INVESTIGATING THE RELATIONSHIP AND POTENTIAL INTERACTIONS OF CD108131 AND SGCE

RHIANNON JAMIESON-WILLIAMS

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
M.Sc. program in Neuroscience, collaborative with
Human and Molecular Genetics specialisation.

Faculty of Medicine
University of Ottawa

Abstract

Myoclonus dystonia (MD) is a rare autosomal-dominant combined dystonia movement disorder characterised by quick, involuntary muscle jerks (myoclonus) paired with sustained muscular contraction (dystonia). Although known to be genetically heterogeneous, the most common genetic factor is mutations within *SGCE*, the gene encoding ε-sarcoglycan, accounting for approximately 45% of cases. Previous linkage analyses conducted on a family displaying inherited MD without *SGCE* mutations lead to the identification of another critical region, DYT15. Preliminary data suggested that mutations within the long non-coding RNA (lncRNA) *CD108131*, found within the DYT15 locus, resulted in decreased expression of both the *SGCE* transcript, as well as the SGCE protein. Validation of the remaining variants of interest yielded no new candidate genes. A low coverage area coinciding with the entire sequence of *TMEM200C* was discovered, however subsequent sequencing data revealed no potential disease-causing variants. Therefore, to further characterise the relationship between *CD108131* and *SGCE* suggested by the preliminary data, a CRISPR-Cas9 knockout was developed in HEK293 cells using a double-cut strategy that allowed for complete excision of the *CD108131* gene. Stable *CD108131* knockout mutant cell lines were examined for differences in gene expression. QRT-PCR analysis was conducted and revealed a significant decrease in *SGCE* expression in the absence of *CD108131*. Additionally, expression also trended towards a decrease for *ZBTB14*, however *ARHGAP28* and *RPPH1* were not significantly altered. This data demonstrates that the lncRNA *CD108131* is likely to have a regulatory effect on *SGCE*, and perhaps *ZBTB14*, transcription.

Acknowledgements

I would like to thank my supervisor, Dr. Dennis Bulman, for all the support and mentorship that he has offered over the course of my journey through this graduate programme. Not only did he see potential in me and chose to hire me as his student, but he has stood by and support me through all the triumphs, trials, and tribulations faced throughout these years – both academically, and personally. I would also like to thank the various individuals that took the time to offer me guidance and training. Most notably: Dr. David Lohnes and his lab, for counsel and assistance involving the design of the CRISPR/Cas9 experiment; Lemuel Racacho, for his optimism, constant "next-step" ideas, and bioinformatic expertise; Ruobing Zou, for her patience in sequencing all the countless samples I sent her way; Thet Fatica, for her direction with regards to the qRT-PCR analysis; and Bruno Fontesca, for his Western blot wizardry (while it might not have worked for me, it was definitely an improvement!). Additional thanks to all the other members of the Research Institute that helped this project in one way or another, as well as the past members of the Bulman lab whose insights laid the foundation for this project.

I would also like to acknowledge my thesis advisory committee members, Dr. Bob Korneluk and Dr. Rashmi Kothary, for their guidance and support throughout my degree. Additionally, thank you to Dr. Bulman, CIHR, and the Dystonia Foundation for funding this research.

Finally, I would like to extend my heartfelt thanks to the friends and family that stood by me, bolstered me, celebrated with me, and believed in me. To my parents, Deborah and Neil, thank you for loving me unconditionally, encouraging me to reach my potential, and making sure that anytime I felt defeated by this project I ended up back on my feet. To my husband, Sam, thank you for making sure every day involved laughter, and for helping me to keep everything in

perspective. To my friends (especially Shannon), thank you for the phone calls, texts, and pep-talks that reminded me I was capable of more than I sometimes felt. I am so lucky to have had a team of cheerleaders in my corner, and appreciate it more than you could ever know.

Table of Contents

List of Abbreviations

µg	microgram
µl	microlitre
µM	micromoles per litre
18p11	chromosome 18, short arm, region 1, band 1
7q21	chromosome 7, long arm, region 2, band 1
ACTB	actin, beta
AD	Alzheimer's disease
ADCY5	adenylate cyclase 5
ANO3	anoctamin 3
ANOVA	analysis of variance
APP	amyloid beta precursor protein
ARGHAP28	rho GTPase activating protein 28
ATCC	American type culture collection
ATP	adenosine tri-phosphate
ATP1A3	ATPase Na/K transporting subunit alpha 3
BLAT	BLAST-like alignment tool
bp	base pair
Cas1, 2, 9	CRISPR-associated protein 1, 2, 9
CCDS	consensus coding sequence
cDNA	complementary DNA
CHART	capture hybridization analysis of RNA
CHIP	chromatin immunoprecipitation
Chr	chromosome
cM	centimorgans
CRISPR	clustered regularly interspersed short palindromic repeats
crRNA	CRISPR RNA
Ctrl	control
DAP	dystrophin-associated protein

DMEM	Dulbecco’s modified eagle medium
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
DRD	dopamine-responsive dystonia
DRD2	dopamine receptor D2
DSB	double-stranded break
DYT1-28	dystonia locus 1-28
EST	expressed sequence tag
FACS	fluorescence activated cell sorting
FBS	fetal bovine serum
FISH	fluorescence in situ hybridisation
FSHD	facioscapulohumeral muscular dystrophy
g	gram
GABA	gamma-aminobutyric acid
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GCH1	GTP cyclohydrolase 1
GFP	green fluorescent protein
GNAL	G protein subunit alpha L
HDR	homology directed repair
HEK293	human embryonic kidney cells 293
Het	heterozygous
Hg38	human genome build 38
Hom	homozygous
IDT	Integrated DNA Technologies
IgG	immunoglobulin G
IMD	inherited myoclonus dystonia
IMP2	IGF2 mRNA-binding protein 2
IP	immunoprecipitation

iPSC	induced pluripotent stem cells
kb	kilobase
kDa	kilodalton
KCTD17	potassium channel tetramerization domain containing 17
KMT2B	lysine methyltransferase 2B
KO	knockout
LB	Lysogeny broth
LINE	long interspersed nuclear element
lncRNA	long non-coding RNA
Mb	megabase
MD	Myoclonus-Dystonia
MeOH	methanol
MF1	myoclonus family 1
mL	millilitres
mM	millimoles per litre
mRNA	messenger RNA
MyoD	myogenic differentiation
nCas9	nickase CRISPR-associate protein 9
ncRNA	non-coding RNA
NEAA	non-essential amino acids
ng	nanogram
NHEJ	non-homologous end joining
nt	nucleotides
PAGE	poly-acrylamide gel electrophoresis
PAM	protospacer adjacent motif
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PET	positron emission tomography

PNKD	paroxysmal nonkinesigenic dyskinesia
PRKRA	protein activator of interferon induced protein kinase EIF2AKA
PRRT2	proline-rich transmembrane protein 2
qRT-PCR	Quantitative Real-Time PCR
RIPA	radioimmunoprecipitation assay buffer
RNA	ribonucleic acid
rpm	revolutions per minute
RPPH1	ribonuclease P RNA component H1
RT-PCR	reverse-transcriptase PCR
SDS	sodium dodecyl sulphate
SGCE	epsilon-sarcoglycan
SGCZ	zeta-sarcoglycan
sgRNA	single guide RNA
SINE	short interspersed nuclear elements
siRNA	small interfering RNA
SLC2A1	solute carrier family 2 member 1
Sno-lncRNA	small nucleolar lncRNA
SNP	single nucleotide polymorphism
TAF1	TATA box-binding protein-associated factor 1
TALE	transcription activator-like effector
TALEN	transcription activator-like effector nuclease
TBS-T	tris-buffered saline with TWEEN20
TH	tyrosine hydroxylase xv
THAP1	Thanatos-associated protein domain containing apoptosis-associated protein 1
TITF-1	thyroid nuclear factor 1
TM	melting temperature
TOR1A	torsin A
tracrRNA	trans-activating crRNA

TSS	transcription start site
UCSC	University of California Santa Cruz
UMI	unique molecular identifier
V	volt
WT	wild-type
Xist	X-inactive specific transcript
ZBTB14	zinc finger and BTB domain containing 14
ZFN	zinc finger nuclease

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Dystonic movement disorders

Dystonia is the third most common movement disorder affecting humans, after Parkinson's disease and essential tremor (Richter et al, 2015). The phenotype was initially documented in 1911, although it wasn't until 1994 that the first genetic link was discovered (Albanese, et al., 2013, Klein, 2014). Over the years, the clinical definition of dystonia as a movement disorder has varied. In 2013, an international panel of experts revised the definition to the following: Dystonia is a movement disorder characterized by sustained or intermittent muscle contractions causing abnormal, often repetitive, movements, postures, or both. Dystonic movements are typically patterned and twisting and may be tremulous. Dystonia is often initiated or worsened by voluntary action and associated with overflow muscle activation (Albanese, et al., 2013). While it can manifest at any age, the earlier the age of onset, the more severe the disorder tends to be (Phukan et al, 2011). The region affected by dystonia can be singular (focal) or can spread connected muscles (segmental) or also unconnected muscles (multifocal). More severe cases can involve generalized dystonia, where often all four limbs can be affected, and hemidystonia, where half the body is symptomatic (Charlesworth, Bhatia, and Wood, 2013).

Dystonia encompasses an incredibly heterogeneous set of disorders. To date, 29 loci have been genetically linked to dystonia: DYT1-28, as well as one unnamed locus. Of these, 15 loci (DYT1, DYT3, DYT5a, DYT5b/DYT14, DYT6, DYT8, DYT10, DYT11, DYT12, DYT16, DYT18, DYT24, DYT25, DYT28, and the unnamed locus) have been narrowed down to their specific causative gene, respectively: *TOR1A*, *TAF1*, *GCH1*, *TH*, *THAP1*, *PNKD*, *PRRT2*, *SGCE*, *ATP1A3*, *PRKRA*, *SLC2A1*, *ANO3*, *GNAL*, *KMT2B*, and *SPR* (Weisheit, Pappas, and Dauer, 2018). These genes are involved in a variety of different pathways, some of which

include: dopamine signalling and metabolism, cytoskeleton and cell movement, vesicle recycling and transport, protein folding and stability, transcriptional regulation, cell cycle progression, nitric oxide synthesis, ion channels, phosphorylation, and glucose transportation and uptake (Lohmann and Klein, 2013). The mode of inheritance for the majority of known monogenic forms of dystonia is autosomal dominant, although 3 are known to be autosomal recessive, and one is X-linked (Weisheit, Pappas, and Dauer, 2018). Interestingly, a predominant feature of autosomal dominant forms of dystonia is that they tend to display reduced penetrance (Lohmann and Klein, 2013).

Furthermore, dystonia can be sub-classified into three categories: complex dystonia, isolated dystonia, and combined dystonia. A diagnosis of complex dystonia occurs when dystonic symptoms arise following damage to any areas of the brain responsible for movement, however typically it is the putamen that is compromised (Weisheit, Pappas, and Dauer, 2018). These cases of dystonia do not have a causative genetic factor involved. Isolated dystonia, formerly referred to as primary dystonia, describes cases of dystonia in which no central nervous system injury has occurred (Weisheit, Pappas, and Dauer, 2018). In these cases, the dystonic condition can have a varied age of onset and body distribution, with persistent dystonia as the only movement disorder (Klein, 2014). Combined dystonia, formerly referred to as dystonia-plus syndrome, involves either persistent or paroxysmal (episodic) dystonia paired with one or more additional movement disorders (Klein, 2014).

Combined dystonia encompasses three different disorders: persistent dystonia-parkinsonism, paroxysmal dystonia with mixed dyskinesia, and persistent myoclonus dystonia (MD). Dystonia-parkinsonism can have a broad phenotype and age of onset, and can be inherited or occur sporadically. In these syndromes, dystonia is paired with non-neurodegenerative symptoms of

Parkinson's disease, such as tremors, bradykinesia, rigidity, and postural instability (Asmus and Gasser, 2010). Dopa-responsive dystonia (DRD) is a subset of this category, whereby patients respond well with dopamine treatment (Asmus and Gasser, 2010). When dystonia is combined with other dyskinesia, the additional symptoms can be one or more of the following: chorea (rapid, flowing, unpatterned movement), athetosis (slow writhing movements), or ballismus (large, flinging movements). Myoclonus dystonia (MD) is the third dystonia-plus syndrome and will be discussed in greater detail.

1.2 Myoclonus Dystonia

MD is a rare autosomal dominant disorder, with an estimated prevalence of 1 in 100 000. Individuals usually become symptomatic within the first two decades of life, with a mean onset age of 6 years of age (Asmus and Gasser, 2010). Myoclonus is typically the predominant clinical sign and is characterized by lightning fast muscle jerks (Asmus and Gasser, 2010). These myoclonic jerks typically present in the arms, shoulders, and neck, while the dystonia presents in the neck (as torticollis) or the hand (as chronic writer's cramp) (Grabowski et al, 2003). Individuals generally do not show progression of their disease state, and live normal lifespans (Grimes et al, 2001).

Myoclonus Dystonia is a genetically heterogeneous disorder (Grimes et al., 2001) with two associated loci, DYT11 and DYT15 (Spatola & Wider, 2012). DYT11 is caused by mutations in the gene epsilon-sarcoglycan (SGCE) (Zimprich et al., 2001). However, mutations in SGCE only account for about 30-50% of MD cases (Grünewald et al., 2008; Han et al., 2007; Ritz et al., 2011). Two separate families, a 5 generation British family and a 2 generation German family, were found to have MD related to a mutation in a conserved, yet functionally uncharacterised, region of *KCTD17*, coding for potassium channel tetramerization domain-containing protein 17

(Mencacci et al, 2015). This presentation of MD differed from the better known DYT11 presentation in that dystonia was the predominant clinical feature and symptoms progressed over time, sometimes spreading to other sites (Mencacci et al, 2015). Little is known about *KCTD17*, but was found to be highly expressed in the putamen and thalamus, and it is thought to have a role in dopaminergic signalling and ciliogenesis (Mencacci et al, 2015). Recently, a potential association has also been made with mutations in the *ADCY5* gene, encoding adenylyl cyclase 5, and MD. *ADCY5*-related mutations have been known to cause various forms of dyskinesia, some of which have been shown to present as an inherited myoclonus dystonia phenotype (Douglas et al, 2017).

Several neurophysiological studies have been performed in an attempt to elucidate the underlying mechanism for the motor symptoms associated with MD. To evaluate sensorimotor integration, functional MRI imaging of DYT11 MD patients and healthy controls was conducted while subjects were performing a simple finger-tapping task. In comparison to the healthy controls, the MD patients were shown to have hyperresponsiveness of the contralateral secondary somatosensory cortex, ipsilateral premotor cortex, primary somatosensory cortex, dorsolateral prefrontal cortex, and ipsilateral cerebellum. A decrease in responsiveness was noted in the contralateral insula (Beukers et al, 2010).

Expanding on these findings, [^{18}F]-fluorodeoxyglucose PET functional imaging was performed on a small sample of DYTY11 MD patients, DYT11 carriers, and age-matched healthy controls, revealing metabolic differences in sensorimotor and brain stem regions (Carbon et al, 2013). When compared to healthy controls, both MD-exhibiting patients and carriers presented with increased metabolic activity in the pontine nuclei and the posterior thalamus, as well as decreased metabolic activity in the ventromedial prefrontal cortex. Additionally, the

affected MD patients presented with hypermetabolism of the parasagittal cerebellum (Carbon et al, 2013).

Structural analyses of brain matter have also been conducted in DYT11 MD patients. Cortical plasticity was inferred by examining the relative volumes of grey and macrostructural white matter, as well as quantification of diffusion of water molecules in microstructural white matter. Results showed that, compared to healthy controls, MD patients had a bilateral increase in white matter volume in the subthalmic area of the brain stem, and decreased diffusivity in white matter near the cortical sensorimotor areas (van der Meer et al, 2012). This structural data corroborates the functional data, suggesting that the brain stem and sensorimotor areas are the main brain regions affected by MD.

In addition to functional, metabolic, and structural abnormalities, psychological disorders are also quite common in patients suffering from MD, who often have diagnoses of depression, anxiety, obsessive-compulsive disorder, alcohol dependence, or panic disorder (Kim et al, 2017). While originally thought to be linked to mutations in *SGCE*, a recent study has found that the frequency of psychiatric disorders does not differ significantly between DYT11 MD patients (61.5%) and those without *SGCE* mutations (54.5%) (Kim et al, 2017). Interestingly, a subset of patients also experiences alleviation of symptoms upon alcohol consumption. However, as blood alcohol levels drop, symptoms rebound with greater severity. Consequently, nearly all those who have tried to self-medicate with alcohol develop a long-term dependency (Ritz et al, 2011).

Alternative treatments for the motor symptoms of MD include several drug treatments aimed to reduce neuronal excitability through GABAergic and anticholinergic agents, such as benzodiazepine or chlorpromazine (Caviness, 2014). Other therapies involve drugs designed to treat other neuromuscular disorders, such as antiepileptic drugs (Caviness, 2014). In severe

cases, where pharmaceuticals have not alleviated symptoms, bilateral deep brain stimulation of the internal globus pallidum can be considered. This involves the insertion of two microelectrodes into the posteroventral portion of the globus pallidum, which will then provide electrical stimulus to the region at a prescribed voltage, pulse width, and frequency. Over 90% of patients that receive this treatment reach an improvement of 50% or more, with the benefits lasting up to ten years (Fernandez-Parjarin et al, 2016).

1.3 Characterization of the gene *SGCE*

Located on chromosome 7q21.3, *SGCE* is the 70 986bp coding gene for the epsilon sargoclycan protein. *SGCE* is highly conserved and is comprised of 13 exons with 5 known protein-coding isoforms formed through the alternative splicing of exons 2, 8, 10, and 11b. Of these exons, 11b is of particular interest as its inclusion denotes the brain specific isoform. While the mouse homologue of *SGCE* has two distinct brain specific isoforms, in humans there exists only one to date (Ritz et al, 2011). Ubiquitous *SGCE* expression is found in a variety of tissues, including skeletal muscle, heart, liver, kidney, and brain. With regards to the brain specific isoform, the highest levels of expression are seen in the primary somatosensory cortex and the motor cortex, with intermediate expression seen in the caudate nucleus and substantia nigra, and the lowest levels of expression seen in the globus pallidus (Ritz et al., 2011). While originally thought to be the only sarcoglycan expressed in the brain, it has since been revealed that zeta-sarcoglycan (*SGCZ*) is also expressed within the brain, with a similar expression pattern to *SGCE* (Shiga et al, 2006). The widespread expression of *SGCE* and *SGCZ* is in direct contrast to the localised expression of the other sarcoglycan genes, which are generally restricted to musculature, although it is now known that all the major sarcoglycans are expressed in the cerebrovascular system (Piras et al, 2000, Boulay et al, 2015). *SGCE* expression, regardless of

isoform, is variable amongst individuals and amongst different tissue types, with the greatest individual variation being seen in cerebellar and putamen tissues (Ritz et al, 2011).

As with most dystonia-related mutations, reduced penetrance is a predominant feature of *SGCE*-related MD. This reduced penetrance can be explained by *SGCE* imprinting in several tissues including human brain (Grabowski *et al*., 2003). Observations in human blood cells, and murine embryogenic fibroblasts, neonatal and embryogenic brain tissue show that the maternal allele of epsilon-sarcoglycan is methylated, resulting in only the paternal allele being expressed (Grabowski et al, 2003; Carbon et al, 2013). Those individuals with a paternally inherited loss-of-function mutation exhibit MD, whereas individuals with a maternally inherited mutation are asymptomatic (Grabowski et al., 2003). While most cases of inherited MD do follow this pattern, it has been noted that in adult mouse brain samples, weak expression of the maternal allele can occur (Piras et al, 2000). A study observing metabolic changes in DYT11 patients, carriers, and controls noted that metabolic abnormalities were common between patients and non-manifesting carriers (Carbon et al, 2013). Gene imprinting is not always binary, and the degree to which any given gene is imprinted can fall anywhere along the spectrum from entirely uniparental to equal biallelic expression (Wang et al, 2008). Without methylation and expression data for adult human neurons, the completeness of the maternal imprinting of *SGCE* remains elusive, as imprinting patterns differ between species, tissue type, and developmental age (Carbon et al, 2013).

1.4 Functional characteristics of the SGCE protein

At the time of discovery, the EST for ε-sarcoglycan was originally assumed to be a homologue of α-sarcoglycan (Ettinger et al., 1997). This was based on the observation that the coding regions of α- and ε-sarcoglycan are 47% homologous in nucleotide sequence and 62%

homologous in amino acid sequence (Ozawa et al., 2005). Epsilon-sarcoglycan is now known to be a distinct member of the N-glycosylated transmembrane sarcoglycan protein family, alongside five other members: α-, β-, γ-, δ-, and ζ-sarcoglycan.

The canonical role of the sarcoglycans occurs within muscle, where they act as an essential component of the dystrophin-associated protein complex (DAP). The DAP complex links the extracellular matrix to the intercellular cytoskeletal actin through the association of syntrophins, dystrobrevins, dystroglycans, and sarcospan with dystrophin (Nishiyama et al, 2004). This complex assists in protecting striated muscle fibres from contraction-induced mechanical stress (Nishiyama et al, 2004). Within the DAP complex exists the sarcoglycan subcomplex, which consists of a functional core formed by β-sarcoglycan and δ-sarcoglycan. This core unit recruits α- and γ-sarcoglycan to complete the subcomplex in skeletal muscle, but in smooth muscle ε- and ζ-sarcoglycan are incorporated instead (Tarakci and Berger, 2016). The εβδγ tetramer has also been seen in Schwann cells and adipose tissue (Waite et al, 2016). Formed in the ER, the sarcoglycan subcomplex functions to strengthen the interaction between dystrophin and dystroglycan (Tarakci and Berger, 2016). Limb-girdle muscular dystrophies result from mutations in α-, β-, γ-, or δ-sarcoglycan (Nishiyama et al., 2004). To date, no human diseases have been linked to mutations in ζ-sarcoglycan (Waite et al, 2016).

Within the brain, ubiquitous ε-sarcoglycan has only been observed in brain-derived capillary endothelial cells and astrocytes, while the brain specific ε-sarcoglycan is believed to be the primary isoform in neurons (Waite et al, 2016). Previous in vitro studies have shown that the βδ-sarcoglycan core is necessary for the assembly and trafficking of the entire sarcoglycan subcomplex, however later studies discovered that while β-sarcoglycan mutations did disrupt the assembly and trafficking of the core unit, ε- and ζ-sarcoglycan were not affected in transfected

human embryonic kidney (HEK293) cells (Waite et al, 2016). Therefore, ε- and ζ-sarcoglycan are able to traffic independently of the sarcoglycan subcomplex. This suggests that where mutations in other sarcoglycans would prevent them from associating as a complex and being trafficked out of the ER, in the brain ε- and ζ-sarcoglycan are able to bypass this. This could be the reason that MD does not display limb-girdle muscular dystrophy pathologies (Waite et al, 2016). A recent study using induced pluripotent stem cell (iPSC)-derived cortical neurons generated from MD patient fibroblasts was able to retain the methylated promoter region that results in maternal imprinting of *SGCE*. Interestingly, they found that mutated ε-sarcoglycan underwent retrotranslation from the ER to the proteasome for subsequent degradation (Grutz et al, 2017). Due to the widespread expression of ε-sarcoglycan, it is clear that the protein has the potential to perform a variety of different roles depending on cell type, isoform, and developmental stage.

Structural analysis of the C-terminal cytoplasmic region of ε-sarcoglycan revealed that the brain specific isoform had a consensus sequence indicating a phosphorylation site of cAMP- and cGMP-dependent protein kinase, while modelling of the N-terminal extracellular region in the mouse homologue revealed a cadherin-like domain and a PDZ-binding domain (Nishiyama et al., 2004; Yokoi et al, 2012). An enrichment of ε-sarcoglycan occurs in the pre- and post-synaptic membranes of murine neuronal cells, suggesting a potential synaptic function within the central nervous system (Nishiyama et al., 2004). Indeed, PDZ-containing proteins are known to assist in the assembly of supramolecular complexes involved in localised signalling, leading some to hypothesise that ε-sarcoglycan may be involved in the formation of such a complex within the synapses (Yokoi et al, 2012).

One particular family with MD was reported to have a point mutation in the dopamine D2 receptor gene, *DRD2*, which, along with the observation of high expression of ε-sarcoglycan within mouse dopaminergic neurons, indicates the possibility of ε-sarcoglycan playing a role in dopamine transmission and signalling (Nishiyama et al., 2004). In addition, a mouse model generated to possess a paternally-inherited heterozygous knockout of ε-sarcoglycan exhibited high serotonin turnover and a hyperdopaminergic striatum (Yokoi et al., 2006). Similarly, a homozygous knockout mouse model presented with reduced levels of D2 receptors in both the pre-synaptic dopaminergic terminals and the post-synaptic medium spiny neurons, as well as an increase in striatal dopamine discharge upon amphetamine injection (Zhang et al, 2012). Conditional mouse models were also made to look at the effects of a lack of ε-sarcoglycan specifically in either the cerebellar Purkinje cells or the striatal medium spiny neurons in two studies conducted by Yokoi et al (Yokoi et al, 2012a; Yokoi et al, 2012b). When ε-sarcoglycan was knocked out in either of these cell types alone, no abnormalities were noted in the nuclear envelopes. However, when ε-sarcoglycan was knocked out globally, nuclear blebbing and other abnormalities were observed in both Purkinje cells and medium spiny neurons. Furthermore, these abnormalities presented well in advance of MD phenotypical motor deficits within the mice, indicating that this could be an early symptom and potential biomarker (Yokoi et al, 2012a, Yokoi et al, 2012b). These data suggest that DYT11 MD is most likely caused by ε-sarcoglycan-related dysfunction in multiple cell types within the brain, however more studies will need to be done to truly understand the complete role of this protein within the human body.

1.5 Discovery of the DYT15 locus

In 2001, a 5-generation, French-Canadian family was described as having inherited myoclonus dystonia (IMD) presenting with an autosomal dominant mode of inheritance with

incomplete penetrance (Grimes et al, 2001). Using two-point and multipoint genome-wide linkage analysis, mutations within *DRD2* and *SGCE* were eliminated as the cause of the disease; however, a novel IMD-associated locus was identified: DYT15. (Grimes et al, 2001, Grimes et al, 2002). DYT15 originally encompassed a 17 cM region on chromosome 18p11, though further refinement of the locus was executed utilising SNPs for fine-mapping, resulting in a smaller 3.18 Mb critical region (Grimes et al, 2001, Han et al, 2007).

In an attempt to determine what the causative mutation was for this family's IMD, the exons and exon-intron boundaries of all of the known REFSEQ genes within that region were sequenced and screened for mutations (Vanstone, 2012). While many novel SNPS and indels were detected, there were no variants that segregated within the family and were also absent in healthy control samples (Vanstone, 2012). Additionally, a genome-wide gene dosage analysis was performed using a 500K SNP array in order to search for potential large deletions or duplications, revealing no significant dosage alterations (Vanstone, 2012). Subsequently, two meiotically distant affected patients were chosen from the pedigree to undergo targeted capture of the disease-linked region, followed by Roche 454 sequencing. Results identified 2292 shared novel variants within the critical region, and further analysis and validation of these results revealed a 3 bp duplication within the lncRNA *CD108131* (Vanstone, 2012).

1.6 Characterization of the gene *CD108131*

There are no published papers discussing the function, structure, or properties of *CD108131*; however, much can be gathered from the UCSC genome browser data base as well as from preliminary data. *CD108131* is located on chromosome 18p11, and is 863bp in length. The transcript contains three regions containing repetitive elements, each distinct from one another. In addition, there is at least partial conservation of this sequence in several species, with high

levels of conservation seen amongst primates. There is also a region identified by DNase I hypersensitivity that indicates a possible site of protein binding. Preliminary data were able to show that *CD108131* is expressed in 20 human tissues, with the highest expression being in the cerebellum (Fig. 1.1). Additionally, *CD108131* does not appear to have an open reading frame that produces a functional protein, solidifying its status as a long non-coding RNA (lncRNA).

1.7 Functional characteristics of lncRNAs

It is estimated that two-thirds of the human genome is transcribed, 80% of which is non-coding (Novikova et al, 2013). LncRNAs are defined by a lack of coding potential and a length of 200 nucleotides or longer (Novikova et al, 2013). Within the human genome, there are 15,778 lncRNA genes which in turn produce 27,908 lncRNA transcripts (Morlando and Fatica, 2018). Most lncRNAs have a cellular localization within the nucleus, different from mRNAs which localize in the cytoplasm (Derrien et al, 2012). LncRNA genes can be intronic, intergenic, bidirectional, or antisense to protein-coding genes (Novikova et al, 2013). Interestingly, some transcripts have been found to function as both coding and non-coding, though none yet have been found in mammalian lineages (Ulitsky and Bartel, 2013). Enhancer RNAs and small nucleolar RNA hosts are also considered to be lncRNAs (Batista and Chang, 2013). Although lncRNAs do not get translated into proteins, many transcripts do undergo post-translational modifications. In fact, up to 30% of non-coding RNAs have at least one alternative transcript, and many undergo polyadenylation (Ponting, Oliver, and Reik, 2009). LncRNAs can form secondary and higher structures with a stability that allows for active roles in organization and regulation. These roles include DNA replication, RNA transcription, protein translation, signalling, development, differentiation, disease, epigenetics, and pluripotency (Novikova et al, 2013, Wang et al, 2017). Their regulatory effects can be seen on both cis and trans genes

(Ponting, Oliver, and Reik, 2009). While lncRNAs are observed in almost all tissues, 40% of known transcripts are specific to the brain (Derrien et al, 2012). Furthermore, lncRNA

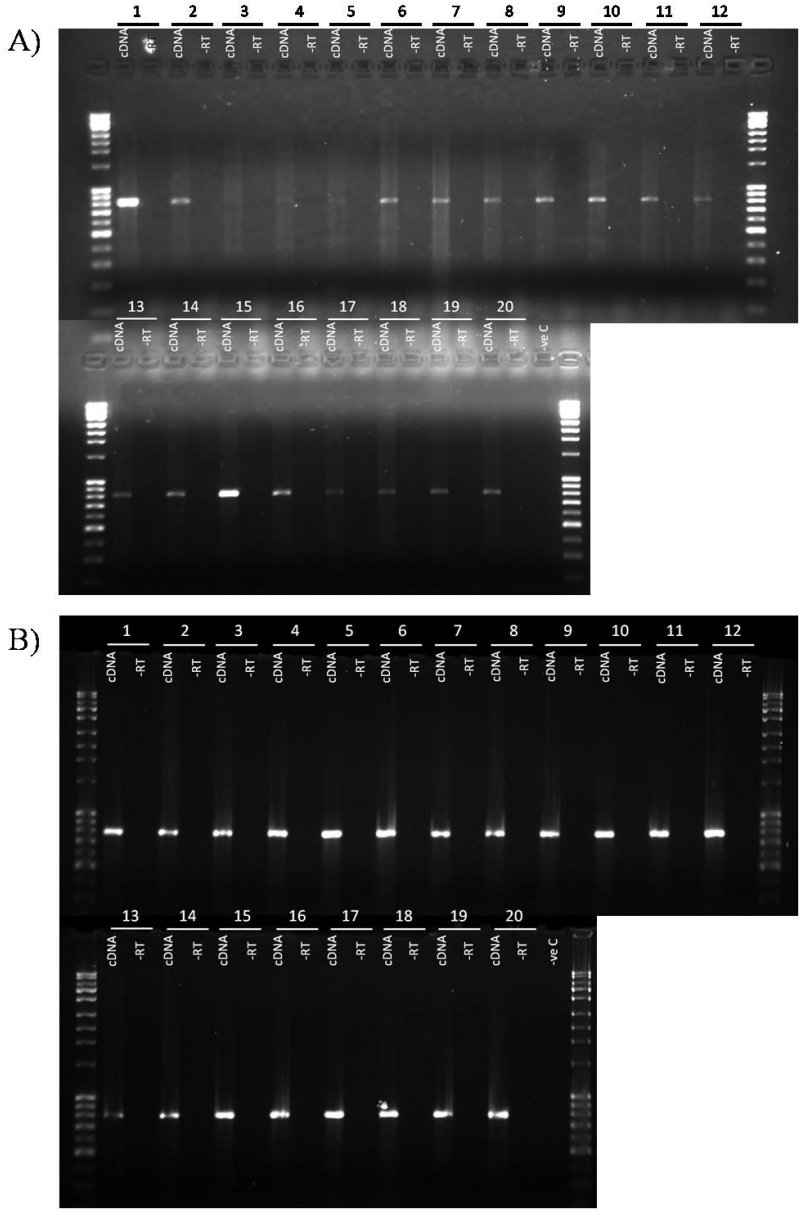

Figure 1.1. *CD108131* expression analysis in 20 human tissues. Gel photos of the PCR products from (A) *CD108131* and (B) *GAPDH* amplification of 20 different cDNA templates. The -RT controls are the total RNA samples prior to cDNA synthesis to ensure a lack of genomic DNA contamination. *GAPDH* serves as a positive control for successful cDNA synthesis. Total RNA was extracted from the following 20 tissues: 1 = cerebellum; 2 = whole brain; 3 = fetal brain; 4 = fetal liver; 5 = heart; 6 = kidney; 7 = adrenal gland; 8 = lung; 9 = placenta; 10 =

prostate; 11 = salivary gland; 12 = skeletal muscle; 13 = spleen; 14 = testis; 15 = thymus; 16 = thyroid gland; 17 = trachea; 18 = uterus; 19 = stomach; 20 = small intestine.

M. Vanstone, Thesis (2012)

transcript expression is more cell-type specific than that seen in protein-coding genes (Batista and Chang, 2013).

Sequence conservation is low between species, originally lending to the hypothesis that non-coding transcripts did not serve integral functions (Ponting, Oliver, and Reik, 2009). However, it has been found that many lncRNAs do show significant, although weak, signatures of natural selection when conservation is examined using whole genome alignments (Ulitsky and Bartel, 2013). For lncRNAs in which the promoter region has been identified, it has been noted that these promoter regions are under selective pressure similar to that of protein-coding gene promoters (Batista and Chang, 2013). Some studies of intergenic lncRNA families have even shown there to be lineage separations between lncRNA conservation, with 30% of transcripts being specific to primates, and 0.7% being specific to humans (Derrien et al, 2012).

LncRNAs can work through several mechanisms to regulate their targets, and the mechanism by which a single lncRNA functions can differ depending on tissue type and localisation (Morlando and Fatica, 2018). Four main models include acting as a decoy, scaffold, guide, or enhancer (Rinn and Chang, 2012). When acting as a decoy, the lncRNA will bind to DNA-binding factors, therefore impairing them from binding to the DNA itself. LncRNAs can also bind several proteins, bringing them into close proximity with one another, by means of forming a scaffold. Similarly, some transcripts will bind proteins as well as DNA, acting to guide proteins to their destination. Finally, it is also possible for lncRNAs to induce chromosome looping, much like enhancers do, to bring DNA and proteins into spatial proximity (Rinn and Chang, 2012). Through these mechanisms, lncRNAs have been shown to bind with DNA, transcription factors,

accessory proteins, and enzymes (Ponting, Oliver, and Reik, 2009). Furthermore, lncRNAs have been shown to play an important role in epigenetics through the recruitment and inhibition of chromatin modifying and remodelling complexes (Morlando and Fatica, 2018).

Genome-wide association studies have demonstrated that 43% of disease or trait-associated SNPs are found outside of protein coding genes, emphasizing the degree to which lncRNAs could be involved in human disease (Batista and Chang, 2013). There are cases in which lncRNAs are involved with disease through epigenetic silencing, as seen with ANRIL and HOTAIR which form scaffolds for chromatin modification complexes. The overexpression of either of these lncRNAs changes the chromatin landscape, enabling the initiation and progression of cancer (Wapinski and Chang, 2011). Alternatively, MALAT-1 is a lncRNA that plays an important role in synaptogenesis by regulating neuronal serine/arginine-rich splicing factors. However, increased expression of MALAT-1 is a marker for non-small-cell lung cancer, and is correlated with poor survival (Wapinski and Chang, 2011). Elevated levels of the lncRNA BACE1-AS have been linked to the pathogenesis of Alzheimer's disease (AD), where it seems that AD-related cell stress leads to the upregulation of BACE1-AS. BACE1-AS regulates the translation of BACE1, an enzyme that cleaves β-site amyloid precursor proteins (APP), therefore its upregulation results in higher APP processivity and toxic accumulation of Aβ plaques (Wapinski and Chang, 2011). Additionally, some lncRNAs, such as Gas5, can contribute to cellular health by regulating mechanisms of apoptosis (Wapinksi and Chang, 2011).

LncRNAs have also been associated with human myopathies and myogenesis. One such example is facioscapulohumeral muscular dystrophy (FSHD), which is caused by a contraction in the number of D4Z4 repeats within the 4q35 region, resulting in a loss of chromatin repression. The lncRNA DBE-T functions as a locus control element within this region that

promotes an active chromatin state. FSHD occurs when mutations to the DBE-T promoter region disrupt its regulation (Batista and Chang, 2013). Within Prader-Willi syndrome, which affects muscle development, the associated 15q11-q13 region contains a multitude of intronic lncRNAs with small nucleolar RNA tails (so called sno-lncRNAs). It has been implicated that these nuclear sno-lncRNAs create a domain whereby the splicing factor Fox2 is bound and enriched. Through this mechanism, the sno-lncRNAs are suggested to play a role in regulating splicing within this region (Batista and Chang, 2013). It has also been shown that lncRNAs located in the enhancer and distal regulatory regions of *MyoD* aid in recruiting transcription machinery to the nearby promoter regions to assist with chromatin reorganization and myogenin activation, respectively (Ballarino et al, 2016). With respect to myogenesis, two other lncRNAs have been shown to participate in human myoblast differentiation: linc-MD1 and lncMyoD. Linc-MD1 improves myogenesis by binding to miR-133 and miR-135 and preventing them from repressing their targets: mastermind-like transcriptional coavtivator-1, and myocyte enhancer factor 2C (Ballarino et al, 2016). Overexpression of linc-MD1 benefits myoblast differentiation, silencing causes a delay, and downregulation has been noted in patients with Duchene muscular dystrophy (Ballarino et al, 2016). LncMyoD is controlled by MyoD, and is necessary for myoblast differentiation. Through binding with IGF2 mRNA-binding protein 2 (IMP2), it disrupts IMP2's translational control and thus enables cell-cycle exit and terminal differentiation (Ballarino et al, 2016).

1.8 Suggestion of an interaction between *SGCE* and *CD108131*

Preliminary data demonstrated that the variant of *CD108131* possessing the 3 bp mutation did not alter the amount of *CD108131* transcript present nor that of the nearest gene, *ZBTB14*,

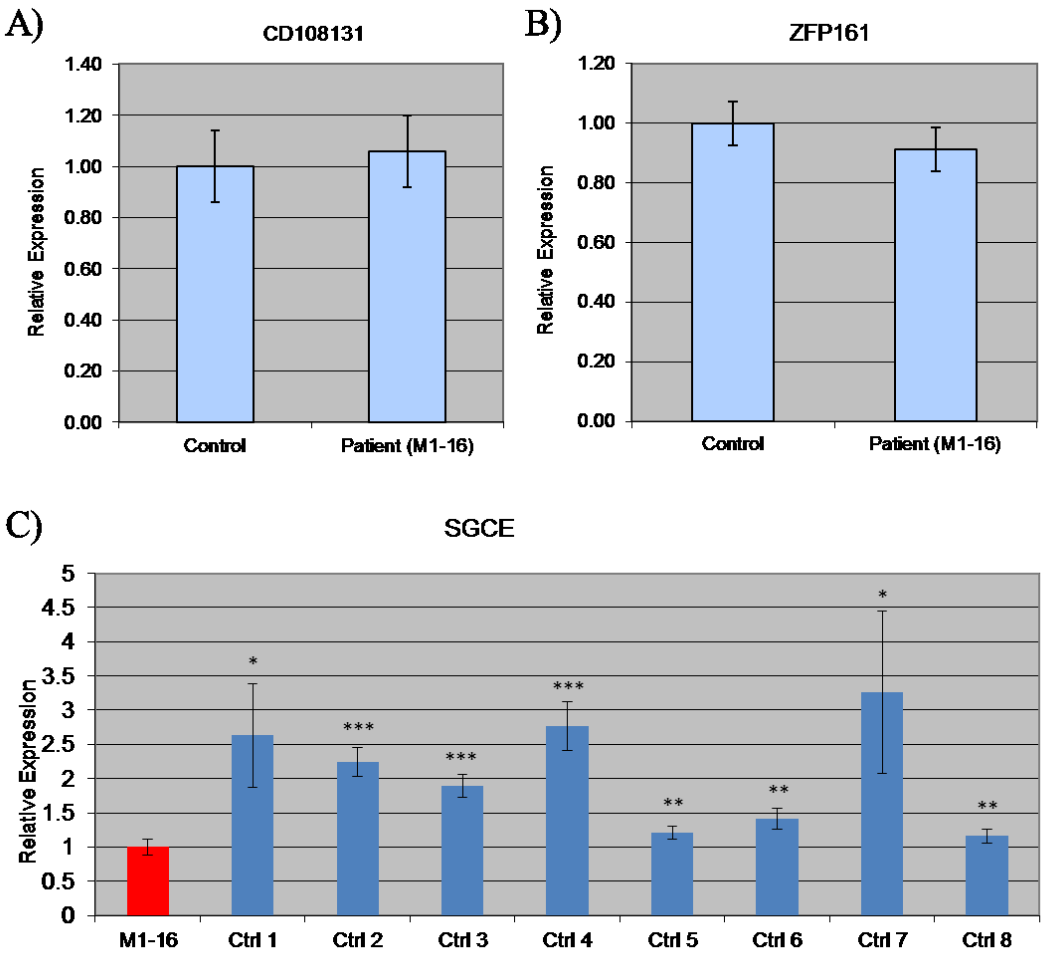

Figure 1.2. Comparing the relative expression of *CD108131*, *ZFP161*, and *SGCE* in patient and control cell lines. Results from qRT-PCR of patient M1-16 and control lymphoblast cDNA to quantify the expression of A) the *CD108131* transcript; B) *ZFP161*; and C) *SGCE*. For the *CD108131* and *ZFP161* analysis, the control and patient values represent an average of three replicates; the patient value is expressed as abundance relative to the control. Expression levels were normalized to that of β-actin. There were no significant changes. For the *SGCE* analysis, all samples in blue are controls external to MF1. The Ctrl 1, Ctrl 2, Ctrl 5, Ctrl 7, and Ctrl 8 values represent an average of three replicates; the Ctrl 3 and Ctrl 4 values represent and average of 4 replicates, and the M1-16 and Ctrl 6 values represent an average of 7 replicates. The control patient values are represented as a relative abundance to the MD patient, M1-16. Expression levels were normalized to a β-actin endogenous control. The significance of the difference in expression between each control and M1-16 is represented by asterix (*): p-value ≤ 0.05 (*); p-value ≤ 0.03 (**); p-value ≤ 0.003 (***).

NB: *ZFP161* was the previous gene name for *ZBTB14*

Vanstone, Thesis (2012)

however it did reduce the levels of SGCE expression (Fig. 1.2). Additionally, when a plasmid containing the antisense *CD108131* transcript was transfected into HEK293 cells, acting as a *CD108131* knockdown condition, SGCE protein levels were nearly completely knocked out. When the sense transcript was over-expressed, SGCE protein expression increased (Fig. 1.3).

It has since been found via the 1000 Genome Project that the 3 bp duplication found in this family is a variant that occurs at a frequency of approximately 1 in 10 000. Since MD occurs at a frequency of 1 in 100 000, this is no longer a likely candidate for causing the disease. Regardless of this, the apparent regulatory effect of *CD108131* on SGCE remains viable.

1.9 Rationale, hypothesis, and aims

1.9.1 Rationale

In December 2013, a new genome build, hg38, became available through the Human Genome Consortium, resulting in altered alignment in comparison to its predecessor. Furthermore, every year the UCSC Genome Browser is updated with new transcripts and annotations. For instance, between 2016 and 2017, the GENCODE gene track alone increased by 2604 transcripts (Tyner et al, 2017). Due to this increase in available information, as well as the multitude of different ways in which data can be analysed, it would be beneficial to re-examine the previous sequencing data. Additionally, although at present a clinical link can no longer be established between *CD108131* and MD, the observed regulatory effect of this lncRNA on SGCE should be explored due to the well-established relationship between SGCE and MD.

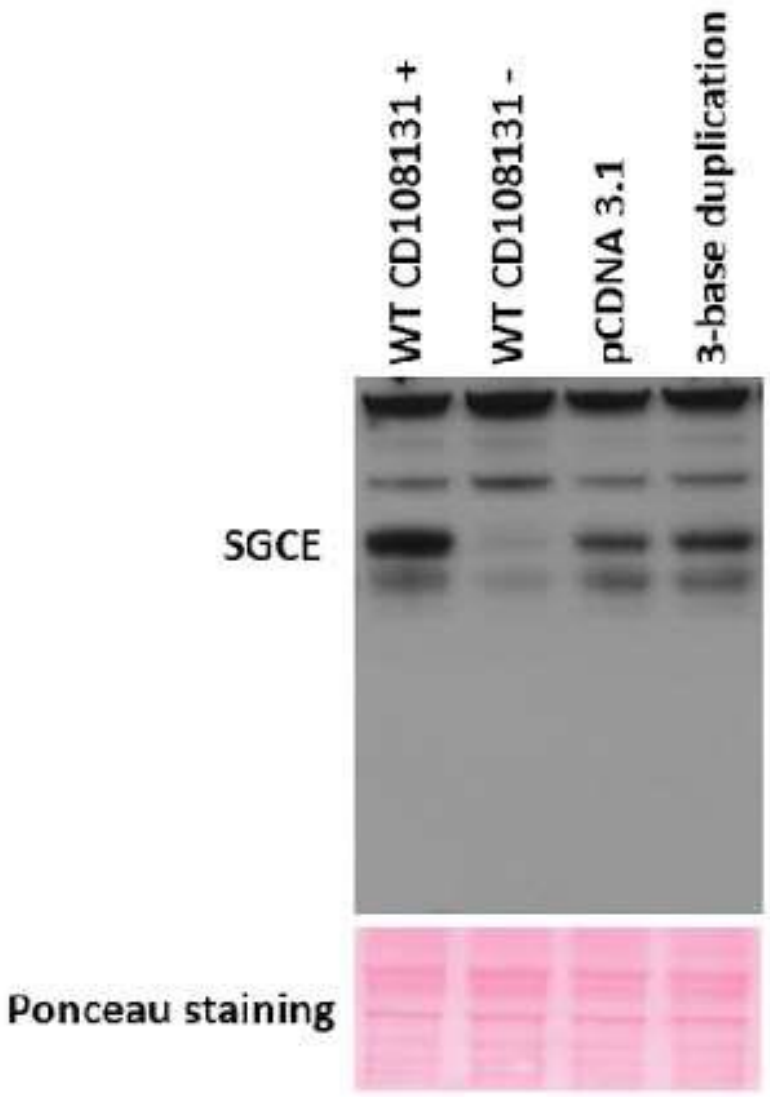

Figure 1.3. Analysis of SGCE protein expression. (Top) Western immunoblot of SGCE in HEK293 cells transfected with a pcDNA3.1 plasmid containing the sense transcript of wildtype *CD108131* (lane 1), the antisense transcript of wildtype *CD108131* (lane 2), no insert (empty vector control, lane 3), or the sense transcript of the mutant *CD108131* (3 bp duplication, lane 4). Transfection with the sense wildtype transcript results in an overexpression of *CD108131*, while transfection with the antisense wildtype transcript results in knockdown of *CD108131*.

(Bottom) Ponceau staining of the same membrane, demonstrating even loading across all lanes. Courtesy of Malcolm MacKenzie

1.9.2 Hypothesis

I hypothesise that through re-aligning and examining the Roche 454 sequencing conducted on the two patients within the family, missed genes or areas of low coverage may be revealed. The remaining potential variants within the critical region should be validated as one could prove to be disease-linked within this family. Additionally, I hypothesise that *CD108131* has a regulatory effect on *SGCE* at a transcriptional and/or translational level.

1.9.3 Aims

In order to address my hypotheses, I will achieve the following objectives:

1. Validate the remaining variants of interest, and re-examine the Roche 454 sequencing against the new Human Genome Consortium build (hg38).

Primers were designed to validate the remaining variants of interest via Sanger sequencing. Sequencing data was analysed using Geneious to determine whether the variant calls were false positives, or true variants. Through the use of the bioinformatics software CLC Genomics Workbench (v9), original Roche 454 sequencing results were realigned to the hg38 Genome build and analysed for regions with a read depth ≤ 10.

2. Develop a stable CRISPR/Cas9 knockout cell line for *CD108131*.

To build upon the preliminary data, and further examine potential effects of *CD108131* on *SGCE,* a double-cut CRISPR/Cas9 experiment was conducted in SH-SY5Y and HEK293 cells. Experimental design was such that a successful integration of the repair template would result in the complete removal of the entire known *CD108131* sequence. Potential clones were PCR genotyped and then confirmed using Sanger sequencing.

3. Determine the effects of a lack of *CD108131* on *SGCE* transcript expression. Additionally, determine the effects of a lack of *CD108131* on the transcript expression of select other genes, namely *ZBTB14* (the nearest coding gene to *CD108131*) and *RPPH1* (a housekeeping gene).

Upon successful identification of *CD108131* knockout (KO) clones, the relationship between *CD108131* and *SGCE* had to be characterised. After extracting RNA from all cell lines, qPCR analysis was conducted to compare relative transcript expression levels of *SGCE, ZBTB14*, and *RPPH1* between KO and control cell lines.

Chapter 2: Screening of mutation calls from original sequencing results

2.1. Introduction

Two independent linkage analyses were conducted on our 5 generation Canadian Myoclonus-Dystonia family (MF1), both revealing a 3.18 Mb disease-linked region on chromosome 18p11 (Fig. 2.1) (Grimes *et al.*, 2002; Han *et al.*, 2007). DNA from two affected patients, MF1-3 (number on pedigree: 39) and MF1-5 (number on pedigree: 47), were chosen from the pedigree to undergo targeted capture of the disease-linked region, followed by sequencing by Roche 454. Meiotically distant patients were selected as they would have fewer common familial variants (Vanstone, 2012). Roche 454 was the chosen sequencing platform because, at the time, it produced longer read lengths (~400 bp) when compared to competing sequencing platforms. To enable comprehensive screening of all potential mutations, the entire critical region was sequenced, including introns and intergenic regions.

NimbleGen Sequence Capture was used to enrich for DNA from the 3.18 Mb critical region, plus 50 kb flanking either side, with an array design which allowed for >97% coverage of the target sequence (Vanstone, 2012). The methodology involved fragmenting the genomic DNA and ligating linkers to the ends of each fragment, before hybridizing the fragments to a custom array containing oligonucleotide probes specific to the region of interest. Non-specific fragments were washed away before the target fragments were eluted and amplified using primers specific to the linker sequences. Amplified samples could then undergo sequencing. During the Roche 454 protocol, the template DNA is fragmented and universal adaptors are ligated to the ends, before fragments are individually bound to independent beads and captured in oil droplets containing all necessary PCR reagents. Each template fragment is then amplified via emulsion polymerase chain reaction (PCR), before breaking the emulsion and loading beads into the wells

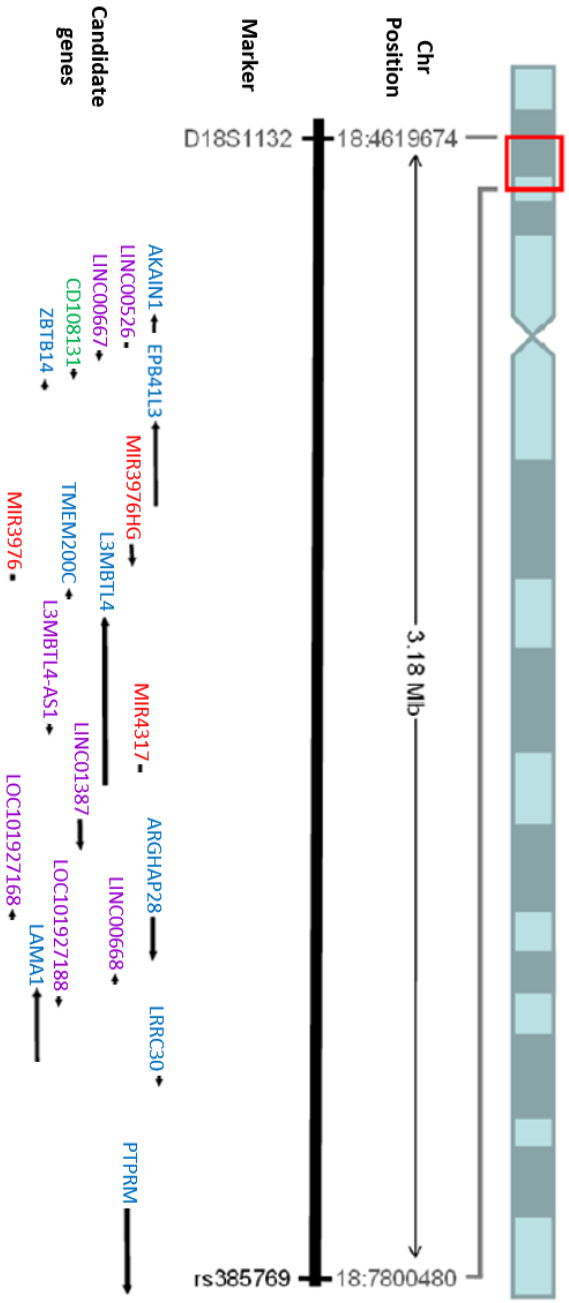

Figure 2.1. Schematic representation of DYT15, a 3.18 Mb region on chr18p11. Diagram of candidate genes, their location, as well as their orientation determined using the UCSC genome browser (http://genome.ucsc.edu/). Genes listed are those that have been reviewed by SwissProt, RefSeq, or both (with the exception of *CD108131*). The genes are categorized based on function: blue denotes protein-coding genes, purple denotes non-coding RNAs, and red denotes microRNAs. *CD108131*, coloured green, is viewed through the Human EST track, and possesses signature features of a lncRNA. This map is based on the GRC38/hg37 build.

of a fibreoptic PicoTitre Plate. At this point separate beads containing the sequencing reagents are added to the wells. Nucleotides are then introduced one at a time, and a successful incorporation will result in an inorganic phosphate ion being released and initiating an enzyme cascade that culminates in the emission of light. The intensity and location of emitted light will be measured, and as such, the template sequence for each fragment can be constructed (Margulies et al., 2005; Moorthie et al., 2011).

Upon sequencing the disease-linked region in both patients, all sequence reads were trimmed of the adaptor sequences, the sequencing primers, and the enrichment primers. The trimmed sequence reads were aligned to the entire human reference sequence Genome Reference Consortium: Human build 37 using NextGENe™ software from SOFTGENETICS. Separate alignments for each patient were performed using the following parameters: matching base percentage ≥ 90.0; mutation percentage ≤ 20.00; allowable mismatched bases = 4; allowable ambiguous alignments = 1; seed = 17 bases; move step = 7 bases (Vanstone, 2012). A mutation report was generated for each patient, and all variant calls were filtered for known SNPs, before being compared between patients. Since MD is an autosomal dominant disorder, all homozygous variants were removed from the shared variant list, leaving only novel, heterozygous, shared variants.

In order for a variant to be concluded as disease-causing, it would have to: validate via Sanger sequencing to determine that the call was not a false-positive, co-segregate with all affected and obligate carrier patients within the pedigree, be absent in unaffected family members and control samples (~300 control chromosomes). To streamLine validation, variants were assigned one of the following priority categories, listed from top priority to lowest priority:

coding variants, variants within splice sites, intronic variants, structural variants, variants occurring within expressed sequence tags, and intergenic variants (Vanstone, 2012).

Upon discovering the segregation of the 3 bp duplication within *CD108131*, the remainder of the variant calls were not validated. As we now know that the duplication found in *CD108131* is unlikely to be disease-causing, the genetic cause of this family's MD remains unknown. In an attempt to further validate the original sequencing results, the remaining variants of interest have been validated by Sanger sequencing, and a coverage analysis was performed.

2.2. Materials and Methods

2.2.1. Annotation of variant calls

There remained 34 variants of interest to be validated. To assess these variants, I examined their locations using the updated University of California Santa Cruz (UCSC) genome table browser (Hg38, Gencode v24) to determine their functional classification, biotype, as well as any other annotations of interest. All variants were intronic, and were either within protein-coding genes, or within lncRNAs that were antisense to protein-coding genes. Variants that were located within ESTs, enhancer regions, promotor regions, or binding sites were prioritised over those that overlapped with repetitive elements (SINEs, LINEs, and homopolymers).

2.2.2. Analysis of low coverage regions

The original sequencing files were re-examined in order to conduct further analysis. All reads were aligned to the new Genome Reference Consortium: Human build 38, and filtered for heterozygous mutations. Using the default parameters of CLC Genomics Workbench (v9), with the exception of changing the default pyro-error homopolymer length from 3 to 6, variant calls that resulted in amino acid changes were examined. Next, University of California Santa Cruz

(UCSC) genome table browser (Hg38, Gencode v24) was used to identify all coding exons within the critical region. A coverage report for each patient was compiled, and custom tracks were created in UCSC of each patient's coverage gaps that were at a read depth ≤ 10. These were further compared to the coding genes annotated by the Consensus Coding Sequence (CCDS) Project.

2.2.3. Primer design and PCR

For each region of interest, the genomic DNA sequence was obtained from the UCSC Genome Brower (Hg38). Primers were then designed using NCBI's Primer-BLAST tool. This tool incorporates the Primer3 PCR primer design tool, with NCBI's BLAST global alignment algorithm to screen primers for non-specific amplification. Primer-BLAST retrieved primers using the standard parameters, with the following changes: amplicon size was set to a maximum of 500 nts, maximum primer length was increased to 28 nts, SNP handling was set to not design primers over known SNPs, and the repeat filter was set to "Human". In cases where the region to be sequenced involved significant and unavoidable repetitive sequences, the low complexity filter was turned off allowing for primers to be designed in areas of biased base compositions.

PCR was performed using recombinant Taq Polymerase, PCR reaction buffer, Mg^2, and dNTPs from Thermo Fischer Scientific. Each primer set was optimized individually using control DNA, by varying the annealing temperature and/or cycle number. After optimization, primer sets were used to amplify the desired region in DNA from a subset of affected patients, asymptomatic carriers, and unaffected married-in controls.

2.2.4. PCR product verification and sequencing

Following PCR, products were separated by agarose gel electrophoresis in order to confirm amplicon size and specificity. For all successful amplifications, approximately 10μL of PCR product then underwent Sanger sequencing in-house using the associated forward and reverse primers at a concentration of 5mM. Sequencing results were aligned with the human reference sequence and analysed using the Geneious™ Sequence Analysis Software from Biomatters Ltd. Each sequencing result was also input into UCSC's BLAT alignment tool to ensure specificity to the chromosomal location of the amplified region.

2.3. Results and Discussion

2.3.1. Status of remaining variants of interest

In an attempt to further validate the original sequencing results, the remaining 34 variants of interest were validated by Sanger sequencing. All of these variants were found in intronic regions of known coding or non-coding genes, and possessed overlapping genomic annotations, such as being within: the exon of an EST, a promoter/ enhancer region, a DNase I hypersensitivity region, a known binding site, a repressor site, a transcription elongation site, or a conserved region (Table 2.1). While BLAT results confirmed that all sequence results were specific to each variants' location on chromosome 18, all validated variants were found to be false positives. Of these 34 variants, seven were unable to be validated due to an inability to design primers for that chromosomal location. All seven of these variants overlapped with repetitive sequences.

Chr	Start hg19	End hg19	Ref	Alt	Classification	Gene ID	Biotype	Overlapping genomic annotations
18	5447474	5447474	-	T	intronic	EPB41L3	protein coding	Within exon of ESTs AW023410 and DB358203
18	5447491	5447491	A	-	intronic	EPB41L3	protein coding	Within exon of ESTs AW023410 and DB358203
18	5568423	5568423	-	A	ncRNA_intronic	RP11-286N3.2	antisense to EPB41L3	Within exon of ESTs AW083859 and CV369160
18	6031807	6031807	-	T	ncRNA_intronic	RP11-793A3.2	antisense to L3MBTL4	Within exon of mRNA BC039316, Within RNA protein binding site ELAVL1, PABPC1, SLBP
18	7598802	7598802	-	T	intronic	PTPRM	protein coding	Within exon of EST H87954, within HMM Weak Enhancer, within TF binding site MAFK, within DNAse I HS, within RNA binding protein site ELAVL1, IGF2BP1, PABPC1, SLBP
18	5452802	5452802	-	TA	intronic	EPB41L3	protein coding	Within HMM Strong Enhancer, STR, overlaps with rs74270590
18	5457549	5457549	-	A	intronic	EPB41L3	protein coding	Within HMM Strong Enhancer, exon of EST BY795823 , overlaps with rs543713220 (C>A/T)
18	6526199	6526199	-	T	intronic	LINC01387	lincRNA	Within CEBPB binding site
18	6742185	6742185	-	T	intronic	ARHGAP28	protein coding	**Within SINE**, within HMM Strong Enhancer
18	6747088	6747088	-	T	intronic	ARHGAP28	protein coding	Within HMM Strong Enhancer, Weak Enhancer, Repressed State
18	7075004	7075004	T	-	intronic	LAMA1	protein coding	Within HMM Weak Enhancer, Weak Promoter, Insulator, within RNA binding protein site PABPC1
18	7609541	7609541	-	T	intronic	PTPRM	protein coding	Within HMM Strong Enhancer, within RNA binding protein sites ELAVL1, IGF2BP1, PABPC1, SLBP
18	7683182	7683182	G	-	intronic	PTPRM	protein coding	Within HMM Weak Enhancer, within RNA binding protein sites ELAVL1, IGF2BP1, PABPC1, SLBP, within SINE
18	6729491	6729491	A	G	ncRNA_intronic	RP11-9118.3	antisense to ARHGAP28	Within HMM active promoter, weak promoter, weak enhancer, **simple repeat (CA)n,** some in-vivo TF binding, DNAse I HS
18	6729500	6729500	C	G	ncRNA_intronic	RP11-9118.3	antisense to ARHGAP28	Overlaps with rs201923574 (ACAG>del), **within STR (CA)n**, Strong POL2RA binding site, slightly conserved
18	6729485	6729486	GT	-	ncRNA_intronic	RP11-9118.3	antisense to ARHGAP28	Within HMM active promoter, weak promoter, weak enhancer, **simple repeat (CA)n,** some in-vivo TF binding, DNAse I HS
18	5556975	5556975	G	T	intronic	EPB41L3	protein coding	Within HMM Weak Enhancer and highly divergent position
18	6038248	6038248	-	T	ncRNA_intronic	RP11-793A3.2	antisense to L3MBTL4	Within RNA protein binding site ELAVL1, PABPC1, SLBP
18	6311112	6311112	-	T	intronic	L3MTLB4	protein coding	**Within LTR**, within HMM TxN elongation, within RNA binding protein site ELAVL1, PABPC1, SLBP
18	6797681	6797681	-	T	intronic	ARHGAP28	protein coding	**Within SINE**, within RNA binding protein site IGFBP2
18	6823107	6823107	T	-	intronic	ARHGAP28	protein coding	Within RNA binding protein site IGFBP2
18	6885818	6885818	-	T	intronic	ARHGAP28	protein coding	Within HMM Weak Enhancer, within RNA binding protein site IGF2BP1
18	6893884	6893884	G	-	intronic	ARHGAP28	protein coding	**Within SINE**, within RNA binding protein site IGF2BP1
18	7041152	7041152	-	A	intronic	LAMA1	protein coding	Within RNA binding protein site PABPC1
18	7075951	7075951	-	A	intronic	LAMA1	protein coding	Within RNA binding protein site PABPC1
18	7092621	7092621	-	A	intronic	LAMA1	protein coding	**Within SINE**, HMM Repressed State, within RNA binding protein site PABPC1
18	7681579	7681579	-	A	intronic	PTPRM	protein coding	Slightly conserved, within RNA binding protein site ELAVL1, IGF2BP1, PABPC1, SLBP
18	7720712	7720712	A	C	intronic	PTPRM	protein coding	**Within LINE**, RNA binding protein site ELAVL1, IGF2BP1, PABPC1, SLBP, slightly conserved position
18	7795466	7795466	-	A	intronic	PTPRM	protein coding	Slightly conserved position, within RNA binding protein sites ELAVL1, IGF2BP1, PABPC1, SLBP
18	7033453	7033453	-	A	intronic	LAMA1	protein coding	Within HMM Transcription Elongation, **within SINE**, within RNA binding protein site PABPC1
18	7056075	7056075	-	A	intronic	LAMA1	protein coding	**Within SINE**, RNA binding protein site PABPC1, **near homopolymer (A)n**
18	7108639	7108639	C	-	intronic	LAMA1	protein coding	**Within SINE**, HMM Repressed State, within RNA binding protein site PABPC1, **near homopolymer (A)n**
18	7108638	7108638	-	A	intronic	LAMA1	protein coding	**Within SINE**, HMM Repressed State, within RNA binding protein site PABPC1, **near homopolymer (A)n**

Table 2.1. List of 34 novel, shared, heterozygous variants of interest tested via Sanger sequencing. Table containing the remaining 34 variants of interest from the previous analysis of the Roche454 sequencing data of of patients MF1-3 (number on pedigree: 39) and MF1-5 (number on pedigree: 47). All variants were filtered to remove any variants that were not novel, heterozygous, or shared between the two patients. Classification, Gene ID, and Biotype was determined using Gencode v24. Rows shaded in grey denote variants for which primers could not be designed, presumably due to overlapping repetitive sequences. White rows denote variants that were successfully sequenced through Sanger sequencing. Within the "Overlapping genomic annotations" column, bolded text relates to repetitive elements, and red text denotes known SNPs. All sequenced variants were found to be false positives.

2.3.2. Status of low coverage regions

Upon analysis of each patient's coverage report, several regions of low coverage were found, many of which were shared between both sequenced individuals (Table 2.2). Once these coverage gaps were compared to the CCDS data, it was revealed that both patients lacked sufficient coverage for the entire coding sequence of *TMEM200C*. To ensure that this had not resulted in any missed variant calls, over-lapping primers were designed to allow for Sanger sequencing of the region. Ultimately, it was found that there were no shared, heterozygous mutations within this gene.

2.4. Conclusion

At this point in time the causative mutation for MD within our family is still unknown. The remaining candidates have been shown to be false positives, or unable to be sequenced using our current methods. The most likely obstacle for complete sequencing of the critical region is the existence of multiple repetitive elements. Repetitive sequences, particularly homopolymer strings, are prone to causing pyro-errors during sequencing. Not only were many of the validated variants located within repetitive sequences, all seven of the variants for which primers could not be designed were as well (as denoted in Table 2.1). For this reason, it can be assumed that these seven variants were also false positives. Additionally, although the original parameters for variant calls required they not overlap with known SNPs, when using the updated human genome build and annotations, three of the validated variants now overlapped with known SNPs. This also could have impacted variant call accuracy at these locations.

	Number of low coverage blocks
Patient 1	148
Patient 2	152
Shared	146

Table 2.2. Identification of low coverage blocks within sequenced patients. Roche454 sequencing data of the critical region for patients MF1-3 and MF1-5 was reanalysed to create coverage reports using CLC Genomics Workbench (v9) and UCSC's Genome Browser (Hg38). The sequencing files were titled "Patient 1" and "Patient 2", and as such those will be the identifiers in this table. The coverage report identified low coverage regions within known coding genes, where low coverage was defined by a read depth of less than 10. Low coverage data from each patient was compared to identify shared blocks of low coverage.

Poor read depth was a potential reason for why the causative mutation had yet to be found within the Roche454 sequencing data of the critical region. However, while it was discovered that the initial sequencing had insufficiently covered the coding sequence of *TMEM200C*, this gene is no longer of interest due to a lack of potential disease-causing variants being identified upon Sanger sequencing. Although the focus has been on exons of coding genes, the coverage data did reveal some shared intronic regions of low coverage. As non-coding RNAs can often be located within intronic regions of coding genes, it may be of future interest to sequence these low coverage introns within the critical region.

Chapter 3: Generation of a *CD108131* knockout using CRISPR-Cas9

3.1. Introduction

For several decades now, researchers have striven to develop efficient and precise methods for genome editing, that is, the ability to alter the genetic sequence of DNA in a controlled and predetermined manner. In the 1980s, targeted gene disruption experiments conducted in yeast and mammalian cells aimed to integrate exogenous sequence into the genome. These attempts yielded low and variable efficiency, as well as poor specificity (Rothstein, 1983, Smithies et al, 1985, Thomas et al, 1986). The discovery that double-stranded breaks (DSB) introduced at the target site resulted in a substantial increase in integration frequency shifted the focus to meganucleases, which would induce DSBs at specific locations. However, the chances of a meganuclease cut-site being available for a desired target were low, and non-homologous end joining (NHEJ) was the predominant repair mechanism – a mechanism known for its propensity to result in random indels at the cut-site (Rouet et al, 1994, Jeggo, 1998). This is in contrast to the more desirable homology directed repair (HDR), whereby the DNA repairs using a homologue piece of DNA as the template, whether endogenous or introduced.

In a continued attempt to improve specificity and accessibility of directed gene editing, two new techniques emerged: zinc finger nucleases (ZFNs), and transcription activator-like effector nucleases (TALENs). ZFNs are derived from zinc finger protein motifs which recognize and bind to sequence-specific 3bp DNA sequences. Multiple zinc finger proteins can be combined to increase DNA binding specificity to the target region. The distinct DNA cleavage domain of the Fok I endonuclease, an enzyme that required homodimerization in order to cleave, was chimerically fused to the zinc finger modules. By designing two separate ZFNs with

proximal binding sites, Fok I is able to homodimerize and induce a DSB at the desired genomic location (Bibikova et al, 2001). TALENs make use of bacterial transcription activator-like effector (TALE) proteins, which recognise and bind DNA at single base-pairs, as opposed to the three needed for ZFNs. Similar to ZFNs, TALENs involve the assembly of multiple TALEs to create a specific target sequence, and the fusion of the Fok I DNA cleavage domain for endonuclease activity (Li et al, 2011).

Concurrently, strides were being made to better understand and manipulate a newly discovered aspect of certain bacterial immune responses: clustered regularly interspaced short palindrome repeats (CRISPRs). CRISPRs were originally discovered as repetitive sequences of 25-50 nts, separated by spacers of about 30 nts (Bolotin et al, 2005). Although initially thought to be unique sequences, it was soon revealed that the spacer sequences derived from pre-existing DNA sequences – specifically, those of bacteriophages and conjugative plasmids (Mojica et al, 2005). Genes encoding for CRISPR-associated (Cas) endonucleases are found in operons adjacent to CRISPR loci (Bolotin et al, 2005). Ultimately, three different CRISPR/Cas systems were identified and shown to function as part of the adaptive immunity within bacteria and archaea (Jinek et al, 2012).

One system in particular, type II CRISPR/Cas9, was found to have mechanisms that could translate to programmable, directed gene editing. Endogenously, this immune mechanism is executed in three phases: adaptation, expression, and interference. During the adaptation phase, a Cas1/Cas2 complex will identify a protospacer adjacent motif (PAM), a sequence that varies depending on the species of origin, within the invading DNA sequence. Upon identification, it will cleave the foreign DNA and incorporate this fragment into the host CRISPR loci as a spacer (Rath et al, 2015). The expression phase then commences with the transcription

of the CRISPR locus containing the spacers, producing the primary CRISPR transcript (pre-crRNA) (Rath et al, 2015). A complex of trans-activating crRNA (tracrRNA), Cas9, and RNase III is then formed to process the pre-crRNA. The 3' terminus of the resultant crRNA binds to the 5' terminus of the tracrRNA, forming a secondary structure that complexes with Cas9. This effector complex then binds to the PAM sequence and complementary DNA target, and induces a DSB (Mohanraju et al, 2016). Cas9 is capable of cleaving both strands of foreign DNA as it possesses domains homologous to both HNH- and RuvC-like endonucleases (Jinek et al, 2012).

Interestingly, chimeric RNAs consisting of the 3' end of the crRNA fused to the 5' end of the tracrRNA were able to bind and direct Cas9 with similar efficiency to wildtype crRNA:tracrRNA (Jinek et al, 2012). These chimeric RNAs were dubbed small guide RNAs (sgRNAs) and provided an alternative programmable genome-editing tool. Where ZFNs and TALENs worked via protein-guided DNA cleavage, CRISPR/Cas9 allowed for cleavage guided by small RNAs. Through the use of sgRNAs and CRISPR/Cas9, any sequence can be targeting so long as a PAM sequence is located directly upstream. The most commonly appropriated CRISPR/Cas9 system is that of *S. pyogenes*, in which the PAM sequence is NGG (where "N" can be any nucleotide) (Jinek et al, 2012). In the absence of a PAM sequence even fully complementary sequences will not be recognised by Cas9, however these NGG PAM sites are located approximately every 8 nts, and as such do not pose a major design limitation (Doudna and Charpentier, 2014, Zhang et al, 2014). The guiding sequence of the sgRNA can then be engineered to consist of approximately 20 nts complimentary to the target sequence, resulting in a DSB. One major advantage of CIRSPR/Cas9 is that it is relatively easy to design and produce multiple sgRNAs, therefore allowing for the simultaneous induction of multiple independent genomic modifications (Zhang et al, 2014). The biggest challenge facing CRISPR/Cas9 is the

propensity for off-target mutations to occur. While the sgRNAs with direct cleavage to the target sequence, large genomes are known to have multiple DNA sequences that match identically or very closely to that of the target (Zhang et al, 2014). SgRNA design tools have been made to help minimise the chance of off-target effects. Nonetheless, CRISPR/Cas9 editing efficiency have been shown to be as high or higher than that of ZFNs or TALENs (Doudna and Charpentier, 2014).

3.2. Materials and Methods

3.2.1 CRISPR-Cas9 *CD108131* knockout design

In order to determine what effect a lack of *CD108131* might have on SGCE, it was first necessary to develop a *CD108131* knockout (*CD108131-KO*) cell line. The most straight-forward applications of CRISPR/Cas9 involve disrupting promoter regions, transcription start sites (TSS), or the known active domains of a gene. The easiest way to do so is by inducing a DSB, and allowing for a random mutation to occur via NHEJ. Alternatively, a repair template containing a specific mutation flanked my homologous sequence can be transfected into the cell along with the CRISPR/Cas9 machinery. When repair occurs via HDR, this known mutation will be incorporated into the target region. However, since the regulatory and active regions of lncRNAs remain largely unknown, a small-scale disruption would be unlikely to affect the function of the transcript. Therefore, to create a *CD108131* knockout, CRISPR/Cas9 was used to delete the entirety of the known *CD108131* gene sequence (Fig. 3.1). To facilitate the excision of the whole gene, two DSBs were induced, one upstream and one downstream of *CD108131*. The sgRNAs were designed using two different design tools, DNA2.0 and CHOPCHOP, using input sequences consisting of 250bp upstream or downstream of *CD108131*, depending on the target site. This assured a long enough stretch of sequence to find the necessary PAM sites, as well as

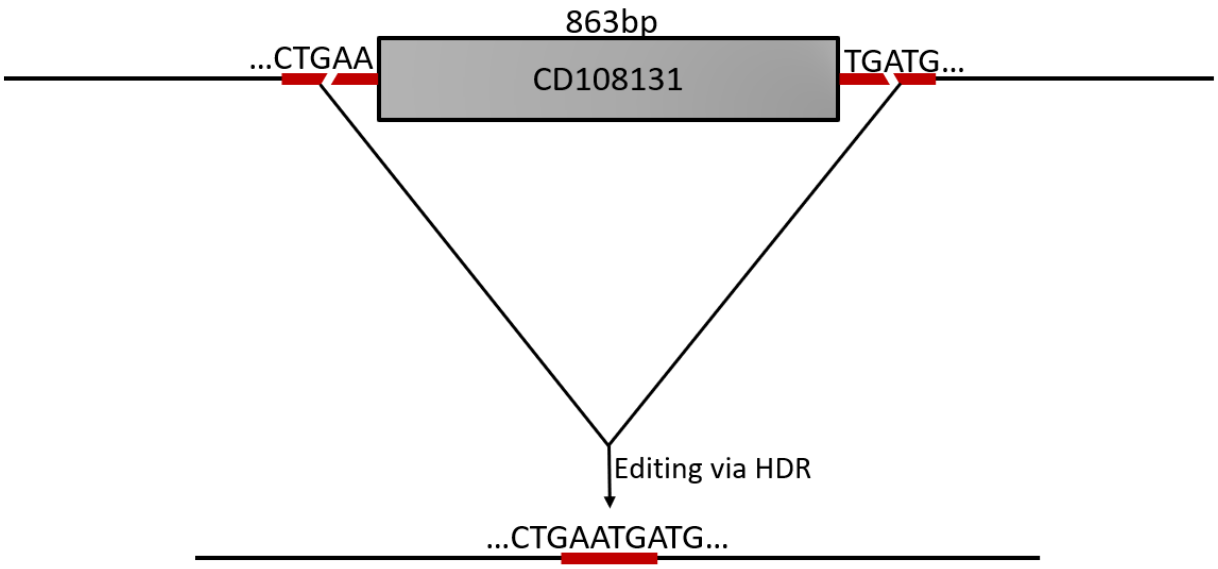

Figure 3.1. Experimental design for CRISPR/Cas9 knockout of *CD108131*. *CD108131*, denoted as the grey rectangle, is 863bp in length. To excise the entire gene, input sequences of 250bp upstream or downstream of *CD108131*, shown in red, were input into sgRNA design software. SgRNA cutsites are represented by the break in the red sequence, and are not to scale. Through HDR, *CD108131* would be removed from the genome, and the gBlock homology amplicon would be inserted. Successful gBlock insertion would return sequencing results of "…CTGAATGATG…" at the former location of the *CD108131* gene.

appropriate target, without interfering with other known gene elements. An upstream and downstream sgRNA were designed using each design tool (Table 3.1). Additionally, to guide HDR, a template comprised of upstream and downstream homology arms was designed and ordered through Integrated DNA Technologies (IDT). While originally aiming for approximately 800bp of homology in either direction, this was not supported by IDT due to complexity issues such as repetitive sequences, hairpin structure formation, or a GC content of over 65%. To conform to IDT design parameters, the synthetic gBlock amplicon was made of 950bp upstream of the *CD108131* gene directly leading into the 450bp downstream of the gene. The gBlock was requested to be phosphorylated to facilitate ligation.

3.2.2 sgRNA vector generation

The PX458 plasmid designed by the Zhang lab and acquired through Addgene, was the backbone used to generate the sgRNA vectors. This ampicillin-resistant, GFP-expressing plasmid contains both the Cas9 gene as well as a sgRNA scaffold. For each sgRNA, the digestion-ligation protocol outlined by Ran et al. (2013) was performed using the BbsI restriction enzyme and the forward and reverse oligos at a concentration of 33nM. Plasmids were transformed into DH5α *E. coli* bacterial cells, and plated onto ampicillin agar plates. After incubating overnight at 37°C, colonies were picked and grown in liquid Lysogeny broth (LB) with 1:1000 ampicillin overnight in a 37°C shaker. Each of the picked colonies was also streaked onto one square of an ampicillin agar grid plate, which was then incubated overnight at 37°C. DNA was subsequently extracted from the bacterial cultures using the QIAprep Spin Miniprep Kit (Qiagen) and digested using BbsI and PsiI restriction endonucleases. Following digestion, DNA was separated by size via agarose gel electrophoresis to determine banding pattern.

Design tool	Cut location	sgRNA oligos
CHOPCHOP	Upstream of CD108131	sgRNA1_a: CACCGGTAATGTAAGAGGAAGGTCA sgRNA1_b: AAACTGACCTTCCTCTTACATTACC
	Downstream of CD108131	sgRNA3_a: CACCGGGAATGCTCTTTAGGAAGCG sgRNA3_b: AAACCGCTTCCTAAAGTGCATTCCC
DNA2.0	Upstream of CD108131	sgRNA2_a: CACCGCATTGTATTGGCACTGATTG sgRNA2_b: AAACCAATCAGTGCCAATACAATGCG
	Downstream of CD108131	sgRNA4_a: CACCGCCCCAGCATCGACACTTTTT sgRNA4_b: AAACAAAAAGTGTCGATGCTGGGC

Table 3.1. SgRNA sequences designed using CHOPCHOP and DNA2.0 design software. For sgRNAs designed to cut upstream of the *CD108131* gene, the 250bp immediate upstream of the *CD108131* gene sequence were input into the software. Similarly, for sgRNAs designed to cut downstream of *CD108131*, the 250bp immediately downstream from the gene were used. Default parameters were used for both design tools. Oligo sequences were ordered as listed, from IDT.

Insertion of the sgRNA oligo disrupts the BbsI cutsite, and as such plasmids with an insert would only be cut by PsiI. Therefore, plasmids with an insert would display as a single band, while those without would display as two bands. All plasmids demonstrating presence of an insert were sequenced using a U6 primer. Sequencing data were reviewed using Geneious software to determine correct insertion of the sgRNA oligos. When a plasmid was found to have the correct insertion, bacteria from the corresponding square of the grid plate was used to grow up a large liquid culture. DNA was then extracted from this culture using the Plasmid Maxi Kit (Qiagen) and concentration was determined via nanodrop.

3.2.3 Repair template vector generation

A standard ampicillin-resistant cloning vector, pBluescript II SK+, was digested with an EcoRI restriction enzyme and then ligated with the gBlock amplicon. The ligation reaction was performed through a short 3-minute incubation at 42°C to denature any looping present that may interfere, followed by overnight incubation at 16°C. Plasmids were transformed into Stbl3 *E. coli* bacterial cells, and plated onto ampicillin agar plates. After incubating for approximately 18 hours at 30°C, colonies were picked and grown in liquid LB with 1:1000 ampicillin for 18 hours in a 30°C shaker. A grid plate of all picked colonies was created and incubated for 18 hours at 30°C. DNA was subsequently extracted from the bacterial cultures using the QIAprep Spin Miniprep Kit (Qiagen) and digested using the EcoRI and HindIII enzymes, before being run on a DNA electrophoresis agarose gel. Plasmids that displaying both the 3 kb band of the vector, as well as the 1.4 kb band of the gBlock insert, were sequencing using T7 and T3 primers. Sequencing data were reviewed using Geneious software to determine correct insertion of the gBlock amplicon. Large liquid cultures of those plasmids found to have the correct insertion

were grown from the grid plate streaks, and DNA was extracted using the Plasmid Maxi Kit (Quagen). DNA concentration was determined via nanodrop.

3.2.4 Cell culture and maintenance

HEK293 cells were obtained from liquid nitrogen storage and thawed in a 37°C water bath. Once thawed, cultures were transferred to 15mL conical tube and centrifuged at 1500 rpm for 2 minutes at 4°C. Freezing media was then aspirated, and the cell pellet was resuspended in 2mL of Dulbecco′s Modified Eagle′s Medium (DMEM), supplemented with 10% fetal bovine serum (FBS), 1% L-glutamine, 1% non-essential amino acids (NEAA), and 1% penicillin/streptomycin. The resuspended cell mixture was added to a 10cm culture plate containing 8mL of DMEM (10% FBS, 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin), and incubated at 37°C and 5% CO_2. Cells were monitored every 2 days, and given 10mL of fresh culture medium when the pH of the current medium dropped (as indicated by a colour change from pink to orange). Whenever cell confluency reached 80-90%, culture media was aspirated and the cells were washed with approximately 3mL of phosphate-buffered saline (PBS) (warmed to 37°C). After aspirating the PBS, 2mL of 37°C trypsin was applied to the cells and left to incubate at 37°C in a 5% CO_2 environment for 5 minutes. Trypsinization allows for the breakdown of the cell surface proteins that adhere the cells to the culture dish. Three mL of culture medium were added to the trypsin-cell mixture in the plate, and pipetted up and down gently to dislodge any cells remaining adhered to the dish. The entire ~5mL mixture was then transferred to a 15mL conical tube and centrifuged at 1500 rpm for 4 minutes at 4°C. The liquid was then aspirated, and the cell pellet was resuspended in 3mL of culture medium. 0.5mL of the resuspended cells were then added to each well of a 6-well culture plate, already containing 2.5mL of culture medium.

Additionally, a human neuroblastoma cell line (SH-SY5Y) was acquired from the American Type Culture Collection (ATCC) via Cederlane Corporation. The cell culture and maintenance process for this cell line was identical to that of the HEK293 cells, except that the DMEM was supplemented with 20% FBS, as well as 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin.

3.2.5 Transfection of constructs into mammalian cells

To execute the *CD108131*-KO experimental design, cells would need to express a vector containing the upstream sgRNA machinery, a vector containing the downstream sgRNA machinery, and the vector containing the gBlock repair template. To test all combinations of sgRNA, four experimental conditions were needed, as well as a control condition (Table 3.2). Once a 6-well cell culture plate of each cell type (HEK293 and SH-SY5Y) had reached 85% confluency, media was aspirated and each well of cells was washed with 1.5mL of 37°C PBS. One mL of trypsin was added to each well, and cells were incubated at 37°C and 5% CO_2 until 90% of the cells had detached. While cells were being treated with trypsin, the DNA to be inserted was aliquoted into 5 separate 1.5mL tubes corresponding with each of the 5 conditions. Each well represented one of the experimental conditions, and cells received 0.5µg of an upstream sgRNA plasmid, 0.5µg of a downstream sgRNA plasmid, as well as 1µg of the gBlock plasmid. After the cells had detached from the surface of the plate, the reaction was neutralised using the culture medium appropriate for the cell type (as described in section 3.2.4). Cells were centrifuged at 90 x g for 10 minutes at room temperature, and the liquid was aspirated. The cell pellets were resuspended in 100µl of 4D-Nucleofector solution (Lonza Bioscience), and the prepared DNA was added. The cell and DNA mixture was transferred to a Nucleocuvette Vessel (Lonza Bioscience), and tapped gently to ensure contents had mixed. After placing the cuvettes

Experimental condition	Upstream sgRNA vector	Downstream sgRNA vector	Contains gBlock vector?
1	sgRNA1	sgRNA3	Yes
2	sgRNA1	sgRNA4	Yes
3	sgRNA2	sgRNA3	Yes
4	sgRNA2	sgRNA4	Yes
Control	Empty PX458	Empty PX458	Yes

Table 3.2. SgRNA combinations that comprise the 4 experimental conditions used to generate CRISPR/Cas9 *CD108131* knockouts. Each experimental condition consisted of one vector containing an sgRNA to cut upstream of *CD108131*, one vector containing an sgRNA to cut downstream of *CD108131*, and the gBlock vector. The control had the empty PX458 vector in pace of the sgRNA-containing vectors.

inside the 4D-Nucleofector X Unit (Lonza Bioscience), programme CA-137 was run for SH-SY5Y samples, and programme CA-130 was run for HEK293 samples. When the procedure was complete, the cuvettes were removed from the machine and incubated for 10 minutes at room temperature. Cells were mixed with 500µl of the appropriate culture medium, and pipette up and down gently before being plated into a new 6-well cell culture plate.

3.2.6 Fluorescence activated cell sorting

Two days after the cells underwent electroporation, they were analysed via fluorescence activated cell sorting (FACS). To prepare the cells for analysis, media was aspirated and the cells were treated with trypsin for 5 minutes at 37°C and 5% CO_2. After neutralising the reaction with the appropriate culture medium, cells were centrifuged at 1500rpm for 4 minutes at 4°C. Cell pellets were then resuspended in 300µl of PBS with 0.5% BSA. It is important for cells to be in single-cell suspension for FACS, so the cells were resuspended using a P1000 pipette and then filtered through a 40µm cell strainer. Resuspended cells were brought to the Flow Cytometry & Virometry Core Facility at the University of Ottawa to be sorted using the MoFlo Astrios Sorter. GFP-positive cells were sorted on a single cell basis into individual wells of 96-well cell culture plates. All wells for all plates, regardless of cell type, were filled with DMEM supplemented with 20% FBS, 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin. The additional FBS was provided to both cell types to assist with recovery. For experimental conditions 1-4, two full 96-well plates were sorted per condition. A single 96-well plate was sorted for the control condition. All plates were briefly centrifuged to facilitate cells binding to the surface of the plate.

3.2.7 Monoclonal cell culture and expansion

After cell sorting, it typically takes 7-10 days for the monoclonal cultures to recover and start proliferating. As individual cultures began to grow, whenever they reached full confluency, they would be treated with trypsin and passed to larger cell culture plates in the following order: 12-well plate, 6-well plate, 10cm dish. During the expansion process, all cells continued to receive DMEM supplemented with 20% FBS, 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin.

3.2.8 Genotyping and sequencing

Once a monoclonal cell culture had been expanded to a confluent 12-well plate, cells were treated with trypsin and the volume of liquid cell culture was divided equally into two conical tubes. Both tubes were pelleted via centrifugation at 1500 rpm for 4 minutes at 4°C. One cell pellet was resuspended and plated back into the original 12-well plate to continue expanding, and the other was used for DNA extraction using the Blood & Cell Culture DNA Mini Kit (Qiagen). Due to the lack of a known *CD10131*-negative control, PCR was performed on DNA from all clonal expansions to test for presence of wildtype *CD108131*, using the following primers:

Forward primer: 5’ – TTCTTTTCTTGAAAGTTGTTTATTCTG – 3’

Reverse primer: 5’ – GGTGAAACTCAGGCTGAATGA – 3’

PCR products were separated by agarose gel electrophoresis. A product containing the wildtype *CD108131* gene would have a band of 973bp, while a product with a *CD108131* deletion would have a band of 110bp. Any cultures that amplified wildtype *CD108131* were discarded. Since the 110bp band was very small compared to the 973bp, it was prone to running off of the gel or

having poor resolution. New primers were designed to amplify the crossover region between the two homology arms:

Forward primer: 5’ – TACACGTGTGTTAGGGCACC – 3’

Reverse primer: 5’ – AAAGTGTCGATGCTGGGGTT – 3’

PCR products were again separated by agarose gel electrophoresis, with sizes of 1438bp indicating wildtype, and 575bp indicating a deletion. Clones that genotyped as having a deletion were Sanger sequenced, and the sequencing data was reviewed using Geneious software.

3.3. Results and Discussion

3.3.1. CRISPR-Cas9 constructs contained correct sequences

For each of the 4 sgRNA vectors, a separate vial of DH5α cells with transfected and plated onto individual ampicillin agar plates. The cells transfected with the sgRNA1 vector grew 21 colonies, 5 of which were set up for liquid culture growth. After DNA extraction and digestion with restriction enzymes, 2 colonies displayed banding patterns consistent with sgRNA1 oligo insertion. Both samples were Sanger sequence verified. For sgRNA2, 12 colonies grew and again 5 were selected, all of which showed the appropriate banding pattern. Of these 5 samples, 4 were sequence verified. SgRNA3 vector transfection resulted in a growth of 16 colonies, of the 5 selected 4 showed the correct banding pattern. Interestingly, the sequencing results revealed that 3 of the samples had the same single nucleotide variation, where the A at position 13 had become a T. The remaining sample had the correct sequence. Finally, sgRNA4 transfection resulted in 25 colonies. Of the 5 colonies picked, 2 demonstrated the banding pattern for potential insertion, and both sequences were verified.

After transfecting the Stbl3 cells with the gBlock vector, 86 distinct colonies grew on the ampicillin agar plate. Of these colonies, 10 were selected and set up for liquid culture growth, and subsequent DNA extraction and restriction enzyme digestion. 6 colonies displayed a banding pattern indicative of gBlock insertion. After sequencing, 4 of the samples had the gBlock inserted in the reverse orientation, and 2 samples had correct insertion.

3.3.2. SH-SY5Y cells did not survive monoclonal cell sorting

The desired cell line for the *CD108131* knockout was the human neuroblastoma-derived cell line, SH-SY5Y. The reason for this being that MD is a neurological disorder, and preliminary demonstrated a potential link between *CD108131* and the known MD-related gene, *SGCE*. Both the SH-SY5Y and HEK293 cells were growing at a comparable rate prior to electroporation, and recovered from electroporation within a similar time frame. As such, by the time the cells underwent FACS, both cell lines were healthy and proliferating. However, after the single-cell sorting stage, the SH-SY5Y cells did not proliferate. This had been a concern, as SH-SY5Y cells are known to grow poorly in the absence of cell-to-cell interaction (Kovalevich and Langford, 2013). Unfortunately, the monoclonal cultures never grew, and eventually suffered contamination. The HEK293 cells grew slowly, and it took 3 weeks before any of the clones were noticeably proliferating. All clones remained healthy throughout the entire expansion process. Although not of neuronal origin, HEK293 cells have been shown to express neuronal-specific genes, and as such have been used previously for neuroscientific studies (Shaw et al, 2002). During embryogenesis, the structure that will become the adrenal gland is situated adjacent to that of the kidney, and gene expression signatures for HEK293 cells more closely align with those of the adrenal gland than the kidney. For this reason, the reigning hypothesis is

that HEK293 most likely originate from the adrenal precursor structure (Shaw et al, 2002, Lin et al, 2014).

3.3.3. Identification of heterozygous and homozygous CD108131 knockouts

Through genotype screening, two clones displayed the correct size of band to suggest a full *CD108131* deletion. Sanger sequencing was able to verify these clones as mutants: one (HOM) had the full *CD108131* deletion, the other (HOM*8) had the full *CD108131* deletion as well as an 8bp deletion 230bp downstream in a region of DNA with no assigned function (Fig. 3.2 and Fig. 3.3).

In an attempt to obtain a third biological replicate, the CRISPR/Cas9 experiment was replicated, again in HEK293 cells. For this second experiment, only conditions 1 and 3 were repeated, as they were the ones that had yielded knockout results previously. Genotype screening did not reveal any potential homozygous *CD108131*-KO clones, but did identify a clone with banding that suggested a potential heterozygous clone (1:1 intensity ratio of both the wildtype and knockout band sizes). Sequencing results confirmed this clone to be a heterozygous *CD108131*-KO (HET) (Fig. 3.4). As MD is an autosomal dominant disease, and therefore patients are heterozygous for any mutations, having a heterozygous clone could provide useful information and comparison.

The control condition clones were expanded and genotyped. All clones genotyped as wildtype, as anticipated.

3.3.4. Examining efficiency of experiment

For all 4 experimental conditions, 192 GFP-positive cells were sorted in a monoclonal fashion. The majority of these never proliferated to the point of expansion, and of those that did

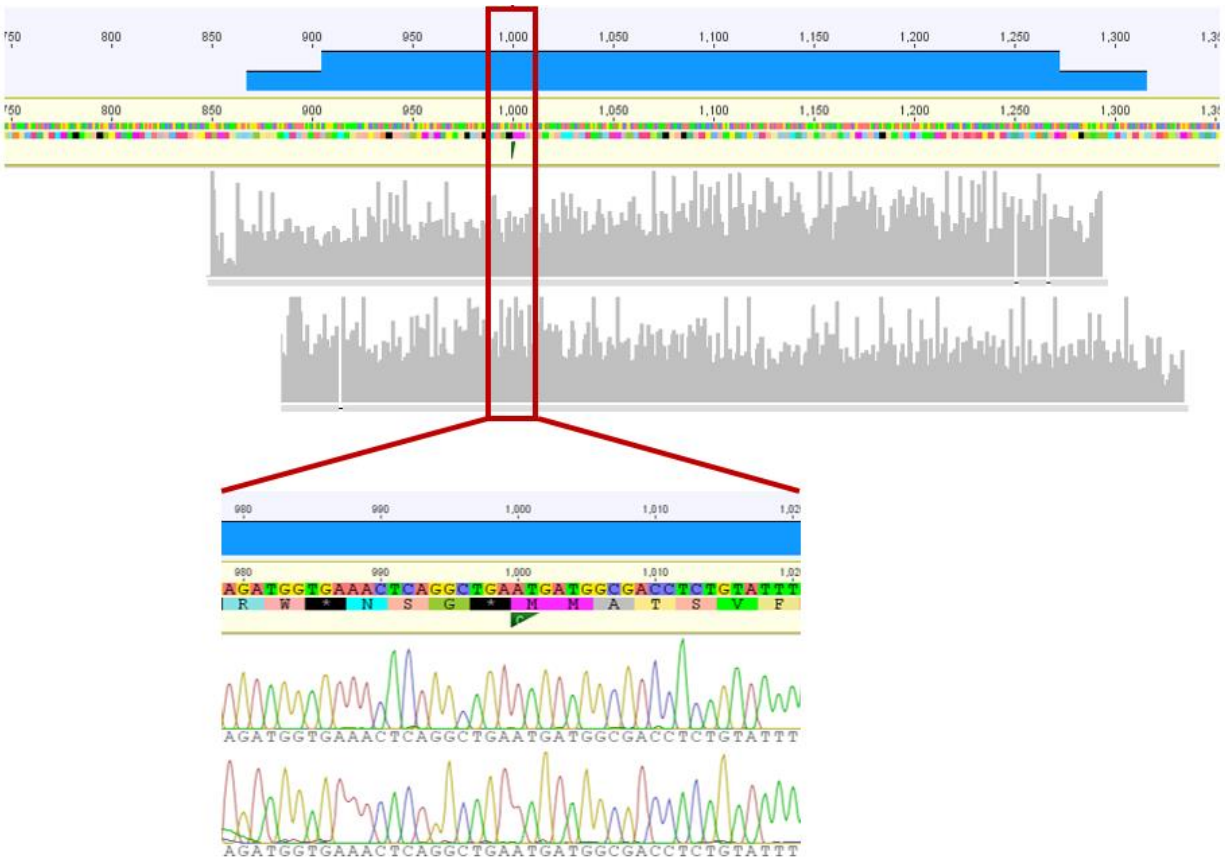

Figure 3.2. Sequencing results demonstrating the complete knockout of *CD108131* in mutant HOM. Sequencing data from clonal CRISPR-Cas9 *CD108131*-KO mutant HOM aligned to template homology sequence. Area outlined by red box, and further demarcated by the green triangle, is the region of cross over between the up- and downstream homology sequences. The sequence of this mutant remains an exact match at this region, so we can conclude that it is a true knockout. Additionally, the single peaks of all the nucleotide calls indicate that this is a homozygous deletion.

Figure 3.3. Sequencing results demonstrating the complete knockout of *CD108131* in mutant HOM*8. Sequencing data from clonal CRISPR-Cas9 *CD108131*-KO mutant HOM*8 aligned to template homology sequence. Area outlined by red box, and further demarcated by the green triangle, is the region of cross over between the up- and downstream homology sequences. The sequence of this mutant remains an exact match at this region, so we can conclude that it is a true knockout. The region outlined by the purple box is an 8bp deletion that was unexpected, but allows for two distinct mutant clones to be analysed. Additionally, the single peaks of all the nucleotide calls indicate that this is a homozygous deletion at both sites.

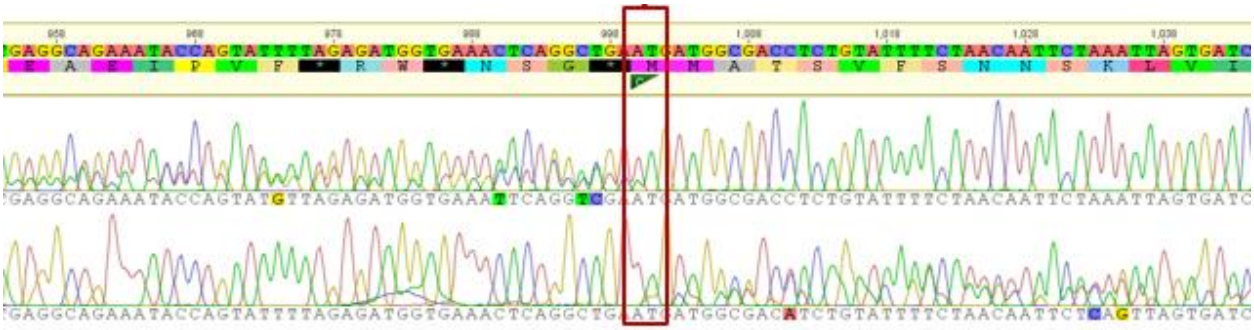

Figure 3.4. Sequencing results demonstrating the heterozygous knockout of *CD108131* in mutant HET. Sequencing data from clonal CRISPR-Cas9 *CD108131*-KO mutant HET aligned to the template homology sequence. Area outlined by red box, and further demarcated by the green triangle, is the region of cross over between the up- and downstream homology sequences. The double-peaks demonstrate the heterozygous state of this clone. It can be seen that for one strand the downstream homology sequence matches exactly, before switching to double-peaks at the deletion site; while in the other strand the upstream homology sequence matches exactly, before again switching to double-peaks at the deletion site.

only 2 generated the intended homozygous deletion (Table 3.3). As HDR efficiencies of less than 20% are considered low, the observed efficiency of generating a complete CD108131 knockout being 0-5.88% is a poor rate (Guo et al, 2018).

Several methods have been reported to assist in increasing the efficiency of CRISPR/Cas9 gene editing via HDR. Recently, it has been reported that so called "cold-shocking" cells through a 24-48 hour incubation at 32°C post-transfection can improve the rate of HDR in CRISPR/Cas9 edited cells by up to ten-fold (Guo et al, 2018). A three-fold increase in HDR efficiency was noted in electroporated cells that were co-transfected with RS-1, a small molecule shown to stabilise the association between RAD51 and DNA. RAD51 is a nucleofilament-forming protein that plays a role in the identification of homology and subsequent strand invasion (Pinder, Salsman and Dellaire, 2015). Additionally, new advances in electroporation technology can also prove beneficial for HDR efficiency. A new tube designed to replace traditional cuvettes has been developed, consisting of two small surface electrodes at the top and bottom. The cell mixture fills the entire tube, therefore eliminating the curved surface meniscus, and minimising air bubble generation. Using this method, an HDR rate of 42.1% was achieved in induced pluripotent stem cells (Xu et al, 2018).

3.4. Conclusions

Although efficiency of HDR was low, three *CD108131* knockout clones were generated by inducing two DSB at target sites situated up- and downstream of the target gene. These clones were created in HEK293 cells, which have been shown to have neuronal gene expression profiles. Two of the clones sequence verified as homozygous *CD108131* deletions, while the other was found to be a heterozygous deletion. All three clones will be maintained for future experiments.

Experimental condition	Number of cells sorted	Number of clones genotyped	Number of clones sequenced	% HDR efficiency
1	192	17	1	5.88
2	192	11	0	0
3	192	21	1	4.76
4	192	19	0	0

Table 3.3. Examination of the efficiency of HDR deletion of *CD108131* in CRISPR/Cas9 experimental conditions. FACS was conducted on each experimental condition, resulting in 192 single cell clones per condition. All monoclonal cultures that survived and proliferated were expanded and genotyped. Any clones that demonstrated a banding pattern indicative of *CD108131* deletion were Sanger sequenced. Condition # and # each resulted in one successful homozygous *CD108131* deletion. Data in the table is from the original experiment only, and does not include data from the experiment resulting in the identification of the clone with the heterozygous *CD108131* deletion.

Chapter 4: The effect on *SGCE, ZBTB14, ARHGAP28,* and *RPPH1* expression, due to the loss of *CD108131.*

4.1. Introduction

The gene responsible for the majority of MD cases is *SGCE* (Ritz et al, 2011). Our preliminary data, using both the *CD108131* variant as well as the *CD108131* knockdown, suggested a regulatory effect of *CD108131* on *SGCE* transcription and translation. To further examine this relationship, I quantified *SGCE* transcript levels in the context of a CRISPR/Cas9-derived cellular *CD108131* knockout.

The nearest coding gene to *CD108131*, separated by 725 bp, is the zinc-finger protein, *ZBTB14*. Little is known about the role of this gene in humans, however its amino acid sequence is 99% identical to that of the mouse homologue, *ZF5*. Within mice, *ZF5* is known to act as a transcriptional repressor of the regulatory oncogene, c-myc, and thymidine kinase (Sugiura et al, 1997). Preliminary data demonstrated that possession of the 3bp duplication within CD108131 had no effect on *ZBTB14* expression, although that variant was shown not to be disease-causing. Since *ZBTB14* is located within the MD critical region, and its proximity to *CD108131* would make it a likely regulatory target, gene expression was worth re-examining with the new *CD108131*-KO model.

Another gene within the critical region, 1.6Mb from *CD108131*, is the Rho GTPase-activating protein *ARHGAP28*. While not impossible, the distance between the two genes makes *ARHGAP28* a less likely regulatory target for *CD108131*. As one of the farthest genes from *CD108131* while still remaining within the critical region, it is expected to act as a control. For a traditional housekeeping gene, expression of *RPPH1*, the gene encoding for a component H1 of human RNase P ribonucleic protein, was analysed.

4.2. Materials and Methods

4.2.1. Cell culture

After the monoclonal expansion detailed in section 3.2.7 of Chapter 3, cells continued to be maintained using the recovery culture medium (DMEM supplemented with 20% FBS, 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin) and passed at 80-90% confluency.

For each of the four conditions (HOM, HOM*8, HET, and Control), a confluent 10cm plate was first passed 1:3 to three new 10cm plates. This created three biological replicates for each condition. Further passages were conducted until each biological replicate had expanded to 18 10cm plates. Of these plates, three were frozen down in a culture medium with 10% dimethyl sulfoxide (DMSO). The remaining 15 plates were used for RNA extraction.

The HEK293 cells that had been transfected with pcDNA3.1 containing the antisense transcript of *CD108131*, used as a knockdown model in one of the preliminary experiments, were retrieved from -80°C and thawed in a 37°C water bath. These cells were also expanded in the manner detailed above for use as a comparison condition. This condition was referred to as over-expression of reverse strand, or OER.

4.2.2. RNA extraction, clean-up, and cDNA synthesis

RNA was extracted from each confluent 10cm culture dish using TRIzol Reagent (ThermoFischer). Cells were first rinsed twice with approximately 2mL of cold PBS. Following this wash, 1mL of TRIzol was added directly to the culture dish, and cells were scraped using a rubber cell scraper. The lysed cells were pipetted up and down using a P1000 pipette until the mixture homogenised. As the TRIzol reagent is translucent pink, and the clumps of lysed cells appear opaque white, pipetting continued until the mixture was uniformly opaque pink. In lieu of

the traditional chloroform treatment used for phase separation, the Direct-zol Miniprep Plus kit (Zymo Research) was used. Not only does this kit circumvent the need for chloroform-based phase separation, but it is also designed to better capture small RNAs. The Direct-zol procedure was conducted using the protocol provided by the manufacturer. In short, an equal volume of ethanol was added to the TRIzol lysed cell mixture, before being transferred to a spin column and centrifuged at 14000 x g for 30 seconds. After a pre-wash and wash step, RNA was eluted from the column using 50μL of nuclease-free water. The optional in-column DNase I treatment was skipped in favour of performing a separate treatment post-RNA extraction, using the DNA-free DNA Removal Kit (ThermoFischer). DNase I and buffer were added to the eluted RNA, and incubated at 37°C for 30 minutes. The inactivation reagent was then added and mixed for 2 minutes, before being pelleted via centrifugation. The supernatant, containing the DNase I-treated RNA, was collected and transferred to a fresh RNase-free tube.

At this point, 15 tubes of RNA had been extracted for each biological replicate of each condition. The 15 tubes were pooled together, generating one tube of extracted and DNase I-treated RNA per biological replicated. After pooling, general concentration of RNA present was determined using a Nanodrop spectrophotometer.

The treated RNA then underwent first-strand cDNA synthesis using the iScript cDNA Synthesis Kit (Bio-Rad). Reverse transcriptase, reaction mix, nuclease-free water, and 1μg of RNA were added to a 0.5mL tube and placed in the thermocycler using the following conditions: 5 minutes at 25°C, 20 minutes at 46°C, 1 minute at 95°C, hold at 4°C. The resulting cDNA samples were stored at 4°C. Remaining RNA samples were stored at -80°C.

4.2.3. Primer design

For each gene of interest, primers were designed to span exon-exon junctions, when applicable. This ensured that the PCR primers amplify mature RNA transcripts, as both genomic DNA and preRNA retain intronic sequences. All primers were designed using NCBI's Primer-BLAST tool. Primer-BLAST retrieved primers using the standard parameters, with the following changes: amplicon size was set to a maximum of 200 nts, maximum primer length was increased to 28 nts, SNP handling was set to not design primers over known SNPs, and the repeat filter was set to "Human".

The primers for *ZBTB14* amplified a region spanning exon 4 and 5. The designed primer pair was as follows:

Forward primer: 5' – TGGTTTCAGACATACTGATGAAAAA – 3'

Reverse primer: 5' – AAGATTTTGGCGTTCAAGGCA – 3'

To amplify *ARHGAP28*, the following primer pair was designed spanning exons 8 and 9:

Forward primer: 5' – CCGCCATCTCTCTCTGATTGA – 3'

Reverse primer: 5' – AGTGGAACTCCAAAAATCCCA – 3'

The forward primer for the amplification of *SGCE* is within exon 3, while the reverse primer spans the junction of exons 4 and 5:

Forward primer: 5' – TGTGGGGAAGCCAACAATCA – 3'

Reverse primer: 5' – TGGCAACGGGAAGTCTTCTG – 3'

RPPH1 is a single exon gene, therefore primers were limited to within this exon, and could not be designed to span a junction:

Forward primer: 5’ – GCGGACGGAAGCTCATCAG – 3’

Reverse primer: 5’ – TCAGACCTTCCCAAGGGACAT – 3’

Additionally, primers were created to amplify β-actin (*ACTB*), as a housekeeping gene for which the genes of interest will be compared. Primers were designed to the junction of exons 3 and 4:

Forward primer: 5’ – CATGTACGTTGCTATCCAGGC – 3’

Reverse primer: 5’ – CTCCTTAATGTCACGCACGAT – 3’

4.2.4. Quantitative Real-Time PCR and

All Quantitative Real-Time PCR (qRT-PCR) experiments were performed using SsoFast EvaGreen Supermix (Bio-Rad). For each reaction, 10μL of SsoFast EvaGreen supermix was added, in addition to 1μL of the forward primer, 1μL of the reverse primer, 6μL of nuclease-free water, and 2μL of the cDNA sample. For every run, reactions were set up in 96-well plates where each biological replicate was tested in triplicate for both the gene of interest and *ACTB*. Plates were centrifuged briefly to ensure that all added components were at the base of the well, and then placed into the MasterCycler Ep RealPlex (Eppendorf) thermocycler. The qRT-PCR was run according to the following protocol: 95°C for 2 minutes; 40 cycles of 95°C for 5 seconds and 58°C for 20 seconds; and then a continuous increase from 65-95°C in 0.5°C increments. Experiments were replicated two to four times.

4.2.5. Data analysis

Gene expression data was consolidated across all replications of the experiment. Then, the expression for each gene of interest was compared to each sample’s β-actin expression, using the ΔΔCT method of analysis. Firstly, the CT value of β-actin is subtracted from the CT value of

the gene of interest, and ΔCT is calculated by raising 2 to the power of this value. The ΔCT expression for each technical replicate is then averaged for each biological replicate. The average ΔCT of each replicate is then divided by that of the biological replicate within the control group that most represents the mean, generating ΔΔCT expression values. The ΔΔCT expression of each biological replicate is averaged for each condition. Thus, this method allows for the detection of changes in expression of the gene of interest relative to β-actin and normalised to the control group. One-way analysis of variance (ANOVA) tests with Tukey's post-hoc analysis was conducted for each gene of interest, providing measures of variance between all conditions in the form of p-values. Standard deviation was also measured based on the ΔΔCT values. For data visualisation purposes, when the average ΔΔCT of the control group was not 1, all ΔΔCT values for all conditions were scaled by an integer that would bring the control group value to 1.

4.3. Results

4.3.1. Lack of *CD108131* significantly reduces *SGCE* expression

The qRT-PCR data for *SGCE* expression relative to that of *ACTB*, normalised to the control, indicate that the lack of *CD108131* significantly reduces *SGCE* transcription levels (Fig. 4.1). Across all knockout and knockdown conditions, *SGCE* expression was reduced by 47-76%. Interestingly, while there was no significant difference between the HOM*8, HET, and OER conditions, the HOM condition differed from the other three. HOM was also the condition that experienced the most variation between biological replicates, resulting in a standard deviation equating to just over 10%.

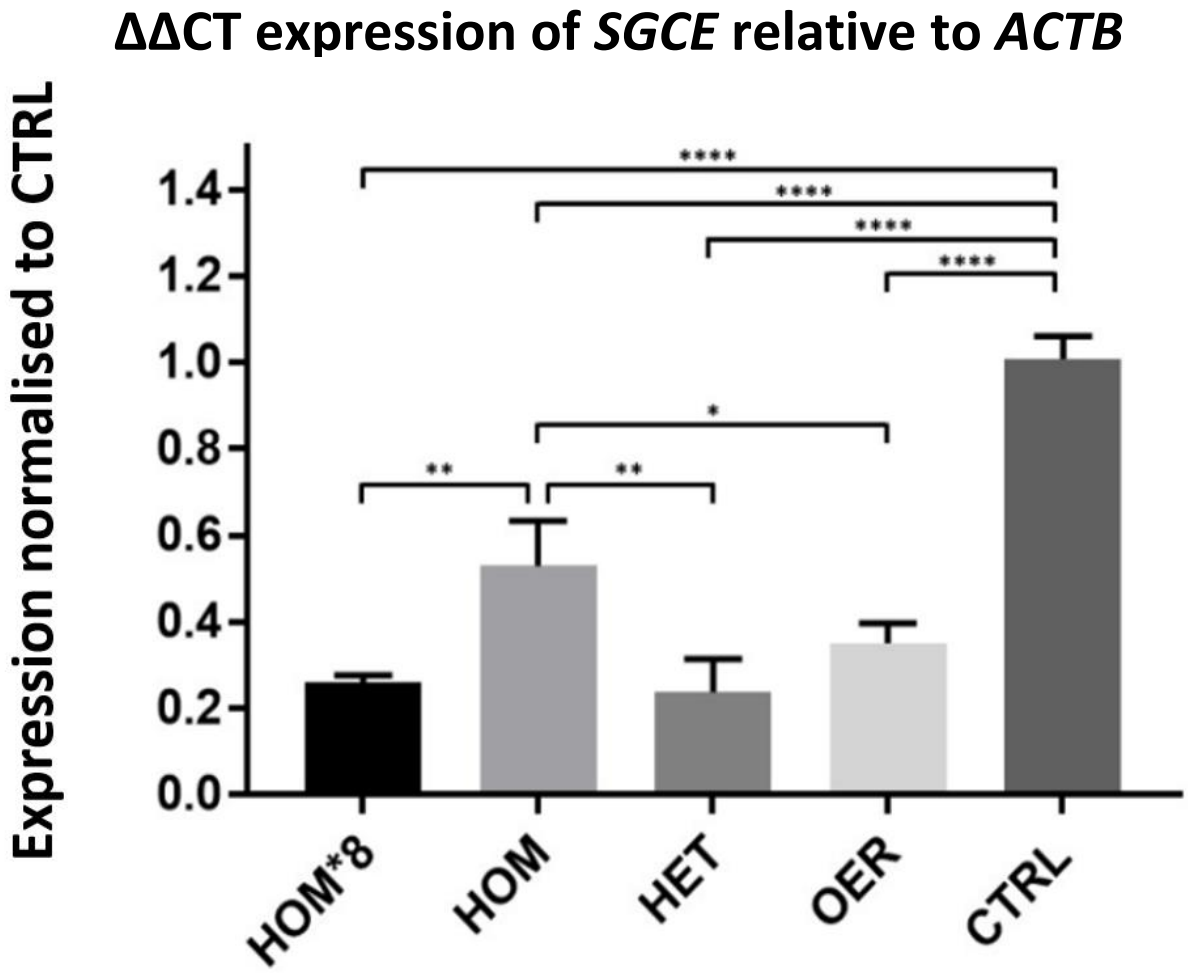

Figure 4.1. ΔΔCT analysis of *SGCE* in *CD108131* knockout and knockdown conditions compared to control. Results of qRT-PCR analysis quantifying levels of *SGCE* expression relative to β-actin, normalised against the control. HOM*8, HOM, and HET represent, respectively: the homozygous *CD108131*-KO clone with the additional 8bp deletion, the additional homozygous *CD108131*-KO clone, and the heterozygous *CD108131*-KO clone. OER (over-expression reverse) represents the knockdown condition cells, which are over-expressing the antisense transcript of *CD108131*. Control cells (CTRL) were transfected with empty CRISPR/Cas9 vectors. All samples are an average of 3 biological replicates. Error bars represent the standard deviation between biological replicates of each condition. Statistical significance was determined using a one-way ANOVA between all conditions, with Tukey's post-hoc analysis. Asterix denote the following p-values: 0.01-0.05 = *, 0.001-0.01 = **, 0.001-0.0001 = ***, <0.0001 = ****.

4.3.2. Lack of *CD108131* potentially reduces *ZBTB14* expression

Unlike the observed reduction to *SGCE* expression, *ZBTB14* expression was only significantly reduced in two of the four conditions: HOM*8 and HET (Fig. 4.2). However, the general trend for all conditions was towards lowered expression, with *ZBTB14* transcript levels being reduced by 32-71%. Due to the high standard deviation of the control group, the potentially expression changes noted in the HOM and OER conditions could jeopardise the detection of a significant difference. Throughout each replication of the experiment, the variation remained high within the control group.

4.3.3. Lack of *CD108131* unlikely to affect *ARHGAP28*

The expression data for *ARHGAP28* garnered unexpected results. While none of the knockdown or knockout conditions had significantly reduced expression compared to the control, HOM*8 visually trended towards such a reduction (Fig. 4.3). Additionally, HOM*8 was significantly reduced compared to both the HET and OER conditions. The HOM and OER conditions presented with an increase in ARHGAP28 expression. Standard deviation was very high for both the HOM and HET conditions, making statistical significance difficult to attain. The general trend seems to suggest that, for the majority of the conditions lacking *CD108131*, no major effects on ARHGAP28 expression occurred.

4.3.4. Lack of CD108131 does not significantly alter *RPPH1* expression

Since *RPPH1* was intended as a housekeeping gene, the anticipated qRT-PCR data would not be expected to show alterations to its expression amongst the experimental conditions. The data did not show any significant increase or reduction of *RPPH1* expression, nor any obvious

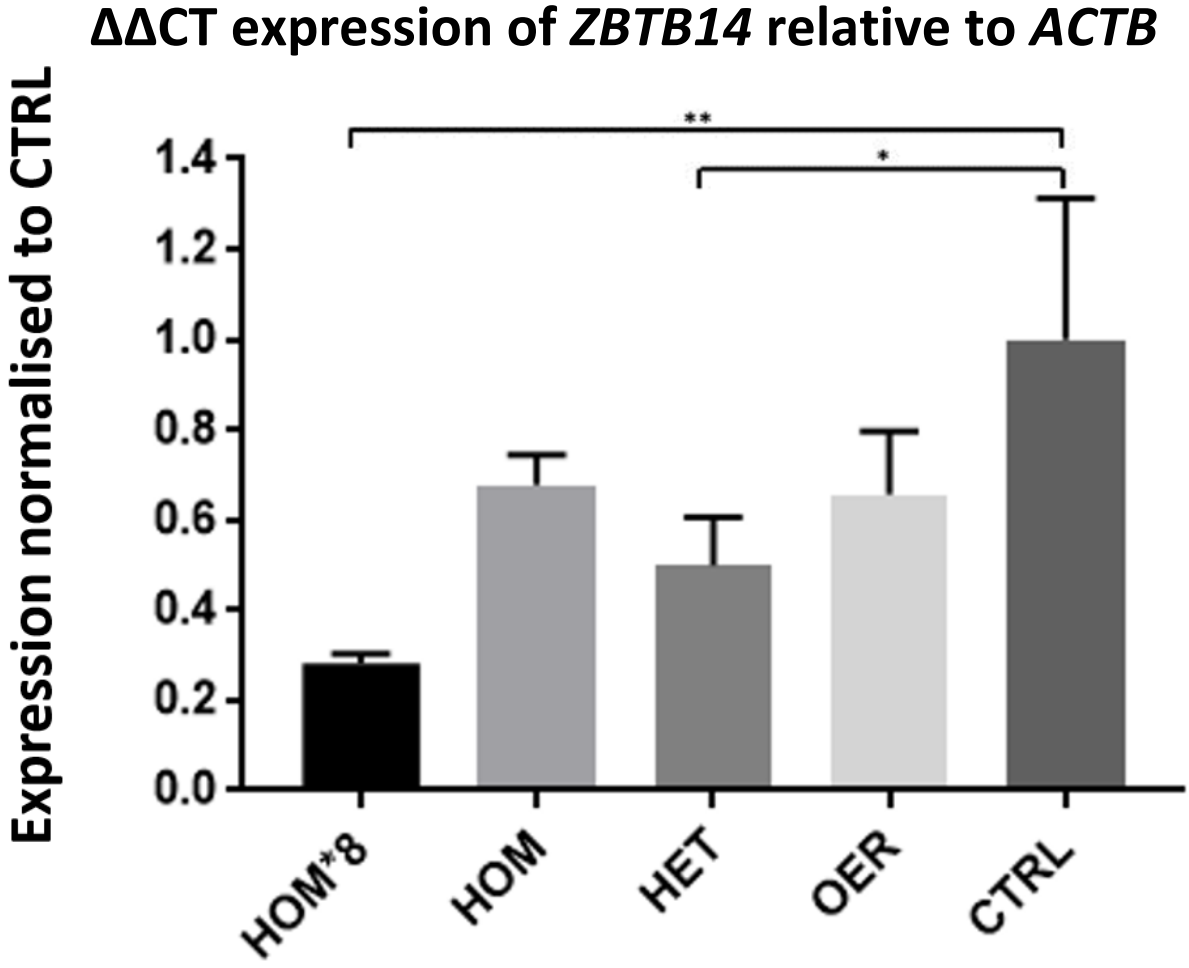

Figure 4.2. ΔΔCT analysis of *ZBTB14* in *CD108131* knockout and knockdown conditions compared to control. Results of qRT-PCR analysis quantifying levels of *ZBTB14* expression relative to β-actin, normalised against the control. HOM*8, HOM, HET, OER, and CTRL represent, respectively: the homozygous *CD108131*-KO clone with the additional 8bp deletion, the additional homozygous *CD108131*-KO clone, the heterozygous *CD108131*-KO clone, the cells over-expressing the antisense transcript of *CD108131*, and the control. All samples are an average of 3 biological replicates. Error bars represent the standard deviation between biological replicates of each condition. Statistical significance was determined using a one-way ANOVA between all conditions, with Tukey's post-hoc analysis. Asterix denote the following p-values: 0.01-0.05 = *, 0.001-0.01 = **, 0.001-0.0001 = ***, <0.0001 = ****.

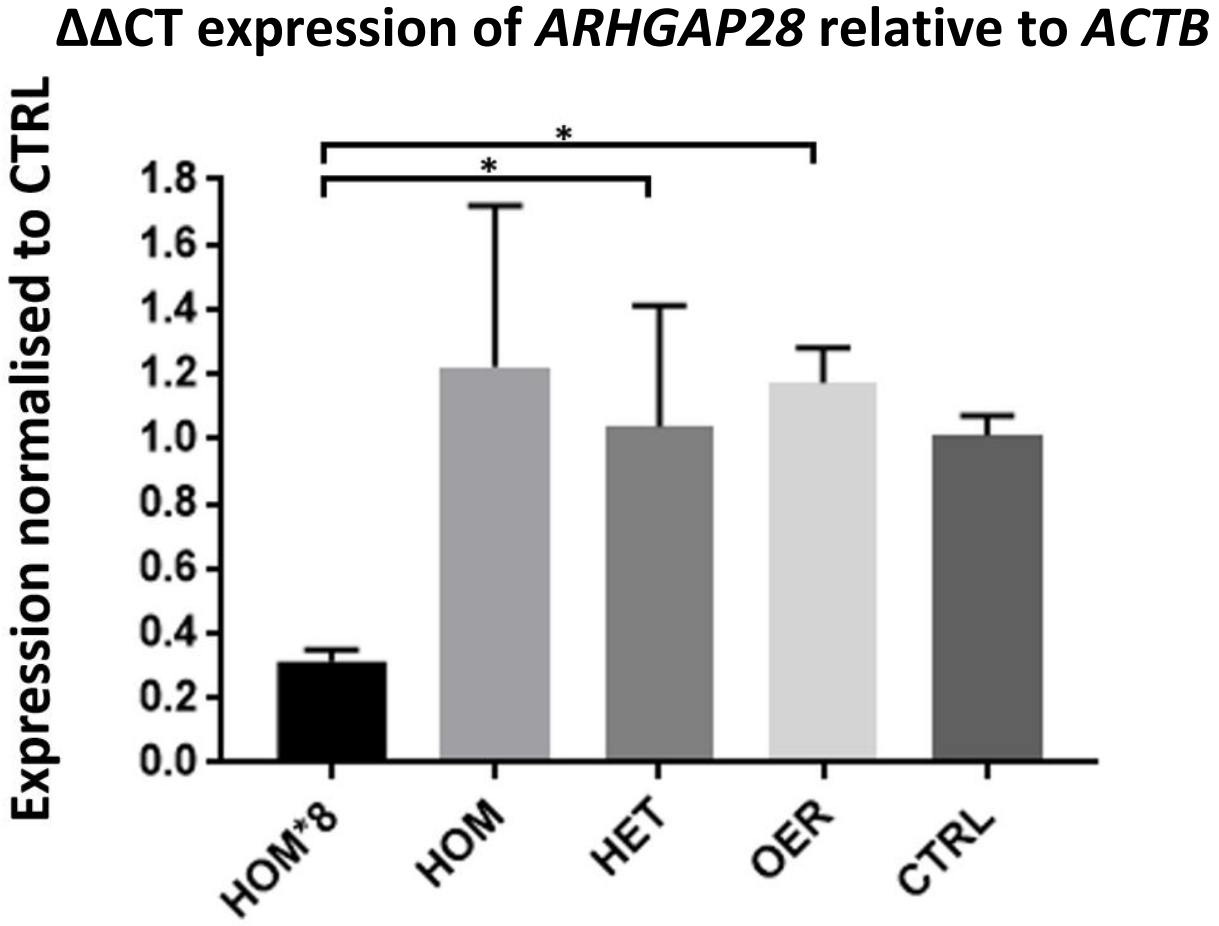

Figure 4.3. ΔΔCT analysis of *ARHGAP28* in *CD108131* knockout and knockdown conditions compared to control. Results of qRT-PCR analysis quantifying levels of *ARHGAP28* expression relative to β-actin, normalised against the control. HOM*8, HOM, HET, OER, and CTRL represent, respectively: the homozygous *CD108131*-KO clone with the additional 8bp deletion, the additional homozygous *CD108131*-KO clone, the heterozygous *CD108131*-KO clone, the cells over-expressing the antisense transcript of *CD108131*, and the control. All samples are an average of 3 biological replicates. Error bars represent the standard deviation between biological replicates of each condition. Statistical significance was determined using a one-way ANOVA between all conditions, with Tukey's post-hoc analysis. Asterix denote the following p-values: 0.01-0.05 = *, 0.001-0.01 = **, 0.001-0.0001 = ***, <0.0001 = ****.

trends (Fig. 4.4). However, the results did present with more variation than expected – both by way of expression level, and standard deviation. Although not the expected results for a housekeeping gene, a specific effect of *CD108131* on *RPPH1* expression is not observed.

4.4. Discussion

The reduced expression of *SGCE* observed in all the knockdown and knockout conditions was to be expected following the results seen in the preliminary data. However, a potential mechanism by which the *CD108131* transcript, located on chromosome 18, alters the expression of *SGCE* transcription on chromosome 7 is unknown. There is no part of the *CD108131* sequence that is complimentary to the sequence of *SGCE*, so any regulatory effect is likely to be indirect. It is possible that the DNase I hypersensitivity region present within *CD108131* could be the binding location of a transcription factor or other intermediate protein needed for proper transcription of *SGCE*. The same reduced expression had previously seen in the patient with the 3bp duplication, which happens to be located only 29bp upstream of the DNase I hypersensitivity region. Were this to be close enough that it would impact potential binding of certain proteins, it could explain why the phenotype was present even when the gene was not knocked out.

With respect to *ZBTB14*, although the results differ from those seen in the preliminary data, this is not entirely unexpected. The preliminary data for *ZBTB14* expression only looked at

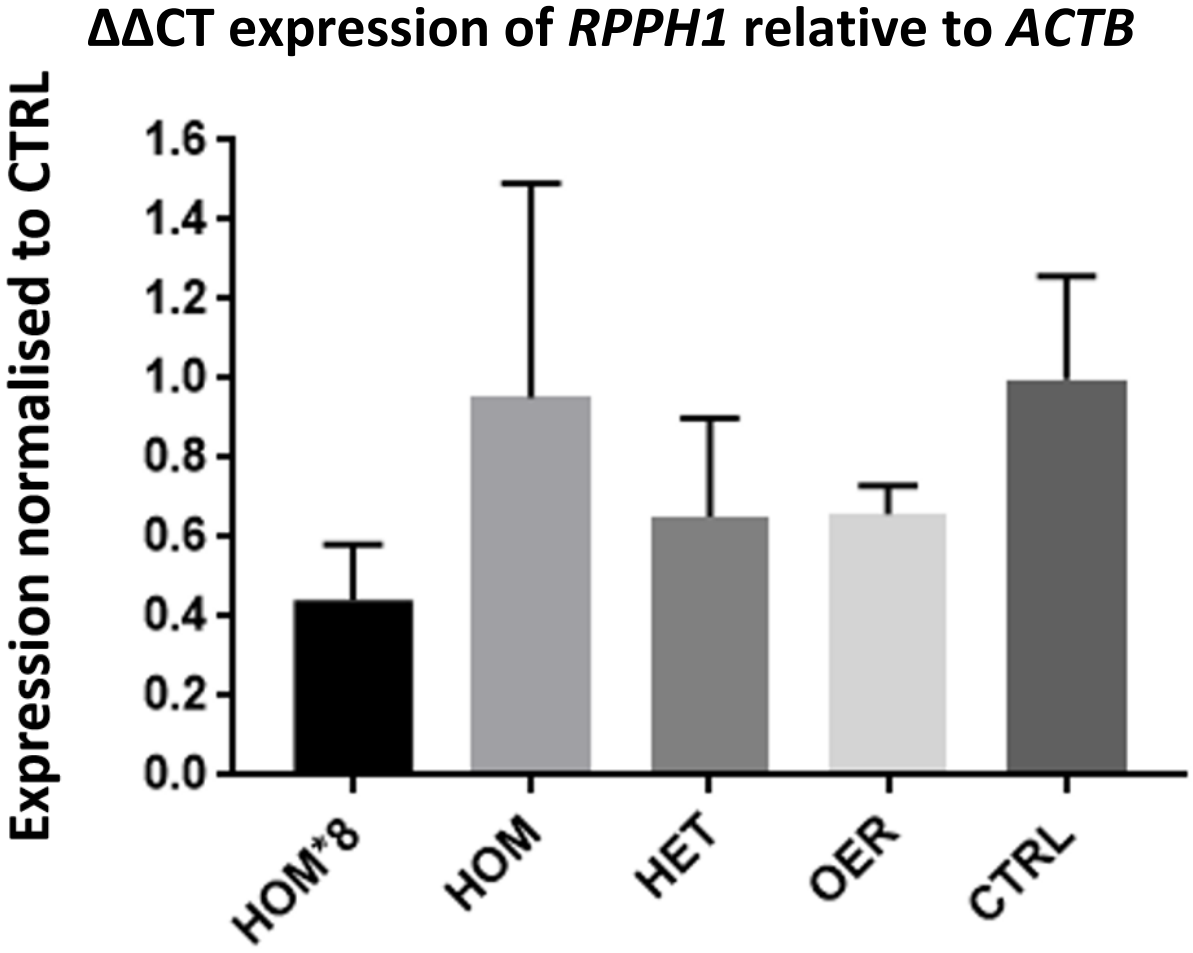

Figure 4.4. ΔΔCT analysis of *RPPH1* in *CD108131* knockout and knockdown conditions compared to control. Results of qRT-PCR analysis quantifying levels of *RPPH1* expression relative to β-actin, normalised against the control. HOM*8, HOM, HET, OER, and CTRL represent, respectively: the homozygous *CD108131*-KO clone with the additional 8bp deletion, the additional homozygous *CD108131*-KO clone, the heterozygous *CD108131*-KO clone, the cells over-expressing the antisense transcript of *CD108131*, and the control. All samples are an average of 3 biological replicates. Error bars represent the standard deviation between biological replicates of each condition. Statistical significance was determined using a one-way ANOVA between all conditions, with Tukey's post-hoc analysis. Asterix denote the following p-values: 0.01-0.05 = *, 0.001-0.01 = **, 0.001-0.0001 = ***, <0.0001 = ****.

the variation in expression between healthy control lymphocytes, and patient lymphocytes with the 3bp duplication. Firstly, expression profiles are known to differ between cell types. Secondly, as the variant was not disease-causing and occurs in healthy individuals, it is likely that little or no effect would be seen in comparison to a full gene deletion. While the mode of potential regulation is also unknown, the proximity of the genes might suggest an enhancer function of *CD108131* on *ZBTB14*. Although *CD108131* is downstream of the start site for *ZBTB14* transcription, enhancers can be located both up- or downstream of their targets (Pennacchio et al, 2013). Additionally, lncRNA can regulate transcription through epigenetic modification of nearby genes in either direction (Yuan et al, 2017).

With the exception of *SGCE*, all other genes analysed had a great degree of variation. This variation was reproducible experiment-to-experiment, and as such lies within the biological replicates themselves. Cell-to-cell variation is a known phenomenon, and one that might explain the variation seen in these experiments. Cell-to-cell variation occurs on two levels: environmental and cellular. In a naturally occurring group of cells, the environmental impacts would involve the access of the cell, or group of cells, to nutrient and chemical gradients. In the experimental process, many of these things are regulated. To this end, every effort was made to maintain all cells and process all samples under identical conditions, however it is possible that small differences might have resulted in detectable variation due to human error. However, it has been shown with bacteria, that even when it is assumed that the environmental conditions are homogenous, variation can occur within isogenic cells (Heins and Weuster-Botz, 2018). Cellular variability can result from a wide range of native and/or engineered pathways, such as: abundance of exogenous material, genetic variation, abundance of endogenous resources, and stochastic gene expression (Wang and Dunlop, 2018). While it is plausible that during the

transfection of the cells with the CRISPR/Cas9 machinery plasmid, some cells received more or less copies of these exogenous plasmids, since they were sorted monoclonal it is safe to assume that all the cells that propagated from one particular clone contained the same number of plasmids. For the same reason, it is also unlikely that genetic variation has occurred within the monoclonal culture. In terms of accessibility and abundance of endogenous resources, this can refer to such things as cellular levels of ATP or ribosomal availability (Wang and Dunlop, 2018). It is possible that cell-to-cell variation of these elements may have occurred. In terms of stochastic gene expression, it has been noted that even genetically identical cells can experience detectable fluctuations in individual gene expression, and that often these variations do not average away (Raj and van Oudenaarden, 2008). The interplay of cell-to-call variation pathways could help to explain the large standard deviations noted between some biological replicates.

In addition to variation among biological replicates, some gene expression data presented with unexpected variation among experimental conditions. It had been expected that the HOM*8 and HOM conditions would behave similarly, as they both possessed homozygous deletions of the *CD108131* gene. Interestingly, for every gene analysed, HOM*8 presented with lower gene expression than HOM. In the case of *SGCE* expression, this difference was particularly acute. As previously mentioned, the number of plasmids transfected into the cell can have an effect on gene transcription. It is possible that variation in this number, between the two transfected cells from which HOM*8 and HOM were derived, could result in the altered expression profiles observed. Additionally, some of the observed results could be due to off-target effects from CRISPR/Cas9 gene editing. It is known that at least one off-target effect, or a second independent mutation after knocking-out *CD108131*, did occur, as the sequencing results used to verify *CD108131* deletion revealed the extra 8bp deletion in HOM*8.

A common strategy for reducing off-target mutations is to utilise the nickase version of CRISPR/Cas9 (Zhang et al, 2014). Since the Cas9 enzyme has two different domains that each cleave one strand of the DNA, introducing a mutation into either domain would result in a mutated variant which is only capable of cleaving one of the strands – these mutated versions are referred to as nickases, or nCas9. When the RuvC domain is mutated, nCas9 can only cut the complementary strand, while mutating the HNH domain results in a nCas9 that can only cut the opposing strand (Doudna and Charpentier, 2014). Using nCas9 with two different sgRNAs set to target nearby, but offset target sites creates a break in both strands.

CRISPR/nCas9 is often used as a single pair of sgRNAs directed to cleave a single target location. The target location is generally small, as the optimal distance between the two sgRNA is 40-70bp, which works well when generating a knockdown or knockout by targeting promoter regions or transcription start sites (TSS). As with the rationale of the current *CD108131*-KO model, since these regions in lncRNAs are mainly unknown, this method is not viable. Indeed, the lack of information on the active and regulatory regions for lncRNA remains a roadblock regardless of which variation of Cas9 is used. To generate the complete excision of *CD108131*, the double-DSB method I used could have been replaced with a double-nickase pair, whereby each of the two target sites would be cut by distinct pairs of sgRNA. It is possible that using this approach may have increased target specificity, and is an avenue that can be explored in the future. Nevertheless, a BLAT search conducted for this experiment's target cut sites revealed that the sequence is specific to only the desired chromosomal location on chr18p11. This, however, does not exclude the possibility that near-identical sequences were recognised and cut by Cas9.

Furthermore, unavoidable secondary effects could have occurred due to the excision of the entire *CD108131* gene. Genomic excision of a larger sequence carries with it the added potential of deleting or disrupting regulatory DNA elements, such as: transcription factor binding sites, splice sites, enhancers, silencers, and promoters. Disruption of these elements can result in changes to the transcription and processing of other genes, given rise to observed phenotypes that are not attributable to the gene in question (Goyal et al, 2016). Additionally, since the TSS of *CD108131* is unknown, deleting the gene without deleting the TSS could result in the formation of a new gene body (Goyal et al, 2016). In fact, in a study aimed at determining the "CRISPRability" of lncRNA, it was determined that the criteria for editing the sequence of a lncRNA without potentially affecting neighbouring genes were as follows: lncRNA could not be transcribed from bidirectional promoters (no opposite direction promoters within 2000bp up- or downstream of the start of the lncRNA), the start of the lncRNA must be at least 2000bp away from its neighbouring genes, the start of the lncRNA could not fall within the gene body of another transcript (coding or non-coding). This study determined that 59% of examined lncRNAs proved to be "non-CRISPRable", demonstrating the difficulty of disrupting lncRNAs without also disrupting the genes around them (Goyal et al, 2016).

4.5. Conclusion

Strong statistical evidence exists to suggest that a lack of *CD108131* results in reduced expression of *SGCE*. These results corroborate with those seen in the preliminary data, where the patient with the 3bp duplication within *CD108131* presented with lower levels of *SGCE* transcript, and cells with knocked down *CD1018131* had reduced SGCE translation. Contrary to the preliminary data on *ZBTB14* expression, there is some statistical evidence, as well as an observed trend, that *CD108131* knockdown and knockout cells have reduced *ZBTB14*

transcription. The data does not suggest that a lack of *CD108131* has any conclusive effect on *ARHGAP28* or *RPPH1*.

Chapter 5: Optimization of SGCE protein quantification analysis

5.1. Introduction

The results of the qRT-PCR revealed a significant decrease in the levels of *SGCE* transcript in knockdown and knockout samples normalised to the control. Since the preliminary data also demonstrated that the knockdown cells had a near-complete knockdown of SGCE protein, it was of interest to examine changes in SGCE protein expression within our CRISPR/Cas9 *CD108131*-KO cell lines.

The original Western blots conducted several years ago utilised a custom-made antibody generated in rabbits against a human SGCE peptide, however all bleeds of this antibody have ceased to work, generating no bands. To move forward with protein analysis, a new custom antibody was ordered against the same SGCE peptide. Commercially available anti-SGCE antibodies were also ordered from Abcam and Sigma.

Additionally, Dr. Derek Blake's lab at Cardiff University published a paper in which they used their own custom-made anti-SGCE antibody for immunoaffinity purification and downstream Western blot analysis. This antibody was generated in rabbits which had been injected with a peptide sequence that corresponded with the C-terminus of the brain specific isoform of SGCE conjugated to keyhole limpet hemocyanin carrier protein (Waite et al, 2016). An aliquot of this antibody was provided by Dr. Blake.

5.2. Materials and Methods

5.2.1. Cell culture

Frozen stocks of the HOM*8, HOM, HET, and control cell lines were thawed and grown following the methods detailed in section 3.2.4 of Chapter 3. Cells were grown in DMEM supplemented with 20% FBS, 1% L-glutamine, 1% NEAA, and 1% penicillin/streptomycin.

5.2.2. Protein extraction

Once 10cm culture plates reached 90% confluency, culture media was aspirated and cells were washed with ~3mL cold PBS. PBS was aspirated, and then plates were left sitting tilted for 1 minute before aspirating a second time to ensure that all PBS has been removed. 1mL of lysis buffer was then applied directly to the plate. Three different lysis buffers were tested: 10% sodium dodecyl sulphate (SDS), radioimmunoprecipitation assay (RIPA), and CHAPS. Plates were incubated for 1 hour at 4°C, before contents was scrapped with a rubber cell scraper and transferred to a 1.5mL tube. Tubes were centrifuged at maximum speed for 10 minutes, and the supernatant was transferred to a new tube. A 150µL aliquot of the supernatant was mixed with 50µL of 5x loading buffer containing β-mercaptoethanol, and heated for 10 minutes at 90°C before being stored at -80°C.

5.2.3. SDS-PAGE and Western immunoblotting

To separate the protein samples, SDS polyacrylamide gel electrophoresis (PAGE) was conducted using a 10% acrylamide/bis-acrylamide gel. Samples were again heated for 10 minutes at 90°C, and then loaded into the wells of the gel. Electrophoresis was run at 100V for 1 hour and 20 minutes. Proteins were then transferred to a 0.22µm nitrocellulose membrane using transfer buffer containing 10% MeOH. This transfer was run for 1 hour and 15 minutes at 100V,

at 4°C. To block the membranes, two timepoints and three mixtures were tested. Blocking occurred for either 1 hour at room temperature, or overnight at 4°C with either 5% skim milk in tris-buffered saline with TWEEN-20 (TBS-T), 5% bovine serum albumin (BSA) in TBS-T, or a 50:50 mixture of the two. After blocking, the membrane was rinsed twice with double-distilled water, and the primary antibody was applied and incubated on a rocker overnight at 4°C. Various concentrations of antibody were tested, and the antibody was prepared in either 5% skim milk in TBS-T or 5% BSA in TBS-T. The membrane was then rinsed twice with double-distilled water, before undergoing three 5-minute washes with TBS-T on a room temperature rocker. The secondary anti-rabbit antibody (Sigma) was prepared at a concentration of 1:10000 in either 5% skim milk in TBS-T or 5% BSA in TBS-T, and incubated with the membrane for 1 hour on a room temperature rocker. The membrane was rinsed and underwent an additional three TBS-T washes, before being patted dry and treated with chemiluminescent substrate for 1 minute. Upon transferring the membrane to a cassette, film sheets were exposed to the chemiluminescence and developed.

5.2.4. Immuno-precipitation

A basic immuno-precipitation (IP) was conducted using Dynabeads Protein G (Thermo Fisher) magnetic beads. After protein lysis, protein concentration was determined via Bradford assay and 500µg of each sample was mixed with PBS (supplemented with 0.04% TWEEN-20 and protease inhibitors) to a total volume of 500µL. Samples were precleared via the addition of 20µL of Dynabeads and 1 hour of end-over-end mixing at 4°C. Using a magnetic rack, transfer sample to a new tube at add 5µL of anti-SGCE primary antibody (courtesy of Dr. Blake) to each sample. Samples were incubated overnight at 4°C with end-over-end mixing. 20µL of Dynabeads were added to each sample and incubated while mixing for 4 hours at 4°C. A

magnetic rack was used to separate the beads, and the supernatant was discarded. The beads underwent three 5-minute washes in PBS-T, rotating at 4°C. Protein was then eluted using 30µL of Laemmli buffer, at heated for 10 minutes at 70°C. Beads were removed, and the supernatant was boiled for 5 minutes at 95°C before being loaded into an SDS-PAGE gel and analysed via Western blot as detailed above.

Additionally, a crosslink IP was also attempted using Protein G Mag Sepharose (GE Healthcare) magnetic beads, following the manufacturer's protocol exactly. Associated buffer solutions were also ordered from GE Healthcare. As with the previous IP, each sample contained 500µg of protein, and 5µL of anti-SGCE primary antibody (courtesy of Dr. Blake) was used. For crosslinking and target protein binding, two steps with variable incubation times, samples were incubated for 30 minutes and 1 hour, respectively. Upon protein elution, protein was stored at -80°C until it was thawed and mixed with loading buffer in preparation for SDS-PAGE and Western blot analysis as detailed above.

5.2.5. SiRNA knockdown of *SGCE*

To facilitate the optimisation of the IP/Western blot experiments, siRNA-mediated *SGCE*-knockdowns were attempted using constructs ordered from Dharmacon. Scrambled siRNA constructs were also ordered to act as a control. HEK293 cells were cultured in 6-well plates, and the siRNA was transfected via Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher) following the manufacturer's protocol. A variety of different conditions were executed: cells were transfected at 60% or 80% confluency, siRNA concentration was 30pmol or 100pmol, and incubation length was 72 or 96 hours. Protein was then extracted using the method outlined above, and these protein samples were included in the crosslink IP experiment.

5.3. Results and Discussion

Prior to the validation of the *CD108131*-KO cell lines, the newly generated custom anti-SGCE antibody, as well as the commercially available antibodies, were tested using wildtype HEK293 protein. With respect to the Abcam and Sigma antibodies, signal was very low regardless of using a higher concentration of primary antibody, stronger chemiluminescent substrates, or more sensitive film. Additionally, these antibodies produced multiple non-specific bands. With the custom antibody, signal was detected, however an extensive amount of background signal and noise was detected. This resulted in large blotches that obscured the bands. In an attempt to rectify this, different concentrations of primary and secondary antibodies were attempted, and the blocking agent (as well as antibody solution agent) and length of incubation were troubleshooted. Ultimately, the best results occurred when the membrane was blocked overnight in a 1:1 mixture of 5% BSA in TBS-T and 5% skim milk in TBS-T, primary antibody was mixed 1:1000 in 5% BSA in TBS-T, and secondary antibody was mixed 1:2000 in 5% BSA in TBS-T. This greatly diminished the amount of background noise appearing on the blot, however, similarly to the commercially available antibodies, multiple non-specific bands were apparent.

Upon receiving the anti-SGCE antibody provided by Dr. Blake's lab, initial optimization using wildtype HEK293 involved determining any potential effects that different lysis buffers may have on the banding pattern observed. All three tested lysis buffers (RIPA, 2% SDS, and CHAPS) generated the same banding pattern. Since the Western blot was still presenting with many non-specific bands, an IP was conducted on protein extracted using RIPA buffer from the *CD108131*-KO cell lines. The subsequent Western blot involved blocking the membrane for 1 hour in 5% skim milk in TBS-T, mixing the primary antibody 1:500 in 5% BSA in TBS-T, and

mixing the secondary antibody 1:10000 in 5% skim milk in TBS-T. The IP resulted in a sizeable decrease in the number of non-specific bands. However, since both SGCE and the heavy chain of IgG are 50 kDa in size, it was impossible to differentiate which protein was responsible for the observed band.

To allow for the visualisation of SGCE without IgG, a crosslink IP was conducted using protein G sepharose beads. Additionally, siRNA-mediated SGCE-knockdowns were attempts to provide a negative control. Unfortunately, this experiment did not provide analysable data, and has not yet been optimised to a degree that bands are visible and quantifiable. The next steps in terms of developing a negative control would involve validating the siRNA constructs via qRT-PCR to ensure successful knockdown of *SGCE*. To improve the IP, various stages of the protocol can be altered, such as: the amount of magnetic beads used, the amount of primary antibody used, and the incubation length for antibody binding, crosslinking, blocking and/or target protein binding. Additionally, buffer solutions can be tested to ensure they remain at the appropriate pH levels, or alternative buffer solutions can be attempted (several are listed by the manufacturer as viable alternatives).

5.4. Conclusion

At this point in time, we are not able to assess the effect of a lack of *CD108131* on SGCE translation. Although progress has been made in terms of optimising the anti-SGCE antibodies, further work needs to be done to be able to confidently quantify changes in protein expression. The antibodies that achieved the greatest clarity were those obtained from Dr. Blake's lab, however the continued presence of multiple non-specific bands make the data difficult to interpret. Due to the size of SGCE, traditional IP experiments are confounded by the IgG heavy chain. Additional optimisation needs to be conducted to troubleshoot the protein G sepharose

crosslink IP, to allow for the visualisation of SGCE without IgG. Furthermore, a negative control would be beneficial, and therefore further testing of the siRNA knockdown cells should be conducted to determine if they were successful.

Chapter 6: Conclusions and Future Directions

6.1. Conclusions

I have validated the remaining variants of interest from the original Roche 454 sequencing data derived from patients in the MD-affected family. All of these variants were determined to be false positives. Reanalysis of the data revealed a shared region of low coverage, however upon Sanger sequencing this region, no shared, heterozygous variants were observed. The sequencing data has been thoroughly examined by multiple individuals, and as such it can be concluded that the disease-causing mutation may not be within that data. Repetitive regions were poorly sequenced, and non-coding regions are under-represented. While the *CD108131* variant is no longer a candidate, lncRNAs have been shown to play important regulatory roles within the cell, and have been increasingly linked to various diseases and disorders. The LncRNADisease database was first launched in 2012, containing 480 experimentally supported lncRNA-disease associations. It now contains 10564 experimentally supported lncRNA-disease associations, and an additional 195395 computationally predicted associations, spanning 529 different diseases (Bao et al, 2018). To that end, I created a stable *CD108131* knockout cell line, in both heterozygous and homozygous states. Unfortunately, the heterozygous cell line did not survive long-term, however the two homozygous cell lines will stay with the lab. These can be used for a multitude of future experiments to further characterise *CD108131* and its interactions. Using these cell lines, I demonstrated *CD108131* has a significant impact on the transcription of the known disease-associated gene, *SGCE*. *CD108131* is a relatively uncharacterised gene, with no known function. This data demonstrates that *SGCE* is a likely regulatory target of *CD108131*, providing novel insight into the role of *CD108131*. A potential interaction was also noted between *CD108131* and *ZBTB14*, suggesting that *CD108131* may be involved in multiple

pathways, or regulate multiple genes. Given that neurological and myopathic disorders have been shown to involve dysregulated lncRNA, these findings could still relate to MD on a pathological level.

6.2. Future directions

The basis of this research was first formed from a clinical perspective, focusing on identifying the genetic cause of MD in a large multiplex family. As the research progressed, the focus became that of a basic science project, focusing to characterise a potential interaction between two genes: *CD108131* and *SGCE*. Due to this overlap, there are many possible directions that this project could be taken, which will be elaborated on further throughout this chapter.

Determining disease-causing variant in MD family

With respect to the family that originally presented to the lab with inherited MD, the disease-causing variant is still unknown. Since the discovery of the DYT15 in 2001, all logical regions of this critical region have been examined via Roche 454 and Sanger sequencing. The remaining unexamined regions tend to be complicated by extensive repetitive sequence elements. With the current technology available, the next step would be to sequence the critical region using barcode identifier tags. Sequential adaptor sequences could be introduced prior to sequencing, used to aligned in the reads in the correct order, and then be removed as part of the post-processing. Use of this methodology would allow repetitive sequence reads to be appropriately matched to their true genomic location.

Determining effects of CD108131 on global gene expression

In the absence of *CD108131*, *SGCE* expression levels are significantly decreased. The data also suggests a potential decrease in *ZBTB14* transcript expression in cells lacking *CD108131*. *CD108131* and *SGCE* are on different chromosomes, and are not known to interact with each other. If *CD108131* is impacting the expression of two distinct genes, then it may be involved in multiple regulatory pathways. In order to determine additional regulatory targets, the next step would be to conduct RNA-seq experiments. This would involve fragmenting the RNA collected from the *CD108131*-KO cell lines, generating cDNA, adding adaptor sequences, and amplifying via PCR. These sequence fragments can then undergo paired-read sequencing, and the resulting sequence reads can be mapped to the genome and analysed for changes in expression. For this experiment, it would be useful to consider implementing unique molecular identifiers (UMIs). These sample-specific tags are added at the same point as adaptors, prior to enrichment and amplification. During a multiplexed experiment, PCR amplification bias occurs can occur, whereby one sample is amplified more than the other. When these PCR products are sequenced and analysed, it can confound the degree to which gene expression is altered. Using the UMI sequences, to compare sample-to-sample read counts, can help approximate the true relative abundance of all cellular mRNAs (Kou et al, 2016). Overall, not only would RNA-seq data provide valuable information about what genes are up- or down-regulated in *CD108131* knockout cells, the data can also be analysed for any determinable patterns of common features among the genes whose expression may be altered, which might give insight to potential pathways that are involved.

Determining effects of CD108131 on SGCE translation

In order to fully describe the functional relationship between *CD108131* and *SGCE*, the effects of knocking out *CD108131* on SGCE translation will have to be characterised. The lack of specificity demonstrated by the available anti-SGCE antibodies is a major obstacle. Two avenues exist to attempt to circumvent this issue. Firstly, a negative control for *SGCE* would be an asset. Generating a stable *SGCE*-negative cell line using either shRNA methods or gene editing would allow for continued optimization of the Western blot experiments, as well as provide an additional control when analysing future expression data. However, even with a negative control, it may not be possible to improve the clarity of the anti-SGCE antibody to an appropriate level. In the event that the SGCE-antibody cannot be optimized, the *CD108131*-KO cell lines could undergo additional gene editing to introduce a His-tag sequence to the *SGCE* gene sequence. Strong and specific anti-His antibodies are well described and easy to come by, allowing for the relative quantification of SGCE protein levels without access to a reliable anti-SGCE antibody.

Characterising the relationship between CD108131 and SGCE

The current conclusion, based on sequence complementarity and gene location, is that *CD108131* is likely to interact with *SGCE* through an intermediary protein. In order to explore this, possible avenues for investigation include localisation of both *CD108131* and *SGCE* transcripts, and the individual interaction partners of each transcript. To assess localisation, a simultaneous RNA-DNA fluorescent in situ hybridization (FISH) assay could be performed for *CD108131* and *SGCE*. This involves culturing cells on cover slips, before performing a paraformaldehyde fixation and hybridizing with fluorescently labeled oligo probes designed to detect and bind *CD108131* or *SGCE*. Fluorescent microscopy analysis is conducted to examine

localisation data. Through this, it could be observed whether *CD108131* and *SGCE* co-localise. If co-localisation is observed, it may suggest that *CD108131* regulates *SGCE* by acting as a scaffold and inducing structural changes to chromosome looping. If co-localisation is not observed, it could suggest that the involvement of *CD108131* in regulating *SGCE* may be in the form of a pathway or cascade mechanism. To gain further insight, capture hybridization analysis of RNA (CHART) could be performed to identify proteins and DNA sequences that interact with *CD108131*. CHART involves cross-linking chromatin, and using capture oligonucleotides designed to hybridize and immobilise *CD108131*-chromatin complexes on to beads. These complexes can then be eluted, purified, and analysed. By sequencing the enriched DNA garnered from this experiment, it could be proven whether *CD108131* and *SGCE* interact directly, and additional regulatory targets or intermediary interactions of *CD108131* would be revealed. In the event that an anti-SGCE antibody has been optimised, or a His-tagged SGCE had been constructed, chromatin immunoprecipitation (ChIP) could be conducted. Mass spectrometry can be used to identify novel protein interactions. In a similar mechanism to CHART, ChIP involves crosslinking protein-DNA interactions. This could be implemented for SGCE, to acquire the DNA sequences that interact with the protein, which could then be sequenced to see if there is any crossover between sequences that interact with SGCE, and those that interact with *CD108131*.

References

Albanese, A., Bhatia, K., Bressman, S. B., DeLong, M. R., Fahn, S., Fung, V. S. C., … Teller, J. K. (2013). Concept and classification of dystonia. *Movement Disorders*, *28*(7), 863–873. https://doi.org/10.1002/mds.25475.Phenomenology

Asmus, F., & Gasser, T. (2010). Dystonia-plus syndromes. *European Journal of Neurology*, *17*, 37–45. https://doi.org/10.1111/j.1468-1331.2010.03049.x

Ballarino, M., Morlando, M., Fatica, A., & Bozzoni, I. (2016). Non-coding RNAs in muscle differentiation and musculoskeletal disease. *Journal of Clinical Investigation*, *126*(6), 2021–2030. https://doi.org/10.1172/JCI84419

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., & Dong, D. (2018). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky905

Batista, P. J., & Chang, H. Y. (2013). Long Noncoding RNAs: Cellular Address Codes in Development and Disease. *Cell*, *152*(6), 1298–1307. https://doi.org/10.1016/j.cell.2013.02.012

Beukers, R. J., Foncke, E. M. J., van der Meer, J. N., Nederveen, A. J., de Ruiter, M. B., Bour, L. J., … Tijssen, M. A. J. (2010). Disorganized Sensorimotor Integration in Mutation-Positive Myoclonus-Dystonia. *Archives of Neurology*, *67*(4), 469–474. https://doi.org/10.1001/archneurol.2010.54

Bibikova, M., Carroll, D., Segal, D. J., Trautman, J. K., Smith, J., Kim, Y.-G., & Chandrasegaran, S. (2001). Stimulation of Homologous Recombination through Targeted

Cleavage by Chimeric Nucleases. *Molecular and Cellular Biology*, *21*(1), 289–297. https://doi.org/10.1128/MCB.21.1.289-297.2001

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, England)*, *151*(Pt 8), 2551–2561. https://doi.org/10.1099/mic.0.28048-0

Boulay, A.-C., Saubaméa, B., Cisternino, S., Mignon, V., Mazeraud, A., Jourdren, L., … Cohen-Salmon, M. (2015). The Sarcoglycan complex is expressed in the cerebrovascular system and is specifically regulated by astroglial Cx30 channels. *Frontiers in Cellular Neuroscience*, *9*, 9. https://doi.org/10.3389/fncel.2015.00009

Carbon, M., Raymond, D., Ozelius, L., Saunders-Pullman, R., Frucht, S., Dhawan, V., … Eidelberg, D. (2013). Metabolic changes in DYT11 myoclonus-dystonia. *Neurology*, *80*(4), 385–391. https://doi.org/10.1212/WNL.0b013e31827f0798

Caviness, J. N. (2014). Treatment of Myoclonus. *Neurotherapeutics*, *11*(1), 188–200. https://doi.org/10.1007/s13311-013-0216-3

Charlesworth, G., Bhatia, K. P., & Wood, N. W. (2013). The genetics of dystonia: New twists in an old tale. *Brain*, *136*(7), 2017–2037. https://doi.org/10.1093/brain/awt138

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096–1258096. https://doi.org/10.1126/science.1258096

Douglas, A. G. L., Andreoletti, G., Talbot, K., Hammans, S. R., Singh, J., Whitney, A., … Foulds, N. C. (2017). ADCY5-related dyskinesia presenting as familial myoclonus-dystonia. *Neurogenetics*, *18*(2), 111–117. https://doi.org/10.1007/s10048-017-0510-z

Ettinger, A. J., Feng, G., & Sanes, J. R. (1997). epsilon-Sarcoglycan, a broadly expressed homologue of the gene mutated in limb-girdle muscular dystrophy 2D. *The Journal of Biological Chemistry*, *272*(51), 32534–32538. https://doi.org/10.1074/JBC.272.51.32534

Fernández-Pajarín, G., Sesar, A., Relova, J. L., Ares, B., Jiménez-Martín, I., Blanco-Arias, P., … Castro, A. (2016). Bilateral pallidal deep brain stimulation in myoclonus-dystonia: our experience in three cases and their follow-up. *Acta Neurochirurgica*, *158*(10), 2023–2028. https://doi.org/10.1007/s00701-016-2904-3

Goyal, A., Myacheva, K., Groß, M., Klingenberg, M., Duran Arqué, B., & Diederichs, S. (2016). Challenges of CRISPR/Cas9 applications for long non-coding RNA genes. *Nucleic Acids Research*, *45*(3), gkw883. https://doi.org/10.1093/nar/gkw883

Grabowski, M., Zimprich, A., Lorenz-Depiereux, B., Kalscheuer, V., Asmus, F., Gasser, T., … Strom, T. M. (2003). The epsilon-sarcoglycan gene (SGCE), mutated in myoclonus-dystonia syndrome, is maternally imprinted. *European Journal of Human Genetics*, *11*(2), 138–144. https://doi.org/10.1038/sj.ejhg.5200938

Grimes, D. A., Bulman, D., George-Hyslop, P. S., & Lang, A. E. (2001). Inherited myoclonus-dystonia: evidence supporting genetic heterogeneity. *Movement Disorders : Official Journal*

of the Movement Disorder Society, *16*(1), 106–110. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11215567

Grünewald, A., Djarmati, A., Lohmann-Hedrich, K., Farrell, K., Zeller, J. A., Allert, N., … Klein, C. (2008). Myoclonus-dystonia: significance of large SGCE deletions. *Human Mutation*, *29*(2), 331–332. https://doi.org/10.1002/humu.9521

Grütz, K., Seibler, P., Weissbach, A., Lohmann, K., Carlisle, F. A., Blake, D. J., … Grünewald, A. (2017). Faithful SGCE imprinting in iPSC-derived cortical neurons: an endogenous cellular model of myoclonus-dystonia. *Scientific Reports*, *7*(1), 41156. https://doi.org/10.1038/srep41156

Guo, Q., Mintier, G., Ma-Edmonds, M., Storton, D., Wang, X., Xiao, X., … Feder, J. N. (2018). 'Cold shock' increases the frequency of homology directed repair gene editing in induced pluripotent stem cells. *Scientific Reports*, *8*(1), 2080. https://doi.org/10.1038/s41598-018-20358-5

Han, F., Racacho, L., Lang, A. E., Bulman, D. E., & Grimes, D. A. (2007). Refinement of the DYT15 locus in myoclonus dystonia. *Movement Disorders*, *22*(6), 888–892. https://doi.org/10.1002/mds.21400

Heins, A.-L., & Weuster-Botz, D. (2018). Population heterogeneity in microbial bioprocesses: origin, analysis, mechanisms, and future perspectives. *Bioprocess and Biosystems Engineering*, *41*(7), 889–916. https://doi.org/10.1007/s00449-018-1922-3

Jeggo, P. A. (1998). Identification of Genes Involved in Repair of DNA Double-Strand Breaks in Mammalian Cells. *Radiation Research*, *150*(5), S80. https://doi.org/10.2307/3579810

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Kim, J.-Y., Lee, W.-W., Shin, C. W., Kim, H.-J., Park, S.-S., Chung, S. J., … Jeon, B. (2017). Psychiatric symptoms in myoclonus-dystonia syndrome are just concomitant features regardless of the SGCE gene mutation. *Parkinsonism and Related Disorders*, *42*, 73–77. Retrieved from https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/13538020/v42icomplete/73_psimsarotsgm.xmL

Klein, C. (2014). Genetics in dystonia. *Parkinsonism and Related Disorders*, *20*(SUPPL.1), S137–S142. https://doi.org/10.1016/S1353-8020(13)70033-6

Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., … Li, S. (2016). Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLOS ONE*, *11*(1), e0146638. https://doi.org/10.1371/journal.pone.0146638

Kovalevich, J., & Langford, D. (2013). Considerations for the use of SH-SY5Y neuroblastoma cells in neurobiology. *Methods in Molecular Biology (Clifton, N.J.)*, *1078*, 9–21. https://doi.org/10.1007/978-1-62703-640-5_2

Li, T., Huang, S., Zhao, X., Wright, D. A., Carpenter, S., Spalding, M. H., … Yang, B. (2011). Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Research*, *39*(14), 6315–6325. https://doi.org/10.1093/nar/gkr188

Lin, Y.-C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., … Callewaert, N. (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications*, *5*(1), 4767. https://doi.org/10.1038/ncomms5767

Lohmann, K., & Klein, C. (2013). Genetics of dystonia: What's known? What's new? What's next? *Movement Disorders*, *28*(7), 899–905. https://doi.org/10.1002/mds.25536

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Mencacci, N. E., Rubio-Agusti, I., Zdebik, A., Asmus, F., Ludtmann, M. H. R., Ryten, M., … Wood, N. W. (2015). A missense mutation in KCTD17 causes autosomal dominant myoclonus-dystonia. *American Journal of Human Genetics*, *96*(6), 938–947. https://doi.org/10.1016/j.ajhg.2015.04.008

Mohanraju, P., Makarova, K. S., Zetsche, B., Zhang, F., Koonin, E. V., & van der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science*, *353*(6299), aad5147. https://doi.org/10.1126/science.aad5147

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, *60*(2), 174–182. https://doi.org/10.1007/s00239-004-0046-3

Moorthie, S., Mattocks, C. J., & Wright, C. F. (2011). Review of massively parallel DNA

sequencing technologies. *The HUGO Journal*, *5*(1–4), 1–12. https://doi.org/10.1007/s11568-011-9156-3

Morlando, M., & Fatica, A. (2018). Alteration of Epigenetic Regulation by Long Noncoding RNAs in Cancer. *International Journal of Molecular Sciences*, *19*(2). https://doi.org/10.3390/ijms19020570

Nishiyama, A., Endo, T., Takeda, S., & Imamura, M. (2004). Identification and characterization of epsilon-sarcoglycans in the central nervous system. *Brain Research. Molecular Brain Research*, *125*(1–2), 1–12. https://doi.org/10.1016/j.molbrainres.2004.01.012

Novikova, I. V., Hennelly, S. P., Tung, C.-S., & Sanbonmatsu, K. Y. (2013). Rise of the RNA Machines: Exploring the Structure of Long Non-Coding RNAs. *Journal of Molecular Biology*, *425*(19), 3731–3746. https://doi.org/10.1016/j.jmb.2013.02.030

Ozawa, E., Mizuno, Y., Hagiwara, Y., Sasaoka, T., & Yoshida, M. (2005). Molecular and cell biology of the sarcoglycan complex. *Muscle & Nerve*, *32*(5), 563–576. https://doi.org/10.1002/mus.20349

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews. Genetics*, *14*(4), 288–295. https://doi.org/10.1038/nrg3458

Phukan, J., Albanese, A., Gasser, T., & Warner, T. (2011). Primary dystonia and dystonia-plus syndromes: Clinical characteristics, diagnosis, and pathogenesis. *The Lancet Neurology*, *10*(12), 1074–1085. https://doi.org/10.1016/S1474-4422(11)70232-0

Pinder, J., Salsman, J., & Dellaire, G. (2015). Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Research*, *43*(19), 9379–9392. https://doi.org/10.1093/nar/gkv993

Piras, G., El Kharroubi, A., Kozlov, S., Escalante-Alcalde, D., Hernandez, L., Copeland, N. G., … Stewart, C. L. (2000). Zac1 (Lot1), a potential tumor suppressor gene, and the gene for epsilon-sarcoglycan are maternally imprinted genes: identification by a subtractive screen of novel uniparental fibroblast lines. *Molecular and Cellular Biology*, *20*(9), 3308–3315. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10757814

Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell*, *136*(4), 629–641. https://doi.org/10.1016/j.cell.2009.02.006

Raj, A., & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, *135*(2), 216–226. https://doi.org/10.1016/j.cell.2008.09.050

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., … Zhang, F. (2013). Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell*, *154*(6), 1380–1389. https://doi.org/10.1016/j.cell.2013.08.021

Rath, D., AmLinger, L., Rath, A., & Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie*, *117*, 119–128. https://doi.org/10.1016/J.BIOCHI.2015.03.025

Richter, A., Hamann, M., Wissel, J., & Volk, H. A. (2015). Dystonia and Paroxysmal Dyskinesias: Under-Recognized Movement Disorders in Domestic Animals? A Comparison with Human Dystonia/Paroxysmal Dyskinesias. *Frontiers in Veterinary Science*, *2*, 65.

https://doi.org/10.3389/fvets.2015.00065

Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, *81*(1), 145–166. https://doi.org/10.1146/annurev-biochem-051410-092902

Ritz, K., van Schaik, B. D., Jakobs, M. E., van Kampen, A. H., Aronica, E., Tijssen, M. A., & Baas, F. (2011). SGCE isoform characterization and expression in human brain: implications for myoclonus-dystonia pathogenesis? *European Journal of Human Genetics : EJHG*, *19*(4), 438–444. https://doi.org/10.1038/ejhg.2010.206

Rothstein, R. J. (1983). One-step gene disruption in yeast. *Methods in Enzymology*, *101*, 202–211. https://doi.org/10.1016/0076-6879(83)01015-0

Rouet, P., Smih, F., & Jasin, M. (1994). Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(13), 6064–6068. https://doi.org/10.1073/PNAS.91.13.6064

SHAW, G., MORSE, S., ARARAT, M., & GRAHAM, F. L. (2002). Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *The FASEB Journal*, *16*(8), 869–871. https://doi.org/10.1096/fj.01-0995fje

Shiga, K., Yoshioka, H., Matsumiya, T., Kimura, I., Takeda, S., & Imamura, M. (2006). Zeta-sarcoglycan is a functional homologue of gamma-sarcoglycan in the formation of the sarcoglycan complex. *Experimental Cell Research*, *312*(11), 2083–2092. https://doi.org/10.1016/j.yexcr.2006.03.011

Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A., & Kucherlapati, R. S. (1985). Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. *Nature*, *317*(6034), 230–234. https://doi.org/10.1038/317230a0

Spatola, M., & Wider, C. (2012). Overview of primary monogenic dystonia. *Parkinsonism & Related Disorders*, *18 Suppl 1*, S158-61. https://doi.org/10.1016/S1353-8020(11)70049-9

Sugiura, K., Muro, Y., Nagai, Y., Kamimoto, T., Wakabayashi, T., Ohashi, M., & Hagiwara, M. (1997). Expression cloning and intracellular localization of a human ZF5 homologue. *Biochimica et Biophysica Acta*, *1352*(1), 23–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9177479

Tarakci, H., & Berger, J. (2016). The sarcoglycan complex in skeletal muscle. *Frontiers in Bioscience (Landmark Edition)*, *21*, 744–756. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26709803

Thomas, K. R., & Capecchi, M. R. (1986). Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene. *Nature*, *324*(6092), 34–38. https://doi.org/10.1038/324034a0

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., … Kent, W. J. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, *45*(D1), D626–D634. https://doi.org/10.1093/nar/gkw1134

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell*, *154*(1), 26–46. https://doi.org/10.1016/j.cell.2013.06.020

van der Meer, J. N., Beukers, R. J., van der Salm, S. M. A., Caan, M. W. A., Tijssen, M. A. J., & Nederveen, A. J. (2012). White matter abnormalities in gene-positive myoclonus-dystonia. *Movement Disorders*, *27*(13), 1666–1672. https://doi.org/10.1002/mds.25128

Vanstone, M. (2012). Indentification, validation and characterization of the mutation on chromosome 18p which is responsible for causing mayclonus-dystonia. University of Ottawa, Department of Biochemistry, Microbilogy and Immunoglogy (Master of Science Thesis).

Waite, A. J., Carlisle, F. A., Chan, Y. M., & Blake, D. J. (2016). Myoclonus dystonia and muscular dystrophy: ε-sarcoglycan is part of the dystrophin-associated protein complex in brain. *Movement Disorders*, *31*(11), 1694–1703. https://doi.org/10.1002/mds.26738

Wang, C., Wang, L., Ding, Y., Lu, X., Zhang, G., Yang, J., … Xu, L. (2017). LncRNA Structural Characteristics in Epigenetic Regulation. https://doi.org/10.3390/ijms18122659

Wang, T., & Dunlop, M. J. (2019). Controlling and exploiting cell-to-cell variation in metabolic engineering. *Current Opinion in Biotechnology*, *57*, 10–16. https://doi.org/10.1016/J.COPBIO.2018.08.013

Wang, X., Sun, Q., McGrath, S. D., Mardis, E. R., Soloway, P. D., & Clark, A. G. (2008). Transcriptome-Wide Identification of Novel Imprinted Genes in Neonatal Mouse Brain. *PLoS ONE*, *3*(12), e3839. https://doi.org/10.1371/journal.pone.0003839

Wapinski, O., & Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends in Cell Biology*, *21*(6), 354–361. https://doi.org/10.1016/j.tcb.2011.04.001

Weisheit, C. E., Pappas, S. S., & Dauer, W. T. (2018). *Inherited dystonias: clinical features and molecular pathways*. *Handbook of Clinical Neurology* (1st ed., Vol. 147). Elsevier B.V. https://doi.org/10.1016/B978-0-444-63233-3.00016-6

Xu, X., Gao, D., Wang, P., Chen, J., Ruan, J., Xu, J., & Xia, X. (2018). Efficient homology-directed gene editing by CRISPR/Cas9 in human stem and primary cells using tube electroporation. *Scientific Reports*, *8*(1), 11649. https://doi.org/10.1038/s41598-018-30227-w

Yokoi, F., Dang, M. T., Zhou, T., & Li, Y. (2012). Abnormal nuclear envelopes in the striatum and motor deficits in DYT11 myoclonus-dystonia mouse models. *Human Molecular Genetics*, *21*(4), 916–925. https://doi.org/10.1093/hmg/ddr528

Yuan, M., Wang, S., Yu, L., Qu, B., Xu, L., Liu, L., … Liu, H. (2017). Long noncoding RNA profiling revealed differentially expressed lncRNAs associated with disease activity in PBMCs from patients with rheumatoid arthritis. *PloS One*, *12*(11), e0186795. https://doi.org/10.1371/journal.pone.0186795

Zhang, F., Wen, Y., & Guo, X. (2014). CRISPR/Cas9 for genome editing: progress, implications and challenges. *Human Molecular Genetics*, *23*(R1), R40–R46. https://doi.org/10.1093/hmg/ddu125

Zhang, L., Yokoi, F., Parsons, D. S., Standaert, D. G., & Li, Y. (2012). Alteration of Striatal Dopaminergic Neurotransmission in a Mouse Model of DYT11 Myoclonus-Dystonia. *PLoS ONE*, *7*(3), e33669. https://doi.org/10.1371/journal.pone.0033669

Zimprich, A., Grabowski, M., Asmus, F., Naumann, M., Berg, D., Bertram, M., … Gasser, T.

(2001). Mutations in the gene encoding ε-sarcoglycan cause myoclonus–dystonia syndrome. *Nature Genetics, 29*(1), 66–69. https://doi.org/10.1038/ng709