

Isolation and Identification of *De Novo* Long
Noncoding RNAs from Mouse Myoblasts and
Embryonic Stem Cells

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Catherine Ann Alsager Lee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Nobuaki Kikyo

November, 2012

© Catherine Ann Alsager Lee 2012
ALL RIGHTS RESERVED

Acknowledgements

I extend my deepest gratitude to Nobuaki Kikyo, my advisor and principal investigator, whose guidance, patience, and expertise aided the writing of this thesis in innumerable ways. I am also grateful to Sue Keirstead, Atushi Asakura, Ying Zhang, and all the members of the Kikyo Lab for their help and support, without whom this work would not have been possible.

Dedication

This work is dedicated to my parents and to Tristan, whose love and encouragement during the long hours of research and writing were integral to the completion of this task. And also to my grandfather, Dr. Kyu Lee, who inspired in me my enthusiasm for science and who taught me the value of hard work and education.

Abstract

Long noncoding RNAs (lncRNAs) are a pervasive class of transcripts whose importance and biological relevance are only beginning to be elucidated. LncRNAs have been detected in nearly every cell type and found to be fundamentally involved in many biological processes; however, studies that characterize lncRNA expression during certain periods of development are largely missing. Here, we demonstrate how a pool of potentially relevant lncRNAs can be identified using a RNA-chromatin immunoprecipitation (RNA-ChIP) technique that pulls down sufficient amount of RNA to send for sequencing. In our initial experiment, we attempted to identify lncRNAs bound to the MyoD protein in myoblast cells; however, the lack of immunoprecipitation-compatible highly specific antibodies against MyoD prevented us from pursuing this project. As an alternative, we successfully identified lncRNAs bound to the histone-modifying complex COMPASS, as well as those bound to the master pluripotency factors Oct4 and Sox2 in mouse embryonic stem cells. This study provides a proof-of-principle to identify lncRNAs potentially involved in chromatin regulation of pluripotency.

Contents

Acknowledgements	i
Dedication	ii
Abstract	iii
List of Tables	vi
List of Figures	vii
1 Introduction	1
2 Methods	6
2.1 Cell Culture	6
2.2 Myoblast Differentiation	6
2.3 Immunostaining	7
2.4 Antibody Testing	7
2.5 DNA Length Optimization for ChIP	10
2.6 ChIP	11
2.7 RNA-ChIP	12
2.8 RNA-sequencing and Data Analysis	14
3 Results	16
3.1 ITS Differentiates Myoblasts into Myotubes	16
3.2 Western Blots Identify Immunoprecipitation- compatible Antibodies	19

3.3	Chromatin Size is Optimized for Immunoprecipitation	24
3.4	ChIP with MyoD Shows Non-specific Binding	28
3.5	RNA-ChIP with MyoD Leads to Re-testing of Antibodies	33
3.6	RNA-seq Data Analysis	34
4	Discussion	36
5	Conclusion	38
6	Glossary of Bioinformatics Terms	39
	References	40

List of Tables

2.1	Primary antibodies used for western blotting.	9
2.2	Secondary antibodies used for western blotting.	9
2.3	Antibodies used for ChIP experiments.	12
2.4	Sequences of primers used for qPCR.	13
3.1	Mapping percentages of RNA-seq 1 samples.	34

List of Figures

1.1	Overview of lncRNA populations based on their location in the genome.	3
3.1	MHC expression in myotubes from differentiated C2C12 cells.	17
3.2	Myogenin expression in myotubes from differentiated C2C12 cells. . . .	18
3.3	Western blots of IP-compatible antibodies against muscle-specific proteins.	20
3.4	Western blots of IP-compatible antibodies against components of the COMPASS complex.	21
3.5	Western blots of IP-compatible antibodies against Oct4.	22
3.6	Western blots of IP-compatible antibodies against Sox2.	23
3.7	Size distribution of DNA after sonication of chromatin ranges from 0.1kb- 12kb.	25
3.8	Size distribution of DNA after sonication of chromatin ranges from 0.1kb- 3kb	26
3.9	Size distribution of DNA after incubation with micrococcal nuclease and 45 pulses of sonication of chromatin ranges from 0.3kb-1kb.	27
3.10	Regulatory regions of the MyoD gene.	29
3.11	Relative expression fold change of DNA from ChIP determined by qPCR	31
3.12	Relative expression fold change of DNA from ChIP determined by qPCR	32
3.13	Representative examples of peaks viewed with IGV.	35

Chapter 1

Introduction

Long noncoding RNA (lncRNA) is operationally defined as RNA longer than 200 bases that does not encode mRNA, rRNA or tRNA [1, 2]. Although several lncRNAs have been sporadically identified and characterized in the past 20 years, genome-wide identification of lncRNAs has only recently become possible with the advent of high-throughput sequencing technologies of cDNA (RNA-seq). Evidence that this field is gaining momentum can be seen in the most recent report of the ENCODE (Encyclopedia of DNA Elements) project published in September 2012, which described 9,640 lncRNA loci in comparison to 20,687 protein-coding genes in 15 human cell lines [3, 4, 5]. This ratio of lncRNAs and protein-coding genes underscores the potential magnitude and diversity of the biological effects mediated by lncRNAs. Indeed, despite the fact that only about 100 lncRNAs have been functionally characterized to date [4], it has become clear that lncRNAs are involved in almost every aspect of cellular and molecular biology. LncRNAs control cell differentiation, development, cancer progression, and cell metabolism, among other cell functions. At the gene expression level, lncRNAs regulate all processes of RNA metabolism including chromatin modification, transcription, splicing, RNA transport, and translation. LncRNAs themselves are transcribed from intergenic regions, exons, introns, and their overlapping regions (Figure 1.1A and 1.1B). At the mechanistic level, lncRNAs serve as "scaffolds" providing platforms to assemble RNA-protein complexes, "guides" to recruit RNA-protein complexes to target genes, and "decoys" by binding to and sequestering regulatory proteins away from their target DNA sequences [1, 2].

The first challenge in studying lncRNAs is how to collect RNA pools that potentially contain lncRNAs of interest. One can prepare RNA pools by simply isolating total RNA from cells or tissues in an unbiased manner; however, immunoprecipitation-based approaches are also commonly used to enrich lncRNAs associated with specific proteins. Cross-linking with UV or formaldehyde followed by fragmentation of chromatin is used to immunoprecipitate RNA-chromatin complexes (RNA-chromatin immunoprecipitation or RNA-ChIP) [6, 7, 8].

For any immunoprecipitation-based approaches, specificity and affinity of the antibodies are decisive factors for the success or failure of the projects. While the specificity of the antibodies is commonly verified by detecting only one band in western blotting, the antibodies may react with other proteins when detergents are used at a low concentration during immunoprecipitation. One solution to address the specificity issue is to use multiple antibodies against the same protein and select reproducibly co-precipitated lncRNAs for further study. Similarly, immunoprecipitation of several different subunits within a single protein complex is also an option to identify lncRNAs that are likely to be genuinely interacting with the complex.

After collection by immunoprecipitation, the sequences of the RNA pool of interest can be obtained by RNA-sequencing (RNA-seq). RNA-seq is a powerful tool based on the principles of next-generation sequencing that can be applied to the detection and quantification of lncRNAs. It works on a genome-wide scale at single nucleotide resolution and is not limited to detecting already known sequences. Thus, it can be used to discover previously unknown lncRNAs in an unbiased manner [9].

For RNA-seq, one must decide whether to use total RNA or polyadenylated RNA. The presence of rRNA (around 80-85% of total RNA) and tRNA (15%) [10, 11] can drastically reduce the diversity of a cDNA library during amplification of cDNAs. Polyadenylated RNA is frequently used for RNA-seq to avoid this problem. However, given the prevalence of non-polyadenylated lncRNA in the genome (around 40% of total lncRNAs), the disadvantage of losing this fraction is not negligible [12]. One solution to this problem is to use commercially available kits to remove rRNA from total RNA without losing non-polyadenylated RNA.

After sequencing, a typical pipeline for RNA-seq analysis is to align the reads generated by sequencing to the UCSC mouse mm9 or human hg19 reference genomes using

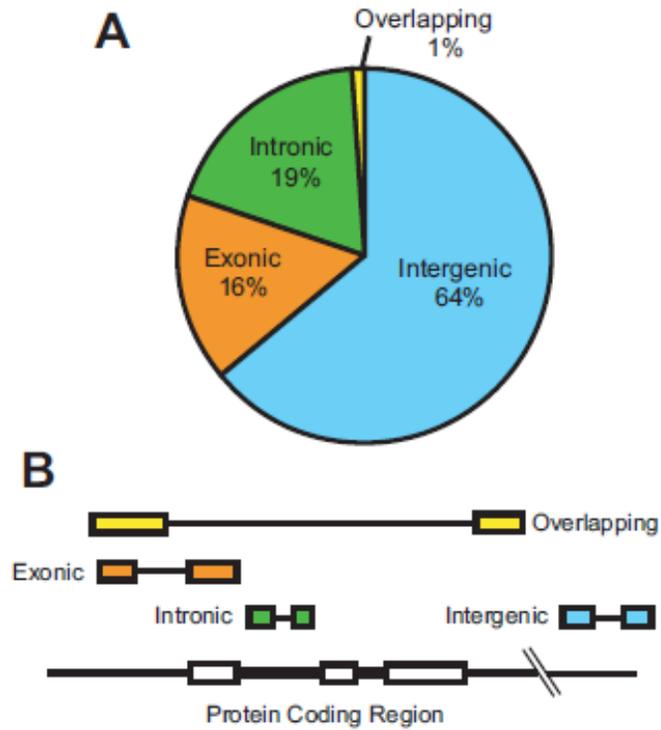


Figure 1.1: Overview of lncRNA populations based on their location in the genome. LncRNAs can be categorized into subgroups of intergenic, exonic, intronic, and overlapping according to where they are found relative to nearby protein-coding genes. (A) Proportion of lncRNA subgroups [3]. (B) Location of each type of lncRNA.

software programs such as the short-read mappers Bowtie 2 [13] and Burrows-Wheeler Aligner [14], and the splice-junction identifier TopHat [15]. Next, the reads are used to assemble a transcriptome and discover previously unannotated transcripts with programs such as Cufflinks [16], which relies on a reference annotation database, or Scripture, which builds the transcriptome *ab initio* [17]. From here, novel lncRNAs can be identified by excluding protein-coding transcripts and annotated lncRNAs based on the databases of RefSeq, ENCODE, and FANTOM (Functional Annotation of the Mammalian Genome) [18], as well as the two databases of experimentally verified lncRNAs generated by the Mattick lab: lncRNAdb [19] and NRED (Noncoding RNA Expression Database) [20].

Novel lncRNAs often undergo further scrutiny to ensure that they are not transcriptional noise and that they indeed do not encode proteins. Finally, behavior of the system of interest following knockdown or overexpression of the lncRNAs is typically a final step in functionally characterizing a novel lncRNA.

The transcription factor MyoD is the master regulator of muscle differentiation. The Kikyo lab recently demonstrated that when the transactivation domain of MyoD is fused to the Oct4 protein, the efficiency of reprogramming of fibroblasts to induced pluripotent stem cells (iPSCs) is increased more than 10-fold [21]. The power of MyoD has been partly attributed to its interaction with the Pbx and Meis homeodomain proteins through its H/C region and helix III region as this allows MyoD to bind to its target promoters [22]. However, this interaction is essential for only a subset of MyoD target genes [23]. It remains to be elucidated how other target genes are activated by MyoD [24] and it is possible that long noncoding RNAs (lncRNAs) interacting with MyoD are playing a role. We set out to identify lncRNAs potentially bound to MyoD in undifferentiated muscle cells (myoblasts).

At the same time, we looked at other groups of proteins potentially interacting with lncRNAs. We thought it likely that lncRNAs could be playing a role in transcriptional activation in addition to their previously known role in transcriptional repression and for this reason we investigated the potential of the proteins WDR5, MLL1, and Rbbp5 to immunoprecipitate lncRNAs. These proteins, in addition to Ash2L, form a complex known as COMPASS [25] in yeast and Trithorax in mouse and human (reviewed in [26]). All four of these proteins are required to catalyze the tri-methylation of histone

3 lysine 4 (H3K4me3), a marker of active transcription [27]. Their opposite is the polycomb repressive complexes (PRC1 and PRC2), which tri-methylate histone 3 lysine 27 (H3K27me3), a marker of transcriptional repression [28, 29, 30]. PRC2 is known to interact with the lncRNAs HOTAIR (*HOX* antisense intergenic RNA) [31], and ANRIL (antisense noncoding RNA in the INK4 locus) [32].

LncRNAs have also been shown to dramatically influence pluripotency and the pluripotent state [33]. Early microarray studies found lncRNAs that were differentially expressed during mouse embryonic stem cell (MESC) differentiation [34] and microarrays have also been used to discover lncRNAs that function in regulating reprogramming to pluripotency of somatic cells [35]. Other studies have shown that lncRNAs are necessary to maintain a state of pluripotency in stem or progenitor cells [36, 37]. In addition, data from chromatin immunoprecipitation (ChIP) experiments have shed light on a population of lncRNAs whose occupancy intersects with the known pluripotency factors Oct4 and Nanog [38]. To date, however, it has not been shown that Oct4 or Sox2 directly interact with lncRNAs.

Chapter 2

Methods

2.1 Cell Culture

C2C12 myoblast cells [39] were grown in Dulbecco's Modified Eagle Medium (DMEM/High Glucose, HyClone SH30243.01) containing 10% fetal bovine serum (FBS). Cells were passaged and expanded when they reached 80-90% confluency. This involved detaching the cells with trypsin and replating them in fresh media at a 1:20 dilution.

For harvesting, trypsin was again used to detach the cells. Cells being used for western blotting were left unfixed. Cells being used for ChIP or RNA-immunoprecipitation (RIP) were collected into 50ml conical tubes and incubated with 1% formaldehyde (paraformaldehyde, Sigma 30525-89-4) for 10 minutes while rotating at room temperature. 1.25M glycine was added to be 10% of the total solution to end the fixation process. The cells were collected by centrifugation at 1000rpm for 5 minutes at 4°C, washed with PBS (137mM NaCl, 2.7mM KCl, 4.3mM Na₂HPO₄, 1.47mM KH₂PO₄, and 10mM phosphate (pH 7.4)) and divided into 2x10⁷ or 1x10⁷ cell aliquots for ChIP or RIP, respectively. The cells were immediately frozen at -80°C.

2.2 Myoblast Differentiation

C2C12 myoblast cells were cultured as previously described. Cells were seeded at the density required to achieve 20% confluency the following day (Day 1). On Day 1, the cells were washed once with PBS and the media was replaced with DMEM containing

1% insulin transferrin selenium [40, 41] (ITS, Gibco 41400-045). The ITS media was changed on Day 3 and the cells harvested and fixed on Day 5.

2.3 Immunostaining

For immunostaining, C2C12 myoblasts or myoblasts treated with ITS were fixed for 10 minutes in 4% formaldehyde and permeabilized with a solution containing 0.05% Triton X-100 (Fisher Scientific BP151-500). All antibodies were diluted in blocking solution (9% FBS and 0.2% Tween 20 (Fisher Scientific BP337-500)). Permeabilized cells were washed with blocking solution and incubated with primary antibodies diluted 1:200 for 1 hour. The following primary antibodies were used: myosin heavy chain (MHC) (MF20, Developmental Studies Hybridoma Bank (DSHB)) and myogenin (F5D, DSHB). After incubation with primary antibodies, the cells were washed with blocking solution and incubated with secondary antibodies diluted 1:200 and Hoechst 33342 dye diluted 1:200 for 1 hour. The following secondary antibodies were used: Alexa 594 (Life Technologies A21207) and Alexa 488 (Life Technologies A11001) After incubation with secondary antibodies, the cells were washed twice with blocking solution and once with PBS. A Zeiss Axiovert 200m fluorescent microscope was used to visualize and photograph the cells.

2.4 Antibody Testing

We used sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-page) followed by western blotting to test our antibodies. 12% SDS gels were poured according to Laemmli [42]. When solidified, the gels were assembled into a gel tank (BioRad Mini Trans-Blot Cell) and immersed in running buffer containing 1% SDS, 1.92M glycine, and 0.25M Tris (Trizma base, Sigma T-6066). C2C12 cells were combined with sample dye containing Bromophenol blue (BioRad 161-0404) and β -mercaptoethanol, vortexed thoroughly and heated at 94°C. The samples were loaded into the wells and a prestained molecular protein marker (Benchmark, Invitrogen 10748-010) was loaded alongside the samples. The gels were run at 10mA/gel through the stacking phase and 20mA/gel through the separating phase.

After electrophoresis, the proteins in the gels were electrotransferred to Immobilon P membranes (Millipore) in a transfer buffer solution containing 15% methanol, 192mM glycine, 25mM Tris, and 0.1% SDS in a Mini-PROTEAN 3 system (Biorad) running at 150mA/tank overnight.

After overnight transfer the cassettes were disassembled and the gel with the membrane was placed on Saran wrap. A ballpoint pen was used to trace the edges of the gel and the molecular marker onto the membrane. The membrane was cut to the traced size and then into 10-12 strips for testing of different antibodies.

The membranes were incubated for 1 hour in a blocking solution containing 2% milk powder (blotting grade blocker non-fat dry milk, BioRad 170-6404) in PBT (0.1% Tween 20 (Fisher Scientific BP151-500) in PBS) on a lab shaker. Excess blocking solution was removed and the membranes were incubated with primary antibodies (Table 2.1) for 1 hour while shaking, washed 3 times for 4 minutes with PBT, and incubated with secondary antibodies (Table 2.2) for 1 hour while shaking. The membranes were washed 6 times for 4 minutes with PBT and incubated in SuperSignal West Dura Extended Duration Substrate (Thermo Scientific 34075) before being placed in a film cassette and exposed to x-ray films in the darkroom.

Antibody	Company	Catalog Number	Lot Number
MyoD	Millipore	MAB3878	JC1628178
MyoD	Santa Cruz	(N-19) sc-31940	I0706
MyoD	BD Pharmingen	554130	04882
MyoD	Santa Cruz	(M-318) sc-760	C0212
MyoD	Santa Cruz	(C-20) sc-302	C0812
MyoD	Santa Cruz	(C-20) sc-304	I0909
MyoD	Santa Cruz	(C-20) sc-304	J2111
MyoD	Santa Cruz	(C-20) sc-304	D0312
Myogenin	DSHB	F5D-MG	unknown
Myogenin	Millipore	MAB3876	1967332
Myf5	Santa Cruz	(C-20) sc-302	C0812
WDR5	R&D Systems	AF5810	CCZK0111111
WDR5	Bethyl Laboratories	A302-429A	A302-429A-1
WDR5	Bethyl Laboratories	A302-430A	A302-430A-1
Rbbp5	Bethyl Laboratories	A300-109A	A300-109A-2
MLL1	Active Motif	61295	17210001
MLL1	Bethyl Laboratories	A300-374A	A300-374A-1
MLL1	Bethyl Laboratories	A300-086A	A300-086A-1
MLL1	Millipore	ABE240	NRG1922437
Oct4	Santa Cruz	(N-19) sc-8628	A2412
Oct4	Santa Cruz	(H-34) sc-9081	L0210
Oct4	Santa Cruz	(H-34) sc-9081	E1011
Oct4	Abcam	ab19857	GR60398-1
Sox2	Millipore	CS204373	DAM1948375
Sox2	Santa Cruz	(Y-17) sc-17320	L0211
Sox2	Santa Cruz	(Y-17) sc-17320	A0312
Sox2	Millipore	CS207294	NRG1928895

Table 2.1: Primary antibodies used for western blotting. Antibodies were diluted at various concentration in blocking solution.

Antibody	Company	Catalog Number
Bovine-anti Goat IgG	Jackson ImmunoResearch Laboratories	085-035-180
Goat-anti Rabbit IgG	Jackson ImmunoResearch Laboratories	211-032-171
Goat-anti Mouse IgG	Jackson ImmunoResearch Laboratories	115-035-174

Table 2.2: Secondary antibodies used for western blotting. Antibodies were diluted at 1:1000 in blocking solution.

2.5 DNA Length Optimization for ChIP

We used agarose gel electrophoresis to test DNA length after sonication and micrococcal nuclease (MNase) incubation. 2×10^7 C2C12 cells fixed with 1% formaldehyde were resuspended in $498.5 \mu\text{l}$ of a cell lysis buffer solution containing 50mM Tris-HCl, 1mM CaCl_2 , and 2.5mM MgCl_2 . The protease inhibitors leupeptin (Sigma L2884, $1 \mu\text{g}/\text{ml}$), Pepstatin A (Sigma P5318, $1 \mu\text{g}/\text{ml}$ methanol), and PMSF (Sigma P7626, $17 \mu\text{g}/\text{ml}$ isopropanol) were added.

The cells were lysed by vortexing for 15 seconds, placing them on ice for 15 minutes, and vortexing for an additional 15 seconds every five minutes. 1000 gel units of MNase (NEB M02475) diluted in $9.5 \mu\text{l}$ of TE (1mM EDTA and 10mM Tris-HCl) was added and the samples were incubated at room temperature for 20 minutes. Addition of $10 \mu\text{l}$ of 0.5M EDTA and placement of the samples on ice stopped the enzymatic reaction. The samples were adjusted to contain 150mM NaCl, 1% NP-40 (MP Biomedicals 198596), 0.5% Na deoxycholate (MP Biomedicals 102906), and 0.1% SDS and sonicated with a Branson 450 sonicator to test the effect of varying the power, duty cycle, and number of pulses on DNA length.

The cell debris was pelleted by centrifugation at 13,000rpm for 15 minutes at 4°C . The supernatant containing the sheared chromatin was transferred to a new tube. $100 \mu\text{l}$ of an elution buffer containing 1% SDS and 0.1M NaHCO_3 with $0.1 \mu\text{g}/\mu\text{l}$ of proteinase K (Invitrogen 25530-015) was added and the samples were incubated for 2 hours at 65°C while rotating in a hybridization oven. Proteinase K was inactivated by incubation at 95°C .

The DNA was purified using a Zymo plasmid miniprep kit (Zymo Research D4020). $100 \mu\text{l}$ of a membrane binding solution containing 4.5M guanidine isothiocyanate and 0.5M potassium acetate was added to each sample and mixed until a cloudy, white precipitate formed. The mixture was then transferred to the column (Zymo Spin II C1008-250) and centrifuged at 12,000rpm for 1 minute at room temperature. $400 \mu\text{l}$ of Zymo wash buffer (Zymo Research D4036-4-48) was added followed by centrifugation at 12,000rpm for 1 minute. The flow-through was discarded and the samples were centrifuged for an additional 5 minutes to dry the filter. The column was transferred to a low retention recovery tube (Fisher Scientific 02-681-320) and $52 \mu\text{l}$ of Zyppey Elution

Buffer (1M Tris-HCl at 10mM and 0.5M EDTA at 0.1mM) was added to the column. After allowing this to incubate at room temperature for 1 minute, the samples were centrifuged at 12,000rpm for 2 minutes. This elution step was repeated once.

150-500ng of DNA from each sample was loaded into a 2% agarose gel alongside a 100bp (Invitrogen 15628-019) and 1kb ladder (Invitrogen 10787-018). The gel was run at 100V for 20-25 minutes in TAE (40mM Tris and 1mM EDTA adjusted to pH 8.3 with glacial acetic acid) on an EMBI Tec Electrophoresis Cell system (Model RunOne). The gels were shaken in ethidium bromide (BioRad 161-0433) for 15 minutes, rinsed once with de-ionized water and visualized with a BioRad gel documentation system.

2.6 ChIP

The same procedure as detailed above for cell lysis and MNase incubation was followed and sonication was performed using the optimized conditions: power 4, duty cycle 50%, for 45 pulses with a rest period of 30 seconds between every 9 pulses.

For the immunoprecipitation, 20 μ l of Dynabeads Protein G bead suspension (Invitrogen 1004D) per sample was washed twice with 100 μ l of RIPA buffer containing 150mM NaCl, 1% NP-40, 0.5% Na deoxycholate (MP Biomedicals 102906), 0.1% SDS, and 50mM Tris-HCl. A magnetic rack (MagnoRack, Invitrogen CS15000) was used to collect the beads. The 20 μ l of bead suspension was combined with 5 μ g of antibody (Table 2.3), 50 μ l of supernatant from the cell extract, and 450 μ l RIPA buffer. The preparation was thoroughly combined by pipetting and incubated overnight at 4°C with rotation.

The beads were washed for five minutes with rotation in 500 μ l of each of the following solutions: RIPA, LiCl wash buffer (1% Triton X-100, 0.1% SDS, 250mM LiCl, 0.2mM EDTA, and 20mM Tris-HCl), and TE.

The samples were decrosslinked and the DNA was purified using the methods described previously. Quantitative PCR (qPCR) was used to test for the presence or absence of DNA transcripts pulled down by our antibodies. Each reaction used 10 μ l of PCR mix (Promega GoTaq qPCR Master Mix 289548) combined with 3 μ l of DEPC treated water, 2 μ l of 2.5 μ M mixed forward and reverse primer (Table 2.3), and 5 μ l of purified DNA. The following PCR Program was used: 95°C for 2 minutes, 40 cycles

(95°C 30 seconds, 58°C 30 seconds, 72°C 30 seconds), 95°C 15 seconds, 60°C 15 seconds, 20 minute ramp up to 95°C, 95°C for 15 seconds, hold at 4°C (Eppendorf Mastercycler realplex2). Reactions were run in triplicate and averages were taken: $Input_{avg}$, IgG_{avg} , and X_{avg} .

Fold change was calculated as:

$$\begin{aligned} &= 2^{-(X_{avg} - IgG_{avg}) - Input_{avg}} \\ &= 2^{(-\Delta C_T - Input_{avg})} \\ &= 2^{(-\Delta\Delta C_T)} \end{aligned}$$

Antibody	Company	Catalog Number	Lot Number
MyoD	Santa Cruz	(C-20) sc-304	I0909
MyoD	Santa Cruz	(C-20) sc-304	J2111
Rabbit IgG	Santa Cruz	sc-2027	C2712
H3K27me3	Upstate	17-622	24440
MyoD	Santa Cruz	(C-20) sc-304	D0312
MyoD	Millipore	MAB3878	JC1628178

Table 2.3: Antibodies used for ChIP experiments.

2.7 RNA-ChIP

The same procedure as previously described was followed for MNase incubation, sonication, and immunoprecipitation. The only differences were that a larger number of C2C12 cells, 2×10^8 , was used as the starting material and $10 \mu\text{l}$ of RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen 10000840) and $10 \mu\text{g}/\text{ml}$ Heparin was added along with the protease inhibitors.

The sample was decrosslinked using reagents from a PureLink RNA Mini Kit (Life Technologies 12183-016). $100 \mu\text{l}$ of lysis buffer (Life Technologies 46-6001) was added and the sample was incubated at 80°C for 1 minute, vortexed briefly, and $0.1 \mu\text{g}/\mu\text{l}$ of proteinase K (Invitrogen 25530-015) was added. The sample was incubated at 56°C for 15 minutes then at 80°C for 15 minutes. Finally, the sample was placed on ice for 1 minute and allowed to come to room temperature.

Name	Sequence
MyoD CER F1	GGG CAT TTA TGG GTC TTC CT
MyoD CER R1	CTC ATG CCT GGT GTT TAG GG
MyoD DRR F1	TCA GGA CCA GGA CCA TGT CT
MyoD DRR R1	CTG GAC CTG TGG CCT CTT AC
Myogenin E2/E1 F1	GAA TCA CAT GTA ATC CAC CTG GA
Myogenin E2/E1 R1	ACA CCA ACT GCT GGG TGC CA
B-actin F3	TTC GCG GGC GAC GAT GCG
B-actin R1	TTC TGA CCC ATT CCC ACC ATC ACA
Oct4 F3	AGG TCA AGG GGC TAG AGG GTG GGA TT
Oct4 R3	TGA GAA GGC GAA GTC TGA AGC CA
Sox2 F11	GCC GGA AAC CCA TTT ATT CCC TGA
Sox2 R11	TCG GGC TCC AAA CTT CTC TCC TTT
MyoD CER ChIP F2	AGC CAG TTA ATC TCC CAG AGT GCT
MyoD CER ChIP R2	TAG AGA AAC CGG AGA AGA CCC AGG AA
MyoD DRR ChIP F2	AAA GTA AGA GGC CAC AGG TCC AGA
MyoD DRR ChIP R2	TCT GGA AAC CGG ATC CAA CTA GCA

Table 2.4: Sequences of primers used for qPCR.

A DNase step was added from the PureLink DNase set (Invitrogen 46-6026) that used a DNase I mixture of 0.09M MnCl₂, 7μl of 2X DNase Buffer (Invitrogen 46-6025), and 10μl of PureLink DNase I (Invitrogen 10002884). After this mixture was added to the sample it was incubated at room temperature for 15 minutes. The beads were removed and 325μl of bead lysis buffer and 200μl of isopropanol was added to the sample and mixed by vortexing briefly. The sample was transferred to a spin column and centrifuged at 10,000rpm for 30 seconds at room temperature. 500μl of wash solution (Invitrogen 46-6003 with ethanol added) was added to the column followed by centrifugation at 10,000rpm for 30 seconds at room temperature. This was repeated once. 30μl of RNase-free water (Invitrogen 46-8000) was added to the column and it was incubated at room temperature for 1 minute followed by centrifugation at 13,000rpm for 2 minutes. The RNA concentration was measured using a Qubit fluorometer (Life Technologies).

2.8 RNA-sequencing and Data Analysis

My colleague prepared cDNA libraries and the BioMedical Genomics Center (BMGC) obtained raw sequence data from the libraries. I then analyzed the sequences in collaboration with the Minnesota Supercomputing Institute (MSI). In preparation for sequencing, the Ovation RNA-seq System V2 (NuGEN 7102) was used to create cDNA from the co-precipitated RNA fragments. The cDNA was sent to BMGC for fragmentation by sonication and size estimation using a DNA High Sensitivity Lab Chip (Agilent 5067-4626). The cDNA was sent back to our lab and the Ovation Ultraflow Library System (NuGEN 0303) was used to generate blunt ends by end repair so that adaptor and barcode sequences could be ligated to the cDNA fragments. The cDNA fragments that contained adaptor sequences at both ends were amplified by PCR to create the final cDNA library. AgencourtRNAClean XP Beads were used to further purify the cDNA from ribosomal RNA and small RNA fragments.

Two rounds of samples were sent. The first contained cDNA from RNA-ChIP experiments using CGR (mouse embryonic stem) cells with antibodies against WDR5, MLL1, and Rbbp5, and an IgG control and the second contained cDNA from RNA-ChIP experiments with antibodies against Sox2 (Millipore), Sox2 (Santa Cruz), Oct4 (Santa Cruz), Oct4 (Abcam), a repeat of WDR5, and an IgG control.

Paired-end sequencing using 50-base-pair reads and 200-base-pair fragments was performed at the BMGC on an Illumina HiSeq 2000. Our cDNA library was washed across a flow cell which binds the adaptor sequences. The cDNA that hybridized to the flow cell underwent bridge amplification to form clusters of cDNA clones. Sequencing primer was added and DNA bases were added one at a time. Each cycle produced a base read for each cluster and the flow cell was imaged after the addition of each base. The bases were labeled with different fluorophores and it was the reading of this fluorescence that produced the sequence information.

The output of this sequence information was in the format of several FastQ files, which were uploaded by BMGC into our project space at MSI. A FastQ file contains the raw reads data. It is given a Phred score, which gives the probability of the accuracy of the base calling. A Phred score of 30 is considered acceptable though, generally, even with a high Phred score it is necessary to use Quality Trimmer or Column Trimmer

to trim reads that fall below the accepted level. The reads can then be mapped to a reference genome, in our case the UCSC mouse mm9 genome build, using a software program developed specifically for RNA-seq data analysis called Bowtie. Another program called TopHat works together with Bowtie to identify exon-exon splice junctions. The mapping process produces two output files. The first is a SAM file, a tab-delimited text file that contains sequence alignment data and the other is a BAM file, the binary version of the SAM file. It is the BAM file that can be opened in the Interactive Genome Browser (Broad Institute) to visualize the mapping of the reads which form peaks when many reads map to one region of the genome. A program called MACS (Model based Analysis of ChIP-seq) is then used to ‘call’ the peaks, or in more general terms, to assign statistical significance to peaks based on their width and height above background (IgG) levels. Additional significance can be attached to certain peaks that are shared between samples known to form a complex (WDR5, MLL1, Rbbp5), samples that are known to bind the same region of the DNA (Oct4, Sox2), or between samples from antibodies against the same protein but from different companies (Sox2 Millipore, Sox2 Santa Cruz).

Chapter 3

Results

3.1 ITS Differentiates Myoblasts into Myotubes

A protocol was developed for the large-scale production of myotubes via the differentiation of C2C12 myoblast cells. We determined early on that insulin transferrin selenium (ITS) [40, 41] was more effective than horse serum (HS) [43, 44] at producing myotubes from myoblasts. Therefore, we chose ITS to develop the large-scale protocol.

Myotube formation was greatest when ITS was added to myoblasts at a confluency of 20%. The number of myoblasts seeded per dish on Day 0 to obtain 20% confluency on Day 1 when ITS was added was determined as 3×10^4 cells/3.5cm dish, 3×10^5 cells/10cm dish, and 2×10^6 cells/15cm dish. On Day 2, the cells became 40% confluent with no perceptible changes. On Day 3, the cells became 60-70% confluent with some elongation and circular patterning of the cells. By Day 4, this elongation and patterning increased and cells reached 80-90% confluency with many dead cells observed. On Day 5, confluency decreased to 50-60% with a further increase in dead cells. At this point, obvious elongation and bundling of the cells into multi-nucleated myotubes was observed. This method typically produced about 3.6×10^6 cells/10cm dish and 7.5×10^6 cells/15cm dish.

Successful myotube formation was assessed by the activation of the myotube-specific genes myosin heavy chain (MHC) and myogenin. Immunofluorescence microscopy confirmed the expression of MHC in 47% of cells treated with ITS (Figure 3.1) and myogenin in 33% of cells treated with ITS (Figure 3.2), compared to 0% expression of MHC or myogenin in untreated myoblast cells.

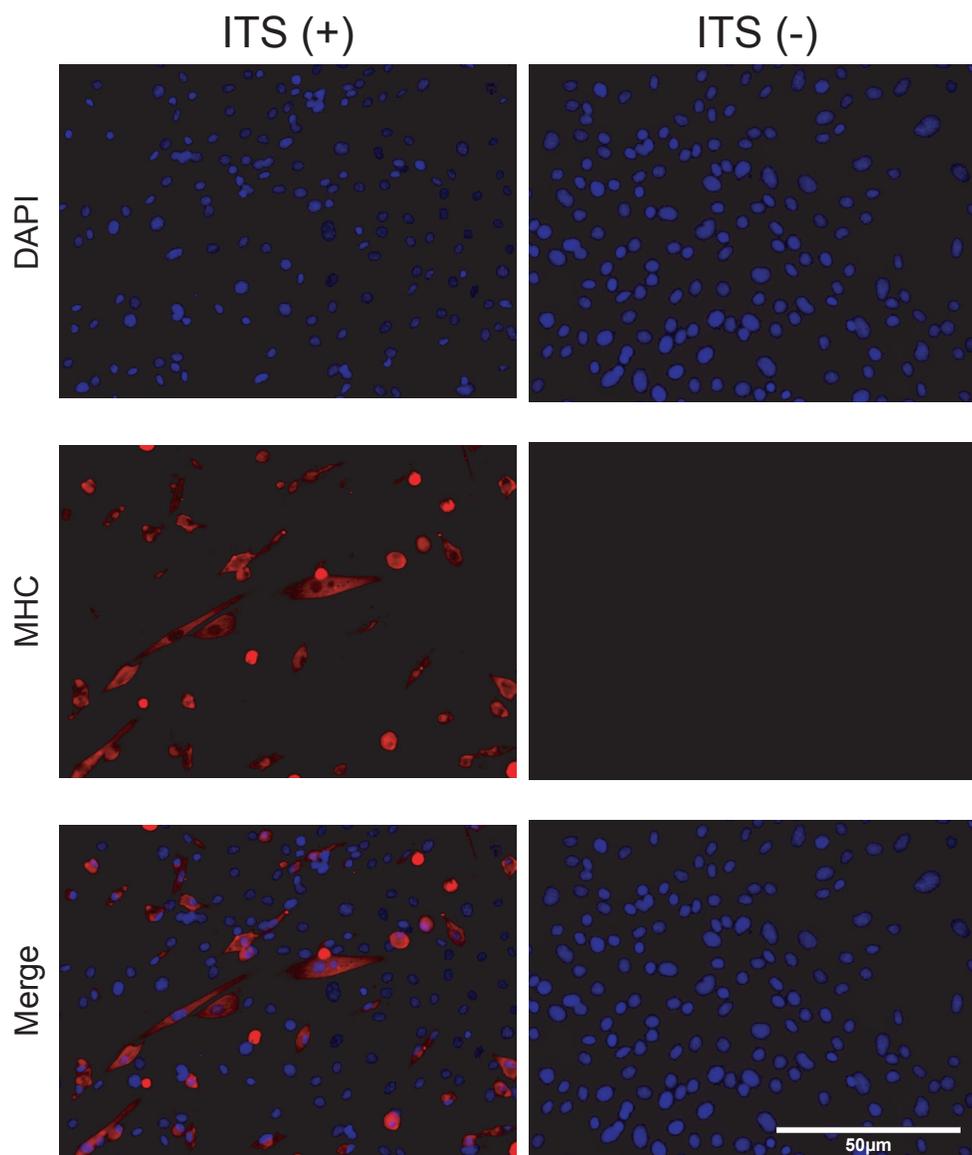


Figure 3.1: MHC expression in myotubes from differentiated C2C12 cells. C2C12 cells were treated with or without ITS and immunostained with antibodies against MHC after fixation with formaldehyde on Day 5. DNA was counterstained with DAPI. Cells were visualized at 20X with a Zeiss Axiovert 200M fluorescent microscope.

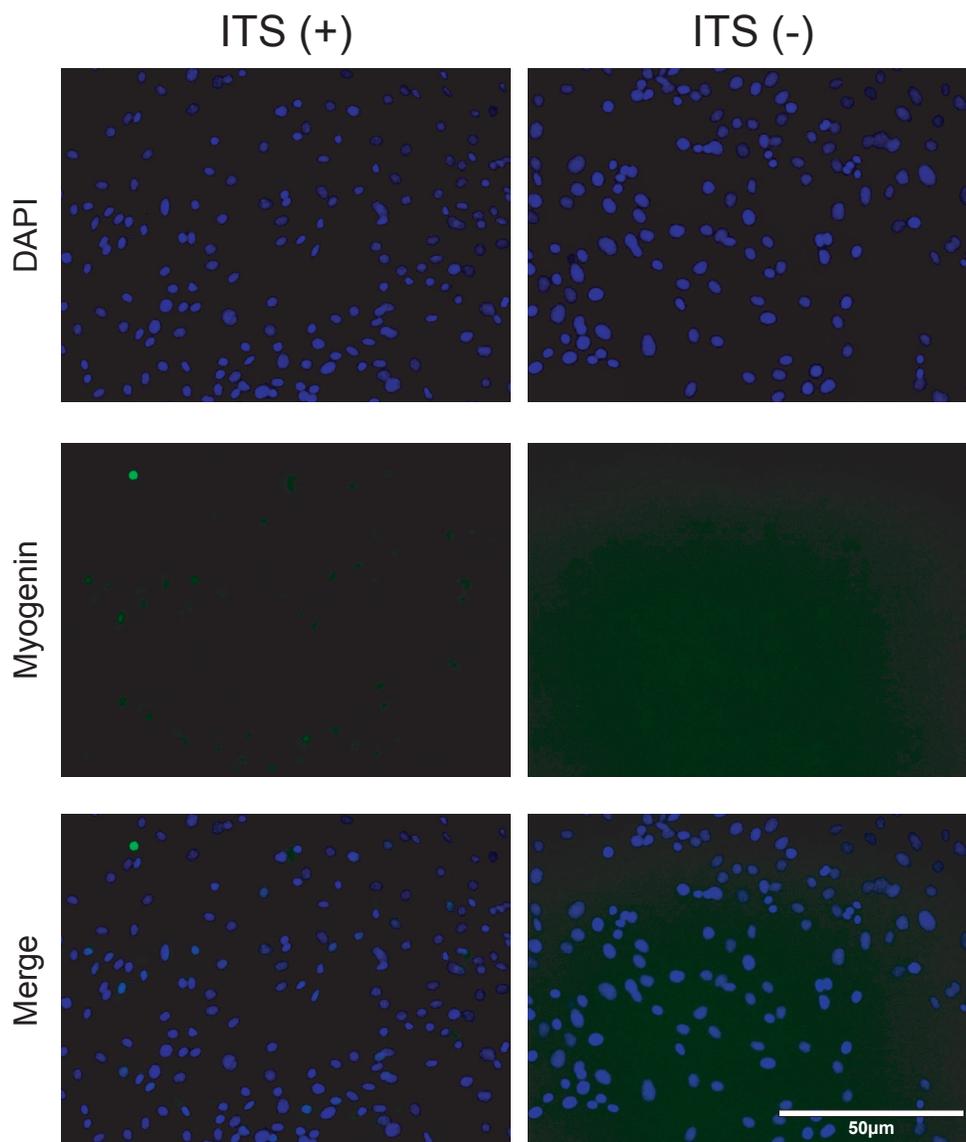


Figure 3.2: Myogenin expression in myotubes from differentiated C2C12 cells. C2C12 cells were treated with or without ITS and immunostained with antibodies against myogenin after fixation with formaldehyde on Day 5. DNA was counterstained with DAPI. Cells were visualized at 20X with a Zeiss Axiovert 200M fluorescent microscope.

3.2 Western Blots Identify Immunoprecipitation-compatible Antibodies

Immunoprecipitation (IP) requires that the antibodies used have high affinity and specificity for the protein being pulled down. Western blots that detect a single band of the expected size indicate that the antibody is appropriate for use in immunoprecipitation. Using this technique, we found several IP-compatible antibodies against muscle-specific proteins, components of the COMPASS complex, and pluripotency proteins for use in our ChIP and RNA-ChIP experiments.

We evaluated eight MyoD antibodies and found six that detected multiple bands or had high background (Millipore MAB3878: lot JC1628178, Santa Cruz sc-31940: lot I0706, BD Pharmingen 554130: lot 04882, Santa Cruz sc-760: lot C0212, Santa Cruz sc-302: lot C0812, Santa Cruz sc-304: lot D0312) and two that detected a single band at 45kDa (Santa Cruz sc-304: lot I0909 and J2111) (Figure 3.3). Testing of two myogenin antibodies found that both detected multiple bands (F5D-MG (Developmental Studies Hybridoma Bank (DSHB), Millipore MAB3876: lot 1967332). The one Myf5 antibody tested (Santa Cruz sc-302: lot C0812) detected multiple bands as well.

For the COMPASS complex, we tested three WDR5 antibodies and found two of them (R&D Systems CCZK0111111: lot AF5810, Bethyl Laboratories A302-429A: lot A302-429A-1) detected multiple bands and one of them (Bethyl Laboratories A302-430A: lot A302-430A-1) showed a strong dominant band at 40kDa. Testing of four Mll1 antibodies revealed three that detected multiple bands (Active Motif 61295: lot 17210001, Bethyl Laboratories A300-374A: lot A300-374A-1 and A300-086A-1) and one that detected a single band at 180kDa (Millipore ABE240: lot NRG1922437). Testing of one Rbbp5 antibody (Bethyl Laboratories A300-109A: lot A300-109A-2) detected a single band at 60kDa (Figure 3.4).

Of the four Oct4 antibodies tested one of them (Santa Cruz sc-8628: lot A2412) detected multiple bands and three of them (Santa Cruz sc-9081: lot L0210 and E1011, Abcam ab19857: lot GR60398-1) detected a single band at 34kDa. Testing of four Sox2 antibodies found one that detected a band of the incorrect size (Millipore CS204373: lot DAM1948375) and three that detected a band at 34kDa (Santa Cruz sc-17320: lot L0211 and A0312, Millipore CS207294: lot NRG1928895) (Figure 3.5 and 3.6).

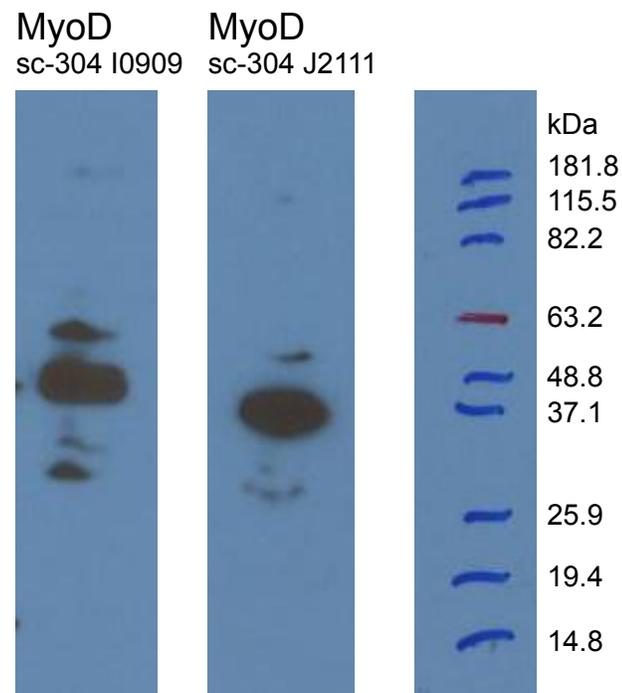


Figure 3.3: Western blots of IP-compatible antibodies against muscle-specific proteins. All antibodies were diluted in blocking solution. Primary antibodies: MyoD sc-204: lot I0909 diluted 1/2000, MyoD sc-304: lot J2111 diluted 1/2000. Secondary antibody: Rabbit IgG diluted 1/1000. Exposure time = 1 minute. Expected size = 45kDa. Due to the nature of our western blotting method, where the membranes are cut into strips instead of left intact, shifting of the strips can cause variation in the orientation of the molecular marker to the band of interest.

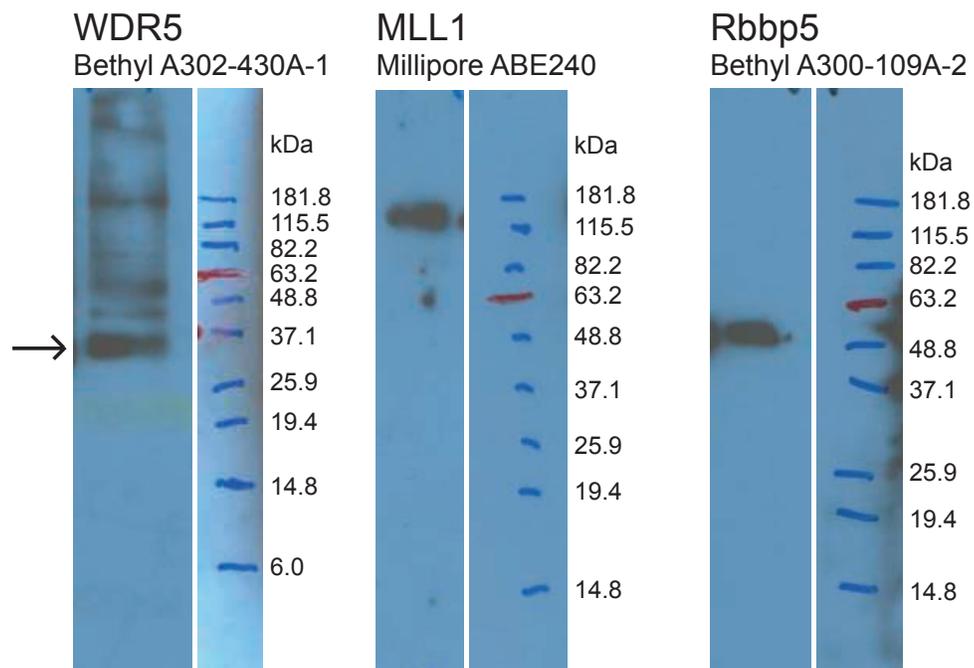


Figure 3.4: Western blots of IP-compatible antibodies against components of the COM-PASS complex. All antibodies were diluted in blocking solution. Primary antibodies: WDR5 Bethyl Laboratories A302-430A lot: A302-430A-1 diluted 1/20,000, MLL1 Millipore ABE240 lot: NRG1922437 diluted 1/600, Rbbp5 Bethyl Laboratories A300-109A lot: A300-109A-2 diluted 1/6400. Secondary antibody: Rabbit IgG diluted 1/1000. Exposure time WDR5, Rbbp5 = 1 second, MLL1 = 1 minute. Expected size = 40kDa WDR5, 180kDa MLL1, 60kDa Rbbp5.

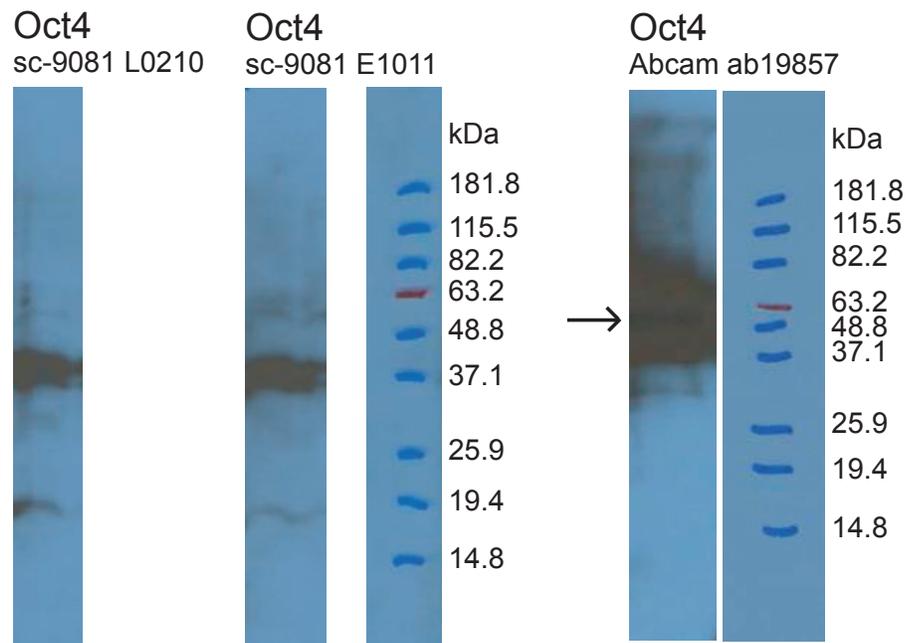


Figure 3.5: Western blots of IP-compatible antibodies against Oct4. All antibodies were diluted in blocking solution. Primary antibodies: Santa Cruz sc-9081 lot: L0210 and E1011 diluted 1/4000, Abcam ab19857 lot: GR60398-1 diluted 1/6400. Secondary antibody: Rabbit IgG diluted 1/1000. Exposure time = 1 second. Expected size = 34kDa.

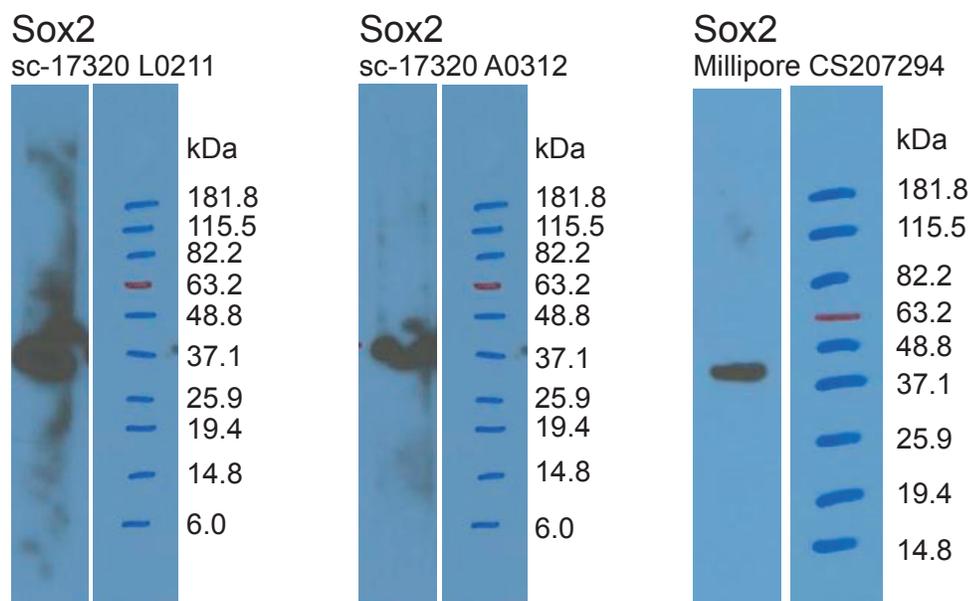


Figure 3.6: Western blots of IP-compatible antibodies against Sox2. All antibodies were diluted in blocking solution. Primary antibodies: Santa Cruz sc-17320 lot: L0211 and A0312 diluted 1/400, Millipore CS207294: lot NRG1928895 diluted 1/6400. Secondary antibodies: Goat IgG diluted 1/1000 for Santa Cruz sc-17320 lot: L0211 and A0312, Mouse IgG diluted 1/1000 for Millipore CS207294: lot NRG1928895. Exposure time = 1 second. Expected size = 34kDa.

3.3 Chromatin Size is Optimized for Immunoprecipitation

For the purpose of ChIP or RNA-ChIP, the DNA needs to be sheared into fragments between 200 and 700 base pairs (bp) in length, centered at 500bp. We established and tested a method that did this using a combination of sonication and incubation with micrococcal nuclease.

At duty cycle 50%, power 4, increasing the number of pulses to 108 or the power level to 6 with our sonicator (Branson 450) had a negligible effect on DNA length and size distribution, producing DNA in lengths ranging from 0.1kb to 12kb (Figure 3.7). Increasing the total number of pulses to 144 at duty cycle 50%, power 4, had a noticeable effect on DNA length, producing DNA between 0.1kb and 3kb in length (Figure 3.8). However, sonication for this number of pulses was labor-intensive and resulted in loss of greater than 20% of the sample volume during the sonication process. Attempts to raise the duty cycle from 50% to 100% resulted in emulsification and degradation of the sample. It was only with the addition of a 20 minute micrococcal nuclease incubation step that a dramatic decrease in DNA length was observed, producing DNA between 0.3kb and 1kb in length, centered at 0.5kb at both 25°C and 37°C (Figure 3.9).

As a result, we added a 20 minute incubation at 25°C with micrococcal nuclease to our preparation of DNA for ChIP followed by sonication at power 4, duty cycle 50% for 45 pulses with a rest period of 30 seconds between every 9 pulses of sonication delivered.

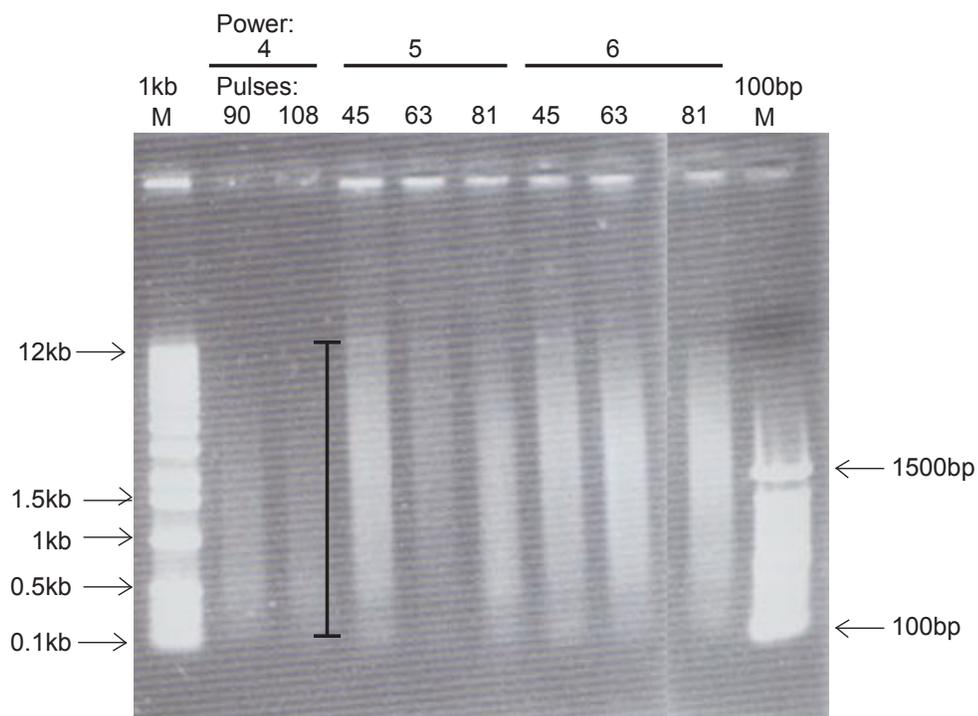


Figure 3.7: Size distribution of DNA after sonication of chromatin ranges from 0.1kb-12kb. Measured range is indicated by bar. Power settings and number of pulses delivered are indicated. Size markers, 1kb and 100bp, are applied to the left and right sides of the gel, respectively.

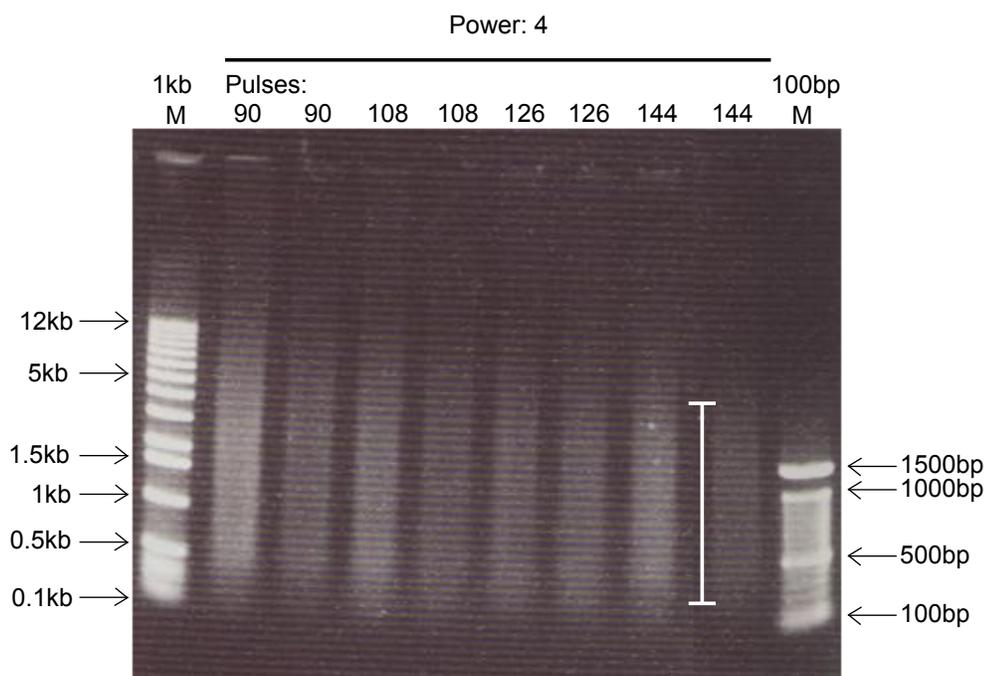


Figure 3.8: Size distribution of DNA after sonication of chromatin ranges from 0.1kb-3kb. Measured range is indicated by bar. Power settings and number of pulses delivered are indicated. Size markers, 1kb and 100bp, are applied to the left and right sides of the gel, respectively.

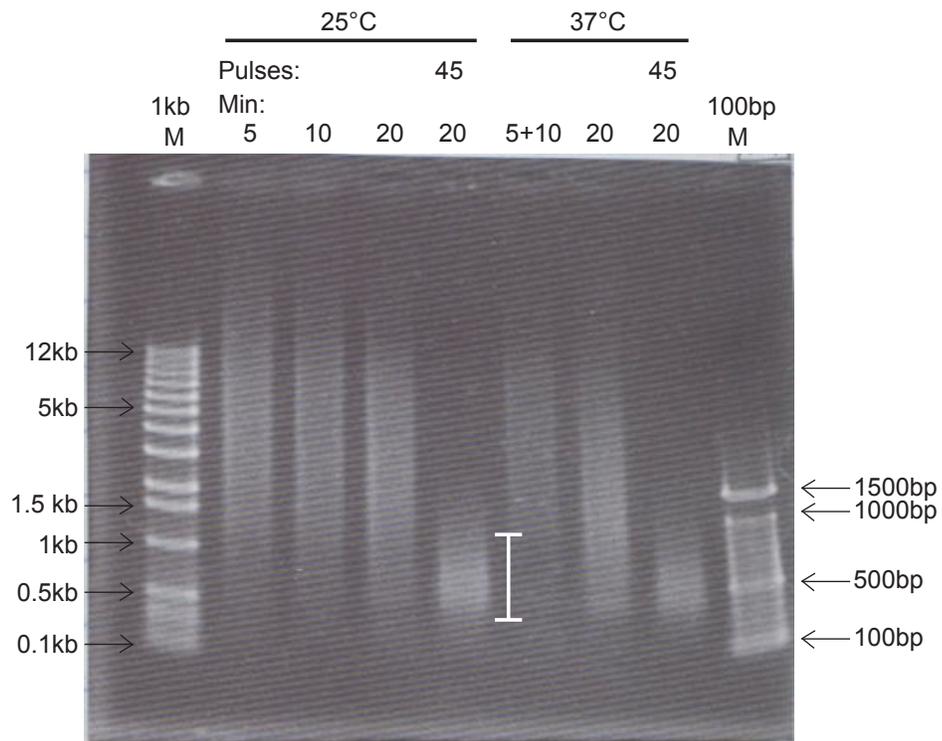


Figure 3.9: Size distribution of DNA after incubation with micrococcal nuclease and 45 pulses of sonication of chromatin ranges from 0.3kb-1kb. Measured range is indicated by bar. Temperatures, power settings, and number of pulses delivered are indicated. Size markers, 1kb and 100bp, are applied to the left and right sides of the gel, respectively.

3.4 ChIP with MyoD Shows Non-specific Binding

Even with optimally sized chromatin and highly specific antibodies, IP experiments can still fail to pull down RNA or DNA previously known to be associated with the protein of interest. To confirm that our MyoD antibodies were IP-compatible, we designed qPCR primers targeting regions where MyoD is known to bind. These included regions upstream of the MyoD gene itself: the proximal regulatory region (PRR), the distal regulatory region (DRR), and the core enhancer region (CER) [45] (Figure 3.10), in addition to the E2/E1 region upstream of the transcription start site (TSS) of the muscle-specific gene, myogenin. Testing of these primers by qPCR and MOPS (3-(N-morpholino)propanesulfonic acid) denaturing gradient gel electrophoresis according to Lerman [46], revealed that the MyoD CER, MyoD DRR, and myogenin E2/E1 primers were compatible with qPCR.

Positive control for our ChIP protocol used antibodies against H3K27me3 (Upstate 17-622 lot: 24440). H3K27me3 is a marker of transcriptional repression. It binds near the TSS of repressed genes which in myoblasts includes the genomic regions of the Oct4 and Sox2 genes targeted by our Oct4 F3/R3 and Sox2 F11/R11 primers. qPCR indicated that a high amount of DNA was precipitated by H3K27me3 at the Oct4 F3/R3 and Sox2 F11/R11 regions, ranging from 66.87-110.92 (Figure 3.11).

ChIP with two MyoD antibodies (sc-304 lot I0909 and J2111) did not precipitate greater amounts of DNA at the regulatory regions of MyoD and myogenin compared to the control IgG (sc-2027 lot: C2712). The recovered DNA amount for MyoD and myogenin genes ranged from 0.61-1.5 and from 0.34-3.28 for the negative controls (Figure 3.11).

Overall, these results indicated that we could not detect MyoD binding the regulatory regions of MyoD and myogenin genes as reported in the literature [45]. It is highly likely this was due to the low quality of the antibodies, as our positive and negative controls precipitated DNA at the expected amounts.

Primers for the PRR, DRR, and CER regions of MyoD were redesigned and ordered in addition to new primers overlapping the TSS of MyoD and myogenin. Testing revealed that the new MyoD CER primer (MyoD CER F2/R2), the new MyoD DRR primer (MyoD DRR F2/R2), and a primer spanning the TSS of MyoD (CD1) were

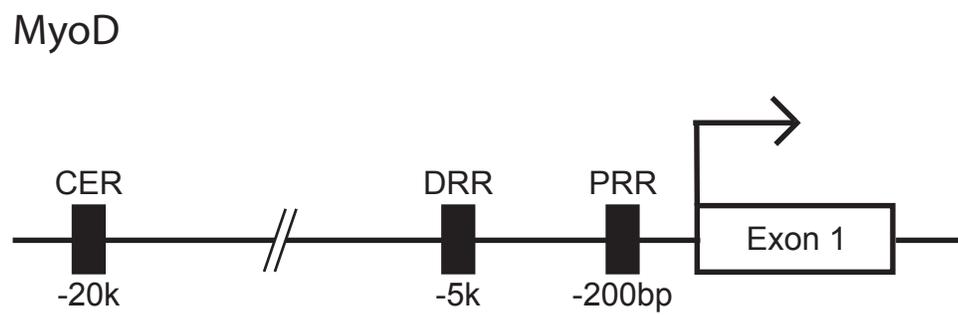
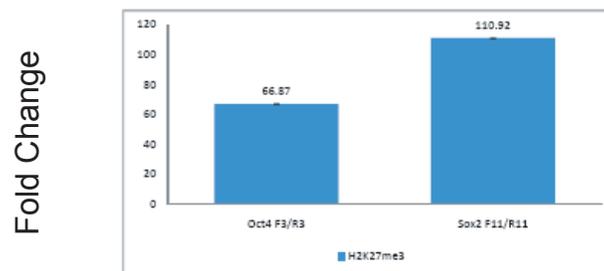


Figure 3.10: Regulatory regions of the MyoD gene. CER = core enhancer region, DRR = distal regulatory region, PRR = proximal regulatory region.

compatible with qPCR. In addition, MyoD sc-304 was reordered (lot:D0312) and ChIP with this antibody and also with MyoD (Millipore MAB3878 lot: JC1628178) was performed. However, qPCR revealed that these MyoD antibodies were also not binding the regulatory regions of MyoD and myogenin genes at a higher level than the control IgG. Fold change for MyoD sc-304 (lot:D0312) ranged from 2.14-3.4 compared to 1.97 for the negative control. Similarly, fold change for MyoD MAB3878 ranged from 0.19-0.51 compared to 0.31 for the negative control, indicating non-specific binding for both MyoD antibodies. Indeed, both MyoD antibodies (MyoD sc-304 lot:D0312 and Millipore MAB3878) showed multiple bands by western blot. As a result of the lack of IP-compatible antibodies against MyoD, we were obligated to suspend the MyoD project.

ChIP 1



ChIP 2

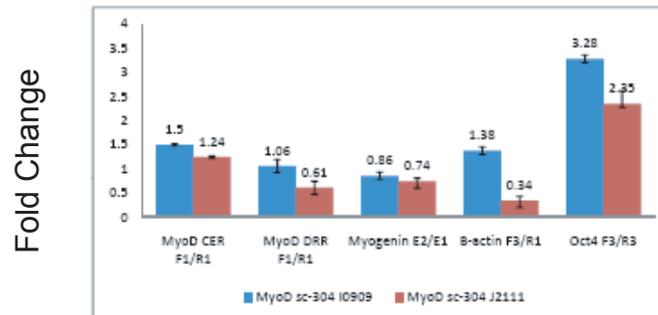


Figure 3.11: Relative expression fold change of DNA from ChIP using antibodies against H3K27me3 (top) and MyoD sc-304 (I0909 and J2111) (bottom), as determined by qPCR. Three technical replicates were used for each antibody. Amount of DNA co-precipitated with IgG was defined as 1.0.

ChIP 3

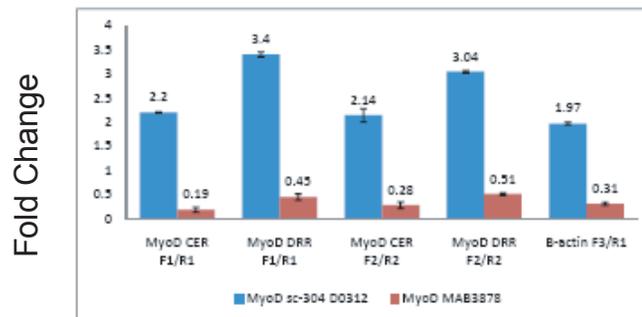


Figure 3.12: Relative expression fold change of DNA from ChIP using antibodies against MyoD sc-304 (lot:D0312) and MyoD MAB3878 as determined by qPCR. Three technical replicates were used for each antibody. Amount of DNA co-precipitated with IgG was defined as 1.0.

3.5 RNA-ChIP with MyoD Leads to Re-testing of Antibodies

The amount of RNA purified from the RNA-ChIP experiments using antibodies against MyoD (sc-304 lot: I0909 and J2111) was too low (<20ng/ml) to be detected by a Qubit fluorometer (Life Technologies). RNA-ChIP with antibodies against Ezh2 and Suz12, members of the polycomb repressive complex 2 (PRC2) previously known to bind lncRNAs, were used as a positive control; however, they also did not recover sufficient RNA to be measured. Re-testing by western blot of the two MyoD antibodies (sc-304 lot: I0909 and J2111) indicated that they had become inactive. Attempts to test other MyoD antibodies by western blot showed multiple dominant bands or high background (Millipore MAB3878 and Santa Cruz sc-760). Re-ordering of MyoD (sc-304 lot: D0312) showed an additional dominant band by western blot. Because of the low specificity of the available antibodies, we were not able to use MyoD antibodies for RNA-ChIP. However, we found highly specific antibodies against WDR5, MLL1, Rbbp5, Oct4, and Sox2. My colleague obtained sufficient amount of RNA (>100ng) with each of these antibodies using the RNA-ChIP protocol we optimized together.

3.6 RNA-seq Data Analysis

Phred scores for the WDR5, MLL1, Rbbp5 and IgG samples were greater than 30, with only one or two positions requiring trimming with Column Trimmer. FastQC was used to check the quality of the samples and adaptor contamination was listed as an overrepresented sequence (ORS) for the Rbbp5 sample. CutAdapt was used to remove the contaminating adaptor sequences from this sample. Ribosomal RNA was listed as an ORS for all 4 of the samples. As a consequence, mapping the reads to the genome using Bowtie in conjunction with Tophat yielded low mapping percentages (Table 3.1). Rescuing reads by allowing a 1bp mismatch for the barcode sequences rescued 5-10% of the unmappable MLL1, Rbbp5, and IgG reads. Unexpectedly, rescuing the reads reduced the mapping percentage of WDR5 to 4% and this sample could not be used for further analysis.

Sample	Mapping %
MLL1	30%
WDR5	54%
Rbbp5	46%
IgG	34%

Table 3.1: Mapping percentages of RNA-seq 1 samples.

After mapping, the BAM files were uploaded to IGV for manual inspection. MACS was used to call the peaks from the MLL1, Rbbp5, and IgG samples. After IgG peaks were subtracted, 6,794 MLL1 peaks and 11,764 Rbbp5 peaks remained and 803 of these peaks overlapped. Representative examples of peaks viewed with IGV are shown in Figure 3.13.

The low mapping efficiency of the first set of samples was attributed to degradation of components of the Ovation RNA-seq System V2. The second set of samples showed higher mapping percentages (75-90%) and a greater number of peaks called per sample (36,871-50,653). Importantly, 92 loci (peaks) were co-precipitated by all 4 pluripotency antibodies (the Oct4 and Sox2 duplicates), generating a robust list of candidate lncRNAs for further analysis.

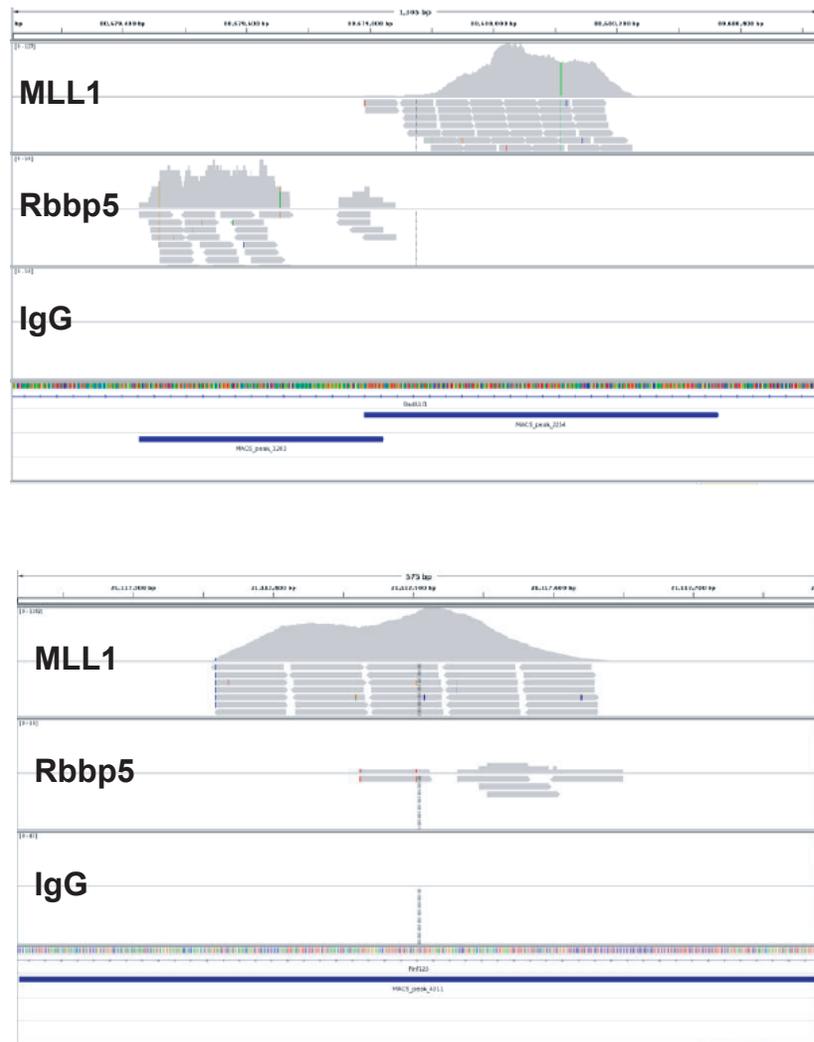


Figure 3.13: Representative example of peaks viewed with IGV. Top: A 1,305 base pair region (chr12:80,679,221-80,680,526) showing a peak shared by MLL1 and Rbbp5 but not IgG. Bottom: A 507 base pair region (chr18:54,082,038-54,082,648) showing an MLL1-specific peak not shared by Rbbp5 or IgG. The height and shape of a peak is determined by the number of reads (square arrows below the peaks) mapping at that location. DNA base pairs are depicted in unique colors and location of the peaks relative to known genes is shown (exon = solid blue bar, intron = hashed blue line).

Chapter 4

Discussion

While establishing a protocol for RNA-ChIP, we immunoprecipitated three sets of proteins: MyoD, the components of the COMPASS complex WDR5, MLL1 and Rbbp5, and two key pluripotency proteins Oct4 and Sox2.

Whether lncRNAs interact with MyoD, the master regulator of muscle differentiation, remains an elusive question. Muscle differentiation is a well-defined system that could be taken advantage of to characterize lncRNAs expressed during progressive stages of muscle development. RNA-ChIP with antibodies against MyoD followed by RNA-seq persists as a valid pursuit should IP-compatible antibodies against MyoD become available.

As for the COMPASS complex, MLL1 contains an RNA binding domain [8]. This makes it and the other proteins of the COMPASS complex likely candidates for interacting with lncRNAs. Generating new WDR5, MLL1, and Rbbp5 samples for sequencing using the replaced Ovation RNA-seq System V2 is highly likely to improve mapping results and in this way overlapping peaks from all three samples could be used to generate a list of candidate lncRNAs.

The high mapping percentage and large number of overlapping peaks co-precipitated by our Oct4 and Sox2 duplicates with the second round of RNA-ChIP indicates that our protocol is indeed effective. This data could lead to an exciting new chapter in the regulation of pluripotency as it has not been previously shown that Oct4 and Sox2 bind RNA. Oct4 and Sox2 are known to form a heterodimer and bind a cis-regulatory element essential for the activation of a third master pluripotency factor, Nanog [47].

These three proteins in turn bind to closely localized genomic sites [48], upregulating genes important for pluripotency and downregulating lineage specific genes. It is conceivable that lncRNAs are playing a role by acting as scaffolds for the assembly of the three pluripotency factors on the chromatin or as guides to recruit these factors to regulatory regions. Data from RNA-ChIP with antibodies against Nanog could be used to augment previously existing Oct4 and Sox2 peak data and assist in the selection of candidate lncRNAs.

Future work with this project involves filtering the list of candidate lncRNAs produced by the RNA-seq data analysis. Subtraction of the regions occupied by previously annotated lncRNAs and protein coding genes can generate a list of candidate novel lncRNAs. These novel lncRNAs can be further scrutinized to verify that they are not transcriptional noise and that they indeed do not encode proteins. For instance, if the candidate is located within a K4-K36 domain and enriched with RNA polymerase II binding sites and DNase I hypersensitivity sites (a sign of open chromatin) as detected with the ENCODE data, the candidate is likely to be a product of active transcription [37, 49, 50]. The protein-coding potential of a candidate lncRNA can be evaluated with the Coding Potential Calculator (CPC) algorithm and other programs [51, 52].

Back at the bench, characterization of lncRNA function typically involves rapid amplification of cDNA ends (RACE) to identify the full length transcript [53]. Knockdown and overexpression of the novel lncRNA can further validate its biological function in a system of interest.

In summary, we established a RNA-ChIP protocol to identify lncRNAs bound to chromatin proteins in embryonic stem cells. Although RNA-ChIP has been previously used to identify lncRNAs that bind to RNA-binding proteins, the novelty of our approach is that it has been applied to proteins that are not previously known to bind RNA. Furthermore, while the ENCODE and FANTOM projects have identified several thousand lncRNAs from human and mouse cells, these groups used a defined number of cells in a very defined situation. The power of this technique is that it can be applied to very specific contexts to detect lncRNAs potentially binding proteins of interest and also that it allows for the discovery of *de novo* lncRNAs .

Chapter 5

Conclusion

Currently, interactions between proteins, microRNAs, and specific regions of the DNA are the main concepts applied to studies of gene regulation. Our approach opens the door to a novel layer of gene regulation that incorporates RNA-protein interactions through the isolation and identification of lncRNAs bound to specific proteins of interest. The proof-of-principle technique established by this project is a useful tool to characterize lncRNA expression during any developmental stage and is widely applicable to the study of chromatin binding proteins in other biological contexts. Furthermore, we expect that additional technological innovations geared toward studying lncRNAs will continuously emerge to support the rapid development of this fascinating research field.

Chapter 6

Glossary of Bioinformatics Terms

- **Fragment** – A cDNA piece 200 base pairs in length generated by sonication and reverse transcription.
- **Flow cell** – A planar optically transparent surface similar to a microscope slide which contains a lawn of oligonucleotide anchors bound to its surface.
- **Adaptor sequences** – Sequences ligated to the ends of fragments that attach to the oligonucleotide anchors bound to the flow cell and simultaneously provide primers during bridge amplification.
- **Multiplexing** – The sequencing of multiple samples on one lane of a flow cell.
- **Barcode** – A unique sequence used to distinguish samples during multiplexing.
- **Read** – A 50 base pair sequence read from the end of a fragment bound to a flow cell.
- **Paired-end sequencing** – Reading of 50 base pairs from both ends of a fragment. This method generates a ‘forward’ and a ‘reverse’ read.
- **Unmappable read** – A read that cannot be unambiguously assigned a location in the genome. In this case, the threshold was set to 5. Thus, any read mapping to greater than 5 locations was labeled as unmappable.

References

- [1] K. C. Wang and H. Y. Chang. Molecular mechanisms of long noncoding rnas. *Mol Cell*, 43(6):904–914, 2011.
- [2] J. L. Rinn and H. Y. Chang. Genome regulation by long noncoding rnas. *Annu Rev Biochem*, 81:145–166, 2012.
- [3] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhataar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome Res*, 22(9):1775–1789, 2012.
- [4] B. Banfai, H. Jia, J. Khatun, E. Wood, B. Risk, Jr. Gundling, W. E., A. Kundaje, H. P. Gunawardena, Y. Yu, L. Xie, K. Krajewski, B. D. Strahl, X. Chen, P. Bickel, M. C. Giddings, J. B. Brown, and L. Lipovich. Long noncoding rnas are rarely translated in two human cell lines. *Genome Res*, 22(9):1646–1657, 2012.
- [5] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shores, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, P. Kheradpour, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi,

- S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. A. Feingold, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. C. Giddings, T. R. Gingeras, R. Guigo, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, J. A. Stamatoyannopoulos, S. A. Tennebaum, Z. Weng, K. P. White, B. Wold, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [6] K. L. Yap, S. Li, A. M. Munoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh, and M. M. Zhou. Molecular interplay of the noncoding rna anril and methylated histone h3 lysine 27 by polycomb cbx7 in transcriptional silencing of ink4a. *Mol Cell*, 38(5):662–674, 2010.
- [7] T. Sanchez-Elsner, D. Gou, E. Kremmer, and F. Sauer. Noncoding rnas of trithorax response elements recruit drosophila ash1 to ultrabithorax. *Science*, 311(5764):1118–1123, 2006.
- [8] S. Bertani, S. Sauer, E. Bolotin, and F. Sauer. The noncoding rna mistral activates hoxa6 and hoxa7 expression and stem cell differentiation by recruiting mll1 to chromatin. *Mol Cell*, 43(6):1040–1046, 2011.
- [9] S. R. Atkinson, S. Marguerat, and J. Bahler. Exploring long non-coding rnas through sequencing. *Semin Cell Dev Biol*, 23(2):200–205, 2012.
- [10] J. Robert E. Farrell. *RNA Methodologies: A Laboratory Guide for Isolation and Characterization*. Elsevier Science, 2005.
- [11] H. Lodish, A. Berk, C.A. Kaiser, M. Krieger, M.P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology*. W. H. Freeman, 2007.
- [12] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun,

- B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakraborty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.
- [13] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9(4):357–359, 2012.
- [14] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.
- [15] C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.
- [16] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, 2010.
- [17] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincnas. *Nat Biotechnol*, 28(5):503–510, 2010.
- [18] H. Kawaji, J. Severin, M. Lizio, A. R. Forrest, E. van Nimwegen, M. Rehli, K. Schroder, K. Irvine, H. Suzuki, P. Carninci, Y. Hayashizaki, and C. O. Daub.

- Update of the fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res*, 39(Database issue):D856–D860, 2011.
- [19] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick. Incrnadb: a reference database for long noncoding rnas. *Nucleic Acids Res*, 39(Database issue):D146–D151, 2011.
- [20] M. E. Dinger, K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond, and J. S. Mattick. Nred: a database of long noncoding rna expression. *Nucleic Acids Res*, 37(Database issue):D122–D126, 2009.
- [21] H. Hirai, T. Tani, N. Katoku-Kikyo, S. Kellner, P. Karian, M. Firpo, and N. Kikyo. Radical acceleration of nuclear reprogramming by chromatin remodeling with the transactivation domain of myod. *Stem Cells*, 29(9):1349–1361, 2011.
- [22] C. A. Berkes, D. A. Bergstrom, B. H. Penn, K. J. Seaver, P. S. Knoepfler, and S. J. Tapscott. Pbx marks genes for activation by myod indicating a role for a homeodomain protein in establishing myogenic potential. *Mol Cell*, 14(4):465–477, 2004.
- [23] I. L. de la Serna, K. A. Carlson, and A. N. Imbalzano. Mammalian swi/snf complexes promote myod-mediated muscle differentiation. *Nat Genet*, 27(2):187–190, 2001.
- [24] H. Hirai, T. Tani, and N. Kikyo. Structure and functions of powerful transactivators: Vp16, myod and foxa. *Int J Dev Biol*, 54(11-12):1589–1596, 2010.
- [25] T. Miller, N. J. Krogan, J. Dover, H. Erdjument-Bromage, P. Tempst, M. Johnston, J. F. Greenblatt, and A. Shilatifard. Compass: a complex of proteins associated with a trithorax-related set domain protein. *Proc Natl Acad Sci U S A*, 98(23):12902–12907, 2001.
- [26] V. Pirrotta. Polycomb the genome: Pcg, trxg, and chromatin silencing. *Cell*, 93(3):333–336, 1998.

- [27] R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone h3 lysine 27 methylation in polycomb-group silencing. *Science*, 298(5595):1039–1043, 2002.
- [28] D. Pasini, K. H. Hansen, J. Christensen, K. Agger, P. A. Cloos, and K. Helin. Coordinated regulation of transcriptional repression by the rbp2 h3k4 demethylase and polycomb-repressive complex 2. *Genes Dev*, 22(10):1345–1355, 2008.
- [29] T. Mahmoudi and C. P. Verrijzer. Chromatin silencing and activation by polycomb and trithorax group proteins. *Oncogene*, 20(24):3055–3066, 2001.
- [30] J. A. Kennison. The polycomb and trithorax group proteins of drosophila: trans-regulators of homeotic gene function. *Annu Rev Genet*, 29:289–303, 1995.
- [31] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, 129(7):1311–1323, 2007.
- [32] E. Pasmant, I. Laurendeau, D. Heron, M. Vidaud, D. Vidaud, and I. Bieche. Characterization of a germ-line deletion, including the entire ink4/arf locus, in a melanoma-neural system tumor family: identification of anril, an antisense non-coding rna whose expression coclusters with arf. *Cancer Res*, 67(8):3963–3969, 2007.
- [33] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander. lincrnas act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364):295–300, 2011.
- [34] M. E. Dinger, P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, B. B. Gardiner, M. E. Askarian-Amiri, K. Ru, G. Solda, C. Simons, S. M. Sunkin, M. L. Crowe, S. M. Grimmond, A. C. Perkins, and J. S. Mattick. Long noncoding rnas in mouse embryonic stem cell pluripotency and differentiation. *Genome Res*, 18(9):1433–1445, 2008.

- [35] S. Loewer, M. N. Cabili, M. Guttman, Y. H. Loh, K. Thomas, I. H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P. D. Manos, S. Datta, E. S. Lander, T. M. Schlaeger, G. Q. Daley, and J. L. Rinn. Large intergenic non-coding rna-ror modulates reprogramming of human induced pluripotent stem cells. *Nat Genet*, 42(12):1113–1117, 2010.
- [36] S. Y. Ng, R. Johnson, and L. W. Stanton. Human long non-coding rnas promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J*, 31(3):522–533, 2012.
- [37] M. Kretz, D. E. Webster, R. J. Flockhart, C. S. Lee, A. Zehnder, V. Lopez-Pajares, K. Qu, G. X. Zheng, J. Chow, G. E. Kim, J. L. Rinn, H. Y. Chang, Z. Siprashvili, and P. A. Khavari. Suppression of progenitor differentiation requires the long noncoding rna ancr. *Genes Dev*, 26(4):338–343, 2012.
- [38] J. Sheik Mohamed, P. M. Gaughwin, B. Lim, P. Robson, and L. Lipovich. Conserved long noncoding rnas transcriptionally regulated by oct4 and nanog modulate pluripotency in mouse embryonic stem cells. *RNA*, 16(2):324–337, 2010.
- [39] H. M. Blau, C. P. Chiu, and C. Webster. Cytoplasmic activation of human nuclear genes in stable heterocaryons. *Cell*, 32(4):1171–1180, 1983.
- [40] K. E. Yutzey, R. L. Kline, and S. F. Konieczny. An internal regulatory element controls troponin i gene expression. *Mol Cell Biol*, 9(4):1397–1405, 1989.
- [41] N. Yoshida, S. Yoshida, K. Koishi, K. Masuda, and Y. Nabeshima. Cell heterogeneity upon myogenic differentiation: down-regulation of myod and. *J Cell Sci*, 111 (Pt 6):769–779, 1998.
- [42] U. K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage t4. *Nature*, 227(5259):680–685, 1970.
- [43] D. Yaffe and O. Saxel. A myogenic cell line with altered serum requirements for differentiation. *Differentiation*, 7(3):159–166, 1977.

- [44] C. H. Clegg, T. A. Linkhart, B. B. Olwin, and S. D. Hauschka. Growth factor control of skeletal muscle differentiation: commitment to terminal. *J Cell Biol*, 105(2):949–956, 1987.
- [45] J. H. Yang, Y. Song, J. H. Seol, J. Y. Park, Y. J. Yang, J. W. Han, H. D. Youn, and E. J. Cho. Myogenic transcriptional activation of myod mediated by replication-independent histone deposition. *Proc Natl Acad Sci U S A*, 108(1):85–90, 2011.
- [46] S. G. Fischer and L. S. Lerman. Length-independent separation of dna restriction fragments in two-dimensional gel electrophoresis. *Cell*, 16(1):191–200, 1979.
- [47] D. J. Rodda, J. L. Chew, L. H. Lim, Y. H. Loh, B. Wang, H. H. Ng, and P. Robson. Transcriptional regulation of nanog by oct4 and sox2. *J Biol Chem*, 280(26):24731–24737, 2005.
- [48] I. Chambers and S. R. Tomlinson. The transcriptional foundation of pluripotency. *Development*, 136(14):2311–2322, 2009.
- [49] R. J. Flockhart, D. E. Webster, K. Qu, N. Mascarenhas, J. Kovalski, M. Kretz, and P. A. Khavari. Brav600e remodels the melanocyte transcriptome and induces bancr to regulate melanoma cell migration. *Genome Res*, 22(6):1006–1014, 2012.
- [50] S. Guil, M. Soler, A. Portela, J. Carrere, E. Fonalleras, A. Gomez, A. Villanueva, and M. Esteller. Intronic rnas mediate ezh2 regulation of epigenetic targets. *Nat Struct Mol Biol*, 19(7):664–670, 2012.
- [51] L. Kong, Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei, and G. Gao. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 35(Web Server issue):W345–349, 2007.
- [52] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, and E. S. Lander. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A*, 104(49):19428–19433, 2007.
- [53] J. Sambrook and D.W. Russell. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 2001.