CTCF is necessary for establishing the pattern of H3K27me3 and gene expression in the *Hoxa* cluster during motor neuron differentiation

Authors: Cora Hyun Jung & Christopher Catalano
Advisors: Professors Manpreet S. Katari & Brian Parker
Group Name: Applied Genomics: Introduction to Bioinformatics & Network Modeling
Submission Date: May 13, 2020

CTCF is necessary for establishing the pattern of H3K27me3 and gene expression in the *Hoxa* cluster during motor neuron differentiation

Cora Hyun Jung & Christopher Catalano

Abstract

An understanding of the mechanisms underlying the differentiation of embryonic stem cells into specific cell types has the potential to allow for targeted differentiation of disease-relevant cell types. Precise regulation of expression of the *Hoxa* family of genes underlies the differentiation of embryonic stem cells into motor neurons. The purpose of this study is to analyze the mechanisms underlying the change in *Hoxa* expression during motor neuron differentiation. This study examines the patterns of H3K27me3 histone modifications in the *Hoxa* cluster and gene expression in embryonic stem cells and motor neurons as they relate to the CRISPR-Cas9-mediated deletion of the CTCF binding site between *Hoxa5* and *Hoxa6* (Δ5|6). ChIP-seq analyses of CTCF and H3K27me3 reveal changes in *Hoxa* chromatin landscape and increased clearance of H3K27me3 in Δ5|6 motor neurons relative to *wt.* RNA-seq differential gene expression analysis and gene set enrichment analysis reveal a parallel increase in *Hoxa7* expression in Δ5|6 motor neurons relative to *wt.* and differentially expressed Gene Ontology gene sets related to ribosome function. These results demonstrate that CTCF binding at the 5|6 locus in the *Hoxa* cluster is required to establish a boundary between chromatin landscapes and the pattern of gene expression in differentiating motor neurons. This study contributes to the current understanding of motor neuron differentiation, which will ultimately facilitate research into motor neuron specific diseases.

Introduction

The ability to generate specific cell types through targeted embryonic stem cell (ESC) differentiation has broad potential application in the modeling of disease mechanisms and the development of therapies. It is necessary to understand the factors that govern and regulate gene expression in order to manipulate and ultimately control cell fate[1].

The *Hox* family of genes encodes for a group of related transcription factors that have conserved roles in specifying the body plans and morphology of animals. *Hox* genes confer segmental identity in developing embryos. Precise regulation of their expression is required to maintain the relative positions of cells, thereby ensuring that structures form in the appropriate locations[2]. During development, differential expression of the *Hox* genes establishes morphological patterning along the rostrocaudal axis, including specifying motor neuron (MN) subtype identity in the vertebrate hindbrain and spinal cord[3].

At the level of the rostral spinal cord, retinoic acid and sonic hedgehog signaling induce the expression of rostral *Hoxa* genes (*Hoxa1-Hoxa6)* in differentiating motor neurons[3]. This change in expression is mirrored by the loss of histone H3 lysine 27 trimethylation (H3K27me3) in the rostral region of the *Hoxa* cluster, whereas H3K27me3 is retained in the caudal region (*Hoxa7-Hoxa13*)[4]. H3K27me3 is a histone modification associated with facultative heterochromatin and transcriptionally repressed genes. Clearance of H3K27me3 is typically associated with euchromatin and transcriptionally active genes. The factors and mechanisms underlying this epigenetic phenomenon during motor neuron differentiation are not well understood.

The structure of chromatin is an important regulator of gene expression. CCCTC-binding factor (CTCF) is a DNA binding protein that functions to establish distinct topologically associated domains (TAD) in chromatin architecture. TAD boundaries are established by CTCF binding. DNA between pairs of CTCF proteins is extruded into

loops that establish TADs, which are discrete functional domains of DNA[5]. DNA within a TAD interacts with itself but is sequestered from DNA outside the TAD or in other TADs.

CTCF has a conserved binding motif that is present at several loci within the *Hoxa* cluster, including in the region that corresponds to the boundary between the rostral and caudal regions[6]. Because of its known function establishing TAD boundaries, we investigated whether the CRISPR/Cas9-mediated deletion of the rostral-most CTCF binding site between *Hoxa5* and *Hoxa6* (5|6) in the *Hoxa* cluster disrupted the pattern of H3K27me3 clearance in differentiating motor neurons[6]. The purpose of this study is to address the experimental question, is CTCF necessary for establishing the observed pattern of H3K27me3 and gene expression in the *Hoxa* cluster during embryonic stem cell differentiation into motor neurons? The null hypothesis of this analysis is that the deletion of the *Hoxa5|6* CTCF binding site will not affect the pattern of H3K27me3 in the *Hoxa* cluster or gene expression. The alternative hypothesis of this analysis is that deletion of *Hoxa5|6* CTCF binding site will significantly change the pattern of H3K27me3 in the *Hoxa* cluster and gene expression.

The analysis of chromatin regulation was performed using Chromatin Immunoprecipitation Sequencing (ChIP-seq) datasets of H3K27me3 and CTCF pulldowns in *Mus musculus* (mouse) embryonic stem cells and motor neurons. The analysis of gene expression during motor neuron differentiation was performed using Ribonucleic Acid Sequencing (RNA-seq) datasets of mouse embryonic stem cells and motor neurons[6]. This experiment is a factorial design of two factors: cell type and genotype. Each factor has two levels. The levels of cell type are embryonic stem cells and motor neurons. The levels of genotype are *wild type* (*wt*.) cells and cells with a CRISPR/Cas9-mediated deletion of the CTCF binding site at the boundary between *Hoxa5* and *Hoxa6* (Δ5|6)[6]. There are two biological replicates for each condition. Single-ended sequencing data was generated on an Illumina HiSeq[6]. We performed an analysis of CTCF and H3K27me3 enrichment in the *Hoxa* cluster from ChIP-seq datasets. We performed a differential gene expression analysis of the *Hoxa* genes as well as a Gene Ontology (GO) gene set enrichment analysis (GSEA) from RNA-seq datasets.

The results of this study provide insight into the role of CTCF in the regulation of embryonic stem cell differentiation into motor neurons through changes in chromatin structure. Motor neurons are implicated in wide range of motor neuron diseases, including amyotrophic lateral sclerosis (ALS), progressive bulbar palsy (PBP), and spinal muscular atrophy[7]. An understanding of how the *Hoxa* genes are regulated during motor neuron differentiation *in vivo* will provide insight that will eventually allow for the precise differentiation of embryonic stem cells *in vitro* to specific disease-relevant cell identities.

Results

ChIP-seq Dataset Preprocessing

Author's note: Code used to generate the below results can be found in the ChIP-seq Analysis section of the Supplementary Materials.

Datasets

Illumina HiSeq ChIP-seq datasets were downloaded from the Sequence Read Archive (SRA) at The National Center for Biotechnology Information (NCBI) using the fasterq-dump command-line function of the SRA-tools module on the Prince cluster of the High Performance Computing (HPC) servers of New York University (NYU) (Table 1)[6]. Each dataset contains sequencing data from two merged biological replicates.

Sample	Run	# of Bases	Published
CTCF.MN.WT	SRR1539508	2.2G	2015-02-27
CTCF.MN. Δ 5|6	SRR1539509	1.7G	2015-02-27
H3K27me3.ESC.WT	SRR1539510	2.8G	2015-02-27
H3K27me3.ESC. Δ 5|6	SRR1539511	2.9G	2015-02-27
H3K27me3.MN.WT	SRR1539512	2.5G	2015-02-27
H3K27me3.MN. Δ 5|6	SRR1539513	2.8G	2015-02-27

Table 1: Illumina HiSeq ChIP-seq datasets. CTCF pulldowns were performed on *wt.* and Δ5|6 differentiated motor neurons. H3K27me3 pulldowns were performed on *wt.* and Δ5|6 embryonic stem cells and *wt.* and Δ5|6 differentiated motor neurons. Each dataset contains sequencing data from two merged biological replicates.

Trimming and Quality Control

A quality control check on the raw ChIP-seq datasets was performed using the FastQC module before alignment to the mouse genome[8]. FastQC identified the presence of Illumina Universal Adapters in three of the six datasets. The Trim Galore module from Babraham Bioinformatics was used to remove adapter sequences from the datasets, as well as short sequences (below 10bp), and low-quality sequences with a Phred score below 20 (Table 2)[9]. FastQC was run on the trimmed FASTA files to confirm the quality of the trimmed sequences and proper adapter content removal.

Sample	Reads with Adapters (%)	Low Quality Reads (%)	Short Reads Trimmed (%)
CTCF.MN.WT	32.9	1.8	30.8
CTCF.MN. Δ 5|6	49.6	1.4	48.0
H3K27me3.ESC.WT	3.3	0.5	0.1
H3K27me3.ESC. Δ 5|6	3.9	0.6	0.7
H3K27me3.MN.WT	10.3	1.1	7.5
H3K27me3.MN. Δ 5|6	4.0	0.6	0.9

Table 2: Summary of run statistics as reported by Trim Galore. CTCF pulldown datasets exhibited significantly higher Illumina adapter content than H3K27me3 pulldown datasets. It is typical to see higher adapter content in transcription factor pulldowns than in histone marker pulldowns. The abundance of short reads (below 10bp) was correlated with adapter content.

Alignment

The December 2011 assembly of the *Mus musculus* genome (mm10, Genome Reference Consortium Mouse Build 38: GRCm38), was downloaded from the University of California Santa Cruz (UCSC) Genomics Institute Genome Browser using the wget command-line function on Prince. A Bowtie index was built of the mouse reference

genome FASTA file using the bowtie2-build indexer of the Bowtie2 module[10]. Bowtie2 was then used to align the trimmed ChIP-seq datasets to the mouse reference genome, generating SAM files (Table 3)[10]. All datasets were successfully aligned to the mouse reference genome with a high overall alignment rate.

Sample	Overall Alignment Rate (%)
CTCF.MN.WT	97.23
CTCF.MN. Δ 5|6	97.12
H3K27me3.ESC.WT	97.71
H3K27me3.ESC. Δ 5|6	98.36
H3K27me3.MN.WT	97.95
H3K27me3.MN. Δ 5|6	96.80

Table 3: Alignment statistics as reported by Bowtie2. After trimming of adapter content, low quality reads, and short reads with Trim Galore, each ChIP-seq data set aligned to the mm10 mouse reference genome with a high overall alignment rate at or above 96.80%.

Preparation of BAM files

The SortSam function of Picard Tools from Broad Institute was used to generate sorted BAM files from each of the SAM files[11]. The ValidateSamFile function of Picard Tools was used to report on the validity of the generated BAM files with respect to the specifications of the BAM format[11]. ValidateSamFile identified a missing read group in each of the BAM files. The AddOrReplaceReadGroups function of Picard Tools was used to replace the missing read group in each of the BAM files[11]. The ValidateSamFile function was run on the new BAM files in order to confirm that the read group had been properly added to each BAM file and that the files conformed to the BAM format. As a final step of preprocessing of the ChIP-seq datasets, the MarkDuplicates function of Picard Tools was used to identify and remove duplicate reads from the BAM files[11]. The index function of the SAMtools module was used to generate an index file for each of the BAM files[12].

ChIP-seq Analysis and Visualization

Author's note: Code used to generate the below results can be found in the ChIP-seq Analysis section of the Supplementary Materials.

Quality check

The plotFingerprint function of the deepTools module from the Bioinformatics Facility at the Max Planck Institute for Immunobiology and Epigenetics, Freiburg was used to generate a read coverage plot for each of BAM files (Figure 1)[13]. These results provide a measure of the quality of each of the antibody pulldowns. Input pulldown datasets, typically generated with a nonspecific antibody or by sequencing a library without an antibody pulldown, were not available. A read coverage plot for an ideal input pulldown would yield a line from the origin with a slope of 1, indicated a uniform distribution of reads. The read coverage plots for each of the antibody pulldowns display a non-uniform distribution of reads, with enrichment in a small fraction of the genome. This is an indication that the antibody pulldowns were successful.

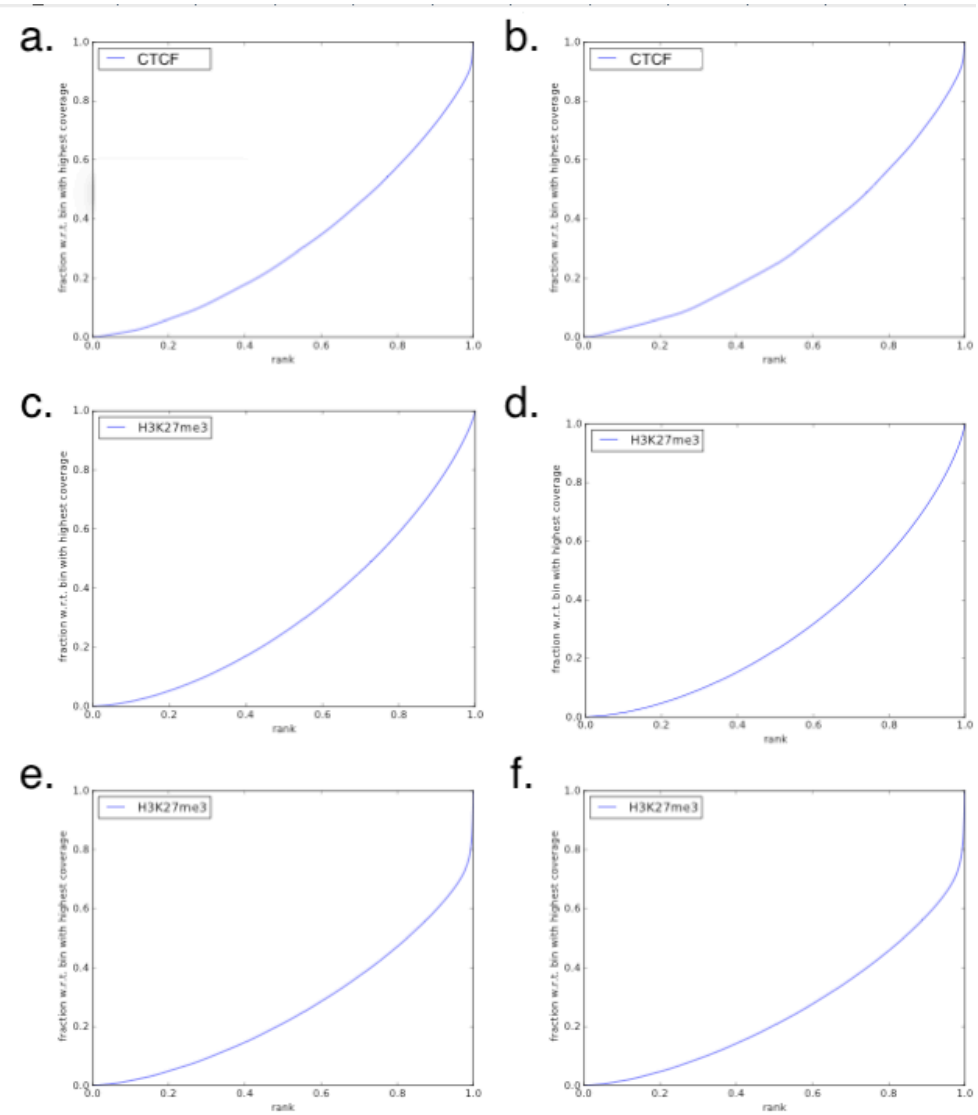

Figure 1: Read coverage plots for each antibody pulldown generated by the plotFingerprint function of deepTools. Input datasets were not available but would be expected to generate a plot with a slope of 1. Each coverage plot displays enrichment of pulldown signal relative to a theoretical input read. This is an indication that the antibody pulldowns were successful. **(a)** CTCF.MN.WT (**b)** CTCF.MN.Δ5|6 **(c)** H3K27me3.ESC.WT **(d)** H3K27me3.ESC.Δ5|6 **(e)** H3K27me3.MN.WT **(f)** H3K27me3.MN.Δ5|6

Visualization of enrichment

The bamCoverage function of the deepTools module was used to generate coverage tracks in bigWig file format from each of the aligned BAM files[13]. Reads were normalized using a Reads Per Kilobase per Million scaling factor (RPKM). The generated bigWig files were opened in the Integrative Genomics Viewer (IGV) to visualize enrichment of CTCF and H3K27me3 in the *Hoxa* cluster (Figure 2). Δ5|6 cells display a loss of CTCF enrichment at the 5|6 locus as well as the locus between *Hoxa6* and *Hoxa7* (6|7). This result suggests that there is an interaction between the CTCF transcription factor that binds at the 5|6 locus and the CTCF transcription factor that binds at the 6|7 locus. This is consistent with the known behavior of CTCF in forming

distinct TADs. Because of the loss of CTCF binding at these two loci, the CTCF bound to the locus between *Hoxa7* and *Hoxa9* (7|9) is the rostral-most CTCF in Δ5|6 cells.

In embryonic stem cells, H3K27me3 decorates chromatin across the entire *Hoxa* cluster. This is observed in both *wt.* and Δ5|6 embryonic stem cells (Figure 2). In motor neurons, H3K27me3 is lost in the rostral region of the *Hoxa* cluster and retained in the caudal region of the *Hoxa* cluster. In *wt.* motor neurons, clearance of H3K27me3 is observed from *Hoxa1* through *Hoxa6.* The CTCF that is bound to the 6|7 locus seems to form a boundary between rostral and caudal regions, delineating the region of H3K27me3 clearance from the region that retains H3K27me3. In Δ5|6 motor neurons, clearance of H3K27me3 is observed from *Hoxa1* through *Hoxa6* as in *wt.* In addition, there is a reduction of enrichment of H3K27me3 from *Hoxa6* through *Hoxa7* in Δ5|6 motor neurons compared to *wt.* motor neurons. This reduction of H3K27me3 enrichment is not observed from the start of *Hoxa9* to the end of the *Hoxa* cluster. This is consistent with the rostral-most CTCF at the 7|9 locus forming a new boundary between rostral and caudal regions within the *Hoxa* cluster in Δ5|6 motor neurons. These results suggest that CTCF binding to the 5|6 locus in the *Hoxa* cluster is necessary to prevent the clearance of H3K27me3 from spreading from the rostral to the caudal regions of the *Hoxa* cluster in differentiating motor neurons.

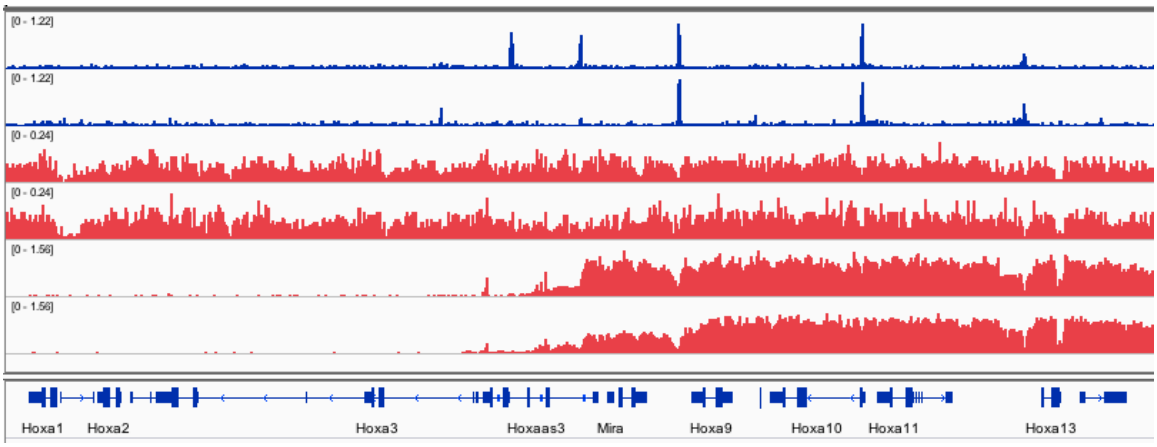

Figure 2: Visualization of enrichment in the *Hoxa* cluster using Integrative Genomics Viewer. Top to bottom: CTCF enrichment in *wt.* motor neurons; CTCF enrichment in Δ5|6 motor neurons; H3K27me3 enrichment in *wt.* embryonic stem cells; H3K27me3 enrichment in Δ5|6 embryonic stem cells; H3K27me3 enrichment in *wt.* motor neurons; H3K27me3 enrichment in Δ5|6 motor neurons; visualization of the mm10 *Hoxa* cluster. Scale normalized by reads per million (RPM). CTCF enrichment is lost at the 5|6 and 6|7 loci in Δ5|6 motor neurons relative to *wt.* motor neurons. H3K27me3 is lost in the rostral region of motor neurons relative to embryonic stem cells, with increased reduction through *Hoxa7* in Δ5|6 motor neurons relative to *wt.* motor neurons.

Enrichment peak calling

The callpeak function of the MACS module was used to identify peaks from the aligned BAM files[14]. Narrowpeak files were generated with called peaks from the CTCF pulldown data corresponding to CTCF enrichment across the mouse genome in *wt.* and Δ5|6 motor neurons. Broadpeak files were generated with called peaks from the H3K27me3 pulldown data corresponding to H3K27me3 histone marker enrichment across the mouse genome in *wt.* and Δ5|6 embryonic stem cells and motor neurons. Narrowpeak and Broadpeak files were visualized in IGV (Figure 3).

Peaks of CTCF enrichment were called at the 5|6 locus and 6|7 locus in *wt.* motor neurons. No peak of CTCF enrichment was called at the 5|6 locus in Δ5|6 motor neurons, confirming that deletion of the CTCF binding site successfully ablated CTCF binding at this locus. A peak of CTCF enrichment was called at the 6|7 locus in Δ5|6 cells, but it was significantly weaker than the peak that was called at the same locus in *wt.* motor neurons (Figure 3). Loss of CTCF binding at the 5|6 locus is correlated with reduction of CTCF binding at the 6|7 locus. This supports the model that the CTCF transcription factor at the 5|6 locus interacts with the CTCF transcription factor at the 6|7 locus.

MACS did not call any peaks of H3K27me3 enrichment in either condition of embryonic stem cells. This may have been a result of the broad enrichment across the *Hoxa* cluster (Figure 2) or a consequence of the lack of an input control. Peaks of H3K27me3 enrichment were detected at *Hoxa7* in both motor neuron conditions. In *wt.* motor neurons, three broad peaks of H3K27me3 were called over *Hoxa7*. In Δ5|6 motor neurons, two weaker broad peaks of H3K27me3 were called over *Hoxa7* (Figure 3). These results together suggest that loss of CTCF binding at the 5|6 and 6|7 loci in the *Hoxa* cluster leads to a loss of H3K27me3 in Δ5|6 motor neurons.

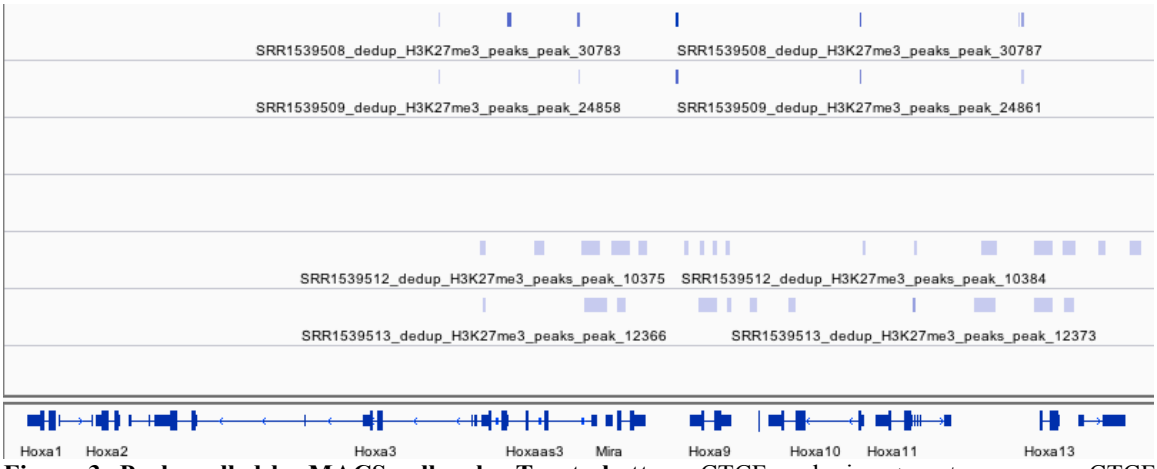

Figure 3: Peaks called by MACS callpeak. Top to bottom: CTCF peaks in *wt.* motor neurons; CTCF peaks in Δ5|6 motor neurons; H3K27me3 peaks in *wt.* embryonic stem cells; H3K27me3 peaks in Δ5|6 embryonic stem cells; H3K27me3 peaks in *wt.* motor neurons; H3K27me3 peaks in Δ5|6 motor neurons; visualization of the mm10 *Hoxa* cluster. CTCF peaks were called at the 5|6 and 6|7 loci in *wt.* motor neurons. No CTCF peak was detected at the 5|6 locus, and a weak peak was detected at the 6|7 locus in Δ5|6 motor neurons. No peaks of H3K27me3 were called in embryonic stem cells. A decreased number and significance of H3K27me3 peaks were detected at *Hoxa7* in Δ5|6 motor neurons relative to *wt.* motor neurons.

Differential peak calling

The bdgdiff function of the MACS module was used to identify regions of differential enrichment of H3K27me3 between *wt.* and Δ5|6 motor neurons[14]. BED files were generated for differentially enriched peaks in *wt.* and Δ5|6, as well as common peaks. BED files were visualized in IGV (Figure 4). Although MACS callpeak detected peaks at *Hoxa7* in both *wt.* and Δ5|6 motor neurons (Figure 3), the peaks detected in Δ5|6 motor neurons were of weaker significance. This finding was confirmed by MACS bdgdiff, which identified several differentially enriched peaks in the *Hoxa* cluster in *wt.* cells that were not present in Δ5|6 cells, specifically at the *Hoxa7* locus (Figure 4). This

differential ChIP-seq analysis reveals a significant loss of H3K27me3 in Δ5|6 motor neurons relative to *wt*. motor neurons.

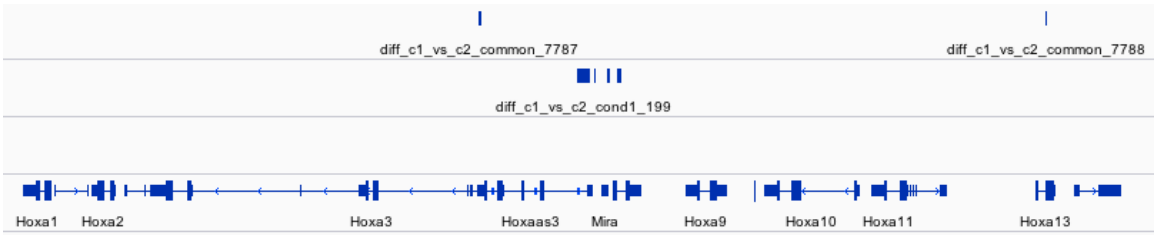

Figure 4: Regions of differential enrichment of H3K27me3 as identified by MACS bdgdiff. Top to bottom: Common peaks between *wt.* and Δ5|6 motor neurons; Differentially enriched peaks of H3K27me3 in *wt.* motor neurons; Differentially enriched peaks of H3K27me3 in Δ5|6 motor neurons. Several peaks of differential H3K27me3 enrichment in were called *wt.* motor neurons clustered at the *Hoxa7* locus.

Profile plots

The bigwigCompare function of the deepTools module was used to compare pairs of bigWig files based on the numbers of mapped reads in each file[13]. The function was run to compare the two CTCF pulldowns in *wt.* and Δ5|6 motor neurons as well as the two H3K27me3 pulldowns in *wt.* and Δ5|6 motor neurons. An output file in bigWig format was generated for each analysis reporting the log2 of the ratio between the two respective conditions.

The computeMatrix function of the deepTools module was used to generate files in BED format containing values of enrichment across the *Hoxa* cluster[13]. This analysis was run on each of the bigWig files corresponding to the antibody pulldowns as well as the each of the bigWig comparison files from bigwigCompare. The plotProfile function of the deepTools module was run with each of the generated BED files from computeMatrix to generate profile plots of each condition across the *Hoxa* cluster (Figure 5)[13].

The CTCF profile plots reveal four sharp peaks of enrichment in *wt.* motor neurons corresponding to the four CTCF binding sites in the *Hoxa* cluster (Figure 5a). The CTCF profile plots display only two sharp peaks of enrichment in Δ5|6 motor neurons, corresponding to the CTCF binding sites at the 7|9 locus and the 10|11 locus (between *Hoxa10* and *Hoxa11*) (Figure 5b). The CTCF profile plot generated from the output of the bigwigCompare function demonstrates the reduction in CTCF enrichment in Δ5|6 motor neurons relative to *wt.* motor neurons (Figure 5c). The two sharp nadirs of reduced enrichment correspond to the CTCF binding sites at the 5|6 and 6|7 loci. These data further confirm that the CRISPR/Cas9-mediated deletion of the 5|6 CTCF binding site led to a significant loss of CTCF binding at both the 5|6 and 6|7 loci in the *Hoxa* cluster in motor neurons.

The H3K27me3 profile plots reveal the sharp boundary between the caudal and rostral regions of the *Hoxa* cluster in motor neurons, which correspond to the regions with and without H3K27me3 enrichment, respectively. In *wt.* motor neurons, H3K27me3 is highly enriched in the region corresponding to *Hoxa7* through *Hoxa13* (Figure 5d). In Δ5|6 motor neurons, H3K27me3 enrichment is observed across this same region, but the enrichment is significantly reduced in the region corresponding to *Hoxa7* (Figure 5e). The H3K27me3 profile plot generated from the output of the bigwigCompare function demonstrates the reduction in H3K27me enrichment in Δ5|6

motor neurons relative to *wt.* motor neurons (Figure 5f). The sharp nadir of reduced enrichment corresponds to the loss of H3K27me at the *Hoxa7* locus. These data confirm that the CRISPR/Cas9-mediated deletion of the 5|6 CTCF binding site led to a significant loss of H3K27me3 at *Hoxa7* in motor neurons.

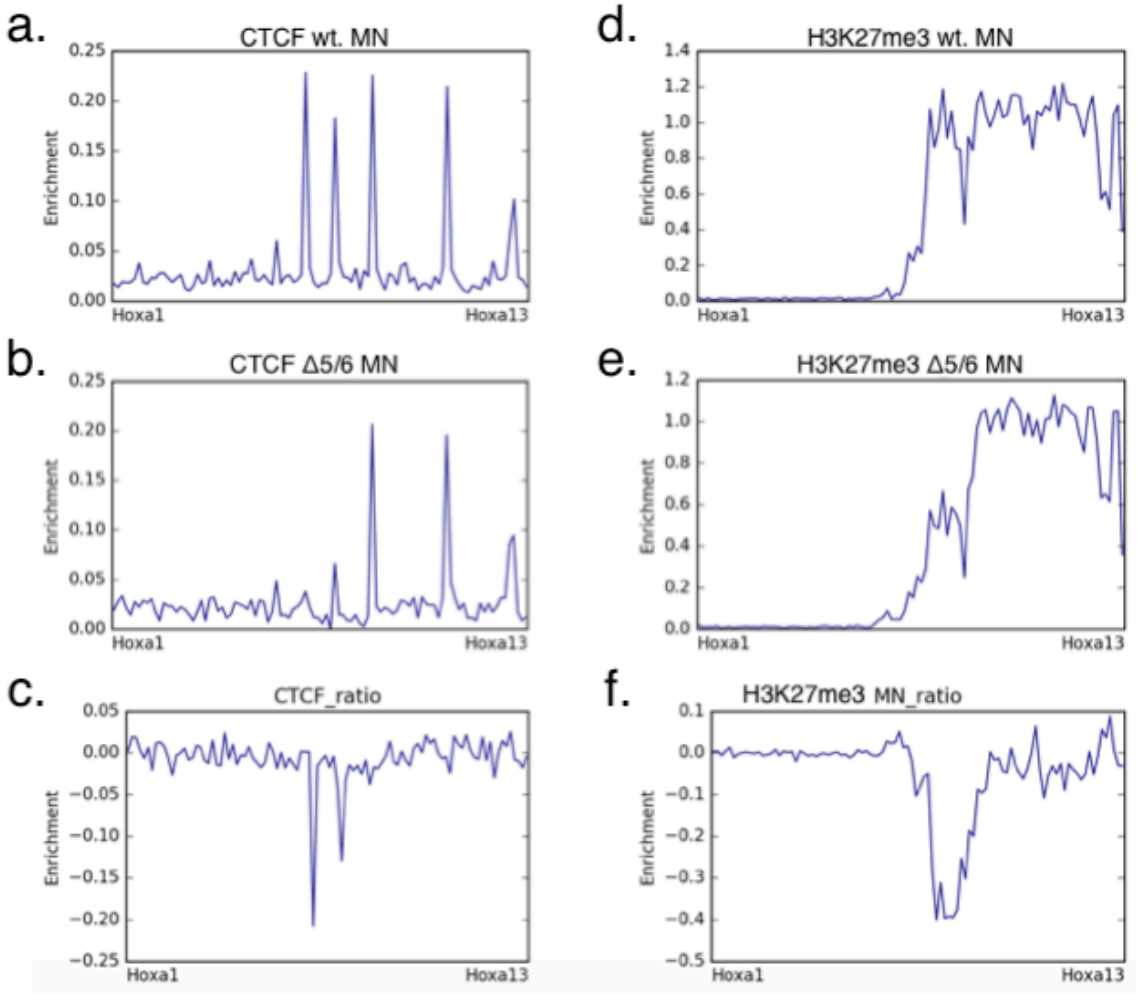

Figure 5: Profile plots across the *Hoxa* cluster. **(a)** CTCF enrichment in *wt.* motor neurons. **(b)** CTCF enrichment in Δ5|6 motor neurons. **(c)** Relative CTCF enrichment in Δ5|6 motor neurons compared to *wt.* motor neurons (log2 ratio). **(d)** H3K27me enrichment in *wt.* motor neurons. **(e)** H3K27me3 enrichment in Δ5|6 motor neurons. **(f)** Relative H3K27me3 enrichment in Δ5|6 motor neurons compared to *wt.* motor neurons (log2 ratio). A significant loss of CTCF enrichment is observed at the 5|6 and 6|7 loci in Δ5|6 motor neurons. A significant loss of H3K27me3 enrichment is observed at the *Hoxa7* locus in Δ5|6 motor neurons.

RNA-seq Dataset Preprocessing

Author's note: Code used to generate the below results can be found in the RNA-seq Workflow section of the Supplementary Materials.

Datasets

Single-ended Illumina HiSeq RNA-seq SRA datasets were downloaded from NCBI and converted into FASTQ formats using fasterq-dump on the Prince cluster of

HPC at NYU (Table 4)[6]. This experiment is a factorial design of two factors: cell type and genotype. Each factor has two levels. The levels of cell type are embryonic stem cells and motor neurons. The levels of genotype are *wt.* cells and cells with a deletion of the CTCF binding site at the *Hoxa5|6* boundary (Δ5|6). Each condition has two replicates.

Sample	Run	# of Bases	Published
ESC.WT_1	SRR1539361	700.66 M	2015-02-27
ESC.WT_2	SRR1539362	797.09 M	2015-02-27
ESC.Δ5|6_1	SRR1539363	543.70 M	2015-02-27
ESC.Δ5|6_2	SRR1539364	873.20 M	2015-02-27
MN.WT_1	SRR1539366	655.70 M	2015-02-27
MN.WT_2	SRR1539367	743.30 M	2015-02-27
MN.Δ5|6_1	SRR1539368	590.20 M	2015-02-27
MN.Δ5|6_2	SRR1539369	1 G	2015-02-27

Table 4: Illumina HiSeq RNA-seq datasets. ESC.WT: wild type embryonic stem cells. ESC.Δ5|6: embryonic stem cells with deletion of CTCF binding site at the Hoxa5|6 boundary. MN.WT: wild type motor neurons. MN.Δ5|6: motor neurons with a deletion of the 5|6 CTCF binding site. Each condition has two replicates.

Quality Control

A quality control was run using FastQC and MultiQC[8, 15]. No adaptor issues were flagged with a high percentage of duplicate reads ranging from 60~80% (Table 5). Because they were confirmed as high quality reads, no trimming was performed on the sequence datasets.

Sample	% Dups	% GC	M Seqs
ESC.WT_1	68.3	52	15.9
ESC.WT_2	76.0	46	17.7
ESC.Δ5|6_1	65.1	53	12.4
ESC.Δ5|6_2	73.1	46	19.4
MN.WT_1	65.6	51	14.9
MN.WT_2	61.0	47	16.2
MN.Δ5|6_1	62.4	52	13.4
MN.Δ5|6_2	65.4	47	22.8

Table 5: Quality control results generated by MultiQC. Dups: % duplicated reads. GC: Average % GC counts. M Seqs: Total sequences (millions).

Alignment

Sequences were indexed and aligned using HISAT2 against the same mouse reference genome that was utilized for ChIP-seq dataset alignment (mm10/GRCm38)[16]. The alignment was successful for each dataset, with a high average overall alignment rate of approximately 90% (Table 6). The resulting SAM files were converted to BAM files using Picard Tools[11].

Sample	Aligned 0 times (%)	Aligned 1 time (%)	Aligned > 1 times (%)	Overall alignment rate (%)
ESC.WT_1	10.49	61.91	27.60	89.51
ESC.WT_2	10.93	51.06	38.01	89.07
ESC.Δ5|6_1	11.80	61.98	26.22	88.20
ESC.Δ5|6_2	16.41	47.02	36.57	83.59
MN.WT_1	10.29	67.65	22.07	89.71
MN.WT_2	8.94	64.74	26.32	91.06
MN.Δ5|6_1	9.61	68.94	21.45	90.39
MN.Δ5|6_2	6.99	64.80	28.21	93.01

Table 6. RNA-seq HISAT2 Alignment results. Aligned 0 times: % reads mapped 0 times. Aligned 1 time: % reads mapped 1 time. Aligned > 1 times: % reads mapped more than 1 times. Overall alignment rate: % total reads mapped.

Annotation

An Ensembl GTF file containing annotations for the mouse genome was downloaded from UCSC Genome Browser. The aligned BAM files were annotated against this GTF file using the featureCounts function from the RSubread library in R[17]. The average assignment rate across all datasets was approximately 60%. This resulted in a low sequencing depth as discussed later (Table 7).

Sample	Assignment Rate (%)
ESC.WT_1	62.1
ESC.WT_2	63.2
ESC.Δ5|6_1	62.3
ESC.Δ5|6_2	58.1
MN.Δ5|6_1	57.9
MN.WT_2	62.4
MN.Δ5|6_1	61.8
MN.Δ5|6_2	64.4

Table 7. RNA-seq Annotation assignment rate results. The average assignment rate across datasets was approximately 60%.

RNA-seq Analysis and Visualization

Author's note: Code used to generate the below results can be found in the RNA-seq Workflow section of the Supplementary Materials.

Exploratory Analysis

Gene counts were imported into R Studio for the further analysis. Ensembl Gene IDs were converted to common gene symbols using the EnsDb.Mmusculus.v79 Ensembl database. Because the primary focus of this analysis is expression of the *Hoxa* genes, gene counts for this set of genes across the *Hoxa* cluster were calculated. The gene count of *Hoxa7* in Δ5|6 motor neurons is approximately nine-fold greater than the gene count of *Hoxa7* in *wt.* motor neurons (Table 8). The upregulation of expression of *Hoxa7* is consistent with the observed downregulation of H3K27me3 in this region in Δ5|6 motor

neurons. Because H3K27me3 is a marker of facultative heterochromatin, loss of this chromatin modification would be expected to result in higher expression of the underlying genes.

Sample	ESC.WT_1	ESC.WT_2	ESC.Δ5|6_1	ESC.Δ5|6_2	MN.WT_1	MN.WT_2	MN.Δ5|6_1	MN.Δ5|6_2
Hoxa1	551	72	532	120	399	381	303	438
Hoxa2	189	14	174	38	180	267	211	292
Hoxa3	194	29	147	13	529	444	513	561
Hoxa4	34	6	33	8	368	325	358	373
Hoxa5	110	42	37	34	927	836	900	895
Hoxa6	1	2	0	0	92	29	58	109
Hoxa7	24	4	24	13	55	40	327	373
Hoxa8	0	1	5	8	4	12	47	41

Table 8. Gene counts of *HoxA* genes. Average gene count of *wt.* motor neurons in *Hoxa7* is 47.5 and the average gene count of *Hoxa7* in Δ5|6 motor neurons is 350. This reveals an upregulation of *Hoxa7* expression in Δ5|6 motor neurons.

A clustering visualization further supports the alternative hypothesis of this study that the CRISPR/Cas9-mediated deletion of the 5|6 CTCF binding site in the *Hoxa* cluster would significantly change the pattern of *Hoxa* expression. *Hoxa7* is more highly expressed after deletion of 5|6 CTCF binding site relative to the expression of the *Hoxa1* through *Hoxa6*. *Wt*. and Δ5|6 motor neurons are clustered together in the genomic region spanning *Hoxa1* through *Hoxa6* (Figure 6, left). However, Δ5|6 motor neurons display higher expression and cluster independently from *wt.* motor neurons in the genomic region spanning *Hoxa7* through *Hoxa9* (Figure 6, right).

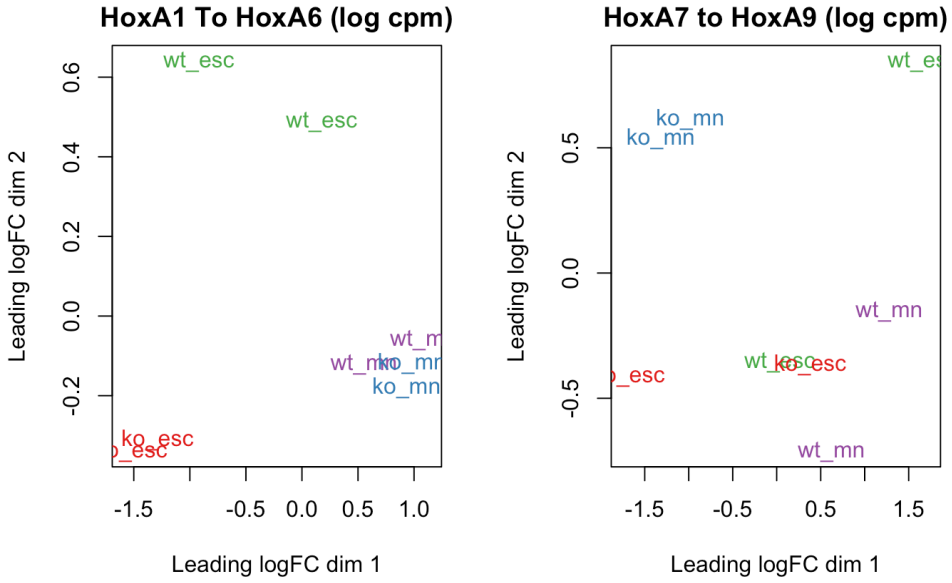

Figure 6. Clustering of samples in the *HoxA* cluster based on log2FC. **Left**: Clustering of samples in the region spanning *Hoxa1* through *HoxA6* in logCPM (counts per million). *Wt.* and Δ5|6 motor neurons cluster together. **Right**: Clustering of samples in the region spanning *Hoxa7* through *Hoxa9* in logCPM. Δ5|6 motor neurons display higher expression and cluster independently from *wt.* motor neurons.

Enrichment of transcript reads in the *Hoxa* cluster was further visualized in IGV. Δ5|6 motor neurons display upregulation of the transcription of *Hoxa7* relative to *wt.* motor neurons (Figure 7, left). A profile plot of the expression in the *Hoxa* cluster in Δ5|6 motor neurons relative to *wt.* motor neurons was created using deepTools (Figure 7, right)[13]. Expression is approximately equivalent between the two conditions in *Hoxa1* through *Hoxa6*. There is a sharp peak of increased expression at the locus corresponding to *Hoxa7*. Both IGV visualization and the profile plot display enrichment of *Hoxa7* in Δ5|6 samples relative to *wt.*, providing further support for a correlation between the loss of CTCF at the 5|6 locus in the *Hoxa* cluster and an increase in *Hoxa7* expression.

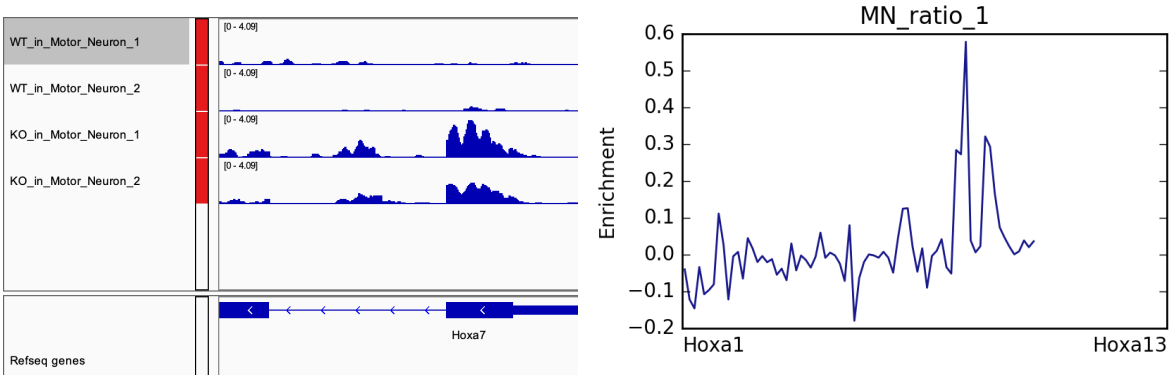

Figure 7. Enrichment of transcription across the *Hoxa* cluster. Left: IGV Visualization of transcript enrichment displaying increased expression at *Hoxa7* region in Δ5|6 motor neurons relative to *wt.* motor neurons. Scale normalized by reads per million (RPM). **Right**: Profile plot of relative expression in the *Hoxa* cluster displaying a sharp peak of increased expression at the *Hoxa7* locus in Δ5|6 motor neurons relative to *wt.* motor neurons.

The log2 fold change (log2FC) values of the relative expression of the *Hoxa* genes between Δ5|6 motor neurons and *wt.* motor neurons were visualized (Figure 8). The log2FC values reveal high expression of both *Hoxa7* and *Hoxa9* in Δ5|6 motor neurons relative to *wt.* motor neurons. These results provide further evidence for the previously observed *Hoxa7* expression pattern between the two experimental conditions. The increase of *Hoxa9* expression suggests that boundary between rostrally expressed genes and caudally repressed genes may have extended further into the caudal region of the *Hoxa* cluster in Δ5|6 motor neurons than previously understood.

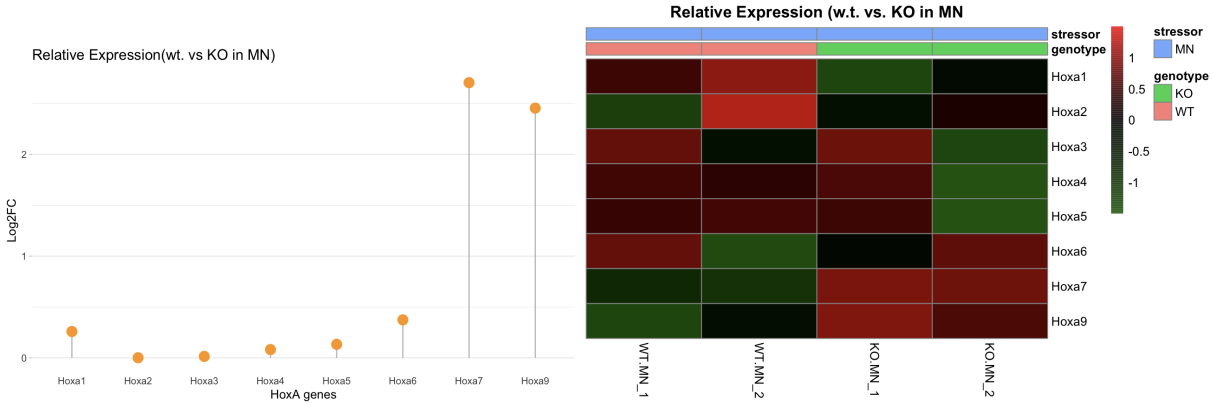

Figure 8. Visualization of log2 fold change of expression throughout *HoxA* Cluster in Δ5|6 motor neurons relative to *wt.* motor neurons. Left: A plot showing high log2FC of expression of *Hoxa7* and *Hoxa9* compared to the rest of the rostral *Hoxa* genes. **Right**: Heatmap showing upregulation of *Hoxa7* and *Hoxa9* in Δ5|6 motor neurons.

Statistical models for differential expression analysis

To determine the optimal statistical model our data, two statistical models, limma voom and DeSeq2 were utilized[18, 19, 20]. Due to the low levels of replication, distribution-free rank and permutation-based methods were ruled out. The Poisson model was considered a good candidate, but ultimately not appropriate for our analysis as a single parameter model. The mean and variance of Poisson distribution are equal, which makes it difficult to explain data that inherently has more variance. Thus, we decided to use a Negative Binomial model to fit our data, which uses the dispersion parameter to model the variability between replicates[18].

Limma-Voom

Negative Binomial methods treat the dispersion estimates as if they were known parameters without accounting for uncertainty of estimation, which may increase the false positive rate (type 1 error) due to overly liberal testing[20]. Thus we are using voom (variance modeling at the observational level) transformation, which uses the variance of genes to create weights for use in linear models[19]. This enables the RNA-seq data to be analyzed as if it were microarray data, so that the linear modeling or gene set testing methods in limma can be applied[18]. This is beneficial particularly because we have noisy data. A design matrix was created employing a factorial design of two factors: cell type and genotype (Figure 9).

	(Intercept)	genotypeKO	stressorMN	genotypeKO:stressorMN
1	1	0	0	0
2	1	0	0	0
3	1	1	0	0
4	1	1	0	0
5	1	0	1	0
6	1	0	1	0
7	1	1	1	1
8	1	1	1	1

Figure 9. Design Matrix for Limma-Voom. The design matrix uses the default R treatment contrasts. Intercept: *wt.* in motor neurons. genotypeKO: Δ5|6 relative to *wt.* in motor neurons. stressorMN: *wt.* in embryonic stem cells relative to motor neurons. genotypeKO:stressorMN: the difference between Δ5|6 in embryonic stem cells relative to motor neurons and wt. in embryonic stem cells relative to motor neurons. Our contrast of main interest is the second column, genotypeKO.

Non-specific filtering was performed to remove low count features. It was set to have at least twenty counts over two samples, which reduced a total number of unique genes from 43432 to 15937. After that, TMM (trimmed mean of M values) normalization by edgeR was performed to remove low count features for inter-sample comparison[21]. Then lmFit and eBayes functions were used to fit our data. TopTable was used to perform FDR (false discovery rate) multiple testing correction to extract the list of most significant genes. TopTable reported that no genes reached sufficient statistical significance (Figure 10). Possible explanations for this result are lack of statistical power due to small datasets and the limited number of replicates. While there is a fold change consistent with the alternative hypothesis, low gene sequencing depth and the limited number of replicates leads to lack of statistical power. The underlying read-count distribution for a gene is a fundamental property of RNA-seq data but without a large number of measurements or replicates, it is not possible to identify the form of this distribution unambiguously. Fewer replicates means the true distribution of read counts for an individual gene is unclear. Many differential gene expression (DGE) tools make strong assumptions about the form of this underlying distribution[21]. Thus, having an unreliable distribution can negatively impact on the ability to correctly identify significantly differentially expressed genes.

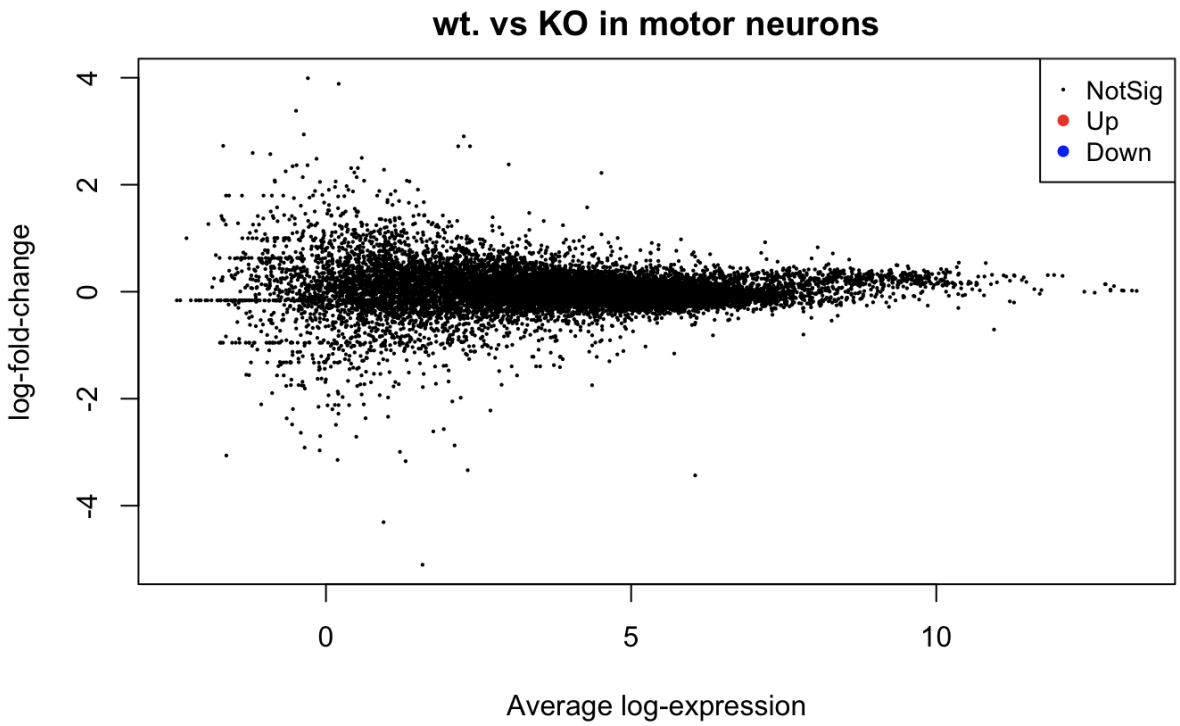

Figure 10. MA Plot created by fitted limma model. No differentially expressed genes achieved statistical significance in Δ5|6 relative to *wt.* motor neurons.

DESeq2

Because limma-voom did not identify any statistically significant differentially expressed genes, DESeq2 analysis was performed. DESeq2 computes a common value, trend, or prior distribution for the dispersions and then shrinks individual gene-wise dispersion estimates toward the chosen anchor[20]. A core challenge was the limited number of replicates (only two per condition). Inferential methods that treat each gene separately suffer here from lack of statistical power, due to the high uncertainty of within-group variance estimates[20]. One sensible solution is to share information across genes. With the employment of DESeq2 to model data under the assumption that genes of similar average expression strength have similar dispersion, a number of genes were identified that reached statistical significance. Analysis was run on three contrasts of interest: (1) Δ5|6 relative to *wt.* in motor neurons, (2) wt. in motor neurons relative to embryonic stem cells, and (3) Δ5|6 in motor neurons relative to embryonic stem cells.

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a given gene set shows statistically significant, concordant differences between two different biological states. Gene Ontology (GO) gene sets were used for this analysis. GSEA was run using the clusterProfiler package in R. The log2FC results from DESeq2 were used to run GSEA with a p-value cutoff of 0.05. The *Mus musculus* annotation file "org.Mm.eg.db" was downloaded from Bioconductor. The dotPlot function from DOSE (Disease Ontology Semantic and Enrichment) package was

used to visualize selected top GO pathways[22]. GSEA was performed utilizing the same methodology for each condition.

Condition 1: Δ5|6 relative to wt. in motor neurons

This analysis examined the log2FC of expression between Δ5|6 relative to *wt.* in motor neurons. DESeq2 analysis results show that there are three genes that reached statistical significance with a False Discovery Rate (FDR) < 0.05. The *Hoxa7* gene falls within FDR < 0.05 and log2FC > 2. Two other genes, *Plagl1* and *Grb10* fall within FDR < 0.05 and log2FC < 2 (Figure 11, left). *Plagl1* encodes for a zinc-finger transcription factor, and *Grb10* encodes for an insulin-receptor binding protein[22, 23]. These two genes have no known role in motor neuron development and are disregarded for the purpose of this analysis. Future research into the possible function of these genes with respect to motor neuron development may be of interest. A volcano plot visualization shows *Hoxa7* as statistically significant and highly expressed in Δ5|6 motor neurons (Figure 11, right). These results reveal that the deletion of the CTCF binding site at the 5|6 locus led to the significant upregulation of *Hoxa7* expression. The top differentially expressed GO gene sets as reported by GSEA were visualized by encoding gene counts as dot size and adjusted p value as colors from red to blue in ascending order (Figure 12). Several of the top upregulated GO gene sets were involved in ribosome production. This may indicate an increased overall level of translation as a result of the upregulation of the transcription factors *Hoxa7* and *Hoxa9*. Several of the downregulated GO gene sets are involved in differentiation into other cell types, including epithelial cell fate commitment, thalamus development, and skeletal muscle cell differentiation. This result is not unexpected in differentiated motor neuron cells.

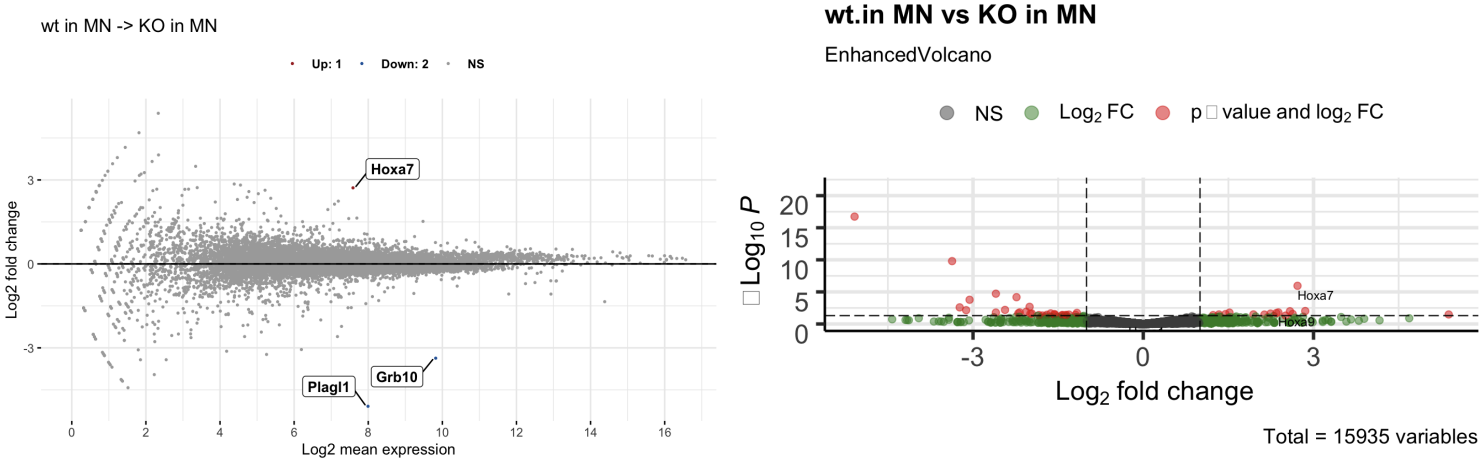

Figure 11: Visualization of relative expression in Δ5|6 and *wt.* motor neurons. Left: MA plot with labeled dots indicating significant differentially expressed genes (FDR < 0.05, logFC > 2). *Hoxa7* displays a significant positive log2FC in Δ5|6 motor neurons relative to *wt.* motor neurons. **Right**: Volcano plot with colored dots indicating significant genes (FDR < 0.05). *Hoxa7* and *Hoxa9* are indicated as statistically significant and have positive log2 FC values.

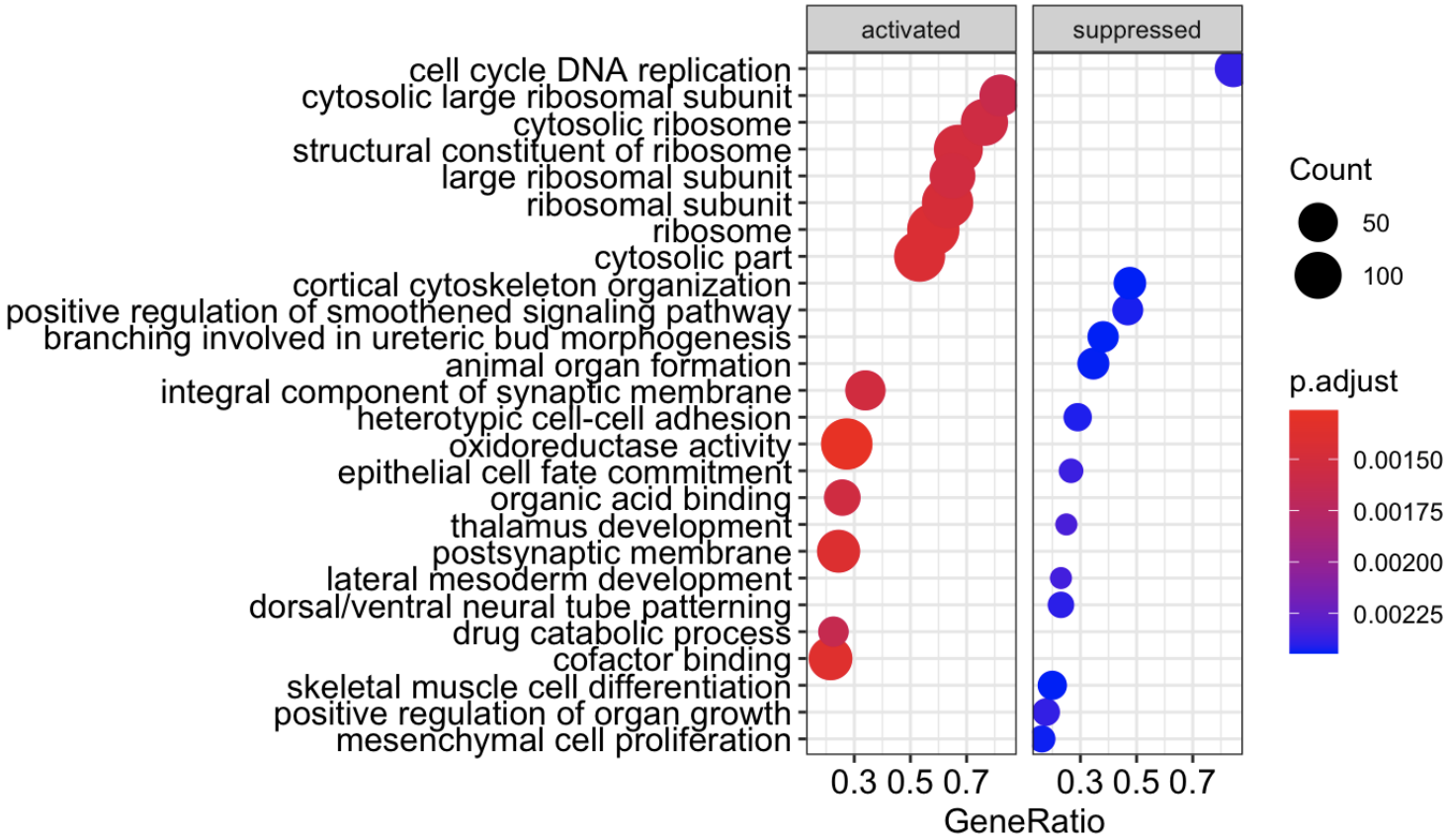

Figure 12. Top differentially expressed GO gene sets in Δ5|6 relative to *wt.* in motor neurons. Left: GO gene sets significantly upregulated. **Right**: GO gene sets significantly downregulated.

Condition 2: wt. motor neurons relative to embryonic stem cells

A total of 979 genes were determined to be differentially expressed between *wt.* motor neurons and embryonic stem cells with an FDR < 0.05 and with a log2FC > 2. 2149 genes are within FDR < 0.05 and log2FC < 2. An MA plot was generated with the top twenty significant genes labeled (Figure 13, left). A volcano plot was generated to show the statistical significance of upregulated and downregulated genes (Figure 13, right). GSEA analysis revealed that many of these statistically significant differentially enriched genes are associated with neuronal development and function, including axonogenesis, glutamatergic synapse, and positive regulation of neurogenesis. This result was not unexpected as the ability to differentiate *wt.* embryonic stem cells into motor neurons through retinoic acid and sonic hedgehog agonist treatment is well established[6]. Several of the top downregulated GO gene sets are again associated with differentiation into different cell types. The top differentially expressed GO gene sets were listed and visualized by encoding gene counts as dot size and adjusted p value as colors from red to blue in ascending order (Figure 14).

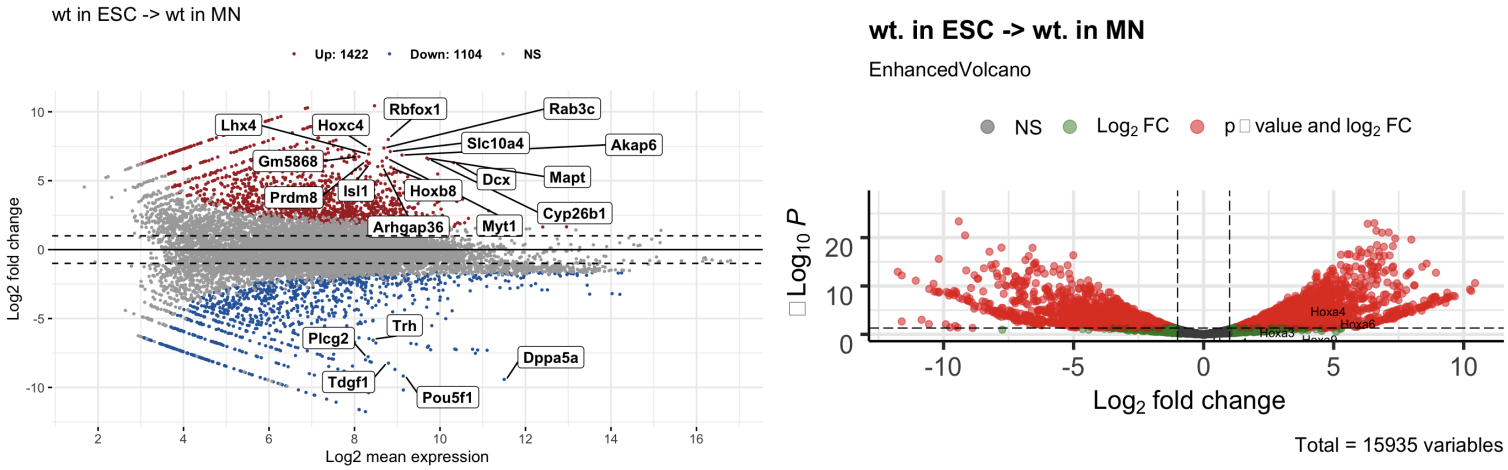

Figure 13. Visualization of relative expression in *wt.* motor neurons and embryonic stem cells, log2FC > 2 and FDR < 0.05. **Left**: MA plot with colored dots indicating significant genes. Top 20 significant genes are labeled. **Right**: Volcano plot with colored dots indicating significant genes. *Hoxa4* and *Hoxa6* are indicated as both statistically significant and have positive log2FC values.

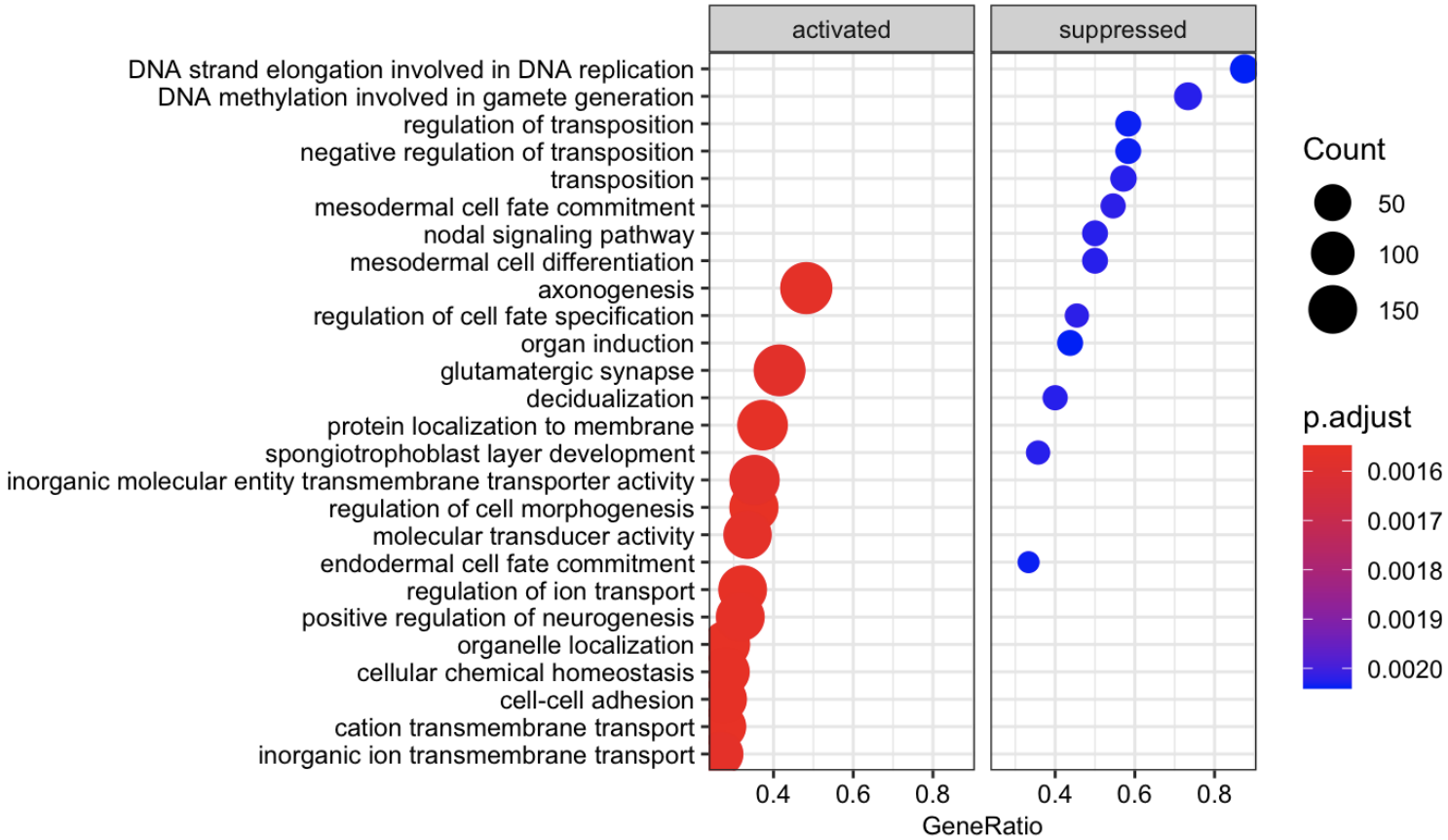

Figure 14. Top differentially expressed GO gene sets in *Wt.* motor neurons relative to embryonic stem cells. **Left**: GO gene sets significantly upregulated. **Right**: GO gene sets significantly downregulated.

Condition 3: Δ5|6 motor neurons relative to embryonic stem cells.

Differential gene expression analysis revealed a total of 1190 genes that fall within an FDR < 0.05 and log2FC > 2. 860 genes are within FDR <0.05 and log2FC < 2. An MA plot was generated with the top twenty significant differentially expressed genes labeled (Figure 15, left). A volcano plot was created to show the statistical significance of upregulated and downregulated genes (Figure 15, right). GSEA analysis revealed that, like in *wt.* motor neurons relative to embryonic stem cells, Δ5|6 motor neurons exhibited many upregulated GO gene sets related to neuron development and function relative to embryonic stem cells. These neuron-related pathways include axon development, presynapse, glutamatergic synapse, and positive regulation of neurogenesis. Top downregulated GO gene sets again are associated with differentiation into different cell types. This is consistent with the observation that Δ5|6 embryonic stem cells retain their ability to differentiate into motor neurons upon exposure to retinoic acid and sonic hedgehog agonist[6]. The top differentially expressed GO gene sets were listed and visualized by encoding gene counts as dot size and adjusted p value as colors from red to blue in ascending order (Figure 16).

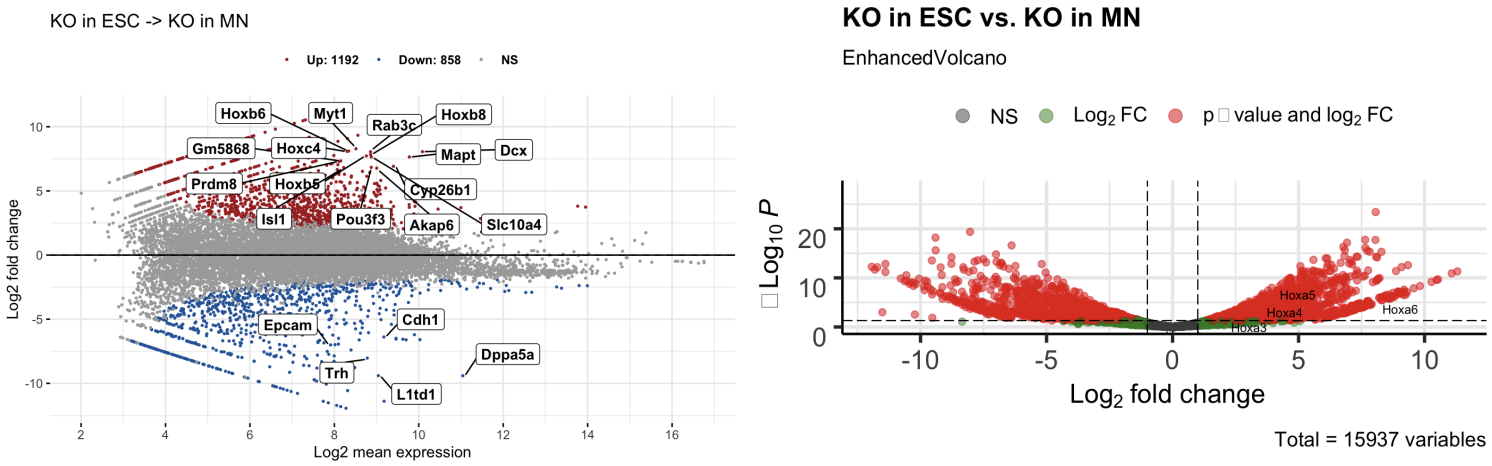

Figure 15. Visualization of relative expression in Δ5|6 motor neurons and embryonic stem cells, log2FC > 2 and FDR < 0.05. Left: MA plot with colored dots indicating significant genes. Top 20 significant genes are labeled. **Right**: Volcano plot with colored dots indicating significant genes. *Hoxa4*, *Hoxa5*, and *Hoxa6* are indicated as both statistically significant and have positive log2FC values.

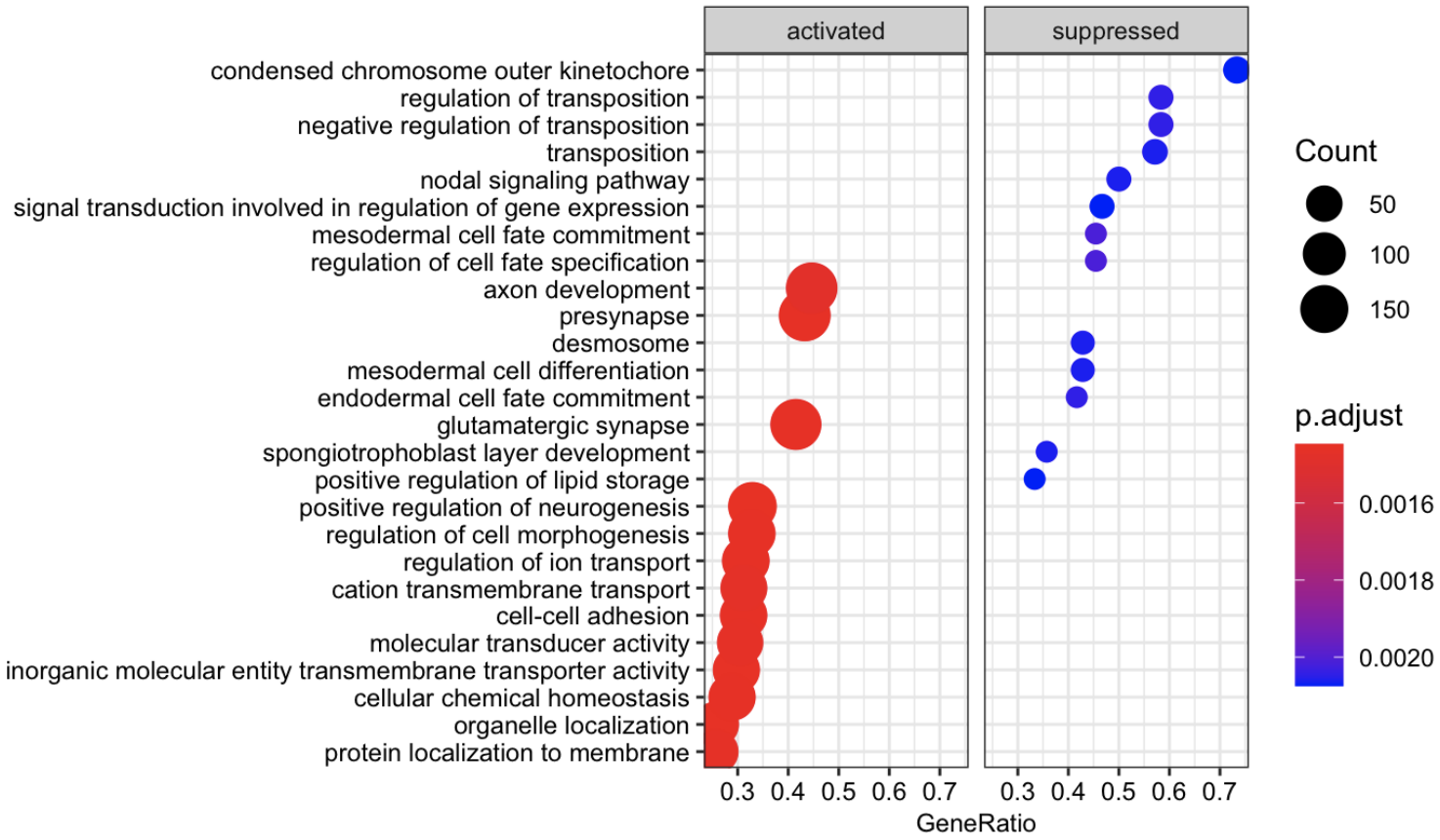

Figure 16. Top differentially expressed GO gene sets in Δ5|6 motor neurons relative to embryonic stem cells. **Left**: GO gene sets significantly upregulated. **Right**: GO gene sets significantly downregulated.

Discussion

These results demonstrate that CTCF binding at the 5|6 locus in the *Hoxa* cluster is necessary to establish the observed boundary between the rostral and caudal *Hoxa* chromatin landscapes in motor neurons. This boundary is flanked by a clearance of H3K27me3 on the rostral side and retention of H3K27me3 on the caudal side. The pattern of gene expression with the *Hoxa* cluster is also determined by the boundary established by the CTCF bound to the 5|6 locus. This directly addresses the experimental question of this study, and as a result we reject the null hypothesis and accept the alternative hypothesis that deletion of *Hoxa5|6* CTCF binding site will significantly change the pattern of H3K27me3 in the *Hoxa* cluster and gene expression.

CRISPR/Cas9-mediated deletion of the 5|6 CTCF binding site in the *Hoxa* cluster led to the loss of CTCF binding at the locus between *Hoxa5* and *Hoxa6* as well as the locus between *Hoxa6* and *Hoxa7*. As a result, the CTCF between *Hoxa7* and *Hox9* became the rostral-most CTCF, establishing a new chromatin boundary between the rostral and caudal landscapes in the *Hoxa* cluster. This was mirrored by a spread of H3K27me3 clearance, typically restricted to the region between *Hoxa1* and *Hoxa6*, through *Hoxa7* upon differentiation of embryonic stem cells into motor neurons.

Loss of the facultative chromatin marker H3K27me3 was coupled with a significant increase in *Hoxa7* gene expression. Pathways affected by the upregulation of

the *Hoxa7* transcription factor were revealed by GO gene set enrichment analysis. Many of these pathways are involved in ribosome synthesis and function, suggesting a possible difference in translation levels in Δ5|6 motor neurons relative to *wt.* GSEA revealed that GO gene sets associated with neuronal development were upregulated in both *wt.* and Δ5|6 motor neurons. This is consistent with the observation that Δ5|6 embryonic stem cells retain their ability to differentiate into motor neurons[6]. These data contribute to the understanding of the mechanisms and regulation underlying motor neuron differentiation.

The primary limitations of this study were related to the small datasets and the limited number of replicates for each of the sequencing conditions. Because of these limitations, the differential gene enrichment analyses were insufficient in their statistical power in some cases. Limma-voom was unable to call any significantly differentially enriched genes, despite many genes exhibiting high logFC of expression between two conditions. Future studies with greater gene sequencing depth and replicates may ameliorate this shortcoming. Another limitation encountered in this analysis was a lack of available input control for the ChIP-seq datasets. This may have impeded the ability of MACS to properly identify broad peaks of enrichment of H3K27me3.

Compared to the original paper that published these datasets, this analysis went further in examining differentially expressed genes and GO gene sets outside of the *Hoxa* cluster across the entire mouse genome[6]. From the DGE and GSEA analyses we were able to identify additional genes that were differentially expressed in Δ5|6 motor neurons relative to *wt.* Further analysis is required to determine if these genes are downstream targets or involved in related pathways as *Hoxa7.* In addition, this analysis confirmed that Δ5|6 motor neurons exhibited many upregulated GO gene sets involved in motor neuron differentiation and function. This confirmed the observation, based on morphology, in the original study that Δ5|6 embryonic cells retained their ability to differentiate into motor neurons[6]. Finally, this analysis utilized both limma and DESeq2 models in gene enrichment analyses, while the original study only utilized DESeq2[6]. The analysis with limma revealed deficiencies in statistical power due to poor sequencing depth and the limited number of replicates that were not addressed in the original paper.

While the data presented in this analysis largely met expectations, further analysis is required to eliminate the possibility of alternative interpretations for some of the results. Although there was a significant reduction in H3K27me3 at the *Hoxa7* locus in Δ5|6 motor neurons, the boundary was not as sharp as the boundary in *wt.* motor neurons. This locus in Δ5|6 motor neurons retained approximately 50% of H3K27me3 enrichment relative to *wt.* This suggests that although CTCF is necessary to establish the boundary between rostral and caudal regions in the *Hoxa* cluster, it is not sufficient to establish this boundary. Furthermore, *Hoxa9* was also determined to be upregulated in Δ5|6 motor neurons, despite the strong enrichment of CTCF at the 7|9 locus. This suggests that the upregulation of gene expression may extend further into the caudal region of the *Hoxa* cluster than does the clearance of H3K27me3, past the rostral-most CTCF. Because of these results, we are unable to reject the alternative interpretation that CTCF is not sufficient to establish a boundary of H3K27me3 or differential gene expression in the *Hoxa* cluster.

These data establish a relationship between the pattern of CTCF binding in the *Hoxa* cluster, the delineation of H3K27me3 between the rostral and caudal regions of the *Hoxa* cluster, and expression patterns of the *Hoxa* genes in differentiating motor neurons.

Loss of CTCF at the 6|7 locus as a result of the deletion of the 5|6 locus suggests an interaction between the CTCF transcription factors that bind to these loci in *wt.* motor neurons. Because CTCF is involved in establishing and delimiting three-dimensional topologically associated domains, a future analysis of chromosome conformation capture (Hi-C) data in embryonic stem cells and motor neurons would be constructive in further understanding the physical DNA interactions that underlie the epigenetic phenomenon of H3K27me3 clearance in the rostral region of the *Hoxa* cluster. Furthermore, CRISPR/Cas9-mediated deletions and subsequent immunoprecipitation and transcriptomic analysis of the CTCF binding sites between *Hoxa7*and *Hoxa9* as well as between *Hoxa10* and *Hoxa11* would lead to a more complete understanding of the function of CTCF binding throughout the entire *Hoxa* cluster. Additional ChIP-seq studies, with antibody pulldowns for RNA polymerase II and the chromatin remodelers Polycomb repressive complex 2 (PRC2) and Trithorax-group (TrxG), would provide further insight into the dynamics of *Hoxa* gene regulation during motor neuron development. These studies may lead to insight into the mechanisms and function underlying how the *Hox* genes affect motor neuron differentiation and development. Such research will aid in the efforts to understand the disease mechanisms underlying motor neuron diseases and promote the development of relevant treatments and therapies.

References

1. Cell Reprogramming: The Many Roads to Success. Aydin B., Mazzoni, E.O., Annual Review of Cell and Developmental Biology 2019 35:1; 433-452.

2. HOX genes: seductive science, mysterious mechanisms. Lappin TR, Grier DG, Thompson A, Halliday HL. Ulster Med J. 2006 Jan;75(1):23-31.

3. Saltatory remodeling of Hox chromatin in response to rostrocaudal patterning signals. Esteban O. Mazzoni, Shaun Mahony, Scott McCuine, Timothy W Danford, Christopher Reeder, P Alexander Rolfe, Robin D. Dowell, Seraphim Thornton, Laurie A Boyer, Thomas M. Jessell, Richard A Young, David K Gifford and Hynek Wichterle-.Nat Neurosci. 2013 Sep;16(9):1191-8.

4. A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. Velasco S, Ibrahim MM, Kakumanu A, Garipler G, Aydin B, Al-Sayegh MA, Hirsekorn A, Abdul-Rahman F, Satija R, Ohler U, Mahony S, Mazzoni EO. Cell Stem Cell. 2017 Feb 2;20(2):205-217.e8.

5. CTCF-mediated topological boundaries during development foster appropriate gene regulation. Narendra V, Bulajić M, Dekker J, Mazzoni EO, Reinberg D.Genes Dev. 2016 Dec 15;30(24):2657-2662.

6. CTCF establishes discrete functional chromatin domains at the Hox clusters during Differentiation. Varun Narendra, Pedro P. Rocha, Disi An, Ramya Raviram, Jane A. Skok, Esteban O. Mazzoni, Danny Reinberg.Science. (2015) Feb 27;347(6225):1017-21. Datasets: https://www.ncbi.nlm.nih.gov/sra?term=SRP045359

7. Evolving Hox Activity Profiles Govern Diversity in Locomotor Systems .Heekyung Jung, Esteban O. Mazzoni, Natalia Soshnikova, Olivia Hanley, Byrappa Venkatesh, Denis Duboule, Jeremy S. Dasen. (2014). Dev. Cell. Apr 28;29(2):171-87.

8. FastQ Screen: A tool for multi-genome mapping and quality control. Wingett SW, Andrews S. F1000Res. 2018 Aug 24 [revised 2018 Jan 1];7:1338.

9. AdapterRemoval: easy cleaning of next-generation sequencing reads. Lindgreen S. BMC Res Notes. 2012;5:337. Published 2012 Jul 2.

10. Fast gapped-read alignment with Bowtie 2. Langmead B, Salzberg SL. Nat Methods. 2012;9(4):357-359. Published 2012 Mar 4.

11. Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. Broad Institute. Available online at: https://broadinstitute.github.io/picard/.

12. The Sequence Alignment/Map format and SAMtools. Li H, Handsaker B, Wysoker A, et al. Bioinformatics. 2009;25(16):2078-2079.

13. deepTools2: a next generation web server for deep-sequencing data analysis. Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. Nucleic Acids Research (2016).

14. Model-based analysis of ChIP-Seq (MACS). Zhang Y, Liu T, Meyer CA, et al. Genome Biol. 2008;9(9):R137.

15. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller Bioinformatics (2016).

16. HISAT: a fast spliced aligner with low memory requirements. Kim, D., Langmead, B. & Salzberg, S. Nat Methods 12, 357–360 (2015).

17. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads, Yang Liao, Gordon K. Smyth, Wei Shi. Nucleic Acids Research.

18. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Ritchie ME, Phipson B, Wu D, et al. Nucleic Acids Res. 2015;43(7):e47.

19. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Law, C.W., Chen, Y., Shi, W. et al. Genome Biol 15, R29 (2014).

20. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Love, M.I., Huber, W. & Anders, S. Genome Biol 15, 550 (2014).

21. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Robinson MD, McCarthy DJ, Smyth GK. Bioinformatics. 2010;26(1):139-140.

22. DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis. G Yu, LG Wang, GR Yan, QY He. Bioinformatics 2015, 31(4):608-609.

23. Grb-IR: a SH2-domain-containing protein that binds to the insulin receptor and inhibits its function. Liu F, Roth RA (1995). Proc. Natl. Acad. Sci. U.S.A. 92 (22): 10287–91.

24. LOT1 (ZAC1/PLAGL1) and its family members: mechanisms and functions. Abdollahi A. J Cell Physiol. 2007 Jan;210(1):16-25.

Acknowledgements

The authors of this paper would like to thank Professor Manpreet Katari, Professor Brian Paker, and Nagashree Srinidhi for their dedicated teaching and mentorship throughout this semester. The course was both extremely interesting and an excellent learning experience. We both look forward to using and continuing to build on the skills and knowledge that we acquired in our future careers as scientists. Thank you, and have a pleasant and safe summer!

Cora Hyun Jung & Christopher Catalano