4Z4 repeat array, a region shown to be critical for both forms of the disease. Our group showed a specific loss of H3K9me3, HP1γ, and cohesin at D4Z4, while others have shown a hypomethylation of DNA near D4Z4. SMCHD1, a protein known for its role in maintaining DNA methylation on the inactivated X chromosome, is bound to D4Z4 and it's binding reduced upon loss of H3K9me3, affecting the degree of DNA methylation. These changes, in concert with altered expression of many genes including that of DUX4—whose expression tends to be upregulated in FSHD—suggest that epigenetics may underlie the disease phenotype.

While much is known about the changes taking place at D4Z4, little is known about the genome-wide changes taking place in FSHD. Many genes and miRNAs appear to be differentially expressed in FSHD, studies show some disagreement on which genes/miRNAs might be differentially expressed. Moreover, how those gene expression differences are connected to the changes at D4Z4 and elsewhere in the genome remains unclear. To address these questions, we have undertaken a large high-throughput sequencing approach to characterize the global differences in gene expression, heterochromatin, and cohesin binding using myoblasts derived from normal and affected individuals. This study will allow us to more directly identify the correlation between the

epigenetic changes and gene expression changes, and to potentially identify a disease-specific signature for FSHD.

5.2 Introduction

FacioScapuloHumeral Muscular Dystrophy (FSHD) is one of the most common forms of muscular dystrophy in the United States. It is characterized by a progressive atrophy of the facial and shoulder muscles, and in some cases the trunk and foot musculature [1]. There are two forms of the disease: 4q-linked (FSHD1) and phenotypic (FSHD2). The most common form of the disease, FSHD1 (95% of cases, [1]), is characterized by a mono-allelic contraction of the D4Z4 repeat array on chromosome 4q. Normal individuals have 11-100 tandem repeats [2], while individuals with the disease have 10 or fewer [3]. Contained within each repeat is the *DUX4* retrogene, whose expression has been correlated with the disease [4].

The less common form of the disease, FSHD2 (< 5%) has been connected to mutations in the gene *SMCHD1* in some individuals [5], with other cases of FSHD2 having an unknown cause. Individuals with mutations in *SMCHD1* still have similar phenotypes as those patients with the monoallelic contraction on chromosome 4q however.

Previous work from our lab has characterized the chromatin at D4Z4 [6]. In patients with either FSHD1 or FSHD2, there is a specific loss of H3K9me3, cohesin, and HP1γ at D4Z4 on chromosome 4q (and the homologous regions on chromosome 10q). Loss of H3K9me3 at D4Z4 results in the upregulation of DUX4 in KD3 myoblasts [7], and also affects the recruitment of SMCHD1 to D4Z4 [7]. However, it is unclear

whether or not other epigenetic changes occur globally, or whether these differences in H3K9me3 are unique to D4Z4.

In order to investigate FSHD1 and FSHD2, we have performed Chromatin Immunoprecipitation coupled with high-throughput sequencing (ChIP-sequencing) to examine the genomic localization of chromatin marks and protein complexes, and RNA-sequencing (sequencing of mRNA transcripts) to identify genes that have altered expression in either form of the disease. In order to find miRNAs that are differentially expressed in FSHD, we have used Nanostring (company info) to study expression of 800 different miRNAs in primary myoblasts.

As a part of this study, we have performed ChIP-sequencing for: cohesin, heterochromatin protein 1 gamma (HP1γ), histone H3 lysine 9 tri-methylation (H3K9me3), histone H3 lysine 27 tri-methylation (H3K27me3), and RNA Polymerase II (RNAPII). We have two primary myoblast samples derived from unaffected individuals, and 3 samples each from individuals with FSHD1 and FSHD2 (Table 1).

The results that are presented here are for the benefit of those researching FSHD. They encompass the most current data obtained from a more extensive project that is ongoing. While covering a wide range of topics, these data help provide a more coherent picture regarding the epigenetic landscape in FSHD and how it changes in comparison to normal individuals. Moreover, the analysis of expression data will aid in determining how the epigenetic changes taking place at D4Z4 and elsewhere in the genome affect the

expression of genes in FSHD, and further characterize the disease etiology. Many advances in our understanding of the disease have occurred in the last few years, and our lab hopes that this project will further refine and extend these.

5.3 Results

5.3.1 H3K9me3 patterns are altered between normal and FSHD

Our lab previously identified the specific loss of H3K9me3 at D4Z4, and the concomitant loss of HP1γ and cohesin in the same region just upstream of the DUX4 transcript [6]. This loss, combined with the DNA hypomethylation present at D4Z4 in FSHD [8], suggests that FSHD is an "epigenetic abnormality disease [6]." Since most studies of FSHD have focused on the marks at or near D4Z4, little is known about the epigenetic differences that are present genome wide in FSHD. To answer this question, Michelle Chen in our lab performed ChIP-Sequencing of H3K9me3 and H3K27me3 in myoblasts from two normal individuals, myoblasts from three patients with FSHD1, and myoblasts from three patients with FSHD2.

After alignment and peak calling using SICER [9], we found 43,000 – 63,000 peaks per sample for H3K9me3 (Table 5.1). Since H3K9me3 is lost at D4Z4 in FSHD1 and FSHD2, we first looked at those peaks that were not present (no called peak) in FSHD. We overlapped the peaks in the both forms of FSHD and normal myoblasts to identify the 546 peaks that only occur in normal myoblasts. To better understand how these peaks might relate to gene expression, we used the GREAT analysis tool [10] with a threshold of 1 Mb to identify nearby genes to these regions with no H3K9me3 peaks called in FSHD, and to calculate the enrichment of different categories of genes. We found that the 548 peaks present only in normal myoblasts are enriched for gene ontology (GO) terms including immune response, FGF receptor expression, and skeletal muscle

development. However, when we examined the expression of the genes in each of these categories, they were not differentially expressed. It is possible that the loss of H3K9me3 at these locations has an impact on differentiation of the myoblasts later on, while not having an impact at the current cell state.

We then examined how much change exists for H3K27me3 between normal and FSHD myoblasts. Here we found 12,000 – 55,000 peaks for H3K27me3. As all samples but FSHD1 sample 10 had more than 35,000 peaks, we excluded it from the downstream analysis. After identifying the peaks from normal myoblasts that had no overlapping peaks in FSHD myoblasts—a total of 1,498—we performed the same analysis using GREAT as we had with the H3K9me3 peaks. This time, we found no enrichment for any GO terms, in contrast to H3K9me3. This suggests that the enrichment of specific GO terms in the H3K9me3 set is specific, while also suggesting that H3K27me3 functions separately in FSHD.

To further this analysis, understanding that peaks may potentially have increased H3K9me3 as well as decreased H3K9me3, we used the combined sets of all H3K9me3 ChIP-seq reads to call peaks and then determine the number of reads contributed from each. After normalization (see methods), the matrix of read counts was used as input to edgeR [11] to determine the differentially methylated regions (DMRs). A total of 238 DMRs were identified in FSHD1, 210 of which had decreased methylation, and 28 of which had increased methylation. In agreement with what is seen at D4Z4, most regions appear to lose H3K9me3. The set of 238 DMRs serve as a subset of the 548 peaks

previously described, with higher stringency required when using edgeR and low between-sample variance to call a region differentially methylated. Unlike FSHD1, only 1 DMR was found in FSHD2 (decreased methylation); the FSHD2 samples show much higher between-sample variance than those of FSHD1, thereby preventing edgeR from calling many regions as being differentially methylated.

We again used the GREAT tool to identify genes nearby these DMRs and found that they were enriched near regulatory sequences. Further examination showed that 5 of the 238 regions sat upstream of D4Z4 and 2 upstream of a D4Z4 homolog on chromosome 10. In both instances, we find a 2 – 4 fold decrease of H3K9me3 upstream of the repeat arrays. However, no specific GO terms were enriched in this smaller set of DMRs. The loss of H3K9me3 upstream of D4Z4 led us to ask the question of whether or not we see other clusters of DMRs elsewhere in the genome. We plotted the presence of these regions across the genome on an ideogram and found several other clusters, with the one on chromosome 6 being the most dense (Figure 5.1). Visualization of this 2 Mb region shows a complete loss of H3K9me3 in FSHD1, while only 1 out of three samples lose H3K9me3 in FSHD2. In contrast, there is little signal for H3K27me3 for either FSHD1 or FSHD2 samples in this region. Genes within this region do not appear to be differentially expressed in the patient myoblasts, though they may be misregulated at other points such as differentiation into myotubes. More research will have to be done to identify the effect of loss of H3K9me3 in these regions.

One exception to this is found upstream of D4Z4, with the upregulation of FRG1 in FSHD1. While several genes lie upstream of D4Z4, including TUB4Q, FRG1, and FRG2, only FRG1 is upregulated. Unlike these other genes, we identified DUX4 binding sites at FRG1; DUX4 expression in FSHD is a likely explanation for FRG1 upregulation, as it was identified as a transcriptional activator with potential to regulate many genes differentially expressed in FSHD [4].

Table 5.1. Sequencing Summary

	Cells	RNAseq	Nanostring	WCE	Rad21	H3k9me3	H3k27me3
Control	Control 2 (F)	180M	Done	7M/4M	19M/16M/105K	85M/57M/54K	51M/39M/47K
	Control 10 (M)	205M	Done	10M/3M	27M/19M/34K	73M/49M/63K 73M/50M/57K	100M/80M/44K
FSHD1	FSHD1 6 (F)	78M	Done	7M/5M	15M/11M/29K	67M/48M/57K	37M/29M/45k
	FSHD1 10 (M)	65M	Done	15M/10M	22M/20M/37K	58M/42M/52K	10M/8M/12k
	FSHD1 4 (M)	78M	Done	10M/6M	15M/14M/28K	55M/34M/50K	29M/22M/45K
FSHD2	FSHD2 5 (M)	63M	Done	10M/4M	18M/12M/56K	69M/49M/50K	35M/28M/35K
	FSHD2 6 (M)	58M	Done	15M/9M	21M/10M/38K	73M/45M/55K	37M/28M/55K
	FSHD2 7 (M)	70M	Done	15M/11M	34M/27M/153K	39M/28M/43k	11M/9M/45k

	Cells	RNAseq	Nanostring	WCE	Rad21	CTCF	H3k9me3	H3k27me3
Rep1	Undiff	74M/62M	Done	8M/5M	12M/3M/27K	12M/4M/40K	42M/21M/50K	28M/17M/32K
	Diff	102M/84M	Done	16M/11M	14M/8M/28K	14M/9M/28K	58M/32M/51K	26M/16M/37K
Rep2	Undiff	78M/60M	Done	11M/4M	15M/13M/32K	In Progress	53M/15M/27K	24M/22M/30K
	Diff	120M/111M	Done	12M/5M	18M/15M/40K	In Progress	51M/23M/30K	21M/19M/34K

Notation: Total Reads/Reads after mapping/Peaks

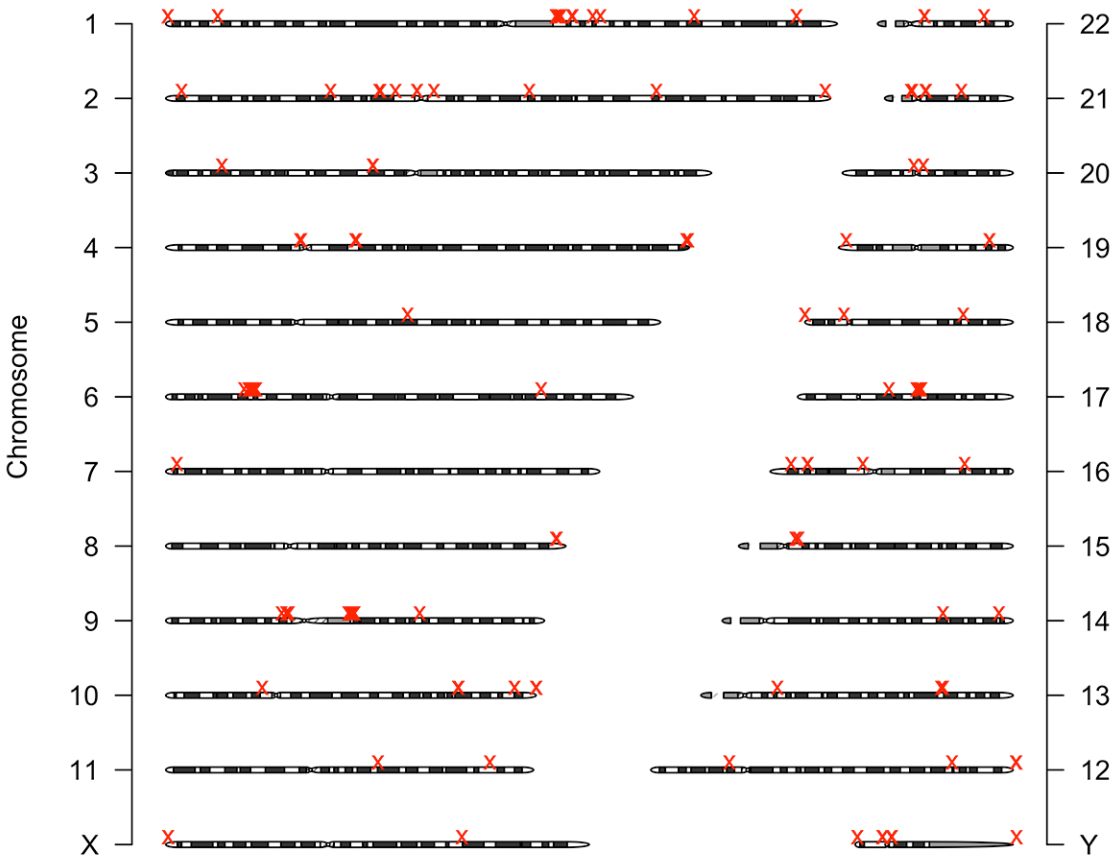


Figure 5.1. Ideogram showing the genomic placement of differentially methylated peaks.

5.3.2 Genes involved in skeletal muscle development are down regulated

In order to identify gene expression changes between normal and FSHD myoblasts, Michelle Chen performed RNA-seq analysis of all samples listed in Table 1. After alignment of reads, RSEM [12] was used to calculate the estimated counts for each transcript (see methods for details). Normalized counts were input into edgeR, and the list of differentially expressed genes (DEGs) for FSHD1 and FSHD2 were determined. A total of 812 genes were differentially expressed in FSHD1, and 128 genes were differentially expressed in FSHD2. Similar to what is seen in the H3K9me3 ChIP-seq data, the between sample variance was higher in FSHD2, producing fewer DEGs. Of the gene sets for each, 76 overlap (Figure 5.2). Using the DAVID bioinformatics website [13], we looked at the enrichment of GO terms in each gene set. For the set of 483 genes downregulated in FSHD1, we find enrichment for genes involved in muscle development (Figure 5.3). Upregulated genes are enriched for those involved in protein folding and cell migration. For FSHD2, we do not find any enriched GO terms. Sample variability is likely hindering the identification of specific gene categories for FSHD2.

5.3.3: Upregulated genes in myoblast differentiation are downregulated in FSHD

In order to identify genes potentially involved in differentiation defects present in FSHD, Michelle Chen performed RNA-seq in undifferentiated and differentiated KD3 myoblasts. KD3 immortalized myoblasts [14], which are easier to culture than primary myoblasts and differentiate more efficiently, were used due to the low differentiation efficiency of the primary myoblasts in our hand. After using eXpress [15] to calculate the estimated counts, we determined the number of DEGs using edgeR. Interestingly, a

heatmap of genes differentially expressed upon differentiation shows a general decrease of these genes in FSHD1 and FSHD2 (Figure 5.4). A t-test indicates that these differences are statistically significant. Two possibilities emerge to explain this. For the first, the genes important for differentiation may be downregulated in FSHD as a result of shifts in epigenetic marks and transcriptional regulators. The lower expression prior to differentiation may then inhibit efficient differentiation. Conversely, the decreased expression of these genes could be due to lower rates of already [spontaneously] differentiating primary myoblasts in the cultures. What specifically is triggering the general decrease of these genes in FSHD is a question for future study.

5.3.4 miRNAs are differentially expressed in FSHD

Previous studies by other groups have sought to identify whether or not miRNAs might be differentially expressed in FSHD [16-18]. However, the results from each group showed no overlap with those of any other group, whether those miRNAs found to be differentially expressed either in patient muscle or myoblasts. In order to correlate the miRNA expression differences with our RNA-seq and ChIP-seq data, we used Nanostring against a panel of 800 known miRNAs to identify miRNAs in all of the samples in Table 1. This data set grants us a comprehensive look at the gene and miRNA expression changes in FSHD, allowing us to identify how the epigenetic changes in FSHD might impact them. Weihua Zeng performed the Nanostring protocol and generated the count information (with replicates) for each of the 8 samples examined. A matrix of the normalized counts was used as input to edgeR, allowing us to identify differentially expressed miRNAs (DEMs).

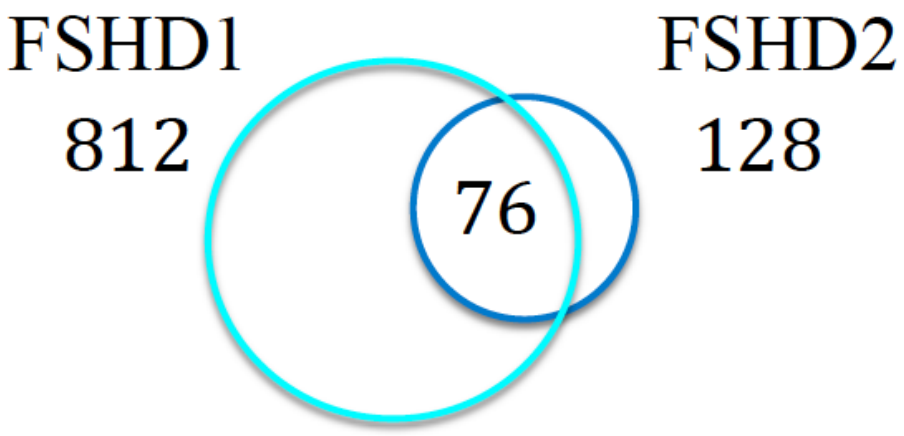
We found that miRs 206 and 145 were upregulated in FSHD1, with miR 145 being upregulated in both FSHD1 and FSHD2. miR-206, a so called "myo-miR," was downregulated in FSHD1 in both replicates. Other miRNAs, such as mir-133a/b were also seen to be downregulated, though with more variation between replicates. Our data also do not overlap with that of other groups, though the disruption of miRNAs known for their role in muscle developmental processes does correlate nicely with the gene expression changes previously discussed.

Figure 5.2. Differentially expressed genes in FSHD1 and FSHD2

A. Overlap between genes differentially expressed in FSHD1 and FSHD2

B. GO terms enriched in differentially expressed genes for FSHD1

A.



B.

Term	Upregulated/ Downregulated	P-value
Myogenesis	Downregulated	1x10-03
Cell Migration	Upregulated	1x10-02
Protein Folding	Upregulated	1x10-02

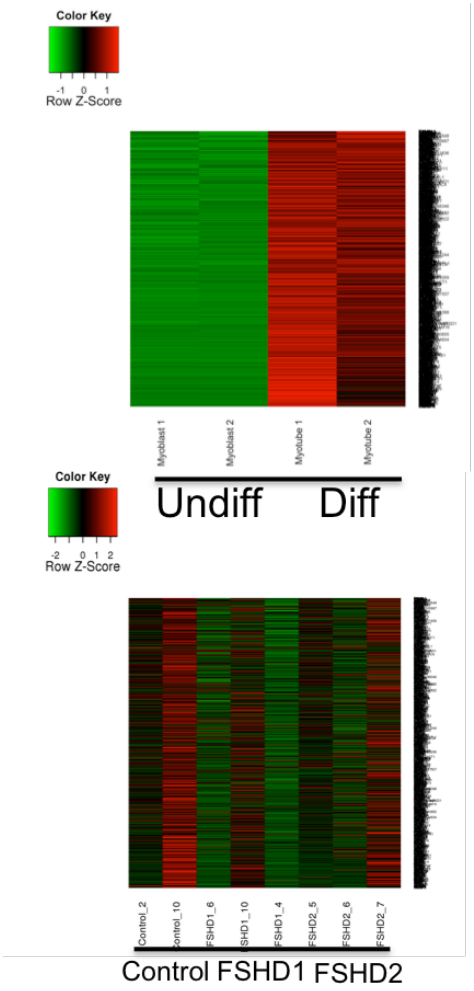
Figure 5.3. Genes upregulated upon differentiation are downregulated in FSHD

Top left: Upregulated gene heatmap for KD3

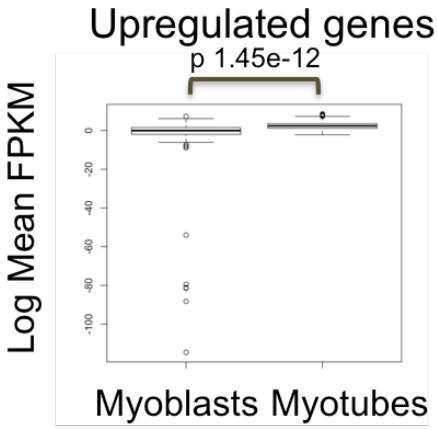
Top right: Boxplot and significance of upregulated genes in KD3

Bottom left: Downregulated gene heatmap for FSHD

Bottom right: Boxplot and significance of genes downregulated in FSHD


Color Key
-1 0 1
Row Z-Score
Myoblast 1
Myoblast 2
Myotube 1
Myotube 2
Undiff
Diff
Color Key
-2 0 1 2
Row Z-Score
Control_2
Control_10
FSHD1_6
FSHD1_10
FSHD1_4
FSHD2_5
FSHD2_6
FSHD2_7
Control FSHD1 FSHD2

KD3


Upregulated genes
p 1.45e-12
Log Mean FPKM
0
-20
-40
-60
-80
-100
Myoblasts Myotubes

FSHD

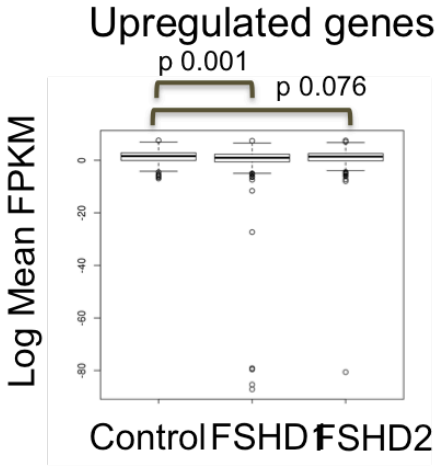

Upregulated genes
p 0.001
p 0.076
Log Mean FPKM
0
-20
-40
-60
-80
Control FSHD1 FSHD2

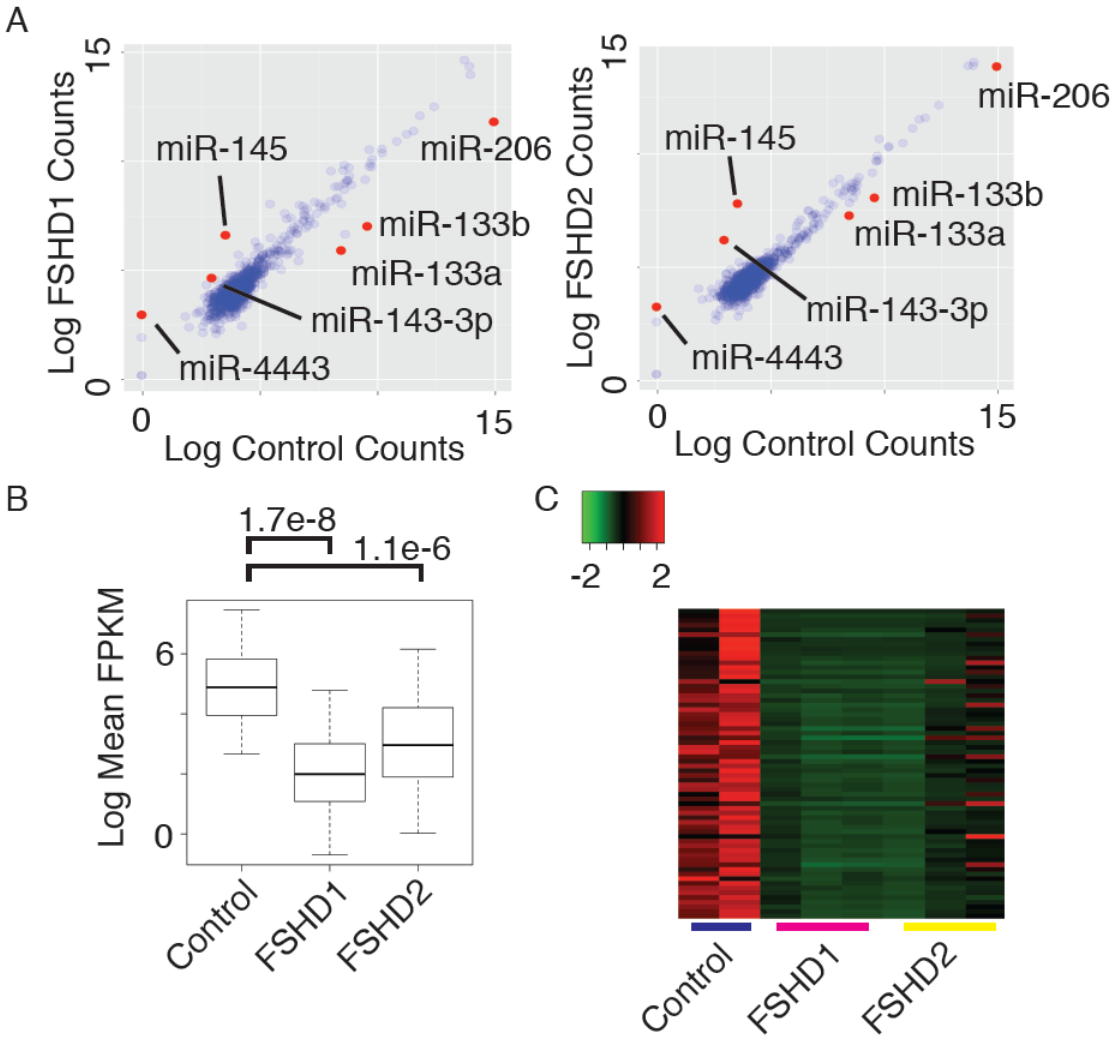
Figure 5.4. miRNA expression in FSHD

A. (left) miRNA expression between normal and FSHD1

   (right) miRNA expression between normal and FSHD2

B. Putative target expression for miR-145

C. Heatmap of expression of putative targets in FSHD1 and FSHD2


A
Log FSHD1 Counts
15
0
0
15
Log Control Counts
miR-145
miR-206
miR-133b
miR-133a
miR-143-3p
miR-4443
Log FSHD2 Counts
15
0
0
15
Log Control Counts
miR-206
miR-145
miR-133b
miR-133a
miR-143-3p
miR-4443
B
1.7e-8
1.1e-6
Log Mean FPKM
6
0
Control
FSHD1
FSHD2
C
-2
2
Control
FSHD1
FSHD2

5.4 Discussion

While incomplete, our initial work in this study has revealed a number of novel findings. Although we had previously known that the loss of H3K9me3, HP1γ, and cohesin occurred at D4Z4, little was known about other regions on other chromosomes. The decrease of H3K9me3 at many sites throughout the genome, particularly near genes that are associated with skeletal muscle development, supports the hypothesis that FSHD is indeed an epigenetic abnormality disease. However, it remains unclear how this H3K9me3 decrease occurs. One hypothesis is that the heterochromatin formation at these regions is dependent on direct interaction with D4Z4; upon loss of heterochromatin at D4Z4 and loss of interactions mediated by cohesin, the heterochromatin at these regions could be compromised. While few genes in these regions are differentially expressed, their proper expression could be hindered by their accessibility later on, with expression at inappropriate times preventing efficient differentiation and other developmental defects. Research on these questions is ongoing.

The expression and miRNA profiling of the normal and FSHD primary myoblasts is the first comprehensive study performed in the same cells along with the profile of heterochromatin state. Our epigenomics approach can allow us to correlate the epigenetic changes taking place genome-wide with altered expression. We have initially found many genes to be differentially expressed in FSHD, and two important miRNAs; both are promising results. New samples being sequenced will give us more statistical power to better characterize the expression patterns in the disease in our ongoing work.

5.5 Methods

RNA-sequencing

Total RNA from patient-derived, primary myoblasts was extracted using a Qiagen RNeasy kit (Qiagen, Germany). Library construction was performed using the Illumina protocol. Libraries were sequenced using an Illumina Hi-Seq sequencer housed in the University of California, Irvine Genomics High-Throughput Core Facility.

Antibodies

ChIP was performed using antibodies directed against RAD21 (cohesin, [6]), HP1γ (Abcam, #10480), RNAPII (Millipore, 8WG16), H3K9me3 (in-house, FAB fragment), and H3K27me3 (Abcam, ab6147).

Chromatin Immunoprecipitation

Chromatin Immunoprecipitation was performed as previously described [19]. For each ChIP experiment, chromatin from 1E07 cells was used. Cells were crosslinked for 10 minutes using 1% formaldehyde diluted in cell culture media, lysed, and sonicated using a Bioruptor sonicator (Diagenode) to a fragment size around 200 bps. Samples were pre-cleared for 1 hr using BSA and Protein A sepharose beads (GE Healthcare). Pre-cleared extracts were incubated with the antibodies overnight, and the IP was performed with Protein A beads and then washed. DNA was eluted off of the beads, reverse-crosslinked overnight, and then purified using a PCR purification kit (Qiagen). ChIP library construction was performed using the Myers protocol (Protocol.pdf).

Samples were submitted to the UCI Genomics and High-Throughput Facility for sequencing using the Illumina protocol and the Illumina Genome Analyzer (GA) system.

Preprocessing

Pre-processing of the sequenced reads was performed to remove low quality sequence and remove adapter sequence. In order to remove adapter sequence, the adapter sequence used for each library was removed from each read using the tool CutAdapt (Python script, https://code.google.com/p/cutadapt/). After adapter removal, sequences are trimmed for low quality sequence (Phred score <= 20) using FastX. If trimmed sequences are less than 20 bp in length, they are removed entirely. Sequences were mapped back to the human genome draft 19 (hg19) from the UCSC genome browser [20]. Mapping was performed using Bowtie version 1.0 [21], with parameters –n 2 –m 3.

Peak calling

Peak calling was done using AREM [22] with settings -no-EM, accepting only reads mapping to fewer than 3 locations, and –pval X, where X is adjusted to produce a FDR lower than 5%. For peak calling of histone modifications, SICER [9] was used to call peaks with width 300, gap size of 600, and FDR less than 0.1%. All other settings were as recommended in the user manual for each package.

Data comparison

All other data processing occurred through the R statistical package [23] and Perl and Python scripts written by the author. Peak overlap was calculated using pybedtools [24].

Transcript analysis

In order to determine transcript abundance, reads were mapped to a transcriptome constructed from the refFlat files available from the UCSC genome browser using bowtie. Settings for bowtie include –n 2 –k 20 –best –strata -S. Output from bowtie (SAM format) was input into RSEM [12] or eXpress [15] using default parameters to generate normalized transcript counts. Data were quantile normalized using Bioconductor and R. Differentially expressed transcripts were found using edgeR with default methods.

Nanostring

Total RNA was extracted and prepared according to manufacturers instructions. RNA was run against sample sets, to identify expression of up to 800 miRNAs. Data was normalized to the top 100 expressed miRNAs, with 2 biological replicates for each sample. Differentially expressed miRNA was determined using the normalized counts input into edgeR.

5.6: References

1. Tawil, R. and S.M. Van Der Maarel, *Facioscapulohumeral muscular dystrophy.* Muscle Nerve, 2006. **34**(1): p. 1-15.
2. van Deutekom, J.C., et al., *Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1.* Hum Mol Genet, 1996. **5**(12): p. 1997-2003.
3. Lunt, P.W., *44th ENMC International Workshop: Facioscapulohumeral Muscular Dystrophy: Molecular Studies 19-21 July 1996, Naarden, The Netherlands.* Neuromuscul Disord, 1998. **8**(2): p. 126-30.
4. Geng, L.N., et al., *DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy.* Dev Cell, 2012. **22**(1): p. 38-51.
5. Sacconi, S., et al., *The FSHD2 Gene SMCHD1 Is a Modifier of Disease Severity in Families Affected by FSHD1.* Am J Hum Genet, 2013.
6. Zeng, W., et al., *Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD).* PLoS Genet, 2009. **5**(7): p. e1000559.
7. Zeng, W., et al., *Genetic and epigenetic characteristics of FSHD-associated 4q and 10q D4Z4 that are distinct from non-4q/10q D4Z4 homologs.* Hum Mutat, 2014. **35**(8): p. 998-1010.
8. de Greef, J.C., et al., *Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD.* Hum Mutat, 2009. **30**(10): p. 1449-59.
9. Xu, S., et al., *Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells.* Methods Mol Biol, 2014. **1150**: p. 97-111.
10. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol, 2010. **28**(5): p. 495-501.
11. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
12. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
13. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
14. Shiomi, K., et al., *CDK4 and cyclin D1 allow human myogenic cells to recapture growth property without compromising differentiation potential.* Gene Ther, 2011. **18**(9): p. 857-66.

15. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
16. Dmitriev, P., et al., *Defective regulation of microRNA target genes in myoblasts from facioscapulohumeral dystrophy patients.* J Biol Chem, 2013. **288**(49): p. 34989-5002.
17. Cheli, S., et al., *Expression profiling of FSHD-1 and FSHD-2 cells during myogenic differentiation evidences common and distinctive gene dysregulation patterns.* PLoS One, 2011. **6**(6): p. e20966.
18. Harafuji, N., et al., *miR-411 is up-regulated in FSHD myoblasts and suppresses myogenic factors.* Orphanet J Rare Dis, 2013. **8**: p. 55.
19. Chien, R., et al., *Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression.* J Biol Chem, 2011. **286**(20): p. 17870-8.
20. Karolchik, D., et al., *The UCSC Genome Browser database: 2014 update.* Nucleic Acids Res, 2014. **42**(Database issue): p. D764-70.
21. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
22. Newkirk, D., et al., *AREM: aligning short reads from ChIP-sequencing by expectation maximization.* J Comput Biol, 2011. **18**(11): p. 1495-505.
23. Team, R.C., *R: A language and environment for statistical computing*. 2013, Vienna, Austria: R Foundation for Statistical Computing.
24. Dale, R.K., B.S. Pedersen, and A.R. Quinlan, *Pybedtools: a flexible Python library for manipulating genomic datasets and annotations.* Bioinformatics, 2011. **27**(24): p. 3423-4.

Chapter 6

Conclusion

6.1 AREM is able to identify cohesin peaks in repetitive sequence.

One of the early challenges in analysis of ChIP-chip and ChIP-sequencing data was the design of algorithms to determine regions of enrichment (protein binding to chromatin). While many groups had successfully developed tools to identify peaks for transcription factors, histone modifications, and other types of ChIP-sequencing data, one area largely unexplored was that of identifying peaks in repetitive sequence. To address this need, we developed AREM, a tool capable of assigning a probability of each read belonging to a peak region, and thereby allowing the users to identify peaks in and near repetitive sequence. Using this data, we showed that we could find more peaks with a minimal increase in the FDR (Table 1 Chapter 2) using the wildtype Rad21 ChIP-seq data later examined in Chapter 3. Importantly, we also showed that cohesin generally binds to the same types of repetitive regions (Figure 3, Chapter 3), with decreased binding in Nipbl (+/-) mutant cells. This data, along with the rate of the CTCF motif contained within the cohesin binding sites, helps validate our approach and shows the utility of being able to call peaks in repetitive regions.

6.2 Cohesin binding decreases genome-wide in Nipbl (+/-) MEFs.

Although other groups had previously shown that many genes in CdLS patients and the CdLS mouse model were misregulated, it was unclear which genes might be directly regulated by cohesin, which Nipbl is required to load. We were able to successfully identify cohesin target genes using ChIP-sequencing of Rad21 in both wildtype and mutant MEFs. To do this, we began by characterizing the cohesin binding sites genome-wide (Figure 1, Chapter 3). Rather than a complete loss of cohesin at many

locations across the genome, we found that cohesin decreases genome-wide by plotting the density of reads at each wildtype cohesin binding site in both the wildtype and mutant samples. Similar to other groups, we found that cohesin binding sites often contain the CTCF motif (Figure 2, Chapter 3). However, peaks called in both wildtype and mutant had the highest frequency of containing the motif. Again similar to previous studies, cohesin was found to be enriched at the promoter region, suggesting that it might have a more direct impact on gene expression—potentially through enhancer-promoter interactions.

6.3 Cohesin binding is correlated with expression changes in Nipbl (+/-) MEFs.

Using a KS plot, we showed the tendency for genes that have altered expression in mutant MEFs to be bound by cohesin in the gene region (Figure 4, Chapter 3). In particular, cohesin binding at the promoter was the most enriched. Moreover, genes downregulated in the mutant were most highly associated with cohesin binding, which suggests that cohesin's role in gene activation is most affected by Nipbl haploinsufficiency. Greater than 50% of the differentially expressed genes in CdLS were bound by cohesin (Figure 5, Chapter 3), indicating that many of the gene expression changes in CdLS may be directly mediated by a decrease of cohesin binding due to Nipbl haploinsufficiency, with others perhaps being regulated by other factors (such as Nipbl as we later show) or indirectly upon the misregulation of transcription factors directly regulated by cohesin.

6.4 Decreased cohesin binding affects chromatin interactions.

One of the ways that cohesin can mediate gene expression is through the establishment of long-range chromatin interactions. Since we see a decrease of cohesin binding in the Nipbl (+/-) MEFs, it was important to ask whether this could affect these interactions. We found indeed that these interactions decrease both in the mutant MEFs and upon depletion of Nipbl in wildtype MEFs. This suggests that the effect of Nipbl haploinsufficiency could be substantial rearrangement of chromatin topology, impacting gene regulation and other nuclear processes.

6.5 NIPBL binds to chromatin near both cohesin and CTCF.

In light of recent studies into NIPBL, the nature of NIPBL binding patterns across the genome was not clear. In MEFs, it was shown that cohesin, CTCF and Nipbl colocalize, while in HeLa neither cohesin nor CTCF colocalize with NIPBL. This contradiction was partially explained by the differences in antibodies used in the most recent study. We have successfully identified NIPBL binding sites in HeLa using an in-house antibody and shown that NIPBL does in fact bind near cohesin and CTCF in most cases (as is expected by NIPBL's role in cohesin loading). About 10% of NIPBL binding sites are free of cohesin and CTCF. In either case, NIPBL is enriched in the promoter, particularly near the TSS.

6.6 NIPBL can regulate gene expression independently of cohesin.

After identification of NIPBL binding sites across the genome, we tested whether or not NIPBL could regulate the expression of genes where NIPBL—but not cohesin—

was bound. We showed that 72 of the 273 genes bound by [only] NIPBL were differentially expressed upon depletion of NIPBL. Moreover, two genes, NSFP1 and FBXL16, were upregulated upon depletion of NIPBL but not upon depletion of cohesin. This data supports the notion that NIPBL can regulate expression of genes independently of cohesin, and that NIPBL can repress as well as activate gene expression upon binding to the promoter of different genes.

6.7 H3K9me3 decreases at many genomic regions in FSHD.

Using ChIP-seq to examine differences in heterochromatin between normal and FSHD myoblasts, we identified over 200 regions with a decrease of cohesin binding across the genome. These regions are enriched near genes important for myoblast differentiation, though the genes contained in these regions are not necessarily upregulated as a result. Instead, these genes may be "poised" for expression at a time when the appropriate factors are present, such as DUX4 at FRG1. Our intial findings suggest that the loss of H3K9me3 may serve as a hallmark of FSHD, both at D4Z4 and elsewhere in the genome.

6.8 Can tag enrichment on tandem repeats be identified?

While we have been able to gather data supporting many of our hypotheses, there are many more questions that remain. We have been able to develop an algorithm that is able to efficiently identify cohesin binding sites in repetitive sequence, with known limitations surrounding large, tandem repeats such as exist at D4Z4. Better approaches to identify binding in these regions will hopefully be developed, though third generation

sequencing techniques may preclude a need for this due to the length of the sequencing reads and the ability to sequence much longer input DNA sequences.

6.9 How does decreased cohesin binding affect each stage of development in CdLS?

Our lab has been able to identify many different cohesin binding sites in MEFs, and to show both correlation and actual regulation of gene expression through specific long range interactions. How CdLS progresses during development, and how different tissues are affected by Nipbl haploinsufficiency still remains unknown in many instances, and further work can be done to identify how the cohesin target genes shift between cell types. Also, predicated on our findings with our in-house antibodies, a better understanding of the global distribution of Nipbl on chromatin in these MEFs may yield a better understanding about what genes Nipbl may target in these MEFs independently of cohesin.

6.10 HCF-1 and YY1 may interact with NIPBL.

The presence of HCF-1 and YY1 near NIPBL binding sites may suggest interactions occur between the two, allowing them to work in concert to regulate gene expression in HeLa and other cell types. Since HCF-1 is known to bind so many different chromatin modifiers, with this recruitment being context-dependent, much more work needs to be done to further identify which genes are bound by these factors, and what chromatin modifiers are recruited when. Upon NIPBL depletion in HeLa, we see both upregulation and downregulation of target genes, indicating that this recruitment could be complex.

6.11 Further characterization of myoblasts in FSHD is needed.

Our results characterizing the decrease of H3K9me3 at locations throughout the genome, and their occurrence near genes involved in skeletal muscle development, along with identifying disease-specific expression patterns, is just a first step. We intend to continue characterizing the epigenomic changes taking place in FSHD using sequencing techniques to examine DNase Hypersensitivity Sites, and further identify expressed genes using ATAC-seq. By having a comprehensive epigenomic and expression profile in these myoblast samples, we hope to better understand the underlying mechanism for FSHD.

Beyond the epigenetic signatures of FSHD, we hope to understand the role of cohesin in FSHD. While we know that cohesin is lost at D4Z4 in FSHD, it's not clear if or how cohesin binding patterns may change globally. The loss of cohesin binding could result in a decreased interaction of D4Z4 with other regions of the genome and affect gene expression. Conversely, the loss of H3K9me3 could precipitate loss of cohesin binding in other regions in a D4Z4-independent manner through some other mechanism. To explore this, we are performing RAD21 ChIP-seq to correlate with the H3K9me3 distribution in normal and FSHD myoblasts.