Exploring the Human Genome for Functional Non-Coding Sequences and Variation: Implications for Understanding Peripheral Nerve Biology

By

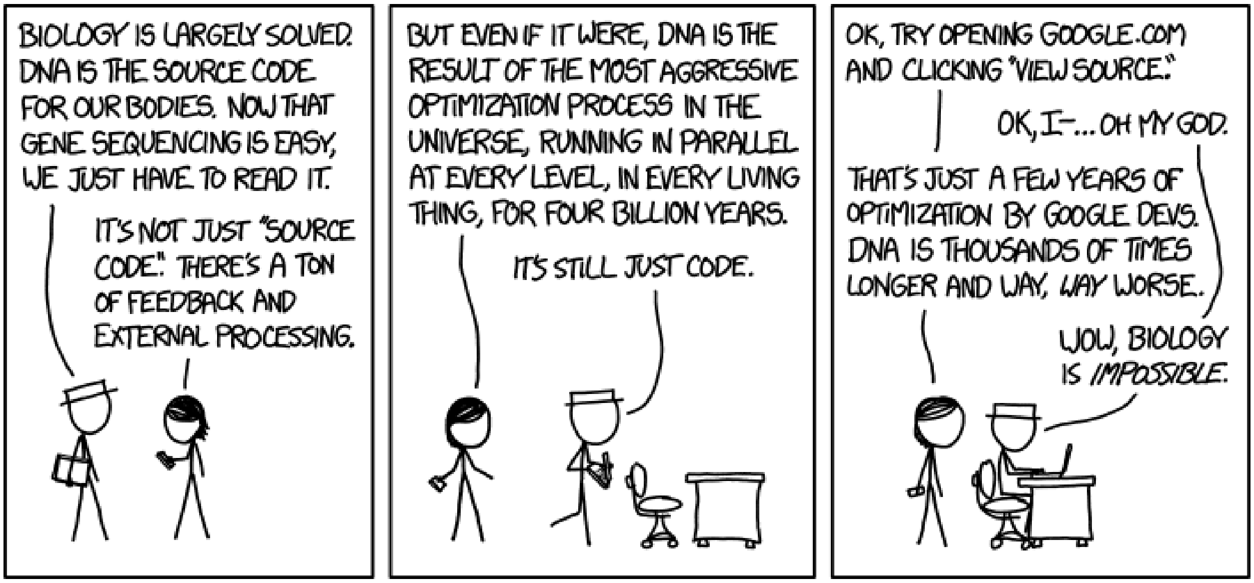
William David Law

**A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in the University of Michigan
2016**

Doctoral Committee:

**Associate Professor Anthony Antonellis, Chair
Associate Professor Scott Barolo
Professor Sally A. Camper
Professor Jeffrey W. Innis
Assistant Professor Jeffrey Kidd**

XKCD 1605: DNA


BIOLOGY IS LARGELY SOLVED. DNA IS THE SOURCE CODE FOR OUR BODIES. NOW THAT GENE SEQUENCING IS EASY, WE JUST HAVE TO READ IT.
IT'S NOT JUST "SOURCE CODE." THERE'S A TON OF FEEDBACK AND EXTERNAL PROCESSING.
BUT EVEN IF IT WERE, DNA IS THE RESULT OF THE MOST AGGRESSIVE OPTIMIZATION PROCESS IN THE UNIVERSE, RUNNING IN PARALLEL AT EVERY LEVEL, IN EVERY LIVING THING, FOR FOUR BILLION YEARS.
IT'S STILL JUST CODE.
OK, TRY OPENING GOOGLE.COM AND CLICKING "VIEW SOURCE."
OK, I—... OH MY GOD.
THAT'S JUST A FEW YEARS OF OPTIMIZATION BY GOOGLE DEVS. DNA IS THOUSANDS OF TIMES LONGER AND WAY, WAY WORSE.
WOW, BIOLOGY IS IMPOSSIBLE.

Title Text: Researchers just found the gene responsible for mistakenly thinking we've found the gene for specific things. It's the region between the start and the end of every chromosome, plus a few segments in our mitochondria.

Munroe, Randall Patrick. "DNA." XKCD. 18 Nov. 2015. <http://xkcd.com/1605/>.

2016

To my loving wife Ashley,

Without you, none of this would be possible.

Acknowledgements

The work presented in this thesis could not have been accomplished without the help and support of many people. I would first like to thank my mentor Tony Antonellis for all the help and guidance throughout the years. Tony came personally recommended from my previous rotation mentor, and upon meeting him, I knew I wanted to perform my thesis research in his lab. We always joke that I never really performed a rotation but mostly just showed up, and to an extent that is true. I believe we only meet a few times in person before Tony was busy with travel and a decision to join his lab had to be made. His exact words (from an e-mail) were, “As far as you joining the lab ... I'm really glad that you're enjoying it and I would be really happy if you want to stay on with us. It's been fun working with you and I think we can carve out at least a few interesting projects. Welcome aboard!” All I can say after these years is that it truly has been fun, and I have thoroughly enjoyed working with you and the lab.

Tony has been the perfect mentor for me. He is able to carefully balance supervision with a “hands-off” approach. Throughout my time here, I’ve always felt like the projects and ideas I worked on were my own but could always get support or help whenever I needed it. I was not only allowed but encouraged to take risks and do things which no one else in the lab (or Michigan) had, such as learning to program in Perl, analyzing RNA-Seq data, and performing CRISPR in S16 cells. I don’t often say it, but I truly thank you for everything you have done for me and hope to one day emulate both the lab environment you created and your mentoring style.

I would also like to thank my committee members, Scott Barolo, Sally Camper, Jeff Innis, and Jeff Kidd. While the scheduling was sometimes challenging, the guidance and suggestions were critical to the success of my thesis. You have all helped shape not only the project but myself as a scientist, and for that I am grateful.

In addition to formal support, I have been fortunate to receive external guidance from many faculty members. I would like to thank John Moran for both scientific and (often needed) comical discussions. In addition, his lab supplied many reagents and equipment that allowed me to perform challenging experiments. I would also like to thank both Miriam Meisler and Jacob Kitzman (and their labs) for vital comments and suggestions at our weekly joint lab meetings. In particular, Jacob provided the necessary guidance for the CRISPR experiments, and I owe much of the success of those experiments to him. I would also like to thank our close collaborators John Svaren and Greg Crawford for sharing both their time and data with us.

In addition to faculty members, I would like to thank our lab. It has been a pleasure to work with each of you for many different reasons. Heather McLaughlin and Megan Brewer welcomed me not only into the lab, but into Ann Arbor, and taught me basic science techniques and where the best happy hours are. Chetna Gopinath and Elizabeth Fogarty have been instrumental in the completion of this work. I have greatly enjoyed having peers working on similar projects with whom I can freely share ideas, and I look forward to seeing both of your projects flourish. I would also like to thank Stephanie Oprescu for her willingness to do any task necessary to help my project succeed. Finally, I would like to thank Laurie Griffin for countless discussions and stress-relieving coffee breaks. I am very fortunate to have had exceptional colleagues and friends to share this adventure with, and I look forward to your continued success in the future.

While graduate school requires a great deal of work, my friends have ensured a balance of leisure. To my college friends Ryan Slinger, Michael McMahon, and Luke Griffith for all the sporting and gaming events, you have kept me humble and laughing. To my hockey team who brought me in even though I had no idea how to play or skate. Finally, to my graduate school friends who have had a bigger impact than they know on both myself and this work. I would especially like to thank Susan and Peter Stamats and the entire balloon crew, specifically Bill McClelland and Dick Frye. I cannot truly describe the impact you all have had on my life. You have been like my second family, and I would not be the person I am today without all the support and guidance. Some of my most treasured memories are from the countless ballooning events I was allowed to join, and I am forever grateful to all of you.

Finally, I would like to thank my family. My parents and sister who have been nothing but supportive of any decision I have made. From driving me to all the sporting and music practices and games to teaching me the value of an education, I am truly thankful for everything you have done for me. To my in-laws, I thank you for welcoming me into your lives and treating me as though I had always been a part of the family. I am truly fortunate to be a part of your lives. Last, but not least, I would like to thank my wife Ashley. Graduate school is challenging for the person experiencing it, but it is twice as difficult for their spouse. From the bad days to the successful experiments, your unwavering support has been essential to completing this work. You are truly the unsung hero of this story, the Sam to my Frodo. I only hope you recognize this is our accomplishment, despite my name being on the diploma. Thank you for everything you have done, and I look forward to a lifetime of adventures with you.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

The vast array of cells and tissues in the human body contain nearly identical genetic information, yet each tissue expresses only a subset of genes present in the genome. The precise spatiotemporal expression patterns of these genes are regulated by transcription factors, which bind to short sequences of DNA called transcription factor binding sites (TFBSs). While it is well-established that mutations within protein-coding sequences cause human disease, sequence variation within TFBSs (regulatory SNPs; rSNPs) can result in allele-specific differences in DNA binding affinity, which can also cause or modify disease phenotypes. Our objective was to identify rSNPs that impact regulatory function at loci important for peripheral nerve function. Such rSNPs represent excellent candidate modifier loci that may explain phenotypic variability in patients with peripheral neuropathies.

A challenge in studying the effects of rSNPs on gene function is absent or incomplete catalogs of TFBSs. To address this, we developed a computational and functional pipeline to identify and characterize putative TFBSs. We utilized genome-wide sequence conservation to prioritize candidate regulatory regions that harbor a SNP. We assessed a pilot set of 159 regions on chromosomes 21, 22, and X for regulatory activity in cells relevant for the peripheral nerve. We identified 28 active regions, of which 13 showed allele-specific differences in regulatory activity. We next incorporated known transcription factor binding site information into our pipeline. The transcription factor SOX10 is essential for Schwann cell function and has a well-characterized consensus site. We assessed the allele-specific activity of 22 prioritized regions that contain a conserved SOX10 consensus site overlapping a SNP. We deeply characterized one region and identified a candidate target gene: *Tubb2b*. Finally, we performed an unbiased search for conserved SOX10-response elements that revealed a previously undescribed potential function for SOX10 in repressing myelination. Importantly, the approach and datasets described here are broadly applicable to studies on SOX protein and Schwann cell biology, regulatory function in the peripheral nerve, and the function of highly conserved sequences in the human genome.

CHAPTER 1

Introduction

Methods to Identify Cis-Regulatory Elements

Patients with peripheral neuropathies (specifically Charcot-Marie-Tooth disease) exhibit large phenotypic variability such as age of onset from the first to seventh decade and severity of sensory loss, often despite identical coding mutations (Thomas et al. 1997; Pareyson et al. 2006; Pareyson and Marchesi 2009). While the cause of the variability is unknown, genetic modifiers in the form of regulatory SNPs may be involved. To this end, the main focus of the thesis work presented here is to identify functional, non-coding variation critical for the peripheral nerve. Throughout this thesis work I will employ a variety of *in silico*, *in vitro*, and *in vivo* experiments to address this aim. In this chapter, I will discuss these techniques, the advantages and limitations of each, and how combinations of these experimental methods can lead to the identification of *cis*-regulatory elements (CREs). In the second half of this chapter, I will discuss a transcription factor SOX10 and the critical function it performs in Schwann cells. The combination of experiments discussed here to identify CREs will be employed with SOX10 binding site information to identify functional SOX10 response elements harboring SNPs which are candidate genetic modifiers of peripheral nerve disease.

Cis-Regulatory Elements

Cis-regulatory elements (CREs) are regions of the genome that regulate the spatial-temporal expression of a gene (or multiple genes). CREs are generally short stretches of DNA that allow for transcription factors to bind. These have been shown to be critical for nearly all cellular processes in almost every species, ranging from humans to bacteria and viruses. CREs can have varying effects on gene function including increasing (enhancers) or decreasing (repressors) gene expression or preventing ectopic CRE activity (insulators). In most cases, for enhancers and repressors, they exert their effects on the promoter element (Figure 1.1 - Adapted from (Noonan and McCallion 2010)). Identification of promoters is relatively easy because the core promoter sequences are generally known (*e.g.*, TATA box and initiator elements), sequences are frequently located immediately upstream (with the exception of downstream promoter elements) of a gene, and usually there is only one promoter per isoform of a gene (Pedersen et al. 1999; Sandelin et al. 2007). Conversely, it is difficult to identify CREs because there are many different transcription factor binding sites, they can exert their effects over great distances, and can reside anywhere in the genome: upstream of a gene, downstream of a gene, or within introns (or exons) of the target gene or other genes. There can also be many CREs per isoform (Vyas et al. 1995). In this chapter, I will discuss a brief history of eukaryotic CREs, methods to identify and functionally assess CREs, their effects on human diseases, and major questions still unresolved in the field.

The first CRE described in detail was in 1981 by Banerji and colleagues (Banerji et al. 1981). In this paper, they transfected HeLa cells with the rabbit β-globin gene, but the transient transfection alone was insufficient to detect rabbit β-globin transcripts; however, upon co-transfection of a plasmid harboring the SV40 viral DNA with the plasmid expressing the rabbit

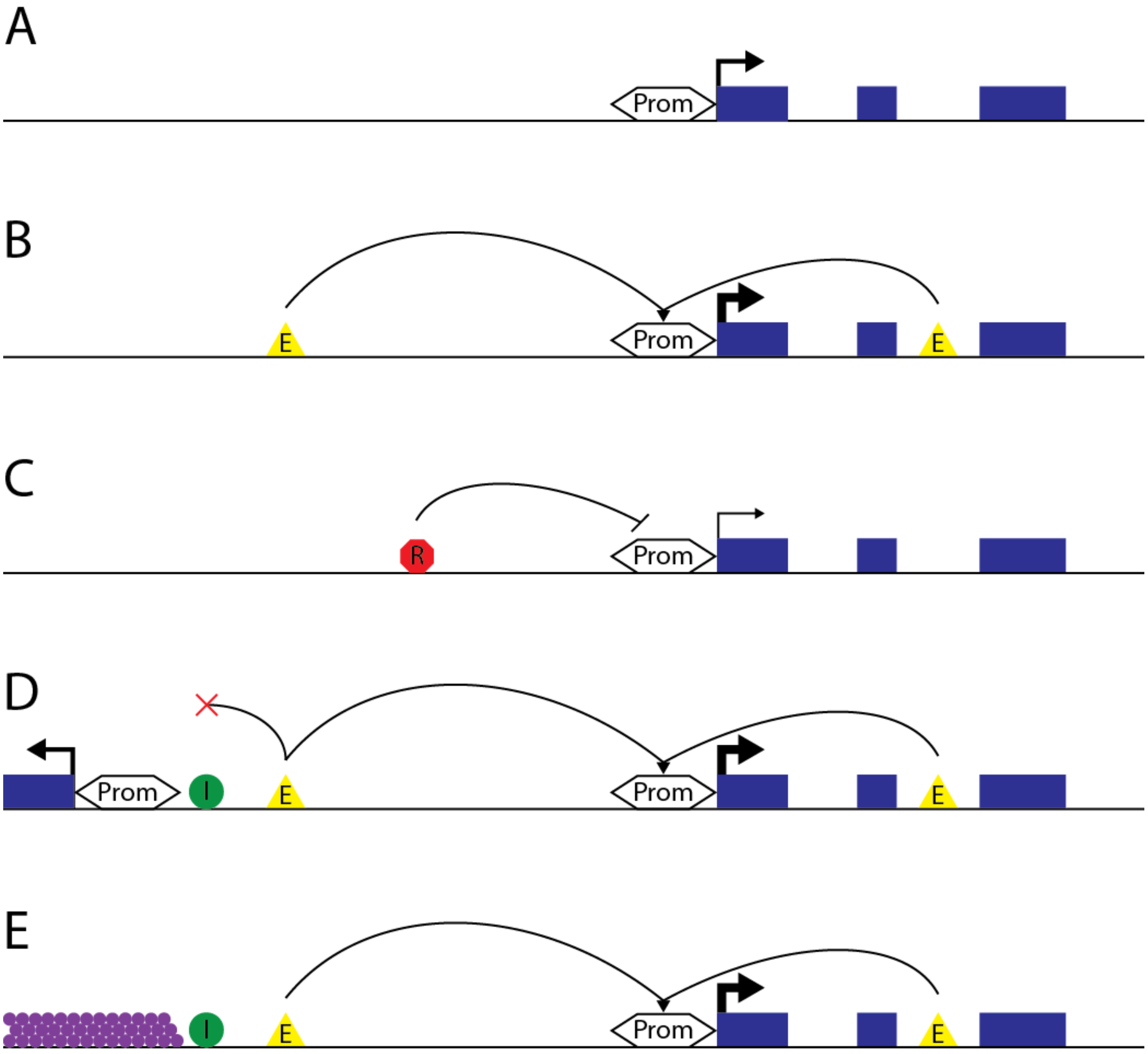


Figure 1.1 *Schematic of the General Function of Cis Regulatory Elements (CREs).* (A) The function of single promoter acting on one gene leads to a certain level of gene expression. (B) Two enhancers, one upstream and one in the intron on the gene, act on one promoter to ultimately increase gene expression. (C) One repressor acts on the promoter to decrease gene expression. (D and E) The effects of an insulator on gene regulation. (D) An insulator element prevents ectopic enhancer activity on a second gene. (E) An insulator element prevents the spreading of heterochromatin to maintain enhancer activity. Blue boxes represent exons, arrows indicate the transcription start site, and size of the arrows reflects gene expression (thicker = higher, thinner = lower). Prom = promoter, E = enhancer, R = repressor, I = insulator, and purple circles = heterchromatin. Figure adapted from Noonan & McCallion, 2010.

β-globin gene, the authors were able to detect 200 times higher transcript levels. They went on to map the region of SV40 that gave the "enhancer effect." The authors narrowed down the region of the SV40 sequence by creating random deletions and reassessing each partially deleted construct in the co-transfection assay. These experiments have since been termed promoter bashing, and while the methods for detecting enhancer activity has changed (*e.g.*, *CAT, LacZ,* or luciferase reporter constructs [discussed below]), this type of assay has been used extensively to characterize many CREs (Kong et al. 1999; Guo et al. 2002; Kobolak et al. 2009). These experiments gave rise to the term 'enhancer' and gave birth to a new field of study in genetics: gene regulation.

Banerji and colleagues went on to identify the first human enhancer at the immunoglobulin heavy chain locus (Banerji et al. 1983), by performing nearly identical experiments to characterize this enhancer as were previously performed on the SV40 enhancer. One experiment involved cloning the enhancer fragment into various locations (*e.g.*, upstream and downstream) of the vector relative to the rabbit *β-globin* gene reporter. They also altered the orientation of the enhancer and in both cases observed a similar level of expression. This lead the authors to conclude that enhancers generally exhibit both location and orientation independence relative to the transcription start site of a gene (Banerji et al. 1983). Since these initial observations, one of the main questions and challenges in the field has been to identify and functionally evaluate CREs.

DNase I Footprinting

One of the earliest methods employed to identify CREs via protein-DNA interactions is DNase I footprinting. In this experiment, DNA is subjected to the enzyme DNase I that cuts exposed DNA; however, regions of DNA that are bound by proteins are generally protected from DNase I activity and leave a ‘footprint’ when the reaction is run out on a polyacrylamide gel (Galas and Schmitz 1978; Carlberg et al. 1988). In the case of CREs, the DNA will be protected by the transcription factor binding the DNA. Perhaps the most well studied CRE that made use of DNase I footprinting is the *β-globin* locus control region (LCR) (Grosveld et al. 1987). The locus control region consists of multiple CREs that regulate the expression of related genes in the appropriate tissue and to physiological levels. For the *β-globin* gene cluster, the LCR regulates the differentially expressed globin genes to ensure proper expression at multiple stages of development. DNase I experiments were able to originally identify the LCR in erythroid cells and subsequent experiments using transgenic mice demonstrated that this region was necessary for endogenous gene expression patterns (Levings and Bungert 2002). This experimental procedure is extremely successful in identifying open regions of the genome but is limited by the number of regions in the genome that can be assessed.

A new genome-wide method termed DNase-Seq (Crawford et al. 2004; Sabo et al. 2004) has been developed to assess open chromatin patterns within the entire genome. DNase-Seq uses the same theory as DNase I footprinting, however the method leverages next-generation sequencing technologies to assess the whole genome in one experiment. Additionally, DNase-Seq is currently one of the most effective single experiments used to predict CREs (Kwasnieski et al. 2014), yet combining multiple genomic datasets (*e.g.* ChIP-Seq, GC content, and transcription factor motifs) can provide higher confidence predictions of CREs and more easily separate active from inactive transcription factor binding sites. In addition to DNase-Seq, other experimental procedures have been employed to search for open chromatin: ATAC-Seq (Buenrostro et al. 2013), FAIRE-Seq (Giresi et al. 2007), and MNase-Seq (Schones et al. 2008). The experiments used to identify open chromatin patterns genome-wide has lead to a plethora of novel putative CREs.

One of the drawbacks of using DNase I footprinting (or DNase-Seq) is the inability to predict what protein is binding to the DNA, as the protein typically is generated from protein lysates of cells or tissues of interest. This problem is being partially alleviated with increased sequencing depth of DNase-Seq experiments which allow high-resolution detection of the transcription factor binding to DNA (Neph et al. 2012). These high-resolution maps demonstrate, on the single base pair level, how a specific transcription factor uniquely binds to DNA. Curating a database of the unique interactions may eventually allow DNase-Seq experiments to also predict the transcription factor binding to the DNA; however, these may be limited by only predicting high affinity binding sites while excluding functionally relevant low affinity binding sites (Rowan et al. 2010; Parker et al. 2011; Ramos and Barolo 2013).

Chromatin Immunoprecipitation (ChIP)

Compared to DNase-Seq that identifies occupied DNA in an unbiased manner, chromatin immunoprecipitation (ChIP) relies on physically crosslinking proteins with the DNA sequences that they bind to and isolating the protein (and bound DNA) of interest with an antibody specific to the protein (*e.g.*, transcription factor) of interest. One of the first experiments describing ChIP was performed by Gilmour and Lis (Gilmour and Lis 1985) to identify RNA-polymerase II occupancy on the *Hsp70* gene in *Drosophila* Schneider line 2 cells. They were able to successfully demonstrate RNA-polymerase II binding throughout *Hsp70* when cells were induced by heat shock but binding was confined to the 5’ end of the gene in uninduced cells. From this early work, it became possible to identify *in vivo* interactions between CREs and the transcription factor of interest; however, traditional ChIP experiments required prior knowledge of both the transcription factor and the predicted CREs being studied.

The advent of ChIP-chip experiments alleviated the need for prior knowledge of CREs. This technique combines ChIP experiments with DNA microarrays into a single method. In ChIP-chip experiments, a standard ChIP assay is performed, but instead of analyzing known sites via labeled probes or PCR, the enriched DNA is hybridized to microarrays (Singh-Gasson et al. 1999; Ren et al. 2000). For organisms with smaller genomes, one chip could contain tiling probes against nearly the entire genome (Ren et al. 2000; Iyer et al. 2001; Zeitlinger et al. 2007). Conversely, for organisms with larger genomes, only selected target regions could be examined (Kim et al. 2005; Akerfelt et al. 2008). In addition, microarrays generally only cover the non-repetitive regions of the genome (Kim et al. 2005). It is known, however, that CREs can also reside within repetitive regions (Bourque et al. 2008), and this signal will be lost using the ChIP-chip method. While ChIP-chip was a vast improvement over traditional ChIP, it left many CREs unidentified.

The newest advancement with ChIP technologies is ChIP-Seq (Barski et al. 2007; Johnson et al. 2007), which relies on next-generation sequencing technologies, similar to DNase-Seq. In contrast to ChIP-chip, ChIP-Seq does not require predefined probes to capture the bound DNA fragments. Rather, the enriched DNA is subjected to linker ligation, amplified using PCR, and sequenced using next-generation sequencing. This method solves many of the limitations of

ChIP-chip, and as such it has become pervasive in the identification of CREs, with thousands of studies utilizing ChIP-Seq experiments. This includes the ENCODE consortium (discussed below) which has conducted hundreds of ChIP-Seq experiments across multiple cell and tissue types (ENCODE Project Consortium 2012; Landt et al. 2012).

The major limitation to all ChIP experiments is the reliance on antibodies as some transcription factors do not have any effective antibodies. One of these issues is the lack of antibodies for every transcription factor. This can be resolved by expressing a tagged version of the transcription factor, however the tagged version must be carefully examined to ensure identical functions to the endogenous, untagged transcription factor. In addition, the level of expression of the tagged version must be controlled because overexpression of transcription factors could alter the occupancy of sites, potentially confounding the interpretation of results (DeKoter and Singh 2000; Fernandez et al. 2003). Traditionally, bacterial artificial chromosomes (BACs; discussed below) were used to conserve as many of the endogenous CREs as possible to express the tagged transcription factor at approximately endogenous expression levels (Hua et al. 2009), however newer methods have been implemented to tag the endogenous locus directly (Savic et al. 2015). Another problem with antibodies is they need to be highly specific. Variation in antibody quality can vary among different batches (even from the same company), and careful control experiments are necessary to prevent confounding results (Landt et al. 2012). Despite the potential limitations of ChIP-Seq and related technologies, these methods have proven extremely successful at demonstrating where transcription factors bind within a genome, ultimately leading to the identification of novel CREs.

Phylogenetic Footprinting

Another method used to identify CREs is phylogenetic footprinting. Phylogenetic footprinting was first used to predict CREs that were necessary for ε and γ globin expression in primates (Tagle et al. 1988). This method relies on the alignment of multiple species sequences at orthologous regions, supported by the theory that conserved non-coding regions, even among divergent species, may be evolutionarily constrained due to necessary function (Hardison 2000). This method proved successful in identifying CREs near specific genes of interest (Aparicio et al. 1995; Loots et al. 2000; Nobrega et al. 2003; Antonellis et al. 2008). A major drawback for this approach was the lack of whole genome sequencing data for multiple species.

This problem was resolved in the early 2000s, when the genomes of multiple species became available, starting with the completion of the human genome project in 2001 (Lander et al. 2001), and the mouse genome project in 2002 (Waterston et al. 2002). With these sequences, as well as those from other species (Gibbs et al. 2004; Hillier et al. 2004), CREs could be identified for the first time using genome-wide phylogenetic footprinting, rather than short alignments of homologous regions near target genes. Many different approaches were used to try to identify CREs, including increased time of divergence between species (Aparicio et al. 1995) and length of the conserved region (Bejerano et al. 2004).

One particular example that demonstrated the strength of phylogenetic footprinting and increasing divergence between species was shown by Nobrega and colleagues at the *Dach* locus (Nobrega et al. 2003). *Dach* is flanked on both the 5’ and 3’ sides by gene deserts, each approximately one million base pairs in length. Alignment of the human and mouse genomes revealed 1,098 regions that were at least 100 base pairs in length and with greater than 70% conservation. The authors were able to narrow down these regions to 32 conserved sequences when including frog, zebrafish, and two pufferfish genomes in the phylogenetic footprinting experiments. Of the 32 regions assessed, the authors tested nine regions in a transgenic mouse assay (discussed below) and reported that seven of them were able to at least partially recapitulate native *Dach* expression (Nobrega et al. 2003). In addition to the conservation approach, this paper was one of the first to illustrate the great distances CREs can reside from the gene(s) on which they act.

While phylogenetic footprinting has been used successfully many times (Zerucha et al. 2000; Goode et al. 2003; Kimura-Yoshida et al. 2004), it is critical to functionally evaluate these predictions, as sequence constraint does not guarantee conserved functions. One example was demonstrated with ultraconserved elements (UCE), which are regions of the genome that are 100% identical in sequence between human, mouse, and rat and at least 200 base pairs in length (Bejerano et al. 2004). These particular UCEs tended to cluster near genes involved in RNA processing, development, or transcriptional regulation, but functional evaluation of 84 UCEs using transgenic mouse reporter assays demonstrated that only 51 of them could direct *LacZ* expression (Pennacchio et al. 2006), leaving 33 UCEs without a visable function. It is important to note, the 33 nonfunctional UCEs may have had function at different developmental timepoints than those assessed by the authors. In another example, four UCEs were deleted *in vivo* from mice (Ahituv et al. 2007), and the resulting mice were viable and had no overt unusual phenotypes. While it is possible the mice could have some phenotype below the level of detection, this does illustrate the fact that conservation (even extreme levels) does not necessarily reflect function.

Similarly, using phylogenetic footprinting can result in false negatives due to conserved functionality without sequence conservation (Fisher et al. 2006; McGaughey et al. 2008). In one case, CREs identified at the human *RET* locus were able to functionally recapitulate appropriate expression patterns in zebrafish, despite lacking sequence conservation between human and zebrafish (Fisher et al. 2006). Some potential explanations given by the authors for this result are small changes within the transcription factor binding sites, coevolution of the transcription factor and the binding site, or rearrangement of individual binding sites within a CRE. Some of these issues may be resolved utilizing stringent conservation over short sequence stretches (*e.g.* tens of base pairs or less) rather than arbitrary conservations thresholds over large distance (*e.g.* 70% conservation over hundreds of base pairs). Indeed, studies of *Drosophila* enhancer evolution demonstrated similar confounding issues facing phylogenetic footprinting (Ludwig et al. 2000; Berman et al. 2004).

Transgenic Animal Models

While accuracy in predicting CREs has been increasing through a variety of methods, the need for functional assessment is clear. One of the earliest methods for functionally assessing CRE transcriptional activity involved fusing upstream flanking DNA near a gene of interest to a reporter cassette, often *LacZ* or *GFP*. These reporter genes could then be injected into a developing embryo such as mouse (DiLeone et al. 1998) or *Drosophila* (Rubin and Spradling 1982; Stanojevic et al. 1991; Malicki et al. 1992), allowed to stably integrate randomly into the genome, and assessed for spatial and temporal reporter gene expression patterns. These experiments were frequently coupled with promoter bashing experiments to try to identify the required base pairs within a CRE(s) (Kong et al. 1999; Guo et al. 2002).

The first experiment to demonstrate successful expression of an exogenous transgene in *Drosophila* laid the groundwork for future studies focusing on CREs (Rubin and Spradling 1982). Some of the earliest work to elucidate the function of CREs was based on transgenic reporter assays in *Drosophila* (Stanojevic et al. 1991; Small et al. 1992). These pioneering studies were able to demonstrate a fundamental concept of CREs: combinatorial control. Combinatorial control refers to the ability of many CREs working together to regulate the expression of a single gene. This concept has been shown for a number of different CREs and has become a defining feature of how CREs establish complex expression patterns (Stanojevic et al. 1991; Small et al. 1992; Ferretti et al. 2005).

One of the major limitations of transgenic reporter assays was the size of the elements that could be tested, generally a few hundred to a few thousand base pairs. This limitation was ameliorated with the development of bacterial artificial chromosomes (BACs). BACs are large (approximately 200 kilobase pairs) stretches of DNA from any organism that can be stably maintained in *E. coli* (O'Connor et al. 1989). Similar to transgenic reporters, BACs can be generated that harbor large portions of a genome, such as human or mouse, which are predicted to harbor CREs and overlap the gene of interest. Once the suspected genomic regions have been cloned into the BACs, a reporter gene can be inserted into the BAC (either downstream or in frame with the gene of interest) which will recapitulate the endogenous gene expression patterns. One of the first uses of BACs to identify CREs was performed by DiLeone and colleagues (DiLeone et al. 2000). Of note, the authors in this study used co-injection of the BACs with a reporter gene plasmid and relied on co-integration of the plasmid at the same site in the genome. The authors used a BAC overlapping the *Bmp5* gene to recapitulate a majority of the expression pattern predicted from previous work using transgenic reporter assays (DiLeone et al. 1998) to describe individual CREs. They went on to study putative CREs in two overlapping BACs and were able to identify novel CREs regulating *Bmp5* greater than 200 kilobase pairs away from the transcription start site. While it was known that enhancers could exert their effects on genes at great distances, this work and studies of a sonic hedgehog enhancer demonstrated the extreme distances (greater than one million base pairs) CREs can be separated from their target genes (Lettice et al. 2002; Lettice et al. 2003).

One of the major drawbacks of these types of transgenic reporter assays is the inability to control the genomic site of integration. Often the injected linearized transgene can integrate at multiple locations in the genome and/or multiple copies can insert into one location in a head-to-tail configuration called concatemers (Wilkie and Palmiter 1987; Hamada et al. 1993; Dai et al. 2010). Additionally, the random nature of the integration can affect the expression of the transgene, such as the transgene integrating into a region of heterochromatin and becoming silenced. These effects have been termed position effects and represent a major problem in the interpretation of the regulatory activity of CREs (Levis et al. 1985).

One method developed to circumvent random integration utilizes homologous recombination to insert a reporter gene into the endogenous locus (Bronson et al. 1996). In this type of experiment, a BAC containing the gene of interest is modified to express an in-frame reporter cassette flanked by unmodified arms of homology. When this modified BAC is injected into an embryo, the arms of homology will allow the BAC to recombine with the wildtype locus to insert the modified version of the gene. The reporter gene is now under control of all the endogenous CREs (Mansour et al. 1990; Guillot et al. 2000). Similar experiments have been performed successfully with yeast artificial chromosomes (Tyas et al. 2006). While single, homology-driven integration experiments can be used to alleviate the random integration problem, the rates of homologous recombination are very low.

A different method to resolve the problem of random integration is site-directed integration, with one of the most common methods employing a modified version of the phage integrase ΦC31 (Thorpe and Smith 1998). This method relies on previously integrated attP sites in the genome of the model organism. A transgene can then be inserted directly into a single attP site, if the plasmid carrying the transgene also carries an attB site, and ΦC31 integrase is expressed. While these initial attP integration sites are random, and subject to the same problems as random integration, many different founders with different insertion sites can be prescreened to determine any detrimental position effects. This method has been used extensively to characterize many CREs in *Drosophila* (Groth et al. 2004; Bischof et al. 2007; Kvon et al. 2014). One of the more exhaustive experiments using this method involved screening 7,705 candidate enhancers (Kvon et al. 2014). The authors placed each putative enhancer upstream of a minimal promoter and used ΦC31 site directed integration to assess for function using a *GAL4* reporter. Of the 7,705 regions tested, 3,557 regions displayed some level of activity during at least one developmental time point (Kvon et al. 2014). This method has also been used to a lesser extent in mice (Tasic et al. 2011).

Luciferase Assays and Next-Generation Sequencing Approaches

While the transgenic animal experiments discussed above identified many CREs, they were limited in a number of ways. Specifically, these experiments were generally focused on a small subset of candidate CREs acting on few target genes. Additionally, they were laborious and expensive. Different methods were devised to increase both the number of regions examined and shorten the length of time for experiments. One of the first *in vitro* methods to functionally assess CRE activity relied on reporter assays. These methods are nearly identical to those described above with a few notable exceptions. First, the transgene is not stably integrated into the genome. The plasmid is transiently expressed, and the cells are assessed shortly after transfection. Second, the type of reporter gene is different. Before, *LacZ* or *GFP* were used to visualize CRE function, while the reporter genes *luciferase* (Wood et al. 1984; Dewet et al. 1985) or *renilla* (Matthews et al. 1977) are generally used to assess regions using this *in vitro* method (Sherf et al. 1996). It is important to note, these assays generally rely on immortalized cells and may not faithfully recapitulate the *in vivo* tissues.

Briefly, the CRE is cloned upstream of *luciferase* (the gene responsible for producing light in fireflies), and the plasmid is transfected into cells. The cells are allowed to grow for a certain (generally 48-72 hours) period of time before being harvested, and the proteins are isolated. If the substrate for luciferase is mixed into the protein lysate, a certain amount of light will be produced which can be detected and quantified by intensity. The amount of light detected for an individual CRE directing *luciferase* is proportional to the amount of luciferase present in the protein lysate. This value is indicative of the strength of the CRE directing expression of the *luciferase* reporter, which may (or may not) be reflective of the strength *in vivo*. These experiments generally include a co-transfected control reporter gene (often *Renilla*) controlled by a standardized CRE (such as CMV) to normalize the activity of *luciferase* to *renilla* and account for different transfection efficiencies and cell viability (Sherf et al. 1996).

The method described above has been used in many different experiments and in many different species (McNabb et al. 2005; Rodda et al. 2005; Antonellis et al. 2008). Despite the prevalence of the use of dual luciferase assays in identifying CREs, there are some limitations, many of which are common between luciferase assay and animal transgenic reporter assay. Similarly, only relatively short sequences can be assessed in luciferase assays, which may result in false positive results. For example, an enhancer being studied may be cloned separated from a repressor, which *in vivo* acts to neutralize the enhancer's effect. Additionally, while the *in vitro* nature of this experiment eludes position effects confounding the results, the CRE on the plasmid does not have endogenous chromatin marks (Reeves et al. 1985; Hebbar and Archer 2008). This can also lead to false positive results, if the CRE is endogenously bound up in heterochromatin and unable to act *in vivo*, but is able to function in the *in vitro* luciferase assays. Another limitation of luciferase assays is the use of a minimal promoter. The aim is to identify a single core promoter motif that is capable of interacting with multiple CREs, since not all promoters can respond to every CRE (Li and Noll 1994; Zabidi et al. 2015).

Despite these limitations, luciferase assays continue to be used frequently to functionally assess CREs. One example exploiting the advantages of luciferase assays was demonstrated by Antonellis and colleagues (Antonellis et al. 2010). In this paper the authors identified a patient variant within a CRE near *myelin protein zero* (*MPZ*). They first used luciferase assays to demonstrate that the wildtype allele induced reporter gene activity in relevant cell lines while the patient variant was significantly less active compared to the wildtype allele. While luciferase assays are by design *in vitro* (and may not reflect *in vivo* regulatory activity), the authors utilized the knowledge gained to assess the region *in vivo* in zebrafish. They observed appropriate expression patterns in the expected tissues, but were unable to conclusively demonstrate that the patient variant altered *MPZ* expression. Nevertheless, this study demonstrated the ability to rapidly assess CREs and putative mutations in relevant cell lines.

While luciferase assays are relatively quick to perform, they are still only able to assess a small number of predictions at one time. This problem has been partially resolved with two similar methods: STARR-Seq (Arnold et al. 2013) and CRE-Seq (Kwasnieski et al. 2012). In these methods, candidate CREs are obtained either by synthetic synthesis of approximately 150 base pair stretches of oligonucleotides (CRE-Seq) or random sheering of genomic DNA followed by size selection (STARR-Seq). The resulting DNA fragments are then cloned into reporter gene plasmids (one CRE per plasmid), and the entire pool of plasmids is transfected into millions of cells. In the case of CRE-Seq, the plasmids contain a unique barcode that is transcribed at the end of the reporter. For STARR-Seq, the CRE is cloned directly within the reporter gene transcript. For both CRE-Seq and STARR-Seq, the cells are harvested, DNA and RNA are isolated, and next generation sequencing of both the DNA and RNA is performed. The ratio of the copies of RNA relative to the copies of DNA can provide a readout of the CRE strength. For CRE-Seq, the barcode and the CRE are matched using the DNA sequencing data. STARR-Seq does not require this step because the RNA already contains the sequence of the CRE within the transcript, and can be used to directly quantify CRE strength.

An additional advantage of both CRE-Seq and STARR-Seq is the ability to test multiple reporter promoter elements, which can help control for promoter-specific affects. One example of this was demonstrated by Zabidi and colleagues (Zabidi et al. 2015), where the authors used STARR-Seq to test an identical set of candidate enhancers with either the *Ribosome protein gene 12* (*RpS12*) core promoter or a synthetic core promoter. The authors were able to demonstrate only 32% of the 9,542 candidate enhancers tested were able to function with either promoter element. The remaining enhancers activated one of the promoter elements at least two fold more than the other promoter (Zabidi et al. 2015).

Electromobility Gel Shift Assays (EMSAs)

While luciferase assays and massively paralleled reporter assays such as STARR-Seq (Arnold et al. 2013) and CRE-Seq (Kwasnieski et al. 2012) continue to help uncover novel CREs, another *in vitro* method used to assess if CREs can bind to a candidate sequence are electromobility gel shift assays (EMSAs) (Fried and Crothers 1981; Garner and Revzin 1981). These assays involve generating a labeled DNA probe (using either radioactivity or more recently using biotinylation) that contains the locus-specific, predicted transcription factor binding site. These probes are then mixed with protein extract from the cell or tissue of interest, the transcription factors are allowed to bind to the probes, the complex is run on a gel, and the probe-transcription factor complex is detected as a high shift relative to unbound probe. The addition of high concentrations of cold (unlabeled and wild-type or mutated) competitor probes are used to demonstrate the specificity of observed DNA:protein interactions. Similar to other methods, the decision about what specific sequence to use can greatly affect the results. EMSAs are more sensitive to this limitation because the probes are much shorter (50-100 base pairs) than regions assessed in other *in vitro* functional assays.

While this traditional form of EMSA can detect proteins interacting with the DNA probe, it was difficult to know what protein is responsible for the interaction. New iterations of EMSAs were developed to alleviate this problem. One method, termed supershift assays, involves the addition of an antibody against the predicted transcription factor (Kristie and Roizman 1986), (Gille et al. 1997). The addition of the antibody against the predicted protein results in a higher shift (supershift) relative to the protein-DNA complex alone. Another potential outcome in supershift assays is that the antibody disrupts the protein’s ability to bind DNA. This results in the loss of a band, rather than the supershift observed in other situations (Ou et al. 2003). A major limitation of supershift assays, similar to ChIP experiments, is that they require highly specific antibodies (see above).

Another variation of EMSA employed to specifically identify the protein of interest uses recombinant protein rather than protein lysate. This results in only one protein being present to bind the probe, and any resulting shift must be due to the recombinant protein (Hunt and Jackson 1974; Zhang et al. 1998). One potential pitfall with using recombinant protein rather than protein lysate is the potential dependence on co-factors. For example, if the recombinant protein requires an additional protein to bind DNA, the EMSA may fail to result in a shift, even though the probe may contain a functional CRE.

General Characteristics of CREs

Through the use of the above-described (and other) functional assays, many CREs have been discovered. These pioneering experiments on CREs have lead to some general guidelines about what defines a CRE. As stated above, CREs can function over great distances, sometimes greater than one megabase pairs away (Lettice et al. 2002; Nobrega et al. 2003). In addition, it has been proposed that CREs can function in orientation-independent fashion relative to the promoter (Banerji et al. 1983; Rogers et al. 1986).

Yet, most of these initial experiments demonstrating CRE orientation independence were performed in an *in vitro* transgenic reporter assay, which may not recapitulate the endogenous activity. Indeed, orientation-dependent enhancer activity has been demonstrated in transgenic mice harboring the wildtype enhancer in either the endogenous or inverted orientations (Swamynathan and Piatigorsky 2002). The authors saw significantly reduced gene expression when the enhancer was in the inverted orientation relative to the wildtype. Another example of orientation-dependent activity was demonstrated when a *GATA-1* enhancer was examined (Nishimura et al. 2000). The authors observed orientation independence of the enhancer when tested *in vitro* using a luciferase assay; however, when these constructs were examined in mice, the orientation of the enhancer was necessary for proper activity. Transgenic mice that carried the enhancer in the endogenous orientation showed *LacZ* expression in six out of nine embryos (the authors attribute position effects to the three negative embryos) but they failed to detect any *LacZ* expression in eight mice that contained the transgene with an inverted enhancer.

While it remains unclear exactly how orientation generally affects CREs, the location independence (Grosveld et al. 1987) of CREs has been repeatedly demonstrated. CREs can function either upstream or downstream of the regulated gene, and in some cases in the introns of non-target genes (ENCODE Project Consortium 2012). How can these CREs exert their effects at long distances from their targets across the genome? Currently, the most supported model to explain CRE function is the looping model.

The looping model was first described by Dunn and colleagues (Dunn et al. 1984). In this paper, the authors studied a repressive element at the L-arbinose operon in *E. coli.* They demonstrated that by altering the position of the CRE relative to the promoter element by increments of 10 base pairs (approximately one turn of the DNA helix) they were able to preserve the repressive effects. Comparatively, when they altered the CRE by increments of five (half the turn of a DNA helix) the repressive effects were lost. From this, the authors deduced that the DNA is looping back on itself to bring the CRE and the promoter into close proximity (Dunn et al. 1984). The DNA looping model was further supported by electron microscopy of λ repressors binding cooperatively to two binding sites and bending DNA when separated by five helical turns. If the binding sites are separated by non-integer turns, such as 4.5 or 5.5 turns, the DNA bending effect was lost (Griffith et al. 1986).

Chromosome Conformation Capture Technologies

In addition to the studies described above, many other experiments have supported the DNA looping model. A new method, chromosome conformation capture (3C), was developed to exploit this model to functionally assess CREs (Dekker et al. 2002) (Figure 1.2). In this method, the DNA and proteins are first crosslinked to maintain the interaction between the CRE and the promoter of the target gene. The underlying principle is based on the DNA looping model, where the CRE element is bound by the transcription factor, the DNA loops to bring the transcription factor into close proximity with the basal transcription machinery and ultimately the promoter of the target gene. The DNA is then digested with restriction enzymes, the protein-DNA crosslink is reversed by degrading the proteins, and PCR using primers designed within the CRE and the promoter can be performed to detect the interaction. This technology has been used extensively to validate many interactions of CREs and their target genes (Tolhuis et al. 2002; Murrell et al. 2004; Spilianakis et al. 2005; Gheldof et al. 2010).

One of the limitations of 3C technologies is that prior knowledge of both the CRE and the target gene are required to design primers. While it has been suggested that 3C can be used to validate predicted CREs and target gene interactions (Gheldof et al. 2010), different iterations of 3C have been developed to eliminate the need for prior knowledge of both interacting partners. The first method developed was circularized chromosome confirmation capture (4C) (Zhao et al. 2006).

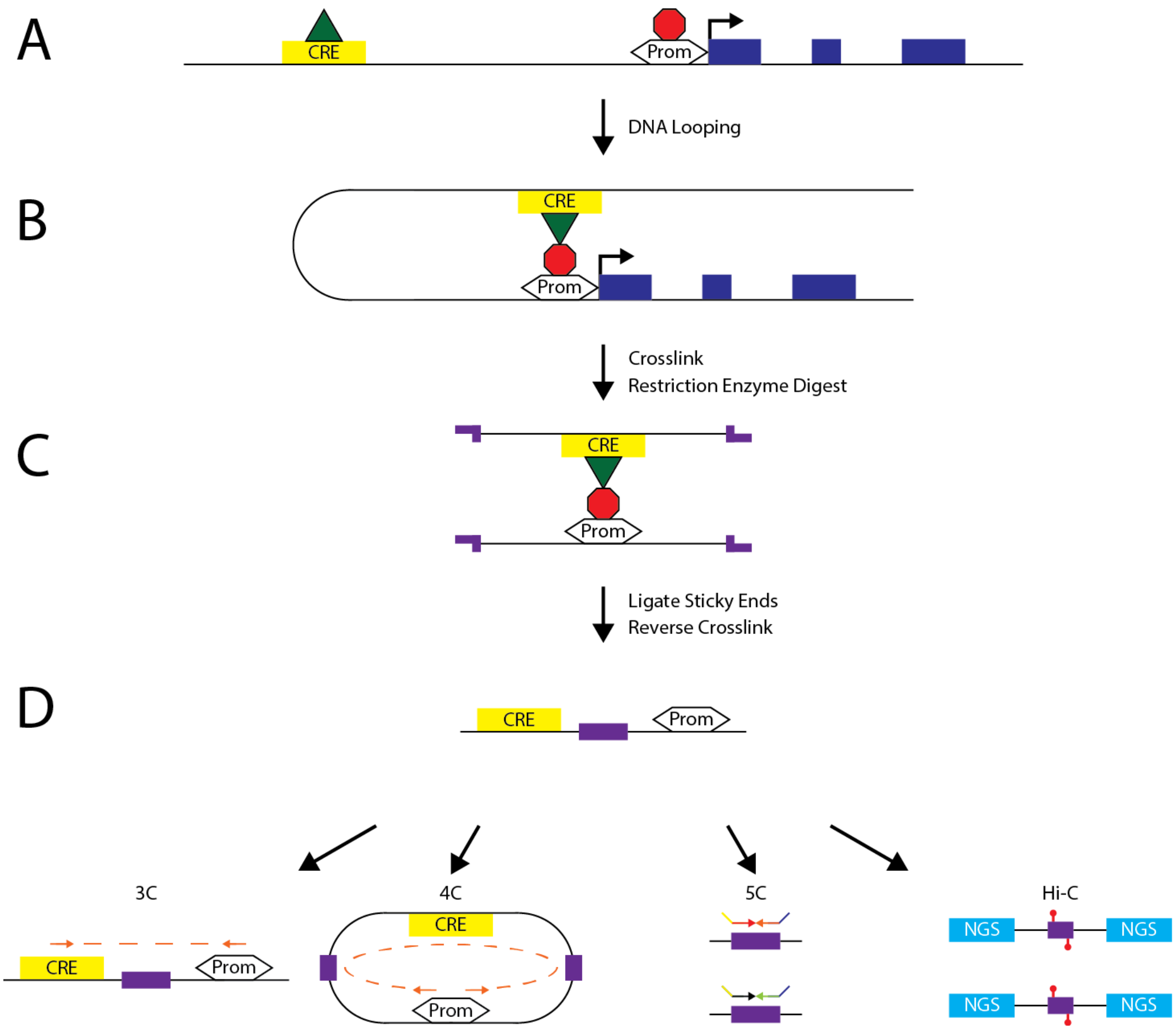


Figure 1.2 *DNA Looping Model and Chromatin Conformation Experiments.* (A) A general schematic of a gene, with a promoter element and upstream cis-regulatory element (CRE). The blue boxes represent exons, the arrow represents the transcription start site, and prom represents the promoter. The green triangle is a transcription factor, and the red hexagon represents the basal transcription machinery. (B) For the CRE to act on the promoter of a gene, the DNA must form a loop to ultimately bring the CRE and promoter into close proximity (DNA looping model). (C) In preparation for a chromosome conformation capture experiment, the proteins and DNA are crosslinked and subjected to restriction enzyme digest. (D) The sticky ends are ligated, and the proteins are digested to link the CRE and the promoter together. The plasmid is then processed for the corresponding technique (see text for details). The purple boxes represent the religated sticky ends. For 3C and 4C the orange arrows represent primers and the dashed line represent amplified PCR product. For 4C, the linear fragment is digest and ligated a second time to form a “plasmid”. For 5C the arrows represent complementary probes, and the green and blue lines are adapters. For Hi-C, the red lines with circles represent biotinylated nucleotides, and the blue boxes are next generation sequencing (NGS) adapters.

The methodology for 4C is identical to 3C, however, rather than performing PCR on the linearized fragment, the DNA is subjected to an additional restriction enzyme digest and the resulting DNA is ligated together. Because the fragment containing the CRE and the promoter has been circularized, it is only necessary to know one interacting partner. Primers are designed within the known sequence, and inverse PCR is used to detect the unknown interacting partner. While in theory 4C can be used to detect novel interactions between CREs and promoters, in practice only a small number of experiments have demonstrated this utility (Verdin et al. 2015).

Similar to 3C, 4C is limited by the requirement of prior knowledge of one interacting partner. The next iteration of 3C technologies to alleviate this problem was carbon-copy chromosome confirmation caption (5C) (Dostie et al. 2006). This method is, again, very similar to 3C and 4C until the protein crosslinking is reversed. In 5C technology, large pools of single-stranded probes are designed immediately flanking the restriction sites used within the genome of interest. The probes are allowed to bind to the DNA and are ligated together. In addition, the probes have adapters on the ends to facilitate additional amplification by PCR and ultimately detection by microarrays or next generation sequencing. 5C technologies remove the requirement for knowing either interacting partner. Similar to 4C, 5C has been used only a limited number of times to detect novel interactions (Fraser et al. 2009; Bau et al. 2011). This may be reflective of the large number of probes necessary for a 5C experiment. Additionally, not all restriction enzyme sites may have a probe flanking the cut site, which will result in false negatives.

While each of the chromosome capture technologies has been used to detect known and novel CRE interactions, none have the resolution to detect all interactions within a cell. The most current improvement in chromosome capture technologies solves this problem with Hi-C (Lieberman-Aiden et al. 2009). Similar to the other chromosome capture technologies, Hi-C involves crosslinking and restriction enzyme digest, however the molecules are not ligated together. Instead, the sticky ends are filled in with nucleotides, where one nucleotide, specifically cytosine, is biotinylated. The fragments are then blunt-end ligated together, the DNA is sheared, and the interaction junction is purified using streptavidin beads in preparation for next generation sequencing. Unlike the other chromosome capture technologies, Hi-C has the resolution to define all DNA-DNA interactions within a cell. Indeed, many studies have used Hi-C (and slight variations) to characterize putative CREs and their interactions within the cell (Lan et al. 2012; Martin et al. 2015).

The Encyclopedia of DNA Elements (ENCODE) Consortium

While each of the prediction methods and functional assays have been used to great success in annotating CREs, the most effective strategies to detect CREs involve combinations of many of the assays previously described. Perhaps the best example for this comes from the encyclopedia of DNA elements (ENCODE) consortium (Birney et al. 2007). One of the aims of the ENCODE consortium is to identify all functional CREs within the human genome. To this end, ENCODE employed a number of relevant assays previously discussed including DNase-Seq, ChIP-Seq, and 5C (or Hi-C) experiments across a large number of cell lines and tissues. In the most recent publications, the ENCODE consortium claimed, "The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. (Birney et al. 2007)" This claim has been meet with criticism, perhaps most chiefly by the definition of 'function' and 'event' as well as the lack of attention paid to evolutionary constraints (Graur et al. 2013). Nonetheless, the ENCODE datasets have increased our knowledge about CREs and their function throughout the genome and have supplied valuable datasets for the scientific community to perform additional physiological relevant studies.

Perhaps the methods that benefitted the greatest from ENCODE's large datasets were the transcription factor binding site prediction algorithms. By aligning ChIP-Seq datasets for one specific transcription factor, a consensus binding site motif or position weight matrix (PWM) (Stormo et al. 1982) can be generated. PWMs were generated prior to the ENCODE datasets, often using a technique called SELEX (systematic evolution of ligands by exponential enrichment) (Tuerk and Gold 1990) to identify sequences transcription factors will bind to. In this experiment, pools of double-stranded DNA oligonucleotides of random sequence are combined with the transcription factor of interest. The transcription factor is allowed to bind to the DNA oligonucleotides, the protein-DNA complex is isolated, the DNA is PCR amplified, and the process is repeated with the enriched sequences. Through multiple rounds of selection, the strongest and most specific DNA-protein interactions will be purified. The resulting pool of oligonucleotides can be sequenced to generate the *in vitro* preferred binding site (Tuerk and Gold 1990).

The limitations of SELEX experiments are similar to EMSA experiments (*e.g.*, potential need for a co-factor, *in vitro* binding sites may not reflect *in vivo* binding, non-specific binding due to *in vitro* conditions), however SELEX has been used successfully to generate PWMs of many transcriptions factors (Wright et al. 1991; Robison et al. 1998; Yagura and Itoh 2006; Jolma et al. 2013). One potential use of the data generated from SELEX experiments is to use the PWM generated to identify novel transcription factor binding sites within a genome of interest. While there are many programs that have been used to accomplish this task, the two most commonly used are TRANSFAC (Wingender 1988; Matys et al. 2003) and JASPAR (Sandelin et al. 2004).

Both TRANSFAC and JASPAR are similar in that they collect functional data from the literature about transcription factor binding sites and combine them together to generate PWMs for individual transcription factors. This allows investigators to extract the PWMs and identify novel transcription factor binding sites genome wide, or conversely, upload a region of interest and allow the programs to predict what transcription factors may be acting on the sequence (Kel et al. 2003). A major limitation of these programs is the reliance on the initial starting datasets to provide accurate PWMs which initially often came from experiments like SELEX, which can result in poor PWMs (Frech et al. 1997). Additionally the low number of datasets used to generate a PWM results in inaccurate predictions of protein-DNA interactions (O'Flanagan et al. 2005). The ENCODE consortium datasets have started to address the limitations of TRANSFAC, JASPAR, and similar programs by providing large datasets generated from multiple experimental conditions and across may diverse cell and tissue types.

CREs in Human Disease

Another interesting observation from the ENCODE consortium was the effect of variation on transcription factor binding sites within humans and how these variations can affect human diseases (Boyle et al. 2012; ENCODE Project Consortium 2012; Schaub et al. 2012). Specifically, many of the significant single nucleotide polymorphisms (SNPs) identified from genome-wide association studies reside within non-coding regions of the genome, rather than gene-coding regions (Schaub et al. 2012). These studies, and others, are currently trying to understand the functional effects (if any) of the disease associated SNPs on CREs.

While the ENCODE consortium has produced vast datasets and has begun to understand the role of CREs in human disease, it is certainly not the first effort to do so. Indeed, many of the first CREs characterized in humans were identified by their phenotypic effects. One of the earliest examples was demonstrated in the human blood disorders thalassemias—a group of inherited blood disorders that disrupt α- and β-globin levels (Marengo-Rowe 2007). These diseases generally result from coding mutations within either the *α-globin* (α-thalassemia) or *β-globin* (β-thalassemia) genes; however, a subset of patients were identified that did not contain coding mutations yet still suffered from β-thalassemia (Kioussis et al. 1983; Driscoll and Dobkin 1989). Upon further inspection, a translocation of the 5' upstream regions (harboring CREs), which was sensitive to DNase I digestion, was the cause of the disease. This region was later termed the locus control region (LCR), which contains many different individual CREs, whose combined input directs appropriate spatial and temporal gene expression of linked genes, in this case the globin genes.

From this initial characterization, many different translocations and mutations have been observed in the LCR. In addition to *β-globin*, mutations have been observed within CREs regulating *α-globin*. Of particular note is a pathogenic SNP located between the α-globin gene cluster and the upstream regulatory elements identified in a patient with α-thalassemia (Gobbi et al. 2006). In this study, the authors were able to identify seven previously uncharacterized SNPs within the disease-associated region, however only one SNP was specific to the α-thalassemia genotype. This SNP creates a novel GATA-1 binding site resulting in a promoter-like element that hijacks the upstream CREs and leads to decreased *α-globin* expression by outcompeting the endogenous *α-globin* promoter, and ultimately causes the α-thalassemia phenotype in patients (Gobbi et al. 2006).

Another example of defects in CREs resulting in human disease was demonstrated at the gene *POU3F4*. Mutations within *POU3F4* result in X-linked deafness in humans (DFN3) (de Kok et al. 1995). It was noticed, however, that a cohort of patients with DFN3 did not contain mutations within *POU3F4* but harbored deletions approximately 400 kilobase pairs upstream of *POU3F4*. This region was further mapped with greater numbers of patients who had variable deletions that overlapped to a minimal eight kilobase pair region approximately 900 kilobase pairs upstream of *POU3F4* (deKok et al. 1996).

While there are many examples where mutations or deletions within CREs result in human diseases, there are a growing number of examples of SNPs that modify either the risk or severity of human disease. These SNPs are generally referred to as regulatory SNPs (rSNPs), and they exert their effect by altering transcription factor binding sites and ultimately disrupting the endogenous protein-DNA interaction (Figure 1.3 - adapted from (Chorley et al. 2008)). This can result in an up- (Feigelson et al. 1998) or down-regulation (Bosma et al. 1995) in the target gene, complete ablation of the binding site (Vasiliev et al. 1999), or creation of a novel transcription factor binding site (Knight et al. 1999; Gobbi et al. 2006).

One example of a rSNP altering disease risk was demonstrated at the *RET* locus. Mutations within *RET* cause Hirschsprung disease - a disorder characterized by the loss of the enteric neurons (Parisi 2015). A subset of patients were identified with Hirschsprung disease that lacked overt coding mutations within *RET*, despite the disease-associated region overlapping *RET* (Emison et al. 2005). The region was narrowed down to the first intron of *RET*, and the authors employed phylogenetic footprinting to identify a functional CRE. This CRE contained a single rSNP that resulted in the mutant minor allele conferring a 2.1-5.7 times greater disease risk

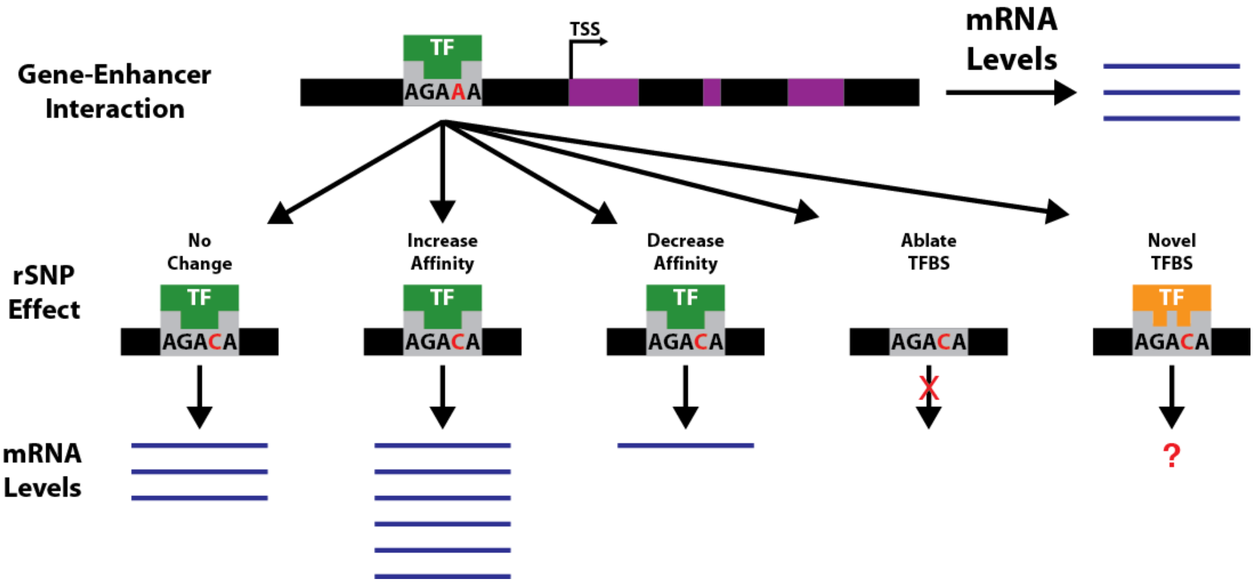


Figure 1.3 *Potential Regulatory SNP (rSNP) Effects on Gene Expression.* (Top) Schematic of a gene with an upstream enhancer element bound by a transcription factor (TF). The purple boxes are exons, the transcription start site (TSS) is marked by the arrow, and the letters in the grey box upstream represent a transcription factor binding site (TFBS). The red letter indicates a rSNP (A to C). In this scenario, a certain amount of mRNA is produced and depicted by the blue lines. (Below) The potential outcomes of a rSNP on TFBS affinity, and ultimately on mRNA production. The number of blue lines indicates an increase (more lines) or decrease (less lines) in mRNA. The red “X” indicates no mRNA expression and the “?” indicates the gene is under control of a novel transcription factor and the spatial and temporal mRNA expression is unknown. Figure adapted from Chorley et al. 2008.

compared to the major allele (Emison et al. 2005). The authors further evaluated this mutation and demonstrated the mutation disrupted a SOX10 binding site (Emison et al. 2010).

The Transcription Factor SOX10

While many of the genome-wide experiments described above to identify CREs can be performed in a transcription factor independent manner (DNase-Seq, STARR-Seq, Cre-Seq, ect.), using a transcription factor-centric approach may lead to a greater understanding of transcription hierarchies in a particular cell line (or tissue) of interest. Since the aim of this thesis work is to identify CREs and rSNPs critical for the peripheral nervous system, utilizing a transcription factor-centric approach for a relevant transcription factor in Schwann cells may provide a deeper understanding of transcriptional hierarchies regulating Schwann cell development. The transcription factor SOX10 resides atop the transcriptional hierarchy and is involved in the regulation of many genes critical for Schwann cell development (discussed in detail below). Therefore, we decided to utilize the knowledge of SOX10 binding sites to identify novel SOX10 responsive CREs harboring rSNPs and SOX10 regulated genes critical for Schwann cells.

SOX10 was first identified from a RT-PCR screen using total RNA isolated from mouse embryos (Wright et al. 1993). This was performed using degenerate primers targeting the high mobility group (HMG) box, a 79 amino acid DNA binding domain, and the resulting products were sequenced to identify novel SOX genes, including *SOX10*. SOX10, and all SOX family members, are grouped together based on homology of the HMG domain with the sex-determining region Y chromosome (*SRY*) and are further subdivided based on homology of the HMG domain to other SOX family members. (Gubbay et al. 1990). The HMG superfamily consists of gene families grouped together based on the variable number of HMG domains and the DNA binding sequence specificity (Laudet et al. 1993). SOX protein family members are a subset of the HMG domain containing superfamily. They are distinct due to the presence of a single HMG domain which binds DNA in a sequence specific manner (Nasrin et al. 1991) and must share at least 50% identity with the *SRY* HMG domain. The SOX family proteins bind to the minor groove of the DNA helix, generally to the sequence 5'-(A/T)(A/T)CAA(A/T)G-3' (Harley et al. 1992; Laudet et al. 1993; Harley et al. 1994).

The SOX family of proteins can be further subdivided into eight currently recognized subfamilies based on primary amino acid sequence (Bowles et al. 2000). SOX10 resides with the SOXE subfamily along with SOX8 and SOX9, which share greater than 90% identity of the HMG domain (Wright et al. 1993). While these proteins are very similar, they are not entirely functionally redundant, as evidenced by the replacement of *Sox10* with *Sox8* in mice (Kellerer et al. 2006). The replacement of *Sox10* with *Sox8* was able to rescue the phenotype of *Sox10* null mice to various degrees in various tissues. For example, the peripheral nervous system was nearly normal, while *Sox8* was unable to fully rescue the defects observed in melanocytes and enteric neurons in *Sox10* null mice (Kellerer et al. 2006).

SOX10 Structure

SOX10 is 466 amino acids long and contains four major domains: dimerization domain, HMG domain, K2 domain, and transactivation domain (Figure 1.4). The dimerization domain encompasses amino acids 61-101 and allows SOX10 to dimerize as both a homodimer and heterodimer (Peirano and Wegner 2000; Schlierf et al. 2002). A homodimeric SOX10 binding site consists of two monomeric sites, oriented in a head-to-head fashion with an intermonomeric sequence. For example, one core monomeric site may be 5’-ACAAA-3’ while the other core monomeric site will be the reverse complement of the consensus site 5’-TGTGT-3’. The two monomeric sites are separated by an intermonomeric sequence spacer that varies in size, however a six base pair spacer is most often observed (Schlierf et al. 2002). This is likely due to both monomeric sites residing on the same face of the DNA helix (*i.e.* 10 total base pairs from position one of one monomer to position one of the second monomer).

The HMG domain consists of amino acids 101-180 and allows SOX10 to bind to DNA in the minor groove (Peirano and Wegner 2000). Upon SOX10 binding to DNA, a bend is induced in the DNA helix of approximately 75-80° or 103-122° for monomeric or dimeric sites respectively (Schlierf et al. 2002). This observation has lead to the hypothesis that SOX10 not only regulates target genes but also alters the three-dimensional chromatin architecture (Giese et al. 1992). While the general consensus sequence for SOX family members is 5’-(A/T)(A/T)CAA(A/T)G-3’, the SOX10 core consensus sequence is altered slightly (all listed 5’ to 3’): ACAAA, ACACA, ACAAT, or ACAAG (Srinivasan et al. 2012; Brewer et al. 2014). The reverse compliment of each sequence can also function as a monomer.

The K2 domain contains amino acids 233-306, and its function remains unclear. The K2 domain has been termed a “cell-specific transactivation domain” (Schreiner et al. 2007). Evidence for this comes from transgenic mouse studies where the endogenous *Sox10* locus was replaced with a *SOX10* mutant lacking the K2 domain. Upon loss of the K2 domain, both early neural crest development and oligodendrocyte differentiation remained unaffected. In contrast, Schwann cell

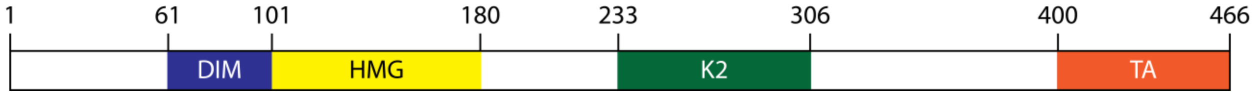


Figure 1.4 *SOX10 Protein Domain Structure.* The four major of SOX10 are displayed: dimerization domain (DIM; blue box), HMG domain (HMG; yellow box), K2 domain (K2; green box), and transactivation domain (TA; orange box). The numbers on the top are the amino acids numbered from the N to C terminus. All domain boxes sizes are accurate relative to size of SOX10.

myelination, melanocyte development, and enteric nervous system development were all disrupted to various degrees (Schreiner et al. 2007). For example in Schwann cells, two myelin genes (*MPZ* and *MBP*) were not expressed in mice lacking the K2 domain. These data suggest the K2 domain can activate gene expression in a cell type-dependent fashion.

Compared to the K2 domain, the transactivation (TA) domain is the main activation domain used to regulate SOX10 target genes. The TA domain resides on the extreme C terminus of SOX10, from amino acids 400-466. Additionally, the TA domain is sufficient to induce reporter gene activity (Pusch et al. 1998). Loss of the TA domain of SOX10 results in complete inability to activate target genes *in vitro* (Bondurand et al. 2000; Potterf et al. 2000). Loss of the TA domain in humans results in a very severe syndrome PCWH (discussed below), a very severe syndrome affecting many different tissue types.

SOX10 Functions in the Neural Crest and Oligodendrocytes

As briefly discussed above, loss of SOX10 (or portions of SOX10) can result in multiple tissue types being affected, specifically tissues derived from the neural crest or oligodendrocytes. Of note, because the main focus of this thesis work is on Schwann cells, the remainder of the chapter will focus on the role of SOX10 in the neural crest or Schwann cells. Perhaps the earliest example of the role of SOX10 in neural crest development was a spontaneous mutation that arose in mice termed *Dom* (Dominant meglacolon) (Lane and Liu 1984; Herbarth et al. 1998; Southard-Smith et al. 1998). This mouse displays symptoms similar to patients with Hirschsprung disease, and the mutation was mapped to *Sox10*. The authors further characterized

Sox10 expression patterns using *in situ* hybridization and observed expression in both the neural crest and the migrating neural crest derivatives (Southard-Smith et al. 1998).

The neural crest is a highly migratory cell population that is formed when the neural plate fuses to generate the neural tube. The neural crest cells first migrate ventrally, and then dorsolaterally giving rise to many distinct tissues including: melanocytes, enteric neurons, sensory neurons and glia of the dorsal root ganglia, and Schwann cells (Erickson and Reedy 1998).

SOX10 plays a critical role in neural crest cells, despite SOX10 being dispensable for neural crest formation. Neural crest cells can be observed in mice containing a homozygous deletion for *Sox10* (Paratore et al. 2001); however, significant cell death is observed later in development, suggesting that SOX10 is necessary for the survival of neural crest stem cells. There also appears to be an additional requirement for SOX10 in the peripheral nervous system. This was observed in neural crest stem cells generated from either *Sox10* heterozygous or homozygous deletion mice exposed to gliogenic conditions. Despite the gliogenic conditions resulting in large numbers of glia cells in wild-type neural crest stem cells, no glial features were detected in the *Sox10* deleted neural crest stem cells in the same conditions (Paratore et al. 2001). Similarly, *Sox10* expression continues in both melanocytes and glia, but its expression is turned off in other neural crest derivatives (Herbarth et al. 1998; Kuhlbrodt et al. 1998; Pusch et al. 1998; Britsch et al. 2001).

SOX10 is Necessary for Schwann cells

Schwann cells are the myelinating cells of the peripheral nervous system, and as previously discussed, they originate from the neural crest. Schwann cells can be divided into two main classes: myelinating and nonmyelinating (Jessen and Mirsky 2005). The Schwann cell precursors differentiate between these two types in a process called radial sorting. In this process, Schwann cell precursor cells associate with axons. If the precursor cell associates with a large (*i.e* greater than 1 μM) diameter axon, then the precursor will become a myelinating Schwann cell and form a one-to-one relationship with the axon (Martin and Webster 1973; Jessen and Mirsky 2005). This is in contrast to oligodendrocytes, the myelinating cells in the central nervous system, where one cell can myelinate many axons at once. Nonmyelinating Schwann cells can associate with multiple small (*i.e.* less than 1 μM) diameter axons and form Remak bundles. While these cells do not generate myelin, they do ensheath small diameter axons with their cytoplasm (Jessen and Mirsky 2005).

The differentiation process from a migrating neural crest cell into either a myelinating or nonmyelinating Schwann cell involves many different transcription factors and genes. Interestingly, *Sox10* remains expressed throughout the differentiation process and is necessary for all stages of myelinating Schwann cell development. Loss of *Sox10* in migrating neural crest cells results in cell death prior to glial differentiation (Britsch et al. 2001; Paratore et al. 2001). Conditional deletion of *Sox10* in the immature Schwann cell stage results in lethality in mice (Finzsch et al. 2010). Additionally, loss of *Sox10* in Schwann cells of adult mice results in severe myelination defects, despite the Schwann cells still remaining in mice (Bremer et al. 2011). This suggests that SOX10 may play a larger role in both the differentiation process and maintenance of the mature differentiated state, rather than survival. Taken together, these data demonstrate SOX10 is necessary for both development and maintenance of Schwann cells.

SOX10 Target Genes and Transcriptional Hierarchy in Schwann Cells

SOX10 regulates many target genes necessary for Schwann cell development. One of the earliest in Schwann cell development is *ErbB3* (Figure 1.5), which allows Schwann cell precursors to respond to Neuregulin 1 (Nrg1). NRG1 signaling is critical for inhibiting neuronal development and allowing Schwann cells to develop normally (Shah et al. 1994). While Sox10 does not regulate *Nrg1*, it does directly regulate a critical Nrg1 receptor, *ErbB3* (Britsch et al. 2001; Prasad et al. 2011). In the absence of ErbB3, Schwann cell precursors are unable to respond to Nrg1 signaling meaning they cannot proliferate and are subsequently lost (Britsch et al. 1998).

Another critical transcription factor for developing Schwann cells is OCT6 (POU3F1). OCT6 is required for the transition from promyelin cells to myelinating Schwann cells (Jaegle and Meijer 1998; Mandemakers et al. 2000). A critical CRE, termed the SCE or Schwann cell enhancer, is both necessary and sufficient to direct appropriate spatial and temporal *Oct6* expression patterns (Mandemakers et al. 2000). Within the SCE is a dimeric SOX10 binding site, which is necessary for SCE function, indicating that SOX10 directly regulates *Oct6* (Jagalur et al. 2011).

Upon upregulation of *Oct6* by SOX10 and other transcription factors, SOX10, Oct6, and Brn2 (Pou3F2), a closely related transcription factor to Oct6 with similar expression patterns, act synergistically to activate *Egr2* (*Krox20*) expression through an enhancer termed the myelinating Schwann cell element (MSE) (Ghislain and Charnay 2006). Egr2 has many characteristics of a “master regulator of myelination” (Ghislain and Charnay 2006), and as such, mutations in *EGR2* have been identified in patients with demyelinating peripheral neuropathies (Warner et al. 1998). Additionally, no myelination is observed in *Egr2* homozygous knockout mice because the Schwann cells are halted at the promyelinating stage (Topilko et al. 1994). Finally, Egr2 has been shown to regulate many genes necessary for myelination (Nagarajan et al. 2001).

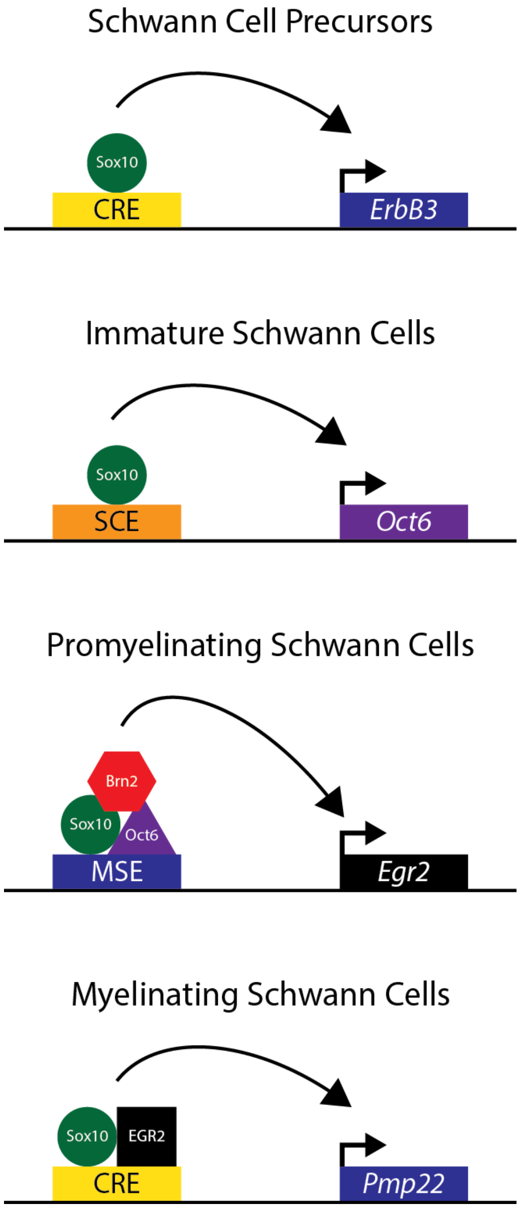


Figure 1.5 *Transcriptional Hierarchies in Schwann Cells.* One of the earliest known targets of SOX10 in Schwann cell development is *ErbB3*. This gene is a receptor for neuregulin 1 signaling that is necessary for inhibiting neuronal differentiation and allows Schwann cell precursors to proliferate. Sox10 upregulates *Oct6* expression through the Schwann cell enhancer (SCE) in immature Schwann cells, which then acts synergistically with Sox10 and Brn2 at the myelinating Schwann cell enhancer (MSE) to upregulate *Egr2*. Upon activation of *Egr2*, promyelinating Schwann cells become mature myelinating Schwann cells and Sox10 and Egr2 regulate many myelination genes including *Pmp22*. Boxes with arrows represent genes, shapes represent transcription factors, colored boxes represent cis-regulatory elements (CRE), and arrows from the CRE to the gene represents activation.

In addition to regulating other transcription factors necessary for myelination, SOX10 also directly regulates, either independently or synergistically with other transcription factors (often EGR2), many myelin-associated genes including *PMP22* (Jones et al. 2011b), *MPZ* (Peirano et al. 2000), *GJB1* (*CX32*) (Bondurand et al. 2001), and *CNTF* (Ito et al. 2006). Combined, these data demonstrate a transcriptional hierarchy in Schwann cells with SOX10 acting as an essential transcriptional regulator at all Schwann cell developmental stages (Figure 1.5).

Mutations in SOX10 and Target Genes

Not surprisingly, mutations within human *SOX10* result in various neurocristopathy symptoms depending on the specific type of mutation. These mutations can manifest as Waardenburg-Hirschsprung disease (WS4) (Pingault et al. 1998) or as a more severe syndrome termed PCWH (peripheral demyelinating neuropathy, central dysmyelinating leukodystrophy, Waardenburg syndrome, and Hirschsprung disease) (Inoue et al. 2002). The discrepancy between these two distinct phenotypes depends on the type of *SOX10* mutation. Mutations that cause a premature stop codon and undergo nonsense-mediated decay, resulting in haploinsufficiency, are associated with the more mild Waardenburg-Hirschsprung disease. This disease is essentially the combination of both Waardenburg syndrome and Hirschsprung disease. It is characterized by hypopigmentation of the hair and skin, heterochromia irides, and impaired hearing (Waardenburg syndrome) combined with aganglionic megacolon (Hirschsprung disease) (Omenn et al. 1979). Comparatively, mutations that cause a premature stop codon, but escape nonsense mediated decay, are associated with the more severe PCWH phenotype. The additional phenotypes observed in patients underscores the additional role of SOX10 in central nervous system myelination, specifically through the expression of SOX10 in oligodendrocytes. The molecular mechanism of PCWH was first elucidated by Inoue and colleagues (Inoue et al. 2004), where they demonstrated mutations associated with PCWH act by a dominant-negative mechanism, ultimately resulting in a more severe (greater than 50%) depletion of SOX10 function.

In addition to the peripheral neuropathy observed in patients with PCWH, mutations within SOX10 target genes result in peripheral neuropathies. One example in particular is Charcot-Marie-Tooth (CMT) disease, which is characterized by distal muscle wasting and sensory loss (Dyck and Lambert 1968; Szigeti and Lupski 2009). CMT can be subdivided into two major classifications based on motor nerve conduction velocities (MNCVs). Patients with CMT type 1 (CMT1) have reduced MNCVs due to the primary defect arising in the Schwann cells. Conversely, patients with CMT type 2 (CMT2) have normal MNCVs, but reduced amplitude because the primary defect is in the axon (Pagon et al. 1993).

Unsurprisingly, many genes critical for proper myelination in Schwann cells also are mutated in CMT1: *PMP22* (Lupski et al. 1991; Raeymaekers et al. 1991), *MPZ* (Kulkens et al. 1993), (Hayasaka et al. 1993), and *GJB1* (Ionasescu and Searby 1994). All of these genes are regulated by SOX10 (discussed above). In addition, mutations within SOX10 binding sites at some of these genes have been reported to cause CMT. One such example was observed at the *PMP22* locus which is duplicated in patients with CMT1A (Lupski et al. 1991; Raeymaekers et al. 1991). Patients with CMT1A were identified that did not harbor the *PMP22* duplication, but rather the upstream genomic regions were duplicated resulting in a milder form of CMT (Weterman et al. 2010). Recent work identified functional EGR2 and SOX10 binding sites within the upstream duplicated regions which were able to direct appropriate peripheral nerve expression patterns within zebrafish (Jones et al. 2011a). Additionally, mutations within the *GJB1* promoter have been shown to cause CMT (Houlden et al. 2004). These mutations disrupt a SOX10 monomeric site and ultimately lead to reduced levels of *GJB1* expression.

Another example was demonstrated at the *MPZ* locus. A rare, non-coding variant was detected in a patient with hypomyelination in the central nervous system (Antonellis et al. 2010). This variant disrupted a SOX10 binding site, and was shown to be less active in luciferase assays; however the authors were unable to implicate the variant in human disease. This variant may be causative for human disease, or perhaps this variant functions as a modifier of disease.

Indeed, large phenotypic variability such as age of onset from the first to seventh decade and severity of sensory loss is observed in patients with CMT, often despite identical coding mutations (Thomas et al. 1997; Pareyson et al. 2006; Pareyson and Marchesi 2009). This variability was even described in two sets of unrelated identical twins with identical duplications of *PMP22* (Garcia et al. 1995). The cause of this variability is unknown, however rSNPs may provide an explanation. For example, if two individuals have a duplication of *PMP22* (resulting in CMT1A), and one individual has the major allele and the other has the minor allele at a specific rSNP, this may account for some of the phenotypic variability observed. The rSNP in this case could be located in the upstream SOX10 binding sites at *PMP22* and thus possessing the allele that disrupts SOX10 binding could be beneficial due to decreased *PMP22* expression. The rSNP however, does not need to act on the same gene that is mutated but could affect a gene in the same genetic pathway. One example was an rSNP identified at the *SH3TC2* locus (Brewer et al. 2014). In this study, Brewer and colleagues identified and functionally evaluated a rSNP which appears to be a modifier of the CMT1A phenotype. Unfortunately, the low minor allele frequency associated with this SNP precluded association studies. Identification of both putative

SOX10 binding sites and rSNPs could elucidate the mechanism of phenotypic variability observed in patients with CMT and could potentially provide therapeutic targets in the future.

Summary

In this chapter, I reviewed some of the methods used to identify and characterize CREs. These included *in silico, in vitro,* and *in vivo* methods, which have been used to great success. Indeed, these pioneering papers helped discover many CREs and define general characteristics of CREs. Current and future studies are applying this knowledge to identify additional CREs, often combining many of the methods discussed above. In this thesis work, we will also use a variety of methods to identify functional, non-coding variation critical for the peripheral nerve. In chapter 2, I will discuss our methods to uncover novel CREs harboring putative rSNPs using an *in silico* phylogenetic footprinting method followed by functional evaluation using luciferase assays.

I also discussed the importance of CREs within human disease, with an emphasis on both the peripheral nervous system and the transcription factor SOX10. As SOX10 is a master regulator of Schwann cell development and maintenance, identification of SOX10 response elements and rSNPs affecting the binding sites may uncover novel mechanisms of phenotypic variation observed in patients with peripheral neuropathies. Additionally, this work may uncover novel SOX10 target genes, allowing for a greater understanding of fundamental Schwann cell function. In chapter 3, I will apply the methods developed in chapter 2 to specifically identify putative SOX10 binding sites harboring rSNPs. I will deeply characterize one of the identified SOX10 response elements to try to uncover the regulated gene(s).

I have discussed how *Sox10* expression is required for all developmental stages of Schwann cells and expression remains consistent throughout Schwann cell development. One question that arises is how can *SOX10* be expressed in all Schwann cell precursors but only activate myelin genes within myelinating Schwann cells? In chapter 4, I will utilize our computational pipeline to identify SOX10 responsive elements residing near genes involved in negative regulation of myelination and discuss some possibilities of how SOX10 (and additional factors) may mediate the switch from negative regulators to positive regulator of myelination. Finally, in chapter 5, I will summarize the findings and impact of the research presented in this thesis and present possible lines of future investigation.

CHAPTER 2

Identification of Regulatory SNPs Relevant the Peripheral Nerve

Introduction

Cis-regulatory elements (CREs) are comprised of promoters, enhancers, repressors, and insulators and are critical for regulating gene expression in a spatial and temporal manner. They are typically short (five to ten) stretches of base pairs that contain one (or more) transcription factor binding sites (TFBS). One functional CRE may contain multiple TFBSs to exert the appropriate regulation of a gene. While it is easy to identify genes and promoters by the mRNA sequence and the 5' location to genes respectively, it remains difficult to identify CREs, due to location independence and lack of knowledge about many TFBS (Chapter 1).

Many methods have been developed and used to successfully identify CREs. One of the largest efforts to discover novel CREs is the encyclopedia of DNA elements (ENCODE) consortium (Birney et al. 2007). The ENCODE consortium has performed many assays across a diverse subset of cell types and has vastly increased the vocabulary of CREs. Despite the immense datasets, the ENCODE project has been limited by the cell types assessed and criticized for the lack of consideration for evolutionary constraints (Graur et al. 2013).

One method to predict CREs and alleviate the limitations of the cell types used in the ENCODE project is phylogenetic footprinting (Tagle et al. 1988). This method relies on the hypothesis that regions conserved across multiple species may imply a function for the sequence, such as a

TFBS. While this method can predict CREs regardless of cell type, a major consideration with using this technique is which species to include in the analysis. Including many closely related species may lead to false positives, while the inclusion of many diverse species can lead to false negatives. In addition, while many studies have used *in silico* techniques to predict CREs, functional evaluation of the predictions is necessary because conservation does not always translate into function (Fisher et al. 2006; Pennacchio et al. 2006).

Identification of CREs does not only reveal novel biology about a system (*i.e.* how a gene is regulated) but can also explain disease etiology. Indeed, disruption of CREs can result in many human diseases including Charcot-Marie-Tooth (CMT) disease. CMT disease is a clinically heterogeneous disease that affects the peripheral nervous system and is characterized by progressive distal muscle wasting (Dyck and Lambert 1968). CMT is subdivided based on whether the primary defect occurs in the Schwann cell (CMT1) or in the axon (CMT2). The most common coding mutation observed in patients with CMT is a duplication of *PMP22* (CMT1A) (Lupski et al. 1991; Raeymaekers et al. 1991). However, a subset of patients with CMT1A were identified that did not harbor a duplication of *PMP22* but did contain a duplication of upstream CREs (Weterman et al. 2010; Jones et al. 2011a). These patients generally have a more mild form of CMT, relative to patients containing a *PMP22* duplication. While in this case the clinical variability between the two patient cohorts may be due to the relative levels of PMP22 protein (Huxley et al. 1998), variability can be observed among patients with molecularly indistinguishable duplications of a 1.4 megabase pair region including *PMP22* (Garcia et al. 1995; Thomas et al. 1997). The cause of the clinical variability is unknown, however one possible explanation may be regulatory single nucleotide polymorphisms (rSNPs).

rSNPs are base pair changes within CREs that alter the DNA binding affinity of a transcription factor. This can result in an increase (Feigelson et al. 1998) or decrease (Bosma et al. 1995) in gene expression, ablate a TFBS (Vasiliev et al. 1999), or create a novel TFBS (Knight et al. 1999) (Figure 1.3). These changes could alter disease risk or modify disease severity leading to clinical variability despite identical coding mutations. For example, a patient with a duplication of *PMP22* that also harbors a rSNP in a CRE that ultimately reduced expression of *PMP22*, may have a more mild phenotype relative to a patient lacking the rSNP.

In this chapter, we developed a novel computation pipeline based on phylogenetic footprinting to predict CREs. This dataset was overlapped with validated SNPs within the human genome to identify putative rSNPs. We functionally assessed a subset of these predicted CREs harboring rSNPs using luciferase assays within three cell lines to approximate a functional peripheral nerve unit: Schwann cells (S16), motor neurons (MN-1), and muscle cells (C2C12). Any region that displayed at least a five-fold increase in luciferase expression relative to an empty control vector was mutated to the minor allele and reassessed in the appropriate cell line. Finally, an *in silico* TFBS prediction program (TRANSFAC) (Matys et al. 2003) was used to predict differential binding of transcription factors. The active regions harboring putative rSNPs identified represent excellent candidate modifiers of CMT disease and other peripheral neuropathies.

All of the work in this chapter was performed by the author with the exception of the generation of the human, mouse, and chicken genome alignments which was performed by Dr. Tony Antonellis and Dr. Arjun Prasad (Antonellis et al. 2006), and a subset of the conserved regions were PCR amplified, cloned into pDONR, and cloned into pE1B forward and reverse by Aimée Vester, Chani Hodonsky, and Chetna Gopinath.

Methods

Computational Identification of Conserved Regions Harboring a SNP

The human (hg18), mouse (mm9), and chicken (Gal3) genomes were downloaded from the UCSC Genome Browser (Kent et al. 2002), aligned using MultiPipMaker(Schwartz et al. 2000), and the alignments were analyzed using ExactPlus (Antonellis et al. 2006) to identify genomic segments that are identical among the three species and at least five base pairs in length (Gopinath and Law, manuscript in preparation). Next, genome-wide SNPs (dbSNP130) were downloaded from the UCSC Table Browser (Karolchik et al. 2004) that were validated "by-frequency." The "by-frequency" validation method requires SNPs to have frequency data about all alleles when they were submitted. A custom Perl script was written to identify the overlap between the two datasets to generate fully conserved regions harboring a SNP. This dataset was uploaded to the UCSC Table Browser (Karolchik et al. 2004), and conserved regions containing SNPs were removed that had any overlap with hg18 RefSeq (Pruitt et al. 2012) exons, to remove regions that were conserved due to gene function.

PCR and Cloning for Each Region

PCR was performed to amplify a region surrounding each of the 144 fully conserved, non-coding regions with SNPs (see Appendix I for primers). The surrounding region to amplify was chosen based on general conservation using the PhastCons 17-way vertebrate alignment dataset (Siepel et al. 2005). For example, our conserved regions (generally short stretches of sequences; tens of base pairs or less) frequently resided within a 'block' of conservation based on PhastCons (generally large stretches of sequences; hundreds of base pairs). Primers were designed to amplify the entire 'block' of conservation, rather than imposing arbitrary sequence length restrictions. Primers were designed using the online Primer3 program (http://bioinfo.ut.ee/primer3/) with default parameters except the primer tm Min = 55°C, Opt = 58°C, and Max = 60°C, and the conserved region was excluded from primer design using the Excluding Regions tool (*i.e.* < and >). The primers were modified to include gateway adapters, ordered from IDT, and diluted to 200 μM in ultrapure water. The primers were diluted 1:10 (20 μM) in ultrapure water prior to PCR reactions. Each region was PCR amplified from mixed human genomic DNA using gateway adapted primers and BP cloned into pDONR221 using the Gateway cloning technology (Life Sciences). For an individual BP reaction: 1 μL of PCR product was mixed with 0.5 μL (150 ng/μL) of pDONR221, 1 μL BP clonase (ThermoFisher cat no 11789-020), and 2.5 μL TE. The reaction was incubated for one hour at room temperature. After incubation, 1 μL of Proteinase K solution was added to stop the reaction by degrading the recombinase. The regions were transformed into Top10 *E. Coli* (ThermoFisher cat no C4040-06): 12.5 μL of bacteria was mixed with 3 μL of the BP reaction and incubated on ice for 25 minutes. The bacteria were heat shocked at 42°C for 45 seconds, 62.5 μL of SOC media (ThermoFisher cat no C4040-06) was added, and the mixture was incubated at 37°C shaking for one hour. All 78 μL was plated on 25 mg/mL kanamycin selective plates and incubated overnight at 37°C. Individual colonies were picked and grown in 6 mL of kanamycin selective media shacking at 225 RPM overnight at 37°C. Plasmid DNA was isolated using the Qiagen miniprep kit as per the manufacturer’s protocol (Qiagen cat no 27106) using 5 mL of the media (1 mL was saved for storage at -80°C). The DNA was assessed for proper recombination using a diagnostic *Bsr*GI digest: 6.75 μL water, 1 μL (150 ng) plasmid DNA, 1 μL BSA, 1 μL NEBuffer 2, 0.25 μL *Bsr*GI (NEB cat no R0575S) (2.5 units). The reaction was incubated for one hour at 37°C, and 1 μL of the reaction was assessed on a 1% agarose in TBE (Fisher cat no 50-751-

7033) gel. The regions were then sequenced to verify the cloned allele and subsequently LR cloned into both pE1B forward and reverse luciferase vectors (Antonellis et al. 2006). pE1B is a Gateway compatible vector with the minimal promoter E1B from adenovirus directing luciferase gene expression. For an individual LR reaction, 1 μL (150 ng/μL) of pDONR221 plasmid was mixed with 0.5 μL (150 ng/μL) of pE1B plasmid, 1 μL LR clonase (ThermoFisher cat no 11791-020), and 2.5 μL TE. The reaction was transformed, plasmid DNA was isolated, and a diagnostic *Bsr*GI digest was performed as above with the exception of 100 mg/mL ampicillin selective plates and media rather than kanamycin.

Cell Culture and Lipofectamine 2000 Transfection

The S16 (Goda et al. 1991) and MN-1 (Salazar-Grueso et al. 1991) cells were cultured at 37°C in 5% CO_2 in general media (GM): Dulbecco's Modified Eagle Medium (DMEM; Invitrogen cat no ILT12430054) and supplemented with 10% fetal bovine serum (FBS; ThermoFisher cat no 26140-079), 2 mM L-glutamine (Corning cat no COR25005CIS), and 1X Penicillin-Streptomycin (50 units of penicillin and 50 μg of streptomycin; ThermoFisher cat no 15070-063). The S16 and MN-1 cells were plated at 10,000 cells per well of a tissue culture treated 96-well plate (Corning cat no 07-200-565). The cells were transfected the following day using Lipofectamine 2000 (ThermoFisher cat no 11668-019). To transfect a single well of a 96-well plate: 0.25 μL Lipofectamine 2000 was mixed with 25 μL OptiMem (ThermoFisher cat no 31985-062) and incubated for 10 minutes at room temperature (cocktail 1). While cocktail 1 incubated, 200 ng of DNA plasmid was mixed with 2 ng of CMV-renilla plasmid and 25 μL of OptiMEM (cocktail 2). Prior to transfection, the plasmid for each region was diluted to 200 ng/μL, and a fresh aliquot of 2 ng/μL of CMV-renilla was prepared. A master mix of CMV-renilla and OptiMEM was made and aliquoted to each tube of DNA. After the 10 minute incubation of cocktail 1, 25 μL of cocktail 1 was added to cocktail 2 (transfection mixture), briefly vortexed, and incubated at room temperature for 20 minutes. During the incubation, the cells were washed with 75 μL of 1X PBS (ThermoFisher cat no 10010-023). After the 20-minute incubation, the PBS was aspirated, and 50 μL of the transfection mixture was added to an individual well. The cells were incubated for 4 hours at 37°C in 5% CO_2. After 4 hours, the transfection mixture was removed, and 75 μL of GM was added. No PBS wash was performed between removal of transfection mixture and addition of GM. After 48 hours, the cells were harvested: cells were washed with 75 μL of 1X PBS, 20 μL of 1X passive lysis buffer (4 μL 5X passive lysis buffer mixed with 16 μL ultrapure water [Promega cat no E1980]) was added, and the plate was shaken on medium speed for one hour at room temperature. After one hour, 10 μL of the lysate was transferred to a white, opaque 96-well plate (Corning cat no CLS3789) and assessed using the Dual-Luciferase Reporter system (Promega cat no E1980): 25 μL of luciferase activating reagent (LAR) was added, 10 seconds of light was recorded, 25 μL of Stop and Glo (SNG) was added, and 10 seconds of light was recorded using a luminometer. Luciferase activity was normalized to renilla activity, and the activity of each region was compared to an empty control vector that does not contain an insert and the activity has been set to a value of '1'. Bar graphs represent at least 8 replicates and statistical calculations were performed using a two-tailed Student's t-test.

The C2C12 cells were maintained and plated in GM at a concentration of 5,000 cells per well of a 96-well plate and transfected as described above. 24 hours post transfection the cells were washed with 1X PBS, and the media was changed to differentiation media (DM): Dulbecco's Modified Eagle Medium (DMEM; Invitrogen) and supplemented with 5% horse serum

(Invitrogen cat no 16050122) (2007), 2 mM L-glutamine, and 1X Penicillin-Streptomycin. The cells were processed and assessed 48 hours post transfection as described above.

Mutagenesis

Major alleles of active regions were mutagenized to the minor allele in pDONR221 and sequenced to verify the integrity of the insert. The mutagenic primers were designed using the online QuikChange Primer Design program with default parameters (http://www.genomics.agilent.com/primerDesignProgram.jsp). The primers were ordered from IDT and diluted to 1,250 ng/μL in ultrapure water. Primers were diluted 1:10 (125 ng/μL) in ultrapure water for mutagenesis reactions. The mutagenesis reaction was performed using the QuikChange II XL Site Directed Mutagenesis Kit (Agilent cat no 200522). One mutagenesis PCR reaction contains the following: 5 μL 10X buffer, 1 μL major allele pDONR template (20 ng/μL), 1 μL forward primer, 1 μL reverse primer, 1 μL dNTPs, 3 μL QuikSolution, 38 μL ultrapure water, and 1 μL Pfu taq. The reaction was performed on a thermocycler with the following conditions: 95°C (2 minutes), 95°C (15 seconds), 50°C (50 seconds), 68°C (10 minutes), repeat step two through four 18 times, 68°C (10 minutes), and 4°C (hold). After amplification, 1 μL of *Dpn*I (10U; from QuikChange Kit) was added to the total reaction and placed at 37°C for two hours. The reaction was then ethanol precipitated: 52 μL QuikChange reaction, 156 μL (3X PCR reaction) 100% ethanol, and 5.2 μL (0.1X PCR reaction) 3M sodium acetate. The reaction was mixed and placed at -80°C for one hour and then spun at maximum speed (13,200 rpm) at 4°C for 30 minutes. After the spin, the supernatant was removed, and the DNA pellet was air dried for 15 minutes. The DNA was then dissolved in 10 μL of ultrapure water and transformed into bacteria as described above. The regions are then sequenced to verify the presence of the desired mutation and the integrity of the surrounding sequence.

TRANSFAC Analysis

The major allele sequence for each active region was obtained from the UCSC genome browser (Kent et al. 2002). The conserved region was centered and additional base pairs were included on both the 5’ and 3’ ends to generate a total of 30 base pairs of sequence. The minor allele for the candidate rSNP was substituted into the major allele sequence to generate the minor allele sequence. Both the major and minor alleles were assessed using the TRANSFAC Match tool (Kel et al. 2003). Default parameters were used with the vertebrate, non-redundant profile minimizing the sum of the false positive and negative error rates. The results were filtered to exclude any predicted binding sites that were identical between the major and minor allele (*i.e.* only binding site predictions that differed between the major and minor allele were included), regardless of the core or matrix scores, and only differential predictions is displayed.

Results

Identification of Genome-wide Conserved Non-Coding Regions Harboring SNPs

To identify conserved non-coding regions harboring SNPs, we developed a novel computational pipeline (Figure 2.1). The human (hg18), mouse (mm9), and chicken (gal4) genomes were downloaded from the UCSC genome browser (Kent et al. 2002). The genomes were aligned using MultiPipMaker (Schwartz et al. 2000), and regions that were identical among the three species and at least five base pairs in length were isolated using ExactPlus (Antonellis et al.

2006). This revealed over two million multi-species conserved sequences (MCSs). Next, all SNPs from dbSNP130 that contained information about allele frequencies (*i.e.* validated 'by-frequency'), regardless of the minor allele frequency (MAF), were downloaded from the UCSC Table Browser (Karolchik et al. 2004). A custom Perl script was designed to overlap the two datasets and generate fully conserved regions harboring SNPs. This dataset was uploaded to the UCSC Table Browser, and regions were filtered out that harbored any overlap with RefSeq exons (Pruitt et al. 2012). The final dataset was comprised of 6,164 fully conserved non-coding regions containing SNPs genome-wide. A similar analysis has been performed with the SNP identified from the 1,000 Genomes Project (Altshuler et al. 2012) from dbSNP137 (Figure 2.2) and all SNPs from dbSNP142 (Figure 2.3).

Functional Assessment of Regions Important for the Peripheral Nerve

To functionally evaluate our computational predictions, we assessed the ability of a pilot set of regions on chromosomes 21 (37 regions), 22 (29 regions), and X (94 regions) to direct luciferase reporter gene expression in three cell lines: Schwann cells (S16), motor neurons (MN-1), and muscle cells (C2C12). These cell lines were chosen to approximate a peripheral nerve and a target tissue (muscle). The S16 cells (Goda et al. 1991) are a rat immortalized Schwann cell line that express many myelin associated genes (*e.g. Pmp22*, *Mpz*, and *Mbp*) and critical Schwann cell transcription factors (*e.g. Sox10* and *Egr2*), and are currently the best model cell line of myelinating Schwann cells (Hai et al. 2002). The MN-1 cells (Salazar-Grueso et al. 1991) were generated by somatic cell fusion between a mouse spinal motor neuron and a mouse neuroblastoma cell, and exhibit traits similar to motor neurons, including the ability to induce neurite projections. The C2C12 cells (Yaffe and Saxel 1977a) were generated following a crush

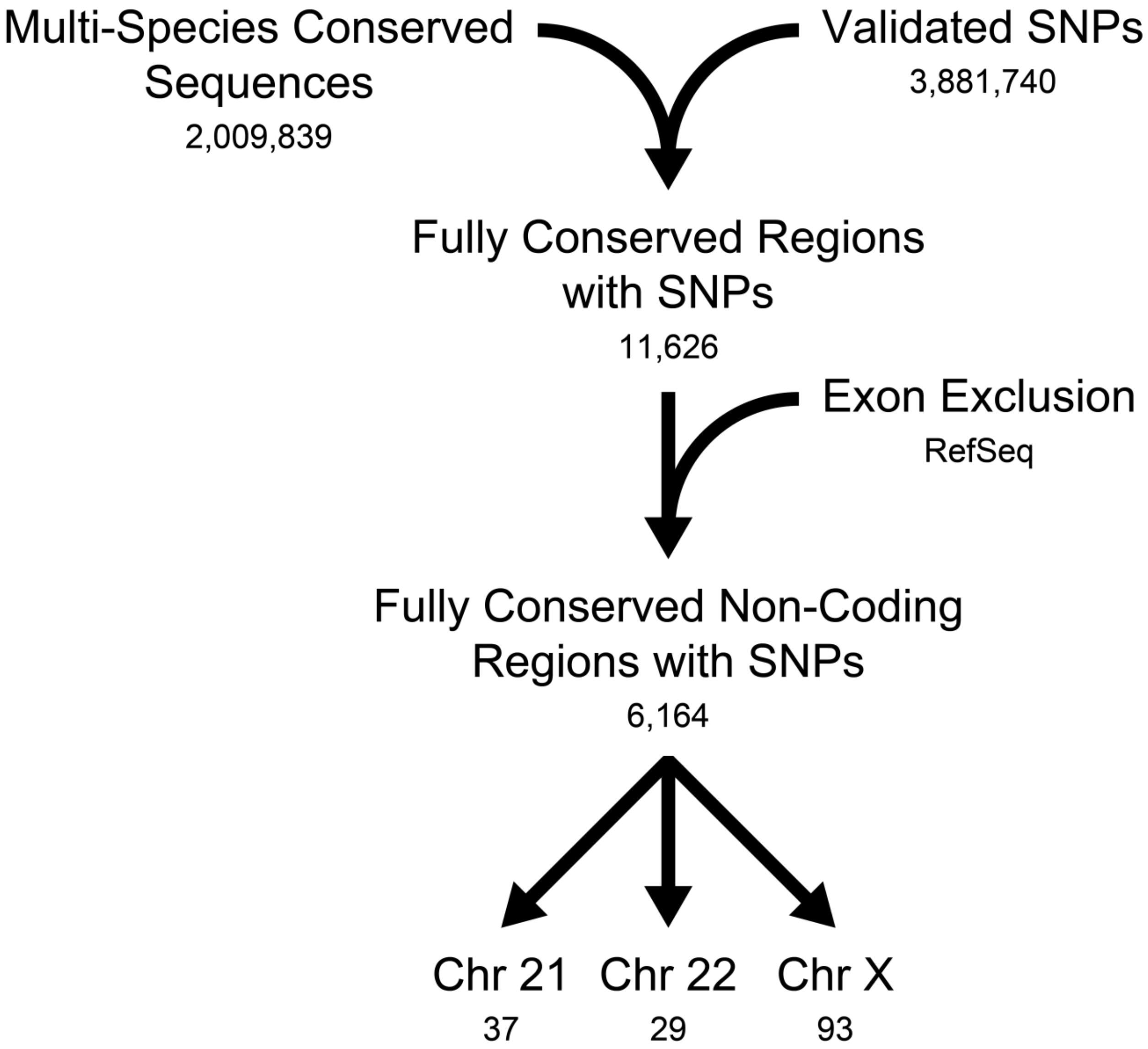

Multi-Species Conserved Sequences
2,009,839
Validated SNPs
3,881,740
Fully Conserved Regions with SNPs
11,626
Exon Exclusion
RefSeq
Fully Conserved Non-Coding Regions with SNPs
6,164
Chr 21
37
Chr 22
29
Chr X
93

Figure 2.1 *A Computational Pipeline to Identify Putative Regulatory SNPs Using Validated 'by-frequency' SNPs from dbSNP130.* The human (hg18), mouse (mm9), and chicken (gal3) genomes were aligned, and genomic segments that are five base pairs in length or greater and identical in all three species were identified to generate a dataset of multiple-species conserved sequences (MCS). Overlap between the MCS dataset and SNPs validated 'by-frequency' from dbSNP130 was determined. Exons were excluded using the RefSeq gene list, and the final dataset was parsed into chromosome 21, 22, and X. Numbers below each dataset label represent the number of entries in that dataset.

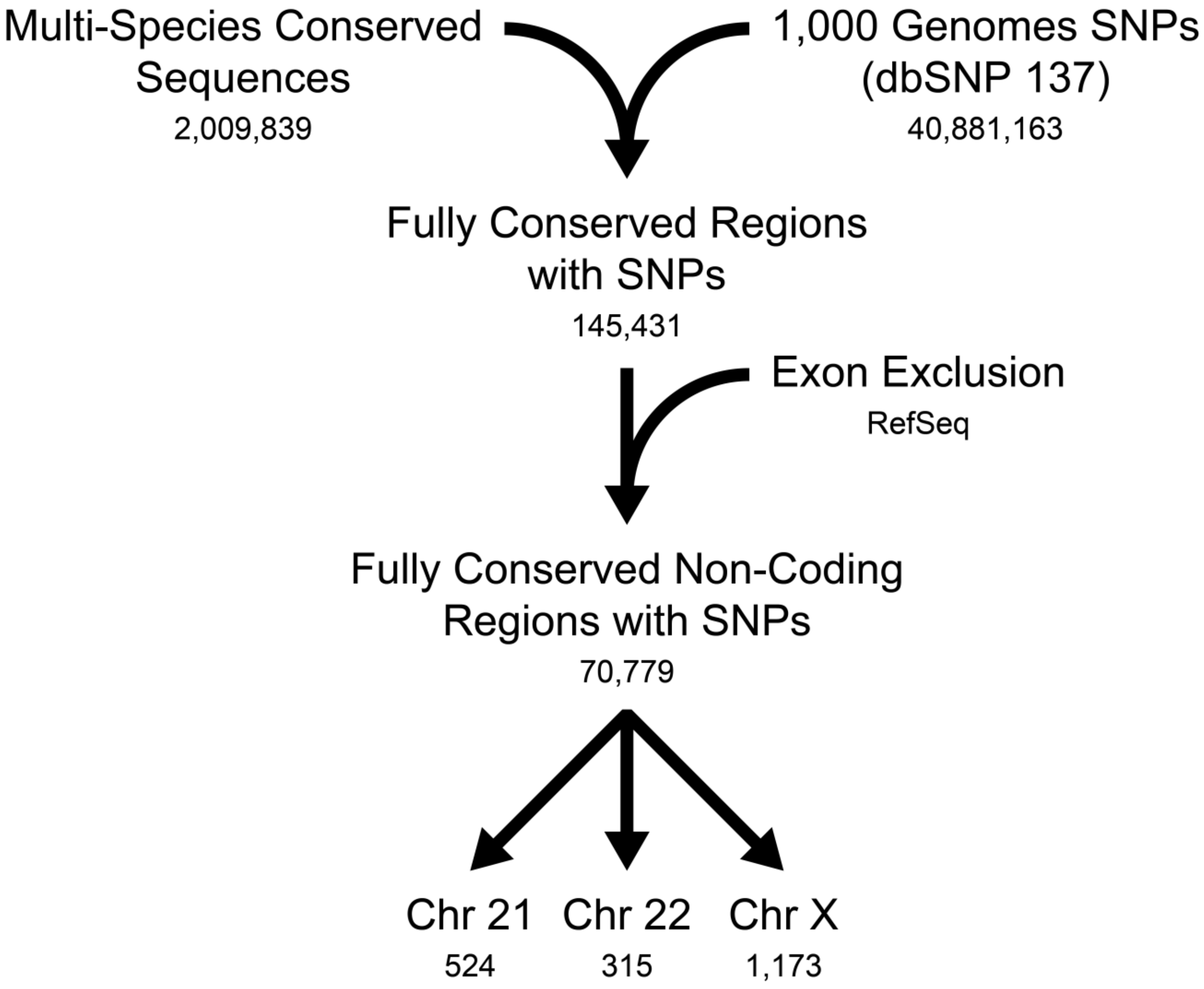

Multi-Species Conserved Sequences
2,009,839
1,000 Genomes SNPs (dbSNP 137)
40,881,163
Fully Conserved Regions with SNPs
145,431
Exon Exclusion
RefSeq
Fully Conserved Non-Coding Regions with SNPs
70,779
Chr 21
524
Chr 22
315
Chr X
1,173

Figure 2.2 *A Computational Pipeline to Identify Putative Regulatory SNPs Using the 1000 Genomes Project SNPs.* The human (hg18), mouse (mm9), and chicken (gal3) genomes were aligned, and genomic segments that are five base pairs in length or greater and identical in all three species were identified to generate a dataset of multiple-species conserved sequences (MCS). Overlap between the MCS dataset and the SNPs identified from the 1000 Genomes Project was determined. Exons were excluded using the RefSeq gene list, and the final dataset was parsed into chromosome 21, 22, and X. Numbers below each dataset label represent the number of entries in that dataset.

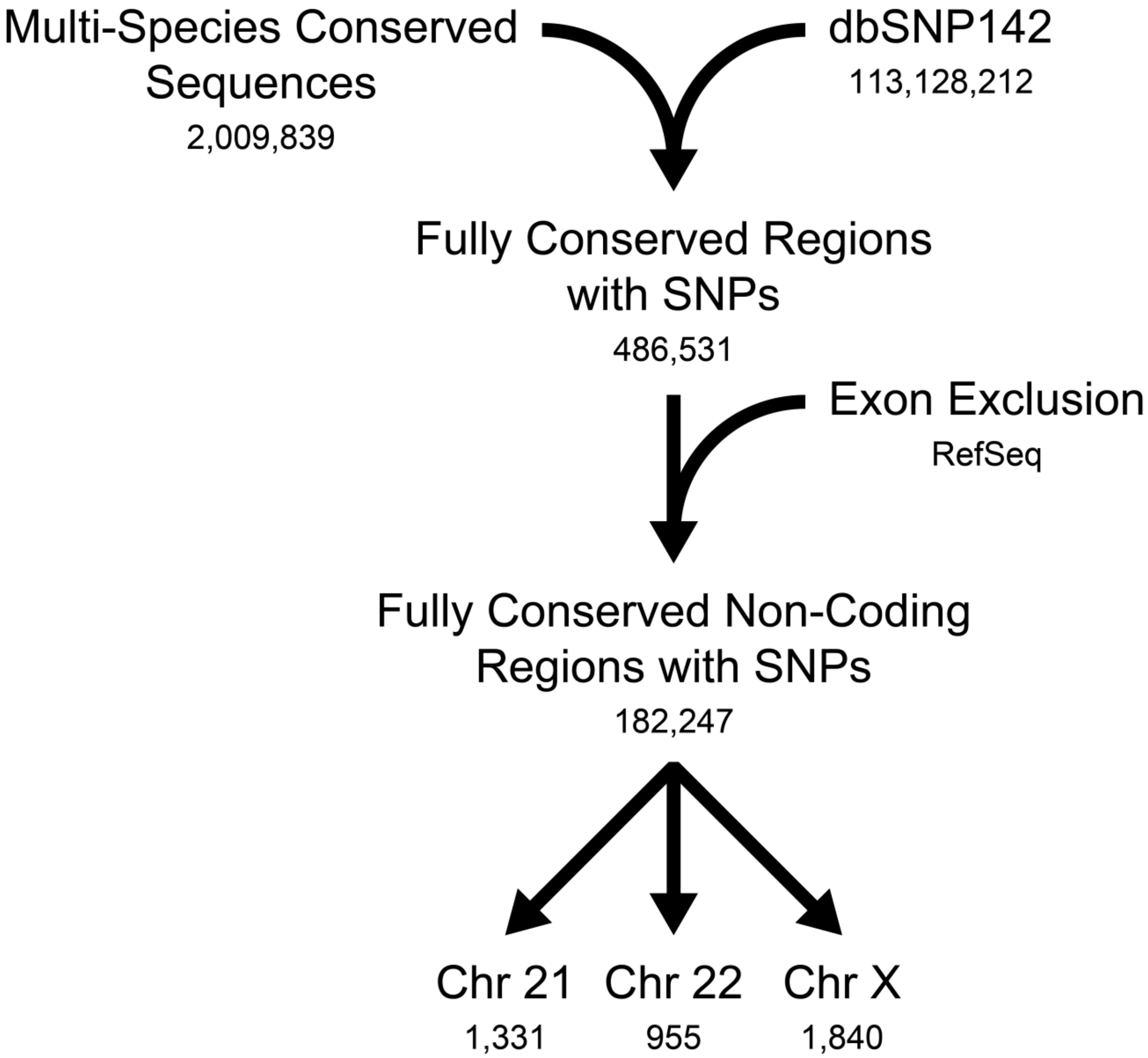

Multi-Species Conserved Sequences
2,009,839
dbSNP142
113,128,212
Fully Conserved Regions with SNPs
486,531
Exon Exclusion
RefSeq
Fully Conserved Non-Coding Regions with SNPs
182,247
Chr 21
1,331
Chr 22
955
Chr X
1,840

Figure 2.3 *A Computational Pipeline to Identify Putative Regulatory SNPs Using All dbSNP142 SNPs.* The human (hg18), mouse (mm9), and chicken (gal3) genomes were aligned, and genomic segments that are five base pairs in length or greater and identical in all three species were identified to generate a dataset of multiple-species conserved sequences (MCS). Overlap between the MCS dataset and all SNPs contained within dbSNP142 was determined. Exons were excluded using the RefSeq gene list, and the final dataset was parsed into chromosome 21, 22, and X. Numbers below each dataset label represent the number of entries in that dataset.

injury to mouse muscle tissue, and the resulting myoblasts were cultured to generate the immortal line. These cells can undergo myogenic differentiation in low serum concentrations to form multinucleated myotubes which model adult skeletal muscle (Yaffe and Saxel 1977b).

Briefly, a region surrounding each putative enhancer element was PCR amplified based on general conservation of the surrounding sequence using the PhastCons 17-way vertebrate alignment dataset (Siepel et al. 2005). These regions were cloned upstream of a minimal E1B promoter sequence directing luciferase gene expression in both the forward and reverse directions (relative to the promoter) and transfected into all three cell lines. Luciferase activity was measured relative to an empty control vector with no insert upstream of the E1B promoter that has been set to a value of '1'. Regions demonstrating a greater than five-fold increase in luciferase activity relative to the empty control vector were considered to have 'strong' activity and were used in further analyses.

We successfully cloned and assessed 144 regions out of the initial 159 prioritized regions. There were 15 regions that were not assessed: six were amplified with another region in the same PCR product (and were thus tested simultaneously), eight failed to PCR amplify, and one could not be cloned into pDONR221. The regions were named SNP conservation ('SC') followed by the chromosome and were numbered from the p-arm to the q-arm. For example, SCX-1 is the most distal region identified on the p-arm of chromosome X. Each of the 144 putative enhancers was subjected to Sanger sequencing to verify the presence of the major allele, cloned into the luciferase expression plasmid, and then transfected into the three cell lines. Additionally, we assessed the activity of each region in both the forward and reverse orientations with respect to the minimal promoter element.

Of the 144 regions tested, 13 demonstrated ‘strong’ (*i.e.* greater than five-fold activity over an empty control) regulatory activity in S16 cells: SC21-13, SC21-16, SC21-20, SCX-3, SCX-4, SCX-21, SCX-39, SCX-58, SCX-60, SCX-65, SCX-67, SCX-78, and SCX-81 (Figure 2.4). In experiments using MN-1 cells, 11 of the 144 regions demonstrated ‘strong’ regulatory activity: SC21-10, SC21-12, SC22-1, SC22-8, SCX-3, SCX-4, SCX-21, SCX-45, SCX-58, SCX-60, and SCX-63 (Figure 2.5). We observed ‘strong’ regulatory activity in 21 of 144 regions in a C2C12 cells: SC21-10, SC21-16, SC21-18, SC21-27, SC21-33, SC21-34, SC22-1, SC22-8, SC22-14, SCX-3, SCX-4, SCX-18, SCX-20, SCX-21, SCX-33, SCX-45, SCX-52, SCX-58, SCX-60, SCX-63, and SCX-67 (Figure 2.6). In sum, we identified 28 unique regions out of 144 regions tested (19.4%) with strong regulatory activity in at least one cell line (Table 2.1).

The SNP Significantly Affects the Regulatory Activity of 13 Regions

To determine if the SNP has any effect on regulatory activity, all 28 regions with ‘strong’ activity in at least one cell line were mutagenized to the minor allele, and reassessed in the relevant cell line(s). If a region displayed strong activity in more than one cell line, then the minor allele was assessed in all cell lines where the major allele demonstrated ‘strong’ activity. Each allele was tested in both orientations regardless of the original orientation activity. The more active allele of each region was normalized to a value of ‘100’, and the less active allele activity is relative to the more active allele. In Schwann cells, seven of the 13 active regions (53.8%) demonstrated allele-specific differences in regulatory activity: SC21-13, SCX-4, SCX-58, SCX-60, SCX-67, SCX-78, and SCX-81 (Figure 2.7). In motor neurons, four of the 11 active regions (36.4%) demonstrated allele-specific differences in regulatory activity: SC21-10, SCX-4, SCX-58, and SCX-60 (Figure 2.8). While in the C2C12 cell line, seven of the 21 active regions

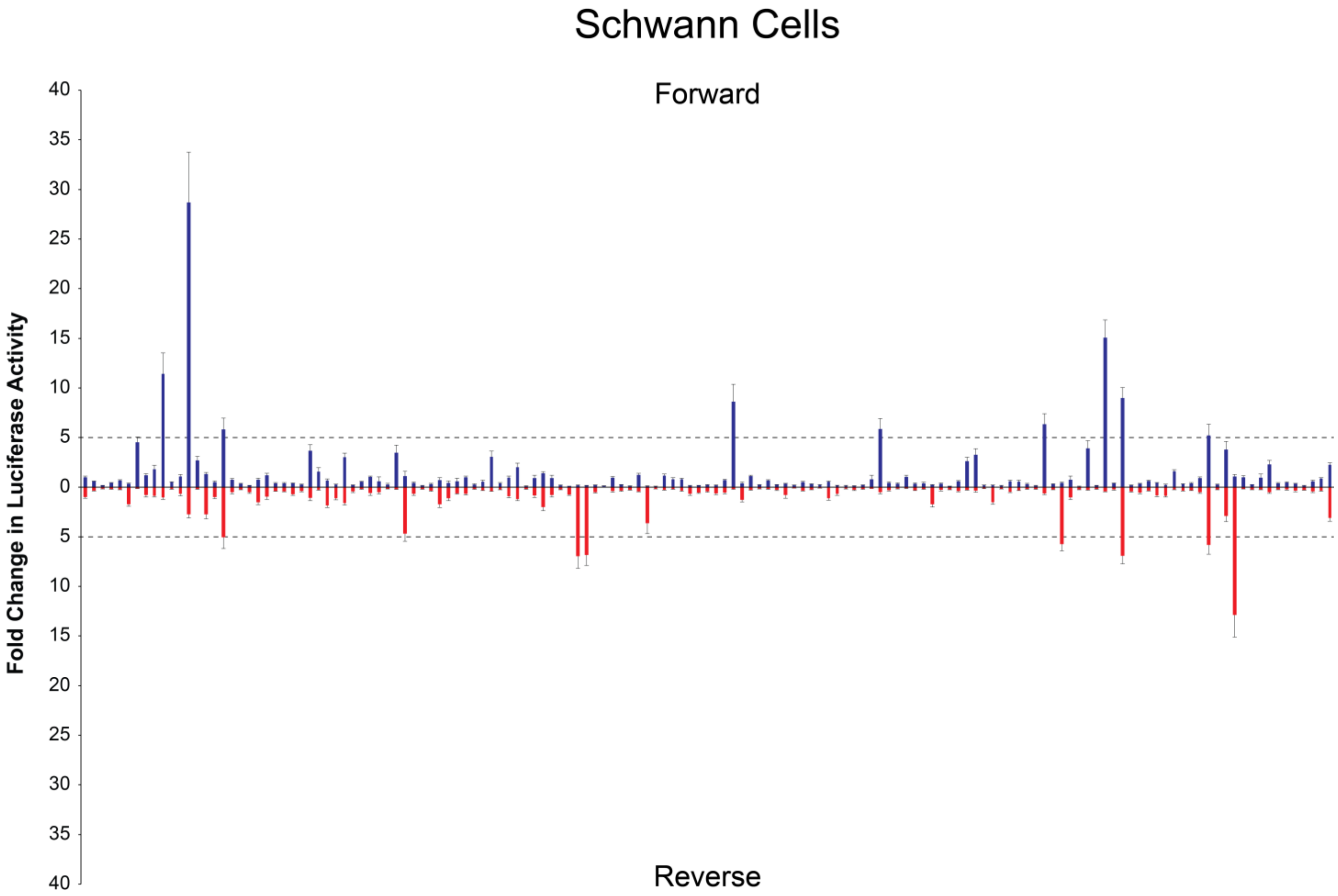


Figure 2.4 *Identification of Regulatory Activity for Regions on Chromosomes 21, 22, and X in Schwann Cells.* All 144 genomic regions containing the major allele were cloned upstream of a luciferase reporter gene and tested in the forward (Top) or reverse (Bottom) orientations. Luciferase activity is measured relative to a renilla control vector. The activity of each genomic segment is expressed relative to a control vector with no insert (‘Empty’) whose activity has been set to ‘1’. A dashed line is set to a five-fold increase in activity over the empty control and indicates ‘strong’ enhancer activity, and error bars represent standard deviation.

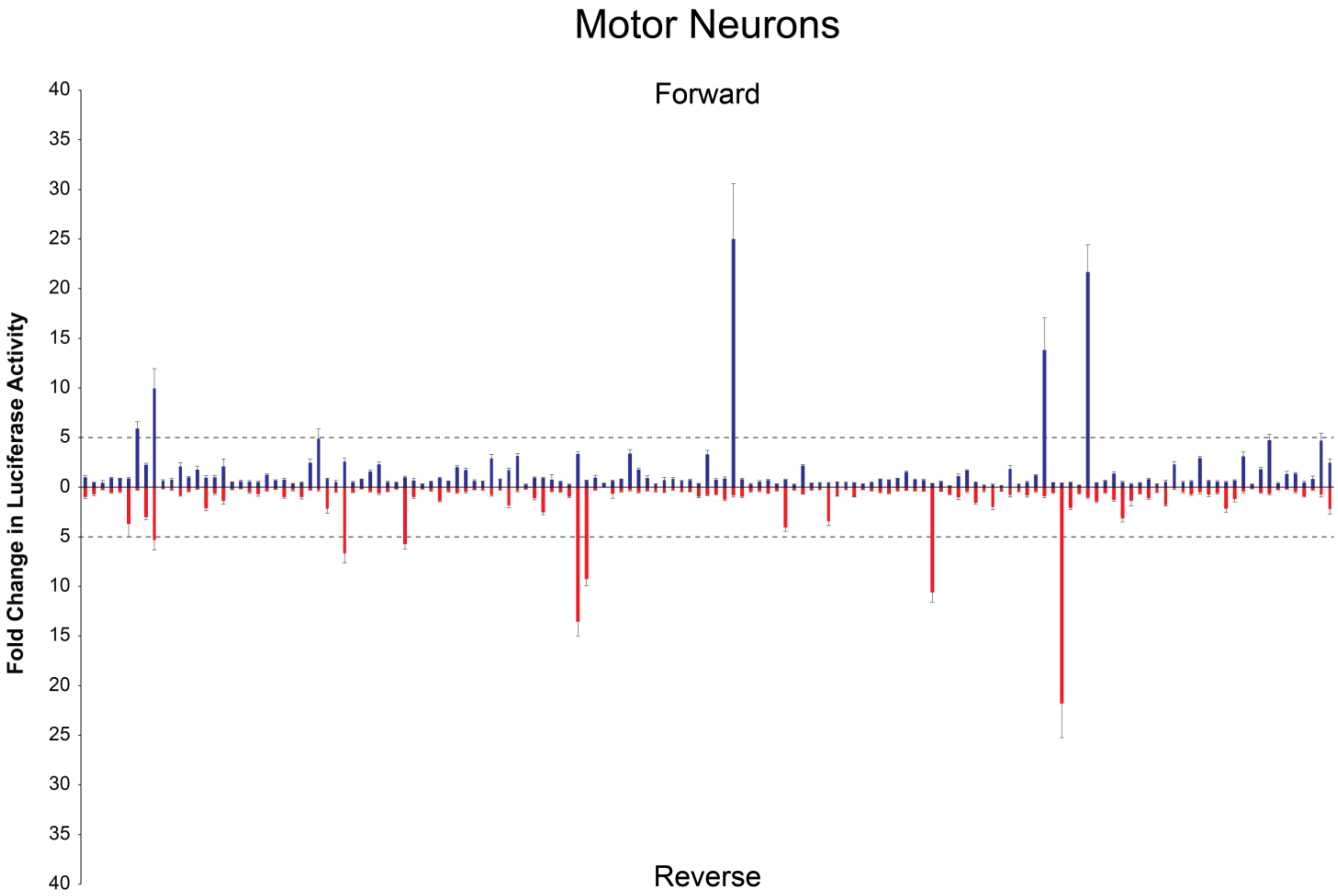


Figure 2.5 *Identification of Regulatory Activity for Regions on Chromosomes 21, 22, and X in Motor Neurons.* All 144 genomic regions containing the major allele were cloned upstream of a luciferase reporter gene and tested in the forward (Top) or reverse (Bottom) orientations. Luciferase activity is measured relative to a renilla control vector. The activity of each genomic segment is expressed relative to a control vector with no insert ('Empty') whose activity has been set to '1'. A dashed line is set to a five-fold increase in activity over the empty control and indicates 'strong' enhancer activity, and error bars represent standard deviation.

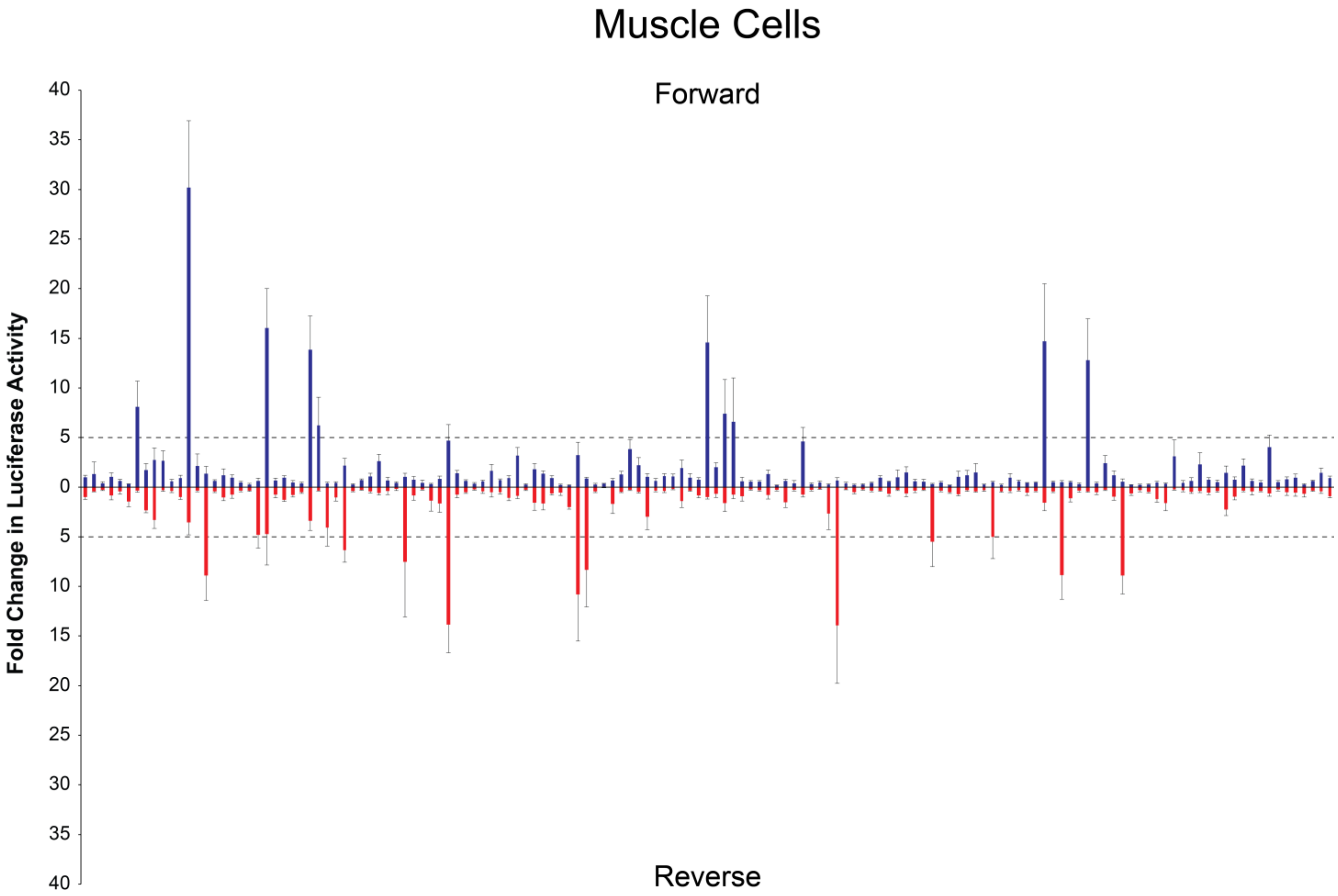


Figure 2.6 *Identification of Regulatory Activity for Regions on Chromosomes 21, 22, and X in Muscle Cells.* All 144 genomic regions containing the major allele were cloned upstream of a luciferase reporter gene and tested in the forward (Top) or reverse (Bottom) orientations. Luciferase activity is measured relative to a renilla control vector. The activity of each genomic segment is expressed relative to a control vector with no insert ('Empty') whose activity has been set to '1'. A dashed line is set to a five-fold increase in activity over the empty control and indicates 'strong' enhancer activity, and error bars represent standard deviation.

Table 2.1 *Luciferase Activity of Regions Displaying 'Strong' Activity.*

Region	Forward[1]	Reverse[1]	Coordinates (hg18)[2]	rs Number
SC21-10	5.91 (M) 8.10 (C)	0.25 (M) 0.33 (C)	chr21:21313182-21313188	rs7277262
SC21-12	9.95 (M)	5.34 (M)	chr21:22535874-22535880	rs2827297
SC21-13	11.42 (S)	1.04 (S)	chr21:27318695-27318702	rs233616
SC21-16	28.67 (S) 30.17 (C)	2.73 (S) 3.56 (C)	chr21:29424534-29424540	rs2832203
SC21-18	1.37 (C)	8.90 (C)	chr21:33139128-33139135	rs2833975
SC21-20	5.81 (S)	5.04 (S)	chr21:33273214-33273219	rs2834040
SC21-27	16.05 (C)	4.71 (C)	chr21:36269198-36269214	rs2835196
SC21-33	13.85 (C)	3.40 (C)	chr21:38940551-38940556	rs16996658
SC21-34	6.22 (C)	0.33 (C)	chr21:38958107-38958118	rs8130434
SC22-1	2.58 (M) 2.17 (C)	6.68 (M) 6.35 (C)	chr22:16689437-16689442	rs5992119
SC22-8	1.00 (M) 1.05 (C)	5.73 (M) 7.53 (C)	chr22:25678449-25678472	rs5761863
SC22-14	4.70 (C)	13.83 (C)	chr22:26146779-26146784	rs733164
SCX-3	0.20 (S) 3.34 (M) 3.21 (C)	6.97 (S) 13.56 (M) 10.79 (C)	chrX:15529186-15529192	rs4646115
SCX-4	0.17 (S) 0.67 (M) 0.86 (C)	6.85 (S) 9.25 (M) 8.31 (C)	chrX:17730099-17730104	rs2187846
SCX-18	14.59 (C)	0.98 (C)	chrX:31252044-31252049	rs7884417
SCX-20	7.39 (C)	1.59 (C)	chrX:31435143-31435149	rs3788892
SCX-21	8.63 (S) 25.00 (M) 6.60 (C)	0.18 (S) 0.84 (M) 0.75 (C)	chrX:31764702-31764707	rs1379871
SCX-33	0.67 (C)	13.92 (C)	chrX:85443058-85443064	rs6623642
SCX-39	5.84 (S)	0.50 (S)	chrX:86429198-86429205	rs16980794
SCX-45	0.37 (M) 0.31 (C)	10.62 (M) 5.49 (C)	chrX:92656893-92656900	rs12687113
SCX-52	0.50 (C)	5.02 (C)	chrX:99454087-99454093	rs7064056
SCX-58	6.33 (S) 13.83 (M) 14.69 (C)	0.64 (S) 0.93 (M) 1.56 (C)	chrX:121682472-121682478	rs17273301
SCX-60	0.43 (S) 0.41 (M) 0.50 (C)	5.73 (S) 21.81 (M) 8.86 (C)	chrX:123382405-123382410	rs2076164
SCX-63	21.68 (M) 12.81 (C)	1.04 (M) 0.42 (C)	chrX:125247437-125247470	rs16998722
SCX-65	15.05 (S)	0.41 (S)	chrX:125925884-125925893	rs5930055
SCX-67	8.97 (S) 0.56 (C)	6.90 (S) 8.91 (C)	chrX:127229700-127229720	rs17266605
SCX-78	5.21 (S)	5.80 (S)	chrX:146960885-146960891	rs6525876
SCX-81	1.09 (S)	12.85 (S)	chrX:147430625-147430635	rs17252118

[1] Activity only shown for cell line(s) where region was active. S = S16, M = MN-1, C = C2C12.
[2] Coordinates for conserved region in hg18.

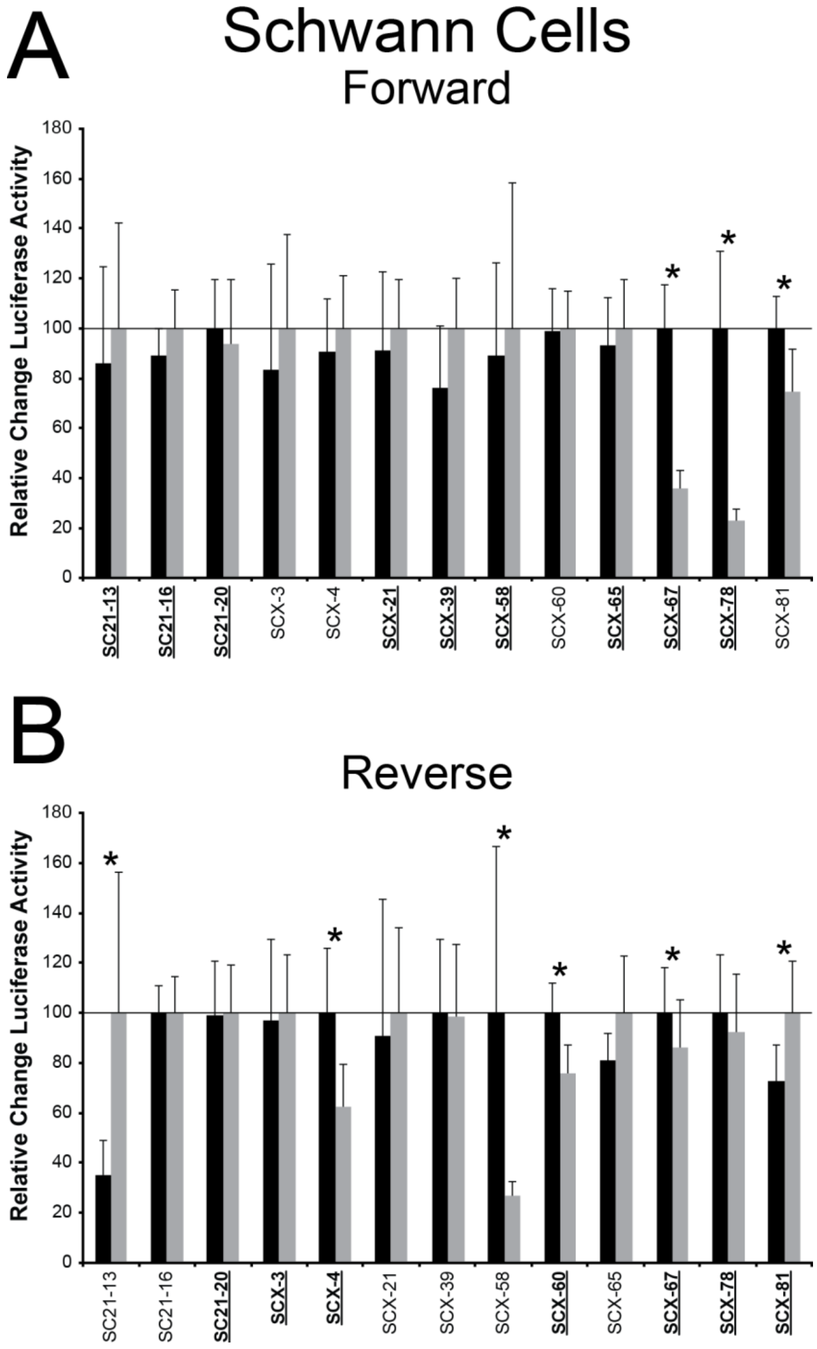


Figure 2.7 *Seven Regions Display a Significant Effect of the SNP on Luciferase Activity in Schwann Cells.* The 13 regions displaying ‘strong’ activity in Schwann cells were mutagenized to the minor allele and reassessed in the forward (A) or reverse (B) orientations. The relative activity between the two alleles is displayed with the more active allele set to a value of ‘100’. Black and grey bars represent the major and minor allele respectively. Bold and underlined regions indicate the more active orientation, and the asterisks represent a significant ($p < 0.05$) difference between the two alleles using the Student’s T test. Bars represent at least eight technical replicates.

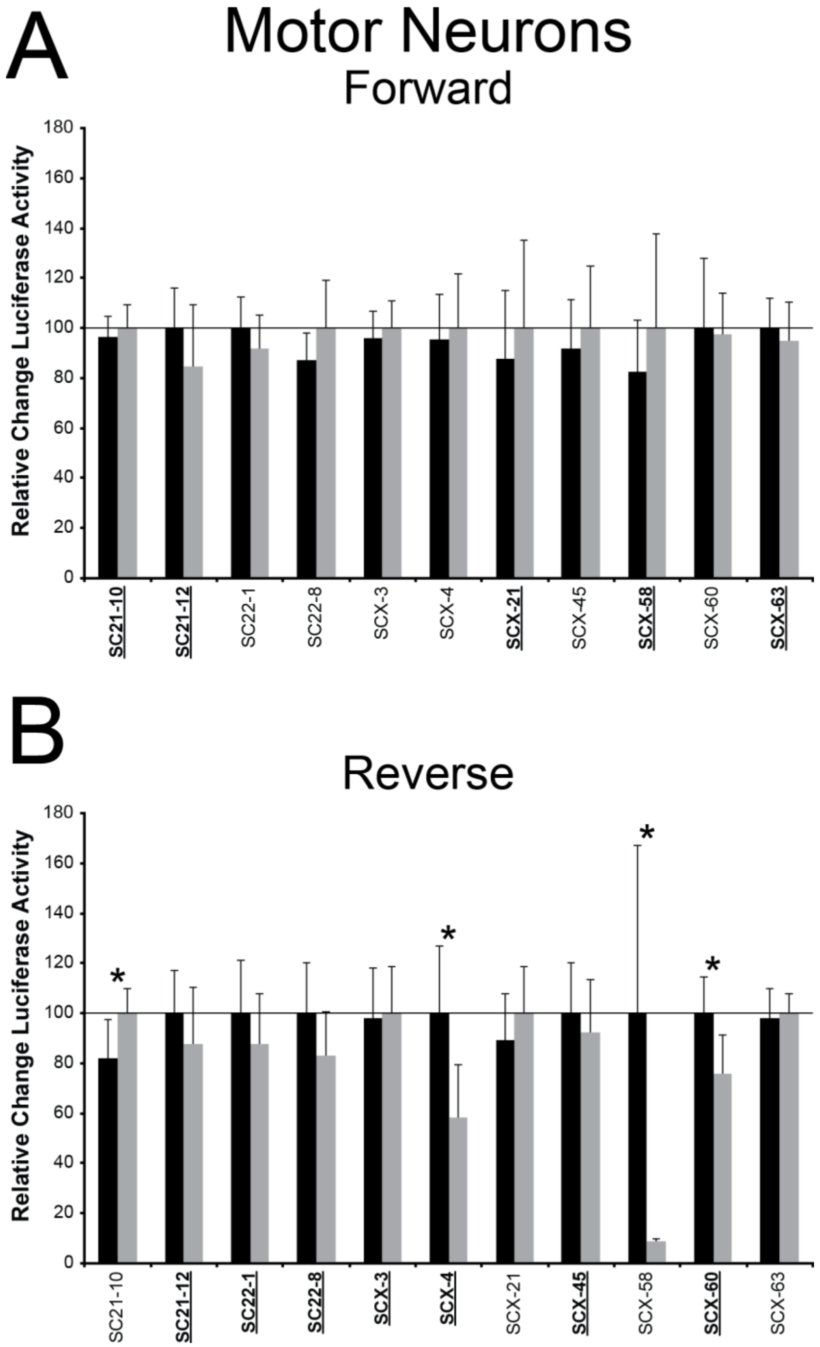


Figure 2.8 *Four Regions Display a Significant Effect of the SNP on Luciferase Activity in Motor Neurons.* The 11 regions displaying ‘strong’ activity in Schwann cells were mutagenized to the minor allele and reassessed in the forward (A) or reverse (B) orientations. The relative activity between the two alleles is displayed with the more active allele set to a value of ‘100’. Black and grey bars represent the major and minor allele respectively. Bold and underlined regions indicate the more active orientation, and the asterisks represent a significant ($p < 0.05$) difference between the two alleles using the Student’s T test. Bars represent at least eight technical replicates.

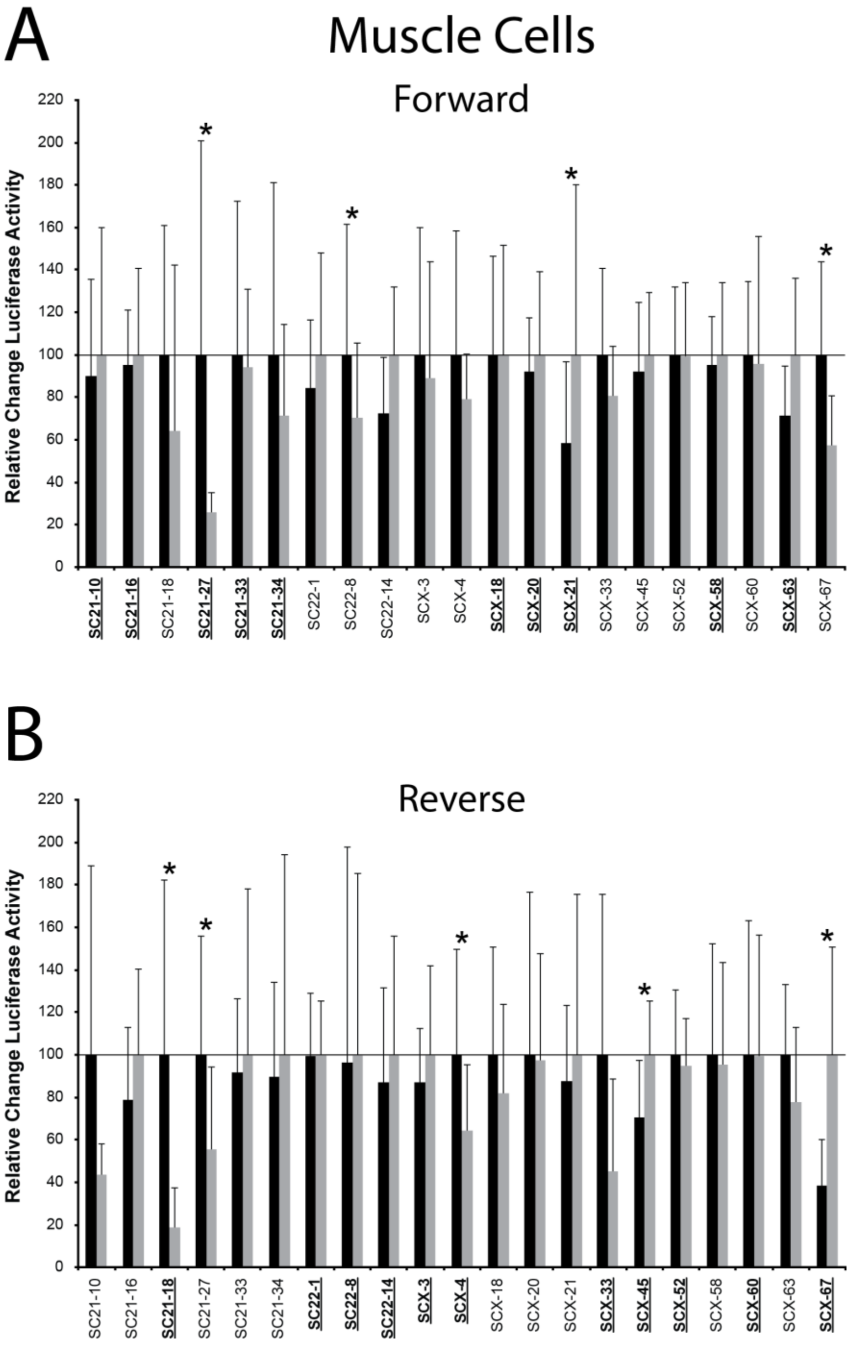


Figure 2.9 *Seven Regions Display a Significant Effect of the SNP on Luciferase Activity in Muscle Cells.* The 21 regions displaying ‘strong’ activity in Schwann cells were mutagenized to the minor allele and reassessed in the forward (A) or reverse (B) orientations. The relative activity between the two alleles is displayed with the more active allele set to a value of ‘100’. Black and grey bars represent the major and minor allele respectively. Bold and underlined regions indicate the more active orientation, and the asterisks represent a significant ($p < 0.05$) difference between the two alleles using the Student’s T test. Bars represent at least eight technical replicates.

(33.3%) demonstrated allele-specific differences in regulatory activity: SC21-18, SC21-27, SC22-8, SCX-4, SCX-21, SCX-45, and SCX-67 (Figure 2.9). Taken together, 13 SNPs of the 28 (46.4%) strong regions demonstrated allele-specific differences in regulatory activity in at least one cell line relevant to the peripheral nerve.

Predicting Differential Transcription Factor Binding to Active Regions Harboring a Putative rSNP Using TRANSFAC

One possible explanation for the allele specific differences is that the SNP may alter transcription factor binding to the regions. To predict transcription factors that may differentially bind to the major or minor allele of the active regions, we used an *in silico* transcription factor binding site prediction program, TRANSFAC (Matys et al. 2003). Briefly, a total of 30 base pairs surrounding (and including) each conserved region containing the major allele was generated. Next, we generated the minor allele sequence by substituting in the minor SNP allele. Both sequences were uploaded to TRANSFAC then analyzed for TFBSs using the TRANSFAC Match algorithm (Kel et al. 2003). We used the vertebrate database of transcription factors and minimized the sum of the false positive and false negative error rates.

Because the different alleles of the SNP altered the regulatory activity of these regions, the results were filtered to display only unique differences in predicted TFBSs. In Schwann cells, all seven regions assessed had at least one predicted TFBS unique to either the major or minor allele (Figure 2.10). Interestingly, none of the three Schwann cell specific regions (SC21-13, SCX-78, and SCX-81) harbored any predicted TFBS of transcription factors known to be important for

Schwann cells. While these results may be due to the limitations of TRANSFAC, they may also illustrate potentially novel roles of the predicted transcription factors in Schwann cells.

Conversely in the motor neurons, one region, SC21-10, did not have any unique predictions (Figure 2.11). There were four predicted TFBS for SC21-10 for the major allele, however all four TFBS were also predicted at the same location and orientation in the minor allele, and therefore no predicted TFBS are displayed. This may indicate that the SNP does not ablate a TFBS, but rather alters binding site affinity. Unlike Schwann cells, no region which demonstrated allele-specific differences was specific to motor neurons. The four regions (SC21-10, SCX-4, SCX-58, and SCX-60) all displayed regulatory activity in at least one additional cell line.

Similar to Schwann cells, all regions assessed in muscle had at least one unique TFBS prediction in either the major or minor allele (Figure 2.12). Only two of the regions, however, were limited in regulatory activity to muscle cells (SC21-18 and SC21-27). Strikingly, a putative LEF-1 binding site is created within the minor allele of SC21-27. *Lef-1* expression has been shown to be upregulated in muscle cells within mice following an injury to the muscle (Amin et al. 2014). Within our data the minor allele of SC21-27 demonstrated significantly less regulatory activity compared to the major allele. This result could be due to loss of an enhancer element within the major allele, or the gain of a repressive element in the minor allele. Indeed, Lef-1 has been shown to act as a transcriptional repressor (Billin et al. 2000; Mao and Byers 2011);however further study will be necessary to determine the role (if any) of this putative LEF-1 binding site.

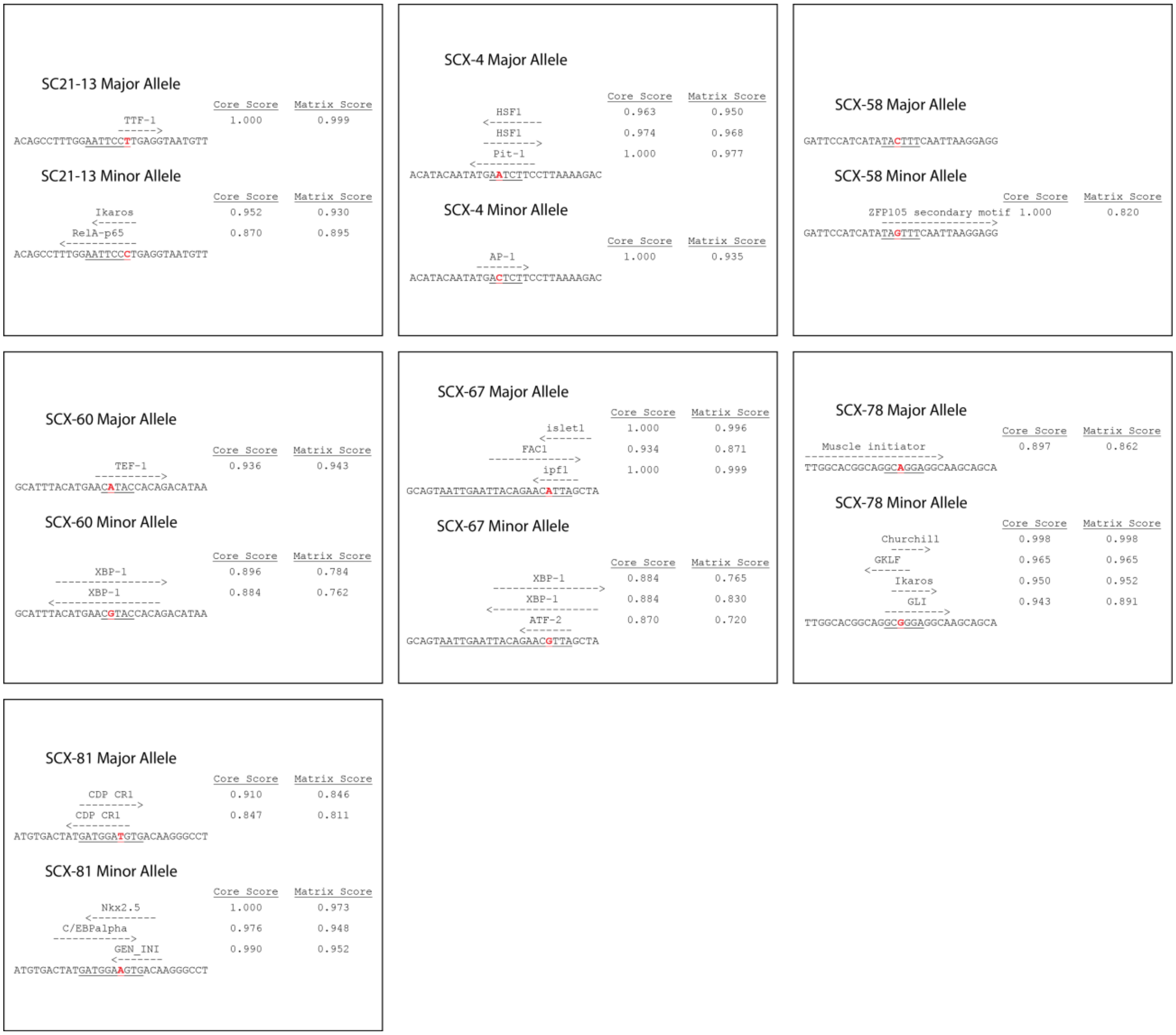


Figure 2.10 *TRANSFAC Predictions of Differential Transcription Factor Binding Sites of the Seven rSNPs in Schwann Cells.* TRANSFAC was used to predict differential TFBSs between the major and minor alleles for SNPs that had a significant effect on luciferase activity. A total of 30 base pairs surrounding the conserved sequences of the seven 'strong' regions were used as input sequence for TRANSFAC. The dashed arrows indicate the position and direction of the predicted TFBS, the transcription factor is indicated above the arrows, and the core and matrix scores are indicated at the right. Only unique differential TFBS predictions between the major and minor alleles are displayed. The underlined base pairs indicate the conserved bases, and the SNP is shown in red and bold.

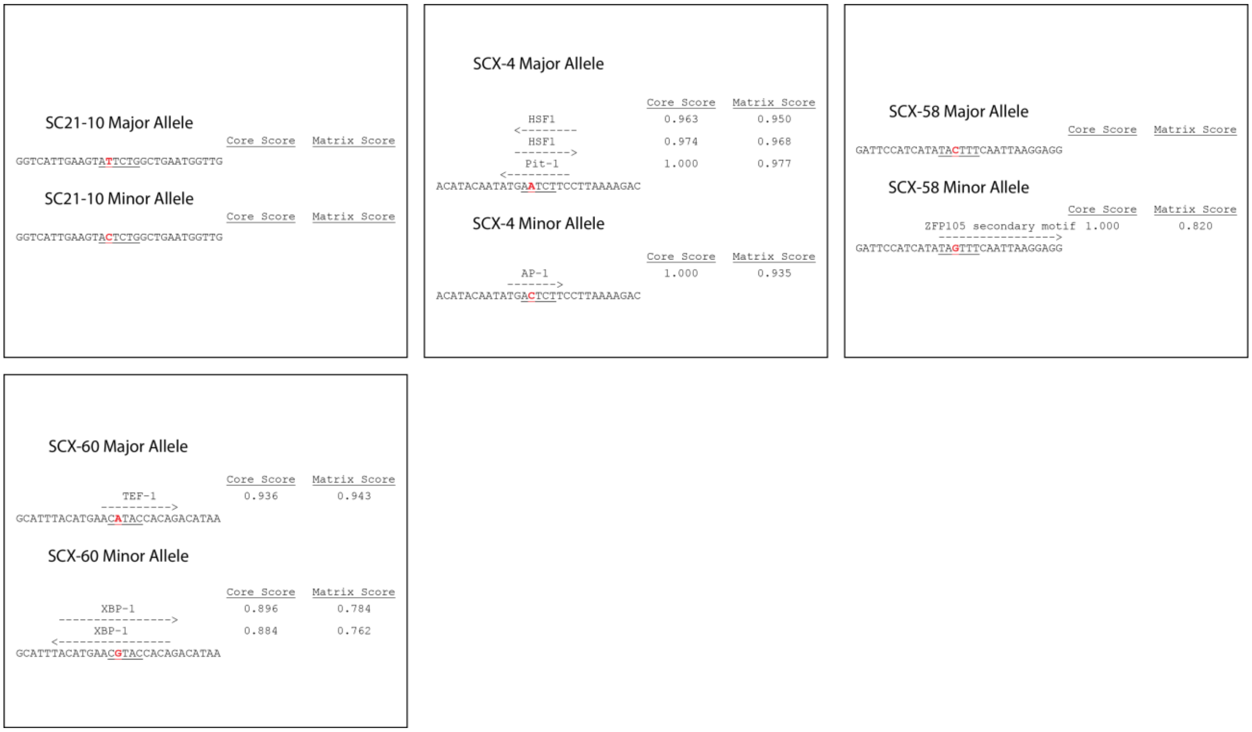
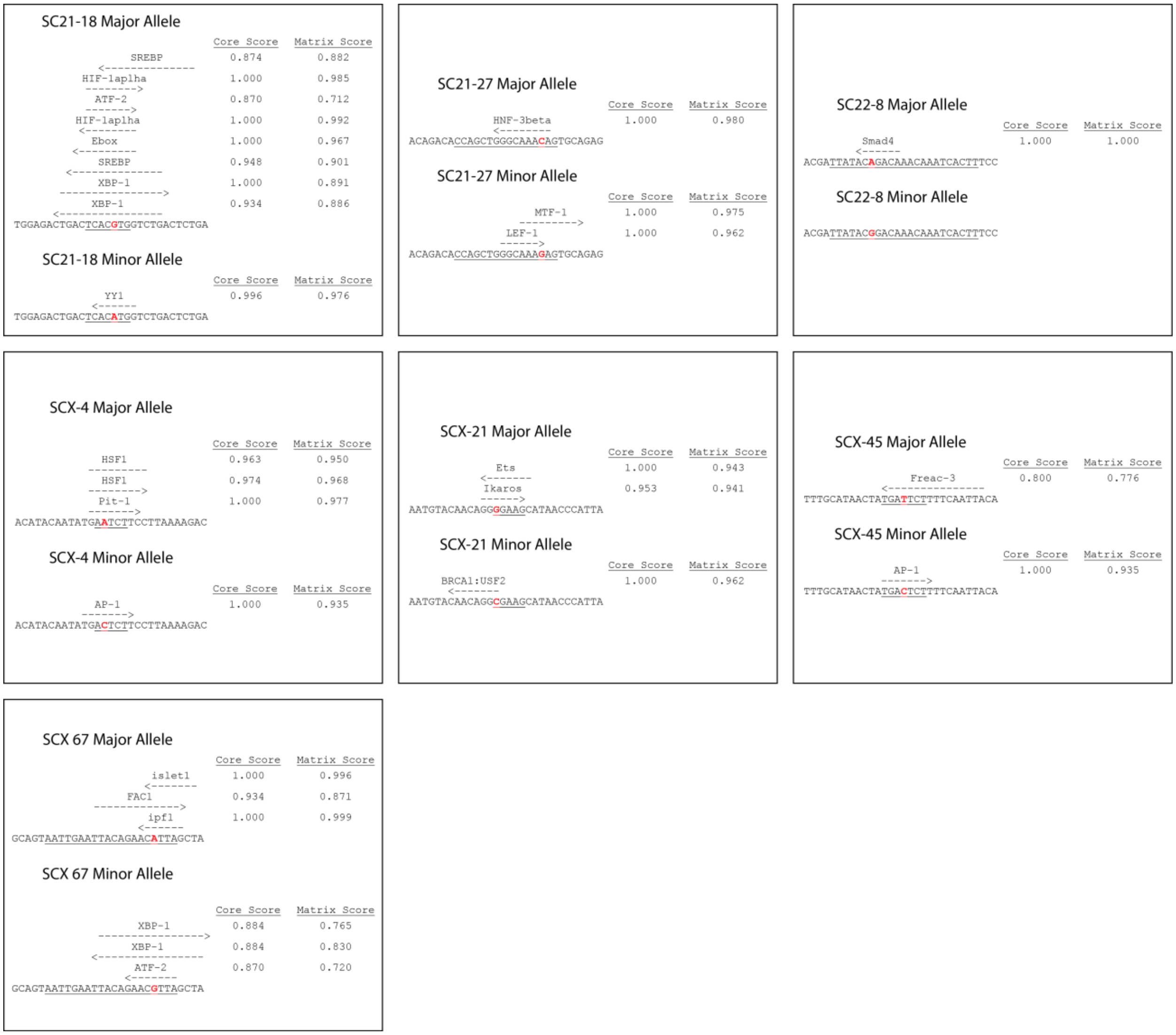


Figure 2.11 *TRANSFAC Predictions of Differential Transcription Factor Binding Sites of the Four rSNPs in Motor Neurons.* TRANSFAC was used to predict differential TFBSs between the major and minor alleles for SNPs that had a significant effect on luciferase activity. A total of 30 base pairs surrounding the conserved sequences of the seven 'strong' regions were used as input sequence for TRANSFAC. The dashed arrows indicate the position and direction of the predicted TFBS, the transcription factor is indicated above the arrows, and the core and matrix scores are indicated at the right. Only unique differential TFBS predictions between the major and minor alleles are displayed. The underlined base pairs indicate the conserved bases, and the SNP is shown in red and bold.



SC21-18 Major Allele

```
                               Core Score   Matrix Score
                  SREBP           0.874         0.882
             <--------------
          HIF-1aplha              1.000         0.985
           -------->
            ATF-2                 0.870         0.712
           ------->
         HIF-1aplha               1.000         0.992
          <--------
            Ebox                  1.000         0.967
         <---------
             SREBP                0.948         0.901
        <--------------
             XBP-1                1.000         0.891
       ---------------->
            XBP-1                 0.934         0.886
      <----------------
TGGAGACTGACTCACGTGGTCTGACTCTGA
```

SC21-18 Minor Allele

```
                               Core Score   Matrix Score
             YY1                  0.996         0.976
           <------
TGGAGACTGACTCACATGGTCTGACTCTGA
```

SC21-27 Major Allele

```
                               Core Score   Matrix Score
           HNF-3beta              1.000         0.980
           <--------
ACAGACACCAGCTGGGCAAACAGTGCAGAG
```

SC21-27 Minor Allele

```
                               Core Score   Matrix Score
                 MTF-1            1.000         0.975
               --------->
            LEF-1                 1.000         0.962
           ------>
ACAGACACCAGCTGGGCAAAGAGTGCAGAG
```

SC22-8 Major Allele

```
                               Core Score   Matrix Score
          Smad4                   1.000         1.000
         <------
ACGATTATACAGACAAACAAATCACTTTCC
```

SC22-8 Minor Allele

```
ACGATTATACGGACAAACAAATCACTTTCC
```

SCX-4 Major Allele

```
                               Core Score   Matrix Score
            HSF1                  0.963         0.950
          ---------
            HSF1                  0.974         0.968
          -------->
           Pit-1                  1.000         0.977
          ------->
ACATACAATATGAATCTTCCTTAAAAGAC
```

SCX-4 Minor Allele

```
                               Core Score   Matrix Score
           AP-1                   1.000         0.935
         ------->
ACATACAATATGACTCTTCCTTAAAAGAC
```

SCX-21 Major Allele

```
                               Core Score   Matrix Score
               Ets                1.000         0.943
             <-------
              Ikaros              0.953         0.941
             ------>
AATGTACAACAGGGGAAGCATAACCCATTA
```

SCX-21 Minor Allele

```
                               Core Score   Matrix Score
        BRCA1:USF2                1.000         0.962
         <-------
AATGTACAACAGGCGAAGCATAACCCATTA
```

SCX-45 Major Allele

```
                               Core Score   Matrix Score
                  Freac-3         0.800         0.776
             <---------------
TTTGCATAACTATGATTCTTTTCAATTACA
```

SCX-45 Minor Allele

```
                               Core Score   Matrix Score
              AP-1                1.000         0.935
             ------->
TTTGCATAACTATGACTCTTTTCAATTACA
```

SCX 67 Major Allele

```
                                Core Score   Matrix Score
                     islet1        1.000         0.996
                    <-------
                FAC1               0.934         0.871
           ------------->
                    ipf1           1.000         0.999
                  <------
GCAGTAATTGAATTACAGAACATTAGCTA
```

SCX 67 Minor Allele

```
                                Core Score   Matrix Score
                  XBP-1            0.884         0.765
            ---------------->
                 XBP-1             0.884         0.830
           <----------------
                  ATF-2            0.870         0.720
                <-------
GCAGTAATTGAATTACAGAACGTTAGCTA
```

Figure 2.12 *TRANSFAC Predictions of Differential Transcription Factor Binding Sites of the Seven rSNPs in Muscle Cells.* TRANSFAC was used to predict differential TFBSs between the major and minor alleles for SNPs that had a significant effect on luciferase activity. A total of 30 base pairs surrounding the conserved sequences of the seven 'strong' regions were used as input sequence for TRANSFAC. The dashed arrows indicate the position and direction of the predicted TFBS, the transcription factor is indicated above the arrows, and the core and matrix scores are indicated at the right. Only unique differential TFBS predictions between the major and minor alleles are displayed. The underlined base pairs indicate the conserved bases, and the SNP is shown in red and bold.

Discussion

Regulatory variation is known to modify the severity or risk of human diseases (Emison et al. 2005) and in some cases, cause genetic diseases (Houlden et al. 2004; Gobbi et al. 2006). Understanding how regulatory variation impacts gene expression may help explain disease etiology; however significant challenges remain in identification of cis-regulatory elements and functional variation. In this chapter, we used a combination of phylogenetic footprinting and whole genome datasets to computationally predict regulatory regions harboring a putative rSNP. The human, mouse, and chicken genomes were used for our conservation analysis because we were interested in identifying functional SNPs critical for the peripheral nervous system. The zebrafish genome was not included for a few reasons, including: (1) the zebrafish genome is partially duplicated, which makes DNA sequence alignments more challenging (Meyer and Schartl 1999; Howe et al. 2013), and (2) we previously identified functional CREs that are conserved between human, mouse, and chicken, but that were not conserved in zebrafish (Antonellis et al. 2008; Hodonsky et al. 2012). Despite these observations, inclusion of a more distally related species, such as zebrafish, has been shown to reduce false positive results when using phylogenetic footprinting assays and may help prioritize future datasets to generate more accurate predictions (Nobrega et al. 2003).

We validated a subset of the genome-wide computational predictions using luciferase assays. Our analysis was performed in a transcription factor blind approach and revealed 6,164 regions at least five base pairs in length and conserved among human, mouse, and chicken that harbored a SNP. A subset of these regions on chromosome 21 (37 regions), 22 (29 regions), and X (93 regions) were assessed in luciferase assays in three cell lines to approximate a peripheral nerve and muscle target: Schwann cells (S16), motor neurons (MN-1) and muscle cells (C2C12). We selected chromosomes 21, 22, and X to ensure our pipeline could accurately predict CREs and putative rSNPs on both the autosomes and sex chromosomes. Additionally in humans, *SOX10* resides on chromosome 22 and as discussed in chapter 1, SOX10 is critical for proper Schwann cell development and function (see chapter 1). Identification of regulatory variation impacting *SOX10* expression could lead to a greater phenotypic variation because SOX10 regulates many critical Schwann cell genes.

We identified 28 regions with 'strong' regulatory activity when assessing the major allele. Surprisingly, only four of the 28 regions displayed orientation independent enhancer activity (*i.e.* 'strong' activity in both orientations). While this orientation dependence may be inherent to our experimental design, such as the inability of the region to interact with the promoter due to steric hindrance, there are examples of orientation dependent enhancer activity (Nishimura et al. 2000; Wei and Brennan 2000; Swamynathan and Piatigorsky 2002). Additionally, there is no relationship between the more active orientation and position of the SNP within the PCR amplicon (Figure 2.13). Further studies of these regions will be necessary to determine if the orientation dependence is inherent to the CRE or an artifact of our luciferase expression plasmid.

It is important to note that we only tested the major allele activity in our luciferase assays, thus excluding the potential for the minor allele to display 'strong' regulatory activity through the creation of a novel TFBS or increasing the binding affinity of an existing TFBS. Our study design also excludes the potential for the SNP alleles to affect a repressive regulatory element as we required the major allele to display 'strong' activity. Future studies could modify the luciferase assay to address these other potential SNP affects, such as replacing the minimal E1B promoter with an active promoter element. This would potentially allow for the identification of

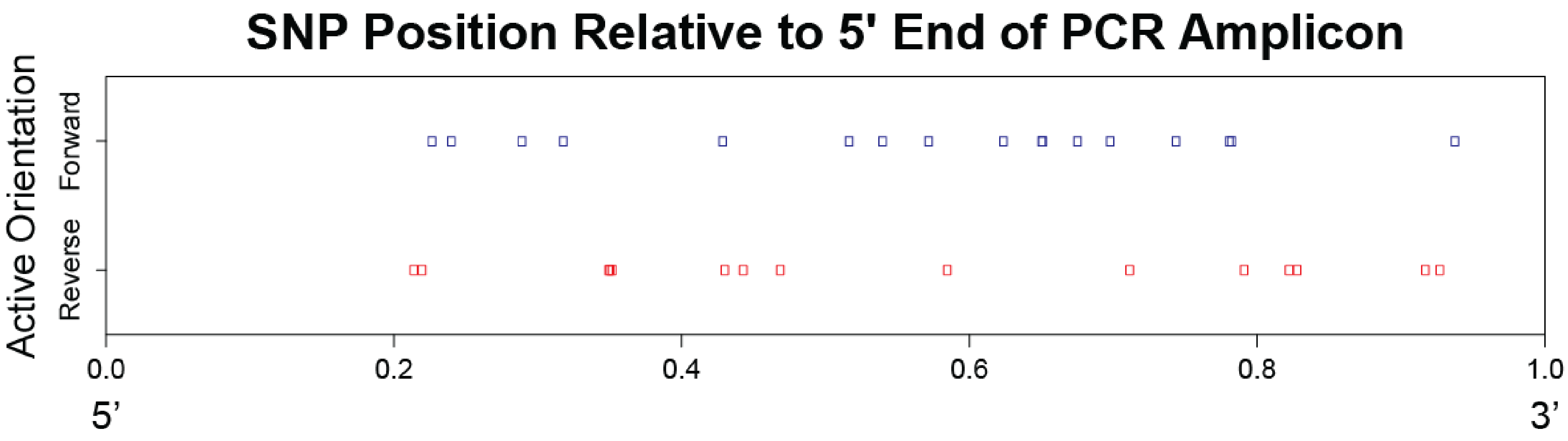


Figure 2.13 *Relative Location of SNPs Within 'Strong' Region PCR Amplicons*. For the 28 active regions, the relative position of the SNP is displayed. The distance from the 5' end of the PCR amplicon to the 5' end of the SNP was determined and divided by the total length of the PCR amplicon. Each square represents an active region, and the regions were separated by active orientation (Forward; blue boxes or Reverse; red boxes). The location on the X-axis represents the percent distance from the 5' end of the PCR amplicon (0 = at the 5' end and 1 = at the 3' end). Four regions were active in both orientations and were included in both the forward and the reverse datasets.

repressive elements within the major allele, and mutation to the minor allele could test for an allele-specific increase in activity.

To assess for potential repressor elements within our regions, we generated lists of the top 20 regions displaying the least regulatory activity for each cell line and orientation. Interestingly, four regions in the forward orientation were in the top 20 least active regions common to all three cell lines (Table 2.2 - Shaded Boxes). An additional five regions were in the top 20 least active regions common to all cell lines in the reverse orientation (Table 2.2 - Unshaded Boxes). One region in particular, SCX-6, was the least active region in the reverse orientation for all three cell lines and was in the top 20 least active regions in the forward orientation in both S16 and MN-1 cells. It was the 24th least active region in the forward orientation in C2C12 cells. This region is a strong candidate for harboring repressive elements within the PCR amplicon. Additional efforts should focus on SCX-6 and other regions that displayed levels of activity much lower than the empty control vector that does not harbor a genomic insert.

While our computational model was successful in predicting CREs, we were only able to test a subset of our predictions. A more high throughput assay such as STARR-Seq (Arnold et al. 2013) or Cre-Seq (Kwasnieski et al. 2012) could be employed to test a greater number of our computational predictions. Based on our rates of identification of active regions (greater than five-fold; 19.4%) and rSNPs within active regions (46.4%), we anticipate identification of 1,195 active regions of which 554 regions would contain a rSNPs. While we should be cautious when extrapolating from data, the previously identified regions in Chapter 2 demonstrate there are a large number of CREs remaining to be identified that are functional in the peripheral nerve or muscle cells. Finally, our five-fold threshold for luciferase activity only allows for identification of relatively strong enhancer elements. While these ’strong’ regulatory regions give greater

Table 2.2 *Nine Regions Common to Top 20 Least Active Regions Across the Three Cell Lines.*

Region	Forward[1]	Reverse[1]	Coordinates (hg18)[2]	rs Number
SCX-2	0.10 (S) 0.26 (M) 0.19 (C)	0.77 (S) 0.95 (M) 1.99 (C)	chrX:13701570-13701575	rs13187
SCX-47	0.10 (S) 0.14 (M) 0.17 (C)	0.24 (S) 0.71 (M) 0.49 (C)	chrX:94777246-94777251	rs5990383
SCX-53	0.14 (S) 0.16 (M) 0.23 (C)	0.17 (S) 0.37 (M) 0.42 (C)	chrX:103711687-103711695	rs1004122
SCX-62	0.07 (S) 0.18 (M) 0.31 (C)	0.23 (S) 0.61 (M) 0.35 (C)	chrX:124339837-124339845	rs3126112
SC21-3	0.17 (S) 0.42 (M) 0.40 (C)	0.16 (S) 0.24 (M) 0.08 (C)	chr21:15652270-15652285	rs16982386
SC22-10	0.15 (S) 0.28 (M) 0.42 (C)	0.18 (S) 0.20 (M) 0.06 (C)	chr22:25769441-25769477	rs17429199
SCX-6	0.14 (S) 0.37 (M) 0.35 (C)	0.02 (S) 0.03 (M) 0.01 (C)	chrX:22421316-22421332	rs5970650
SCX-31	0.24 (S) 0.43 (M) 0.46 (C)	0.10 (S) 0.22 (M) 0.05 (C)	chrX:85404361-85404385	rs6653101
SCX-34	0.15 (S) 0.48 (M) 0.38 (C)	0.14 (S) 0.24 (M) 0.04 (C)	chrX:85764274-85764279	rs16980611

[1] Activity only shown for cell line(s) where region was active. S=S16, M=MN-1, C=C2C12.
[2] Coordinates for conserved region in hg18.

The top 20 least active regions for each cell line was determined for both orientations independently. The regions displayed are in the top 20 least active and are common to all three cell lines for a given orientation. The shaded rows indicate regions in the forward orientation that were within the top 20 least active and common in all three cell lines, while the unshaded rows indicate the same criteria for the reverse orientation.

confidence for assessing a functional role *in vivo*, it is known that low-affinity binding sites (which may not exceed our threshold) can be critical for endogenous gene expression (Rowan et al. 2010; Ramos and Barolo 2013). Despite these limitations, we were able to detect an allele-specific difference in 13 of the 28 active regions. Interestingly, if a region was active in multiple cell lines, the SNP did not necessarily affect them all similarly. For example, the major alleles of SCX-58 and SCX-60 demonstrated strong luciferase activity in all three cell lines, but the minor allele only decreased activity in Schwann cells and motor neurons with no effect in muscle cells. Additionally, SCX-67 was active in both orientations in Schwann cells, and the minor allele significantly decreased activity in both the forward and reverse orientations. The major SNP allele of SCX-67 also displayed 'strong' activity in the reverse orientation in muscle cells, but unlike Schwann cells, the minor SNP allele significantly increased luciferase activity relative to the major allele in the reverse orientation. Conversely other regions displayed identical allele-specific differences across all cell lines where the region was active. For example, the major allele of SCX-4 demonstrated 'strong' luciferase activity in the reverse orientation in all three cell lines, and the minor allele significantly reduced luciferase activity by approximately the same relative change in all three cell lines.

The allele-specific similarities and differences between different cell lines could be informative as to which transcription factor(s) is binding to the region. For instance, SC21-13 demonstrates significantly greater regulatory activity in the minor allele compared to the major allele, and the region only displayed regulatory activity in Schwann cells. From this, we would predict the minor allele either creates a novel TFBS or increases the binding affinity of an existing TFBS. Additionally, since the effect was only observed in Schwann cells, we would anticipate the transcription factor to be expressed in Schwann cells but not in motor neurons or muscle cells.

Conversely, the major allele of SCX-4 displayed a nearly identical decrease across all cell lines. From this, we would predict a more ubiquitous (*i.e.* expressed in at least these three cell lines) TFBS to be disrupted.

Using TRANSFAC, we were able to predict potential TFBSs for each region displaying strong activity. While we only considered whether a transcription factor binding site was predicted in one of the alleles, it may in fact be the effect of the SNP is to alter binding affinity, but not ablate the TFBS. This may be reflected in the difference of either (or both) the core and matrix scores. Further analysis of the TRANSFAC prediction and additional studies, such as overexpression of the predicted transcription factor, will be necessary to elucidate which (if any) of the putative TFBS are functional.

While the primary focus of this chapter was on the identification of regulatory SNPs, it should be noted that the additional 15 regions (from the 28 active regions) which did not display significant differences in regulatory activity between the two alleles, still demonstrated 'strong' regulatory activity in a cell line relevant for the peripheral nerve and are promising candidate enhancer elements. Indeed, some of these regions reside within appealing candidate genes such as SCX-18 and SCX-20. These regions only displayed regulatory activity within muscle cells and both regions reside within *DMD* which when mutated in humans, results in Duchenne muscular dystrophy (Nowak and Davies 2004). Further study will be necessary to elucidate the functional significance of these regions.

In this chapter, we developed a novel computational pipeline to predict CREs harboring SNPs and functionally assessed these predictions in cell types relevant to the peripheral nerve. Our method was successful by identifying 28 unique regions with 'strong' luciferase activity in at least one cell line and 13 of these regions demonstrated significant allele-specific differences in regulatory activity. Both the CREs and rSNPs identified here represent excellent candidate modifiers of peripheral nerve diseases. In Chapter 3, we will modify our existing pipeline to identify rSNPs in a transcription factor centric approach, specifically identifying rSNPs affecting SOX10 binding sites.

CHAPTER 3

Identification of Regulatory SNPs in SOX10 Response Elements

Introduction

Cis-regulatory elements (CREs) have been implicated in many human diseases. These diseases can be caused by duplications (Jones et al. 2011a), deletions (Balemans et al. 2002), or point mutations (Gobbi et al. 2006) that affect CREs. The last class, point mutations, have become increasingly noteworthy because many genome-wide association studies (GWAS) have identified candidate SNPs associated with a disease that do not reside within coding regions of the genome but rather in non-coding regions. Interestingly, many of the variants identified from GWAS reside within genomic features associated with CREs such as DNase I hypersensitivity sites or have histone marks associated with CREs (ENCODE Project Consortium 2012; Maurano et al. 2012). The current prevailing hypothesis is that the SNPs are disrupting transcription factor binding sites (TFBSs) and are either causing disease or modifying disease risk or severity.

SNPs that affect CREs are called regulatory SNPs (rSNPs) and can impact TFBSs in a number of ways, including increasing (Feigelson et al. 1998) or decreasing (Bosma et al. 1995) gene expression, ablating a TFBS (Vasiliev et al. 1999), or creating a novel TFBS (Knight et al. 1999), (Gobbi et al. 2006) (Figure 1.3). While rSNPs may cause disease, they can also modify the risk or severity of disease. For example, a rSNP was identified within the first intron of *RET* that increases the risk of Hirschsprung disease ~2.1-5.7 times compared to the major allele (Emison et al. 2005). It was found in this case that the rSNP disrupted a single SOX10 binding site (Emison et al. 2010).

SOX10 is a member of the SOX protein family and contains four major domains: dimerization domain, HMG domain, K2 domain, and the transactivation domain (Figure 1.4). It is a member of the SOXE subfamily (Wright et al. 1993) and has been shown to be necessary for proper neural crest development (Lane and Liu 1984; Herbarth et al. 1998; Southard-Smith et al. 1998). While SOX10 is downregulated in many neural crest derived tissues, its expression remains consistent in both melanocytes and Schwann cells. Indeed, SOX10 is necessary for all stages of Schwann cell development, and conditional loss of SOX10 at any stage results in demyelination (Paratore et al. 2001; Finzsch et al. 2010; Bremer et al. 2011). Additionally, SOX10 regulates many genes critical for Schwann cell function such as *MPZ* (Peirano et al. 2000), *PMP22* (Jones et al. 2011b), and *GJB1* (Bondurand et al. 2001). Not surprisingly, mutations within SOX10 target genes can result in demyelinating peripheral neuropathies such as Charcot-Marie-Tooth (CMT) disease (Dyck and Lambert 1968; Lupski et al. 1991; Raeymaekers et al. 1991; Hayasaka et al. 1993; Kulkens et al. 1993; Ionasescu and Searby 1994).

CMT disease is characterized by distal muscle wasting and sensory loss (Dyck and Lambert 1968; Szigeti and Lupski 2009). One characteristic of CMT is the large amount of clinical variability that is observed within patients, such as age of onset from the first to seventh decade and severity of sensory loss, often despite identical coding mutations (Thomas et al. 1997; Pareyson et al. 2006; Pareyson and Marchesi 2009). Clinical variability of CMT was even demonstrated within two sets of identical twins with CMT (Garcia et al. 1995). The cause of the clinical variability is unknown, however one potential explanation is rSNPs altering CREs, leading to disruption of endogenous gene expression and ultimately modifying disease severity.

For example, patients with CMT1A generally harbor a duplication of *PMP22,* although duplication of upstream CREs can cause a mild form of CMT1A (Weterman et al. 2010). If a rSNP disrupted CREs that regulate *PMP22* and caused a decrease in expression, it may result in less severe symptoms relative to a patient who does not have this SNP allele.

Despite recent technological advances, major challenges remain in both the identification and functional assessment of CREs and rSNPs (Chapter 1). In Chapter 2, we developed a novel pipeline combining computational predictions with functional assays to discover novel CREs harboring a SNP. In this chapter, we modify our pipeline to include SOX10 consensus sequence information and ultimately identify four active SOX10 response elements that harbor a SNP, two of which demonstrate significant allele-specific differences in activity. We deeply characterized one of the two SOX10 elements by deleting it from our model Schwann cell line (S16) using CRISPR, which helped develop a hypothesis regarding the target gene of this element. Specifically, we performed RNA-Seq experiments on the knockout cells compared to unmodified S16 cells and identified one candidate gene, *Tubb2b*. Both *Tubb2b* and the rSNP identified may give novel insights into Schwann cell function and development, and both represent potential modifiers of disease or potentially causative mutations in patients with neurocristopathies.

All of the work presented in this chapter was performed by the author except the DNase-Seq experiments which were performed by Dr. Lingyun Song and Dr. Gregory E. Crawford (Duke University) and the analysis and peak calling was performed by Weisheng Wu (University of Michigan Bioinformatics Core). The human, mouse, and chicken genome alignments were performed by Dr. Tony Antonellis and Dr. Arjun Prasad (Antonellis et al. 2006).

Methods

Computational Identification of SOX10 Consensus Sequences

The human reference genome (hg18) was downloaded from the UCSC genome browser (Kent et al. 2002), and a custom Perl script was written to identify all SOX10 consensus sequences within the human genome. SOX10 binds to the core consensus sequences (5'-3'): ACAAA, ACACA, ACAAT, or ACAAG (Peirano and Wegner 2000; Srinivasan et al. 2012). The reverse complement of each sequence was also identified (5'-3'): TTTGT, TGTGT, ATTGT, or CTTGT. This identified ~33.5 million SOX10 consensus sequences, and this dataset was overlapped with the regions conserved among human, mouse, and chicken (MCS; Chapter 1). SNPs residing within a conserved SOX10 consensus sequence were identified, and regions overlapping RefSeq exons were excluded as previously described in Chapter 2 - Methods.

PCR, Cloning, and Mutagenesis

An identical procedure was used to PCR amplify regions surrounding the conserved SOX10 consensus sequences harboring SNPs as described in Chapter 2. The primers (Appendix 2) all contained gateway adapter sequences (Life Sciences) to clone into pDONR221. The regions were sequence verified to ensure the presence of the major allele and were then cloned into pE1B forward and reverse as described in Chapter 2 - Methods. Regions demonstrating activity were mutagenized to the minor allele or to delete the SOX10 consensus site using an identical procedure as described in Chapter 2 - Methods.

Overexpression Studies

The S16 and MN-1 cells were cultured at 37°C in 5% CO_2 in general media (GM): Dulbecco's Modified Eagle Medium (DMEM; Invitrogen cat no ILT12430054) and supplemented with 10% fetal bovine serum (FBS; ThermoFisher cat no 26140-079), 2 mM L-glutamine (Corning cat no COR25005CIS), and 1X Penicillin-Streptomycin (ThermoFisher cat no 15070-063). The S16 and MN-1 cells were plated at 10,000 cells per well of a 96-well plate. The cells were transfected the following day using Lipofectamine 2000 (ThermoFisher Cat no 11668-019) (described in detail in Chapter 2).

For overexpression studies, an additional 100 ng (per well) of overexpression plasmid was added in addition to the 200 ng of the SOX10 region in pE1B and 2 ng of CMV-renilla. The dominant-negative SOX10 plasmid harbors a truncated *SOX10* cDNA that contains a premature stop codon (E189X) driven by a CMV promoter (Inoue et al. 2004). This mutation creates a truncated SOX10 protein which has the ability to dimerize with endogenous SOX10, but lacks the transactivation domain resulting in the mutant-wildtype dimer to be nonfunctional. The wildtype SOX10 overexpression plasmid contains the *SOX10* gene with nearly full length 5' and 3' UTRs driven by a CMV promoter (Inoue et al. 2004). Both the plate reading and statistical analysis were identical to Chapter 2.

DNase Hypersensitivity Site Identification

DNase-Seq was performed with three biological replicates of the S16 cells at passage numbers five, eight, and 14. Each replicate contained ~20 million cells frozen into 1 mL of recovery cell culture freezing media (Invitrogen Cat no. 12648010). Cells were thawed and DNase-Seq libraries were generated as previously described (Song and Crawford 2010) with the exception of adding a 5’ phosphate to linker 1 to increase the ligation efficiency. DNase-Seq libraries from three replicates were pooled into one lane of an Illumina Hi-Seq 2000. Raw reads were aligned to the rat rn5 genome using Bowtie (Langmead et al. 2009) and mapping allowing up to two mismatches. For the three samples, 69.2% (36,295,401), 70.8% (43,564,606), and 67.9% (39,579,719) of the reads mapped to rn5. Peaks were called using F-Seq and the default settings (Boyle et al. 2008). For the three samples: 502,787 (sample 1), 438,254 (sample 2), and 412,267 (sample 3) peaks were identified. 149,342 peaks were shared among all three samples. We used sample 2 as a representative experiment and compared all genomic regions to sample 2 peaks.

Generation of Homologous Repair Templates (Gibson Assembly)

The drug resistant repair templates for homologous recombination were generated using Gibson assembly (Gibson et al. 2009). All homologous repair templates were cloned into the *Bam*HI restriction site of the multiple cloning site of pUC19. Primers were designed to amplify an approximately one kilobase pairs 5’ (rn5; chr17:42265085-42266075) and 3’ (rn5; chr17:42266727-42267691) arms of homology surrounding rSOX-4 (rn5; chr17:42266012-42266902) from S16 (rat) genomic DNA using Primer3 (http://bioinfo.ut.ee/primer3/) with the default parameters except the primer tm Min = 55°C, Opt = 58°C, and Max = 60°C (Appendix 3). Once the primers were designed, 30 base pair adapter sequences were added to the primers that were homologous to either the linearized pUC19 backbone or the drug resistance cassette. For example on the 5’ arm of homology, the forward primer contained a 30 base pair adapter sequence that was homologous to the 30 base pairs immediately upstream (5’) of the *Bam*HI cut site, while the reverse primer contained a 30 base pair adapter sequence homologous to the first

30 base pairs of either the blasticidin or neomycin resistance cassette and a LoxP site (Appendix 3). For the 3' arm of homology, the forward primer contained a 30 base pair adapter sequence homologous to the last 30 base pairs of either the blasticidin or neomycin resistance cassette and a LoxP site, while the reverse primer contained a 30 base pair adapter sequence that was homologous to the 30 base pairs immediately downstream (3') of the *Bam*HI cut site. Primers were also designed to amplify either the blasticidin or neomycin resistance cassettes, however no additional adapter sequences were added. The blasticidin resistance cassette was PCR amplified from pCMV/Bsd (ThermoFisher - Cat no. V510-20). The neomycin resistance cassette was PCR amplified from the hCas9 expression plasmid backbone. The hCas9 expression plasmid was a gift from George Church (Addgene plasmid #41815) (Mali et al. 2013).

Prior to the Gibson assembly reaction, pUC19 was linearized using *Bam*HI: 15 µL pUC19 (~1.5 µg), 1 µL *Bam*HI (NEB Cat no. R0136S) (20 units), 3 µL NEBuffer 3, 3 µL BSA, and 8 µL ultrapure water. The *Bam*HI reaction was incubated at 37°C for 90 minutes and assessed on a 1% agarose in 1X TBE (Fisher cat no 50-751-7033) gel. A single Gibson assembly reaction contains the following: 10 µL 2X Gibson Assembly Master Mix, 1 µL (50 ng) of *Bam*HI linearized pUC19, 1 µL (100 ng) drug resistant PCR template, 1 µL (100 ng) 5' arm of homology, 1 µL (100 ng) 3' arm of homology, 6 µL ultrapure water. For some reactions, the volume of DNA was varied (the mass of each DNA fragment was maintained), and the water volume was modified to ensure a total volume of 20 µL. Reactions were incubated at 50°C for 15 minutes and transformed into NEB 5-alpha Competent *E Coli* (NEB Cat no. C2987I) as per manufacturer's protocol: 2 µL of Gibson reaction was added to one tube of NEB 5-alpha Competent *E Coli*, incubated on ice for 30 minutes, heat shocked at 42°C for 45 seconds, 950 µL of SOC media was added, and the tubes were placed at 37°C shaking at 900 rpm for one hour. After one hour, all 1 mL of bacteria were plated on 100 mg/mL ampicillin selection plates and placed in a 37°C incubator overnight. The following day, colonies were picked and placed into 6 mL of 100 mg/mL ampicillin selection media and placed at 37°C shaking overnight at 225 RPM. Plasmid DNA was isolated using the Qiagen miniprep kit as per the manufacturer's protocol (Qiagen cat no 27106).

The drug resistance template in pUC19 was subjected to a diagnostic *Eco*RI restriction enzyme digest: 1 μL drug resistance template in pUC19 (~300 ng), 1 μL *Eco*RI (NEB Cat no. R0101S) (10 units), 1 μL NEBuffer 2, and 7 μL ultrapure water. The *Eco*RI reaction was incubated at 37°C for one hour and assessed on a 1% agarose in 1X TBE (Fisher cat no 50-751-7033) gel. Plasmids that digested appropriately were sent for sequencing using custom sequencing primers designed using Primer3 (http://bioinfo.ut.ee/primer3/) with the default parameters except the primer tm Min = 55°C, Opt = 58°C, and Max = 60°C that tiled the entire insert (spaced ~600 base pairs apart). Properly recombined plasmids were used in subsequent steps as homologous repair templates.

S16 Cell Death Curves

To determine the effective drug concentrations of blasticidin and G418 on S16 cells, death curves were performed. S16 cells were plated in 6-well dishes at 50,000 and 100,000 cells per well. Both blasticidin (Thermo Cat no. A11139-03) and G418 (Invitrogen Cat no. ILT10131035) were tested individually by adding the drug at the indicated concentration into standard growth media. Five drug concentrations of blasticidin and one negative control were tested: 0, 2, 4, 6, 8, and 10 μg/mL. Blasticidin effectively prevented all cell growth within five days after addition of selection media at all concentrations, except 0 and 2 µg/mL. Five concentrations of G418 and one negative control were tested: 0, 500, 600, 700, 800, 1000 µg/mL. G418 effectively prevented all cell growth within 14 days after addition of selection media at concentrations of 700 µg/mL and above. Selection media was made that contained standard growth media plus either 4 µg/mL blasticidin (blast media) or 4 µg/mL blast and 700 µg/mL G418 (blast-neo media).

CRISPR in S16 Cells

Guide RNAs (gRNAs) were designed against rSOX-4 by scrutinizing the Sox10 consensus site and surrounding sequences for the protospacer adjacent motif (PAM; 5'-NGG-3') (Mali et al. 2013). Two gRNAs (Figure 3.9) were identified with cut sites within 30 base pairs of the Sox10 consensus site and flanked by the PAM. The 20 base pairs immediately 5' to the PAM were ordered as overlapping primers with adapter sequences to perform Gibson assembly (Gibson et al. 2009) into the gRNA cloning vector (a gift from George Church; Addgene plasmid #41824). An additional 5'-G was added between the adapter sequence and the gRNA. If the gRNA already contained a 5'-G, no additional modification was added. The overlapping gRNAs were diluted to 200 µM and were PCR amplified to generate a double stranded DNA fragment with the gRNA in the middle, flanked by 40 base pair homologous sequences for Gibson assembly. One PCR reaction contained 5 µL gRNA-1 forward primer, 5 µL gRNA-1 reverse primer, and 40 µL of PCR supermix (Invitrogen Cat no. 10572-014).

The gRNA cloning vector was digested using *Afl*II: 5 µL (~500 ng) gRNA cloning vector, 2 µL CutSmart buffer, 2 µL *Afl*II (NEB Cat no. R0520S; 40 units), and 11 µL ultrapure water. The reaction was incubated at 37°C for one hour, and the gRNAs were cloned into the *Afl*II cut gRNA expression plasmid using Gibson assembly: 2.5 µL (~50 ng) *Afl*II cut gRNA expression plasmid, 1.5 µL (~375 ng) double stranded gRNA insert, 10 µL 2X Gibson Assembly Master Mix, and 6 µL ultrapure water. The reaction was incubated at 50°C for 15 minutes before the reaction was transformed and plasmids isolated as described above (*Gibson Assembly*). Plasmids were assessed by *Eco*RI digestion and sent for sequencing with custom sequencing primers.

S16 cells were plated at 100,000 cells per well in a 6-well dish 24 hours prior to transfection. The drug repair templates were linearized by *Hind*III digestion: 15 µL (~5 µg) of drug repair template, 3 µL NEBuffer 3, 1 µL *Hind*III (NEB Cat no. R0104S); 20 units), and 11 µL ultrapure water. The reaction was incubated at 37 °C for two hours and purified using the Qiagen PCR purification kit (Qiagen Cat no. 28104). The S16 cells were transfected with Lipofectamine 2000 (ThermoFisher Cat no. 11668-019) and 3 µg of total DNA per well: 1 µg of hCas9 (a gift from George Church; Addgene plasmid #41815) (Mali et al. 2013), 1 µg of rSOX-4 targeting gRNA expression plasmid, and 1 µg of linearized drug repair template (~1:1:1 molar ratio). For an individual well, 5 µL of Lipofectamine 2000 was mixed with 500 µL of Optimem (ThermoFisher cat no 31985-062) and incubated for 10 minutes at room temperature (cocktail 1). The 3 µg of total DNA was mixed with 500 µL of Optimem and combined with the first cocktail after the 10 minute incubation (transfection mixture). The transfection mixture was vortexed for 3 seconds and incubated for 20 minutes at room temperature. During the final incubation, the S16 cells were washed with 2mL of 1X PBS (ThermoFisher Cat no. 10010-023), and 1 mL of the transfection mixture was added. The cells were incubated for 4 hours at 37°C in 5% CO_2. After 4 hours, the transfection mixture was removed and replaced with 3 mL of standard growth media. The cells were incubated for 72 hours at 37°C in 5% CO_2, and then the cells were passaged 1:3 into a new 6-well dish with standard growth media and incubated overnight at 37°C in 5% CO_2.

The following morning, the standard growth media was removed and drug selection media (4 μg/mL blasticidin (blast media) or 4 μg/mL blast and 700 μg/mL G418 (blast-neo media)) was added. Selection times varied from 3-5 days for blasticidin and 10-14 days with G418, however an untransfected positive control well was used to assess the drug effectiveness. The selection media was replaced every three days to maintain appropriate drug concentrations. Once the untransfected control cells died, the CRISPR transfected cells were expanded into a T-75 flask and maintained in drug selection media.

Flow Cytometry and Cell Expansion

A confluent T-75 flask of CRISPR-modified S16 cells was prepared for flow cytometry. The cells were harvested into a 15 mL conical tube, spun at 2,000 RPMs for two minutes, the media was removed, and the cell pellet was resuspended in 2 mL of 1X PBS. Then the cells were passed through a 40 μm cell strainer (Falcon Cat no. 352340) by pipetting the solution through the filter into a new 15 mL conical tube. Five 96-well tissue culture dishes (Fisher Cat no. 07-200-565) were filled with 75 μL of drug selection media per well. The cells were flow sorted by the University of Michigan flow cytometry core using front and side scatter (or for GFP positive cells when transfected with Cre:GFP) to one cell per well of the 96-well plate.

The cells were incubated at 37°C in 5% CO_2 for approximately 14 days. The drug selection media was changed every 4-5 days to maintain drug concentrations. The 96-well plates were assessed after 14 days for the presence of cell colonies in individual wells. These wells were expanded into 48-well plates (BioExpress Cat no. 677180) and incubated at 37°C in 5% CO_2 for 3-5 days until the cells were confluent. Each well was passaged from a 48-well dish into a 6-well dish incubated at 37°C in 5% CO_2 for 3-5 days until the cells were confluent. Finally, the cell clones were expanded into a T-75 flask and partially harvested for genomic DNA isolation using the Wizard Genomic DNA Purification Kit (Promega Cat no. A1120) according to the manufacturer's protocol. The cells were assessed by diagnostic PCR to assess for proper recombination of the drug repair template into the rSOX-4 locus (Appendix 4). Intermediate expansion vessels were used as necessary depending on cell growth, such as 12-well plates or T-25 flasks. Additional growth time or media changes were also performed depending on cell growth, however throughout clonal expansion, drug selection media was always used.

Cre:GFP Lipofectamine 2000 Transfection into T-75 Flask

The floxed drug resistance cassettes were removed from the CRISPR modified S16 cells by transient transfection with a Cre:GFP plasmid (a gift from Connie Cepko; Addgene plasmid # 13776). Due to the low transfection efficiency, an approximately 70% confluent T-75 flask was transfected with Cre:GFP. An identical method was used as previously described (Chapter 2), with increased amounts of reagents: first mixture contained 60 μL Lipofectamine 2000 in 1.5 mL OptiMem, second mixture contained 24 μg of Cre:GFP in 1.5 mL OptiMem, and the transfection mixture contained a total of 3 mL of mixture one (1.5 mL) and mixture two (1.5 mL). After the 20 minute room temperature incubation of the transfection mixture, the cells were washed with 10 mL of 1X PBS, 3 mL of the transfection mixture was added, the cells were incubated for four hours at 37°C in 5% CO_2, and the transfection mixture was removed and 10 mL of standard media was added (no PBS wash between transfection mixture and standard media addition). The cells were incubated for 72 hours at 37°C in 5% CO_2. After 72 hours, the cells were processed for flow cytometry and GFP positive cells were sorted to individual cells within wells of a 96-

well plate that were gradually expanded into a T-75 flask for subsequent analysis (see above). Standard growth media was used for all expansion stages after Cre:GFP transfection.

TRIzol RNA Isolation

Cells were harvested from a confluent T-75 flask, placed in a 15 mL conical vial, spun at 2,000 RPM for 2 minutes, and resuspended in 5 mL of TRIzol (ThermoFisher Cat no. 15596-018). 1 mL of the TRIzol solution was used for RNA isolation, while the rest was stored at -80°C. 200 µL of chloroform was added to the 1 mL of cells in TRIzol and vortexed for one minute. The solution was incubated at room temperature for three minutes and then spun at 12,000g for 15 minutes at 4°C. The top, aqueous phase was removed (~1 mL) and transferred to a clean tube. 500 µL of 100% isopropanol was added to the aqueous phase, mixed by inversion 2-3 times, and incubated at room temperature for 20 minutes. The tube was then spun at 12,000g for 10 minutes at 4°C to pellet the RNA. The supernatant was removed and discarded, and 1 mL of 75% ethanol was added. The tube was mixed by inversion 2-3 times, spun at 7,500g for five minutes at 4°C, and the supernatant was removed and discarded. The ethanol wash was performed a second time, before the RNA pellet was air dried for 5-10 minutes (or until no ethanol remained in the tube). The pellet was resuspended in 50 µL of ultrapure water and used in subsequent analyses.

RNA-Sequencing Analysis

RNA was isolated from the three rSOX-4 mutants and two unmodified S16 cells (one parental cell line of the rSOX-4 mutants [passage 9], and one older passage [passage 39]) using TRIzol. PolyA selected mRNA libraries were generated using the TruSeq kit (illumina Cat no. RS-122-

2001) and subjected to 50 base pair single end sequencing on the HiSeq2000 platform. Two technical replicates of the five cells were pooled and run across two sequencing lanes which resulted in ~21.5 million reads per cell line. The quality of the reads was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were aligned against the rat Rnor_5.0 (Ensembl) reference genome downloaded from the illumina iGenomes (http://support.illumina.com/sequencing/sequencing_software/igenome.html) using STAR (Dobin et al. 2013). Default parameters were used, except only uniquely mapped reads were allowed (~82% of total reads mapped uniquely). HTSeq (Anders et al. 2015) was used to count the number of reads per gene using default parameters except the stranded reads function was disabled. Finally differential gene expression between the rSOX-4 mutants and unmodified S16 cells was determined using DESeq2 (Love et al. 2014). All programs were run on the ARCTS flux servers at the University of Michigan. An identical analysis was performed with the five unmodified S16 clones derived from the original parental cell line of the rSOX-4 mutants.

Digital Droplet PCR

cDNA was synthesized from RNA extracted from the rSOX-4 mutant and wildtype S16 cells using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems Cat no. 4368814). A single reaction contained 3.2 μL ultrapure water, 2 μL 10X random primers, 2 μL 10X RT buffer, 1 μL RNase inhibitor, 0.8 μL dNTPs, 1 μg RNA in 10 μL of water, and 1 μL MultiScribe reverse transcriptase. The reaction was incubated for 10 minutes at 25°C, then 2 hours at 37°C, then 85°C for 5 seconds, and finally held at 4°C.

The resulting cDNA was diluted 1:333 in ultrapure water prior to the digital droplet PCR (ddPCR) reaction. A single ddPCR reaction contained 11.5 μL 2X ddPCR supermix (BioRad Cat no. 1863010), 1 µL FAM probe, 1 µL HEX probe, 6 µL cDNA template (diluted 1:333), and 3.5 µL water. FAM probes were designed against the gene of interest while HEX probes were against the control gene (*Gapdh*). The probes and primers were ordered from IDT PrimeTime predesigned assays for rat except for the FAM probe for *Aldh5a1* that was custom designed using IDT PrimerQuest tool with default parameters for two primers one probe (Appendix 5). All probes and primer sets were diluted to a stock 20X concentration in ultrapure water (500 µL for standard size and 100 µL for mini size). Droplets were generated using a droplet generation cartridge (BioRad Cat no. 1864008). Only 20 µL of the reaction mixture was used in droplet generation to avoid any air bubbles, which was combined with 70 µL of oil for probes (BioRad Cat no. 1863005). A gasket (BioRad Cat no. 1863009) was placed over the droplet generation cartridge and placed into the droplet generator (BioRad Cat no. 1864002). 40 µL of the resulting droplets were carefully transferred to a 96-well semi-skirted PCR plate (Eppendorf Cat no. 951020346) and the plate was heat-sealed with foil (BioRad Cat no. 1814040). The following PCR program used for the ddPCR reactions: 95°C for 10 minutes, 95°C for 30 seconds, 60°C for one minute, repeat steps 2 and 3 40 times, 98°C for 10 minutes, and 4°C hold. The plate was then transferred to the droplet reader (BioRad Cat no. 1864003), and the results were analyzed using the QuantaSoft (BioRad Cat no. 1864011) software to determine the absolute quantification of each gene of interest. Significant differences between the rSOX-4 and unmodified S16 cells was performed using a Student's T-test. Each cDNA sample was assessed in four technical replicates, and the average was used in subsequent analysis. For rSOX-4, one cDNA sample was used for each mutant cell line, and all three were combined for the final analysis. For unmodified S16 cells, three independent biological replicates were assessed and combined for the final analysis.

Transient Transgenic LacZ Mice

rSOX-4 was PCR amplified with gateway adapter primers to clone the region into the reverse orientation of the HSP70:*LacZ* (Pennacchio et al. 2006) gateway compatible vector (*i.e.* the gateway adapters were put onto the opposite primer). The plasmid was sequence verified with custom sequencing primers and maxi-prepped (Zymo Cat no. D4203) according to the manufacturer's protocol. The rSOX-4:*LacZ* transgene was liberated from the plasmid backbone using a *Sal*I digestion: 40 µL (~50 µg) rSOX-4:*LacZ* plasmid, 5 µL NEB CutSmart buffer, and 5 µL *Sal*I-HF (NEB Cat no. R3138T); 500 units). The reaction was incubated at 37°C overnight, and a small aliquot of the cut product was assessed on a 1% agarose gel. The digested plasmid was submitted to the University of Michigan transgenic animal core, the transgene was gel purified, and the transgene was injected into mouse zygotes. The zygotes were injected into pseudopregnant mice, and embryos were harvested 11 days later (~E11.5).

The embryos were dissected out and placed into 15 mL of ice cold fixative solution: 0.2% glutaraldehyde and 2% formaldehyde in PBS. The embryos were fixed for 20 minutes on ice and then washed three times for 10 minutes with wash buffer: 2mM $MgCl_2$, 5mM EGTA, and 0.02% NP40 in PBS. After washing, the embryos were transferred to a 50 mL conical vial containing ~25 mL of LacZ staining solution: 5 mM potassium ferrocyanide, 5 mM potassium ferricyanide, and 1 mg/mL X-gal (Invitrogen Cat no. 15520-034) in wash buffer. The X-gal was diluted to 50 mg/mL in dimethylformamide prior to addition to the staining solution. The embryos were gently rocked at 37°C overnight in the dark. The next morning, the embryos were washed three times for 10 minutes each with wash buffer and imaged using a dissecting scope with QImaging camera (QICAM FAST1394) and software (Qcapture Pro Version 6).

Results

Computational Identification of Conserved SOX10 Consensus Sites Containing SNPs

We modified the computational pipeline developed in Chapter 2 to include information about SOX10 consensus sites. A custom Perl script was written to identify all SOX10 consensus sequences in the human genome (hg18) using the core SOX10 consensus sites (5' to 3'): ACAAA, ACACA, ACAAT, or ACAAG. The reverse complement of each sequence was also assessed, and this analysis revealed over 33 million consensus sequences in the human genome (Figure 3.1). These regions were overlapped with the multi-species conserved sequences (*i.e.* five base pair long regions identical between human, mouse, and chicken) to identify fully conserved SOX10 consensus sites. The conserved regions were overlapped with the same SNPs validated 'by-frequency' used in Chapter 2, and finally regions with any overlap with RefSeq exons were removed (Pruitt et al. 2012) to exclude regions conserved due to gene function. This generated a dataset of 224 fully conserved non-coding SOX10 monomers with SNPs.

This dataset was prioritized for functional studies by using two additional filters. The first filter required regions to overlap with *in vivo* Sox10 ChIP-Seq peaks identified in nuclei from rat sciatic nerve, which mainly contains Schwann cell nuclei (Srinivasan et al. 2012). The 224 regions described above were converted from the human genome to the rat genome using the liftOver utility on the UCSC genome browser (Kent et al. 2002), and overlap between the two datasets was determined using the UCSC table browser (Karolchik et al. 2004). This analysis revealed nine conserved SOX10 consensus sites that overlap with ChIP-Seq peaks.

We also prioritized the 224 genome-wide SOX10 predictions by identifying dimeric SOX10 binding sites. A custom Perl script was written to identify dimeric SOX10 binding sites where

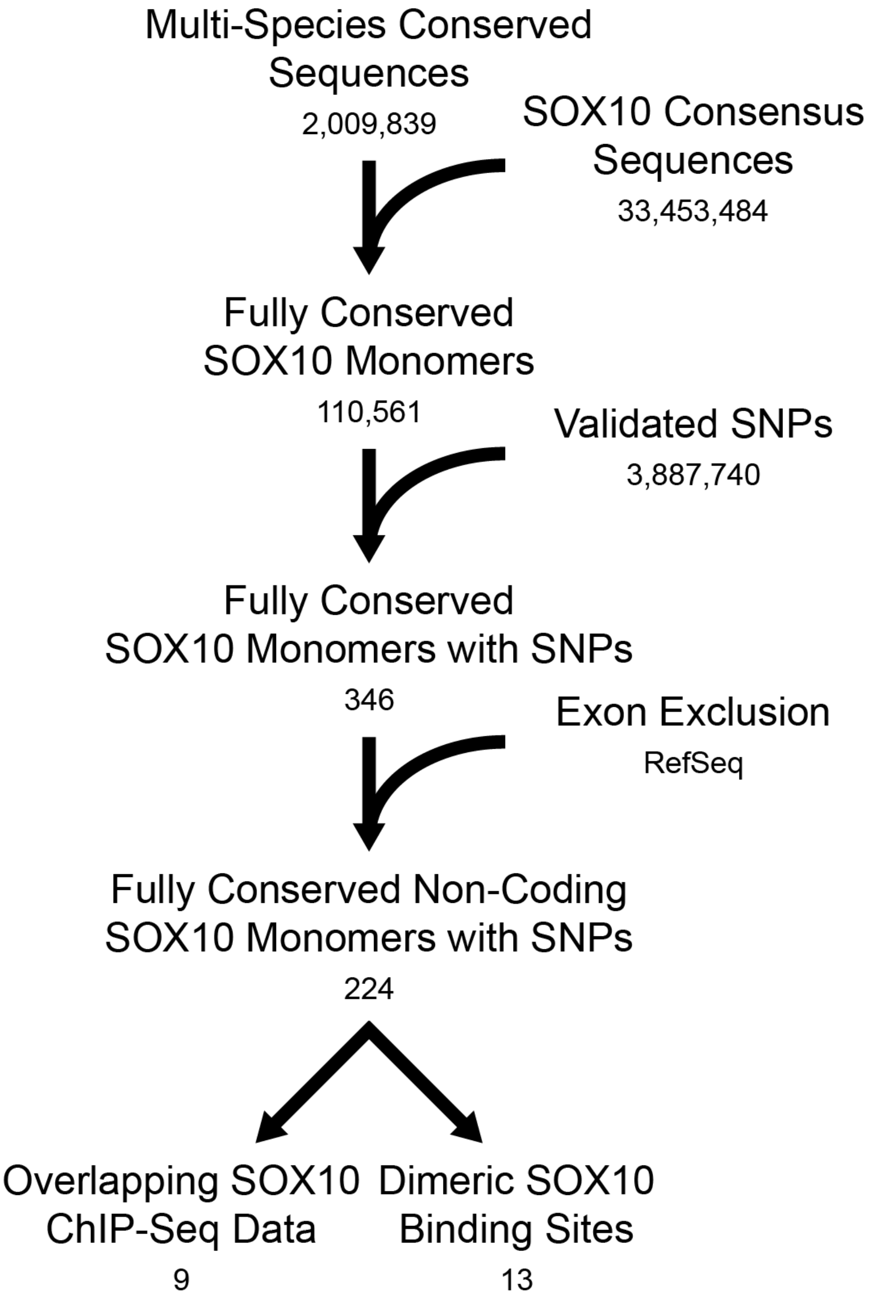

Multi-Species Conserved Sequences
2,009,839
SOX10 Consensus Sequences
33,453,484
Fully Conserved SOX10 Monomers
110,561
Validated SNPs
3,887,740
Fully Conserved SOX10 Monomers with SNPs
346
Exon Exclusion
RefSeq
Fully Conserved Non-Coding SOX10 Monomers with SNPs
224
Overlapping SOX10 ChIP-Seq Data
9
Dimeric SOX10 Binding Sites
13

Figure 3.1 *Computational Identification of Conserved SOX10 Consensus Sites.* The human genome was analyzed for all SOX10 consensus sequences and overlapped with the multi-species conserved sequences (five base pairs in length or greater and identical across the human (hg18), mouse (mm9), and chicken (gal3) genomes). Overlap between the fully conserved SOX10 monomers and SNPs validated ‘by-frequency’ from dbSNP130 was determined, and exons were excluded using the RefSeq dataset to generate fully conserved non-coding SOX10 monomers harboring SNPs. This dataset was further prioritized by identifying regions overlapping Sox10 ChIP-Seq peaks or that formed a putative dimeric SOX10 binding site with the fully conserved SOX10 monomers (intermonomeric spacer varied 1-30 base pairs). Numbers below each dataset label represent the number of regions in the dataset.

one monomer was from the 224 conserved non-coding SOX10 monomers with a SNP, and the other monomeric site was a fully conserved SOX10 binding site (110,561; Figure 3.1). The intermonomeric spacer was allowed to vary between one and 30 base pairs to minimize false negatives. While both monomers were required to be conserved, the intermonomeric spacer was not necessarily conserved. This analysis revealed 13 dimeric SOX10 sites, and combined with the nine overlapping ChIP-Seq peaks, resulted in a prioritized dataset of 22 putative SOX10 binding sites for further analysis. The regions were called rSOX (regulatory SOX) and were numbered incrementally from the p- to q-arm of chromosome 1 to 22, X, and Y. The first nine regions correspond to the SOX10 consensus sites overlapping ChIP-Seq peaks, while rSOX-10-22 correspond to the dimeric SOX10 consensus sites.

Identification of Two SOX10 Enhancers Containing rSNPs

To evaluate the 22 prioritized SOX10 elements, a region surrounding the SOX10 consensus site was amplified based on general conservation using the PhastCons 17-way vertebrate alignment (Siepel et al. 2005). The regions were amplified from human genomic DNA, cloned into pDONR221, and sequence verified to ensure both the presence of the major allele and integrity of the sequence. Properly integrated regions were then cloned into the luciferase expression plasmid pE1B in both the forward and reverse orientations. All 22 regions were assessed in the S16 cell line, which is a rat immortalized cell line that is the best cellular model of a myelinating Schwann cell and expresses endogenous *Sox10* (Goda et al. 1991; Hai et al. 2002; Hodonsky et al. 2012). All regions were compared back to an empty vector that does not contain any sequences upstream of the minimal E1B promoter, and its activity has been set to ‘1’. To be considered for further analysis, regions were required to display five-fold or higher activity compared to the empty vector. Four of the 22 regions met this criteria in at least one orientation (Fwd = forward and Rev = Reverse): rSOX-1 (Fwd = 2.57, Rev = 8.29), rSOX-4 (Fwd = 4.83, Rev = 25.96), rSOX-6 (Fwd = 0.33, Rev = 11.01), and rSOX-22 (Fwd = 14.12, Rev = 4.44) (Figure 3.2 and Table 3.1). To assess the effect of the SNP, the four active regions were mutagenized to the minor allele and reassessed in the S16 cells. Of the four active regions, two regions, rSOX-4 and rSOX-22, demonstrated a statistically significant effect between the two SNP alleles on luciferase activity in both orientations (Figure 3.3).

These two regions were further assessed by deleting the SOX10 consensus site. For rSOX-4, the monomeric site was deleted; however, since rSOX-22 contains a dimeric SOX10 consensus site, both SOX10 monomers and the intermonomeric spacer were deleted. These regions were only assessed in the more active orientation, specifically rSOX-4 in the reverse orientation and rSOX-22 in the forward orientation. For both rSOX-4 and rSOX-22, deleting the SOX10 consensus site further reduced luciferase activity compared to the minor allele in Schwann cells (Figure 3.4). Additionally, all three alleles were assessed in the MN-1 cells, a mouse motor neuron cell line that does not express endogenous Sox10, to determine if these regions may be SOX10 responsive elements (Salazar-Grueso et al. 1991; Hodonsky et al. 2012). None of the alleles for either rSOX-4 or rSOX-22 demonstrated any luciferase activity in the MN-1 cell line, suggesting SOX10 may be important for the regulatory activities of both elements.

To evaluate the necessity of SOX10 protein for the activity of these regions, a plasmid expressing a dominant-negative version of SOX10 (Inoue et al. 2004) was co-transfected with either rSOX-4 or rSOX-22 in Schwann cells (Figure 3.5). This plasmid contains a premature stop codon that results in a truncated form of SOX10 that maintains the dimerization and HMG (DNA

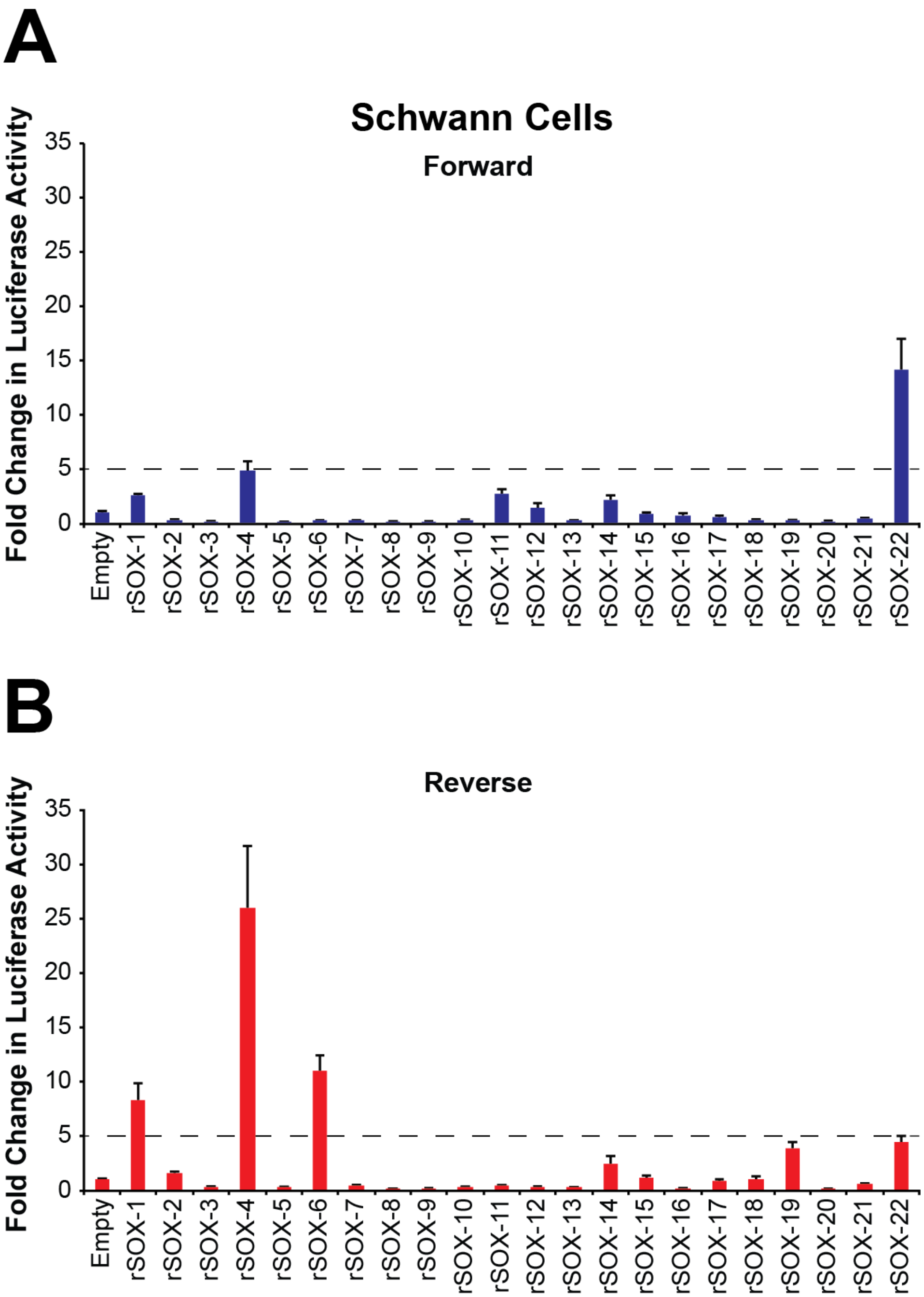


Figure 3.2 *Four Putative SOX10 Binding Sites Display Strong Activity in Schwann Cells.* The 22 conserved SOX10 consensus sites were cloned upstream of a luciferase reporter gene and tested in the forward (A) or reverse (B) orientation in S16 cells. The activity of each genomic segment is expressed relative to a control vector with no insert (‘Empty’), which was arbitrarily set to ‘1’. Four regions display greater than five-fold activity in at least one orientation. A dashed line is set to a five-fold increase in activity over the empty control to indicate ‘strong’ enhancer activity, and error bars show standard deviation.

Table 3.1 *Four Active SOX10 Regions in Schwann Cells.*

Region	**Forward**	**Reverse**	**Coordinates (hg18)**[1]	**rs Number**	**MAF**[2]
rSOX-1	2.57	8.29	chr2:44834620-44834625	rs3738980	0.3035
rSOX-4	4.83	25.96	chr6:22818474-22818479	rs16886790	0.2001
rSOX-6	0.33	11.01	chr6:98692222-98692227	rs17814604	0.1014
rSOX-22	14.12	4.44	chr16:53169712-53169728	rs1186802	0.4219

[1] Coordinates for conserved SOX10 monomeric site for rSOX-1, -4, and -6 and dimeric site for rSOX-22

[2] MAF is the minor allele frequency from the 1000 Genomes Project (Altshuler et al. 2012)

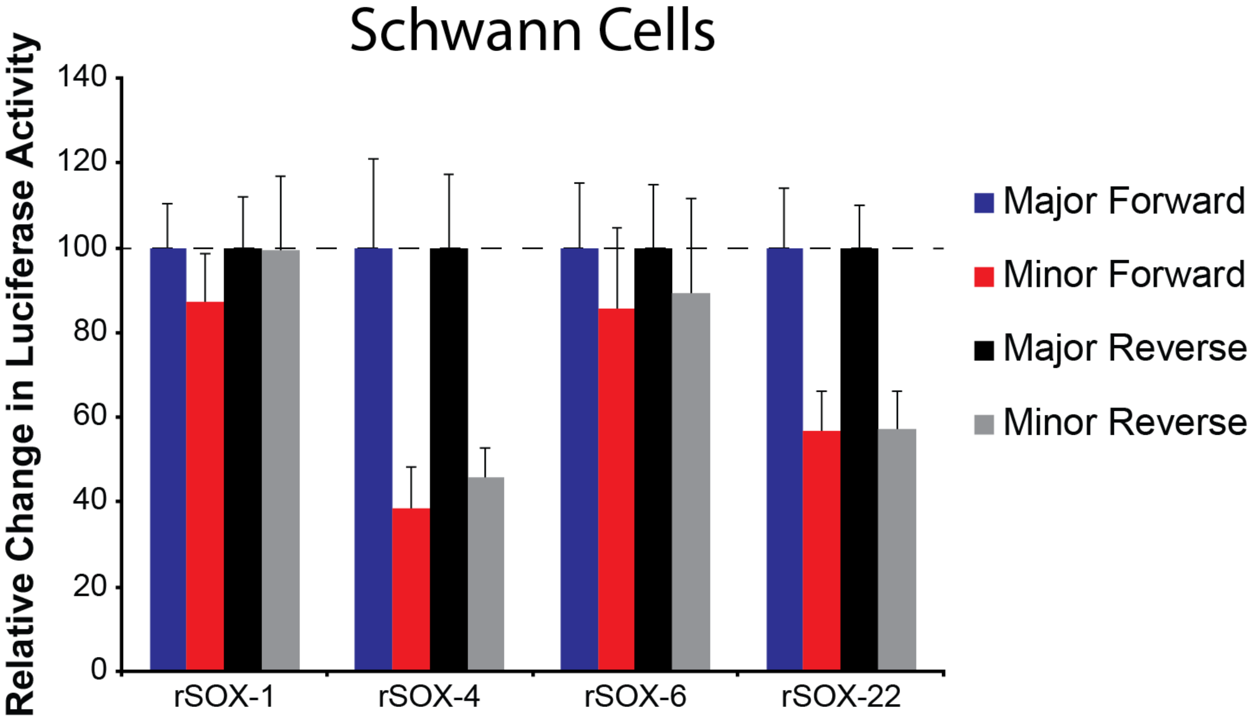


Figure 3.3 *Two SOX10 Regions Display a Significant Effect of the SNP on Luciferase Activity*. All four active regions were mutagenized to the minor allele and reassessed in both orientations. Two regions, rSOX-4 and rSOX-22, demonstrate a significant decrease in activity in the minor allele (red and grey bars) relative to the major allele (blue and black bars). The major alleles for both orientations have been set to '100', a dashed line indicates the normalized major allele activity, and the minor allele activity is relative to the major allele. The error bars represent standard deviation.

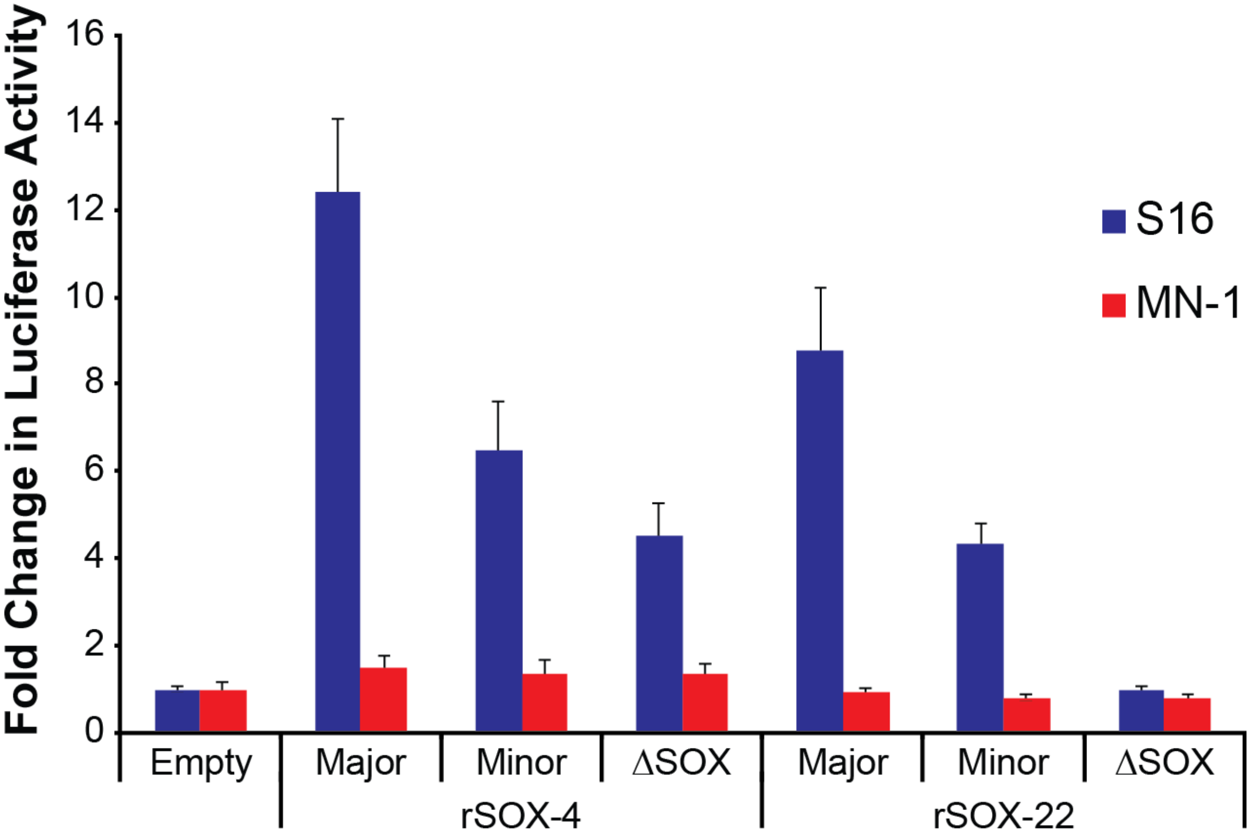


Figure 3.4 *SOX10 is Required for Luciferase Activity of Both rSOX-4 and rSOX-22*. The SOX10 consensus site was deleted from both rSOX-4 (monomeric) and rSOX-22 (monomers and intermonomeric spacer) and reassessed in Schwann cells (blue bars; S16). The activity is compared to an empty control plasmid arbitrarily set to'1', and only the more active orientation, reverse for rSOX-4 and forward for rSOX-22, is assessed. All three alleles for both regions were assessed in motor neurons (red bars; MN-1). The error bars for both S16 and MN-1 cells represent standard deviation, and ΔSOX represents the region with SOX10 consensus sites deleted.

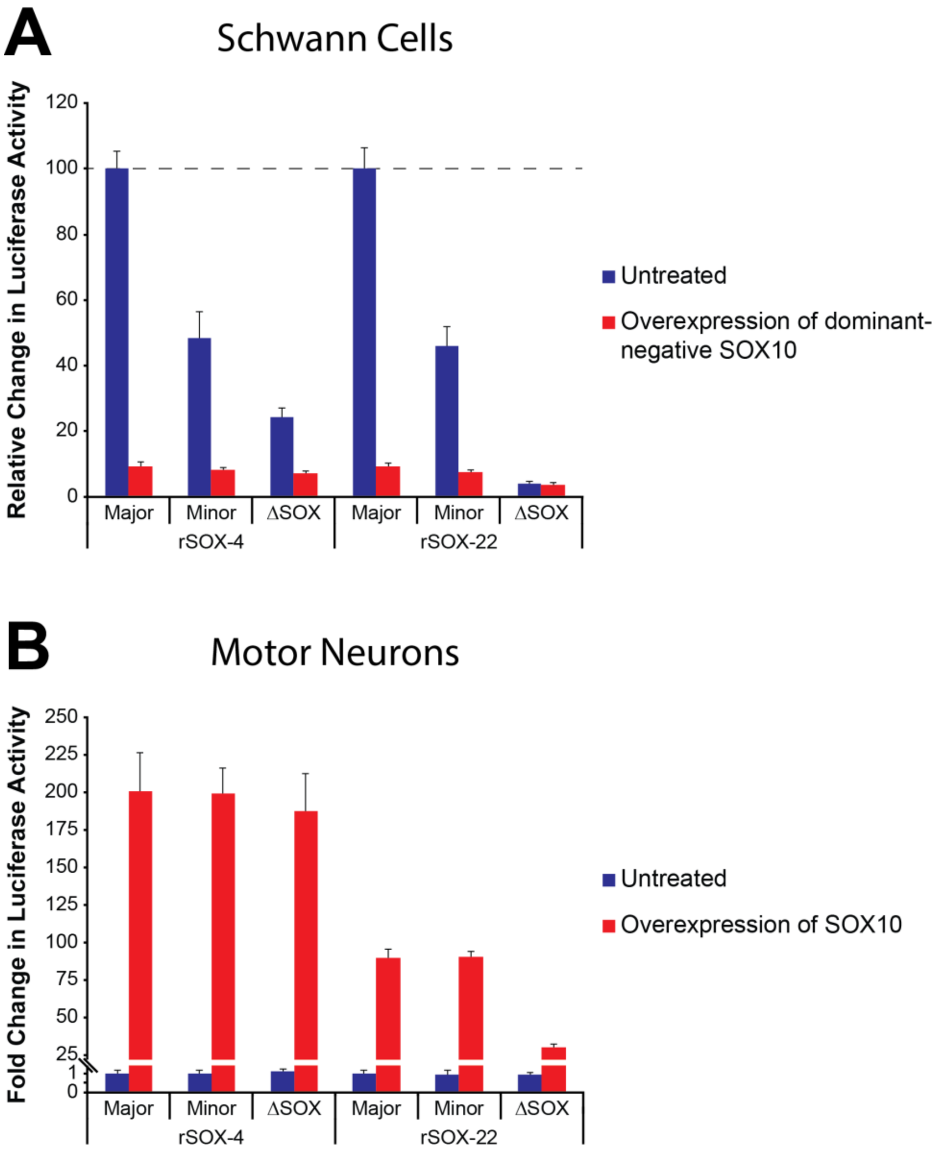


Figure 3.5 *SOX10 is Necessary and Sufficient for Luciferase Activity of rSOX-4 and rSOX-22.* (A) rSOX-4 and rSOX-22 were transfected into Schwann cells either independently (blue bars) or co-transfected with a dominant-negative SOX10 expression plasmid (red bars). The major untreated allele for both rSOX-4 and rSOX-22 has been set to ‘100’, and all other bars represent relative activity. (B) rSOX-4 and rSOX-22 were transfected into motor neurons either independently (blue bars) or co-transfected with a wildtype SOX10 expression plasmid (red bars). The major untreated allele for both rSOX-4 and rSOX-22 has been set to ‘1’, and all other bars represent relative activity. The double slashed line and broken red bars represent a break in the Y-axis. Error bars represent standard deviation, and ΔSOX represents the region with SOX10 consensus sites deleted.

binding) domains, but lacks the transactivation domain. When this mutant SOX10 interacts with the wildtype SOX10, the resulting dimeric complex is not functional, ultimately resulting in a depletion of wildtype SOX10 within the cell. For both regions, co-transfection of dominant-negative SOX10 reduced activity of all three alleles by ~90%. Conversely, when rSOX-4 or rSOX-22 was co-transfected with a wildtype SOX10 expression plasmid into the MN-1 cells, all three alleles displayed high levels of luciferase activity (Figure 3.5). Taken together, these data suggest that SOX10 is both necessary and sufficient for the regulatory activity of rSOX-4 and rSOX-22 in our *in vitro* cell models.

The Genomic Landscape of rSOX-4 and rSOX-22

Both rSOX-4 and rSOX-22 reside within gene deserts, with the nearest genes approximately one megabase pairs away in both the 5’ and 3’ directions with the exception of *Hdgfl1* which is ~150 kilobase pairs 5’ of rSOX-4 (Figure 3.6 and Figure 3.7). To better understand the genomic features of both regions and identify potential candidate target genes, DNase-Seq was performed on the S16 cell line. Three biological replicates of the S16 cell lines were used, and DNase-Seq was performed by Dr. Greg Crawford at Duke University. The data was analyzed at the bioinformatics core at the University of Michigan using F-Seq (Boyle et al. 2008). We identified approximately 450,000 peaks on average in each of the three cell lines. Using the sample 2 cell line as a representative sample, we identified all peaks with an F-Seq score greater than 0.08 (31,845 peaks, 7.3% of all peaks). Using these high confidence peaks, we observed rSOX-4 resided within a DNase hypersensitivity site (DHS) (Figure 3.6), while rSOX-22 does not reside within a DHS (Figure 3.7).

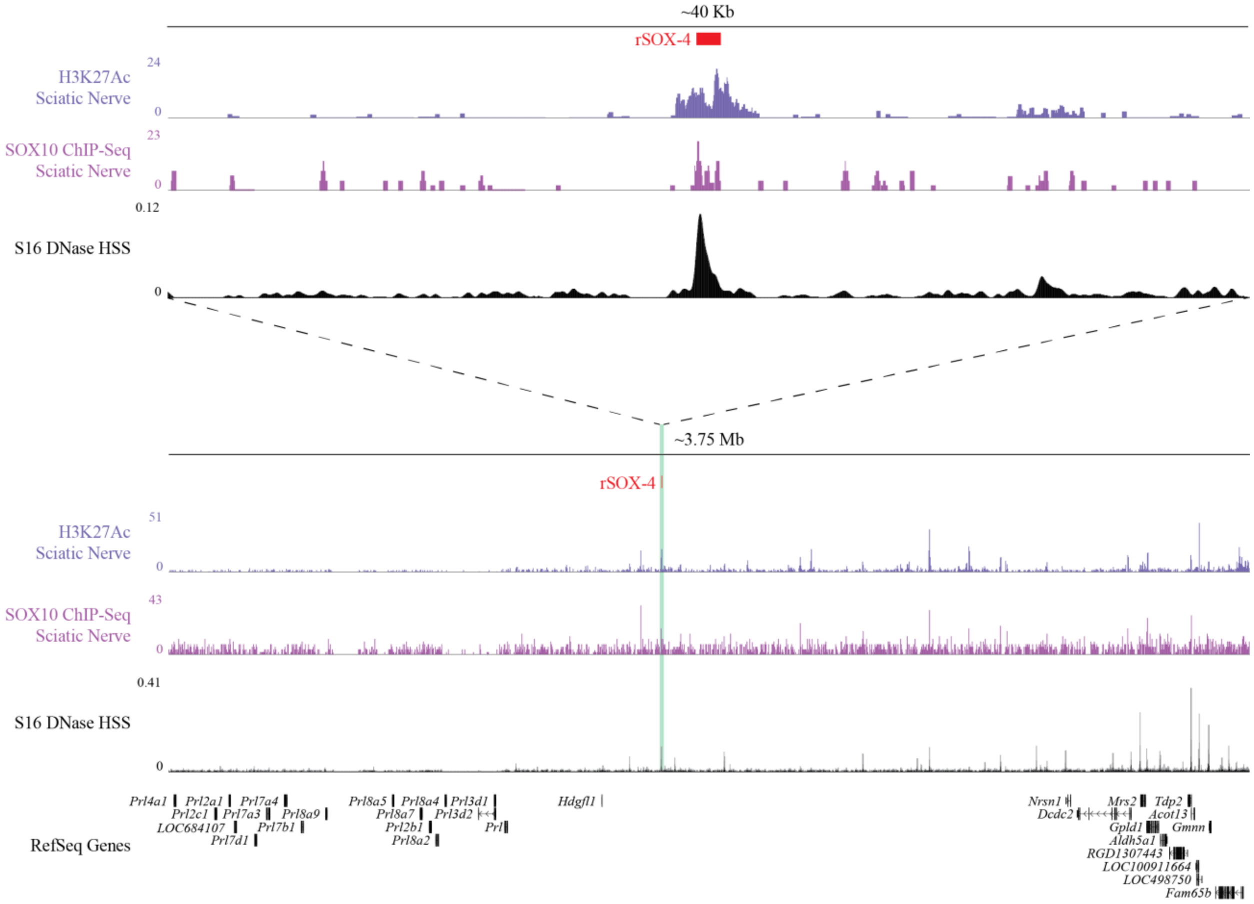


Figure 3.6 *rSOX-4 Overlaps Genomic Features Associated with Enhancers.* (Top) rSOX-4 (red box) overlaps histone 3 lysine 27 acetylation peaks (H3K27Ac; purple track), Sox10 ChIP-Seq peaks from rat sciatic nerve (pink track), and S16 DNase hypersensitivity peaks (S16 DNase HSS; black track). (Bottom) Zoomed out browser from above to show surrounding RefSeq genes. Green bar demonstrates rSOX-4 resides within peaks from all genomic features assessed. For both top and bottom, track names are at the left, the scale for each track is indicated, and the width of each browser window is indicated at the top (Kb = kilobase pairs and Mb = megabase pairs).

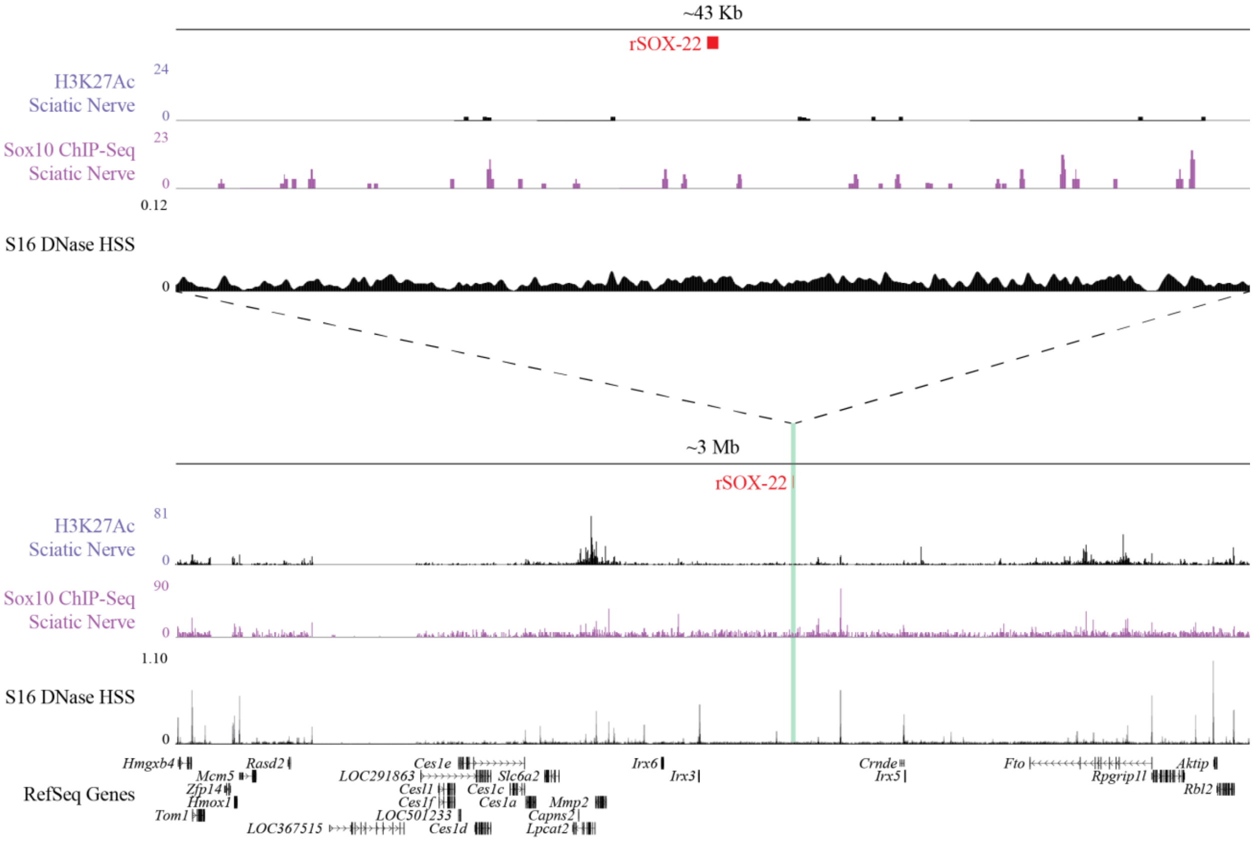


Figure 3.7 *rSOX-22 Does Not Overlaps Genomic Features Associated with Enhancers.* (Top) rSOX-22 (red box) does not overlap histone 3 lysine 27 acetylation peaks (H3K27Ac; purple track), Sox10 ChIP-Seq peaks from rat sciatic nerve (pink track), or S16 DNase hypersensitivity peaks (S16 DNase HSS; black track). The Y-axis scale for all tracks was matched to Figure 3.6 for comparisons. (Bottom) Zoomed out browser from above to show surrounding RefSeq genes. Green bar demonstrates rSOX-22 location across all three tracks. For both top and bottom, track names are at the left, the scale for each track is indicated, and the width of each browser window is indicated at the top (Kb = kilobase pairs and Mb = megabase pairs).

Additionally, we scrutinized each region for the presence of histone 3 lysine 27 acetylation (H3K27Ac) , which has been shown to correlate with active enhancers (Creyghton et al. 2010). We utilized previously published H3K27Ac data (Hung et al. 2015) generated from rat sciatic nerve to determine if our regions correlated with H3K27Ac peaks. Similar to the DHS data, rSOX-4, but not r-SOX-22, resides within a H3K27Ac peak from sciatic nerve (Anido et al. 2015). Finally, we examined each region for the presence of a SOX10 ChIP-Seq peak from sciatic nerve (Srinivasan et al. 2012). rSOX-4 was prioritized because it overlapped a SOX10 ChIP-Seq peak, while rSOX-22 had no evidence for SOX10 occupancy. Taken together, rSOX-4 is associated with many enhancer marks and SOX10 ChIP-seq data (Figure 3.6), while rSOX-22 is not (Figure 3.7). Based on the overlapping genomic features supporting rSOX-4 as an enhancer, we further pursued this region to determine the candidate target gene(s).

CRISPR Deletion of rSOX-4 from S16 Cells

To identify the candidate target gene(s) regulated by rSOX-4, we used the CRISPR-Cas9 system to delete rSOX-4 from S16 cells (Wiedenheft et al. 2012; Mali et al. 2013). Briefly, repair templates were generated with approximately one kilobase pair arms of homology flanking rSOX-4 with a floxed blasticidin or neomycin resistance cassette. The S16 cells were transfected with the blasticidin repair template, human codon optimized Cas9 (hCas9) (Mali et al. 2013), and one of two guide RNAs (gRNAs) targeting rSOX-4. Next, the cells were selected for stable integration of the blasticidin repair template and flow sorted to generate clonal cell lines. We were able to generate two clones, one using gRNA-1 (rSOX-4 Clone 1-B) and one using gRNA-2 (rSOX-4 Clone 2-B) that had properly recombined in the blasticidin resistance cassette in place of rSOX-4. Integration was assessed using a diagnostic PCR with one primer outside the arms of homology and one primer within the drug repair template (Figure 3.8). These PCR products were sequence verified to ensure the integrity of the genomic DNA.

This process was repeated once more using the neomycin resistance template to generate double drug resistant clonal cell lines (Figure 3.8); however, the gRNAs had to be reversed due to InDels (Figure 3.9) in the remaining wildtype alleles (*i.e.* gRNA1 was used in rSOX-4 Clone 2-B and gRNA2 was used in rSOX-4 Clone 1-B). Additionally, we were unable to amplify the blasticidin repair template from rSOX-4 Clone 2 following the second round of clone generation. Nevertheless, after two rounds of CRISPR, there were no remaining wildtype alleles as assessed by a wildtype specific PCR with both primers within the deleted rSOX-4 region (Figure 3.10).

The clones were then transiently transfected with a Cre:GFP plasmid to remove the drug resistance cassettes, and GFP positive cells were flow sorted to generate clonal populations. A final diagnostic PCR was performed to assess for complete removal of the drug resistance cassettes, and the products were sequence verified to ensure only a single loxP scar remained (Figure 3.10). We were able to generate three clonal cell lines: two cell lines shared a parental cell line prior to Cre:GFP transfection (*i.e.* rSOX-4 Clone 2-1 and 2-2 were derived from rSOX-4 clone 2) while the other was independently generated (rSOX-4 Clone 1). The generation of rSOX-4 deleted cell lines will allow us to profile their gene expression and compare them to the gene expression profile of an unmodified S16 cell, to determine the difference between the two cells population and ultimately identify the candidate target gene(s) regulated by rSOX-4.

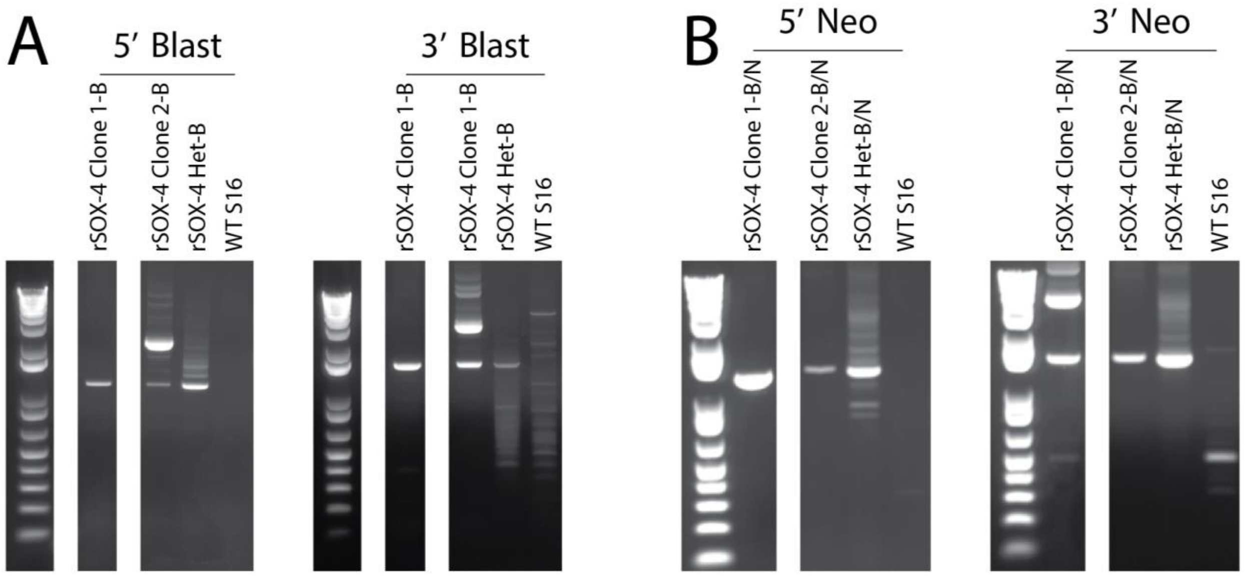


Figure 3.8 *Blasticidin and Neomycin Resistance Cassettes Stably Replaced rSOX-4 in S16 cells.* A blasticidin (A) or neomycin (B) repair template was used to delete rSOX-4 from S16 cells. Two clones were generated using different gRNAs targeting different sites within rSOX-4. rSOX-4 Clone refers to flow sorted clonal cells, rSOX-4 Het are the heterogeneous parental populations prior to flow sorting (-B = blasticidin resistant, -B/N = blasticidin and neomycin resistant), and WT S16 refers to the unmodified original parental S16 cell line. The 5’ and 3’ indicate across which arm of homology the diagnostic PCR was performed, Blast indicates a diagnostic PCR for the blasticidin resistance cassette, and Neo indicates a diagnostic PCR for the neomycin resistance cassette.

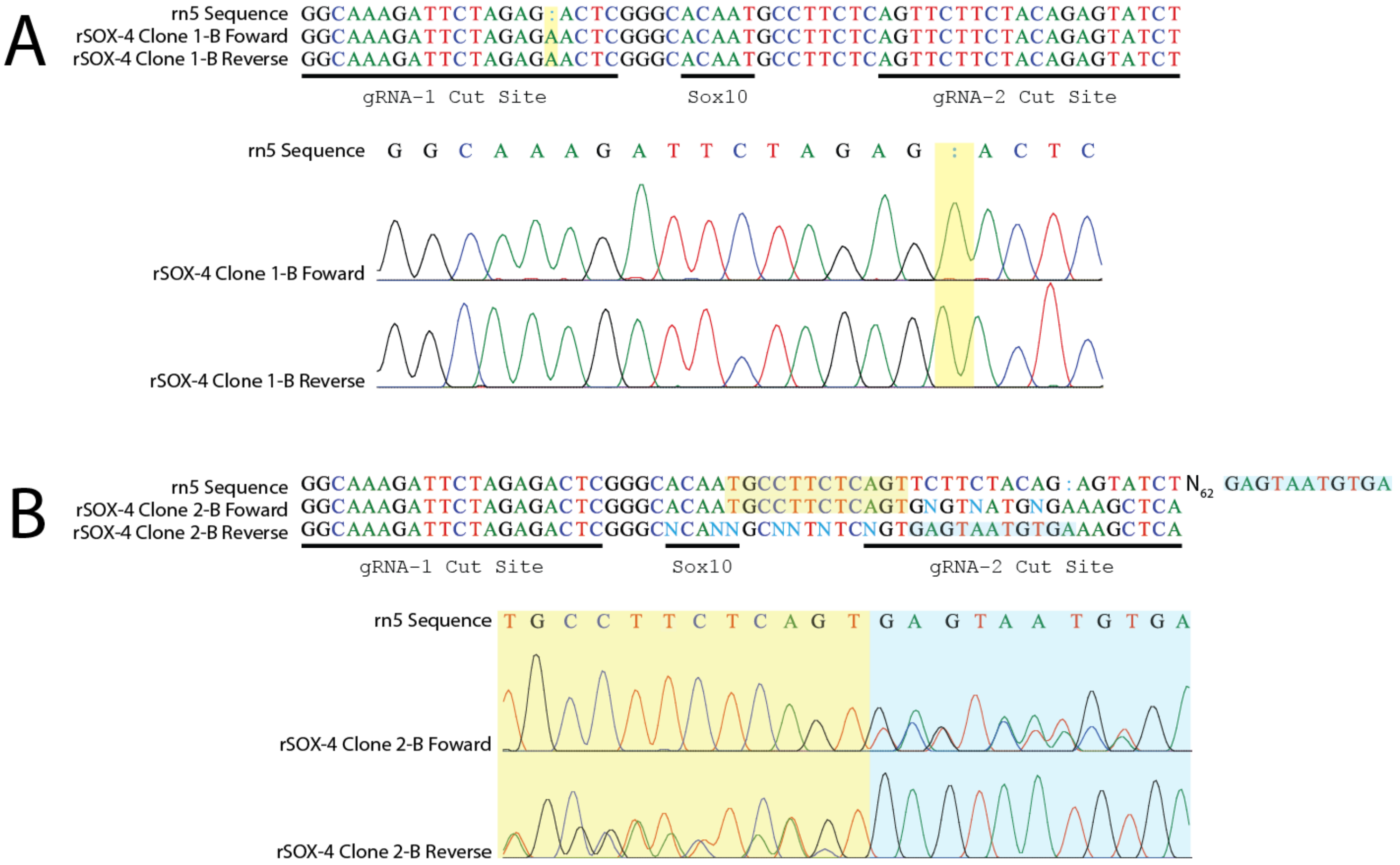


Figure 3.9 *CRISPR Cuts All Alleles in S16 Cells*. A wildtype specific diagnostic PCR was performed on both rSOX-4 Clone 1-B and 2-B, and the products were sequenced. (A) A single base pair insertion (yellow box) of an adenine was identified in the remaining nonrecombined alleles of rSOX-4 Clone 1-B that resides directly within the gRNA-1 cut site. (B) A 79 base pair deletion was detected encompassing the gRNA-2 cut site in rSOX-4 Clone 2-B. The yellow box represents the expected 5’ sequence while the blue box represents the 3’ sequence. For both A and B, the rn5 sequence is the expected sequence from the rn5 genome, and gRNA-1, Sox10 consensus site, and gRNA-2 are labeled and indicated by the line under the specific nucleotide sequences.

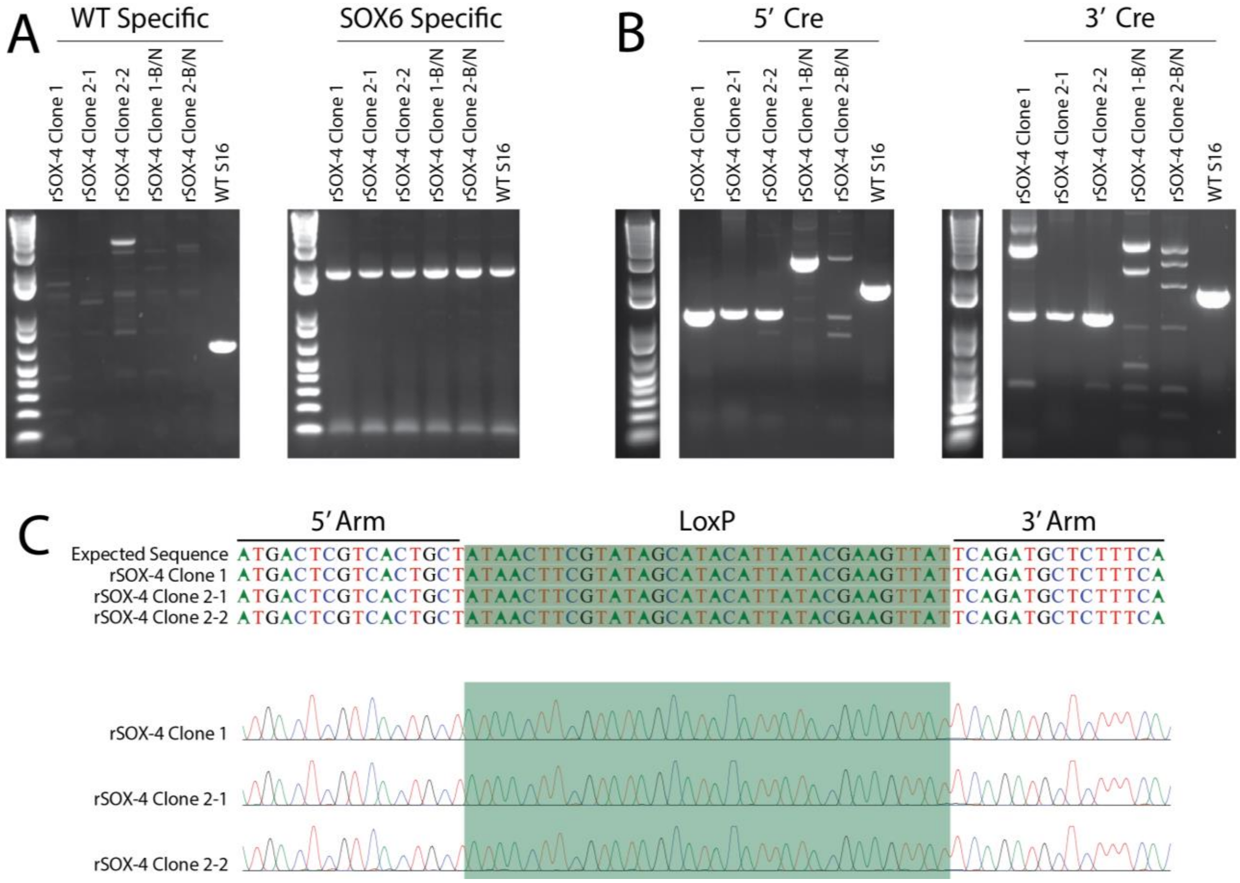


Figure 3.10 *No Wildtype Alleles of rSOX-4 Remain in the CRISPR Modified S16 Cells.* (A) A wildtype specific PCR was performed on the three Cre:GFP transfected clonal cell lines (rSOX-4 Clone 1, 2-1, and 2-2), the parental cell line prior to Cre:GFP transfection (rSOX-4 Clone 1-B/N and -2B/N), and unmodified S16 cells (WT S16). A SOX6 specific PCR was performed as a positive control. (B) A diagnostic PCR was performed across the site of integration (rSOX-4) for the three rSOX-4 clones, the parental cell lines, and unmodified S16 cells (WT S16). (C) Sequencing results from the bands of rSOX-4 Clone 1, 2-1, and 2-2 in B 5' Cre. The expected sequence was generated *in silico* based on the expected result for proper recombination and Cre excision. The 5' and 3' arms of homology are labeled above the sequence by the line, and the loxP scar is shaded in green. Only the sequencing of the 5' PCR product from B is shown for clarity, however sequencing from the 3' PCR confirmed these results.

Tubb2b is a Candidate Target Gene of rSOX-4

We performed RNA Sequencing (RNA-Seq) on the rSOX-4 clonal cell lines and two unmodified S16 cell lines to determine any gene expression changes genome-wide between the two populations. Briefly, RNA was extracted from the three rSOX-4 deleted cell lines and two unmodified S16 cell lines (one was the parental cell for all three rSOX-4 clonal cell lines). The RNA was subjected to 50 base pair single end reads, aligned to the rn5 genome using STAR (Dobin et al. 2013), and reads were counted for each gene using HTSeq (Anders et al. 2015). We then used DESeq2 (Love et al. 2014) to analyze differential gene expression between the mutant rSOX-4 cells and the unmodified S16 cells. This analysis identified 197 genes significantly differentially expressed between the two cell populations (Figure 3.11). We decided to filter these genes based on two criteria: (1) the target gene likely resides on the same chromosome as rSOX-4. While it is possible the target gene(s) reside on an different chromosome (Spilianakis et al. 2005), the vast majority of described CREs act in *cis* (Chapter 1). (2) The target gene(s) will demonstrate decreased expression in the rSOX-4 deleted cell lines compared to the unmodified S16 cells. Based on our luciferase data, we anticipate rSOX-4 acts as a transcriptional activator (enhancer) and deletion of this element would result in decreased gene expression. Of the 197 genes, only six genes were on the same chromosome (Chr 17) as rSOX-4, and only two genes showed decreased expression in the rSOX-4 mutants relative to unmodified S16 cells (fold change shown is relative to unmodified S16 cells): *Tubb2b* (-7.41) and *Gmnn* (-1.61) (Table 3.2).

To validate the RNA-Seq results, we performed digital droplet PCR (ddPCR) (Figure 3.12). While we were able to recapitulate the RNA-Seq results for *Tubb2b*, we were unable to validate the decrease in *Gmnn* expression. Across three independent experimental replicates, the expression of *Gmnn* in the rSOX-4 deleted cells relative to the unmodified S16 cells

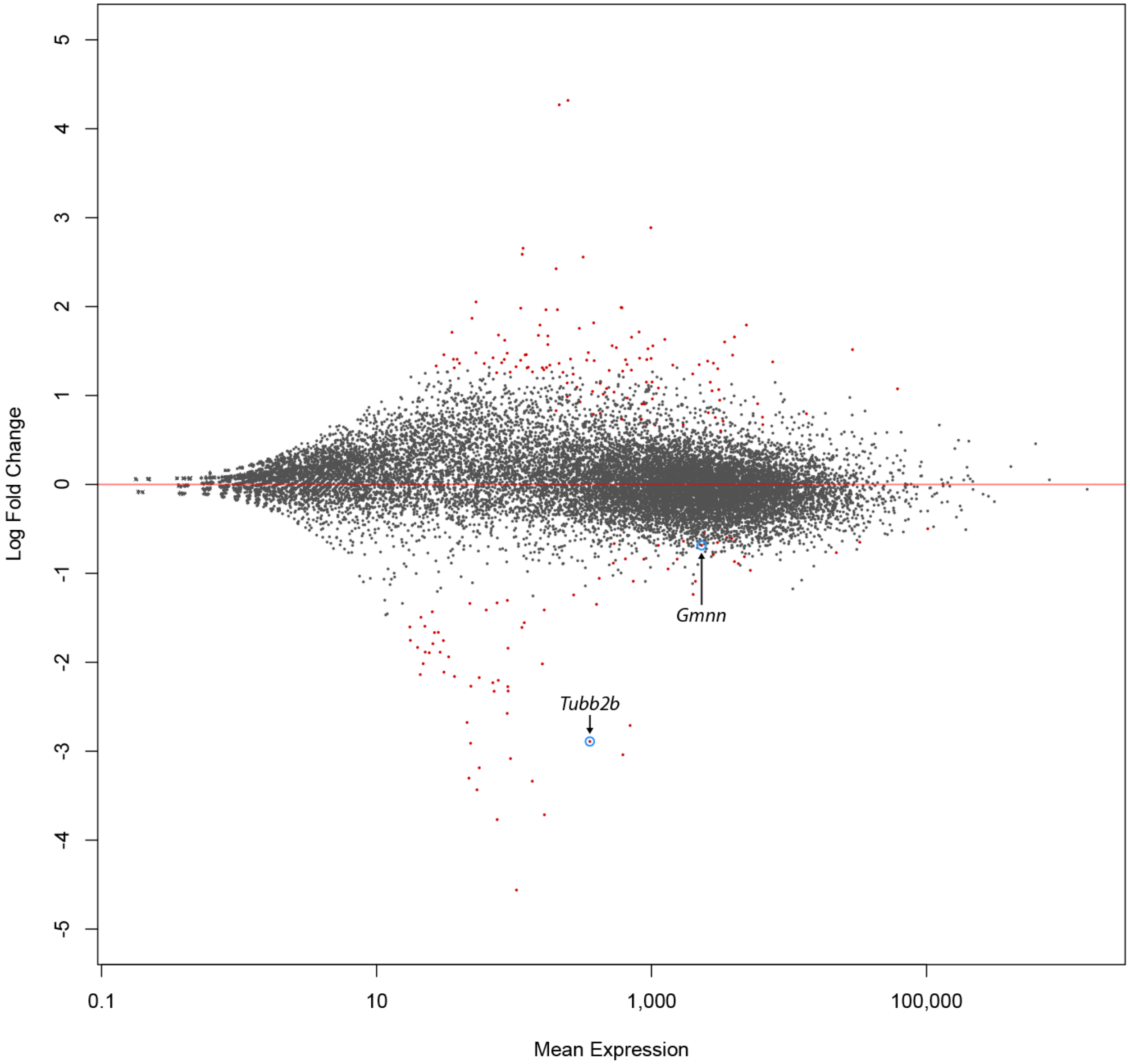


Figure 3.11 *Gmnn and Tubb2b are Significantly Decreased in Expression in the rSOX-4 Mutant Cells*. MA plot of the mean expression of every gene (dots) against the log 2 fold change. The mean expression is calculated as the mean of the normalized counts across all samples, and the log 2 fold change is relative to the unmodified S16 cells. Any gene above the red line (‘0’) indicates higher expression in the rSOX-4 mutant cells, and any gene below the red line indicates lower expression in the rSOX-4 mutant cells. The red dots indicate genes significantly differentially expressed between rSOX-4 and the unmodified S16 cells ($p < 0.05$). *Gmnn* and *Tubb2b* are labeled and indicated by the blue circles.

Table 3.2 *Six Genes are Significantly Differentially Expressed on Chromosome 17.*

Gene	**p-value**[1]	**Mean Expression**[2]	**Fold Change**[3]	**Distance**[4]
Tubb2b	1.48E-10	355.60 (51.5%)	-7.41	8.9
Rbm24	0.025	111.63 (44.2%)	2.63	23
Gmnn	0.042	2,309.92 (75.8%)	-1.61	1.9
Etl4	0.044	376.98 (52.0%)	1.73	47
Mylip	0.044	1,059.61 (63.2%)	1.59	20.5
Akr1c19	0.045	103.36 (43.8%)	2.50	29
Aldh5a1	0.949	494.33 (54.5%)	-1.09	1.7

[1] Benjamini-Hochberg adjusted p-value for multiple testing.

[2] Mean expression is the average of total counts across all cell lines. Number in parentheses is percentile rank of these genes relative to genes expressed in S16 cells (*i.e.* at least one read counted for the gene to be included).

[3] Fold change calculated relative to wild type cells. Negative value means lower expression in rSOX-4 mutant cells, and positive values means higher expression in rSOX-4 mutant cells.

[4] Distances from rSOX-4 to the gene are given in millions of base pairs (Mb).

demonstrated a significant increase, significant decrease, and no significant difference. This may be due to role of *Gmnn* in cell cycle regulation (McGarry and Kirschner 1998) which is confounding our results (discussed below); however, because we could not validate the RNA-Seq results for *Gmnn* we focused our efforts on *Tubb2b* as the likely candidate target gene of rSOX-4. As a control, *Aldh5a1* was included because it was not expected to display a difference based on RNA-Seq results, is expressed at approximately the same level as *Tubb2b*, and resides between rSOX-4 and both *Tubb2b* and *Gmnn*. In our ddPCR assay, *Aldh5a1* did not display significant differences in gene expression between the two cell populations.

One potential confounding variable in our experimental design is the affects of clonal expansion from single cells on gene expression. To account for this, we performed RNA-Seq and differential gene expression analysis using five additional unmodified S16 clones. The original, unmodified parental S16 cell line was flow sorted to generate clonal populations using front and side scatter. Five clones were selected, expanded into T-75 flasks, and RNA was isolated and sent for RNA-Seq. Identical procedures were used for RNA-Seq, mapping the reads, and determining differential gene expression as were used for the rSOX-4 mutants. We included these five clones with the previous unmodified S16 RNA-Seq dataset and reanalyzed the differential gene expression between the rSOX-4 mutants and all unmodified S16 cell lines. In agreement with the ddPCR results, *Tubb2b*, but not *Gmnn*, was significantly decreased in expression in the rSOX-4 mutants compared to all unmodified S16 samples (Figure 3.13 and Table 3.3). Combined, these results reveal *Tubb2b* as a candidate target gene of rSOX-4.

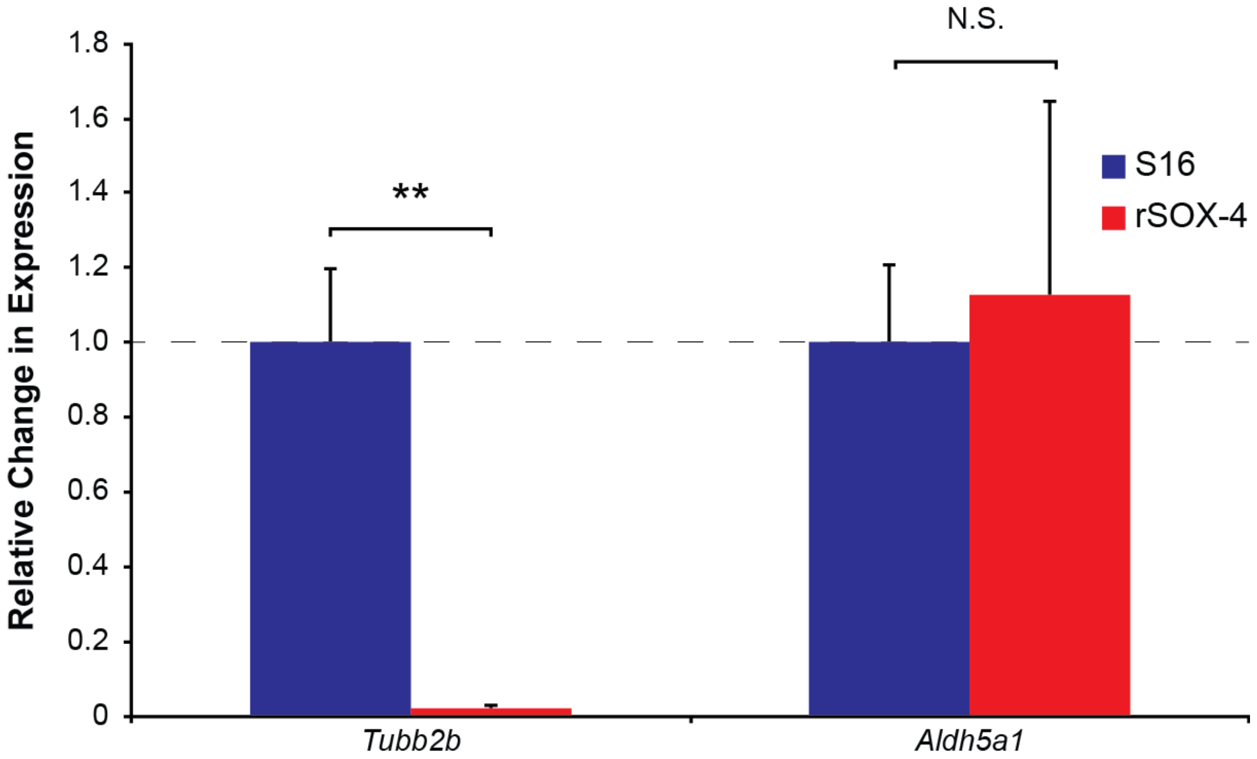


Figure 3.12 *Tubb2b is Significantly Decreased in Expression in ddPCR Assays*. Digital droplet PCR (ddPCR) was used to validate the results obtained from the RNA-Seq analysis. The expression of *Tubb2b* and *Aldh5a1* in unmodified S16 cells has been set to ‘1’, and the relative expression of the gene in the rSOX-4 mutant cells is displayed. *Aldh5a1* was used as a negative control. The blue bars are the unmodified S16 cells, red bars are the rSOX-4 mutant cells, and error bars represent standard deviation. Each bar represents four technical replicates of three biological replicates. ** = p < 0.05, and N.S. = not significant.

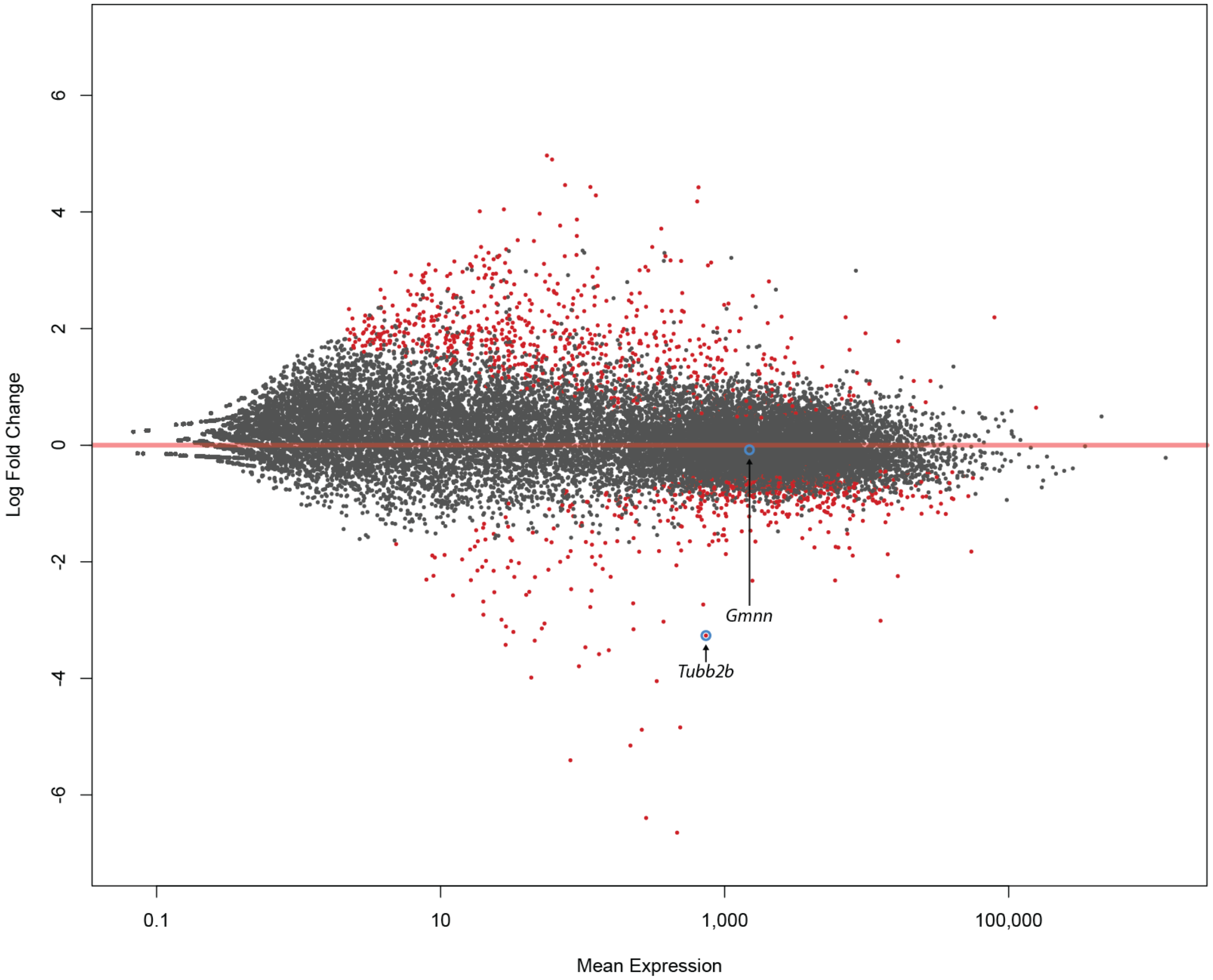


Figure 3.13 *Tubb2b is Significantly Decreased in Expression in the rSOX-4 Mutant Cells Compared to All Unmodified S16 Cells*. MA plot of the mean expression of every gene (dots) against the log 2 fold change. The mean expression is calculated as the mean of the normalized counts across all samples, and the log 2 fold change is relative to the unmodified S16 cells. Any gene above the red line ('0') indicates higher expression in the rSOX-4 mutant cells, and any gene below the red line indicates lower expression in the rSOX-4 mutant cells. The red dots indicate genes significantly differentially expressed between rSOX-4 and the unmodified S16 cells ($p < 0.05$). *Gmnn* and *Tubb2b* are labeled and indicated by the blue circles.

Table 3.3 *Tubb2b is Significantly Decreased in Expression in rSOX-4 Mutants Compared to All Unmodified S16 Cells.*

Gene	**p-value[1]**	**Mean Expression[2]**	**Fold Change[3]**	**Distance[4]**
Tubb2b	1.33E06	738.43 (65.3%)	-9.62	8.9
Gmnn	0.93	1,495.23 (74.2%)	-1.06	1.9

[1] Benjamini-Hochberg adjusted p-value for multiple testing.

[2] Mean expression is the average of total counts across all cell lines. Number in parentheses is percentile rank of these genes relative to genes expressed in S16 cells (*i.e.* at least one read counted for the gene).

[3] Fold change calculated relative to wild type cells. Negative value means lower expression in rSOX-4 mutant cells, and positive values means higher expression in rSOX-4 mutant cells.

[4] Distances from rSOX-4 to the gene are given in millions of base pairs (Mb).

rSOX-4 Directs LacZ Expression In Vivo in Mice at E11.5

To determine the physiological relevance of rSOX-4 in an *in vivo* mouse model, we assessed its ability to direct *LacZ* expression. Because the regulatory activity of rSOX-4 is dependent on SOX10, we anticipated rSOX-4 would demonstrate activity in tissues which express SOX10. Additionally, based on the RNA-Seq data suggesting *Tubb2b* as a candidate target gene, we selected embryonic day 11.5 (E11.5). This was due to the known role of *Tubb2b* in neuronal migration (Jaglin et al. 2009), and we reasoned it may be performing a similar role in migratory neural crest cells and their derivatives. By E11.5 in mice, the neural tube has closed, and neural crest cells and derivative cells have begun to migrate (Serbedzija et al. 1990; Serbedzija et al. 1992; Osumiyamashita et al. 1994; Wilson et al. 2004).

Primers were designed to amplify and clone rSOX-4 into pDONR221 in the reverse orientation because this was the more active orientation. The region was sequence verified to ensure the presence of the major allele and then LR cloned upstream of the minimal *Hsp68* promoter directing *LacZ* expression (Pennacchio et al. 2006). The rSOX-4:LacZ transgene was liberated from the plasmid backbone, gel purified, and injected into mouse zygotes. The zygotes were implanted into pseudopregnant mice and grown to embryonic day 11.5 (E11.5). The embryos were then harvested, and LacZ expression was detected. We isolated 54 embryos from ten mice and identified six LacZ positive mice (Figure 3.14).

Of the six embryos, one displayed blue expression throughout the embryo and most likely represents the transgene integrating into a ubiquitously expressed locus. The remaining five embryos all demonstrated LacZ staining in the dorsal root ganglia and migrating melanoblasts, both tissues which express SOX10 at this developmental timepoint (Britsch et al. 2001; Sonnenberg-Riethmacher et al. 2001). In addition, three of the five embryos displayed staining in

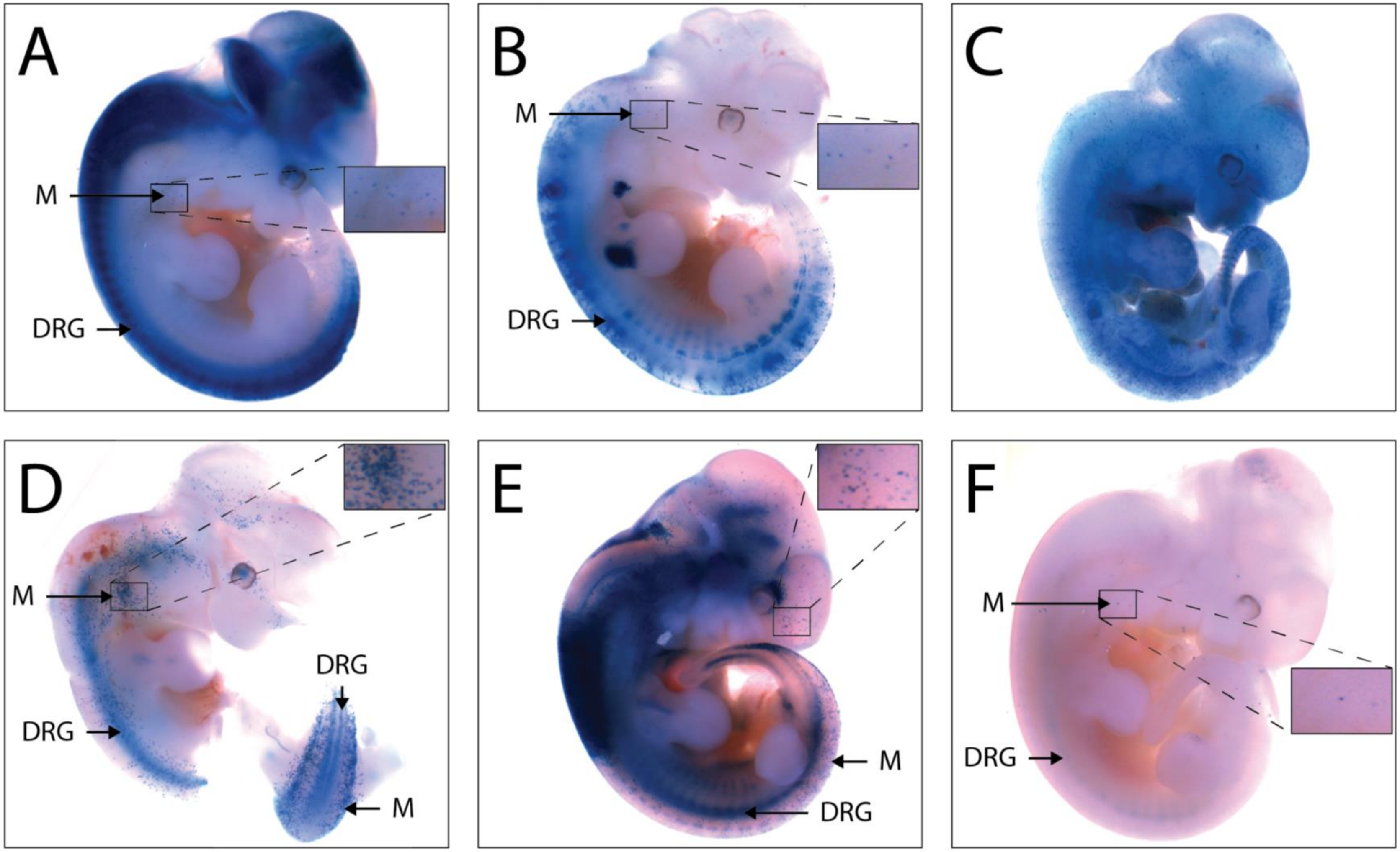


Figure 3.14 *rSOX-4 Demonstrates Enhancer Activity in vivo in Melanoblasts and the Dorsal Root Ganglia in Mice at E11.5.* Five of six mice displayed LacZ expression in a tissue-specific manner. One mouse (C) demonstrated near ubiquitous expression. All five tissue-specific expression mice demonstrated expression in the migrating melanoblasts (M) and dorsal root ganglia (DRG) indicated by the arrow. An enlarged section of the melanoblasts is shown in the upper right for each of the five mice.

the head and brain regions (Figure 3.14 - A, B, E, and F). These experiments demonstrate rSOX-4 is active *in vivo* in mice and combined with the CRISPR RNA-Seq data, suggest that rSOX-4 is regulating *Tubb2b*, which is playing an important role in neural crest cells and melanoblast migration.

Discussion

While utilizing an unbiased approach can identify many CREs important for relevant tissues, as was demonstrated in Chapter 2, the incorporation of transcription factor binding site information allows for a more focused approach to generate transcriptional hierarchies. A transcription factor centric approach also allows for the use of genome-wide datasets such as ChIP-Seq, which ultimately revealed rSOX-4 and *Tubb2b* as strong candidate enhancers and genes relevant for SOX10 expressing tissues, respectively. We therefore modified our computational pipeline from Chapter 2 to incorporate the well-characterized TFBS for SOX10. SOX10 was selected due to its importance in Schwann cells, the identification of mutations within characterized SOX10 binding sites causing peripheral neuropathies (Houlden et al. 2004; Jones et al. 2011a), and the identification of regulatory SNPs within SOX10 binding sites that alter disease severity (Emison et al. 2005; Emison et al. 2010). We identified 224 conserved SOX10 binding sites harboring a SNP. These were prioritized for analysis by using SOX10 ChIP-Seq data (Srinivasan et al. 2012) and dimeric SOX10 binding sites to generate 22 prioritized regions. Surprisingly, of the 22 assessed, only four regions demonstrated strong luciferase activity, despite all 22 regions harboring a conserved SOX10 monomer or dimer.

There are several possible explanations for this observation, such as the role of SOX10 in other tissue types. Indeed, SOX10 is critical for neural crest development, and mutations within SOX10 target genes can affect multiple tissues including Schwann cells (i.e *PMP22*; CMT1) (Jones et al. 2011a), enteric neurons (*i.e. RET*; Hirschsprung's disease) (Emison et al. 2010), and melanocytes (*i.e. MITF*; Waardenburg syndrome) (Bondurand et al. 2000). These 18 regions may be functional, but the S16 cells are missing additional co-factors necessary for activity such as PAX3 in melanocytes (Bondurand et al. 2000). One caveat to this explanation is that seven of the 18 regions overlap a Sox10 ChIP-Seq peak generated from rat sciatic nerves (Srinivasan et al. 2012). While this discrepancy may be a limitation of our *in vitro* Schwann cell model compared to the *in vivo* dataset, ChIP-Seq peaks do not necessarily predict a functional regulatory element (Fisher et al. 2012; Kvon et al. 2012; Leonid et al. 2013). Indeed, rSOX-22 displayed strong luciferase activity and was validated as a SOX10 response element despite not overlapping a SOX10 ChIP-Seq peak.

Another possible explanation for the lack of activity observed in the 18 regions is that they function at a different developmental time point in Schwann cell maturation. For example, these regions may only be active in neural crest cells or early Schwann cell precursors and require additional co-factors such as AP2α (Stewart et al. 2001; Jessen and Mirsky 2005; Wahlbuhl et al. 2012). Additional studies will be necessary to understand and characterize the roles (if any) of these 18 less active regions.

From the four active regions identified, we focused on two regions, rSOX-4 and rSOX-22, because of the significant differences in allele-specific regulatory activity. Through overexpression studies using dominant-negative SOX10 in the S16 cells and wildtype SOX10 in the MN-1 cells, as well as deletion of the SOX10 consensus site, we were able to show that

SOX10 is both necessary and sufficient for luciferase activity of both regions in our cell culture models. Interestingly, upon deletion of the SOX10 consensus site in rSOX-4, there was still significant (~4.5 fold increase relative to empty) activity suggesting there may be additional SOX10 binding sites. Indeed, there are two additional SOX10 consensus sites conserved among human, mouse, and chicken; however deletion of these sites independently or in tandem did not further reduce luciferase activity compared to the single deletion of the original rSOX-4 SOX10 consensus site (Figure 3.15). These data suggest there may be additional TFBSs that are responsible for the residual activity observed in the SOX10 deletion constructs. Using the TRANSFAC Match tool (Kel et al. 2003) to analyze the rSOX-4 region, binding sites for two transcription factors critical for early Schwann cell development were identified: two AP2α (Stewart et al. 2001; Jessen and Mirsky 2005; Wahlbuhl et al. 2012) and one YY1 (He et al. 2010). Deletion of these predicted TFBS individually and in tandem with the SOX10 binding site may explain the residual activity observed in the SOX10 deleted rSOX-4 construct.

Based on the additional support of genomic features (*i.e.* ChIP-Seq, DNase HSS, and H3K27Ac peaks) we focused our efforts on identifying the target gene of rSOX-4. To do this, we utilized CRISPR to delete the region in S16 cells. We performed two rounds of drug selection to target all wildtype alleles, but we observed that all wildtype alleles were cut by Cas9 in a single round of transfection, despite not necessarily being repaired by homologous recombination. This suggests that simultaneous targeting of all wildtype alleles could be performed in a single experiment by transfecting both the blasticidin and neomycin repair templates. This strategy may be hampered by multiple factors including the efficiency of transfection, Cas9 DNA editing, homologous recombination rates, and the actual number of wildtype alleles in a polyploid immortalized cell line. In fact, while the karyotype of our S16 cells is unknown, we did perform

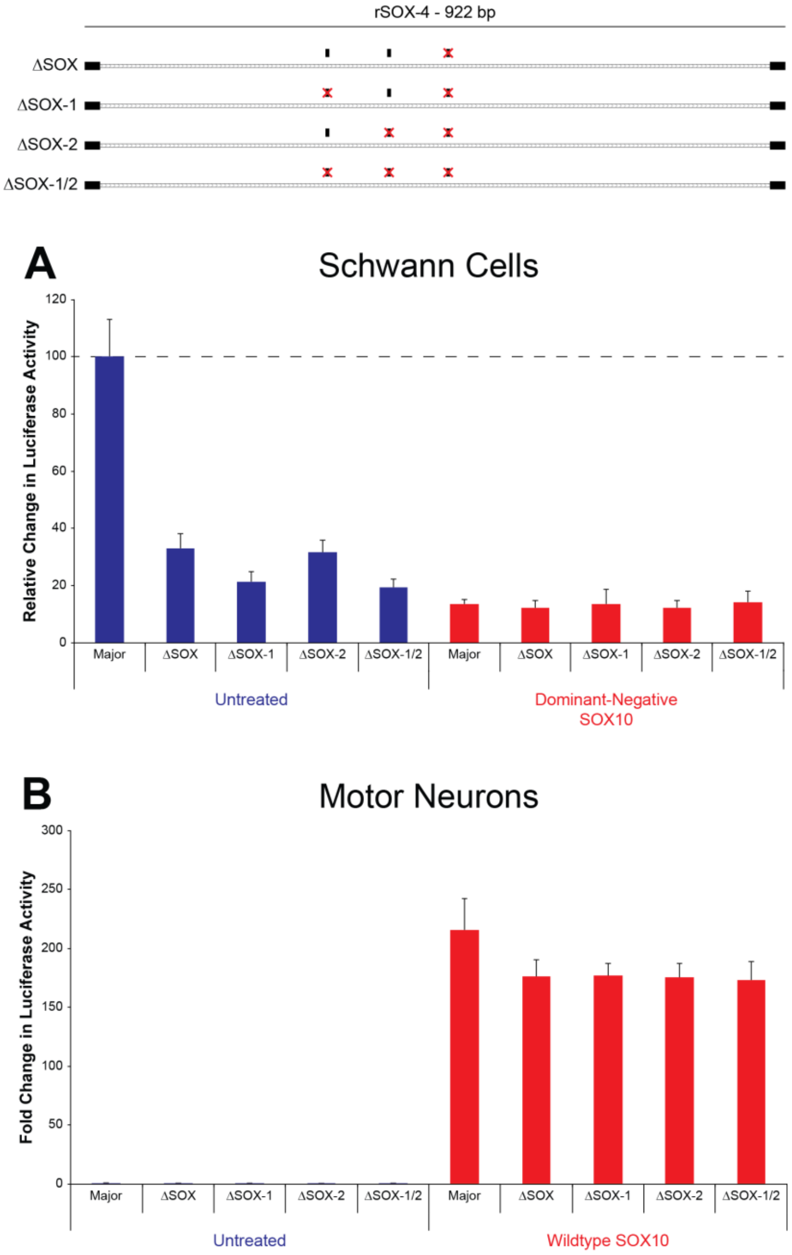


Figure 3.15 *Deleting Additional SOX10 Consensus Sites in rSOX-4 has No Effect on Luciferase Activity*. (Top) A schematic of the three SOX10 consensus sites (black boxes) with each site deleted (indicated by the red Xs). The name of each construct is shown at the left, and a scale bar is shown at the top. (A) The four SOX10 deletion constructs were assessed in the reverse orientation only in the S16 cells either untreated (left side; blue bars) or co-transfected with dominant-negative SOX10 (right side; red bars). All activity is relative to the untreated major activity which as been set to '100'. (B) The four SOX10 deletion constructs were assessed in the reverse orientation only in the MN-1 cells either untreated (left side; blue bars) or co-transfected with wildtype SOX10 (right side; red bars). All activity is relative to the untreated major activity which as been set to '1'.

metaphase spreads and observed approximately triploid chromosome counts (data not shown). This is in agreement with spectral karyotyping (SKY) (Schröck et al. 1996) of different subcultures of S16 cells (unpublished data) which detected near triploid chromosome counts. Interestingly, one of the more common structural abnormalities observed was complete loss of chromosome 17 (where rSOX-4 resides). Our S16 subcultures do not appear to be missing chromosome 17, since two rounds of CRISPR targeting and drug selection were required to modify all alleles present. Regardless, we were able to target all wildtype alleles in the S16 cells to remove rSOX-4 and leave only a single loxP scar behind.

One major problem of using CRISPR in cell lines is the potential for off-target effects (Fu et al. 2013; Pattanayak et al. 2013). This was partially mitigated in our analysis by the independent generation of two clonal cell lines; however due to the InDels generated in the non-recombined alleles, we were unable to use a single gRNA for each clone throughout both rounds of targeting. Because of this, it is possible the two clonal cell lines share similar off target affects. Current methods to detect off target effects such as GUIDE-Seq (Tsai et al. 2015) are only able to predict candidate off target sites prior to CRISPR editing and would be unable to determine off target effects in our edited cell lines. This procedure could be employed to assess candidate sites which could be screened in our modified cells.

A different method to detect off target effects could employ a modified version of L1-Seq (Iskow et al. 2010). L1-Seq was developed to detect novel integration sites of LINE1 elements in cell lines and could potentially identify off target sites where the repair template integrated. This method, however, would be unable to detect small InDels similar to what we observed in the non-recombined alleles of the rSOX-4 cell lines between rounds of CRISPR editing. A final method to detect off target effects would be whole genome sequencing of our CRISPR modified cell lines. This is currently limited in three ways: (1) the S16 cells are immortalized cells with unknown genomic rearrangements and an unknown karyotype (2) the rat genome is poorly annotated making distinctions between SNPs and novel mutations caused by CRISPR challenging and (3) the prohibitive cost of sequencing the rSOX-4 cells to identify novel InDels. The second issue could be partially mitigated by sequencing the unmodified parental cells, but significant problems still arise when mapping the reads back to the rat reference genome, due to unknown genomic rearrangements. In this study, we attempted to minimize false positive target genes by leveraging two independent rSOX-4 mutant cell lines. Future studies should use unique gRNAs for each clonal cell line, potentially designing additional unique gRNAs for every clonal line as necessary.

To detect any gene expression changes based on the loss of rSOX-4, we performed RNA-Seq analysis on the three rSOX-4 mutant clones and two unmodified S16 cell lines. One of the unmodified S16 cell lines was the original founder for all three rSOX-4 mutant clones, while the other was an older passage (passage 39) stock. The older passage stock was chosen to account for potential effects of additional passages on gene expression. After mapping, read counting, and using DeSeq2 to determine gene expression differences, we identified 197 genes that were significantly differentially expressed between the two groups. These 197 differentially expressed genes may be the result of off target effects. Indeed, we observed three distinct clusters of genes (genes that are either direct neighbors or separated by a single gene) that were significantly differentially expressed. Interestingly, all genes within a cluster demonstrated the same direction of the fold change (*i.e.* all were either upregulated or downregulated), and for two of the three clusters this resulted in increased gene expression in the rSOX-4 mutants relative to the unmodified S16 cells (Table 3.4). This may suggest a nearby insertion of the CMV promoter, that was used to direct blasticidin or neomycin expression, is affecting nearby gene expression.

Conversely, 35 of the 197 genes were also significantly differentially expressed between the five unmodified S16 clones. Within the five modified clones, RNA was isolated from three actively dividing cell lines while the other two were isolated from fully confluent flasks (*i.e.* not actively dividing). This suggests that these 35 genes may depend on the cell cycle and are confounding interpretation of the RNA-Seq results. For example, *Ube2c* is a known cell cycle regulator (Townsley et al. 1997), and is one of the 35 genes significantly differentially expressed in both the rSOX-4 compared to unmodified S16 dataset and the S16 clone dataset. Another gene common to both datasets is *S100a6* which is also involved in cell cycle progression and senescence (Słomnicki and Leśniak 2010; Bao et al. 2012).

Despite potential disruption of cell cycle regulators, and other unknown variables, we anticipated the deletion of rSOX-4 would result in decreased gene expression in the rSOX-4 mutant cell lines relative to the unmodified S16 cells based on the luciferase data that rSOX-4 is acting as an enhancer. In addition, while it is possible that the target gene of rSOX-4 resides on a different chromosome, we focused our efforts on downregulated genes in *cis* with rSOX-4. Only two dysregulated genes met this criteria: *Gmnn* and *Tubb2b*. Upon further examination, only *Tubb2b* could be validated using ddPCR with independent cDNA samples and by additional RNA-Seq results using five clonal unmodified S16 cell lines.

Interestingly, *Gmnn* is a known cell cycle regulator (McGarry and Kirschner 1998), and although it was not one of the 35 genes discussed above, it may explain why *Gmnn* was a false positive. It is important to note that *Gmnn* has additional roles that make it an attractive candidate target

Table 3.4 *Tubb2b is Significantly Decreased in Expression in rSOX-4 Mutants Compared to All Unmodified S16 Cells.*

Gene	**p-value[1]**	**Mean Expression[2]**	**Fold Change[3]**	**Distance[4]**
Tubb2b	1.33E06	738.43 (65.3%)	-9.62	8.9
Gmnn	0.93	1,495.23 (74.2%)	-1.06	1.9

[1] Benjamini-Hochberg adjusted p-value for multiple testing.

[2] Mean expression is the average of total counts across all cell lines. Number in parentheses is percentile rank of these genes relative to genes expressed in S16 cells (*i.e.* at least one read counted for the gene).

[3] Fold change calculated relative to wild type cells. Negative value means lower expression in rSOX-4 mutant cells, and positive values means higher expression in rSOX-4 mutant cells.

[4] Distances from rSOX-4 to the gene are given in millions of base pairs (Mb).

gene of rSOX-4. Recent studies where *Gmnn* expression was decreased in mice resulted in inhibition of the epithelial to mesenchymal transition required for neural crest delamination (Emmett and O'Shea 2012). Additionally, when *Gmnn* is specifically deleted in neural crest cells of mice, the enteric nervous system fails to develop, and the mice exhibit severe Hirschprung's disease phenotypes (Stathopoulou et al. 2016). While further experiments such as deletion of rSOX-4 *in vivo* in mice may be necessary to conclusively exclude *Gmnn* as the target gene of rSOX-4, our data at present were unable to validate the initial RNA-Seq results.

Tubb2b is a critical component of microtubules, and mutations of this gene can result in polymicrogyria (Jaglin et al. 2009). There is no known role of *Tubb2b* in either the neural crest nor Schwann cells; however, *Tubb2b* expression is highest in migratory neurons (Jaglin et al. 2009; Breuss et al. 2015). Since rSOX-4 directed LacZ expression in migratory melanoblasts, this may suggest that *Tubb2b* has a similar, uncharacterized role in these and other neural crest derived cells; specifically, in cell migration. A mouse harboring a GFP transgene reporter inserted into the endogenous *Tubb2b* locus has been generated (Breuss et al. 2015). The earliest time point the authors looked at was E14.5, which is after melanoblast migration has completed (Mort et al. 2015); however, the expression patterns are strikingly similar, as the authors observe GFP expression within the presumptive spinal cord and GFP expression in the head and brain region. Using this mouse to investigate earlier time points, such as E11.5 when melanoblasts are migrating, may uncover a novel role for *Tubb2b*.

Interestingly, in the most recent release of the human genome (GRCh38), rSOX-4 resides within the first intron of a long non-coding RNA (lncRNA), LOC105374972. At this time there is no known function of the lncRNA, but it has been classified as 'validated'. Converting the coordinates of the lncRNA from human to rat, we were unable to detect any RNA-Seq reads corresponding to the lncRNA exons. This may be due to generating a polyA selected RNA library for RNA-Seq, and being unable to capture this lncRNA because it may not be polyadenylated (Yang et al. 2011). Another reason for the lack of RNA-Seq reads is that the lncRNA may also not be present in Schwann cells. Based on our *in vivo* rSOX-4:LacZ mice which demonstrated that rSOX-4 is active in the dorsal root ganglia and melanoblasts, the lncRNA may play a role in these other tissues. Additional studies will be necessary to elucidate any functional role of the lncRNA and determine if rSOX-4 regulates the lncRNA.

In this chapter, we modified the unbiased rSNP computational identification program developed in Chapter 2 to include specific TFBS information. We used SOX10 because it has a well characterized binding site (Peirano and Wegner 2000; Srinivasan et al. 2012), and it is critical for Schwann cell function (Chapter 1). While we were able to successfully identify two SOX10 response elements with rSNPs, the pipeline outlined here is very modular and can be used for the identification of other TFBSs. Indeed, we modified our pipeline to predict THAP1 binding sites by including information about THAP1 binding sites, removing the SNP filter, and using conservation among human, mouse, and dog at candidate genes identified through additional experiments. THAP1 is a transcription factor that, when mutated, causes torsion dystonia in humans (Fuchs et al. 2009). In collaboration with Dr. William Dauer, we were able to use this modified pipeline to quickly identify and functionally assess 13 regions near the candidate genes. One of these regions demonstrated decreased expression when the THAP1 consensus site was deleted, and additional experiments are being performed to determine the functional role of this response element.

The pipeline established in this chapter also allows for the rapid characterization of variants of unknown significance identified in patients. Recently we were contacted by a physician regarding a variant of unknown significance identified in a 13-year-old boy with CMT1. The variant resides in the intron of *MPZ*, and the patient had no coding mutations in any other known CMT disease genes. This variant converted one predicted SOX10 monomer to a different SOX10 monomer, and was located 117 base pairs downstream (3') of a previously characterized SOX10 element (Antonellis et al. 2010). Upon testing the wildtype allele, patient variant, and a deleted SOX10 consensus site in our luciferase assays in S16 cells, we observed a large increase in activity of the patient variant compared to the wildtype allele in Schwann cells (Figure 3.16). We also observed low luciferase activity in motor neurons consistent with this being is a SOX10 response element. Interestingly, increased *MPZ* expression has been shown to cause CMT1 in patients (Maeda et al. 2012). Additional studies will be necessary to demonstrate that this variant is causative, however the functional pipeline developed in this chapter allowed rapid assessment of this variant and can be used for similar studies in the future.

As demonstrated in this chapter and in Chapter 2, our computational pipeline can successfully identify TFBSs harboring SNPs. We were able to identify 224 fully conserved SOX10 binding sites and prioritized 22 using additional genomic datasets. We rapidly functionally evaluated the regions using luciferase assays to identify two novel rSNPs within SOX10 response elements. We focused on one region, rSOX-4, and deeply characterized the region through CRISPR knock out and *in vivo* transient transgenic mouse reporter experiments. Through these (and other) experiments, *Tubb2b* was identified as a strong candidate target gene of rSOX-4. While additional experiments will be necessary to confirm our results the rSNP within rSOX-4, rSOX-4 itself, and *Tubb2b* represent excellent candidate modifiers of neurocristopathies. Additionally, understanding and characterizing the role of *Tubb2b* in neural crest cells and in particular melanoblasts will give novel insight into the development of these cell populations.

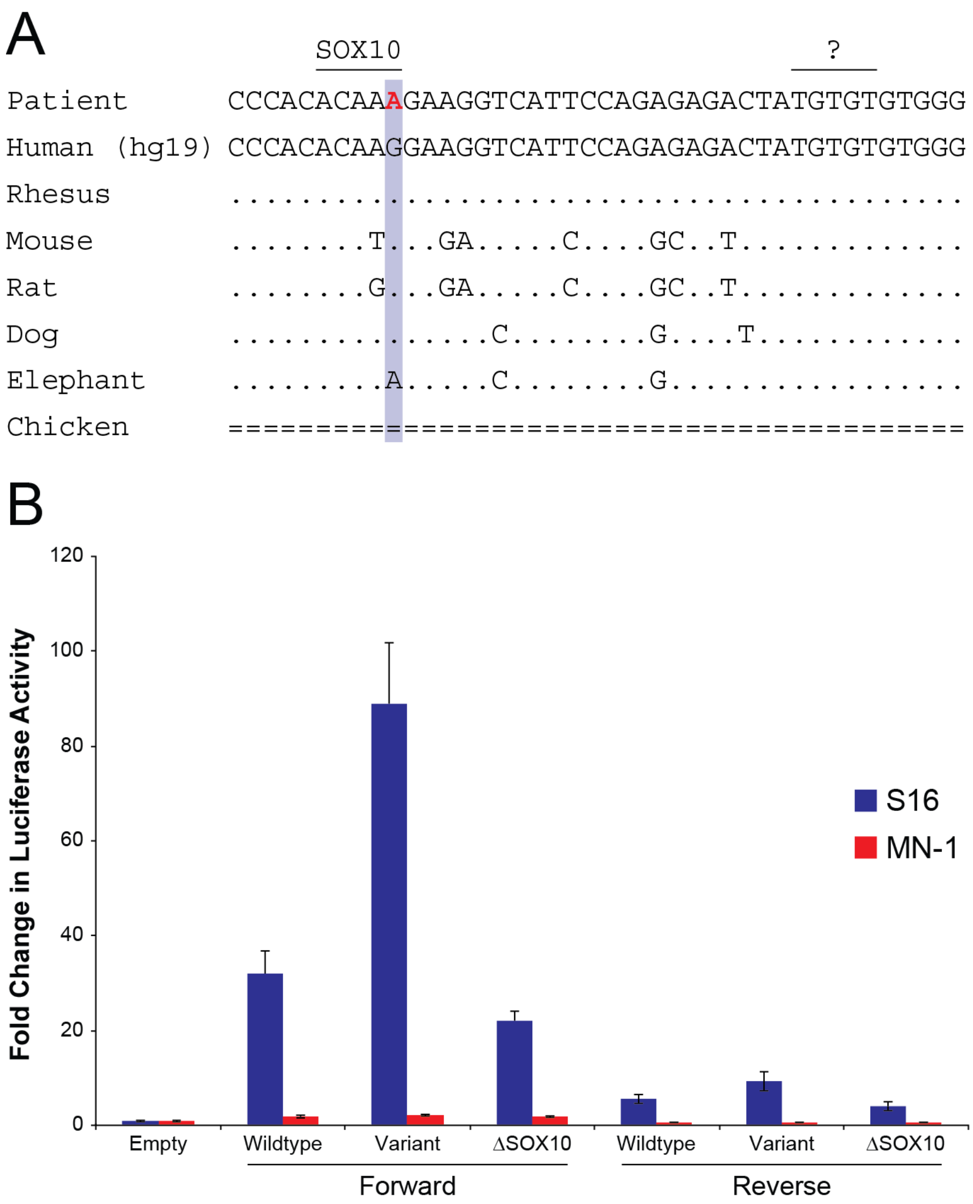


Figure 3.16 *A Variant Allele Identified in a Patient with CMT1 Increases Luciferase Activity Compared to the Wildtype Allele*. (A) The sequence surrounding the patient variant (red letter; purple box) within a SOX10 consensus site indicated by the line at the top. The other line (labeled with '?') indicates another potential SOX10 monomeric or dimeric site. Only changes in sequence between the indicated species and humans is display. The '.' Indicates identical base pair as in humans and '=' indicates no sequence alignment. (B) A region surrounding the patient variant was amplified and assessed in luciferase assays. Three constructs were generated harboring either the wildtype, patient variant, or deletion of the SOX10 monomer (ΔSOX). The regions were assessed in either the forward (left bars; Forward) or reverse (right bars; Reverse) in both Schwann cells (S16; blue bars) or motor neurons (MN-1; red bars). The activity of each allele is compared back to a control plasmid that contains no insert ('Empty') that has been set to '1', and error bars represent standard deviation.

In the next chapter, we will utilize the modularity of our computational pipeline to remove the SNP filter and identify novel dimeric SOX10 binding sites within known genes associated with Schwann cell function. We also identify a subset of SOX10 response elements near genes implicated as negative regulators of myelination and start to address how SOX10 expression remains constant even in non-myelinating Schwann cells.

CHAPTER 4

Stringent Comparative Sequence Analysis Reveals SOX10 as a Putative Inhibitor of Glial Cell Differentiation

Introduction

Schwann cells produce the myelin sheath in the peripheral nervous system (PNS), which allows rapid saltatory conduction and long-range communication between the central nervous system and innervated muscles and sensory organs. Schwann cell development is directed by a transcriptional hierarchy that promotes the expression of proteins important for migration along peripheral nerves, radial sorting of axons, and the initiation of myelination (Jessen and Mirsky 2005; Stolt and Wegner 2015). Atop this hierarchy sits the transcription factor SOX10, which is critical for the development and long-term function of Schwann cells (Kuhlbrodt et al. 1998) and is expressed during all stages of Schwann cell development (Kuhlbrodt et al. 1998; Britsch et al. 2001).

Three major lines of evidence underscore the importance of SOX10 for the function of Schwann cells. First, loss of Sox10 at any developmental stage results in dramatic phenotypes in mice: ablation of Sox10 activity during early development results in a lack of Schwann cells (Britsch et al. 2001), conditional deletion of *Sox10* in the immature Schwann cell stage results in lethality in mice (Finzsch et al. 2010), and conditional deletion of *Sox10* in terminally differentiated myelinating Schwann cells results in demyelination of the peripheral nerves (Bremer et al. 2011). Second, dominant-negative *SOX10* mutations cause an autosomal dominant disease characterized by peripheral demyelinating neuropathy, central dysmyelinating leukodystrophy, Waardenburg-Shah syndrome, and Hirschsprung disease (PCWH) (Inoue et al. 1999; Inoue et al. 2004); the non-PNS phenotypes reflect the role of SOX10 in other neural crest derivatives (*e.g.* melanocytes and enteric neurons) and in oligodendrocytes. Finally, mutations in SOX10 target genes, including those encoding peripheral myelin protein 22 (*PMP22*) (Lupski et al. 1991; Raeymaekers et al. 1991), myelin protein zero (*MPZ*) (Kulkens et al. 1993), early growth response 2 (EGR2) (Warner et al. 1998), and gap junction beta 1 (*GJB1*) (Ionasescu and Searby 1994), cause demyelinating peripheral neuropathy.

The identification of additional SOX10 response elements and target loci will provide important information on the process of myelination in the peripheral nerve as well as novel target sequences to scrutinize for mutations and modifiers of peripheral neuropathy. Indeed, genome-wide analyses have been essential for characterizing SOX10 biology in Schwann cells (Lee et al. 2008; Srinivasan et al. 2012); Chapter 3); however, these efforts have primarily focused on identifying positive regulators of myelination by examining tissues or cells in a myelinating state. Less-biased approaches are needed to complement the above studies and to identify functions of SOX10 outside of the regulation of promyelinating loci.

Here, we modify our stringent computational strategy developed in the previous chapters to rapidly identify SOX10 response elements in the human genome. Combined with molecular functional studies, this strategy revealed SOX10 response elements residing near *SOX5*, *SOX6*, *NOTCH1*, *HMGA2*, *HES1*, *MYCN*, *ID4*, and *ID2*. Interestingly, each of these genes has a known role in the negative regulation of glial cell differentiation. As such, we have identified a potentially novel role for SOX10 in Schwann cells and present a model where SOX10 activates the expression of negative regulators of myelination to temper the pro-myelinating program during non-myelinating stages of Schwann cell development; however additional *in vivo* functional studies will be necessary to test this proposed model.

All the work in this chapter was performed by the author with the exception of the following. A portion of the computational analyses were performed by Tony Antonellis and Arjun Prasad (NIH/NCBI). A portion of the cloning and functional assessment of the 65 regions presented in this chapter were performed by Chetna Gopinath. The follow-up studies focusing on the alternative promoter at *Sox6* (*i.e.* RT-PCR, 5'-RACE, ect.) was performed by Chetna Gopinath as part of her thesis research. The work presented in this chapter resulted in a co-first author (Law and Gopinath) manuscript, which is currently under review. Finally, the siRNA experiments in S16 cells (Figure 4.9 A), the mRNA expression analysis at multiple developmental timepoints from sciatic nerves in rat (Figure 4.9 B), and the siRNA experiments using primary Schwann cells (Figure 4.10) were performed by José F. Rodríguez-Molina in John Svaren's laboratory (University of Wisconsin-Madison).

Methods

Computational Identification and Prioritization of SOX10 Consensus Sequences

To identify all SOX10 consensus sequences in the human genome, we downloaded individual text files for each human chromosome (hg18) from the UCSC Human Genome Browser (Kent et al. 2002) and wrote a Perl script (available upon request) that examines each file for the SOX10 consensus sequences (using a regular expression analysis; 5'-3'): ACACA, ACAAA, ACAAT, ACAAG. To identify two SOX10 consensus sequence monomers that are oriented in a head-to-head manner (and that may represent a dimeric SOX10 binding site), we wrote a second Perl script that examines the human chromosome text files and reports each ACACA, ACAAA, ACAAT, or ACAAG consensus sequence that is five to 10 base pairs 5’ to the reverse complement of this consensus sequence (5’-3’): TGTGT, TTTGT, ATTGT, or CTTGT.

To identify genomic sequences that are identical between human, mouse, and chicken, we downloaded the vertebrate (44 species) multiz alignment (maf) files from the UCSC Human Genome Browser (hg18) and extracted the alignments for human, mouse, and chicken. Next, we utilized the program ExactPlus (Antonellis et al. 2006) to identify all human sequences that are at least five base pairs long and identical across all three species. All subsequent computational analyses that assess for overlap between these and other datasets were performed using the UCSC Table Browser (Karolchik et al. 2004) and the ‘intersection’ tool. For these analyses we employed UCSC Genome custom tracks containing each: (**1**) human RefSeq (hg18) protein-coding sequence to exclude coding sequences; (**2**) human RefSeq entry (hg18) plus 2.5 kb upstream and 2.5 kb downstream of the transcriptional unit to identify regions that map to known genes; (**3**) SOX10 ChIP-Seq peak in the rat genome (rn5) using HOMER analysis (Heinz et al. 2010) of previously described P15 sciatic nerve datasets; and (**4**) DNase-Seq peak in the rat genome (rn5) that has an F-Seq (Boyle et al. 2008) score of at least 0.08 (see below).

To identify the 57 loci (Appendix 6) with a known or predicted role in peripheral nerve myelination, we performed the following PubMed searches in September 2014: (**1**) each gene name plus ‘Schwann’; and (**2**) each gene name plus ‘Myelin’. We also searched for each gene name plus ‘Schwannoma’ in the GEO Profiles database at NCBI to determine if gene expression is depleted upon treatment with SOX10 siRNA (Lee et al. 2008).

Standard and Quantitative RT-PCR

Total RNA was isolated from S16 and MN-1 cells. 100,000 cells were plated in a 6-well plate and incubated overnight at 37°C in 5% CO_2. For an individual well, 5 μL of Lipofectamine 2000 was mixed with 500 μL of Optimem (ThermoFisher cat no 31985-062) and incubated for 10 minutes at room temperature (cocktail 1). 4 μg of wildtype SOX10 or E189X SOX10 (Inoue et al. 2004) plasmid DNA was mixed with 500 μL of Optimem and combined with the first cocktail after the 10 minute incubation (transfection mixture). The transfection mixture was vortexed for three seconds and incubated for 20 minutes at room temperature. During the final incubation, the S16 cells were washed with 2 mL of 1X PBS (ThermoFisher Cat no. 10010-023), and 1 mL of the transfection mixture was added. The cells were incubated for four hours at 37°C in 5% CO_2. After four hours, the transfection mixture was removed and replaced with 3 mL of standard growth media (Chapter 2). The cells were incubated for 72 hours at 37°C in 5% CO_2. Mock transfections were performed in the absence of DNA.

After 72 hours, total RNA was isolated from the transfected cells using the RNeasy kit (Qiagen Cat no. 74104). The cells were harvested, centrifuged for two minutes at 2,000 RPMs, the supernatant was discarded, and the cells were resuspended in 350 μL RLT buffer plus beta-mercaptoethanol (BME; 1 mL of RLT buffer add 10 μL BME). Next, 350 μL of 70% ethanol was added and the mixture was vortexed for 30 seconds. The entire mixture (including precipitate) was transferred to the spin column, centrifuged at 15,000 RPMs for 30 seconds, and the flowthrough was discarded. Subsequently, 700 μL of RW1 buffer was added to the sample, centrifuged at 15,000 RPMs for 30 seconds, and the flowthrough was discarded. The sample was washed by added 500 μL of RPE buffer, centrifuged at 15,000 RPMs for 30 seconds, and the flowthrough was discarded. A second wash was performed, but the sample was centrifuged at

15,000 RPMs for two minutes for the second wash step. The sample was then centrifuged at 15,000 RPMs for one minute to remove any remaining liquid (dry spin). The column was then transferred to a clean collection tube, 50 μL of ultrapure water was added, and the column was centrifuged at 15,000 RPMs for one minute. The flowthrough was added back to the column, the column was placed into the same collection tube, and centrifuged a second time at 15,000 RPMs for one minute to increase the RNA yield. The RNA was then quantified using the Nanodrop Lite and used in subsequent experiments.

cDNA was synthesized using 1 μg of total RNA and the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems Cat no. 4368814). A single reaction contained 3.2 μL ultrapure water, 2 μL 10X random primers, 2 μL 10X RT buffer, 1 μL RNase inhibitor, 0.8 μL dNTPs, 1 μg S16 RNA in 10 μL of water, and 1 μL MultiScribe reverse transcriptase. The reaction was incubated for 10 minutes at 25°C, then two hours at 37°C, then 85°C for 5 seconds, and finally held at 4°C. RT-PCR was performed on isolated cDNA using gene specific primers (Appendix 7). A PCR for *β-actin* served as a positive control. All PCR products were subjected to DNA sequencing to confirm specificity.

RNA was purified from three independent rat sciatic nerves at the P1, P15, and adult timepoints using the RNeasy Lipid kit (Qiagen Cat no. 74804), and quantitative RT-PCR was performed by our collaborator Dr. John Svaren as previously described (Gokey et al. 2012).

5' Rapid Amplification of cDNA Ends

First strand cDNA libraries were synthesized using total RNA isolated from S16 cells (see above) and a primer designed within exon 5 of *Sox6* (Appendix 8): 1.25 μL rnSOX6 GSP1 (2 μM), 6.3 μL (~5 μg) S16 RNA, and 7.95 μL ultrapure water were added. The reaction was incubated at 70°C for 10 minutes, chilled on ice for one minute, briefly centrifuged, and the following components were added (in order): 2.5 μL 10X PCR buffer, 2.5 μL 25mM $MgCl_2$, 1 μL 10 mM dNTP mix, and 2.5 μL 0.1 M DTT. The reaction was gently mixed, briefly centrifuged, incubated for one minute at 42°C, and 1 μL of SuperScript II RT was added. The reaction was incubated for 50 minutes at 42°C, and then incubated at 70°C for 15 minutes to terminate the reaction. The mixture was briefly centrifuged and placed at 37°C for 30 minutes.

The cDNA was then purified using S.N.A.P. column purification. First, 120 μL of binding solution (6M NaI) was added to the cDNA mixture (see above). Next, the entire mixture was transferred to a S.N.A.P. column, centrifuged at 13,000g for 20 seconds, and the flowthrough was saved until the recovery of the cDNA was ensured. The cDNA was washed by adding 400 μL of 4°C 1X wash buffer to the column, centrifuged at 13,000g for 20 seconds, and the flowthrough was discarded. The wash step was repeated three additional times. After washing, 400 μL of 4°C 70% ethanol was added, centrifuged at 13,000g for 20 seconds, and the flowthrough was discarded. The ethanol wash step was repeated one additional time (two total ethanol washes). Finally, the column was transferred to a clean recovery tube, 50 μL of 65°C ultrapure water was added, and the column was centrifuged at 13,000g for 20 seconds.

The cDNA was TdT-tailed using the 5’RACE System (Invitrogen Cat no. 18374058) by adding 6.5 μL ultrapure water, 5 μL 5X tailing buffer, 2.5 μL 2mM dCTP, and 10 μL S.N.A.P.-purified cDNA. No additional quantification was performed on the S.N.A.P.-purified cDNA prior to the TdT-tailing reaction. The reaction was incubated for three minutes at 94°C, chilled on ice for one minute, and briefly centrifuged. Next, 1 μL TDT was added, gently mixed, and incubated at 37°C for 10 minutes. Finally, the reaction was quenched by incubation at 65°C for 10 minutes.

Subsequently, two nested PCRs were performed first using a reverse primer designed within exon 4 of Sox6 (Appendix 8): 31.5 μL ultrapure water, 5 μL 10X PCR buffer, 3 μL 25 mM $MgCl_2$, 1 μL 10 mM dNTPmix, 2 μL 10 μM rnSOX6 GSP2, 2 μL Abridged anchor primer, 5 μL dC-tailed cDNA, and 0.5 μL Taq polymerase (NEB Cat no. M0273S). A second nested PCR was performed using a reverse primer designed within exon 3 of *Sox6* (Appendix 8): 33.5 μL ultrapure water, 5 μL 10X PCR buffer, 3 μL 25 mM $MgCl_2$, 1 μL 10 mM dNTPmix, 1 μL 10 μM rnSOX6 GSP3, 1 μL 10 μM AUAP primer, 5 μL of PCR product from first nested reaction, and 0.5 μL Taq polymerase (NEB Cat no. M0273S). The nested PCR products were separated on a 1% agarose gel, excised, and purified using the QIAquick gel extraction kit (Qiagen Cat no. 28704). Gel purified PCR products were TA cloned (Invitrogen Cat no. 450071): 1 μL 5' RACE PCR product, 1 μL Salt solution (1.2 M $NaCl_2$ and 0.06 M $MgCl_2$), 1 μL (10ng) pCR4 TOPO vector, and 3 μL ultrapure water. The resulting plasmids were transformed into *E. coli*, plated on kanamycin selective plates, colonies were selected and grown, and plasmid DNA was isolated (see Chapter 2 for details). 48 clones were subjected to Sanger sequencing; 44 of the resulting sequences correctly mapped to the rat *Sox6* locus.

siRNA-mediated Depletion of SOX10

Control siRNA (siControl 1, Ambion Cat no. AM4611) or Sox10 siRNA (siSox10 1, Life Technologies Cat no. s131239) were transfected into S16 cells as described using the Amaxa Nucleofection system following the manufacturer's instructions. At 48 hours post-transfection, RNA was isolated using Tri-Reagent (Ambion) and analyzed by quantitative RT-PCR as described (Gokey et al. 2012).

DNase Hypersensitivity Site Identification

DNase-Seq was performed with three biological replicates of the S16 cells at passage numbers five, eight, and 14. Each replicate contained ~20 million cells frozen into 1 mL of recovery cell culture freezing media (Invitrogen Cat no. 12648010). Cells were thawed, and DNase-Seq libraries were generated as previously described (Song and Crawford 2010) with the exception of adding a 5’ phosphate to linker 1 to increase the ligation efficiency. DNase-Seq libraries from three replicates were pooled into one lane of an Illumina Hi-Seq 2000. Raw reads were aligned to the rat rn5 genome using Bowtie (Langmead et al. 2009) and mapped allowing up to two mismatches. For the three samples, 69.2% (36,295,401), 70.8% (43,564,606), and 67.9% (39,579,719) of the reads mapped to rn5. Peaks were called using F-Seq and the default settings (Boyle et al. 2008). For the three samples: 502,787 (sample 1), 438,254 (sample 2), and 412,267 (sample 3) peaks were identified. 149,342 peaks were shared among all three samples. We used sample 2 as a representative experiment and extracted all DNase-Seq peaks with an F-Seq score of at least 0.08. This revealed a set of 31,845 peaks (7.3%) that were used to prioritize SOX10 response elements.

Results

Genome-wide Prediction of SOX10-responsive Transcriptional Regulatory Elements

SOX10 binds to a well-defined consensus sequence (5’-3’; ACACA, ACAAA, ACAAT, or ACAAG) as a monomer or as a dimer when two consensus sequences are oriented in a head-to-head fashion (Peirano and Wegner 2000; Srinivasan et al. 2012). To identify all putative SOX10 binding sites in the human genome, we wrote a Perl script to scan each human chromosome and report all occurrences of the above SOX10 consensus sequence (Chapter 3). This revealed over

33 million monomeric consensus sequences and ~549,000 dimeric consensus sequences with an intermonomeric sequence of five to 10 base pairs.

Multiple-species conservation analysis is an effective approach for predicting non-coding DNA sequences with a role in transcriptional regulation (See Chapters 2 and 3). Importantly, functionally validated SOX10 binding sites have been identified in non-coding genomic sequences that are conserved between human and chicken (Antonellis et al. 2008; Gokey et al. 2012; Hodonsky et al. 2012); Chapter 3). To prioritize the large dataset of SOX10 consensus sequences, we aligned the human, mouse, and chicken genomes and identified all genomic sequences that are five base pairs or longer (the length of the monomeric SOX10 consensus sequence) and that are identical between these three species. This revealed over two million conserved coding and non-coding genomic segments.

To develop a panel of prioritized SOX10 consensus sequences for functional studies, we used the rationale that focusing on: (**1**) conserved dimeric SOX10 consensus sequences will enrich for *bona fide* SOX10 binding sites; (**2**) non-coding sequences will deprioritize sequences that are conserved due to the function of the gene product; and (**3**) proximal promoter and intronic sequences will provide a candidate target gene for further studies. Thus, we compared the above datasets to identify dimeric SOX10 consensus sequences that are conserved between human, mouse, and chicken (including the intermonomeric sequence), reside in non-coding sequences, and map to an intron or 2.5 kb upstream or downstream of a known (RefSeq) human gene. This revealed 238 genomic sequences at 160 loci for further study. To determine the efficacy of our modified approach, we further prioritized the above 238 genomic segments by identifying the subset that map to loci with a known or predicted role in myelination (see methods for details). This revealed 57 genomic sequences at 32 loci with a conserved, dimeric SOX10 consensus sequence that resides within an intron or directly upstream of a myelin-related transcriptional unit; we named these elements SOX10 Conserved Consensus Sequences (SOX10-CCS; Appendix 6).

Seven Conserved SOX10 Consensus Sequences Display Regulatory Activity in Schwann Cells

Using our computational pipeline, we identified 57 regions that harbor conserved head-to-head SOX10-CCSs at loci with a known or predicted role in myelination. To test if these sequences are active in Schwann cells *in vitro*, a region surrounding each consensus sequence (Appendix 6) was amplified from human genomic DNA and cloned upstream of a minimal promoter directing the expression of a luciferase reporter gene in the same orientation as the direction of transcription of the gene. The regulatory activity of each genomic segment was tested in cultured rat Schwann cells (S16) (Goda et al. 1991), which express endogenous Sox10 (Hodonsky et al. 2012). The luciferase expression directed by each genomic segment was determined in luciferase activity assays relative to a control vector with no genomic insert ('Empty'), with the activity arbitrarily set to '1'. Seven of the 57 genomic segments demonstrated a greater than 2.5-fold (differs from five-fold increase used in previous chapters) increase in luciferase activity compared to the empty vector in S16 cells (Figure 4.1): SOX10-CCS-01 (3.7-fold increase; maps to *PAX7*), SOX10-CCS-13 (54-fold increase; maps to *SOX6*), SOX10-CCS-18 (82-fold increase; maps to *SOX5*), SOX10-CCS-19 (49-fold increase; maps to *SOX5*), SOX10-CCS-39 (5.9-fold increase; maps to *TCF7L2*), SOX10-CCS-43 (25-fold increase; maps to *BCAS3*), and SOX10-CCS-51 (2.6-fold increase; maps to *NFIB*). These data suggest that these seven genomic sequences (Table 4.1) are potential SOX10 response elements.

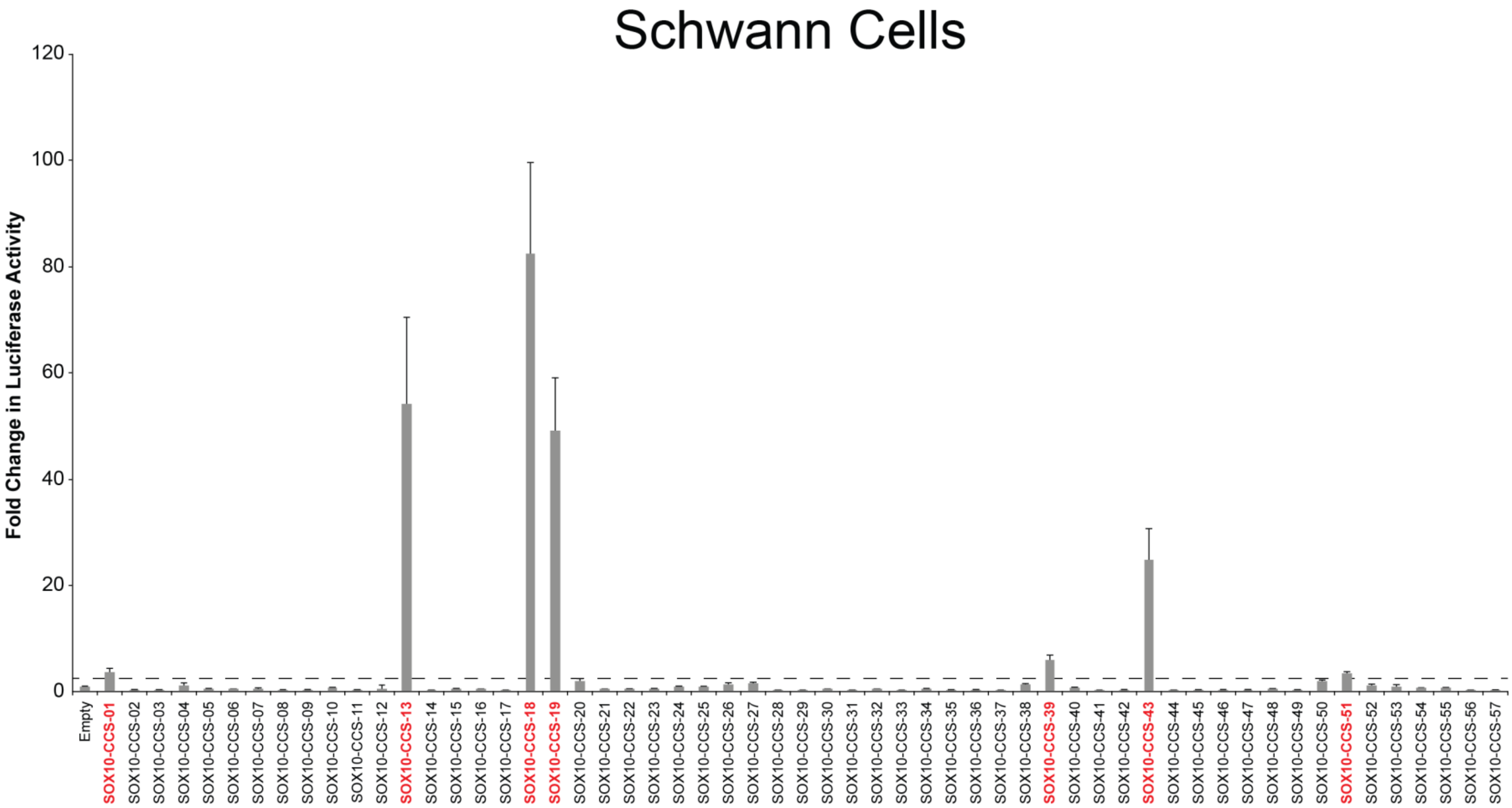


Figure 4.1 *Seven Regions Demonstrate Regulatory Activity in Schwann Cells*. Each of the 57 genomic segments containing prioritized SOX10 consensus sequences were cloned upstream of a luciferase reporter gene in the same orientation as the gene and tested for enhancer activity in cultured Schwann (S16) cells. Luciferase data are expressed relative to a control vector that does not harbor a genomic insert (‘Empty’) with activity arbitrarily set to ‘1’. Regions that demonstrated a greater than 2.5-fold increase (dashed black line) in luciferase activity are indicated in red text, and error bars indicate standard deviations.

Table 4.1 *Seven Genomic Segments with Regulatory Activity in Schwann Cells.*

Element ID	**Locus**	**UCSC Coordinates**[1]	**SOX10 Consensus Sequence**[2]
SOX10-CCS-01	*PAX7*	chr1:18854774-18854793	**ACAAA**CTCATTAAA**CTTGT**
SOX10-CCS-13	*SOX6*	chr11:16334769-16334784	**ACAAT**CAAGC**ATTGT**
SOX10-CCS-18	*SOX5*	chr12:24059368-24059383	**ACAAA**AATGT**ATTGT**
SOX10-CCS-19	*SOX5*	chr12:24059689-24059706	**ACACA**GAACATT**ATTGT**
SOX10-CCS-39	*TCF7L2*	chr10:114895622-114895642	**ACAAT**CCCCAAGATT**TTTGT**
SOX10-CCS-43	*BCAS3*	chr17:56684299-56684319	**ACACA**TTAATAACGT**TTTGT**
SOX10-CCS-51	*NFIB*	chr9:14299587-14299605	**ACAAT**CTGTTCTT**TGTGT**

[1] Coordinates refer to the March 2006 UCSC Genome Browser Human assembly (hg18).

[2] SOX10 consensus sequences are indicated in red letters and bold text.

The SOX10 Consensus Sequence is Required for the Orientation-Independent Activity of Three Regulatory Elements at SOX5, SOX6, and NFIB

To determine if the regulatory activity of the seven genomic segments is dependent on the orientation of the DNA sequence, we retested the activity of each segment in both the 'forward' and 'reverse' orientation relative to the minimal promoter within our reporter gene construct in S16 cells. This revealed three genomic segments that enact a greater than 2.5-fold increase in luciferase activity in both orientations (Figure 4.2): SOX10-CCS-13 (72-fold forward and 9-fold reverse), SOX10-CCS-19 (70-fold forward and 33-fold reverse), and SOX10-CCS-51 (4-fold forward and 9-fold reverse). To assess the specificity of these results to Schwann cells, we tested each of the seven genomic segments in both orientations in cultured mouse motor neurons (MN-1 cells) (Salazar-Grueso et al. 1991), which do not express endogenous SOX10 (Hodonsky et al. 2012). None of the genomic segments enact a greater than 2.5-fold increase in luciferase activity in both orientations in MN-1 cells suggesting that our data in S16 cells is Schwann-cell specific; however, three genomic segments displayed low levels of activity in only the forward orientation in MN-1 cells (Figure 4.2): SOX10-CCS-39 (5.5-fold), SOX10-CCS-43 (6.7-fold), and SOX10-CCS-51 (4-fold).

To test the necessity of the conserved SOX10 consensus sequence for the observed activity associated with the seven genomic segments described above, we deleted the dimeric SOX10 consensus sequence along with the intervening sequence in each construct (ΔSOX10) and compared the activity to the wildtype genomic segment using the more active orientation. This revealed three genomic segments that display at least a 50% reduction in activity upon deleting the SOX10 consensus sequence (Figure 4.3): SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51. Combined, our data are consistent with these three genomic segments—at the *SOX6*,

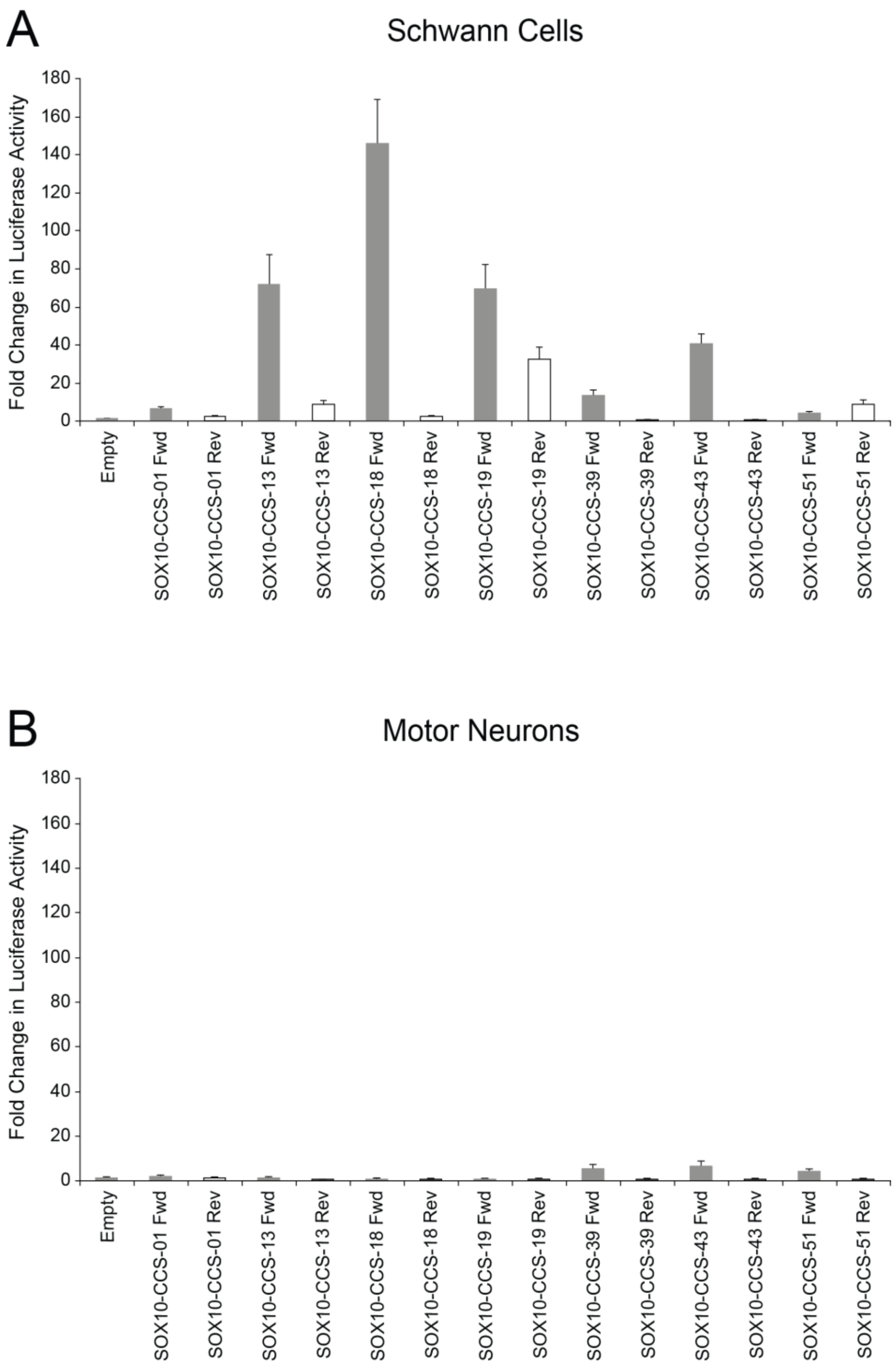


Figure 4.2 *Three Genomic Segments Demonstrate Orientation-independent Luciferase Activity in Schwann Cells but Not Motor Neurons*. The seven active regions from Figure 4.1 were tested in forward (grey bars) and reverse (white bars) orientation in rat Schwann cells (S16; Top) or motor neurons (MN-1; Bottom). Luciferase data are expressed relative to a control vector without a genomic segment ('Empty') with activity arbitrarily set to '1' and error bars indicate standard deviations.

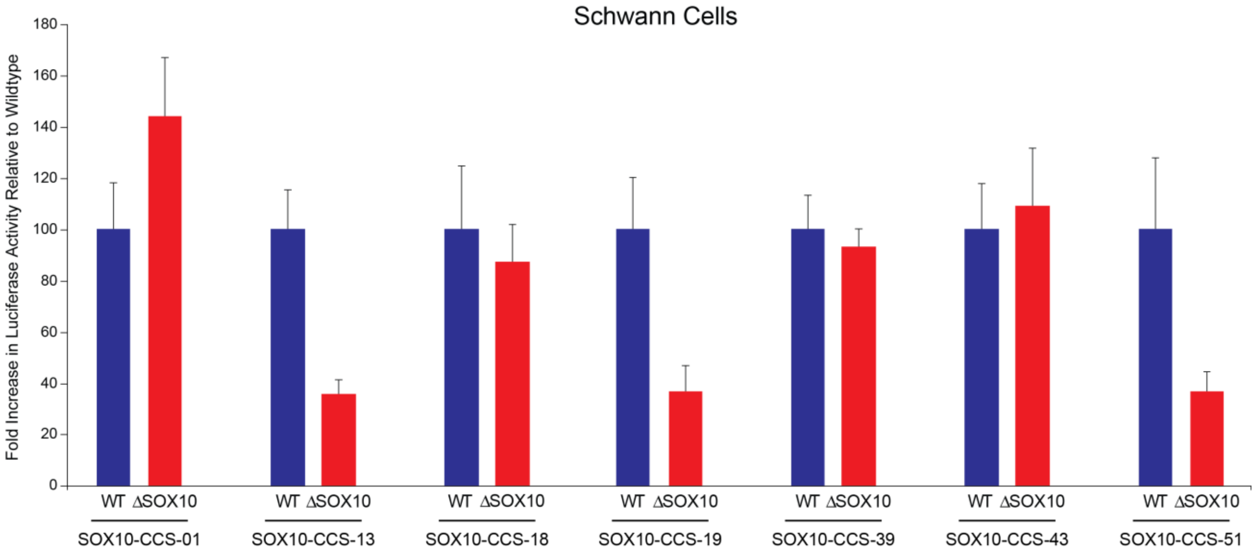


Figure 4.3 *Three Genomic Segments Require the SOX10 Consensus Sequence for Luciferase Activity in Schwann Cells*. Luciferase reporter gene constructs containing either the wildtype sequence (WT) or the sequence lacking the SOX10 consensus sequence(s) (ΔSOX10) were transfected into S16 cells and assessed by luciferase assays. The luciferase activity associated with each ΔSOX10 construct (red bar) is expressed relative to the respective wildtype construct (blue bar), with activity arbitrarily set to ‘100’ and error bars indicate standard deviations.

SOX5, and *NFIB* loci, respectively—representing Schwann cell enhancers that harbor required SOX10 consensus sequences.

SOX10 is Necessary and Sufficient for the Activity of the Three Regulatory Elements at SOX5, SOX6, and NFIB

To test if SOX10 is sufficient for the activity of SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51, we co-transfected each reporter gene construct with or without a construct to express wildtype SOX10 (Inoue et al. 2004) in MN-1 cells. Subsequently, we compared the activity of each region in the presence or absence of SOX10 expression. There was a ~1,000-fold increase in the activity of SOX10-CCS-13 and a ~200-fold increase in the activity of SOX10-CCS-19 and SOX10-CCS-51 in the presence of SOX10 (Figure 4.4).

To determine if SOX10 is necessary for the activity of SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51 in Schwann cells, S16 cells were transfected with each SOX10-CCS luciferase reporter gene construct along with a construct to express a dominant-negative mutant of SOX10 (E189X), which interferes with the function of endogenous SOX10 (Inoue et al. 2004). Importantly, E189X SOX10 has been shown to specifically reduce the activity of genomic segments harboring SOX10 binding sites in luciferase assays (Brewer et al. 2014). We observed a greater than 85% reduction in the activity of all three genomic segments upon co-transfection with E189X SOX10 (Figure 4.4). Combined, our data indicate that SOX10 is both necessary and sufficient for the *in vitro* enhancer activity of SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51.

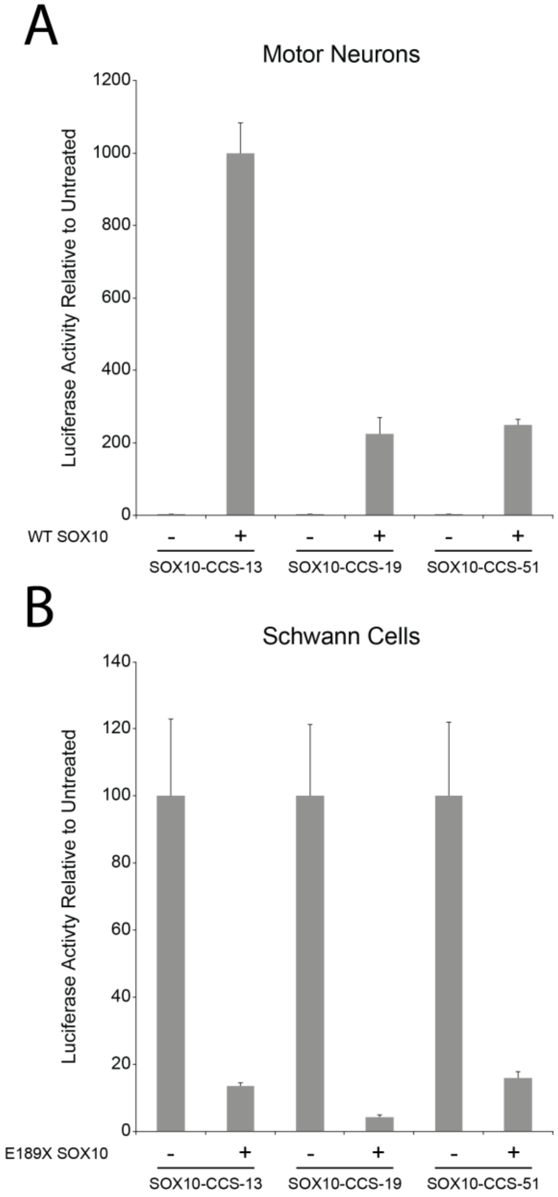


Figure 4.4 *SOX10 is Required for the Regulatory Activity of SOX10-CCS-13, SOX10-CCS-19 and SOX10-CCS-51*. (A) Luciferase reporter gene constructs harboring SOX10-CCS-13, SOX10-CCS-19 or SOX10-CCS-51 were transfected into mouse motor neurons (MN-1) with or without a construct to express wildtype SOX10. The luciferase activity associated with each construct in the presence of SOX10 is expressed relative to that of the construct in the absence of SOX10 with activity set arbitrarily to ‘1’. (B) Luciferase reporter gene constructs harboring SOX10-CCS-13, SOX10-CCS-19 or SOX10-CCS-51 were transfected into rat Schwann (S16) cells with or without a construct to express dominant-negative (E189X) SOX10. The luciferase activity associated with each construct in the presence of E189X SOX10 is expressed relative to that of the construct in the absence of E189X SOX10 with activity arbitrarily set to ‘100’. Error bars indicate standard deviations in both panels.

SOX10-CCS-13 is a Previously Unreported, Alternative Promoter at Sox6

Examination of SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51 on the UCSC Genome Browser revealed that the SOX10 consensus sequence within SOX10-CCS-13 and SOX10-CCS-19 are also conserved between human and zebrafish (Figure 4.5) further suggesting the these sequences may be important in jawed vertebrates. Additionally, analysis of published SOX10 ChIP-Seq data generated from rat sciatic nerve nuclei (Srinivasan et al. 2012) and our own DNase hypersensitivity site (HSS) data generated from S16 cell nuclei (Chapter 3) revealed evidence of SOX10 occupancy and open chromatin at each region with the highest peaks for both datasets at SOX10-CCS-13 relative to the two other active SOX10-CCS regions (Figure 4.5). Thus, to validate the efficacy of our approach, we pursued additional analyses of SOX10-CCS-13, which resides at the *SOX6* locus.

Closer scrutiny of the *SOX6* locus on the UCSC Genome Browser revealed seven unique *SOX6* mRNA isoforms in human, mouse, or rat, distinguished by alternative, non-coding first exons. Interestingly, SOX10-CCS-13 maps directly upstream of the 3'-most alternative first exon, which we named *SOX6* exon 1G (Figure 4.6). We therefore hypothesized that SOX10-CCS-13 acts as an alternative promoter at *SOX6*. To test this, we performed 5'-rapid amplification of cDNA ends (5'-RACE). Briefly, a cDNA library was generated using RNA isolated from cultured rat Schwann (S16) cells and a reverse primer in exon 5 of rat *Sox6*. Subsequently, nested PCR was performed using reverse primers in exon 4 and then exon 3 of rat *Sox6* (see Appendix 8 for primer sequences). The PCR products were cloned, sequenced, and aligned to the rat *Sox6* locus. These analyses revealed the presence of five unique *Sox6* transcription start sites in cultured Schwann cells with 14 of the 44 *Sox6*-specific sequences mapping directly downstream of SOX10-CCS-13 (Figure 4.6). Analysis of RNA-Seq data generated in S16 cells (Chapter 3)

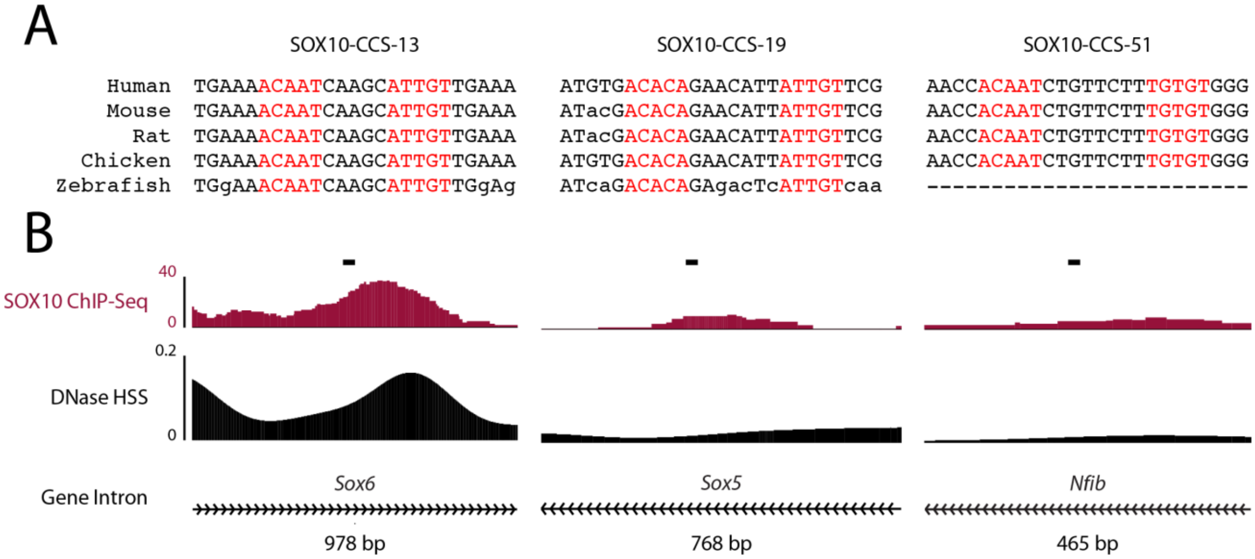


Figure 4.5 *SOX6, SOX5, and NFIB Harbor Intronic SOX10 Response Elements*. (A) Multiple-species sequence analysis was performed using human, mouse, rat, chicken, and zebrafish genomic sequences surrounding the SOX10 consensus sequence (red text) within SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51. Uppercase text indicates a nucleotide identical to the human sequence while lowercase text indicates a non-conserved nucleotide. Dashes indicate the absence of a detectable orthologous sequence in zebrafish. (B) SOX10 ChIP-seq and DNase-seq peaks are shown at SOX10-CCS-13, SOX10-CCS-19, and SOX10-CCS-51. Y-axes represent normalized sequence read depth (SOX10 ChIP-seq) and F-Seq score (DNase-seq). Black bars indicate the position of the SOX10 consensus sequence in each genomic segment (not to scale). The associated RefSeq intron and base pair size of each genomic segment is indicated along the bottom. Arrowheads indicate the direction of transcription.

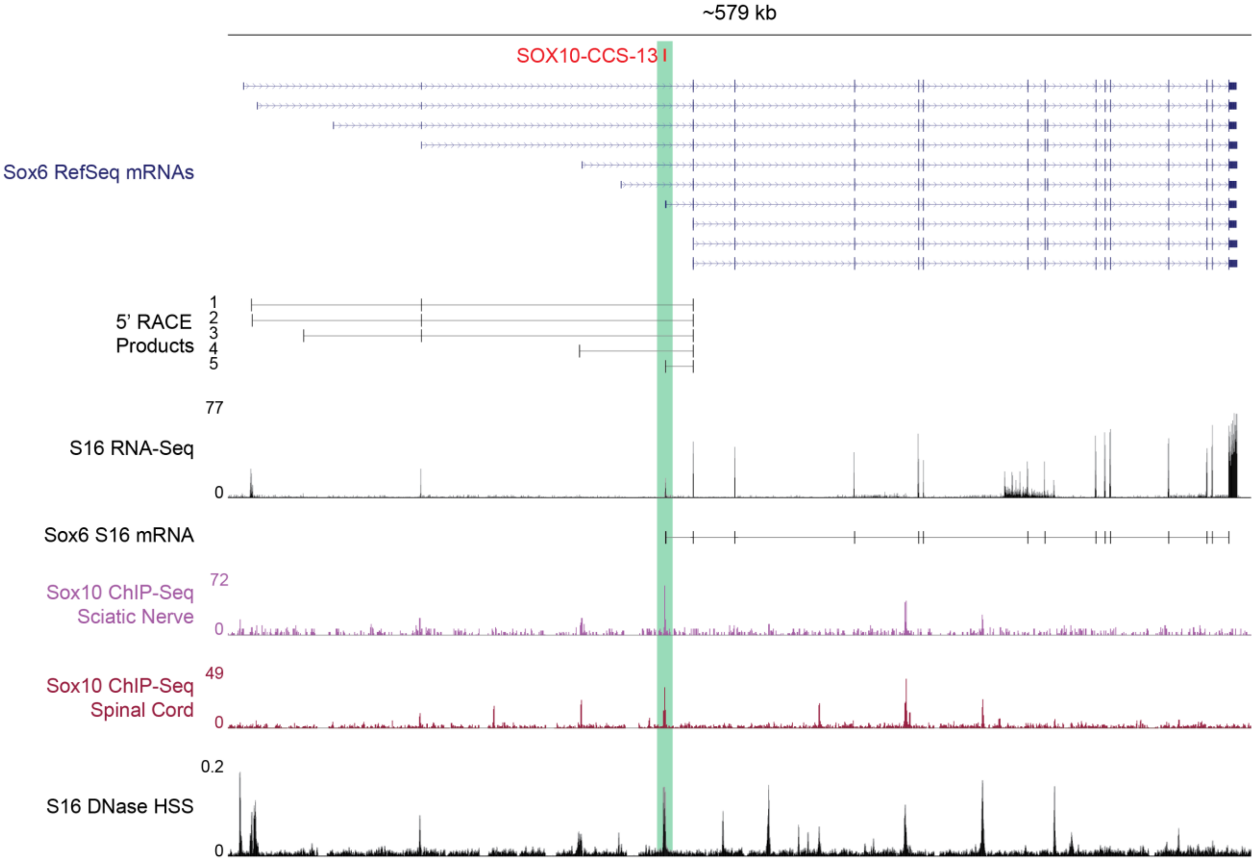


Figure 4.6 *SOX10-CCS-13 is an Alternative Promoter at the Sox6 Locus*. The ~579 kb rat *Sox6* locus is shown on the UCSC Rat Genome Browser. SOX10-CCS-13 is indicated in red along with the seven human, mouse, and rat *SOX6* RefSeq mRNAs (blue). *Sox6*-specific 5' RACE was performed on RNA from S16 cells, and the five distinct *Sox6* sequences were mapped to the rat genome. Please note that SOX10-CCS-13 maps to both the 5' end of the seventh *Sox6* mRNA and the fifth unique 5' RACE-generated sequence. RNA-Seq data from S16 cells were mapped to *Sox6* (the y-axis indicates sequence read depth) as was a PCR-amplified, full-length mRNA that contains *Sox6* exon 1G. Genome-wide regulatory marks were also mapped to *Sox6* with the Y-axes indicating normalized sequence read depths (both SOX10 ChIP-seq data sets) and F-Seq scores (DNase-seq).

also revealed reads that map to *Sox6* exon 1G, with split reads into downstream exons, but no split reads into upstream exons. Additionally, we were able to amplify and sequence-verify a full length *Sox6* mRNA that originates at exon 1G in S16 cells and contains all the expected exons with the exception of one alternatively spliced exon (Figure 4.6). Combined, our data indicate that SOX10-CCS-13 represents an internal, alternative promoter at *Sox6*.

SOX10 is Necessary and Sufficient for the Expression of Sox6 Transcripts Harboring Exon 1G

To determine if SOX10 is sufficient to direct the expression of *Sox6* transcripts, we performed RT-PCR using primers designed in *Sox6* exon 1G and exon 2 in regions conserved between rat and mouse (see Appendix 7 for primer sequences). While these primers amplify *Sox6* transcripts containing exon 1G from a cDNA library generated from S16 RNA, we were not able to amplify these transcripts from a cDNA library generated from cultured mouse motor neurons (MN-1 cells), which do not express endogenous SOX10 (Figure 4.7). However, when MN-1 cells were transfected with a construct to express wildtype SOX10, *Sox6* transcripts containing exon 1G were detected and verified by DNA sequence analysis. Mock transfection or transfection with a construct to express a non-functional mutant version of SOX10 (E189X) (Inoue et al. 2004) did not allow amplification of *Sox6* transcripts containing exon 1G (Figure 4.7). Thus, SOX10 is sufficient to activate the expression of *Sox6* transcripts harboring exon 1G in MN-1 cells.

To determine if SOX10 is necessary for the expression of *Sox6* transcripts containing exon 1G in Schwann cells, we treated S16 cells with a previously validated siRNA against *Sox10* (Gokey et al. 2012; Anido et al. 2015) and tested for an effect on total *Sox6* mRNA levels and for an effect on the level of transcripts containing exon 1G. This analysis revealed a ~70% decrease in both

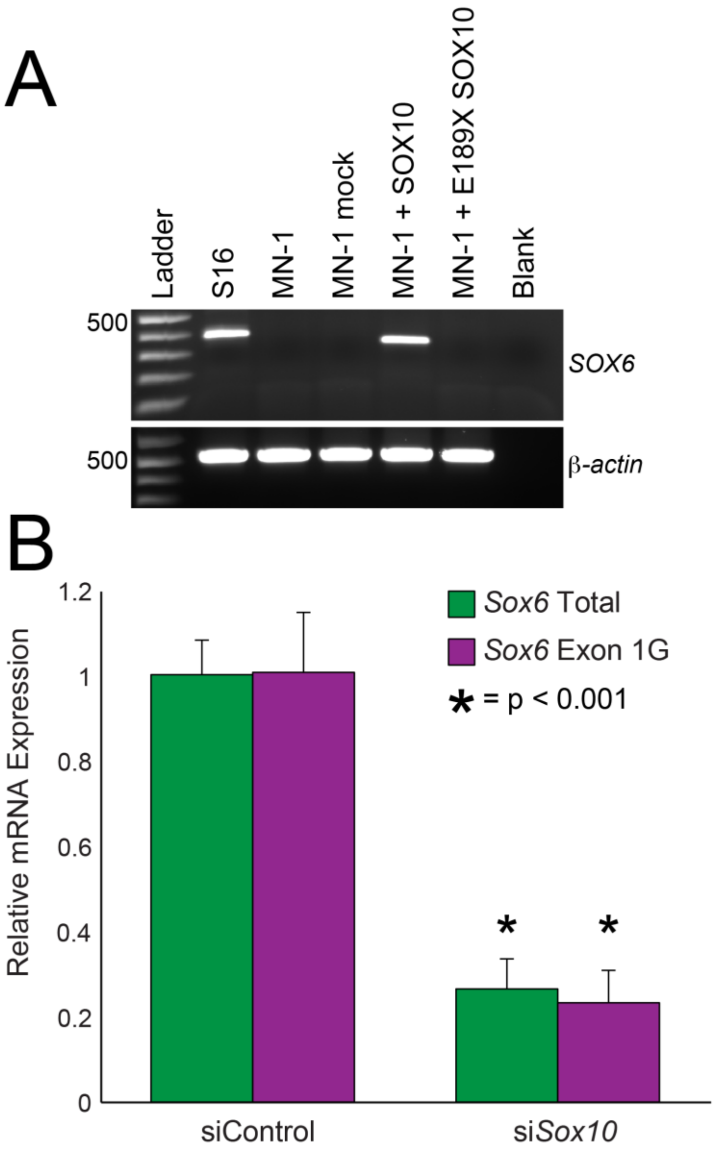


Figure 4.7 *SOX10 is Necessary and Sufficient for SOX6 Expression*. (A) RT-PCR was performed to detect the expression of *Sox6* transcripts harboring exon 1G using cDNA isolated from S16 cells, MN-1 cells, or MN-1 cells transfected with no expression construct (mock), a construct to express wildtype SOX10, or a construct to express dominant-negative (E189X) SOX10. Base pair ladders are indicated on the left. RT-PCR for *β-actin* and samples including no cDNA ('Blank') were employed as positive and negative controls respectively. Please note that while the same primers were used for each reaction, the rat (S16) PCR product was 402 base pairs and the mouse (MN-1) PCR product was 349 base pairs; the rat genome harbors a 53 base pair rat-specific insertion, which we confirmed via DNA sequence analysis. (B) Rat Schwann (S16) cells were treated with a control siRNA (left side) or a siRNA targeted against *Sox10* (right side). Quantitative RT-PCR was used to measure expression levels of total *Sox6* (green bars) or *Sox6* exon 1G-containing (purple bars) transcripts. Error bars indicate standard deviations. Data for panel B was collected by our collaborator Dr. John Svaren at the University of Wisconsin.

total *Sox6* expression and in the expression of transcripts containing exon 1G (Figure 4.7), consistent with SOX10 regulating the promoter activity of SOX10-CCS-13 in Schwann cells. Combined, our data indicate that SOX10 is both necessary and sufficient for the expression of *Sox6* mRNA isoform 7 (Figure 4.6) in our *in vitro* cell culture model systems.

SOX10 Activates the Expression of Genes that Inhibit Myelination

Our stringent computational and functional analyses rapidly identified a previously unreported SOX10-responsive promoter at the *SOX6* locus. Importantly, this finding was facilitated by the knowledge of a well-defined SOX10 consensus sequence and reports that SOX10 binding sites can be conserved among vertebrate species including human and chicken (Antonellis et al. 2008; Gokey et al. 2012; Hodonsky et al. 2012); Chapter 3). While our computational analysis was successful, one limitation of our first approach was only identifying conserved SOX10 consensus sites near genes previously implicated in myelination. Utilizing a less-biased approach may uncover novel functions of SOX10 in Schwann cells. We therefore removed the requirement of the SOX10 consensus site residing near genes involved in myelination and converted our conserved non-coding SOX10 monomers using the liftOver utility from the UCSC genome browser (Kent et al. 2002) from the human (hg18) genome to the rat (rn5) genome (61,133/67,482 were successfully converted) and overlayed them with two datasets: (**1**) previously generated SOX10 ChIP-Seq from rat Schwann cell nuclei *in vivo* (Srinivasan et al. 2012); (**2**) DNase-Seq on cultured rat Schwann (S16) cell nuclei (Chapter 3). Intersecting these three datasets revealed 214 rat genomic segments that harbor conserved SOX10 consensus sequences and map to both SOX10 ChIP-Seq and S16 DNase-Seq peaks. To determine if this approach enriches for loci important for myelination, we identified the rat RefSeq gene closest to each region—the 214 genomic segments map to 191 known genes—and performed a gene ontology (Ashburner et al. 2000) search using the overrepresentation test for biological processes. This analysis revealed 183 biological processes with a p-value less than 0.05 and 37 biological processes that showed a greater than five-fold enrichment compared to the human genome. Ten of the identified biological processes directly relate to myelinating glia, which all resided in the top 14 enriched terms (Table 4.2). Therefore, this combined strategy provided a highly confident set of 214 SOX10-response elements at 191 loci for future functional studies aimed at better understanding the biological process of myelination.

Interestingly, three of the 10 gene ontology biological processes that relate to myelination specifically relate to negative regulation of gliogenesis, which was due to the presence of six genes: *NOTCH1*, *HMGA2*, *HES1*, *MYCN*, *ID4*, and *ID2* (Table 4.2). Computational analyses revealed eight SOX10 consensus sequences at these six loci (Table 4.3). To determine if *NOTCH1*, *HMGA2*, *HES1*, *MYCN*, *ID4*, and *ID2* harbor *bona fide* SOX10 response elements, we amplified genomic regions surrounding the SOX10 consensus sequences using rat genomic DNA and cloned each genomic segment in both the forward and reverse orientation upstream of a minimal promoter directing luciferase expression. The regulatory activity of each genomic segment was tested in S16 cells as described above. This revealed five genomic segments that directed reporter gene activity at least 2.5-fold higher than the empty control vector in both orientations: *Notch1-R1* (4.7-fold forward and 56-fold reverse), *Hmga2-R2* (93.7-fold forward and 87-fold reverse), *Hes1-R1* (22-fold forward and 7.6-fold reverse), *Mycn-R1* (28-fold forward and 16-fold reverse) and *Id2-R1* (8.9-fold forward and 4.1-fold reverse) (Figure 4.8). In the nomenclature for these regions, R refers to the number of regions identified at each locus, and does not reflect the orientation. Regions *Notch1-R2* (7.6-fold) and *Id4-R1* (8.6-fold) directed

Table 4.2 *Gene Ontology Annotations of Loci Harboring Conserved SOX10 Consensus Sites.*

GO Biological Process	**Human**[1]	**Our List**[2]	**Expected**[3]	**P-value**	**Loci**[4]
Negative regulation of oligodendrocyte differentiation	12	4	0.1	2.74E-02	*HES1, ID2, NOTCH1, ID4*
Negative regulation of glial cell differentiation	25	6	0.21	6.34E-04	*HES1, ID2, NOTCH1, HMGA2, MYCN, ID4*
Regulation of astrocyte differentiation	25	6	0.21	6.34E-04	*HES1, ID2, NOTCH1, HMGA2, MYCN, ID4*
Regulation of oligodendrocyte differentiation	28	5	0.23	3.29E-02	*HES1, TCF7L2, ID2, NOTCH1, ID4*
Negative regulation of gliogenesis	34	6	0.28	3.77E-03	*HES1, ID2, NOTCH1, HMGA2, MYCN, ID4*
Oligodendrocyte differentiation	60	10	0.5	9.43E-07	*SOX6, NTRK2, PTPRZ1, SOX8, SOX10, TCF7L2, ID2, NOTCH1, SOX5, ID4*
Regulation of gliogenesis	74	10	0.61	6.96E-06	*PTPRZ1, SOX8, HES1, SOX10, TCF7L2, ID2, NOTCH1, HMGA2, MYCN, ID4*
Regulation of glial cell differentiation	54	7	0.45	3.23E-03	*HES1, TCF7L2, ID2, NOTCH1, HMGA2, MYCN, ID4*
Glial cell differentiation	135	13	1.12	1.22E-06	*SOX6, PTPRZ1, NTRK2, SOX8, HES1, SOX10, TCF7L2, ID2, NOTCH1, PPAP2B, SOX5, ID4, PARD3*
Gliogenesis	168	13	1.39	1.66E-05	*SOX6, PTPRZ1, NTRK2, SOX8, HES1, SOX10, TCF7L2, ID2, NOTCH1, PPAP2B, SOX5, ID4, PARD3*

[1] Number of genes from the human genome (hg19) that correspond to the biological process (column 1).

[2] Number of genes from our list of 191 genes harboring conserved SOX10 consensus sites overlapping SOX10 ChIP-Seq(Srinivasan et al. 2012) and S16 DNase-Seq peaks.

[3] Number of genes expected by chance for each biological process (column 1).

[4] The names of the genes identified from our list (column 4) that and involved in each biological process (column 1).

Table 4.3 *Eight Genomic Segments Within Loci that Inhibit Glial Cell Differentiation.*

Element ID	**UCSC Coordinates**[1]	**SOX10 Consensus Sequence**[2]
Notch1-R1	chr3:9307946-9307962	**ACAAT**GGGGCC**TCTGT**
Notch1-R2	chr3:9308648-9308662	**ACAAT**CGGC**TTTGT**
Hmga2-R1	chr7:65390829-65390834	CTTAG**ACACA**GCACTT
Hmga2-R2	chr7:65428331-65428349	**ACACA**GGCCCCTC**TTTGT**
Hes1-R1	chr11:77415711-77415729	**TGTGT**GAGCGCCA**TGTGT**
Mycn-R1	chr6:51230285-51230310	**ACAAT**GGCCTC**TTTCT**ACAG**ACAAT**
Id4-R1	chr17:18701782-18701812	**ACAAA**AACAGCAGTAAATGGAGGCC**TTTGT**
Id2-R1	chr6:53091198-53091214	**ACAAG**AAACAC**ATTGT**

[1] Coordinates refer to the March 2012 UCSC Genome Browser Rat assembly (rn5).

[2] SOX10 consensus sequences are indicated in red letters and bold text.

reporter gene activity at least 2.5-fold higher than the empty control vector only in the forward orientation (Figure 4.8). *Hmga2-R1* was not active in either orientation and was excluded from further analysis. Thus, we identified seven genomic sequences at six loci (*NOTCH1*, *HMGA2*, *HES1*, *MYCN*, *ID4*, and *ID2*) that display regulatory activity in Schwann cells.

To determine if the identified SOX10 consensus sequences are important for the regulatory activity of the seven active regions described above (Figure 4.8), we deleted the SOX10 consensus sequence from each construct (termed 'ΔSOX10') and compared the activity to the wildtype construct using the more active orientation. *Notch1-R1*, *Notch1-R2*, *Hmga2-R2*, *Hes1-R1*, and *Id2-R1* contain dimeric SOX10 consensus sequences, which were deleted along with the intermonomeric sequence. *Mycn-R1* contains a monomeric consensus sequence (ΔSOX10-1) and a dimeric consensus sequence (ΔSOX10-2), which were independently deleted. *Id4-R1* contains a dimeric consensus sequence with a 20 base pairs intervening sequence. Since this intervening sequence is longer than those previously observed for validated dimeric SOX10 binding sites (Peirano and Wegner 2000; Jones et al. 2011a; Gokey et al. 2012; Hodonsky et al. 2012; Brewer et al. 2014) we studied each monomer independently. Specifically, we deleted the dimeric consensus sequence along with intermonomeric sequence (ΔSOX10-1), the first monomer only (ΔSOX10-2), and the second monomer only (ΔSOX10-3). Deleting the SOX10 consensus sequences in regions *Notch1-R1*, *Hmga2-R2*, *Mycn-R1* (ΔSOX10-2), *Id4-R1* (ΔSOX10-1 and ΔSOX10-3), and *Id2-R1* reduced luciferase activity in S16 cells by at least 50% (Figure 4.8), indicating that the SOX10 consensus sequences in these five regions are required for the full regulatory activity of the genomic segment. In contrast deleting the SOX10 consensus sequences in *Notch1-R2* and *Hes1-R1* did not reduce the enhancer activity associated with these genomic segments (Figure 4.8).

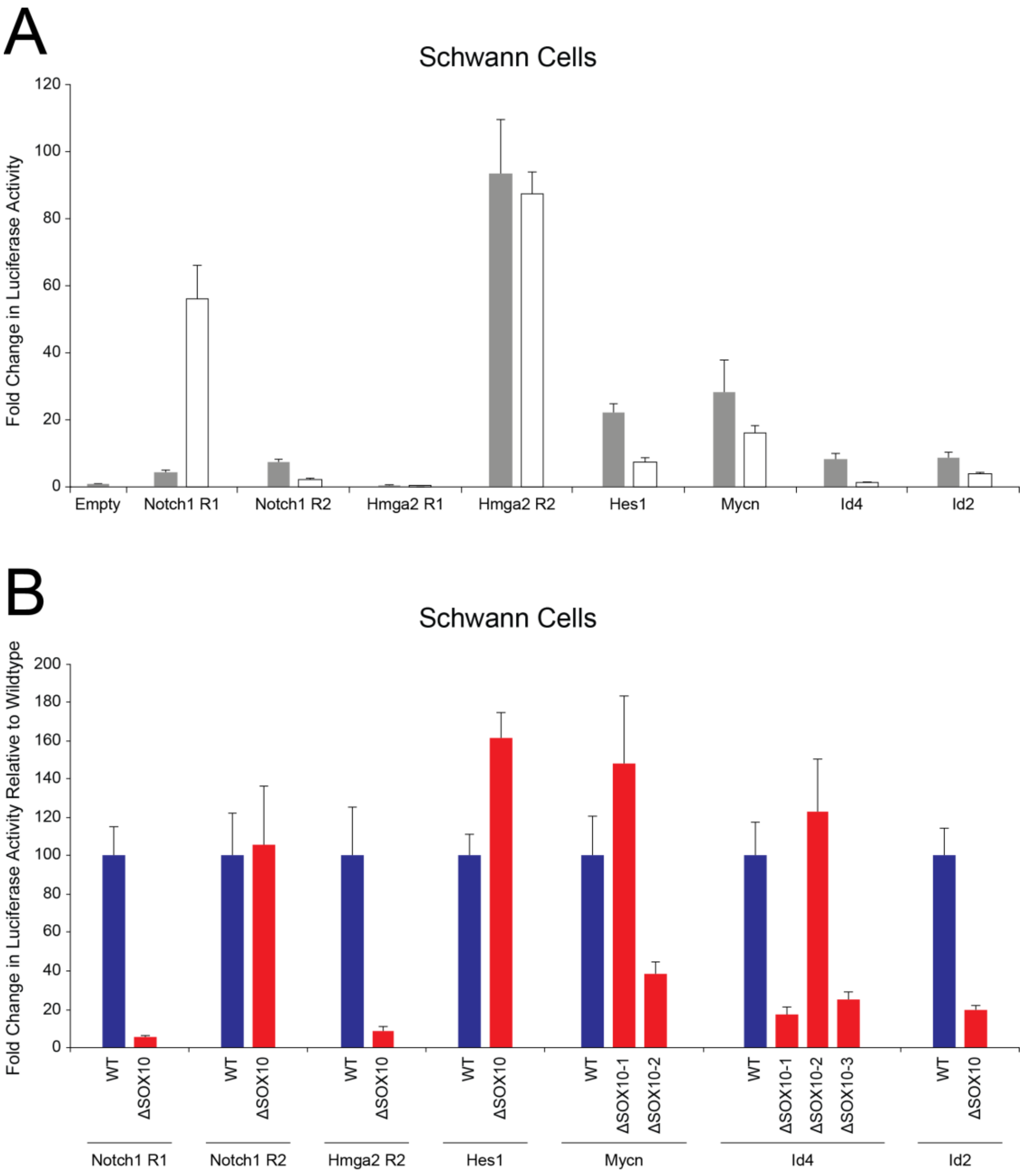


Figure 4.8 *SOX10 Consensus Sequence is Necessary for the Luciferase Activity of Five of the Seven Regions Active in Schwann Cells*. (A) Eight genomics segments at the rat *Notch1*, *Hmga2*, *Hes1*, *Mycn*, *Id4*, and *Id2* loci were cloned upstream of a luciferase reporter gene in both the forward (grey bars) and reverse (white bars) orientations and tested for luciferase activity in rat Schwann (S16) cells. Luciferase data are expressed relative to a control vector with no genomic insert (‘Empty’). Error bars represent standard deviations. (B) The conserved SOX10 consensus sequence(s) were deleted in each of the seven regions that were active in Figure 4.8A (see text for details). Luciferase reporter gene constructs containing the wildtype sequence (WT; blue bars) or the sequence lacking the SOX10 consensus sequence(s) (ΔSOX10; red bars) were transfected into S16 cells and luciferase assays performed. Luciferase activities are expressed relative to the wildtype expression constructs, and error bars represent standard deviations.

We next wanted to determine if SOX10 positively regulates the expression of *Notch1*, *Hmga2*, *Hes1*, *Mycn*, *Id4*, *Id2*, and *Sox5* in cultured rat Schwann (S16) cells, and again utilized the *Sox10* siRNA that has been shown to efficiently down-regulate *Sox10* expression (Gokey et al. 2012; Anido et al. 2015). After isolation of mRNA at 24 hours post-transfection, qRT-PCR shows that *Sox10* depletion in S16 cells results in the reduced expression of all of the above genes except for *Hmga2* (Figure 4.9). Similar findings—including the absence of reduced *Hmga2* expression—were observed upon repressing SOX10 function *in vivo* using primary Schwann cells; however, this system is prone to variability due to heterogeneous cell populations (Figure 4.10).

To directly test if *Notch1*, *Hmga2*, *Hes1*, *Mycn*, *Id4*, *Id2*, *Sox5*, and *Sox6* are developmentally regulated during myelination *in vivo*, we examined mRNA levels at three timepoints in rat sciatic nerve (n=3 at each timepoint). P1 corresponds to the onset of myelination, P15 is a peak timepoint of myelination in the PNS, and adult sciatic nerve is a timepoint where active myelination has subsided. Interestingly, the expression of all seven genes tested (*Notch1*, *Hmga2*, *Hes1*, *Mycn*, *Id4*, *Id2*, *Sox5*, and *Sox6*) are highest at P1 and then decreased at P15 and adult, consistent with a role in repressing precocious myelination (Figure 4.9). In summary, our preliminary data suggests a potentially novel role for SOX10 in positively regulating genes important for inhibiting glial cell differentiation.

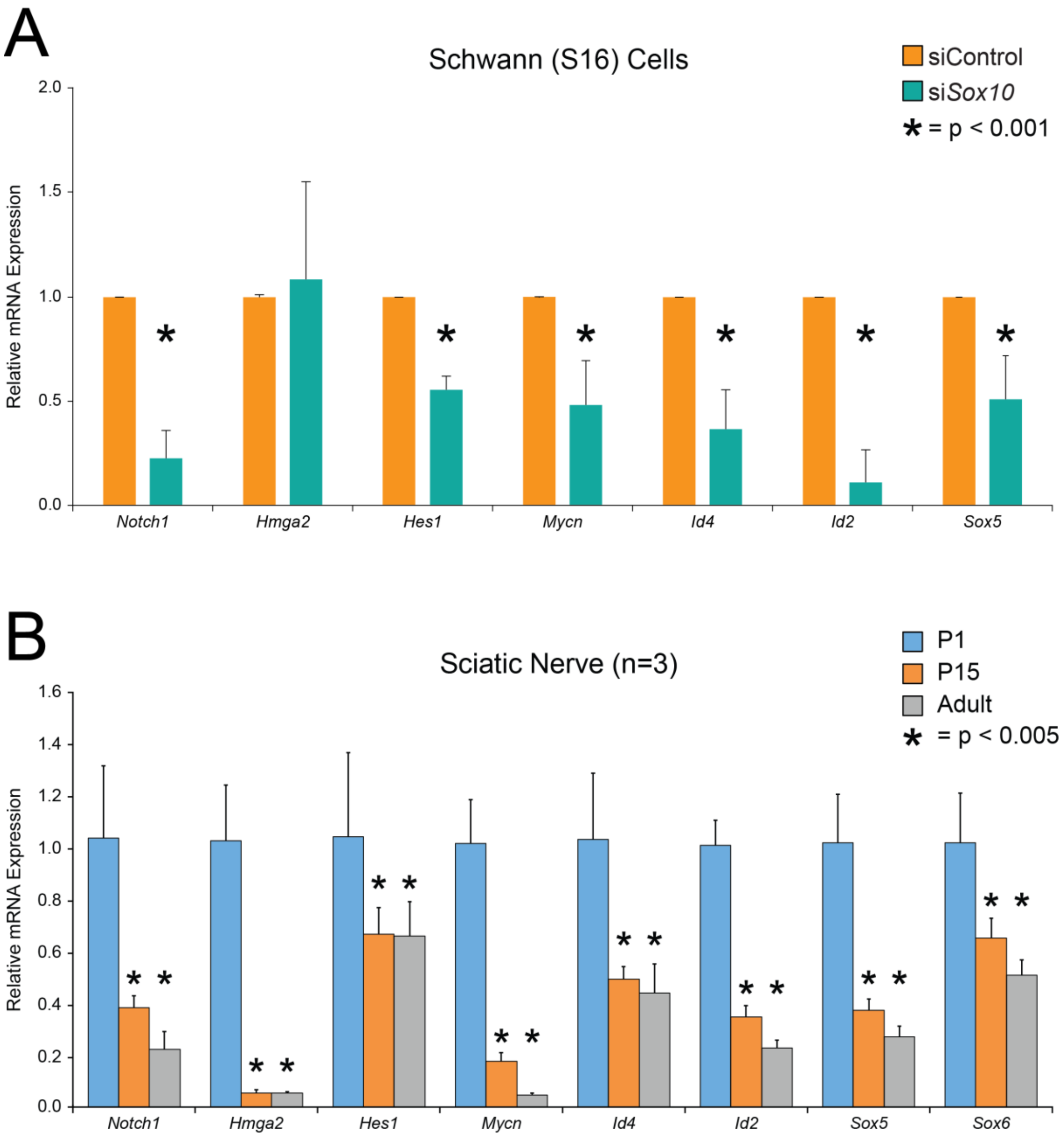


Figure 4.9 *SOX10 Regulates the Expression of Genes that Inhibitor Glial Cell Differentiation.* (A) Rat Schwann (S16) cells were treated with a control siRNA (orange bars) or a siRNA targeted against *Sox10* (green bars). Quantitative RT-PCR was used to measure expression levels of each indicated gene. Asterisks indicate a p-value less than 0.001, and error bars indicate standard deviations. (B) RNA was purified from three independent rat sciatic nerves at the P1 (blue bars), P15 (orange bars), and adult (grey bars) timepoints. Quantitative RT-PCR was used to measure expression levels of each indicated gene with values expressed relative to expression levels at P1. Asterisks indicate a p-value less than 0.005, and error bars indicate standard deviations. Data for this figure was collected by our collaborator Dr. John Svaren at the University of Wisconsin.

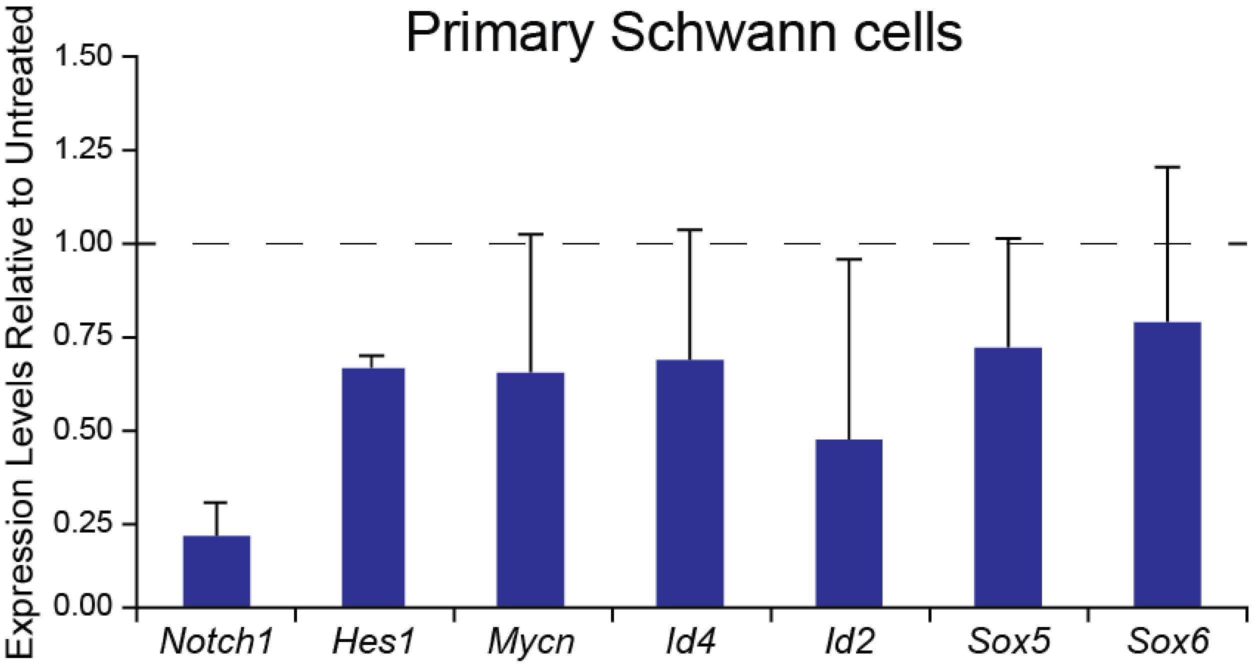


Figure 4.10 *SOX10 Regulates Genes that Inhibit Glial Cell Differentiation In Vivo*. Primary Schwann cells were extracted and grown from three independent rat adult sciatic nerves. Cells were treated with a control siRNA or a siRNA targeted against *Sox10* as in Figure 4.9B. Quantitative RT-PCR was used to measure expression levels of each indicated gene. The effect on expression of each gene (indicated across the bottom) is expressed relative to the control siRNA, and error bars indicate standard deviations. Please note that, consistent with the *in vitro* data, *Hmga2* did not show a decrease in expression levels *in vivo* (data not shown). Data was collected by our collaborator Dr. John Svaren at the University of Wisconsin.

Discussion

In the previous chapters, we developed a versatile computational pipeline to predict functional, non-coding regions of the genome important for peripheral nerve function in both a transcription factor blind (Chapter 2) and transcription factor centric (SOX10; Chapter 3) approach. In this chapter, we again modified the computational pipeline to identify conserved SOX10 binding sites that also resided within genes known to be involved in myelination in glia cells. Using this alternate filter, we were able to identify a novel SOX10-responsive promoter element at *Sox6*. This isoform appears to be Schwann cell specific, although the precise role of this isoform is unclear. Further studies, such as deletion of this novel first exon at *Sox6*, may elucidate a Schwann cell specific function.

Of note, we only required the regions to exceed a 2.5-fold threshold, compared to the five-fold increase used in previous chapters. The reduction was chosen specifically for this chapter to reduce the chance for false-negatives. Despite reducing the threshold for activity, many of the identified regions exceeded a five-fold threshold, confirming that this threshold (in our system) is most likely appropriate for detecting functional SOX10 binding sites.

Additionally from our dataset, we identified SOX10 responsive elements near eight loci with a known or predicted role in repressing glial cell development: *Notch1*, *Hmga2*, *Hes1*, *Mycn*, *Id2*, *Id4*, *Sox5*, and *Sox6*. These findings were unexpected due to the known role of SOX10 in regulating the expression of genes that encode pro-myelination proteins (*e.g.*, *MBP*, *MPZ*, and *PMP22*) (Peirano et al. 2000; Wei et al. 2004; LeBlanc et al. 2006; Li et al. 2007; Jones et al. 2011a; Jones et al. 2011b). We showed that all eight loci are developmentally regulated during myelination *in vivo* in a manner consistent with a role in inhibiting glial cell differentiation. We also validated a SOX10 binding site at seven of the eight loci. We identified a SOX10 ChIP-Seq peak at *HES1* and luciferase assays demonstrated that this genomic segment has strong enhancer activity (Figure 4.8). However, deletion of the predicted SOX10 binding sites in *Hes1-R1* (Table 3) did not reduce luciferase activity. Further mutagenesis of this genomic segment will be required to identify sequences necessary for the observed activity, which may reveal a degenerate SOX10 consensus sequence. Another possibility is that these are false positive results, which ChIP-Seq can be susceptible to, similar to some of the SOX10 regions overlapping ChIP-Seq peaks identified in Chapter 3 (Fisher et al. 2012; Kvon et al. 2012; Leonid et al. 2013).

When we depleted SOX10 activity in Schwann cells *in vitro* and *in vivo* seven of the eight loci were downregulated; while *Hmga2* harbors a validated SOX10 response element (Figure 4.8), depletion of SOX10 activity did not reduce *Hmga2* expression. Further analysis will be required to determine if this SOX10 response element regulates an adjacent locus or if depletion of SOX10 at specific developmental timepoints results in reduced *HMGA2* expression. Consistent with our findings, previous global analyses of SOX10 function revealed that two of the above eight loci are downstream of SOX10: *Id2* and *Notch1* (Srinivasan et al. 2012). Our analysis now localizes at least some of the SOX10-dependent enhancers responsible for the regulation of *Id2* and *Notch1*.

SOX5 and SOX6 are members of the SOXD family of transcription factors and act as negative regulators of myelination in the central nervous system (Stolt et al. 2006); these proteins inhibit the expression of SOX10 target genes (*e.g., MBP*) in oligodendrocytes by competing with SOX10 for DNA binding at sites within cis regulatory elements. SOXD family member lack a transcriptional activation (or repression) domain and are unable to stimulate gene expression upon binding to DNA (Hagiwara 2011). To allow oligodendrocyte differentiation and myelin production, *SOX6* mRNA is targeted for degradation by two microRNAs (miR) in these cells:

miR-219 and miR-338 (Zhao et al. 2010). It was recently reported that SOX13 (the third and final member of the SOXD subgroup) also has an antagonistic effect on the ability of SOX10 to activate the expression of myelin genes in the central nervous system (Baroti et al. 2015). Indeed, *SOX13* is among the group of 191 loci at which we identified a highly confident SOX10 binding site - a single genomic segment within SOX10 ChIP-Seq and DNase-Seq peaks approximately 62 kilobase pairs upstream of *Sox13* (rn5 coordinates chr13:55425486-55425636). Interestingly, a relationship between SOXD and SOXE (SOX8, SOX9, and SOX10) transcription factors has been previously proposed because ablation of SOX8 or SOX9, but not SOX10, reduces *Sox6*, but not *Sox5*, expression in the developing spinal cord (Stolt et al. 2006).

In addition to genes that encode SOXD proteins, we identified *NOTCH1*, *HES1*, *MYCN*, *ID2*, and *ID4* as SOX10 target genes. NOTCH1 is a transmembrane receptor that regulates Schwann cell proliferation, inhibits Schwann cell differentiation in perinatal nerves, and facilitates dedifferentiation of Schwann cells after nerve injury (Woodhoo et al. 2009). HES1 is an effector of NOTCH signaling, acts as a transcriptional repressor (Sasai et al. 1992; Jarriault et al. 1995), and is highly expressed during early stages of Schwann cell development (Woodhoo et al. 2009). In cultured mouse oligodendrocytes, HES1 maintains cells in an immature state, and overexpression of HES1 results in reduced expression of myelin related genes (*Mbp* and *Plp*) (Ogata et al. 2011). MYCN is a proto-oncogene and is known to inhibit astrocyte differentiation from neural precursor cells (Sanosaka et al. 2008); however, the role of MYCN during Schwann cell myelination has not been studied.

Inhibitors of differentiation 2 and 4 (ID2 and ID4) proteins are known to inhibit oligodendrocyte differentiation and the lack of both proteins results in premature oligodendrocyte differentiation (Kondo and Raff 2000; Wang et al. 2001; Marin-Husstege et al. 2006). Furthermore, *Id2* and *Id4* expression declines in Schwann cell development, and ID2 limits induction of *myelin protein zero* (*Mpz*) expression in primary rat Schwann cells (Stewart et al. 1997; Mager et al. 2008). Consistent with our findings, RNA-Seq of oligodendrocytes isolated at various stages of mouse brain development (Zhang et al. 2014) show that *Sox5*, *Sox6*, *Notch1*, *Hes1*, *Mycn*, *Id2*, and *Id4* are developmentally regulated in the central nervous system—*Hmga2* does not appear to be expressed in the cells assessed in this study. Therefore, these genes are predicted to play a role in preventing premature glial cell differentiation in the central nervous system and likely perform a similar role in the peripheral nervous systems.

Combined with previous findings, our data suggest a model (Figure 4.11) where SOX10 activates the expression of genes that inhibit Schwann cell differentiation during early stages of Schwann cell development, thus preventing the precocious expression of myelin proteins. Subsequently, EGR2, NAB, and microRNAs (see below) inhibit the expression of the negative regulators of myelination (*e.g.*, SOXD proteins), which, in part, allows SOX10 to activate the expression of pro-myelination proteins. In addition to the data presented in this study, previous reports support specific aspects of this model. For example, EGR2 likely represses the expression of many of the eight loci reported here. Egr2 and Nab repress *Id2* and *Id4* before myelination via Nab binding to Chd4 (Mager et al. 2008) [conditional ablation of Chd4 in Schwann cells causes increased expression of immature Schwann cell genes including *Id2* and delayed myelination, radial sorting defects, hypomyelination, and the persistence of promyelinating Schwann cells in conditional knockout mice (Hung et al. 2012)]. Furthermore, a comparison of Sox10 and Egr2 binding with expression profiles in Schwann cells treated with siRNA for Sox10 and Egr2-deficent peripheral nerves (Srinivasan et al. 2012) revealed that *Notch1* and *Id2* are Sox10-activated and Egr2-repressed, and *Id2*, *Hmga2*, *Sox5*, and *Id4* remain

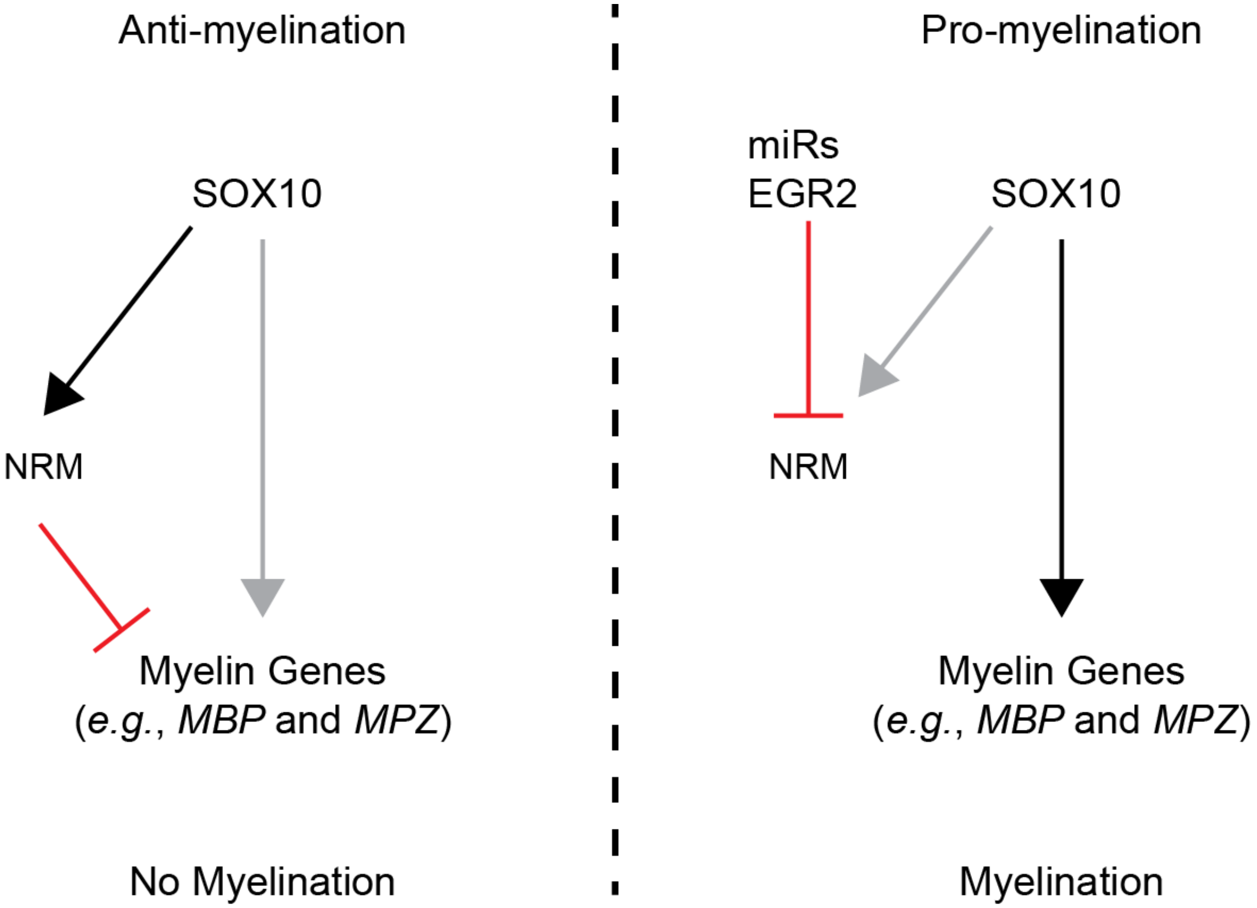


Figure 4.11 *A Simplified Model for the Role of SOX10 in Maintaining a Pre-myelinating State in Developing Schwann Cells*. Previous to myelination (anti-myelination; left side), SOX10 activates the expression of negative regulators of myelination, which inhibit the expression of myelin genes such as *MBP* and *MPZ*. During the activation of the myelination program (pro-myelination; right side), EGR2 and micro RNAs (miRs) inhibit the expression of negative regulators of myelination, which allows SOX10 (and EGR2) to positively regulate the expression of myelin genes.

high in peripheral nerves from *Egr2*- or *Nab*-deficient mice (Le et al. 2005; Mager et al. 2008). Finally, SOX10 directly regulates the expression of *EGR2* (Ghislain and Charnay 2006) and miR-338 (Gokey et al. 2012). In sum, these findings indicate that SOX10 is directly responsible for maintaining a premyelinating state, the switch to a myelinating state, and the expression of myelin proteins (Kelsh 2006). While our data and the data of others supports the proposed model (Figure 4.11) there are additional, non-mutually-exclusive possibilities including: (**1**) decreased expression of SOX10 during or prior to myelination (D'Antonio et al. 2006) despite the fact that it is required by terminally differentiated Schwann cells (Bremer et al. 2011); and/or (**2**) the activity of histone deacetylases, which are known to inhibit another negative regulator of Schwann cell differentiation, NF-κB (Chen et al. 2011).

While previous efforts have been successful in globally identifying SOX10 binding sites and target genes (Srinivasan et al. 2012; Anido et al. 2015), our computational strategy afforded a glimpse of SOX10 function that is not dependent on gene activity in cultured cells or in tissues at specific developmental stages. In fact, this less-biased (albeit less biologically relevant) approach is likely the reason that we were able to identify specific repressors of myelination.

In this chapter, we modified our computational and functional pipeline that resulted in expanding the panel of known SOX10 response elements and target loci. These efforts revealed a potentially novel function for SOX10 in repressing myelination in early stages of Schwann cell (and possibly neural crest) development; although additional *in vivo* functional studies will be necessary to support the proposed model. Furthermore, we provided useful datasets for the scientific community and expanded the mutational screening space for disease-causing mutations—or modifiers of disease—in patients with peripheral neuropathy and other SOX10-related phenotypes. In the final chapter, I will provide a summary of the major findings from this thesis and provide new questions and future experiments that arise from the knowledge generated from this body of work.

CHAPTER 5

Conclusions and Future Directions

Summary

The major aim of this thesis was to discover novel, functional non-coding *cis*-regulatory elements (CREs) important for the peripheral nervous system. We first developed a combined computational and functional pipeline to predict and rapidly evaluate candidate regulatory SNPs (rSNPs). This pipeline is highly versatile and capable of combining multiple different computational datasets to reprioritize candidate rSNPs. We next exploited a modified version of this pipeline to identify novel CREs, regardless of the presence of potential rSNPs. Throughout this work, we utilized this versatility to identify many novel CREs and rSNPs relevant for peripheral nerve and, more specifically, Schwann cell biology.

In Chapter 2, we developed the computational pipeline based on sequence conservation between human, mouse, and chicken. This method, termed phylogenetic footprinting, relies on the hypothesis that non-coding regions conserved among diverse species may imply some function (Hardison 2000). While phylogenetic footprinting has been utilized successfully many times (Zerucha et al. 2000; Goode et al. 2003; Kimura-Yoshida et al. 2004; Antonellis et al. 2008), a limitation to these studies is that the presence of sequence conservation does not always result in function (at least in the assays used) (Bejerano et al. 2004; Pennacchio et al. 2006; Ahituv et al. 2007). Similarly, a lack of sequence conservation does not necessarily mean lack of conserved function (Fisher et al. 2006; McGaughey et al. 2008).

Despite these limitations, we identified over two million regions that were five base pairs or greater in length and identical among the three species: human, mouse, and chicken. We looked for overlap between these conserved regions with SNPs and excluded exons to ensure the regions were not conserved due to protein function. This analysis revealed 6,164 conserved non-coding regions harboring a SNP genome-wide. We prioritized regions located on chromosomes 21, 22, and X to test the efficacy of our approach, which yielded a pilot dataset of 159 regions. The regions were cloned upstream of a minimal promoter directing luciferase expression in both orientations relative to the promoter (Antonellis et al. 2006) and assessed in three cell lines that provide an in vitro model of a peripheral nerve (neurons and glia) and a target tissue (muscle): S16 (Schwann cells; (Goda et al. 1991)), MN-1 (motor neurons; (Salazar-Grueso et al. 1991)), and C2C12 (muscle cells; (Yaffe and Saxel 1977a)).

Out of the 159 regions, we successfully assessed 144 regions in both orientations, which revealed 28 unique regions that displayed a greater than five-fold increase in luciferase activity compared to the empty control vector whose activity had been set to ‘1’ (S16 = 13 regions, MN-1 = 11 regions, C2C12 = 21 regions). Interestingly, we observed regions that displayed orientation-dependent activity in the luciferase assays. While this may be a consequence of our artificial system, orientation-dependent enhancers have been described previously (Nishimura et al. 2000; Wei and Brennan 2000; Swamynathan and Piatigorsky 2002), and further studies are needed to distinguish if the regulatory activity of these regions is dependent on orientation.

Of the 28 active regions, 13 displayed a significant allele-specific differences in at least one orientation and in at least one cell line. Using TRANSFAC (Matys et al. 2003), we predicted transcription factors that were predicted to have differential binding to the major and minor alleles. Combing these predictions with the different activity of the regions could give strong candidate transcription factors. For example, regions that were active in all three cell lines and displayed similar significant allele-specific differences may be bound by a more ubiquitously expressed transcription factor. Conversely, regions that displayed cell-specific activity would be predicted to be regulated by a transcription factor that is expressed and restricted to that cell type. For example, SC21-27 displayed regulatory activity exclusively in muscle cells and the minor allele was significantly less active relative to the major allele. Our TRANSFAC analysis predicts the creation of a novel LEF-1 binding site, which has been shown to exert a repressive affect (Billin et al. 2000; Mao and Byers 2011). While additional work will be necessary to validate any of the TRANSFAC predictions, one of the limitations of our TRANSFAC analysis was only including binding sites that were either created or ablated, with no regard for predicted transcription factor binding affinity, which can be necessary for appropriate gene expression (Rowan et al. 2010; Ramos and Barolo 2013).

Despite these limitations, the computational and functional pipeline developed in Chapter 2 was able to identify novel CREs and rSNPs. We wanted to further assess the functionality of this pipeline by incorporating transcription factor binding site information to determine if our computational pipeline can predict binding sites in a transcription factor centric manner. We choose the transcription factor SOX10 for three main reasons: (1) it is critical for Schwann cell function (Britsch et al. 2001; Finzsch et al. 2010; Bremer et al. 2011), (2) mutations within SOX10 binding sites can cause peripheral neuropathies (Houlden et al. 2004; Jones et al. 2011a), and (3) rSNPs within SOX10 binding sites can alter disease severity (Emison et al. 2005; Emison et al. 2010). Additionally, SOX10 has a well-characterized consensus sequence (Peirano and Wegner 2000; Srinivasan et al. 2012), and functional binding sites have been observed that are conserved between human, mouse, and chicken (Antonellis et al. 2008; Gokey et al. 2012; Hodonsky et al. 2012).

Using the SOX10 consensus site information, we were able to identify 224 conserved, non-coding SOX10 monomers containing SNPs within the human genome. This dataset was further prioritized using SOX10 ChIP-Seq (nine regions) (Srinivasan et al. 2012) and dimeric SOX10 sites where both monomers were conserved but only one monomer contained a SNP (13 regions). These 22 regions were assessed in our luciferase assay, and we identified four regions that displayed strong activity; however, only two regions displayed significant allele-specific differences in regulatory activity.

We further characterized one region, rSOX-4, to determine the gene(s) regulated by this enhancer. rSOX-4 was selected because it overlaps many genomic features associated with enhancers (*i.e.*, ChIP-Seq, DNase HSS, and H3K27Ac). Upon deleting this region from S16 cells, we performed RNA-Seq and ddPCR to ultimately identify one candidate target gene: *Tubb2b*. We further interrogated the function of rSOX-4 by generating transient transgenic mice harboring a rSOX-4:*LacZ* transgene. This revealed that rSOX-4 is active *in vivo* in both the dorsal root ganglia and melanoblasts. Since *Tubb2b* has been implicated in neuronal migration (Jaglin et al. 2009; Breuss et al. 2015), it is reasonable to predict it may perform a similar function in migratory cell populations derived from the neural crest. While these data strongly suggest that the target gene of rSOX-4 is *Tubb2b*, additional studies will be necessary to determine the role of *Tubb2b* in Schwann cells and other neural crest derivatives.

In Chapter 4, we utilized the versatility of our pipeline to identify SOX10 regulated regions that resided near genes involved in myelination, which provided the surprising result that SOX10 potentially regulates genes that inhibit glial cell differentiation. Compared to Chapter 3, we identified conserved dimeric SOX10 binding sites (and also required the intermonomeric spacer to be conserved) that resided within genes (or 2.5 kilobase pairs upstream or downstream) with a known role in Schwann cells or myelination. We also included regions which resided within genes whose expression is significantly altered upon SOX10 depletion (Lee et al. 2008). By overlapping these datasets, we identify 57 regions that we assessed in our luciferase assays, of which seven demonstrated a greater than 2.5-fold increase compared to an empty vector control that has the activity set to '1'. Interestingly, one of these regions (SOX10-CCS-13) resided within an intron of *Sox6*. Using SOX10 expression in MN-1 cells combined with RT-PCR, 5'-RACE experiments, and our RNA-Seq dataset, we were able to demonstrate that SOX10-CCS-13 acts as an alternative promoter.

Since our computational strategy was successful in identifying novel SOX10 binding sites in this chapter (and in Chapter 3), we utilized the versatility of our pipeline and additional genome-wide datasets to repriortize our SOX10 binding sites to identify regions that overlapped both SOX10 ChIP-Seq (Srinivasan et al. 2012) and S16 DNase HSS peaks (Chapter 3). We wanted to determine if combining all three datasets would enrich for strong candidate SOX10 binding sites. By overlapping all three datasets (*i.e.* sequence conservation, ChIP-Seq, and DNase HSS) we revealed 214 putative SOX10 binding sites that reside within 191 known genes. Using gene ontology (Ashburner et al. 2000), we identified 10 significantly enriched biological processes. Strikingly, of the 10 processes, three related to negative regulation of gliogenesis, which was due to the presence of six genes: *NOTCH1*, *HMGA2*, *HES1*, *MYCN*, *ID4*, and *ID2*.

We assessed each predicted SOX10 binding site in our luciferase assay and were able to demonstrate functional SOX10 binding to all regions, except *Hes1*. Interestingly, *Hes1* is developmentally downregulated at the onset of myelination in rats, suggesting additional SOX10 binding sites reside within the locus or additional transcription factors regulate this region. Additionally, experiments utilizing siRNA knockdown of *Sox10* in S16 cells demonstrated decreased expression for all genes, except for *Hmga2*. This data, combined with *in vivo* timecourse experiments assessing gene expression levels suggested that five of the six genes (not *Hmga2*) are regulated by SOX10. This data lead us to a model where SOX10 positively regulates both pro- and anti-myelinating genes. While the exact mechanisms governing the switch between these two states is unclear, one hypothesis involves repression of these genes by EGR2. This is supported by the observation that *Notch1* and *Id2* are SOX10-activated and EGR2-repressed, and *Id2*, *Hmga2*, *Sox5*, and *Id4* remain highly expressed in peripheral nerves from *Egr2*-deficient mice (Le et al. 2005; Mager et al. 2008).

Throughout this thesis, we generated many datasets which we feel will be useful to other investigators studying comparative genomics, SOX protein function, and Schwann cell biology. First, the conserved sequences we identified could be used to similarly prioritize consensus sequences for other transcription factors important for vertebrate development (*e.g.*, THAP1; Chapter 3). Second, the SOX10 consensus sequences we identified could be used to prioritize putative binding sites in other SOX10-positive cells including oligodendrocytes, melanocytes, and developing enteric nervous system neurons (Kelsh 2006). Finally, our DNase-Seq dataset from rat Schwann (S16) cells will be useful for anyone studying transcriptional regulatory elements, highly expressed genes, or any other nuclear structure characterized by open chromatin in myelinating Schwann cells; S16 cells express many myelin-related genes (*e.g.*, *PMP22*, *MPZ*, *MBP*, and *MAG*) and transcription factors (*e.g.*, SOX10 and EGR2) and are biochemically similar to myelinating Schwann cells (Hai et al. 2002).

In addition to the datasets generated, the computational pipeline developed and employed during the thesis research is extremely malleable. This allows for easy incorporations of additional transcription factor binding site information, as well as future genome-wide datasets, which can be used to identify CREs. Finally, the results from this thesis have not only identified many novel enhancers and rSNPs important for the peripheral nerve but potentially implicated a new gene, *Tubb2b*, in Schwann cell, and more generally neural crest biology.

Future Directions

Massively Parallel Reporter Assays

While the work in this thesis generated a computational and functional pipeline that was successful in both the identification and validation of CREs critical for the peripheral nerve and muscles, many questions and future experiments remain. In Chapter 2, we only assessed a small subset of the 6,164 conserved, non-coding regions harboring SNPs. When this work began, there was no effective way to assess thousands of regions simultaneously for regulatory activity; however, recent methods have been developed which address this problem: STARR-Seq (Arnold et al. 2013) and Cre-Seq (Kwasnieski et al. 2012).

Both of these methods utilize next-generation sequencing technologies that allow assessment of thousands (potentially millions) of putative CREs simultaneously. Both methods involve cloning the regions into a reporter construct and transfecting millions of cells. Through sequencing of RNA to ascertain the reporter expression and normalizing these counts to the transfected DNA plasmid, it is possible to obtain the activity of CREs from multiple candidate regions. One limitation is the starting materials. STARR-Seq generally relies on fractionation of the entire genome and is unable to assess a small handful of candidate regions. Comparatively, CRE-Seq generally utilizes synthetically synthesized oligonucleotides which are currently limited to approximately 150 base pairs. Recently, however, a modified approach for capturing candidate regions has been used to successfully identify CREs in conjunction with CRE-Seq (Shen et al. 2016).

This method leverages the advantages of both genome fractionation with targeted region capture. In the first step, biotinylated synthetic oligonucleotide probes are designed against the regions of interest. These probes are 80 base pairs in length, but are designed to overlap additional adjacent probes to cover the entire region targeted to be captured. For example, a 300 base pair target region may have five probes that overlap each other by approximately 20 base pairs. Next, the entire genome is fractionated to a predetermined size. This can be based on the average length of DNase-Seq peaks (approximately 300 base pairs) (Natarajan et al. 2012) or the average length of the regions assessed in Chapter 2 (approximately 900 base pairs).

Once the genome has been fractionated, it is incubated with the targeted biotinylated probes, which hybridize to the targeted regions of the genome, and the complex can be isolated from the background (not targeted) genomic DNA by streptavidin-coated magnet beads. The RNA probes are then degraded away, linkers for PCR amplification and cloning are attached, double-stranded DNA molecules are generated through PCR, and finally the regions are cloned into the reporter construct. CRE-Seq requires an additional step to incorporate unique barcodes with each candidate region. This process occurs randomly but is ultimately resolved through DNA sequencing of the plasmid DNA.

An experiment similar to the one described above would allow us to assess all of our predicted regions containing SNPs. One potential problem would be the inability to control the sequence integrity (*i.e.*, PCR induced mutations, allele-specific capture biases, additional SNPs in cis to the one under study, etc.). While these problems arose during the cloning procedures used in Chapter 2, they were not a major issue due to our ability to detect them during the cloning process. Depending on the number of confounding sequence changes, select regions could be assessed using tradition cloning methods, because the DNA sequencing of the plasmid DNA would inform us of any variants within the region.

Based on our rates of identification of active regions (greater than five-fold; 19.4%) and rSNPs within active regions (46.4%), we anticipate the identification of 1,195 active regions (of the 6,164 predicted regions). From these predicted regions displaying regulatory activity, we anticipate identifing 554 regions would contain a rSNP. While we should be cautious when extrapolating from data, the previously identified regions in Chapter 2 demonstrate there are a large number of CREs remaining to be identified that are functional in the peripheral nerve or muscle cells.

A major advantage of using either CRE-Seq or STARR-Seq is the ability to clone the captured regions into multiple reporter constructs. This allows every region to be assessed with multiple promoter elements, which may uncover novel active CREs that are only active with a specific promoter. For example, utilizing a promoter derived from a gene that is specific for muscle cells may uncover novel, muscle cell-specific CREs that are unable, for example, to interact with the minimal E1B promoter. This effect has been observed before, but it is currently unclear what the mechanistic factors are that determine enhancer-promoter specificity (Zabidi et al. 2015).

CRE-Seq or STARR-Seq could also be leveraged to identify repressive elements. As discussed in Chapter 2, our method at present is unable to identify repressors. Utilizing the versatility of massively parallel reporter assays would allow the regions to be assessed with an extremely active promoter. Similarly, multiple active promoters could be assessed to examine potential repressor-promoter specificity. For example, a ubiquitously active promoter, such as the CMV promoter, element could be compared with the SOX10 promoter in Schwann cells. We would expect both promoters to direct high levels of reporter transcripts, and any element that caused a significant reduction in expression would be a candidate repressive element.

SOX10 Binding Site Affinity

In Chapter 3, we utilized our computational and functional pipeline to identify rSNPs disrupting conserved SOX10 consensus sites. While we are confident in the minimal SOX10 consensus sites used throughout the thesis, we can not be certain that SOX10 is incapable of binding to addition sequences, the effect of a SNP interconverting different SOX10 monomers, nor if any additional sequence information is required beyond the ‘core’ motif for SOX10 binding. For example, SOX proteins generally bind to the sequence 5’-(A/T)(A/T)CAA(A/T)G-3’ (Harley et al. 1992; Laudet et al. 1993; Harley et al. 1994). While SOX10 has a slightly different binding site (Chapters 2 and 3), we only considered the ‘core’ (five internal base pairs) binding site for predictions. In Chapter 3 we noticed a difference in the GC content of the intervening sequence of all SOX10 dimeric regions tested (GC content = 35%) and that of the active dimeric sites (GC content = 61%). While the number of sequences evaluated (57 regions total) was small, these data are consistent with the high GC content of the intervening sequence within other validated dimeric SOX10 binding sites (Peirano et al. 2000; Antonellis et al. 2008; Gokey et al. 2012;

Hodonsky et al. 2012; Brewer et al. 2014), and with a ‘G’ nucleotide being the most commonly observed nucleotide after the ‘core’ motif as seen via SOX10 ChIP-seq analysis (Srinivasan et al. 2012).

One approach to generate higher quality SOX10 consensus sites information and obtain greater predictive power of the effect of SNPs on SOX10 binding sites is through the use of programmed allelic series (PALS; (Kitzman et al. 2015)). PALS allows for the generation of every possible combination of mutations within a specific sequence. It has generally been applied to protein-coding sequences; however, the method is applicable to determining the sequence requirements of *bona fide* SOX10 binding sites and could simultaneously assess the effect of individual SNP alleles within those regions. Using PALS to mutagenize previously characterized SOX10 binding sites, such as the dimeric SOX10 sites within the *MPZ* promoter (Peirano et al. 2000) and rSOX-22 (Chapter 3), and the monomeric SOX10 site at the *SH3TC2* promoter (Brewer et al. 2014) and rSOX-4 (Chapter 3), may uncover additional sequence requirements for SOX10 binding.

The mutagenic primers could be designed to mutate each base pair of the ‘core’ motif, plus one additional base pair on both the 5’ and 3’ ends. For the dimeric SOX10 sites, it is currently not possible to mutagenize both monomers concurrently, due to both the length and number of primers required. To partially alleviate this problem, each monomer would be mutagenized independently, resulting in complete mutagenesis of six monomeric sites each seven base pairs in length. Once the SOX10 binding sites were, clonal libraries would be constructed and assessed using massively parallel reporter assays as described above.

These data would be valuable in future predictions of the effect of both rSNPs and patient variants on SOX10 binding sites. For example, the patient variant assessed in Chapter 3 interconverts one predicted SOX10 monomeric site to another predicted SOX10 monomeric site. We were unsure what the functional consequence of the patient variant would be due to the lack of knowledge about the sequence constraints of SOX10. Additionally, while our computational pipeline predicted rSNPs which ablated the SOX10 ‘core’ motif, we identified two regions in Chapter 3 that overlapped SOX10 ChIP-Seq data, which were active in S16 cells, but the SNP had no effect. Understanding the effects of SNPs on the binding site would lead to more confident predictions. One caveat to our analyses is that it is unclear if these regions are truly SOX10 response elements, as we did not perform the requisite functional studies. Despite this, identification of the sequence constraints of SOX10 binding sites at multiple characterized CREs will lead to greater predictive power of the functional relevance of both rSNPs and patient variants that map to SOX10 binding sites.

Tubb2b Function in Dorsal Root Ganglia, Melanoblasts, and Schwann Cells

In Chapter 3, we identified *Tubb2b* as a candidate target gene of the rSOX-4 enhancer. We generated transient transgenic mice that harbored a *LacZ* reporter directed by rSOX-4 and observed expression within the dorsal root ganglia and melanoblasts at E11.5. There was also some variable expression observed in the dorsal portion of the head and brain, which may represent cranial neural crest cells. As it is known that *Tubb2b* plays a critical role in neuronal migration (Jaglin et al. 2009), we anticipate *Tubb2b* may be performing a similar function in migratory neural crest and derivative cells.

To assess if *Tubb2b* is indeed the target gene of rSOX-4, a *Tubb2b* reporter mouse could be used. If rSOX-4 is regulating *Tubb2b*, we would anticipate the expression patterns of *Tubb2b* at E11.5

in mice to overlap the patterns observed in our rSOX-4:*LacZ* transgenic mice. Fortunately, a mouse reporter model harboring a BAC spanning *Tubb2b* and including approximately 150 kilobase pairs upstream to 40 kilobase pairs downstream has already been generated with *GFP* knocked into the coding sequence of *Tubb2b* (Breuss et al. 2015). Interestingly, the authors observe GFP expression in the spinal cord and the developing head/brain at E14.5; however, the authors do not present any data at E11.5.

Similarly, the generation of rSOX-4:*GFP* stably integrated mice into the endogenous rSOX-4 locus would allow us to examine the spatial and temporal expression patterns globally. By constructing a homologous repair template harboring a floxed rSOX-4 allele directing *GFP* expression, we would be able to also assess the effects of removing rSOX-4 in any tissue through Cre recombination. We would anticipate the expression patterns of such a mouse to recapitulate our results observed in the transient transgenic mouse at E11.5. Furthermore, since this region was identified originally in our S16 Schwann cell model, we would predict the region to be active in Schwann cells, although the precise developmental window is unclear. This is due to the fact that rSOX-4 is a SOX10 response element, and SOX10 is necessary throughout all stages of Schwann cell development (Britsch et al. 2001; Finzsch et al. 2010; Bremer et al. 2011). Based on the known role of *Tubb2b* in migratory cells, we may predict rSOX-4 would be active in migratory Schwann cell precursor cells; however the S16 cells are a model of mature myelinating Schwann cells. Interestingly, *Tubb2b* is highly expressed in both oligodendrocyte precursor and 'newly formed' oligodendrocytes, but the expression levels are greatly reduced in mature myelinating oligodendrocytes (Zhang et al. 2014). It is possible for rSOX-4 to be active within both populations, as well as other cell types, and additional studies, such as the one described above, will be necessary to understand the *in vivo* role of rSOX-4.

If the mouse model was made as described above, it would be possible to use the same mouse to study the effects of rSOX-4 deletion in mice. Based on our current data, we would anticipate defects to arise in the dorsal root ganglia and melanoblasts; however, these predictions may change based on the previously described experiments. Interestingly, deletion of rSOX-4 from the S16 cells resulted in nearly complete loss of *Tubb2b* expression in the cell lines. This suggests that a mouse lacking rSOX-4 may display similar phenotypic defects observed in *Tubb2b* deleted mice. Unfortunately, a *Tubb2b* knockout mouse does not currently exist, and the only mouse model available to study *Tubb2b* harbors a point mutation that does not disrupt *Tubb2b* expression levels (Stottmann et al. 2013).

To further study the role of *Tubb2b* in the tissues where rSOX-4 is active, a knockout mouse model would need to be generated. Through the Knockout Mouse Project (KOMP; www.komp.org), a conditional knockout allele of *Tubb2b* has been generated, and is available as both embryonic stem cells or incorporated into a vector. The generation of these conditional knockout mice would allow us to determine the functional role (if any) of *Tubb2b* in any rSOX-4 affected tissue, through use of specific Cre expressing mouse lines. We may anticipate that a global knockout of *Tubb2b* will be lethal, since mice harboring a point mutation within a critical amino acid motif do not survive beyond the neonatal stage (Stottmann et al. 2013); however, since this mutation does not appear to affect the expression levels of *Tubb2b,* it may be acting as a gain of function mutation, and loss of *Tubb2b* may be tolerated due to redundant function of other tubulin genes (Tischfield et al. 2010).

While generation of a *Tubb2b* conditional knockout mouse may uncover novel functions in migratory neural crest cells, we may conversely observe no phenotype in our candidate tissues (*i.e.* DRG, melanoblasts, or Schwann cells). Nevertheless, the generation of a conditional knockout mouse would be of use to the tubulin field, as this mouse could decipher any compensatory actions of other tubulins and help elucidate the unique function of *Tubb2b* in cortical neurons.

In addition to studying the potential function of *Tubb2b* in neural crest cells and tissue derivatives, assessing the rSNP within human populations may help elucidate any potential functional effects. While we have not applied minor allele frequency thresholds in our experiments, the minor allele frequency of the rSNP (rs16886790) within rSOX-4 is 0.2001, making it feasible to study in human populations. Studies are currently being conducted by Dr. Stephan Züchner (University of Miami) that are collecting DNA from patients with CMT caused by molecularly indistinguishable duplications of *PMP22*. The patients are being classified based on their disease severity, and SNP genotyping is being performed. We have been in contact with the researchers, and our SNP did not have any association within their preliminary results; however currently the sample size is small, and additional patients are being included in the study that may lead to an association between our SNP of interest and CMT disease severity.

While we originally identified regulatory function of rSOX-4 within Schwann cells, based on our rSOX-4 transient transgenic mouse data and the predicted role of *Tubb2b* in migration defects, utilizing patients with CMT may not be the most appropriate patient sample. To address this, generating mice harboring the human minor allele would allow for direct functional assessment of the rSNP. By definition, the SOX10 binding site is conserved within mice, but no mouse strain assessed harbors the human SNP. Utilizing CRISPR and a short homologous repair template harboring the human minor allele, we could knock-in the human minor allele into the endogenous mouse locus (Singh et al. 2015). From these mice, *Tubb2b* expression could be determined in multiple tissues and developmental time points. We would predict the minor allele would result in decreased *Tubb2b* expression in tissues where rSOX-4 demonstrated *in vivo* regulatory activity (*i.e.* dorsal root ganglia, melanoblasts, and potentially Schwann cell precursors [based on the regulatory activity in S16 cells]).

Global Characterization of SOX10 Binding Sites and Regulated Transcripts Important for Schwann Cells

One of the main goals of this thesis was to uncover novel CREs critical for both Schwann cells and the peripheral nerve in general. While initially we were limited to sequence conservation, and later SOX10 ChIP-Seq (Srinivasan et al. 2012), to identify a small number of putative SOX10 binding sites and rSNPs, the era of genomics has allowed us to greatly refine our computational predictions. We have started to utilize genome-wide approaches such as DNase-Seq and RNA-Seq to more globally characterize Schwann cells. Reassuringly, these new datasets generally overlap with the previous regions that were selected based exclusively on conservation, while also identifying novel candidates that are not conserved.

One of the more striking observations that arose from these detailed studies are novel SOX10 regulated (potentially Schwann cell specific) alternative promoters. Indeed, we identified a previously uncharacterized promoter at *Sox6* (Chapter 4), and work from our lab has previously identified novel alternative promoters at *Sh3kbp1* (Hodonsky et al. 2012) and *Mtmr2* (Brewer et al, manuscript under review at Human Molecular Genetics).

From these observations, we hypothesized that there may be additional uncharacterized promoters regulated by SOX10. To identify these using a genome-wide approach, our lab is employing CAGE-Seq on the S16 cell line (Shiraki et al. 2003; Takahashi et al. 2012). The preliminary results have confirmed the presence of alternative promoters at *Sox6* (Chapter 4), *Sh3kbp1*, and *Mtmr2*. Additional work is being performed to optimize the CAGE-Seq protocol, and methods are being developed to utilize existing datasets to identify novel alternative promoters in Schwann cells.

Another major finding from Chapter 4 was preliminary data suggesting that SOX10 influences expression of genes involved in negative regulation of myelination. While we were able to demonstrate functional SOX10 response elements near these genes, further studies will be necessary to link these binding sites to the genes and to determine the importance of these findings during peripheral nerve development. One experiment to assess this could be deletion of these elements in S16 cells using the CRISPR and RNA-Seq approaches developed in Chapter 3. One caveat of this approach is that the S16 cells are a model of mature myelinating Schwann cells, and would not be predicted to express negative regulators of myelination. To alleviate this potential problem (and utilize an *in vivo* system), our lab has begun to isolate pre and post-myelinating Schwann cells from mice. Once pure populations of Schwann cells have been isolated, we will perform RNA-Seq experiments on both myelination states to compare what genes are up or downregulated. To identify genes regulated by SOX10, these cell populations could be reisolated from mice harboring a floxed *Sox10* allele (Finzsch et al. 2010). The cell populations could then be transduced with lentiviral Cre:GFP (Ahmed et al. 2004) to excise *Sox10*, flow sorted for GFP positive cells to select for the removal of *Sox10*, and RNA-Seq performed to compare the *Sox10* knockout cells to the wildtype cells. This analysis would reveal genes that are regulated by SOX10 in both pre and post-myelinating cells and may support our preliminary work suggesting that the genes identified in Chapter 3 are indeed SOX10 regulated.

The results from this work could greatly expand our knowledge of the transcriptional hierarches in Schwann cell development.

Concluding Remarks

Throughout this thesis, we developed and employed a pipeline utilizing both computational predictions and functional evaluations. When these studies began, we were limited to assessing only conserved regions harboring SNPs. As we continued our efforts, genome-wide datasets became available either though collaborations (Sox10 ChIP-Seq; (Srinivasan et al. 2012)) or through our own efforts (DNase-Seq and RNA-Seq; Chapters 3 and 4). These datasets were used to reprioritize our predictions and lead us to novel Schwann cell biology. In the future, we plan to continue to refine our predictions through additional genome-wide datasets. While the computational and functional pipeline has lead to novel and exciting results, it is the hope of the author that the work performed in this thesis will be applicable not only in a basic biological context towards understanding both Schwann cell development and genetic modifiers of peripheral neuropathies, but also as a translational step towards potential therapeutic options for patients.

Appendices

Appendix 1 *Genomic Information and Primers for Regions Assessed in Chapter 2.*

Name	**Conserved Region (hg18)**[1]	**rs Number**	**Left Primer (Forward)**[2]	**Right Primer (Reverse)**[2]	**Size**[3]
SC21-1	chr21:14647869-14647874	rs13048016	TCAATTTCCCAGAGGAGAGG	ATCCACCAGGACAGAAAAGC	511
SC21-2	chr21:15299507-15299521	rs7280064	PCR Failure	PCR Failure	
SC21-3	chr21:15652270-15652285	rs16982386	TTGCACCATTATTGCTCAGG	CATTGGAGTTTCCACCATCC	913
SC21-4	chr21:15729880-15729886	rs2823280	TAGGGCTCCTTGAGAACTCTG	TCACAGCCTTAGGTGACTTCC	670
SC21-5	chr21:16880688-16880703	rs3803997	AGGGATGCTTTTGTATTGATGA	AGTTCTCTGGCATCTGTGTCC	619
SC21-6	chr21:16880718-16880725	rs3803998	Same primers as region SC21-5	Same primers as region SC21-5	
SC21-7	chr21:16887088-16887093	rs2823898	PCR Failure	PCR Failure	
SC21-8	chr21:19230871-19230876	rs460825	PCR Failure	PCR Failure	
SC21-9	chr21:19661559-19661570	rs2825543	TGTGAATGAAATGGCAGTCC	GTGCACACGGATGTAGATTG	700
SC21-10	chr21:21313182-21313188	rs7277262	AAAGCATTTCCATATTTTCAGG	TGAGCATGGAACAGAACTGG	503
SC21-11	chr21:21778651-21778662	rs233760	GCCATTCTTGTTTTGTATGAAAG	GGTTGATAATTGGGGGAAGG	483
SC21-12	chr21:22535874-22535880	rs2827297	TTACCGTTGTTTCCAAAGTGC	TGGTCTCTGTTTTTCCCTAGC	1085
SC21-13	chr21:27318695-27318702	rs233616	CGCTGTTCTAAACACGTCCA	ATGATGGCTCCTATGTAAAACC	520
SC21-14	chr21:28550903-28550909	rs2206849	TTCAGAGGCTTTGGAACTGC	AAAATGGACTGGCTTTCCTC	958
SC21-15	chr21:28681566-28681571	rs2831741	GAAATGTGGGCACAGTGAAG	TCATGCCCATGTTCTACAGG	538
SC21-16	chr21:29424534-29424540	rs2832203	AACCTCAAAAATCACAATCCA	AAAACCCTCCTCCTGTCAGA	477
SC21-17	chr21:29449389-29449394	rs9305393	GAAAGCACCAGACGTAGCTG	CGGTTCATAGCAAGCTCCTC	789
SC21-18	chr21:33139128-33139135	rs2833975	CTCCTTCCCATCTCACATCC	TTGTCCCTTGAGGCTTTGG	554
SC21-19	chr21:33213365-33213372	rs7281293	GGAGCAGACAGACCACACTC	CAGCTCCTTAAGCCCAACTG	1383
SC21-20	chr21:33273214-33273219	rs2834040	CTCTCTTCTCCACCCCAAGC	CTCCAGCAACCAGTCTCTCC	1254
SC21-21	chr21:33431020-33431029	rs8132254	ATGCTGTGAGTCTGGCTGTG	CCTTCCACCCCAAACCTATC	642
SC21-22	chr21:33431167-33431196	rs8132292	Same primers as region SC21-21	Same primers as region SC21-21	
SC21-23	chr21:35234522-35234528	rs8130590	ATGAGGCTGGAGCATTTCAG	TGAGCATTGCGCTATCAGAG	494
SC21-24	chr21:35366984-35366996	rs2834747	ACTCCAGGTGAGGATTGTGC	GGTGATGGTTGGAGATCAGG	1236
SC21-25	chr21:35367060-35367117	rs17227266	Same primers as region SC21-24	Same primers as region SC21-24	
SC21-26	chr21:36033307-36033313	rs2835112	TGGAAGCTCCAGAGAACTCC	TGCAATTTTCCCTGATTTGG	1022
SC21-27	chr21:36269198-36269214	rs2835196	GGCATTTGTGTGATTCTTTCC	CGGTTAGATCAAAAGGATCTGC	361
SC21-28	chr21:36281865-36281876	rs2249599	CCAGCACCTTCTGACAAACC	TTGCAAACCACATTGAGAGG	368

SC21-29	chr21:36317420-36317426	rs4817761	ATGTCCTGCTCTGGAAAAGC	TCTCTTTCCTGTGGCTTTGG	986
SC21-30	chr21:36846163-36846171	rs376590	AGACAAGCCATCCTTCTTGG	TCTTTGCAGTTGCCAGAACC	415
SC21-31	chr21:37916579-37916585	rs878307	TGCACTTGGTTTGCTAATGG	GTCTGGTTGGAAGAGCAAGC	513
SC21-32	chr21:38112515-38112521	rs2835986	PCR Failure	PCR Failure	
SC21-33	chr21:38940551-38940556	rs16996658	TAGGACTGCAGGTGATGAGG	GGACAAAAGCAGAAGGTTGC	620
SC21-34	chr21:38958107-38958118	rs8130434	TTGAGAGGCCGAAAGAAAAC	TAGGAATGGGACAGGACAGG	967
SC21-35	chr21:40499095-40499101	rs16999481	PCR Failure	PCR Failure	
SC21-36	chr21:41127253-41127260	rs2837850	TCCTCTTCAAGCGTTTCTCC	TCTCAGTTGCAATTCCTTTGG	526
SC21-37	chr21:41133606-41133612	rs8131683	CTTCTCATCCCCAGCTTCC	TCAGCTTTTGATTTTTGGACA	563
SC22-1	chr22:16689437-16689442	rs5992119	AAACCTGCCTGTGTCTGTCC	CATACAAAAGGGGCATTTCC	709
SC22-2	chr22:25493989-25493999	rs2051623	AGGGAATCTGGGGTACTTGG	GGTGGCTTTTGTCTTCTTGC	631
SC22-3	chr22:25540690-25540696	rs713974	TTTAGCCCCTTCTCTTGTGC	AATGGAGCACCAGGTTTCTG	720
SC22-4	chr22:25556945-25556960	rs6005195	CAAAGTGACCCGAGATGTCC	CCCCTCCATCTGAATAAAGG	726
SC22-5	chr22:25581021-25581028	rs16982950	TACACAGACCCCTCCCTCTC	GTGCAGCACCTGTCTCTCC	868
SC22-6	chr22:25588565-25588572	rs136557	GCCTCCATCTCTGTGAATCC	CAAACACCTTGGGAATTTGG	920
SC22-7	chr22:25653869-25653874	rs739251	TGCCTCATTCCTCAGAAACC	TGTGTTTCTTTATGCCCTTCG	888
SC22-8	chr22:25678449-25678472	rs5761863	ATGTCAACGAGGGAGCTAGG	AAGGGAGGAGGAGAATGAGG	492
SC22-9	chr22:25749225-25749250	rs17343778	GCATGTCACCATCAATCAGC	AGCTTTCCTGCTTCAACAGC	896
SC22-10	chr22:25769441-25769477	rs17429199	TCAGTGGGAACTCACCATAGC	ATTTGAGACCCTGATTTCTTAGC	901
SC22-11	chr22:25884083-25884096	rs4822861	CGGTTTGGCTACACAGAAGG	GAACCCTTAGGAACCCTTGTC	945
SC22-12	chr22:26012528-26012599	rs17173861	CCTGGGGAAACAGACATCC	TGCGCACACACGTATTTACC	773
SC22-13	chr22:26012622-26012627	rs7293113	Same primers as region SC22-12	Same primers as region SC22-12	
SC22-14	chr22:26146779-26146784	rs733164	CCCCTAAATAGCCCTGATCC	GAGCTCCTGGCTTTGAACC	881
SC22-15	chr22:26247799-26247804	rs17466256	GGCCACAGTTGATAGTCTGG	GAGGATGAAGGGGCTTGG	266
SC22-16	chr22:26274397-26274404	rs16985013	GGAACTTATAGGCCCCAAGG	TTACACCTCGTTCCCTCTGC	422
SC22-17	chr22:26913737-26913746	rs7284814	TCCAGGTGGGTTACATTTCC	TCAACTGAATAAACTTTGCTTTGC	422
SC22-18	chr22:26995094-26995113	rs16986240	CACCCTTCACTTCATCATGG	TGGGAAGAGTGTTGAGACACC	611
SC22-19	chr22:27167830-27167847	rs16986429	PCR Failure	PCR Failure	
SC22-20	chr22:27544643-27544648	rs2301429	GTGTTTCAGCTCTCCCTTGG	TTCTCAGGAAGGCCACTGC	515
SC22-21	chr22:28015405-28015410	rs16987366	AATGCAGCTGTTTTCCACAG	GCCTGTTGTCAAATGACTTTC	684
SC22-22	chr22:33205230-33205249	rs130593	TGGTTGGAGGAGAAGTTTGG	TTTCCCACTTGGATTGATGG	482
SC22-23	chr22:33363804-33363815	rs11089687	CAGCAGGATGTCATTGTTGG	GATCTGGTCCTTTGCTCTGC	479
SC22-24	chr22:35426848-35426890	rs2284017	CCAACCCACCCATTTCTG	AAAACTCCTGTGGGTCATCC	932
SC22-25	chr22:35430194-35430211	rs2267361	GGACGGGTAATTACAAACACG	AAACATGAACCCATCTCATGC	628
SC22-26	chr22:41020252-41020259	rs6002672	BP Failure	BP Failure	
SC22-27	chr22:42090126-42090141	rs695648	TGGATTTGGATTTTGGATCG	GTTGGAGGAATGCTCAGAGG	848
SC22-28	chr22:42438983-42438993	rs6006534	TGAAATTAGAGTGTGGCGTTC	TAGCGTGGGCTCACAGTAGG	340
SC22-29	chr22:42438997-42439002	rs17568513	Same primers as region SC22-28	Same primers as region SC22-28	

SCX-1	chrX:10519616-10519621	rs10284156	ATCCCCAAGGGAGTTGACTT	ACCCTAAAGACGCTCGATCA	1159
SCX-2	chrX:13701570-13701575	rs13187	GGGTGAGCATGATGGCTAGT	TGTTAGAGCCCACAAAATTGG	866
SCX-3	chrX:15529186-15529192	rs4646115	TTCATCCTGGAGAGGACAGA	GGGAAAATGTTGCCCAAGTA	344
SCX-4	chrX:17730099-17730104	rs2187846	CTGCAACACAGAACATACAA	GCAACTTGGACTGAAACTTCG	343
SCX-5	chrX:18560920-18560932	rs34122505	GCTCAGAGCTATGCTGTCAGTCT	TCACAAAAGTATGAGTCTGTGGA	545
SCX-6	chrX:22421316-22421332	rs5970650	TGCAGTCAAAATGCACTCAT	CATAATCTGGGAAAAACTAAAGCA	747
SCX-7	chrX:23089136-23089146	rs17343199	CCCCTTAGCTGCTTCTTTCA	CTAGTGCCTCCCCAAAGGTA	777
SCX-8	chrX:24566571-24566586	rs1921918	CAGGAATGGGCTTTCAACAT	ATGGCAGTATGGGCTTGAAC	793
SCX-9	chrX:24688106-24688116	rs5944676	CACCCCTCACCAGAATGAGT	AAAGGGAAAGCCTGTGAGGT	477
SCX-10	chrX:25170446-25170451	rs5986762	AAGGAAATGTGGAGGCCTTT	TTCCTGTGTGAGACAATGCAG	1057
SCX-11	chrX:25376363-25376368	rs5944057	GGGCACTTTCATGAGGCTTA	TCATGGCAGAGGAACAGACA	936
SCX-12	chrX:25411586-25411596	rs17286168	GCAAGACTCCTTTTGATTTTGA	CTTCAAAACATTAAATATGGGTTC	1044
SCX-13	chrX:29185519-29185525	rs16988485	CATTTGAGTGCATGGGAAGA	ATTCCAGCGGTGAAAACTTG	438
SCX-14	chrX:29403341-29403347	rs7065816	TTTCAGCACTTAGAATGACATGG	AAGACACGGATGCCCTTATG	1271
SCX-15	chrX:29847765-29847774	rs7063049	CAGGGACTCAAGCAAACACA	CAAAAATCCACACAAACAGACA	1375
SCX-16	chrX:30959319-30959330	rs444301	TGAGAGACAGCATTGGTTGG	GACCTTGATGGCTGCTGTTT	772
SCX-17	chrX:31226760-31226769	rs11797901	TTCACCAGAGGATCTAACAGCA	CGAACAATTAGTTGACTTTTCTTCAGT	898
SCX-18	chrX:31252044-31252049	rs7884417	AGAAAGGCAAGGAGGTCAAA	GCATCAGCTGTGCTTCAAAT	1037
SCX-19	chrX:31420373-31420379	rs16989672	TCCCCATAAAAAGCATCAGC	ACTTGTAGTTTGGCGCAATG	576
SCX-20	chrX:31435143-31435149	rs3788892	TCCAAAGAAAGCTGGCACAT	AGCAAAAAGGCCACAAATGA	613
SCX-21	chrX:31764702-31764707	rs1379871	TGGACATTAAGCTCAGGTGC	GCCACTCAGCCAGTGAAGG	1116
SCX-22	chrX:31803223-31803228	rs1800275	TGATACCAAATGAGAAAATTCAGTG	CCAAGAAGGACCATTTGACG	105
SCX-23	chrX:39008367-39008377	rs5917286	CCTGTCTTTGGCCCATTCTA	CATACATTGTACCCATAAAACATACA	833
SCX-24	chrX:39175198-39175203	rs2029475	GCTTCAGCCAGATTTCATCC	CAACCACTGCATTTCTGGTG	688
SCX-25	chrX:39190395-39190404	rs6610273	CTGGTATGCAGAGCCCACTT	CACTTGAGGTTCTCAGGACAG	872
SCX-26	chrX:43712864-43712876	rs12833438	GCAAATGACTGCCAGAGACA	ACCAATTCCCCTCCACTACC	501
SCX-27	chrX:82478122-82478130	rs1299087	ATGCCACACTGCCTTCAGTA	CAAGAAAACAGGGGAGTCTGA	2497
SCX-28	chrX:82637870-82637946	rs12860283	TTTCACATTTAGGCCCGAAG	TCCTGTAGCTTCTCCCATTGA	1384
SCX-29	chrX:85043549-85043617	rs16980331	GATAAGCATACCATAAAAGTTCA	AGGCAAGAGCTCTAGTTCAATG	1430
SCX-30	chrX:85287194-85287201	rs242849	TTCAGTTGGCTGAGGGTTTC	GGCCATTGATCATTGAAAGG	940
SCX-31	chrX:85367482-85367493	rs16980456	CACATTTGGAAGCCAGGAGT	TTGGAATACTTGGCTTTTCTTTG	1433
SCX-32	chrX:85404361-85404385	rs6653101	TCCCATATAGATCCACAAAACTGA	GGGATGGGGGTTGTTTTAAT	587
SCX-33	chrX:85443058-85443064	rs6623642	TTTGGCATGGGAGAGAAAAG	ACCCCATGGAAATGTTTGAA	1256
SCX-34	chrX:85764274-85764279	rs16980611	GAAGGCTCTTTGCCATTTACA	TGCCAACTGCAACTTAACCA	1464
SCX-35	chrX:85809671-85809677	rs5922268	CAGCAATGTCTTCCCTGGAT	GCAAACGAATGCAACATGAC	1427
SCX-36	chrX:85913194-85913210	rs1419032	CAGTGTCATATGCCCCAATG	TGGAATGATGAGGCTTTGGT	1068
SCX-37	chrX:85961901-85961908	rs2185879	PCR Failure	PCR Failure	
SCX-38	chrX:85971813-85971836	rs11092827	CAGGCAGTGCTGTGCTAAAG	CCAGGGACCAGAAGAAAACA	1254

SCX-39	chrX:86429198-86429205	rs16980794	GCTGAATCTGAGGCACCTTC	TGGAATGGCTCTCCTTTCAC	1380
SCX-40	chrX:86890323-86890341	rs5922480	AATCATGCCTTTCTCGGATG	TCTTCCCATACCCAATTCCA	1383
SCX-41	chrX:86890482-86890495	rs5924110	CAGTAAAAAGATTTGTTGGCAAT	TGTAGTAATTTACCGGTTTAGTAACCT	977
SCX-42	chrX:87641465-87641470	rs6522112	TTAGGCCCTTTAATGCTTGC	CAGATTAGTAGCTTCCCACAGTAGC	2126
SCX-43	chrX:91308657-91308662	rs4021810	CATGTCTGTTATAAGGAATTCATCTG	TTTCCAAATTTTGCCGATTC	1175
SCX-44	chrX:91336826-91336843	rs6618904	GGGACAAGAATGATGCCAAT	CAACAAATCACCAGGTGGAA	1977
SCX-45	chrX:92656893-92656900	rs12687113	GGCGACAGCATCAAAGAAAT	GAAGTGCAGAGGGCAAAGAC	1034
SCX-46	chrX:93640694-93640700	rs6619561	TGTAATTAGTACAAGGGTCTGATTT	TTGCCATTAACTTCCTGATGC	953
SCX-47	chrX:94777246-94777251	rs5990383	CAGAGGCTCCCTTTCTACCA	TTTCCTGTTGCTGGGGTTAG	930
SCX-48	chrX:96687553-96687558	rs5967326	TGTGGCCATCTGCTAAAATG	TTTCCCCACCACTGAGAAAG	524
SCX-49	chrX:96740407-96740415	rs5921859	AAGGGATTTTCACCCCACAT	TGTAAGACAAACAGAAAAAGGA	783
SCX-50	chrX:98164479-98164486	rs16982964	TGAATGCTTTGGCATTGGTA	GAAGCTAAATAGATTATTTTTCCTCCA	2370
SCX-51	chrX:99217165-99217172	rs985251	AGGCAAGCAGACATCACCTC	GGCAAGTGCATCTATCAGCA	781
SCX-52	chrX:99454087-99454093	rs7064056	GGGTGGGGAAGCTAAGAAAC	AAGAGCACTGGGACAAGCAT	1082
SCX-53	chrX:103711687-103711695	rs1004122	CATCCATTGATGTGCAGGAC	TGGCAGGCAAGGATTAGAAC	780
SCX-54	chrX:104030456-104030461	rs1343409	TTTTCCCAACAAGTCCTCCA	GCCCAAGGGAAAACAACTTT	640
SCX-55	chrX:104181005-104181031	rs16984615	GAGACCTGTGGCATCTTGTG	GGAAGCAAAGCATCCAAGAA	921
SCX-56	chrX:120221829-120221834	rs1861522	GGATATTCCGTTGGTTTTGC	CCATGTTTTATTGTTTCATTCACA	718
SCX-57	chrX:120428002-120428008	rs6608251	TATCTGGCCACTTTCCCTGT	GCCAAGTATCCTTTCCCACA	1425
SCX-58	chrX:121682472-121682478	rs17273301	CCCTCCATAGAGGCCTTGTT	TGTCAGTGGCAGAATTGCTC	572
SCX-59	chrX:122196457-122196480	rs7890100	AAATTGACTGGGTGGCAATC	GGTGATCTCTGGCTTTCAGG	607
SCX-60	chrX:123382405-123382410	rs2076164	GTGGGAGAAAGACTTGATTTTAAC	TTAGGCCAGAGGAGTCAGGA	626
SCX-61	chrX:123612555-123612568	rs16999342	CCCCACTGAGCCTGTCAATA	TAAGCTGCTTTGCCTAATATG	1179
SCX-62	chrX:124339837-124339845	rs3126112	TGATACGTCTGCTATTAGTGAAAGA	GCTGAAAAATGCTGATGGAA	1500
SCX-63	chrX:125247437-125247470	rs16998722	AGGAAGGGTGGCTGGTTATT	TGATTTCACTATGAAACCCACTC	2296
SCX-64	chrX:125729275-125729280	rs17303490	ATCCAGCTCTTCCTGAACCA	CCAACCTCAGGACAAGTTGC	1394
SCX-65	chrX:125925884-125925893	rs5930055	TTGCTTGCTTTCGAGTTGTTT	ACGGAACAAATGTCCTCACC	814
SCX-66	chrX:125954394-125954405	rs17332319	AGAGGGCTTCTTTGGCATTT	GGGGTCTGCATTAATGATGTG	1336
SCX-67	chrX:127229700-127229720	rs17266605	GGGTATCCTCCAGGTCTAGCA	TTGATCTGGCACTGGTTTCA	681
SCX-68	chrX:127874030-127874040	rs9887026	GGGCCAGTCAGATCCCTAGT	TGTCTGCCCCATTTATGTGA	681
SCX-69	chrX:128107907-128107918	rs722439	AAGTGGTTGGTGGCTGAATC	GGCAGTGGAGACAAGTGGTT	1076
SCX-70	chrX:131733694-131733707	rs7878720	GAAATGGGGAAGCACATCAC	CTTGAGACGGCATGGAAAAT	690
SCX-71	chrX:131733885-131733892	rs5933189	Same primers as region SCX-70	Same primers as region SCX-70	
SCX-72	chrX:135102713-135102725	rs2300913	TCTGATGGGTATGCCATGAA	GGGCAAGAGGCTGATAACAA	436
SCX-73	chrX:136737672-136737678	rs708697	GTTTTGCTTTGGGGTCGATA	TATCCTGCCTTTGAGGGATG	1222
SCX-74	chrX:136795666-136795675	rs6635500	CACCTTGGGAAAGAAGGACA	TTCCCTTTTCCACATCACCT	864
SCX-75	chrX:136916407-136916412	rs6633954	GCAGGAAACATGGCTCAAAT	GTTCCCAAGTGTCCCATACG	620
SCX-76	chrX:137579834-137579839	rs17510193	CTTCTGCCTCTCCCCTTCTC	AAATTTCAAAGTAGCGAAATTGG	588

SCX-77	chrX:139502708-139502777	rs5954039	TAAACCAAGCCAACCCAGAG	GGCTACAATCCTGCAAATCA	1425
SCX-78	chrX:146960885-146960891	rs6525876	TCCCATCCAGTCTTCCAAAC	CTCCCAAAGGGCTCTCTCTT	465
SCX-79	chrX:147139773-147139778	rs16994500	TGAAAAGAAGTTCATAGAAGGGAAA	GAAAAGTTTGCATTTTGTTTGAA	723
SCX-80	chrX:147303281-147303289	rs6641405	AAATAGCCCCCGTGTGATTA	TTCAGTTTGGCCCTTGGTAG	671
SCX-81	chrX:147430625-147430635	rs17252118	TGGACATCCTTCAGGAAAGC	CCAAACATAAAAGAGCATGGTG	952
SCX-82	chrX:147466469-147466477	rs5936216	TCTAGGGCTGCTCAGTCACA	CCGAGGAAGATCCCCACTAT	774
SCX-83	chrX:147612978-147612994	rs5980583	AACGTCCTCTGGCAAAAATG	GGCACAGCAATCTTCCTAGC	967
SCX-84	chrX:147637143-147637154	rs12686890	CTTGAGCACTCACGCAAAAA	TATGGGCTGACACTCATGGA	1137
SCX-85	chrX:147661284-147661318	rs17252278	AGCCTGATCTTGGCCTGTTA	TTGCCATTCTTTCCTCTTGG	623
SCX-86	chrX:147714683-147714691	rs1372593	GTGAGGGAGCTTTGTTCCTG	ACCCTCTGAGAATCCACTGC	1075
SCX-87	chrX:147759881-147759889	rs16994786	AGCCCATTTCCTGAATTTCC	CACAGGCAAAATGGGACTCT	1431
SCX-88	chrX:147991975-147991988	rs9308376	TCCATGGGAAAAATGCTTCT	ATAGGGATTGCTTGCTGCTC	1216
SCX-89	chrX:148185049-148185065	rs764908	CCTGGCACAAATACCGATCT	TGACTCCTTTGACCGTGTGA	1809
SCX-90	chrX:149322258-149322264	rs5924948	TGAGGCAAGAAAAGATTTGTG	AGCCTGAAACCAGATGTTGG	961
SCX-91	chrX:149632616-149632621	rs6627325	AAAACTGGCTGCACTGAAAAA	ATTTGGACAGGAGGGCACTA	1324
SCX-92	chrX:153231488-153231493	rs2070819	CGTGTTCACGACGAACTCAG	GTCTGCTTACGGAGCAGGTC	230
SCX-93	chrX:153239964-153239969	rs5987247	PCR Failure	PCR Failure	

[1] Regions are given in bed format for the human genome (hg18) and only encompass the base pairs identical among human (hg18), mouse (mm9), and chicken (Gal3).

[2] Primers sequences are displayed in the 5’ to 3’ orientation and do not include the gateway adapter sequences. The adapter sequences for the forward and reverse primers are 5’-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3’ and 5’-GGGGACCACTTTGTACAAGAAAGCTGGGT-3’ respectively.

[3] Sizes are given in base pairs and do not reflect the additional size of the gateway adapter sequences. An additional 58 base pairs should be added to the size to account for the gateway adapters.

Regions without primers represent the regions that were dropped due to technical errors or amplified with another conserved region. In place of primer sequences, the technical failure is listed: ‘Same primers as’ represents two conserved regions amplified in a single PCR product, ‘PCR Failure’ represents multiple primers tested without proper amplification, and ‘BP Failure’ represents proper PCR amplification (and product sequence verified) but unable to clone into pDONR221.

Appendix 2 *Genomic Information and Primers for SOX10 Regions Assessed in Chapter 3.*

Name	**Conserved Region (hg18)[1]**	**rs Number**	**Left Primer (Forward)[2]**	**Right Primer (Reverse)[2]**	**Size[3]**
rSOX-1	chr2:44834620-44834625	rs3738980	ACAGGAAGTTGCCAGAGTGC	TTGAGAAGAGAGCAGAATCC	321
rSOX-2	chr3:62426974-62426979	rs6445273	TGGCAGAATTCCTTATTACCG	CAGGTGAAATGTTTCATTGTGA	319
rSOX-3	chr4:146183931-146183936	rs34577920	AGATTTTAAAGGGCAAACATCA	GCAGCAGATATCAGCCTTCA	1340
rSOX-4	chr6:22818474-22818479	rs16886790	CCACTTCTATCTGGGCAAGG	TGTGAGTCCACTTGCAGAGC	922
rSOX-5	chr6:98577359-98577364	rs12524696	TCTTGCCAATTTAAGGTGCTC	CCATTTATCCAGACATGCACA	1084
rSOX-6	chr6:98692222-98692227	rs17814604	TCTTCCCAGTGTGGTCCAGT	GGCAGGGAATATGACAAACC	677
rSOX-7	chr7:131601713-131601718	rs1364510	ATCAATGTGTGGCTCAGCAG	TGGAGAAACCACCCAAGTTC	678
rSOX-8	chr13:66589464-66589469	rs17082112	AGGTGTCAAACCCATTCTGG	TGTGATTCAGAGCTCCAGTG	821
rSOX-9	chr15:55214627-55214632	rs2703617	TGTGCAAGTTTAAAGCAAAATC	GCTTGGCGAAATAACAAACC	1834
rSOX-10	chr1:48904340-48904354	rs1966247	ACACTGACCCCATCTTCCAG	AGTGTGCCCTTGTACCCTTG	624
rSOX-11	chr1:168449895-168449935	rs16863114	GTTGCTTGGGTGAAATGGAC	AAGACCAGGAAGGAGGTGCT	411
rSOX-12	chr6:118599910-118599921	rs17335828	GTGTGTCCCAGTGGACTCCT	TACCATGGAACCCAAAATCC	442
rSOX-13	chr7:9853948-9853966	rs12702949	ATGCTAAATGAGAATGCTGG	GTGCAATTCCAGTGCATGTG	615
rSOX-14	chr7:95350538-95350566	rs10249566	TATGTCAGCTGCCCAAAATG	TGCCAACATATTGCTGGTGT	596
rSOX-15	chr8:37281347-37281384	rs17333409	CACACCACCTCCCTCTTTGT	TAACCCAACTGCATGCTCAG	623
rSOX-16	chr8:138460036-138460047	rs16907090	CACAGACCCCTTTGCCTCTA	GAGTGGGGAGTGGTAATGGA	665
rSOX-17	chr9:80243161-80243189	rs17788061	AAAACAAGGCACGCTCTGAT	TCCTCAAATGAGCCACACTG	747
rSOX-18	chr10:78070796-78070831	rs17469556	TTGCACTTCTGTCTGCATCC	TCTCTCACCTCTCCCCTCAA	635
rSOX-19	chr10:130596685-130596718	rs11819115	CCAAGCCAGCTCTGGTAAAG	AATGCCTGCATGGTAAGGTC	605
rSOX-20	chr11:31400150-31400183	rs1376362	GGGGATGTGATGATCATGTG	AAACAAAATACGGCCATTCG	980
rSOX-21	chr11:125354123-125354139	rs11607720	TTCCCTTTCTTGGCAGTCAG	CTGTTGCTGGTTGGATCAGA	600
rSOX-22	chr16:53169712-53169728	rs1186802	GGCCTGAGCTGTATTTGAGC	CTTCAACTATCCGGCATTGG	489

[1] Regions are given in bed format for the human genome (hg18) and only encompass the SOX10 consensus site. For dimeric sites the coordinates include both monomers and the intermonomeric region.

[2] Primers sequences are displayed in the 5’ to 3’ orientation and do not include the gateway adapter sequences. The adapter sequences for the forward and reverse primers are 5’-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3’ and 5’-GGGGACCACTTTGTACAAGAAAGCTGGGT-3’ respectively.

[3] Sizes are given in base pairs and do not reflect the additional size of the gateway adapter sequences. An additional 58 base pairs should be added to the size to account for the gateway adapters.

Appendix 3 *Primers Used in Gibson Assembly to Construct Drug Resistance Repair Templates in Chapter 3.*

Names	**Primer used in Gibson Assembly**[1]	**Size**[2]
CSOX4 5 Hom Arm For	AGCTTGCATGCCTGCAGGTCGACTCTAGAGTTCAGGGAAGGACAAATTTCA	1085
CSOX4 5 Hom Arm Rev	AACCAATAGGCCGAAATCCCCAAAATCCCTATAACTTCGTATAATGTATGCTATACGAAGTTATAGCAGTGACGAGTCATAACCAG	
CSOX4 5 Blast Rev	TACCGTAAGTTATGTAACGGACCTCGAGCTATAACTTCGTATAATGTATGCTATACGAAGTTATAGCAGTGACGAGTCATAACCAG	
CSOX4 3 Hom Arm Rev	CGGCCAGTGAATTCGAGCTCGGTACCCGGGTCAAATGGTAACTGTGCCTACC	1059
CSOX4 3 Hom Arm For	CTCTAGCTAGAGCTTGGCGTAATCATGGTCATAACTTCGTATAGCATACATTATACGAAGTTATTCAGATGCTCTTTCACATAGCC	
CSOX4 3 Blast For	CAAACTCATCAATGTATCTTATCATGTCTGATAACTTCGTATAGCATACATTATACGAAGTTATTCAGATGCTCTTTCACATAGCC	
NeoR hCAS9 For	AGGGATTTTGGGGATTTCG	1620
NeoR hCAS9 Rev	GACCATGATTACGCCAAGC	
Bsd For	AGCTCGAGGTCCGTTACATA	1273
Bsd Rev	CAGACATGATAAGATACATTGATGAG	

[1] Primers used in in Gibson assembly to generate the drug resistant repair templates. Both blasticidin and neomycin resistance templates used identical 5' forward and 3' reverse primers for the 5' and 3' arms respectively. The Gibson assembly primers have been color coded to reflect different portions of the primers: black represents unique primer sequences to amplify the product, red represents homologous overlap with pUC19 backbone, blue represents homologous overlap with the drug resistance cassette, and green represents loxP sites.

[2] Size is given in base pairs and does include additional size of adapter sequences.

The 5' and 3' arms of homology were amplified from unmodified rat S16 genomic DNA, neomycin resistance template was amplified from the hCas9 expression plasmid (Addgene plasmid #41815), and the blasticidin resistance template was amplified from pCMV/Bsd plasmid (ThermoFisher - Cat no. V510-20).

Please note that CSOX-4 and rSOX-4 are the same region. The names were changed for the thesis, but the primers were called by the original CSOX-4 nomenclature.

Appendix 4 *Screening Primers Used to Verify Proper Recombination of rSOX-4 Mutant Cells in Chapter 3.*

Name	**Primer[1]**	**Size[1]**			
C-SOX4 5 SCRN For	TGATCATCTCATCTCCATTTTAGAG	1180			
Blast Seq Rev	GCGTTACTATGGGAACATACG				
C-SOX4 3 SCRN 3R	ATGTAGGCTCAAGGTGACAC	1414			
Bsd Scrn For	CTGCCCTCTGGTTATGTGTG				
C-SOX4 5 SCRN For	TGATCATCTCATCTCCATTTTAGAG	1253			
NeoR Seq Rev	GACTTTCCACACCTGGTTGC				
C-SOX4 3 SCRN 3R	ATGTAGGCTCAAGGTGACAC	1479			
Neo SCRN 3F	CTATGAAAGGTTGGGCTTCG				
C-SOX4 WT For	CCGTCTTCCTCTGACTCTCC	634			
C-SOX4 WT Rev	CAGATCTCCTCTCTCCTTAGCC				
SOX6 MCS 3 For	CAGGGGAGTCTAAGCCACAG	1759			
SOX6 MCS 3 Rev	CCAGGTGTCTGTCCTGTCC				
Name	**Primer[1]**	**LoxP[3]**	**WT[3]**	**Blast[3]**	**Neo[3]**
C-SOX4 5 SCRN For	TGATCATCTCATCTCCATTTTAGAG	1181	1798	2488	2835
C-SOX4 Cre 2R	CCTCAACACTTAACATAGCC				
C-SOX4 Cre 2F	GGATGCCGCAGAAATCATTG	1256	1873	2563	2910
C-SOX4 3 SCRN 3R	ATGTAGGCTCAAGGTGACAC				

[1] Primer sequences are given in the 5' to 3' direction and were designed against the rat genome (rn5) or drug resistance repair templates.

[2] Size of the PCR product is given in base pairs.

[3] Varying sizes of PCR product depending on the allele amplified: LoxP is the post Cre transfection allele (*i.e.* loxP scar), WT is the unmodified original S16 allele, Blast is the blasticidin resistance cassette inserted into rSOX-4 genomic location allele, and Neo is the neomycin resistance cassette inserted into rSOX-4 genomic location allele.

Shading indicates matching PCR primer pairs for a given reaction.

Please note that CSOX-4 and rSOX-4 are the same region. The names were changed for the thesis, but the primers were called by the original CSOX-4 nomenclature.

Appendix 5 *Probe and Primer Sets Used for ddPCR in Chapter 3.*

Name[1]	**Sequence**[2]
Gmnn For	TCATCTTCAGCGTTCTCCTTG
Gmnn Rev	GAATCTCAGTATGAAGCAGAAACAG
Gmnn Probe	/56-FAM/ACTCTTCAC/Zen/GTTCTCTTGGGCTCC/3IABkFQ/
Gmnn Probe Seq Only	ACTCTTCACGTTCTCTTGGGCTCC
Tubb2B For	TCGATACCATGCTCATCACTG
Tubb2B Rev	GCAAGAAGCTAACGAGGCA
Tubb2B Probe	/56-FAM/AGTGCGGCA/Zen/ACCAGATCGGT/3IABkFQ/
Tubb2B Probe Seq Only	AGTGCGGCAACCAGATCGGT
Aldh5a1 For	GATACACCCTATTCTGCCCTG
Aldh5a1 Rev	GGAGATTTTGGACACGAGGG
Aldh5a1 Probe	/56-FAM/AAGGAAGTG/Zen/GGAGAGGTGCTGTG/3IABkFQ/
Aldh5a1 Probe Seq Only	AAGGAAGTGGGAGAGGTGCTGTG
Gapdh For	GTAACCAGGCGTCCGATAC
Gapdh Rev	TCTCTGCTCCTCCCTGTTC
Gapdh Probe	/5HEX/CACACCGAC/Zen/CTTCACCATCTTGTCT/3IABkFQ/
Gapdh Probe Seq Only	CACACCGACCTTCACCATCTTGTCT

[1] Primers are indicated by the suffix ‘For’ and ‘Rev’ while probes are indicated by the suffix ‘Probe’. The probe is listed twice to indicate the modifications (‘Probe’) and with the modifications removed (‘Probe Seq Only’).

[2] Sequence are listed for both the primers and the probes in the 5’ to 3’ direction and were designed against the rat genome. For probes the modifications are: 5’ contains a 6-FAM fluorophore, an internal quencher Zen, and a 3’ quencher Iowa Black FQ.

Shading indicates probe and primer sets.

Appendix 6 *Genomic Information and Primers for Regions Assessed in Chapter 4.*

Name[1]	**Locus**	**Coordinates (hg18)**[2]	**Forward Primer**[3]	**Reverse Primer**[3]	**Size**[4]
CCS-01	*PAX7*	chr1:18854774-18854793	ATTCCAGTTTCCACGGTCAG	GCTCAAGTCTGATGCTGCAA	756
CCS-02	*ZEB2*	chr2:144876412-144876428	AGCCCAGTTTTTCCTGAGGT	GAGAACAGCTCATGTAAATTATTCCA	942
CCS-03	*ZEB2*	chr2:144901336-144901354	GCTCAATGTGTGAAAATGAAACA	TGGCTATTTGGACCAAGAAACT	940
CCS-04	*ZEB2*	chr2:144944487-144944503	GAGTTGCAAGCAACCTGTGA	TGAAGAACAACAGGCTTTGG	938
CCS-05	*ZEB2*	chr2:144974425-144974443	TATGGCAGCATTTGTTCAGC	GAGCAATCCTTCCATTTCCA	948
CCS-06	*ZEB2*	chr2:144989286-144989301	AAGCAATGGACAGGCTTGAT	TCCCCAAGATTCAGTTCAGG	807
CCS-07	*PAX3*	chr2:222845536-222845551	CGCCACTGTTTATCCCAG	GAGAAACCTGGCAAGGG	805
CCS-08	*SLIT2*	chr4:20089434-20089450	TGCCCCTTCATATGAGTAACC	CCTTGATGTCATGCAAATGG	953
CCS-09	*PPARGC1A*	chr4:23484867-23484883	CAACAAGAAAGCTTGCCAGAG	AGTGTGTGCCTGTGTATGTG	891
CCS-10	*SOX6*	chr11:16100023-16100041	CCAGTTTTCAGCTTACTTTGG	CTGGAAATAAGACAGGGTGG	787
CCS-11	*SOX6*	chr11:16268685-16268703	CGGTTACTACCCTCAGAATGGA	TTATTGGTGGCCAAAGCACT	999
CCS-12	*SOX6*	chr11:16273199-16273219	TGAAGTTGCCAGTTTTAATGC	GGGAGTTCTGTTTTGGGACA	987
CCS-13	*SOX6*	chr11:16334769-16334784	TGACACCTTCCCAAATCACA	TTCGTGCCAATGATGACCT	978
CCS-14	*SOX6*	chr11:16383201-16383216	CAACCAGGTTTCACCATCAA	CTGGCTGAGAGTGTTCTGGA	946
CCS-15	*SOX6*	chr11:16420069-16420089	GGTCAGCACCTCTCCAACAT	TTCCAGAGGCAGGTTTCATT	935
CCS-16	*HTATIP2*	chr11:20351219-20351236	TGTCTGTCCACATGGTTAGG	AGCAAGATTGATTGGAAGG	946
CCS-17	*NTM*	chr11:130816648-130816666	ACAGCTCTTTTTGGTCATGCAG	TTTTCTCCAGGCCTCCAGTG	752
CCS-18	*SOX5*	chr12:24059368-24059383	GACTCCTTAAATTCACAATCTGG	GGCCCTGCTACTTTATCAGC	885
CCS-19	*SOX5*	chr12:24059689-24059706	AAGCGAGTGTCGCCTAGGTA	TCCTCCCTCTGTGCTGTCTT	768
CCS-20	*SOX5*	chr12:24064859-24064874	CATTAACCAACCCCTGATGC	TCCATGCACTTCCTTTGTGT	953
CCS-21	*IGF1R*	chr15:97238999-97239016	TTCCTGGTAAACAGTTCTGCTG	CCCCAGTACTGTGAGCAACA	796
CCS-22	*TCF4*	chr18:51243898-51243915	TCTTAGCATGGGCCCTATC	GGGTTGTATCCATCTCAGAGC	750
CCS-23	*AKT3*	chr1:241943595-241943614	ACATGAATAAGGGAGAGAAGAGGA	TGTGCCTTAACTTAGAAACACTCC	1101
CCS-24	*FOXP1*	chr3:71182835-71182854	GCCACTCCCTTCCCAAACTC	CCTGGAGTCCTGTTGAGCAG	1239
CCS-25	*FOXP1*	chr3:71373671-71373689	GTCTGACTTAGGGGCGAGTG	TGCTTGTTCGAGACAGGTCA	791
CCS-26	*FOXP1*	chr3:71441367-71441382	ACACACTGTTGACTTCACAAGT	ACTGCATTGTGTAAATTTGCTGTG	319
CCS-27	*FOXP2*	chr7:113841891-113841906	AGTCAGTTCTTGCAATAGGAGG	CTTTGGTGTGCAACGTGAGG	993
CCS-28	*FOXP2*	chr7:113853298-113853316	CACAGCCAGGTTGTTTCTGC	CAAGATGTCCCTCTCTGCCA	1123
CCS-29	*FOXP2*	chr7:113860154-113860173	AGAAATGGGAAAATGTGGCATCT	ATGGACTAGGACACAAATGCTCA	592
CCS-30	*FOXP2*	chr7:113930143-113930163	AGAAACTGACAGTGTTTTGGAAGT	TGCTTGAGGAGAAAGGGGATC	875
CCS-31	*FOXP2*	chr7:114082514-114082532	AGACATGTATCTTTTTGAATCTGACA	TGGCACATTCAGAACCCAGA	1204
CCS-32	*LRPPRC*	chr2:44053319-44053338	TGTGGTTCCAAAACACTGGGT	TGGTCATTTTCTTTGTGGGCC	685
CCS-33	*NFIA*	chr1:61419393-61419408	CGGGGCTGGCATATAAGAGC	TCCATCTTACAGACTTTCACAATGA	530
CCS-34	*NFIA*	chr1:61482461-61482477	TGGGGTGTATGTGTATGCTGG	ACAGCTAAACCCCTAGCCCT	463
CCS-35	*NFIA*	chr1:61686010-61686028	CACCCAGAAAATCCGGCAGT	TTCTGGAGCCGCTTATGACG	306

CCS-36	*ROBO2*	chr3:77704023-77704038	CACTGAAGTGTGCAAGTGTGC	TGAAATAAGGCAACCAAGAGGC	391
CCS-37	*ST18*	chr8:53370592-53370610	TACCTCTAAGGAGCCTGCCA	AGGGGGAAGTCAGAGATATGTCA	553
CCS-38	*TCF7L2*	chr10:114809423-114809439	GCTTTCAAGGCTGGACCACT	AGGAGAAAACAATCTGCTCTTTTCC	557
CCS-39	*TCF7L2*	chr10:114895622-114895642	TGAACATGAGCTTGTGACCCA	GGGGTGTCTGAATCCTCCTG	829
CCS-40	*ZFP536*	chr19:35553939-35553959	ATCCAGGCAAAACAGAGGGG	ATACAGCAGGGAGGCAGATG	1106
CCS-41	*ZFP536*	chr19:35584775-35584790	TTTGTGGGTGGTAGGTGTGT	GCTGGGAGAGGTAGAACAGG	833
CCS-42	*BCAS3*	chr17:56264318-56264334	GTCATTTGTCAAACGAAGCAGC	GCACACTTTAAGATCCAAATTCTCC	685
CCS-43	*BCAS3*	chr17:56684299-56684319	AGACTGCTAGGTTCCCAGCT	CTCTGGAGCCCTGGGTTATG	753
CCS-44	*CELF4*	chr18:33340240-33340259	TGCCTTCGTGTCTTGAAGCC	CCATGGGCTTGACCTACAGG	246
CCS-45	*CNTLN*	chr9:17440492-17440507	CAGATTGGCATTTTCAGACCCA	TCTGAAAAATCCACTGAGTTACTGC	636
CCS-46	*EHBP1*	chr2:63084089-63084109	TCCTACAAGTTGCATTCTGAACT	CAGCATCAAGATGGTATTGTCTCAC	951
CCS-47	*EHBP1*	chr2:63115438-63115454	ATGGCTTTCAATATTGTATGTCTTGAA	AGGACACATTACTCATTGCTTCAC	1016
CCS-48	*HAT1*	chr2:172529073-172529092	TGTGAATGAGTTGCAAGGACTG	GTGACACAATTCTTACAGACCTGG	889
CCS-49	*LRBA*	chr4:151496934-151496949	CCATGTAATACGGCCTTCTTCC	TGCTAAAGTAACTCAGATTCACTGC	549
CCS-50	*NFIB*	chr9:14293789-14293808	CCCAAGAATCATTGGACGTCT	ATGTCTCCCTGCACTTCACC	477
CCS-51	*NFIB*	chr9:14299587-14299605	GGAAGGAGTACATGTCCCATCC	GGAAGTGAGTTTCCAAAGCACA	465
CCS-52	*NFIB*	chr9:14302131-14302147	CCAGCCGATGGGTAATATTAATGG	AAGTGTCAGCCAGTCTTGGG	454
CCS-53	*POLA1*	chrX:24774936-24774951	CCCTGGTCCTTGTTGGTTCC	TGTGGCTGCTTCTTGGATGG	680
CCS-54	*SORBS2*	chr4:186930629-186930648	TGCTTGCAATGTTCCCTTGG	GTTTGTAGCCGTGGGATCGA	331
CCS-55	*TLE4*	chr9:81473523-81473540	TGACAGGCATGACGTTGAGG	ACAATCCTAAGCCAGGGAGAC	428
CCS-56	*ZFHX3*	chr16:71426880-71426895	GGAGGGTGGGATGTTTGAGG	TTTCCCACCTGCTTCAGTGG	477
CCS-57	*MPP7*	chr10:28530972-28530987	ATACAGAGCCAGCTCACCAC	TTGGCATGTTCCAGCTGTCA	416

[1] The SOX10- has been omitted for clarity. Full names of regions are SOX10-CCS-(number).

[2] Regions are given in bed format for the human genome (hg18) and encompass the SOX10 consensus sequence and intermonomeric sequences identical among human (hg18), mouse (mm9), and chicken (Gal3).

[3] Primers sequences are displayed in the 5' to 3' orientation and do not include the gateway adapter sequences. The adapter sequences for the forward and reverse primers are 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3' and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGT-3' respectively.

[4] Sizes are given in base pairs and do not reflect the additional size of the gateway adapter sequences. An additional 58 base pairs should be added to the size to account for the gateway adapters.

Appendix 7 *Primers Used in RT-PCR in Chapter 4.*

Name	**Sequence**[1]	**Size**[3]
rnmmB-Actin RTPCR_F[2]	CGCGGGCGACGATGCTCC	532 - Rat
rnmmB-Actin RTPCR_R[2]	GTAGCCACGCTCGGTCAGG	532 - Mouse
rnSOX6_RTPCR_Fwd[2]	GTGCTGTATCTCCCCACAGG	402 - Rat
rnSOX6_RTPCR_Rev[2]	TGGGTCATTGTTTCCTCTCC	349 - Mouse
rnSox6_RTPCR_F3[4]	GCCAGGAGTCTTCACTGCTCC	3062
rnSox6_3'UTR_R2[4]	GGGAGCGAAATGTCAGAGTG	

[1] Primers written in the 5’-3’ directions and designed against the rat (rn5) genome.

[2] Both primer sets were designed to amplify both the rat (rn5) and mouse (mm9) sequences.

[3] Size is written in base pairs and are displayed for either the rat or mouse genome.

[4] Primer set was designed to amplify full length transcript starting within exon 1G from S16 cells.

Please note that while the same primers were used for the SOX6 RT-PCR reaction, the rat (S16) PCR product was 402 bp and the mouse (MN-1) PCR product was 349 bp; the rat genome harbors a 53 base pair rat-specific insertion, which was confirmed via DNA sequence analysis.

Shading indicates primer sets used.

Appendix 8 *Primers Used in 5’RACE in Chapter 4*.

Name	**Primer[1]**
rnSOX6_GSP1	GATTTCTCCAAGAAGTTCACTCG
rnSOX6_GSP2	TCTCCTCCAGCTTCTTCTGC
rnSOX6_GSP3	TGGGTCATTGTTTCCTCTCC

[1]Primers written in the 5’-3’ direction and designed against the rat (rn5) genome.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of Ultraconserved Elements Yields Viable Mice. PLoS Biol 5: 1906–1911.

Ahmed BY, Chakravarthy S, Eggers R, Hermens WTJMC, Zhang JY, Niclou SP, Levelt C, Sablitzky F, Anderson PN, Lieberman AR, Verhaagen J. 2004. Efficient delivery of Cre-recombinase to neurons in vivo and stable transduction of neurons using adeno-associated and lentiviral vectors. BMC Neurosci 5:.

Akerfelt M, Henriksson E, Laiho A, Vihervaara A, Rautoma K, Kotaja N, Sistonen L. 2008. Promoter ChIP-chip analysis in mouse testis reveals Y chromosome occupancy by HSF2. PNAS 105: 11224–11229.

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Amin H, Vachris J, Hamilton A, Steuerwald N, Howden R, Arthur ST. 2014. GSK3β inhibition and LEF1 upregulation in skeletal muscle following a bout of downhill running. J Physiol Sci 64: 1–11.

Anders S, Pyl PT, Huber W. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

Anido CL, Sun G, Koenning M, Srinivasan R. 2015. Differential Sox10 genomic occupancy in myelinating glia. Glia.

Antonellis A, Bennett WR, Prasad, AB, Lee-Lin SQ, Green ED, Paisley D, Kelsh RN, Pavan WJ, Ward A, Sequencing NC. 2006. Deletion of long-range sequences at Sox10 compromises developmental expression in a mouse model of Waardenburg-Shah (WS4) syndrome. Human Molecular Genetics 15: 259–271.

Antonellis A, Dennis MY, Burzynski G, Huynh J, Maduro V, Hodonsky CJ, Khajavi M, Szigeti K, Mukkamala S, Bessling SL, Pavan WJ, McCallion AS, et al. 2010. A Rare Myelin Protein Zero (MPZ) Variant Alters Enhancer Activity In Vitro and In Vivo. PLoS ONE 5:.

Antonellis A, Huynh JL, Lee-Lin S-Q, Vinton RM, Renaud G, Loftus SK, Elliot G, Wolfsberg TG, Green ED, McCallion AS, Pavan WJ. 2008. Identification of Neural Crest and Glial Enhancers at the Mouse Sox10 Locus through Transgenesis in Zebrafish. PLoS Genet 4:.

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S.

1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci USA 92: 1684–1688.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339: 1074–1077.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

Balemans W, Patel N, Ebeling M, Van Hul E, Wuyts W, Lacza C, Dioszegi M, Dikkers FG, Hildering P, Willems PJ, Verheij JBGM, Lindpaintner K, et al. 2002. Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease. J. Med. Genet. 39: 91–97.

Banerji J, Olson L, Schaffner W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33: 729–740.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27: 299–308.

Bao L, Odell AF, Stephen SL, Wheatcroft SB, Walker JH, Ponnambalam S. 2012. The S100A6 calcium-binding protein regulates endothelial cell-cycle progression and senescence. FEBS Journal 279: 4576–4588.

Baroti T, Schillinger A, Wegner M, Claus Stolt C. 2015. Sox13 functionally complements the related Sox5 and Sox6 as important developmental modulators in mouse spinal cord oligodendrocytes. J. Neurochem.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.

Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. 2011. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 18: 107–114.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304: 1321–1325.

Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE. 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 5:.

Billin AN, Thirlwell H, Ayer DE. 2000. Beta-catenin-histone deacetylase interactions regulate the transition of LEF1 from a transcriptional repressor to an activator. Molecular and Cellular Biology 20: 6882–6890.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. PNAS 104: 3312–3317.

Bondurand N, Girard M, Pingault V, Lemort N, Dubourg O, Goossens M. 2001. Human Connexin 32, a gap junction protein altered in the X-linked form of Charcot–Marie–Tooth disease, is directly regulated by the transcription factor SOX10. Human Molecular Genetics 10: 2783–2795.

Bondurand N, Pingault V, Goerich DE, Lemort N, Sock E, Le Caignec C, Wegner M, Goossens M. 2000. Interaction among SOX10 PAX3 and MITF, three genes altered in Waardenburg syndrome. Human Molecular Genetics 9: 1907–1917.

Bosma PJ, Chowdhury JR, Bakker C, Gantla S, Deboer A, Oostra BA, Lindhout D, Tytgat G, Jansen P, Elferink R, Chowdhury NR. 1995. The Genetic-Basis of the Reduced Expression of Bilirubin Udp-Glucuronosyltransferase-1 in Gilberts-Syndrome. N. Engl. J. Med. 333: 1171–1175.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, Liu ET. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research 18: 1752–1762.

Bowles J, Schepers G, Koopman P. 2000. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Dev. Biol. 227: 239–255.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. 24: 2537–2538.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22: 1790–1797.

Bremer M, Froeb F, Kichko T, Reeh P, Tamm ER, Suter U, Wegner M. 2011. Sox10 Is Required for Schwann-Cell Homeostasis and Myelin Maintenance in the Adult Peripheral Nerve. Glia 59: 1022–1032.

Breuss M, Morandell J, Nimpf S, Gstrein T, Lauwers M, Hochstoeger T, Braun A, Chan K, Sánchez Guajardo ER, Zhang L, Suplata M, Heinze KG, et al. 2015. The expression of tubb2bundergoes a developmental transition in murine cortical neurons. J. Comp. Neurol. 523: 2161–2186.

Brewer MH, Ma KH, Beecham GW, Gopinath C, Baas F, Choi B-O, Reilly MM, Shy ME, Zuechner S, Svaren J, Antonellis A, INC. 2014. Haplotype-specific modulation of a SOX10/CREB response element at the Charcot Marie Tooth disease type 4C locus SH3TC2.

Human Molecular Genetics 23: 5171–5187.

Britsch S, Goerich DE, Riethmacher D, Peirano RI, Rossner M, Nave KA, Birchmeier C, Wegner M. 2001. The transcription factor Sox10 is a key regulator of peripheral glial development. Genes Dev 15: 66–78.

Britsch S, Li L, Kirchhoff S, Theuring F, Brinkmann V, Birchmeier C, Riethmacher D. 1998. The ErbB2 and ErbB3 receptors and their ligand, neuregulin-1, are essential for development of the sympathetic nervous system. Genes Dev 12: 1825–1836.

Bronson SK, Plaehn EG, Kluckman KD, Hagaman JR, Maeda N, Smithies O. 1996. Single-copy transgenic mice with chosen-site integration. Proc Natl Acad Sci USA 93: 9067–9072.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth 10: 1213–1218.

Carlberg K, Ryden TA, Beemon K. 1988. Localization and Footprinting of an Enhancer Within the Avian-Sarcoma Virus Gag Gene. J Virol 62: 1617–1624.

Chen Y, Wang H, Yoon SO, Xu X, Hottiger MO, Svaren J, Nave KA, Kim HA, Olson EN, Lu QR. 2011. HDAC-mediated deacetylation of NF-κB is critical for Schwann cell myelination. Nat. Neurosci. 14: 437–441.

Chorley B, Wang X, Campbell M, Pittman G, Noureddine M, Bell D. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. Mutation Research/Reviews in Mutation Research 659: 147–157.

Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, Bouffard G, Young A, Masiello C, Green ED, Wolfsberg TG, Collins FS. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci USA 101: 992–997.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. PNAS 107: 21931–21936.

D'Antonio M, Michalovich D, Paterson M, Droggiti A, Woodhoo A, Mirsky R, Jessen KR. 2006. Gene profiling and bioinformatic analysis of Schwann cell embryonic development and myelination. Glia 53: 501–515.

Dai J, Cui X, Zhu Z, Hu W. 2010. Non-Homologous End Joining Plays a Key Role in Transgene Concatemer Formation in Transgenic Zebrafish Embryos. Int J Biol Sci 6: 756–768.

de Kok YJ, van der Maarel SM, Bitner-Glindzicz M, Huber I, Monaco AP, Malcolm S, Pembrey ME, Ropers HH, Cremers FP. 1995. Association between X-linked mixed deafness and mutations in the POU domain gene POU3F4. Science 267: 685–688.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. Science 295: 1306–1311.

deKok Y, Vossenaar ER, Cremers C, Dahl N, Laporte J, Hu LJ, Lacombe D, FischelGhodsian N, Friedman RA, Parnes LS, Thorpe P, BitnerGlindzicz M, et al. 1996. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. Human Molecular Genetics 5: 1229–1235.

DeKoter RP, Singh H. 2000. Regulation of B lymphocyte and macrophage development by graded expression of PU.1. Science 288: 1439–1441.

Dewet JR, Wood KV, Helinski DR, DeLuca M. 1985. Cloning of Firefly Luciferase Cdna and the Expression of Active Luciferase in Escherichia-Coli. PNAS 82: 7870–7873.

DiLeone RJ, Marcus GA, Johnson MD, Kingsley DM. 2000. Efficient studies of long-distance Bmp5 gene regulation using bacterial artificial chromosomes. Proc Natl Acad Sci USA 97: 1612–1617.

DiLeone RJ, Russell LB, Kingsley DM. 1998. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. Genetics 148: 401–408.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Research 16: 1299–1309.

Driscoll MC, Dobkin CS. 1989. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5'beta-globin gene activation-region hypersensitive sites. In: Proceedings of the ….

Dunn TM, Hahn S, Ogden S, Schleif RF. 1984. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. PNAS 81: 5017–5020.

Dyck PJ, Lambert EH. 1968. Lower Motor and Primary Sensory Neuron Diseases with Peroneal Muscular Atrophy .I. Neurologic Genetic and Electrophysiologic Findings in Hereditary Polyneuropathies. Archives of Neurology 18: 603–618.

Emison ES, Garcia-Barcelo M, Grice EA, Lantieri F, Amiel J, Burzynski G, Fernandez RM, Hao L, Kashuk C, West K, Miao X, Tam PKH, et al. 2010. Differential Contributions of Rare and Common, Coding and Noncoding Ret Mutations to Multifactorial Hirschsprung Disease Liability. The American Journal of Human Genetics 87: 60–74.

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. 2005. A common sex-dependent mutation in a RET enhancer underlies

Hirschsprung disease risk. Nature 434: 857–863.

Emmett LSD, O'Shea KS. 2012. Geminin Is Required for Epithelial to Mesenchymal Transition at Gastrulation. Stem Cells Dev. 21: 2395–2409.

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

Erickson CA, Reedy MV. 1998. Neural crest development: The interplay between morphogenesis and cell differentiation. Growth Factors in Development 40: 177–209.

Feigelson HS, Shames LS, Pike MC, Coetzee GA, Stanczyk FZ, Henderson BE. 1998. Cytochrome P450c17alpha gene (CYP17) polymorphism is associated with serum estrogen and progesterone concentrations. Cancer Res. 58: 585–587.

Fernandez PC, Frank SR, Wang LQ, Schroeder M, Liu SX, Greene J, Cocito A, Amati B. 2003. Genomic targets of the human c-Myc protein. Genes Dev 17: 1115–1129.

Ferretti E, Cambronero F, Tümpel S, Longobardi E, Wiedemann LM, Blasi F, Krumlauf R. 2005. Hoxb1 enhancer and control of rhombomere 4 expression: complex interplay between PREP1-PBX1-HOXB1 binding sites. Molecular and Cellular Biology 25: 8541–8552.

Finzsch M, Schreiner S, Kichko T, Reeh P, Tamm ER, Boesl MR, Meijer D, Wegner M. 2010. Sox10 is required for Schwann cell identity and progression beyond the immature Schwann cell stage. J Cell Biol 189: 701–712.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312: 276–279.

Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, Bickel PJ, Biggin MD, et al. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. PNAS 109: 21330–21335.

Fraser J, Rousseau M, Shenker S, Ferraiuolo MA, Hayashizaki Y, Blanchette M, Dostie J. 2009. Chromatin conformation signatures of cellular differentiation. Genome Biol 10:.

Frech K, Quandt K, Werner T. 1997. Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem. Sci. 22: 103–104.

Fried M, Crothers DM. 1981. Equilibria and Kinetics of Lac Repressor-Operator Interactions by Polyacrylamide-Gel Electrophoresis. Nucleic Acids Res 9: 6505–6525.

Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature Biotechnology 31: 822–826.

Fuchs T, Gavarini S, Saunders-Pullman R, Raymond D, Ehrlich ME, Bressman SB, Ozelius LJ.

2009. Mutations in the THAP1 gene are responsible for DYT6 primary torsion dystonia. Nat Genet 41: 286–288.

Galas DJ, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157–3170.

Garcia CA, Malamut RE, England JD, Parry GS, Liu P, Lupski JR. 1995. Clinical Variability in 2 Pairs of Identical-Twins with the Chareot-Marie-Tooth Disease Type 1a Duplication. Neurology 45: 2090–2093.

Garner MM, Revzin A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 9: 3047–3060.

Gheldof N, Smith EM, Tabuchi TM, Koch CM, Dunham I, Stamatoyannopoulos JA, Dekker J. 2010. Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res 38: 4325–4336.

Ghislain J, Charnay P. 2006. Control of myelination in Schwann cells: a Krox20 cis-regulatory element integrates Oct6, Brn2 and Sox10 activities. EMBO Rep 7: 52–58.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Meth 6: 343–345.

Giese K, Cox J, Grosschedl R. 1992. The Hmg Domain of Lymphoid Enhancer Factor-I Bends Dna and Facilitates Assembly of Functional Nucleoprotein Structures. Cell 69: 185–195.

Gille J, Swerlick RA, Caughman SW. 1997. Transforming growth factor-alpha-induced transcriptional activation of the vascular permeability factor (VPF/VEGF) gene requires AP-2-dependent DNA binding and transactivation. EMBO J 16: 750–759.

Gilmour DS, Lis JT. 1985. In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. Molecular and Cellular Biology 5: 2009–2018.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Research 17: 877–885.

Gobbi MD, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, Jong P de, Cheng J-F, Rubin EM, et al. 2006. A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. Science 312: 1215–1217.

Goda S, Hammer J, Kobiler D, Quarles RH. 1991. Expression of the Myelin-Associated Glycoprotein in Cultures of Immortalized Schwann Cells. J. Neurochem. 56: 1354–1361.

Gokey NG, Srinivasan R, Lopez-Anido C, Krueger C, Svaren J. 2012. Developmental Regulation of MicroRNA Expression in Schwann Cells. Molecular and Cellular Biology 32: 558–568.

Goode DK, Snell P, Elgar G. 2003. Comparative analysis of vertebrate Shh genes identifies novel conserved non-coding sequence. Mamm Genome 14: 192–201.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. Genome Biology and Evolution 5: 578–590.

Griffith J, Hochschild A, Ptashne M. 1986. DNA loops induced by cooperative binding of lambda repressor. Nature 322: 750–752.

Grosveld F, Vanassendelft GB, Greaves DR, Kollias G. 1987. Position-Independent, High-Level Expression of the Human Beta-Globin Gene in Transgenic Mice. Cell 51: 975–985.

Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic Drosophila by using the site-specific integrase from phage phi C31. Genetics 166: 1775–1782.

Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Munsterberg A, Vivian N, Goodfellow P, Lovellbadge R. 1990. A Gene-Mapping to the Sex-Determining Region of the Mouse Y-Chromosome Is a Member of a Novel Family of Embryonically Expressed Genes. Nature 346: 245–250.

Guillot PV, Liu L, Kuivenhoven JA, Guan J, Rosenberg RD, Aird WC. 2000. Targeting of human eNOS promoter to the Hprt locus of mice leads to tissue-restricted transgene expression. Physiological Genomics 2: 77–83.

Guo JY, Xu J, Mao DQ, Fu LL, Gu JR, De Zhu J. 2002. The promoter analysis of the human C17orf25 gene, a novel chromosome 17p13.3 gene. Nature Publishing Group 12: 339–352.

Hagiwara N. 2011. Sox6, jack of all trades: a versatile regulatory protein in vertebrate development. Dev Dyn 240: 1311–1321.

Hai M, Muja N, DeVries GH, Quarles RH, Patel PI. 2002. Comparative analysis of Schwann cell lines as model systems for myelin gene transcription studies. J. Neurosci. Res. 69: 497–508.

Hamada T, Sasaki H, Seki R, Sakaki Y. 1993. Mechanism of chromosomal integration of transgenes in microinjected mouse eggs: sequence analysis of genome-transgene and transgene-transgene junctions at two loci. Gene 128: 197–202.

Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 16: 369–372.

Harley VR, Jackson DI, Hextall PJ, Hawkins JR, Berkovitz GD, Sockanathan S, Lovellbadge R, Goodfellow PN. 1992. Dna-Binding Activity of Recombinant Sry From Normal Males and Xy Females. Science 255: 453–456.

Harley VR, Lovell-Badge R, Goodfellow PN. 1994. Definition of a consensus DNA binding site for SRY. Nucleic Acids Res 22: 1500–1501.

Hayasaka K, Himoro M, Sato W, Takada G, Uyemura K, Shimizu N, Bird TD, Conneally PM, Chance PF. 1993. Charcot-Marie-Tooth neuropathy type 1B is associated with mutations of the myelin P0 gene. Nat Genet 5: 31–34.

He Y, Kim JY, Dupree J, Tewari A, Melendez-Vasquez C, Svaren J, Casaccia P. 2010. Yy1 as a molecular link between neuregulin and transcriptional modulation of peripheral myelination. Nat. Neurosci. 13: 1472–1480.

Hebbar PB, Archer TK. 2008. Altered histone H1 stoichiometry and an absence of nucleosome positioning on transfected DNA. Journal of Biological Chemistry 283: 4595–4601.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38: 576–589.

Herbarth B, Pingault V, Bondurand N, Kuhlbrodt K, Hermans-Borgmeyer I, Puliti A, Lemort N, Goossens M, Wegner M. 1998. Mutation of the Sry-related Sox10 gene in Dominant megacolon, a mouse model for human Hirschsprung disease. PNAS 95: 5161–5165.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen M, Delany ME, Dodgson JB, Chinwalla AT, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716.

Hodonsky CJ, Kleinbrink EL, Charney KN, Prasad M, Bessling SL, Jones EA, Srinivasan R, Svaren J, McCallion AS, Antonellis A. 2012. SOX10 regulates expression of the SH3-domain kinase binding protein 1 (Sh3kbp1) locus in Schwann cells via an alternative promoter. Molecular and Cellular Neuroscience 49: 85–96.

Houlden H, Girard M, Cockerell C, Ingram D, Wood NW, Goossens M, Walker RWH, Reilly MM. 2004. Connexin 32 promoter P2 mutations: A mechanism of peripheral nerve dysfunction. Ann Neurol 56: 730–734.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498–503.

Hua S, Kittler R, White KP. 2009. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. Cell 137: 1259–1271.

Hung H, Kohnken R, Svaren J. 2012. The nucleosome remodeling and deacetylase chromatin remodeling (NuRD) complex is required for peripheral nerve myelination. Journal of Neuroscience 32: 1517–1527.

Hung HA, Sun G, Keles S, Svaren J. 2015. Dynamic Regulation of Schwann Cell Enhancers after Peripheral Nerve Injury. Journal of Biological Chemistry 290: 6937–6950.

Hunt T, Jackson RJ. 1974. The rabbit reticulocyte lysate as a system for studying mRNA. Hamatol. Bluttransfus. 14: 300–307.

Huxley C, Passage E, Robertson AM, Youl B, Huston S, Manson A, Saberan-Djoniedi D, Figarella-Branger D, Pellissier JF, Thomas PK, Fontes M. 1998. Correlation between varying levels of PMP22 expression and the degree of demyelination and reduction in nerve conduction velocity in transgenic mice. Human Molecular Genetics 7: 449–458.

Inoue K, Khajavi M, Ohyama T, Hirabayashi S-I, Wilson J, Reggin JD, Mancias P, Butler IJ, Wilkinson MF, Wegner M, Lupski JR. 2004. Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. Nat Genet 36: 361–369.

Inoue K, Shilo K, Boerkoel CF, Crowe C, Sawady J, Lupski JR, Agamanolis DP. 2002. Congenital hypomyelinating neuropathy, central dysmyelination, and Waardenburg-Hirschsprung disease: Phenotypes linked by SOX10 mutation. Ann Neurol 52: 836–842.

Inoue K, Tanabe Y, Lupski JR. 1999. Myelin deficiencies in both the central and the peripheral nervous systems associated with a SOX10 mutation. Ann Neurol 46: 313–318.

Ionasescu V, Searby C. 1994. Point mutations of the connexin32 (GJB1) gene in X-linked dominant Charcot—Marie—Tooth neuropathy. Human Molecular Genetics.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. Cell 141: 1253–1261.

Ito Y, Wiese S, Funk N, Chittka A, Rossoll W, Bommel H, Watabe K, Wegner M, Sendtner M. 2006. Sox10 regulates ciliary neurotrophic factor gene expression in Schwann cells. PNAS 103: 7871–7876.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533–538.

Jaegle M, Meijer D. 1998. Role of Oct-6 in Schwann cell differentiation. Microsc. Res. Tech. 41: 372–378.

Jagalur NB, Ghazvini M, Mandemakers W, Driegen S, Maas A, Jones EA, Jaegle M, Grosveld F, Svaren J, Meijer D. 2011. Functional Dissection of the Oct6 Schwann Cell Enhancer Reveals an Essential Role for Dimeric Sox10 Binding. Journal of Neuroscience 31: 8585–8594.

Jaglin XH, Poirier K, Saillour Y, Buhler E, Tian G, Bahi-Buisson N, Fallet-Bianco C, Phan-Dinh-Tuy F, Kong XP, Bomont P, Castelnau-Ptakhine L, Odent S, et al. 2009. Mutations in the beta-tubulin gene TUBB2B result in asymmetrical polymicrogyria. Nat Genet 41: 746–752.

Jarriault S, Brou C, Logeat F, Schroeter EH, Kopan R, Israel A. 1995. Signalling downstream of activated mammalian Notch. Nature 377: 355–358.

Jessen KR, Mirsky R. 2005. The origin and development of glial cells in peripheral nerves. Nat. Rev. Neurosci. 6: 671–682.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, et al. 2013. DNA-Binding Specificities of Human Transcription Factors. Cell 152: 327–339.

Jones EA, Brewer MH, Srinivasan R, Krueger C, Sun G, Charney KN, Keles S, Antonellis A, Svaren J. 2011a. Distal enhancers upstream of the Charcot-Marie-Tooth type 1A disease gene PMP22. Human Molecular Genetics 21: 1–11.

Jones EA, Lopez-Anido C, Srinivasan R, Krueger C, Chang L-W, Nagarajan R, Svaren J. 2011b. Regulation of the PMP22 Gene through an Intronic Enhancer. Journal of Neuroscience 31: 4242–4250.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: 493–496.

Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576–3579.

Kellerer S, Schreiner S, Stolt CC, Scholz S, Bösl MR, Wegner M. 2006. Replacement of the Sox10 transcription factor by Sox8 reveals incomplete functional equivalence. Development 133: 2875–2886.

Kelsh RN. 2006. Sorting out Sox10 functions in neural crest development. Bioessays 28: 788–798.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler AD. 2002. The Human Genome Browser at UCSC. Genome Research 12: 996–1006.

Kim TH, Barrera LO, Zheng M, Qu CX, Singer MA, Richmond TA, Wu YN, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. Nature 436: 876–880.

Kimura-Yoshida C, Kitajima K, Oda-Ishii I, Tian E, Suzuki M, Yamamoto M, Suzuki T, Kobayashi M, Aizawa S, Matsuo I. 2004. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. Development 131: 57–71.

Kioussis D, Vanin E, deLange T, Flavell RA, Grosveld FG. 1983. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature 306: 662–666.

Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. 2015. Massively parallel single-amino-acid mutagenesis. Nat Meth 12: 203–206.

Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, Marsh K, Kwiatkowski D. 1999. A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. Nat Genet 22: 145–150.

Kobolak J, Kiss K, Polgar Z, Mamo S, Rogel-Gaillard C, Tancos Z, Bock I, Baji AG, Tar K, Pirity MK, Dinnyes A. 2009. Promoter analysis of the rabbit POU5F1 gene and its expression in preimplantation stage embryos. BMC Mol Biol 10:.

Kondo T, Raff M. 2000. The Id4 HLH protein and the timing of oligodendrocyte differentiation. EMBO J 19: 1998–2007.

Kong XF, Zhu XH, Pei YL, Jackson DM, Holick MF. 1999. Molecular Cloning, Characterization, and Promoter Analysis of the Human 25-Hydroxyvitamin D_3-1α -Hydroxylase Gene. Proc Natl Acad Sci USA 96: 6988–6993.

Kristie TM, Roizman B. 1986. Alpha-4, the Major Regulatory Protein of Herpes-Simplex Virus Type-1, Is Stably and Specifically Associated with Promoter-Regulatory Domains of Alpha-Genes and of Selected Other Viral Genes. PNAS 83: 3218–3222.

Kuhlbrodt K, Herbarth B, Sock E, Hermans-Borgmeyer I, Wegner M. 1998. Sox10, a Novel Transcriptional Modulator in Glial Cells. The Journal of Neuroscience 18: 237–250.

Kulkens T, Bolhuis PA, Wolterman RA, Kemp S. 1993. Deletion of the serine 34 codon from the major peripheral myelin protein P0 gene in Charcot–Marie–Tooth disease type 1B. Nature.

Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A. 2014. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 512:.

Kvon EZ, Stampfel G, Yanez-Cuna JO, Dickson BJ, Stark A. 2012. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev 26: 908–913.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. Genome Research 24: 1595–1602.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci USA 109: 19498–19503.

Lan X, Witt H, Katsumura K, Ye Z, Wang Q, Bresnick EH, Farnham PJ, Jin VX. 2012. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res 40: 7690–7704.

Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22: 1813–1831.

Lane PW, Liu HM. 1984. Association of megacolon with a new dominant spotting gene (Dom) in the mouse. J Hered 75: 435–439.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:.

Laudet V, Stehelin D, Clevers H. 1993. Ancestry and diversity of the HMG box superfamily. Nucleic Acids Res.

Le N, Nagarajan R, Wang JYT, Svaren J, LaPash C, Araki T, Schmidt RE, Milbrandt J. 2005. Nab proteins are essential for peripheral nervous system myelination. Nat. Neurosci. 8: 932–940.

LeBlanc SE, Jang S-W, Ward RM, Wrabetz L, Svaren J. 2006. Direct regulation of myelin protein zero expression by the Egr2 transactivator. Journal of Biological Chemistry 281: 5453–5460.

Lee KE, Nam S, Cho E-A, Seong I, Limb J-K, Lee S, Kim J. 2008. Identification of direct regulatory targets of the transcription factor Sox10 based on function and conservation. BMC Genomics 9:.

Leonid TB, Thurtl DM, Rine J, van Oudenaarden A. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. PNAS 110: 18602–18607.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Ben A Oostra, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 12: 1725–1735.

Lettice LA, Horikoshi T, Heaney S, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, et al. 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci USA 99: 7548–7553.

Levings PP, Bungert J. 2002. The human beta-globin locus control region - A center of attraction. Eur J Biochem 269: 1589–1599.

Levis R, Hazelrigg T, Rubin GM. 1985. Effects of Genomic Position on the Expression of Transduced Copies of the White Gene of Drosophila. Science 229: 558–561.

Li H, Lu Y, Smith HK, Richardson WD. 2007. Olig1 and Sox10 interact synergistically to drive Myelin Basic Protein transcription in oligodendrocytes. Journal of Neuroscience 27: 14375–14382.

Li XL, Noll M. 1994. Compatibility Between Enhancers and Promoters Determines the Transcriptional Specificity of Gooseberry and Gooseberry Neuro in the Drosophila Embryo. EMBO J 13: 400–406.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288: 136–140.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.

Lupski JR, de Oca-Luna RM, Slaugenhaupt S. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell.

Maeda MH, Mitsui J, Soong B-W, Takahashi Y, Ishiura H, Hayashi S, Shirota Y, Ichikawa Y, Matsumoto H, Arai M, Okamoto T, Miyama S, et al. 2012. Increased gene dosage of myelin protein zero causes Charcot-Marie-Tooth disease. Ann Neurol 71: 84–92.

Mager GM, Ward RM, Srinivasan R, Jang S-W, Wrabetz L, Svaren J. 2008. Active gene repression by the Egr2.NAB complex during peripheral nerve myelination. Journal of Biological Chemistry 283: 18187–18197.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-Guided Human Genome Engineering via Cas9. Science 339: 823–826.

Malicki J, Cianetti LC, Peschle C, McGinnis W. 1992. A Human Hox4b Regulatory Element Provides Head-Specific Expression in Drosophila Embryos. Nature 358: 345–347.

Mandemakers W, Zwart R, Jaegle M, Walbeehm E, Visser P, Grosveld F, Meijer D. 2000. A distal Schwann cell-specific enhancer mediates axonal regulation of the Oct-6 transcription factor during peripheral nerve development and regeneration. EMBO J 19: 2992–3003.

Mansour SL, Thomas KR, Deng CX, Capecchi MR. 1990. Introduction of a Lacz Reporter Gene Into the Mouse Int-2 Locus by Homologous Recombination. Proc Natl Acad Sci USA 87: 7688–7692.

Mao CD, Byers SW. 2011. Cell-Context Dependent TCF/LEF Expression and Function: Alternative Tales of Repression, De-Repression and Activation Potentials. Crit. Rev. Eukaryot. Gene Expr. 21: 207–236.

Marengo-Rowe AJ. 2007. The thalassemias and related disorders. Proceedings (Baylor University. Medical Center) 20: 27–31.

Marin-Husstege M, He Y, Li J, Kondo T, Sablitzky F, Casaccia-Bonnefil P. 2006. Multiple roles of Id4 in developmental myelination: Predicted outcomes and unexpected findings. Glia 54: 285–296.

Martin JR, Webster H. 1973. Mitotic Schwann Cells in Developing Nerve - Their Changes in Shape, Fine-Structure, and Axon Relationships. Dev. Biol. 32: 417–431.

Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, Cooper NJ, Barton A, Wallace C, Fraser P, Worthington J, Eyre S. 2015. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun 6: 10069–10069.

Matthews JC, Hori K, Cormier MJ. 1977. Purification and Properties of Renilla-Reniformis Luciferase. Biochemistry 16: 85–91.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos D-U, Land S, et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195.

McGarry TJ, Kirschner MW. 1998. Geminin, an inhibitor of DNA replication, is degraded during mitosis. Cell 93: 1043–1053.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Research 18: 252–260.

McNabb DS, Reed R, Marciniak RA. 2005. Dual luciferase assay system for rapid assessment of gene expression in Saccharomyces cerevisiae. Eukaryot Cell 4: 1539–1549.

Meyer A, Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol 11: 699–704.

Mort RL, Jackson IJ, Patton EE. 2015. The melanocyte lineage in development and disease. Development 142: 620–632.

Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat Genet 36: 889–893.

Nagarajan R, Svaren J, Le N, Araki T, Watson M, Milbrandt J. 2001. EGR2 mutations in inherited neuropathies dominant-negatively inhibit myelin gene expression. Neuron 30: 355–368.

Nasrin N, Buggs C, Kong XF, Carnazza J, Goebl M, Alexander-Bridges M. 1991. DNA-binding properties of the product of the testis-determining gene and a related protein. Nature 354: 317–320.

Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Research 22: 1711–1722.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83–90.

Nishimura S, Takahashi S, Kuroha T, Suwabe N, Nagasawa T, Trainor C, Yamamoto M. 2000. A GATA box in the GATA-1 gene hematopoietic enhancer is a critical element in the network of GATA factors and sites that regulate this gene. Molecular and Cellular Biology 20: 713–723.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. Science 302: 413.

Noonan JP, McCallion AS. 2010. Genomics of Long-Range Regulatory Elements. Annu Rev Genomics Hum Genet 11: 1–24.

Nowak KJ, Davies KE. 2004. Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment - Third in Molecular Medicine Review Series. EMBO Rep 5: 872–876.

O'Connor M, Peifer M, Bender W. 1989. Construction of Large Dna Segments in Escherichia-Coli. Science 244: 1307–1312.

O'Flanagan RA, Paillard G, Lavery R, Sengupta AM. 2005. Non-additivity in protein-DNA binding. Bioinformatics 21: 2254–2263.

Ogata T, Ueno T, Hoshikawa S, Ito J, Okazaki R, Hayakawa K, Morioka K, Yamamoto S, Nakamura K, Tanakab S, Akai M. 2011. Hes1 Functions Downstream of Growth Factors to Maintain Oligodendrocyte Lineage Cells in the Early Progenitor Stage. Neuroscience 176: 132–141.

Omenn GS, McKusick VA, Gorlin RJ. 1979. The association of Waardenburg syndrome and Hirschsprung megacolon. Am. J. Med. Genet. 3: 217–223.

Osumiyamashita N, Ninomiya Y, Doi H, Eto K. 1994. The Contribution of Both Forebrain and Midbrain Crest Cells to the Mesenchyme in the Frontonasal Mass of Mouse Embryos. Dev. Biol. 164: 409–419.

Ou XM, Lemonde S, Jafar-Nejad H, Bown CD, Goto A, Rogaeva A, Albert PR. 2003. Freud-1: A neuronal calcium-regulated repressor of the 5-HT1A receptor gene. Journal of Neuroscience 23: 7415–7425.

Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, Bird TD, Fong C-T, Mefford HC, Smith RJ, Stephens K, Bird TD. 1993. Charcot-Marie-Tooth Neuropathy Type 2.

Paratore C, Goerich DE, Suter U, Wegner M, Sommer L. 2001. Survival and glial fate acquisition of neural crest cells are regulated by an interplay between the transcription factor

Sox10 and extrinsic combinatorial signaling. Development 128: 3949–3961.

Pareyson D, Marchesi C. 2009. Diagnosis, natural history, and management of Charcot-Marie-Tooth disease. Lancet Neurol 8: 654–667.

Pareyson D, Scaioli V, Laura M. 2006. Clinical and electrophysiological aspects of Charcot-Marie-Tooth disease. Neuromolecular Medicine 8: 3–22.

Parisi MA. 2015. Hirschsprung Disease Overview.

Parker DS, White MA, Ramos AI, Cohen BA, Barolo S. 2011. The cis-Regulatory Logic of Hedgehog Gradient Responses: Key Roles for Gli Binding Affinity, Competition, and Cooperativity. Sci Signal 4: –ra38.

Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. 2013. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nature Biotechnology 31: 839–843.

Pedersen AG, Baldi P, Chauvin Y, Brunak S. 1999. The biology of eukaryotic promoter prediction - a review. Computers & Chemistry 23: 191–207.

Peirano RI, Goerich DE, Riethmacher D, Wegner M. 2000. Protein zero gene expression is regulated by the glial transcription factor Sox10. Molecular and Cellular Biology 20: 3198–3209.

Peirano RI, Wegner M. 2000. The glial transcription factor Sox10 binds to DNA both as monomer and dimer with different functional consequences. Nucleic Acids Res 28: 3047–3055.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.

Pingault V, Bondurand N, Kuhlbrodt K, Goerich DE, Préhu MO, Puliti A, Herbarth B, Hermans-Borgmeyer I, Legius E, Matthijs G, Amiel J, Lyonnet S, et al. 1998. SOX10 mutations in patients with Waardenburg-Hirschsprung disease. Nat Genet 18: 171–173.

Potterf SB, Furumura M, Dunn KJ, Arnheiter H, Pavan WJ. 2000. Transcription factor hierarchy in Waardenburg syndrome: regulation of MITF expression by SOX10 and PAX3. Hum Genet 107: 1–6.

Prasad MK, Reed X, Gorkin DU, Cronin JC, McAdow AR, Chain K, Hodonsky CJ, Jones EA, Svaren J, Antonellis A, Johnson SL, Loftus SK, et al. 2011. SOX10 directly modulates ERBB3 transcription via an intronic neural crest enhancer. BMC Dev Biol 11:.

Pruitt K, Brown G, Tatusova T, Maglott D. 2012. The reference sequence (RefSeq) database.

Pusch C, Hustert E, Pfeifer D, Sudbeck P, Kist R, Roe B, Wang ZL, Balling R, Blin N, Scherer G. 1998. The SOX10/Sox10 gene from human and mouse: sequence, expression, and transactivation by the encoded HMG domain transcription factor. Hum Genet 103: 115–123.

Raeymaekers P, Timmerman V, Nelis E, De Jonghe P, Hoogendijk JE, Baas F, Barker DF, Martin JJ, De Visser M, Bolhuis PA. 1991. Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. Neuromuscul. Disord. 1: 93–97.

Ramos AI, Barolo S. 2013. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos Trans R Soc Lond B Biol Sci 368:.

Reeves R, Gorman CM, Howard B. 1985. Minichromosome assembly of non-integrated plasmid DNA transfected into mammalian cells. Nucleic Acids Res 13: 3599–3615.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, et al. 2000. Genome-wide location and function of DNA binding proteins. Science 290: 2306–2309.

Robison K, McGuire AM, Church GM. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. Journal of Molecular Biology 284: 241–254.

Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, Robson P. 2005. Transcriptional regulation of Nanog by Oct4 and Sox2. Journal of Biological Chemistry 280: 24731–24737.

Rogers BL, Sobnosky MG, Saunders GF. 1986. Transcriptional enhancer within the human placental lactogen and growth hormone multigene cluster. Nucleic Acids Res 14: 7647–7659.

Rowan S, Siggers T, Lachke SA, Yue Y, Bulyk ML, Maas RL. 2010. Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. Genes Dev 24: 980–985.

Rubin GM, Spradling AC. 1982. Genetic-Transformation of Drosophila with Transposable Element Vectors. Science 218: 348–353.

Sabo PJ, Humbert R, Hawrylycz M, Wallace JC, Dorschner MO, McArthur M, Stamatoyannopoulos JA. 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc Natl Acad Sci USA 101: 4537–4542.

Salazar-Grueso EF, Kim S, Kim H. 1991. Embryonic mouse spinal cord motor neuron hybrid cells. NeuroReport 2: 505–508.

Sandelin A, Alkema W, EngstroEm P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: 91–94.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8: 424–436.

Sanosaka T, Namihira M, Asano H, Kohyama J, Aisaki K, Igarashi K, Kanno J, Nakashima K. 2008. Identification of genes that restrict astrocyte differentiation of midgestational neural precursor cells. Neuroscience 155: 780–788.

Sasai Y, Kageyama R, Tagawa Y, Shigemoto R, Nakanishi S. 1992. Two mammalian helix-loop-helix factors structurally related to Drosophila hairy and Enhancer of split. Genes Dev 6: 2620–2634.

Savic D, Partridge EC, Newberry KM, Smith SB, Meadows SK, Roberts BS, Mackiewicz M, Mendenhall EM, Myers RM. 2015. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. Genome Research 25: 1581–1589.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. Genome Research 22: 1748–1759.

Schlierf B, Ludwig A, Klenovsek K, Wegner M. 2002. Cooperative binding of Sox10 to DNA: requirements and consequences. Nucleic Acids Res 30: 5509–5516.

Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887–898.

Schreiner S, Cossais F, Fischer K, Scholz S, Boesl MR, Holtmann B, Sendtner M, Wegner M. 2007. Hypomorphic Sox10 alleles reveal novel protein functions and unravel developmental differences in glial lineages. Development 134: 3271–3281.

Schröck E, Manoir du S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, Garini Y, Ried T. 1996. Multicolor spectral karyotyping of human chromosomes. Science 273: 494–497.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. 2000. PipMaker - A Web server for aligning two genomic DNA sequences. Genome Research 10: 577–586.

Serbedzija GN, Bronnerfraser M, Fraser SE. 1992. Vital Dye Analysis of Cranial Neural Crest Cell-Migration in the Mouse Embryo. Development 116: 297–307.

Serbedzija GN, Fraser SE, Bronnerfraser M. 1990. Pathways of Trunk Neural Crest Cell-Migration in the Mouse Embryo as Revealed by Vital Dye Labeling. Development 108: 605–612.

Shah NM, Marchionni MA, Isaacs I, Stroobant P, Anderson DJ. 1994. Glial growth factor restricts mammalian neural crest stem cells to a glial fate. Cell 77: 349–360.

Shen SQ, Myers CA, Hughes AEO, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Research 26: 238–255.

Sherf BA, Navarro SL, Hannah RR, Wood KV. 1996. Dual-Luciferase® reporter assay: An advanced co-reporter technology integrating firefly and Renilla luciferase assays. Promega Notes.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. PNAS 100: 15776–15781.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050.

Singh P, Schimenti JC, Bolcun-Filas E. 2015. A mouse geneticist's practical guide to CRISPR applications. Genetics.

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nature Biotechnology 17: 974–978.

Small S, Blair A, Levine M. 1992. Regulation of Even-Skipped Stripe-2 in the Drosophila Embryo. EMBO J 11: 4047–4057.

Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010:.

Sonnenberg-Riethmacher E, Miehe M, Stolt CC, Goerich DE, Wegner M, Riethmacher D. 2001. Development and degeneration of dorsal root ganglia in the absence of the HMG-domain transcription factor Sox10. Mech Dev 109: 253–265.

Southard-Smith EM, Kos L, Pavan WJ. 1998. Sox10 mutation disrupts neural crest development in DOM Hirschsprung mouse model. Nat Genet 18: 60–64.

Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA. 2005. Interchromosomal associations between alternatively expressed loci. Nature 435: 637–645.

Srinivasan R, Sun G, Keles S, Jones EA, Jang S-W, Krueger C, Moran JJ, Svaren J. 2012. Genome-wide analysis of EGR2/SOX10 binding in myelinating peripheral nerve. Nucleic Acids Res 40: 1–12.

Stanojevic D, Small S, Levine M. 1991. Regulation of a Segmentation Stripe by Overlapping Activators and Repressors in the Drosophila Embryo. Science 254: 1385–1387.

Stathopoulou A, Natarajan D, Nikolopoulou P, Patmanidi AL, Lygerou Z, Pachnis V, Taraviras S. 2016. Inactivation of Geminin in neural crest cells affects the generation and maintenance of enteric progenitor cells, leading to enteric aganglionosis. Dev. Biol. 409: 392–405.

Stewart H, Brennan A, Rahman M, Zoidl G, Mitchell PJ, Jessen KR, Mirsky R. 2001. Developmental regulation and overexpression of the transcription factor AP-2, a potential regulator of the timing of Schwann cell generation. Eur J Neurosci 14: 363–372.

Stewart HJ, Zoidl G, Rossner M, Brennan A, Zoidl C, Nave KA, Mirsky R, Jessen KR. 1997. Helix-loop-helix proteins in Schwann cells: a study of regulation and subcellular localization of Ids, REB, and E12/47 during embryonic and postnatal development. J. Neurosci. Res. 50: 684–701.

Stolt CC, Schlierf A, Lommes P, Hillgaertner S, Werner T, Kosian T, Sock E, Kessaris N, Richardson WD, Lefebvre V, Wegner M. 2006. SoxD proteins influence multiple stages of oligodendrocyte development and modulate SoxE protein function. Dev Cell 11: 697–709.

Stolt CC, Wegner M. 2015. Schwann cells and their transcriptional network: Evolution of key regulators of peripheral myelination. Brain Res.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982. Use of the “Perceptron” algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 10: 2997–3011.

Stottmann RW, Donlin M, Hafner A, Bernard A, Sinclair DA, Beier DR. 2013. A mutation in Tubb2b, a human polymicrogyria gene, leads to lethality and abnormal cortical development in the mouse. Human Molecular Genetics 22: 4053–4063.

Swamynathan SK, Piatigorsky J. 2002. Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alpha B-crystallin and Mkbp/HspB2 genes. Journal of Biological Chemistry 277: 49700–49706.

Szigeti K, Lupski JR. 2009. Charcot–Marie–Tooth disease. European Journal of Human Genetics 17: 703–710.

Słomnicki LP, Leśniak W. 2010. S100A6 (calcyclin) deficiency induces senescence-like changes in cell cycle, morphology and functional characteristics of mouse NIH 3T3 fibroblasts. J Cell Biochem 109: 576–584.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. 1988. Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus). Journal of Molecular Biology 203: 439–455.

Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5 ' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7: 542–561.

Tasic B, Hippenmeyer S, Wang C, Gamboa M, Zong H, Chen-Tsai Y, Luo L. 2011. Site-specific integrase-mediated transgenesis in mice via pronuclear injection. Proc Natl Acad Sci USA 108: 7902–7907.

Thomas PK, Marques W, Davis MB, Sweeney MG, King R, Bradley JL, Muddle JR, Tyson J, Malcolm S, Harding AE. 1997. The phenotypic manifestations of chromosome 17p11.2 duplication. Brain 120: 465–478.

Thorpe HM, Smith MC. 1998. In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. Proc Natl Acad Sci USA 95: 5505–5510.

Tischfield MA, Baris HN, Wu C, Rudolph G, Van Maldergem L, He W, Chan W-M, Andrews C, Demer JL, Robertson RL, Mackey DA, Ruddle JB, et al. 2010. Human TUBB3 Mutations Perturb Microtubule Dynamics, Kinesin Interactions, and Axon Guidance. Cell 140: 74–87.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10: 1453–1465.

Topilko P, Schneidermaunoury S, LEVI G, Baronvanevercooren A, Chennoufi, AB, Seitanidou T, Babinet C, Charnay P. 1994. Krox-20 Controls Myelination in the Peripheral Nervous-System. Nature 371: 796–799.

Townsley FM, Aristarkhov A, Beck S, Hershko A, Ruderman JV. 1997. Dominant-negative cyclin-selective ubiquitin carrier protein E2-C/UbcH10 blocks cells in metaphase. PNAS 94: 2362–2367.

Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, Aryee MJ, Joung JK. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature Biotechnology 33: 187–197.

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249: 505–510.

Tyas DA, Simpson TI, Carr CB, Kleinjan DA, Van Heyningen V, Mason JO, Price DJ. 2006. Functional conservation of Pax6 regulatory elements in humans and mice demonstrated with a novel transgenic reporter mouse. BMC Dev Biol 6: 21–21.

Vasiliev GV, Merkulov VM, Kobzev VF, Merkulova TI, Ponomarenko MP, Kolchanov NA. 1999. Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY-1 transcription factor binding site. FEBS Lett. 462: 85–88.

Verdin H, Fernández-Miñán A, Benito-Sanz S, Janssens S, Callewaert B, Waele KD, Schepper JD, François I, Menten B, Heath KE, Gómez-Skarmeta JL, Baere ED. 2015. Profiling of conserved non-coding elements upstream of SHOX and functional characterisation of the SHOX cis-regulatory landscape. Sci Rep 5: 17667–17667.

Vyas P, Vickers MA, Picketts DJ, Higgs DR. 1995. Conservation of Position and Sequence of a Novel, Widely Expressed Gene Containing the Major Human Alpha-Globin Regulatory Element. Genomics 29: 679–689.

Wahlbuhl M, Reiprich S, Vogl MR, Bösl MR, Wegner M. 2012. Transcription factor Sox10 orchestrates activity of a neural crest-specific enhancer in the vicinity of its gene. Nucleic Acids Res 40: 88–101.

Wang SL, Sdrulla A, Johnson JE, Yokota Y, Barres BA. 2001. A role for the helix-loop-helix protein Id2 in the control of oligodendrocyte development. Neuron 29: 603–614.

Warner LE, Mancias P, Butler IJ, McDonald CM, Keppen L, Koob KG, Lupski JR. 1998. Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies. Nat Genet 18: 382–384.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Wei Q, Miskimins WK, Miskimins R. 2004. Sox10 acts as a tissue-specific transcription factor enhancing activation of the myelin basic protein gene promoter by p27Kip1 and Sp1. J. Neurosci. Res. 78: 796–802.

Wei W, Brennan MD. 2000. Polarity of transcriptional enhancement revealed by an insulator element. PNAS 97: 14518–14523.

Weterman MAJ, van Ruissen F, de Wissel M, Bordewijk L, Samijn JPA, van der Pol WL, Meggouh F, Baas F. 2010. Copy number variation upstream of PMP22 in Charcot-Marie-Tooth disease. European Journal of Human Genetics 18: 421–428.

Wiedenheft B, Sternberg SH, Doudna JA. 2012. RNA-guided genetic silencing systems in bacteria and archaea. Nature 482: 331–338.

Wilkie TM, Palmiter RD. 1987. Analysis of the Integrant in Myk-103 Transgenic Mice in Which Males Fail to Transmit the Integrant. Molecular and Cellular Biology 7: 1646–1655.

Wilson YM, Richards KL, Ford-Perriss ML, Panthier JJ, Murphy M. 2004. Neural crest cell lineage segregation in the mouse neural tube. Development 131: 6153–6162.

Wingender E. 1988. Compilation of transcription regulating proteins. Nucleic Acids Res 16: 1879–1902.

Wood KV, de Wet JR, Dewji N, DeLuca M. 1984. Synthesis of active firefly luciferase by in vitro translation of RNA obtained from adult lanterns. Biochem Biophys Res Commun 124: 592–596.

Woodhoo A, Alonso MBD, Droggiti A, Turmaine M, D'Antonio M, Parkinson DB, Wilton DK, Al-Shawi R, Simons P, Shen J, Guillemot F, Radtke F, et al. 2009. Notch controls embryonic Schwann cell differentiation, postnatal myelination and adult plasticity. Nat. Neurosci. 12: 839–847.

Wright EM, Snopek B, Koopman P. 1993. Seven new members of the Sox gene family expressed during mouse development. Nucleic Acids Res 21: 744–744.

Wright WE, Binder M, Funk W. 1991. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Molecular and Cellular Biology 11: 4104–4110.

Yaffe D, Saxel O. 1977a. Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. Nature 270: 725–727.

Yaffe D, Saxel O. 1977b. A Myogenic Cell Line with Altered Serum Requirements for Differentiation. Differentiation 7: 159–166.

Yagura M, Itoh T. 2006. The Rep protein binding elements of the plasmid ColE2-P9 replication origin. Biochem Biophys Res Commun 345: 872–877.

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen L-L. 2011. Genomewide characterization of non-polyadenylated RNAs. Genome Biol 12:.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature 518: 556–559.

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev 21: 385–390.

Zerucha T, Stühmer T, Hatch G, Park BK, Long Q, Yu G, Gambarotta A, Schultz JR, Rubenstein JL, Ekker M. 2000. A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. Journal of Neuroscience 20: 709–721.

Zhang WQ, Shields JM, Sogawa K, Fujii-Kuriyama Y, Yang VW. 1998. The gut-enriched Kruppel-like factor suppresses the activity of the CYP1A1 promoter in an Sp1-dependent fashion. Journal of Biological Chemistry 273: 17917–17925.

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, et al. 2014. An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. Journal of Neuroscience 34: 11929–11947.

Zhao X, He X, Han X, Yu Y, Ye F, Chen Y, Hoang T, Xu X, Mi Q-S, Xin M, Wang F, Appel B, et al. 2010. MicroRNA-mediated control of oligodendrocyte differentiation. Neuron 65: 612–626.

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. Nature.

2007. Genome Research.