Spring 5-2020

Identifying the Link Between Non-Coding Regulatory RNAs and Phenotypic Severity in a Zebrafish Model of gmppb Dystroglycanopathy

Grace Smith

IDENTIFYING THE LINK BETWEEN NON-CODING REGULATORY RNAS AND PHENOTYPIC SEVERITY IN A ZEBRAFISH MODEL OF *GMPPB* DYSTROGLYCANOPATHY

by

Grace Smith

A Thesis Submitted in Partial Fulfillment
of the Requirements for a Degree with Honors
(Molecular & Cellular Biology)

The Honors College

University of Maine

May 2020

Advisory Committee:
Benjamin King, Assistant Professor of Bioinformatics, Advisor
Clarissa Henry, Associate Professor of Biological Sciences
Melissa Ladenheim, Associate Dean of the Honors College
Sally Molloy, Assistant Professor of Genomics and NSFA-Honors Preceptor of Genomics
Kristy Townsend, Associate Professor of Neurobiology

ABSTRACT

Muscular Dystrophy (MD) is characterized by varying severity and time-of-onset by individuals afflicted with the same forms of MD, a phenomenon that is not well understood. MD affects 250,000 individuals in the United States and is characterized by mutations in the dystroglycan complex. *gmppb* encodes an enzyme that glycosylates dystroglycan, making it functionally active; thus, mutations in *gmppb* cause dystroglycanopathic MD[1]. The zebrafish (*Danio rerio*) is a powerful vertebrate model for musculoskeletal development and disease. Like human patients, *gmppb* mutant zebrafish present both mild and severe phenotypes. In order to understand the molecular mechanisms involved, we performed high-throughput RNA Sequencing (RNA-Seq) and small RNA Sequencing at 4 and 7 days-post-fertilization (dpf) in mild and severe *gmppb* mutants and controls. We hypothesize that variable phenotypes in *gmppb* mutants are due to differences in gene regulation; therefore, we identified differentially expressed (DE) long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) - both potent genetic regulators. We identified "MD-relevant" DE Ensembl-annotated genes involved in cell cycle regulation, the immune response, neural development and maturation, and skeletal muscle atrophy. We identified DE miRNAs that regulate these DE genes in the 4dpf severe mutants – identifying 55 of these interactions. We utilized a novel method of visualizing gene expression networks by generating co-expression networks of miRNAs and subsequently removing miRNA nodes to identify important miRNAs. We identified 95 potential lncRNAs for further analysis. By integrating analyses of both coding *and* non-coding genes, we contributed towards the understanding of the molecular mechanisms of Dystroglycanopathy, highlighting potential phenotypic modulators.

ACKNOWLEDGEMENTS

None of this work would have been possible without the support of countless professors, teachers, and other mentors who believed in me, inspired me, and pushed me to reach out of my comfort zone. My love of Molecular Biology first sprouted my sophomore year in Honors Biology with Mr. Rand. As both my science professor and Cross Country/Track Coach, he challenged me to excel in the classroom and beyond. He was the reason I applied to my first research internship at the University of Maine. That is where I was first mentored by Dr. Eric Gallandt and Sonja Birthisel who nurtured me and encouraged me to strive to become a more independent thinker. Not to mention, Sonja was the first to introduce me to R!

Many thanks to Dr. Molloy who was a pivotal component of my freshman year (and beyond). She encouraged me to strive for personal growth, become a better team member, let go of my competitiveness (through bribing me with hot cocoa), and fearlessly make mistakes - as long as I learned from them.

To Dr. King who took a chance on me and allowed me to join his lab after getting rejected from *every single* summer internship I applied for freshman year. I've been so blessed to have your support all throughout my undergraduate - you've challenged me to strive for every opportunity, whether I think I have a chance or not. Being a part of your lab has been my favorite experience during this time and has catalyzed so much personal and scientific growth - not to mention it made me fall in love with computer science! I can only aspire to mentor in the way you do one day.

Thanks to Dr. Jason Thomas, Jennifer Lipps, and Zeus whose support has remained unwavering and have challenged me to defend the defenseless and stand up for what I believe in.

Thanks to Dr. Aadel Chaudhuri, Pradeep Chauhan, and Peter Harris. Washington University had such a profound effect on my academic path, and I cannot thank you enough for allowing me to explore all the opportunities there - and mentoring and supporting me throughout.

Thanks to Nives Dal Bo-Wheeler for her unwavering support of the students at the University of Maine - we are so incredibly lucky to have you!

Many thanks to the Honors College. The Honors College and the liberal arts education I received through it, completely changed my view of the world around me - and I cannot thank you all enough for that. Dr. Ladenheim and Dr. Amar have helped me pursue my dreams and share my honors experience with others. Furthermore, Dr. Ladenheim has been a constant source of support and encouragement - as well as a phenomenal reading list advisor.

Finally, to my parents who have supported me in everything I do and loved me unconditionally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Muscular Dystrophy

Muscular Dystrophy (MD) is a group of debilitating musculoskeletal disorders that affects 250,000 individuals in the United States[1]. At the moderate end, MD symptoms can develop during late adulthood and be characterized by weak, but still functional muscles (Appendix A, Table A1). At the more severe end of the spectrum, symptoms can develop prenatally and consist of severe brain and eye abnormalities which often lead to miscarriage of the developing fetus[2]. To some extent, symptoms are related to the form of MD an individual has. There are nine main forms of MD (Appendix A, Table A1); each caused by mutations in dozens of different genes encoding the protein subunits of the dystroglycan complex, or in some cases, mutations in enzymes involved in post-transcriptional modification of these subunits.

Mutations in any of the protein components of the dystroglycan complex can cause MD (Figure 1). This complex is essential in skeletal muscle as it provides structural stability to the sarcolemma by linking the cells to one another, resulting in long, robust fibrils. The dystroglycan complex consists of the transmembrane β-dystroglycan complex which is bound to α-dystroglycan. On the intracellular end of β-dystroglycan, dystrophin subunits polymerize and are eventually linked to the cytoskeleton. The sarcoglycans provide increased stability to the complex and help mediate the strong connection between α and β dystroglycan[3]. α-dystroglycan is a ligand for laminin 2 which links the complex to the extracellular matrix. This interaction is crucial during muscle contraction, where impediment of this complex leaves muscle fibers more susceptible to damage. Two of the most common forms of MD, Becker MD (BMD) and Duchenne's MD (DMD) are caused by mutations in dystrophin, the intracellular tether that links the dystroglycan complex to the cytoskeleton. Generally, the more severe phenotype of DMD is attributed to genetic mutations that results in premature termination of dystrophin whereas BMD is the result of missense or frameshift mutations which do not majorly impact the length of the protein[4–6]. As stated previously, mutations in any of these protein subunits can lead to MD; however, mutations in genes associated with post-transcriptional modification of these subunits can also cause MD.

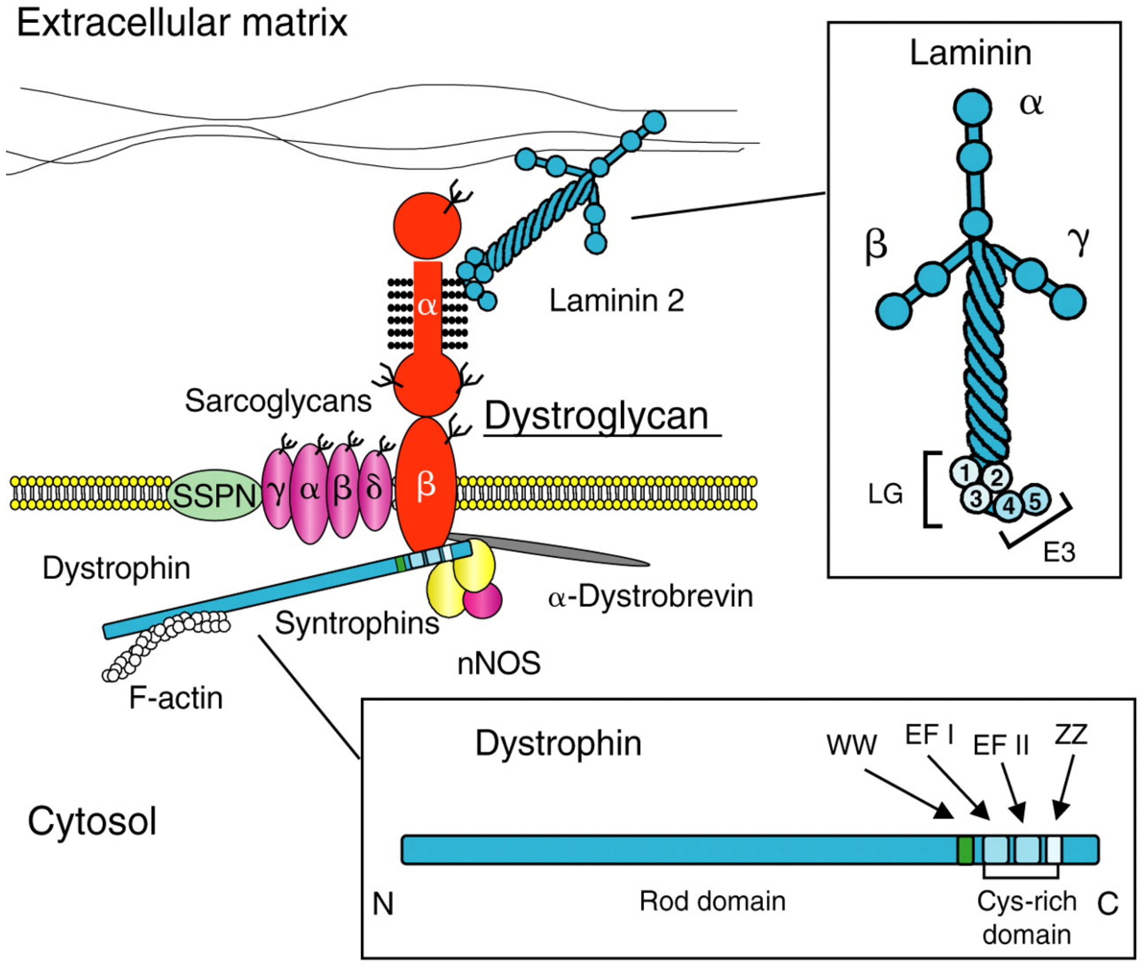
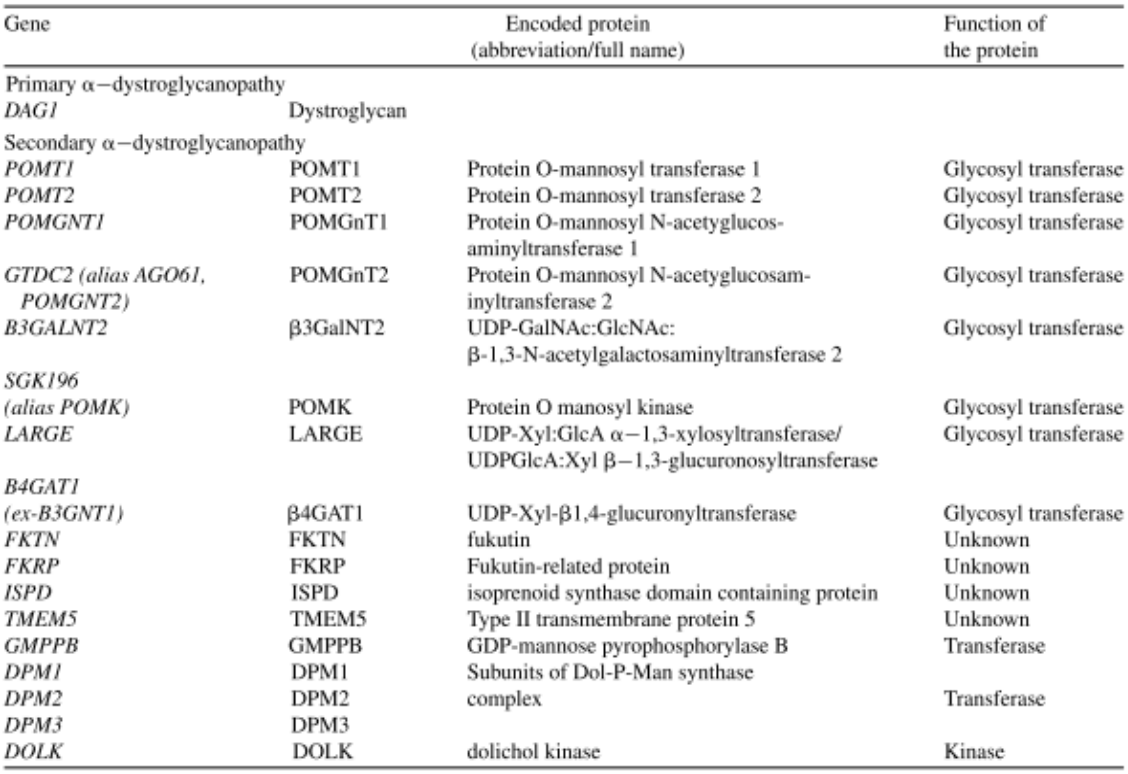


Figure 1. The dystroglycan complex consists of multiple protein subunits that connect the extracellular matrix to the cellular cytoskeleton. Figure taken from Barresi, 2018[2].

1.2 Dystroglycanopathies

Dystroglycanopathies include two forms of MD: Congenital MD (CMD) and Limb Girdle MD (LGMD). They are caused by improper glycosylation of the protein subunits composing the dystroglycan complex, a post-transcriptional modification. The α and β dystroglycan subunits are derived from the same gene - the transcribed dystroglycan mRNA is split into two mRNAs prior to translation[7]. Mutations in dystroglycan lead to decreased muscle fiber strength, similar to the result of mutations in dystrophin. Besides being involved in muscle fiber strength, the α-dystroglycan subunit is also involved in signaling pathways. Under certain conditions, it inhibits survival signaling in muscle cells via caspase activation, leading to muscle cell apoptosis[8], another characteristic result of MD. In addition, the dystroglycan complex serves as a node in the signal transduction pathway that leads to activation of STAT3 (signal transducer and activator of transcription 3). STAT3 plays an important role in regulating satellite cell self-renewal, and inducing expression of Interleukin 6, a cytokine that acts as an anti-inflammatory myokine[9]. Thus, improper formation of dystroglycan leads to multiple abnormal regulatory pathways. As stated before, for dystroglycan to be functional, it must be glycosylated. There are 17 genes many of which are enzymes that have been implicated in dystroglycan glycosylation (Table 1). Even within dystroglycanopathies, the symptom severity and time of onset varies greatly - even within individuals with mutations in the *same* genes, a perplexity that is not well understood[2]. This phenotypic complexity warrants further studies to understand the genetic basis behind these differences.

Table 1. Genes Associated with Dystroglycanopathic Muscular Dystrophy[10].



Gene	Encoded protein (abbreviation/full name)		Function of the protein
Primary α−dystroglycanopathy			
DAG1	Dystroglycan		
Secondary α−dystroglycanopathy			
POMT1	POMT1	Protein O-mannosyl transferase 1	Glycosyl transferase
POMT2	POMT2	Protein O-mannosyl transferase 2	Glycosyl transferase
POMGNT1	POMGnT1	Protein O-mannosyl N-acetyglucos-aminyltransferase 1	Glycosyl transferase
GTDC2 (alias AGO61, POMGNT2)	POMGnT2	Protein O-mannosyl N-acetyglucosam-inyltransferase 2	Glycosyl transferase
B3GALNT2	β3GalNT2	UDP-GalNAc:GlcNAc: β-1,3-N-acetylgalactosaminyltransferase 2	Glycosyl transferase
SGK196 (alias POMK)	POMK	Protein O manosyl kinase	Glycosyl transferase
LARGE	LARGE	UDP-Xyl:GlcA α−1,3-xylosyltransferase/ UDPGlcA:Xyl β−1,3-glucuronosyltransferase	Glycosyl transferase
B4GAT1 (ex-B3GNT1)	β4GAT1	UDP-Xyl-β1,4-glucuronyltransferase	Glycosyl transferase
FKTN	FKTN	fukutin	Unknown
FKRP	FKRP	Fukutin-related protein	Unknown
ISPD	ISPD	isoprenoid synthase domain containing protein	Unknown
TMEM5	TMEM5	Type II transmembrane protein 5	Unknown
GMPPB	GMPPB	GDP-mannose pyrophosphorylase B	Transferase
DPM1	DPM1	Subunits of Dol-P-Man synthase	
DPM2	DPM2	complex	Transferase
DPM3	DPM3		
DOLK	DOLK	dolichol kinase	Kinase

1.3 Skeletal Muscle Structure

Skeletal muscle is one of the three major muscle types. It consists of striated muscle tissue controlled voluntarily via the somatic nervous system. Skeletal muscle is attached to bones via bundles of collagen and its rigidity comes from the fractal arrangement of subunits (Figure 2). The muscle is composed of multiple muscle fascicles lined up in a parallel fashion, wrapped in a connective tissue sheath called the epimysium, which is surrounded by an outer connective tissue layer, called the fascia. The muscle fascicles are composed of multiple muscle fibers surrounded by another connective sheath called the perimysium; blood vessels and nerves are dispersed between the fascicles. The muscle fibers are cylindrical, multinucleated cells, with a sarcolemma cell membrane that result from fusion of multiple cells during myogenesis, or muscle development. The muscle fibers are surrounded by another connective tissue layer called the endomysium. Muscle fibers themselves consist of multiple myofibrils aligned in an organized parallel fashion with abundant mitochondria dispersed throughout. The myofibrils are composed of myofilaments, sarcomere-based structures composed of thin actin filaments and thick myosin filaments that shrink for muscle contraction and stretch for muscle relaxation[11,12]. Neuromuscular junctions (NMJ) are sites where a motor neuron meets a muscle fiber. Excitation signals travel from the brain down the neuron to the muscle fiber. The muscle fiber is surrounded by the sarcoplasmic reticulum, which upon exposure to acetylcholine provided by the action potential, opens up sodium channels that allow for an influx of sodium into the cell. This signal propagates, causing voltage sensitive calcium channels to open, allowing for an influx of calcium into the myofilaments which allows myosin to bring the actin filaments closer together via the sliding filament model, initiating muscle contraction[13].

Exercise and repeated use of muscles leads to structural changes in the muscle fibers. Some of these changes include angiogenesis - formation of more extensive capillary networks to meet the oxygen needs of the muscle, increased production of mitochondria, hypertrophy - increasing the diameter of the muscle fibers, and changes in the proportions of slow oxidative (SO), fast oxidative (FO), and fast glycolytic (FG) fibers[14]. Strenuous exercise can cause muscle fiber damage via overstretching of sarcomeres, leading to inflammation and damage to the connective tissue layers of the muscle. Overall, this leads to necrosis, that peaks 48 hours after strenuous activity or overuse. Healthy exercising persons undergo this process constantly and can regenerate the damaged muscle fibers through satellite cell proliferation and differentiation. Satellite cells are found underneath the basal lamina of muscle fibers[15] and begin proliferating when exposed to signals derived from damaged fibers and infiltrating immune cells. Following proliferation, they differentiate into myoblasts which through fusion replace the damaged muscle fibers[16]. Notch signaling is thought to play an important in role in stimulating this process[17]. In individuals with Muscular Dystrophy, regeneration after muscle fiber damage is impeded due to satellite cell depletion[16].

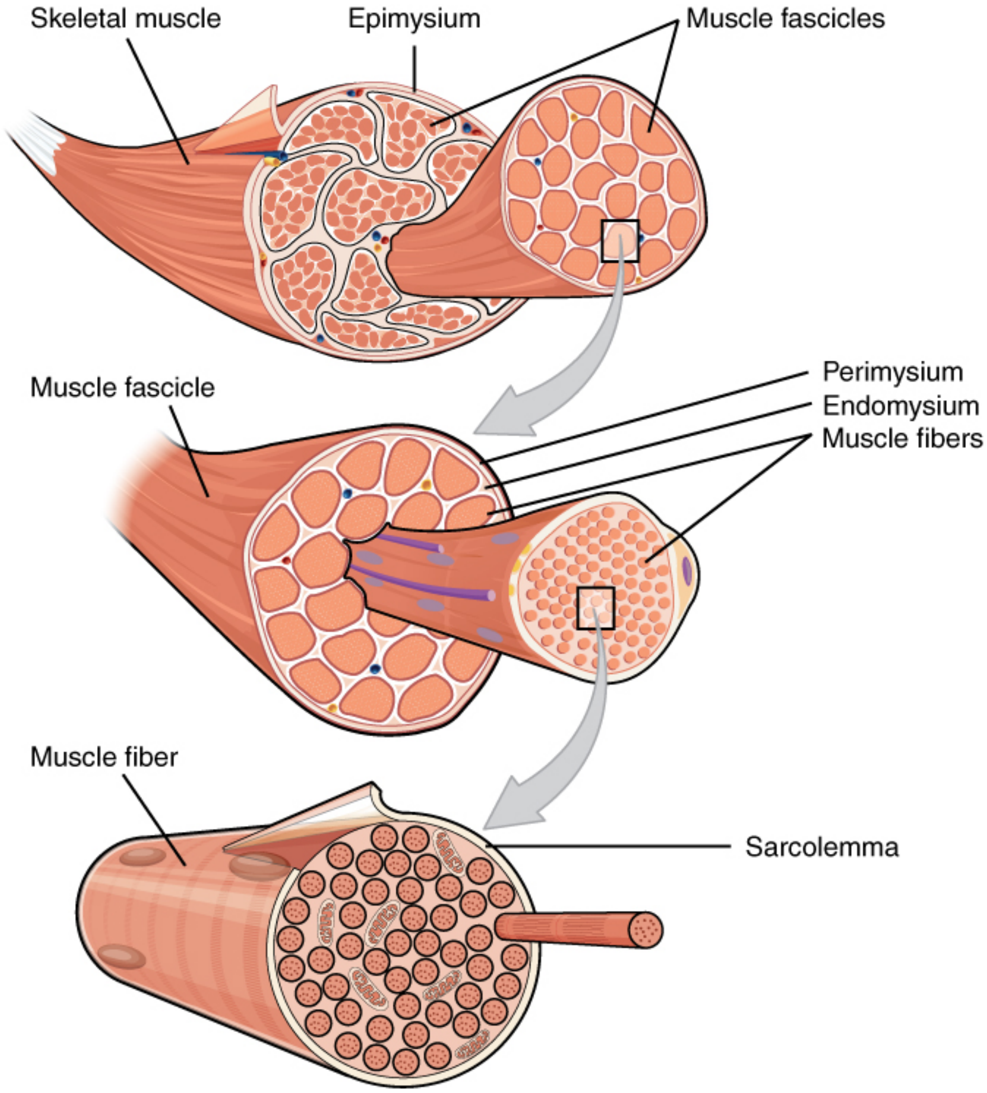


Figure 2. Skeletal Muscle Structure[12].

1.4 Zebrafish as a model organism for Skeletal Muscle Development

The zebrafish (*Danio rerio*) is a well-established vertebrate model organism that has been used to study neurological diseases including Alzheimer's and Parkinson's disease as well as musculoskeletal disorders. Zebrafish development is much faster than in mice models, and they are cheaper and require less physical space to grow. Zebrafish have 70% of genes conserved with humans[18]. Structurally, they have many of the same organs and systems that humans have, including a heart, brain, spinal cord, various types of musculature, blood, and both an innate and adaptive immune system. Several methods, including CRISPR/Cas9, can be used to develop transgenic zebrafish where specific mutations can be introduced.

Zebrafish skeletal muscle structure closely resembles that of humans, making them good models for musculoskeletal development. Using zebrafish, precursors to slow and fast twitch muscles have been identified and observed during development[19]. Additionally, the interaction between the muscle fibers attachment to the cytoskeleton and the extracellular matrix and how it relates to muscle rigidity has been determined[20], and the molecular mechanism of muscle cell contraction has been elucidated[21]. Of course, these are but a few of the discoveries that zebrafish models of skeletal muscle development have unraveled.

1.5 Zebrafish as a model organism for dystroglycanopathy

To better understand the complexity of dystroglycanopathy, accurate and useful model organisms are required. Most models for MD are mice models. In fact, mouse models of dystroglycanopathy exist that accurately represent the phenotypes of patients.

These mutants have mutated genes earlier presented in Table 1 including *large*[22], *pomt1*[23], *dag1* [24], and others. Unfortunately, similar zebrafish glycosylation mutants are still in development; current mutants exist for *ispd*[25], but most mutants are created transitively using morpholino knockdowns. While morphants are useful in some contexts, the need exists for stable zebrafish lines. A zebrafish model for Duchenne's Muscular Dystrophy (DMD) with mutated dystrophin exist which displays similar phenotypes as DMD human patients, including myofiber atrophy, immune cell infiltration into skeletal muscle, and abnormally shaped myofibrils[20]. Therefore, zebrafish have the potential to be an accurate, convenient, and fast-growing model for dystroglycanopathy research. The Henry lab at the University of Maine has been working on developing zebrafish with mutations in each of the genes listed in Table 1.

1.6 A *gmppb* mutant model of dystroglycanopathy

One of the dystroglycanopathic zebrafish models that has been successfully established by the Henry lab is a GDP-mannose Pyrophosphorylase (*gmppb*) mutant. The protein product of *gmppb* catalyzes the conversion of mannose-1-phosphate and GTP to GDP-mannose, a reaction involved in the production of N-linked oligosaccharides. These sugars are produced in the Golgi Apparatus and then are subsequently attached to a-dystroglycan as a post-translational modification. Mutations in *gmppb* have been associated with Limb Girdle MD and Congenital MD due to hypo-glycosylated α-dystroglycan[26]. As of 2018, 81 MD patients worldwide have been described with mutated *gmppb*[2,26–28]. The low number of documented cases may be based on a lack of screening. Astrea et al. tested this hypothesis, screening 73 Italian individuals with genetically unidentified forms of Congenital MD and α-dystroglycan hypoglycosylation for *gmppb* mutations[2]. Thirteen cases of *gmppb* biallelic mutations were identified in which seven novel mutations in *gmppb* were revealed: all leading to highly variable phenotypes from congenital clubfoot, seizures, neurodevelopmental abnormalities and autism spectrum disorders[2].

Zebrafish *gmppb* mutants were engineered by the Henry Lab using a CRISPR/Cas9 system previously described by Gagnon et. al[29]. Primers were designed using CHOPCHOP[29] with the intention of inserting a stop cassette in exon three of *gmppb*. CRISPR/Cas9 was performed in one cell stage AB zebrafish embryos (F0 generation) and the resulting fish were crossed to form the F1 generation which was similarly crossed to form the F2 generation. Data presented in this thesis is from the F2 generation and subsequent generations. Generation of the mutant line was done by the the Henry Lab[30]. The Henry Lab's *gmppb* mutant presents variable phenotypes, in a similar manner to human patients with *gmppb* mutations. The mutants display muscular atrophy, decreased muscle density, and disorganized muscle fibers (Figure 3), and can be classified by those with either severe or mild phenotype.

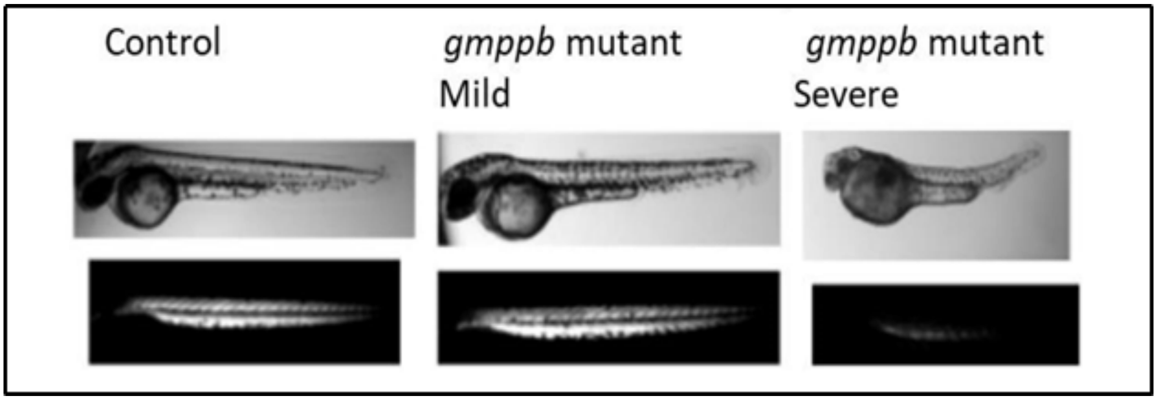


Figure 3. Differences in severity of MD phenotypes in *gmppb* mutant zebrafish at two days post-fertilization (2df). Control is *gmppb* wild type whereas mild and severe are homozygous *gmppb* mutants. The bottom figure of each zebrafish shows the birefringence which indicates skeletal muscle organization[3]. Figures courtesy of C. Henry lab.

1.7 Long non-coding RNAs

LncRNAs are regulatory genes located in regions of the genome previously termed "junk DNA" that provide a novel lens to view physiological processes and diseases[31]. These genes are transcribed into RNA, but not translated; thus, existing as RNA intermediates that regulate gene expression through diverse, uncharacterized mechanisms in both the nucleus and cytoplasm. For example, the lncRNA *XIST* directly interacts with DNA, signaling the condensations of one of the X-chromosomes in mammalian females, forming a barr body[31]. LncRNAs can also form secondary and tertiary structures that aid in their mechanism of action. The lncRNA HOTAIR is a repressor of tumor repressor and metastasis genes. It forms an intricate structure consisting of 56 helical segments, 38 terminal loops, 34 internal loops, and 19 junction regions[32]. LncRNAs are also able to regulate gene expression by hybridizing with mRNA gene transcripts to signal mRNA degradation, decreasing protein expression[31]. For example, the lncRNA α-HIF is a natural antisense transcript of hypoxia-inducible factor 1 alpha (HIF-1α) that binds to HIF-1a based on sequence similarity. When it does so, it exposes AU-riches elements present in the 3'UTR of the HIF-1α mRNA, increasing the speed of mRNA degredation[33]. Alterations in protein expression becomes much more complex when one considers the plethora of interactions that a single protein can have, thus lncRNAs act as essential nodes in a complex map of physiological processes. LncRNAs are considered the most functionally diverse and numerous classes of RNAs[34], yet their regulatory roles in the majority of processes and diseases is not well understood. They even have proposed hypothetical roles in MD[35], but have yet to be experimentally investigated in this context.

1.8 MicroRNAs

MicroRNAs (miRNAs) are another class of non-coding regulatory genes, who, in contrast to lncRNAs, have well understood mechanisms of action. Additionally, miRNAa are more conserved than lncRNAs with humans, and over 400 miRNAs are annotated by Ensembl[36] in the zebrafish. The biogenesis of miRNAs follows a conserved processing pathway. Following transcription by RNA Polymerase II or III, the pri-miRNA is cleaved by Drosha, a ribonuclease, to form a pre-miRNAs which is then able to exit the nucleus via export by Exportin 5/RanGTP. Next, it is bound by Dicer, another ribonuclease, which cleaves the hairpin structure such that the ~21 nucleotide fragment can be complexed with Argonaut to form the RNA-induced silencing complex (RISC complex)[37,38]. Once the complex is formed, miRNAs can post-transcriptionally modify gene expression through a number of mechanisms. Based on the 4-9nt "seed sequence" of the miRNA complex, complementary base-pairing of target mRNA transcripts can occur. This may lead to cleavage of the target mRNA, poly-A-tail shortening, or blockage of the ribosome binding site, preventing translation etc. In vertebrates, miRNAs primarily function by degrading target mRNAs[39]. Currently, the specific role miRNAs play in MD is under-investigated. MiR-188 has been identified as a biomarker of Duchenne's MD, but it is unclear if this miRNA might be contributing towards the phenotype[40]. Moreover, numerous miRNAs have been shown to modulate apoptosis, regeneration, cell growth and organization: processes that are likely pertinent to MD. Thus, investigating both miRNAs and lncRNAs is a necessary step towards a better understanding of the genetic pathways involved in this disease, and may have implications on the range of phenotypic severity and lifespan MD patients display. By incorporating protein-coding gene expression, miRNA expression, and lncRNA expression in genetic regulatory pathways, treatments could emerge that target specific pathways to treat resulting symptoms, informing and advancing our understanding of MD.

1.9 RNA sequencing to measure gene expression

RNA sequencing (RNA-Seq) is a high-throughput method to determine gene expression. It can be used to identify differentially expressed genes, genes that are turned on or off in certain situations and in response to stimuli, which can be used to answer numerous research questions. RNA-Seq can also be used to examine alternative splicing where exon usage may vary in different tissues or samples.

Illumina high-throughput RNA sequencing begins with isolation of RNA transcripts. Since total RNA recovered using standard procedures contains >80% ribosomal RNA (rRNA)[41], standard protocol includes selection for poly-adenylated sequences using magnetic beads or cellulose coated with oligo-dT molecules, a process that removes most of the rRNA. Next, the transcripts are fragmented and converted to cDNA with ligated adapters used for next generation sequencing (NGS). To ensure proper removal of rRNAs, rRNA depletion is performed. Multiple strategies for rRNA depletion exist, but most utilize rRNA probes that target rRNA transcripts, signaling them for degradation. For example, Roche's KAPA RNA HyperPrep Kit with Riboerase utilizes rRNA DNA probes that hybridize to rRNA fragments, forming RNA-DNA hybrids, that are targeted by RNase H for degradation. This method has less off-targets and preserves a higher proportion of non-coding RNAs than other strategies for rRNA depletion[42].

Samples are then sequenced using Next Generation Sequencing (NGS). Following this, a workflow is performed that utilizes multiple software to assemble and align reads, annotate genes, and calculate gene expression. The output of sequencing is typically two FASTQ files for each sequenced sample, one for forward reads and another for reverse reads. First, these files are checked using FastQC[43] which determines the read length and quality. Next, the two FASTQ files are concatenated into one file, trimmed of the adapters used for sequencing, and aligned to a given genome using either HISAT2[44] or BowTie[45]. If one is interested in identifying novel transcripts, Stringtie[46] is run to align the transcripts to genes and identify ensembl-annotated genes. The bam file that is the result of BowTie or HISAT2 is run through HTSeq2[47] which counts the number of transcripts that align to each of the genes, and then DESeq2[48] can be used to identify differentially expressed genes.

Small RNA sequencing allows for the preferential sequencing of microRNAs (miRNAs). It selects miRNAs using bead or gel-based size selection. Unlike most cellular RNAs, mature miRNAs possess both a 3' hydroxyl and a 5' phosphate which allows for preferential adapter ligation[49]. From there, cDNA preparation using reverse transcriptase, sequencing, and transcript alignment and annotation occur as would in traditional RNA Sequencing.

RNA Sequencing provides information about the expression of protein coding genes which account for a mere 2% of the genome, as well as non-protein coding like long non-coding RNAs (lncRNAs) and microRNAs (miRNAs). A relatively newer view of the molecular mechanisms that lead to phenotype is that the way protein-coding gene are regulated, where they are expressed and when, is just as important as the protein-coding genes themselves. Therefore, traditional sequencing that excludes non-coding genes, potent genetic regulators of protein-coding genes, excludes a plethora of valuable information. Non-coding regulatory genes are severely under investigated in MD, thus, we aimed to emphasize them in our analysis.

1.10 MiRNA Network Attack

Co-expression networks are used to understand large genetic networks. In the context of gene expression, they can be used to identify possible interactions between genes. In these networks, the nodes represent individual genes, and edges between them indicates that two genes are correlated – that their levels of expression are similar i.e. r2 >0.75). These co-expression networks can then be used for further analyses to identify gene candidates.

Network attack models are a computational method of modeling a network's vulnerability in response to removal of nodes or edges. This is used for identification of the most important edges or nodes, those that contribute most towards the network stability[50]. They have been used to model social networks, identify key players in criminal organizations[51], and model power grids[52], and have proposed benefits in modelling relationships from large datasets generated from biological research[50].

Network stability is defined by multiple parameters including the characteristic path length, the degree of separation, and the network size. The characteristic path length is the shortest path (the least number of edges) between two nodes. Nodes or edges that are most important to the network stability, cause relatively large changes in the characteristic path length upon removal. Another factor to consider in network attack graphs is the degrees of separation. The value of the degrees of separation defines the number of nearly independent networks, those with relatively few connections to other networks. An example of this could be the social activities of people worldwide. If the nodes were people, and the edges were interactions, one would expect that each continent would have its own relatively independent network that would consist of far fewer edges connecting nodes between different continents than nodes within the same continent. Thus, the degree of separation would be six, based on the number of *inhabited* continents. Moreover, if you repeatedly selected edges that were transcontinental, after removing the last transcontinental edge, a huge change in the characteristic path length would occur. The network size is another useful parameter in finding important nodes. By repeatedly removing nodes, large jumps in network size can be used to pinpoint nodes that contribute to network stability.

Network attack has been used to model miRNA networks to compare miRNA expression in different forms of cancer[53]. In the context of miRNA expression, nodes represent individual miRNAs and edges connect the distinct miRNAs if they show similar patterns of expression between different timepoints, treatments, etc. These co-expression networks can identify miRNAs that are controlled by the same transcriptional pathway or functionally related[54]. Furthermore, by attacking the network through removal of miRNAs, nodes can be identified that cause large changes in characteristic path length or network size. These nodes are thus important in maintaining the structure of the network and are interesting candidates for further investigation.

1.11 MiRNA gene target analysis

MiRNAs recognize mRNA targets based on complementary base pairing of the seed sequence of the miRNA with the mRNA transcript. TargetScanFish[55] (version 6.2) is a database with lists of zebrafish mRNAs targeted by zebrafish miRNAs that can be used in combination with differential gene expression to predict miRNA/mRNA relationships. If a miRNA is differentially expressed and it is targeting a specific mRNA transcript, it would be expected that the mRNA transcript would be differentially expressed in an opposite direction of the miRNA, since most miRNAs cause degradation of their targets. The relationship between miRNAs and their mRNA targets can be used to identify upstream putative pathways that lead to biological differences.

1.12 Research Objectives

The primary goal of this research is to construct genetic regulatory networks that link miRNAs, lncRNAs, and protein-coding genes to cellular processes implicated in dystroglycanopathy phenotypes. We will do so with the following objectives:

1. Identify previously annotated lncRNAs and miRNAs that are differentially expressed in the mutants.
    a. Predict miRNA targets
2. Identify candidate *novel* lncRNAs via de-novo analysis from unannotated transcripts that are differentially expressed in the *gmppb* mutants.
    a. Characterize these lncRNAs and the adjacent genes to determine if they are relevant to MD.
3. Incorporate lncRNA, miRNA, and protein-coding gene expression to construct a genetic regulatory network that may contribute towards our understanding of the molecular pathways involving MD phenotype.
4. Characterize the different types of mutations that were induced in the *gmppb* mutants through Polymerase Chain Reaction (PCR) amplification of *gmppb* mutants and subsequent sequencing.

2. MATERIALS AND METHODS:

2.1 Overview of the experimental design

Four biological replicate samples of each treatment at each timepoint were submitted for rRNA depletion, followed by small RNA sequencing and mRNA sequencing (Figure 4). This included four samples of heterozygous *gmppb* mutants, four samples of homozygous *gmppb* mutants that displayed mild phenotypes, and four samples of homozygous *gmppb* mutants that displayed severe phenotypes. Each sample consisted of total RNA made from homogenizing three embryos (as described in section 2.2). The two timepoints included were 4 and 7 days post fertilization (dpf). For the small RNA Sequencing, only 3 samples were submitted for each treatment due to insufficient total RNA quantity.

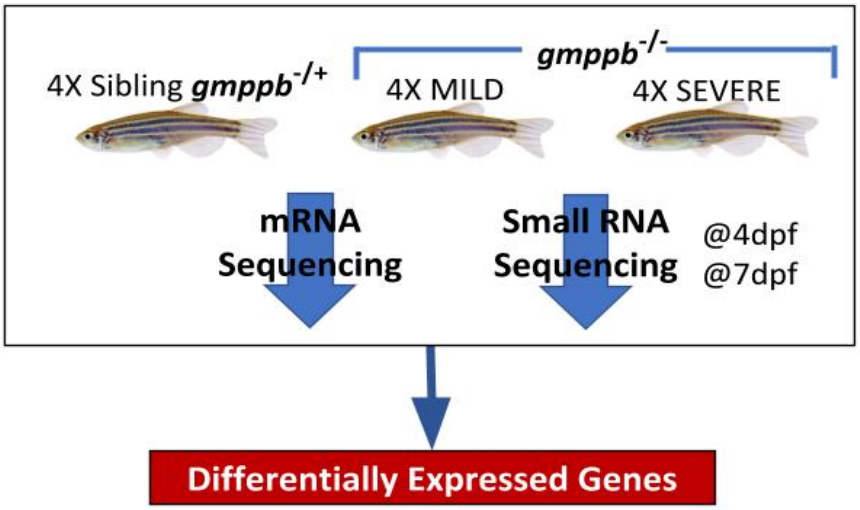


Figure 4. Experimental overview of RNA Sequencing and small RNA Sequencing in gmppb$^{+/-}$ controls and gmppb$^{-/-}$ mutants exhibiting mild or severe phenotypes.

2.2 RNA extraction

Homozygous mutant and sibling control embryos were raised to the specified time point (4 or 7 days post fertilization), and then the mutants were classified based on birefringence using a confocal microscope in the Henry Lab. Birefringence was quantified using FIJI as previously described[20]. Mutants with a percent area and a percent Mean Gray Value at 85% or higher were classified as mild mutants; mutant embryos that did not meet this standard were classified as severe[30]. Zebrafish were segregated into separate tubes based on this classification system, euthanized via tricane, and preserved in 300 μL of trizol by the Henry Lab following an approved University of Maine IACUC protocol. To obtain sufficient RNA for RNA Sequencing, small RNA Sequencing, and quantitative Polymerase Chain Reaction (qPCR), each sample consisted of 3 zebrafish embryos.

Samples were defrosted and homogenized using a Fisher PowerGen 125 (Fisher Scientific, Waltham, MA) mechanical homogenizer. RNA Extraction was performed using a Quick-RNA MicroPrep kit from Zymo following the manufactures protocol (Zymo Research, Irvine, CA). Samples were centrifuged at 12,000xg for 1 minute and the supernatant was removed and placed into a clean test tube. One volume ethanol (95-100%) was added to each sample, mixed well, and then the mixture was transferred to a Zymo-Spin IC column with a collection tube. The column was centrifuged 30 seconds at 12,000xg and the flow through was discarded. The column was washed with 400 μL RNA Wash Buffer, centrifuged for 30 seconds at 12,000xg, and flow through was discarded. A DNase I Mastermix was prepared in an RNase free tube with 5 μL DNase I and 35 μL DNA Digestion buffer per sample. The master mix was mixed via gentle inversion. Next, the column was washed with 400 μL of RNA Prep Buffer followed by 700 μL RNA Wash

Buffer, each time centrifuging for 30 seconds at 14,000xg and discarding the supernatant. Next, the column was washed with 400 µL RNA Wash Buffer and centrifuged 2 minutes, then placed into a new RNase free tube. 10 µL of DNase/RNase Free Water was added to the center of the column and the sample was eluted via centrifugation for 30 seconds at 14,000xg. The eluted RNA quality and concentration was immediately read by a Thermo Scientific NanoDrop OneC Spectrophotometer (Waltham, MA) and then samples were stored in a -80C freezer. The Zymo protocol is available online at https://files.zymoresearch.com/protocols/_r1050_r1051_quick-rna_microprep_kit.pdf.

2.3 RNA sequencing and small RNA Sequencing

Samples were submitted for RNA sequencing at QuickBiology in Pasadena, California. Total RNA samples were assayed for quality using an Agilent Bioanalyzer 2100 (Agilent Technologies, San Francisco, CA) by QuickBiology.

Libraries for RNA-seq were prepared with a KAPA Stranded RNA-Seq Kit with a RiboErase (KAPA Biosystems, Wilmington, MA) system. Final library quality and quantity were analyzed by Agilent Bioanalyzer 2100 (Agilent Technologies, San Francisco, CA) and Life Technologies Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA), respectively. The RNA Sequencing was performed on a HiSeq 4000 Illumina Sequencer (Illumina Inc, San Diego, CA) with 150 base paired end reads.

The same total RNA samples were submitted for small RNA Sequencing. The library was prepared according to Qiagen QIAseq miRNA library kit (Qiagen Inc, Germantown, MD) using 100 ng total RNA as input. Final library quality and quantity was analyzed by Agilent Bioanalyzer 2100 and Life Technologies Qubit 3.0 Fluorometer. The

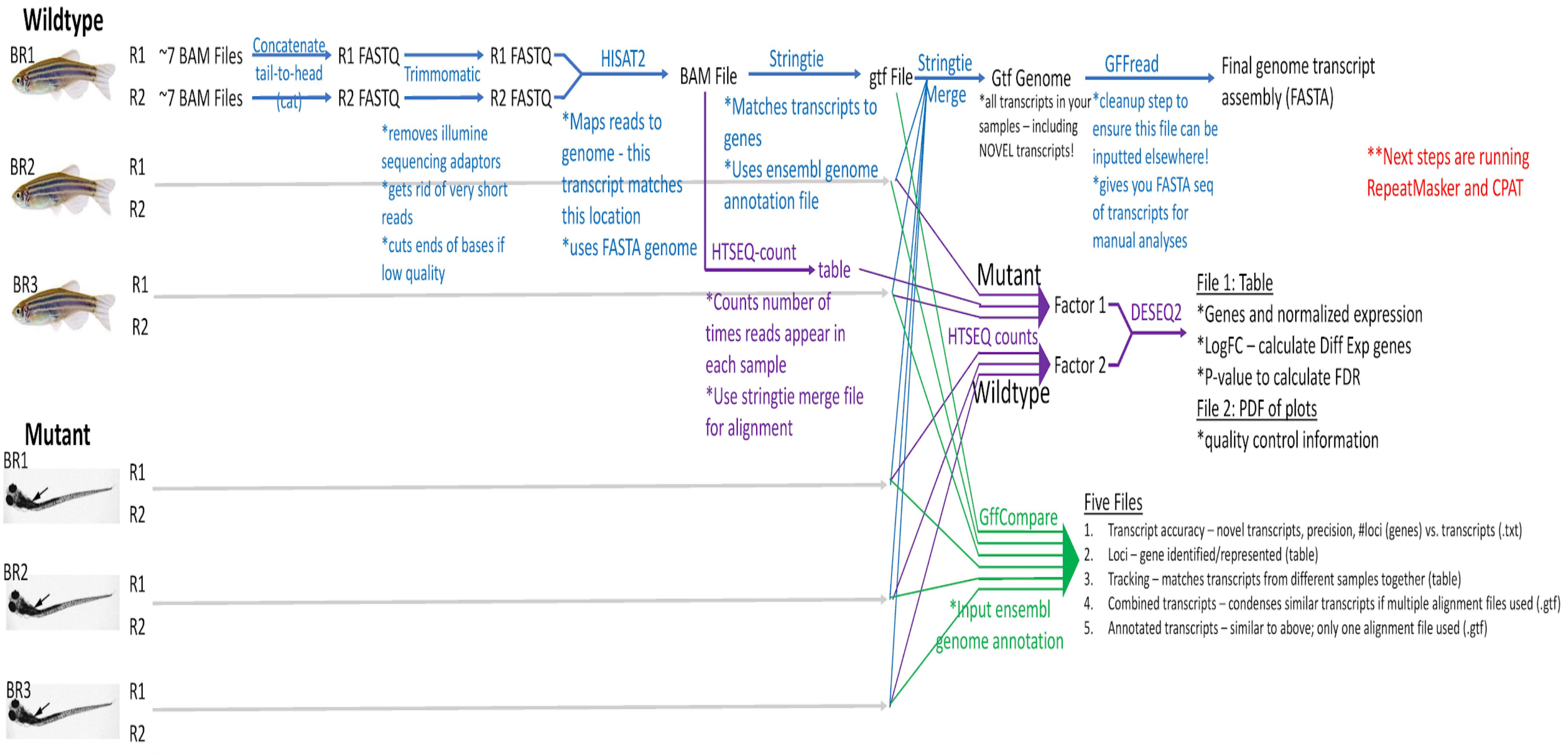
Small RNA Sequencing was performed on an Illumina NextSeq 500 (Illumina Inc, San Diego, CA) with 75 base paired end reads.

2.4 RNA sequencing annotation workflow

FastQC version 0.11.9[43] was used to verify the quality of the RNA Sequencing reads prior to further analyses. All analyses were performed using Galaxy (https://usegalaxy.org) and the histories of analyses be accessed through Galaxy upon request.

Following FastQC diagnostic analyses, each of the FASTQ files were concatenated tail-to-head to produce a single set of Forward (R1) and Reverse (R2) FASTQ files per sample. FASTQ files were then trimmed using Trimmomatic version 0.38.0[56] which removes the sequencing adapters, and low quality bases. The forward and reverse reads were mapped to the GRCz11/danRer11 (May 2017) zebrafish genome assembly (https://www.ncbi.nlm. nih.gov/assembly/ GCA_000002035.4/) using HISAT2 version 2.1.0[44]. The resulting BAM (binary alignment and mapping) file for each sample was then used to develop gene models using StringTie version 3.1.6 and the GRCz11 Ensembl (version 98) GTF annotation file. The GTF file for each sample was combined using StringTie Merge to produce a single GTF annotation file. A FASTA formatted sequence file for each transcript in the GTF file was produced using GFFread from within Cufflinks version 2.1.1.2[57]. Next, the BAM files were run through HTSeq version 0.9.1[47] to count the number of reads that map to exons of genes in the GTF annotation file. Read counts per gene per sample were analyzed using DESeq2 version 2.11.40.6[48]. Four different pairwise sample group comparisons were made to determine which genes were differentially expressed. These comparisons included: 4dpf sibling vs. 4dpf mild, 4dpf sibling vs. 4dpf severe, 7dpf sibling vs. 7dpf mild, and 7dpf sibling vs. 7dpf severe. For each pairwise comparison, the normalized expression across all samples, $\log_2$ Fold Change ($\log_2$FC), log ratio statistic, p-value, and false-discovery rate (FDR) adjusted p-value were computed for each gene using DESeq2. After merging the DESeq2 files with the Ensembl annotation, genes were subset based on gene type (i.e protein-coding, long non-coding RNAs, miRNAs, etc.). Additionally, a script was run to annotate un-annotated transcripts that were structurally similar to other annotated transcripts. Finally, GffCompare was used to identify unannotated transcripts that were used for novel lncRNA identification.

Below is a pictorial example of this workflow using triplicate wildtype and triplicate *gmppb* mutants (Figure 5).


Wildtype
BR1
R1
R2
~7 BAM Files
~7 BAM Files
Concatenate
tail-to-head
(cat)
R1 FASTQ
R2 FASTQ
Trimmomatic
R1 FASTQ
R2 FASTQ
*removes illumine sequencing adaptors
*gets rid of very short reads
*cuts ends of bases if low quality
HISAT2
*Maps reads to genome - this transcript matches this location
*uses FASTA genome
BAM File
Stringtie
*Matches transcripts to genes
*Uses ensembl genome annotation file
gtf File
HTSEQ-count
table
*Counts number of times reads appear in each sample
*Use stringtie merge file for alignment
Stringtie
Merge
Gtf Genome
*all transcripts in your samples – including NOVEL transcripts!
GFFread
*cleanup step to ensure this file can be inputted elsewhere!
*gives you FASTA seq of transcripts for manual analyses
Final genome transcript assembly (FASTA)
**Next steps are running RepeatMasker and CPAT
BR2
R1
R2
BR3
R1
R2
Mutant
Factor 1
HTSEQ counts
Factor 2
Wildtype
DESEQ2
File 1: Table
*Genes and normalized expression
*LogFC – calculate Diff Exp genes
*P-value to calculate FDR
File 2: PDF of plots
*quality control information
Mutant
BR1
R1
R2
BR2
R1
R2
BR3
R1
R2
GffCompare
*Input ensembl genome annotation
Five Files
1. Transcript accuracy – novel transcripts, precision, #loci (genes) vs. transcripts (.txt)
2. Loci – gene identified/represented (table)
3. Tracking – matches transcripts from different samples together (table)
4. Combined transcripts – condenses similar transcripts if multiple alignment files used (.gtf)
5. Annotated transcripts – similar to above; only one alignment file used (.gtf)

BR = Biological Replicate; R1/R2 = Read 1 or Read 2; FDR = False Discovery Rate

Figure 5. Gene annotation workflow for identifying differentially expressed ensembl annotated genes and unannotated novel transcripts.

2.5 Small RNA sequencing workflow

Analysis of differentially expressed miRNAs was performed using miRExpress version 2.0[58] using only the read 1 (R1) FASTQ files for each of the three biological replicate samples profiled using small RNA sequencing. MiRExpress was used to trim adapters and then align the trimmed reads to precursor miRNA sequences provided by miRGeneDB version 2.0[59]. The number of aligned reads to the 5p- or 3p-ends of the precursor miRNA sequences were reported. These read counts for the mature miRNAs were analyzed using DESeq2 to perform the same four pairwise comparisons done for the RNA sequence described above.

2.5 Splicing analysis of *gmppb* mutants

To verify incorporation of the STOP cassette in the *gmppb* homozygous mutants, gDNA hotshot extractions were performed based on the procedure described by Gagnon et. al[29]. One 4dpf embryo was placed in a PCR tube per sample in non-lethal tricane. Once the embryo appeared to be asleep, the liquid was removed and 20 µL of 50mM NaOH was added. The sample was heated in a PCR machine for 20-30 minutes at 95C and then cooled to 4C. 2 µL 1M TrisCl, Ph 8.0 was added and mixed via pipetting up and down. Samples were frozen at 4C prior to Polymerase Chain Reactions.

To amplify the *gmppb* target region, PCR was performed with the reverse primer (TGAAAGCTCTGATTCTTGTCGGTG) and the forward primer (CTGGTGGAACTTG AGCATGTCGT). The genomic DNA was spun down on a bench top centrifuge and then 2 µL was added to each reaction tube in a 96 well PCR plate. To the reaction each of the following were added: 1 µL of 20 uM primer (forward + reverse mixed), 0.125 µL Taq

Polymerase (5000 Units/mL, New England Biological Laboratory, Ipswich, MA), 0.5 μL of 10mM dNTPs, and 21.375 μL DI water -to bring the total volume to 25 μL. Samples were quickly mixed via inversion, spun down, and placed in the PCR machine with the following cycle temperatures and times:

Step 1: 95 degrees 3 minutes initial denaturing
Step 2: 95 degrees 20 seconds denature
Step 3: 60 degrees 25 seconds anneal
Step 4: 68 degrees 30 seconds extend
Step 5: Repeat steps 2-4 34 more times
Step 6: 68 for 5' final extension
Step 7: Hold samples at 4C

The samples were submitted for sequencing at 10ng/mL and then the resulting forward and reverse reads were aligned with wildtype *gmppb* and the STOP cassette using Basic Local Alignment Search Tool (BLAST version 2.10.0)[60]. Mutants were categorized based on the types of mutations that were present and the degree of and location of stop cassette insertion. Mutated reads were analyzed using Open Reading Frame Finder (ORF Finder) to validate whether stop codons were present in all possible reading frames. Finally, the RNA Sequencing data was aligned based on the results and categorization of different mutations found via qPCR analysis, to attempt to categorize the mutants based on their *gmppb* mutations.

2.6 Identification and characterization of differentially expressed novel and previously annotated lncRNAs and nearby genes

Previously annotated differentially expressed lncRNAs were subset from the annotated genes as described previously in section 2.4. Unannotated transcripts were also subset as described in section 2.4. Fasta Sequences for each of the transcripts were generated using GFFread[57]. To identify novel potential lncRNAs, the un-annotated sequences were run through Coding Potential Assessment Tool (CPAT version 1.2.2)[61] and then RepeatMasker (version 4.1.0) to identify transcripts with high coding potential and highly repetitive transcripts, respectively. A perl script was used to subset transcripts with a coding potential less than 0.38. The RepeatMasker results were compiled to generate a ratio of repetitive bases per transcripts (Appendix B, Section B.2), and transcripts were subset based on a threshold of less than 50% repetitive bases.

2.7 Characterization of differentially expressed protein coding genes

Protein coding genes were subset from the annotated DESeq files generated in section 2.4. Differentially expressed genes are defined as those with either adjusted or non-adjusted p-values less than 0.05. Information about each of the genes was gathered using Ensembl Biomart[66], including the gene name and description. Venny 2.0 (https://bioinfogp.cnb.csic.es/tools/venny/) was used to subset genes based on temporal and sample dependent expression. Enriched gene sets with Gene Ontology annotations were determined using David[67,68] and Panther[69]. Additional functional information about genes was obtained from GeneCards[64] and AmiGO[70,71].

2.8 Characterization of differentially expressed miRNA and network attack

Differentially expressed miRNAs were subset from the small RNA Sequencing data based on an adjusted p-value less than 0.05. An R script was generated to determine the predominantly expressed miRNA based on expression, thus, the non-predominantly expressed miRNA arm (otherwise known as the passenger strand) was discluded from future analyses (Appendix B, section B.1). Venny 2.0 was used to look for trends in expression across multiple samples and time points.

Network attack plots were generated using a combination of R and python scripts to better understand the robustness of the networks present in the 4dpf/7dpf siblings, 4dpf/7dpf mild mutants, and 4dpf/7dpf severe mutants. Scripts to generate these plots can be requested. Networks were imported into and visualized with Cytoscape version 3.7.2[72].

2.9 miRNA target analysis

MiRNA mRNA targets were identified from TargetScanFish[55]. The MirBase IDs present in the TargetScanFish files were converted to MirGeneDB[59] IDs using MiRExpress[58] which are the IDs present in the DESeq2 data. MiRNAs in the small RNA Sequencing dataset that were differentially expressed at 4dpf in the severe mutants were subset according to an adjusted P value < 0.05. Only the predominant strand of each miRNA was included in the analyses, the passenger strand was excluded from the data. Ensembl annotated genes from RNA sequencing that were differentially expressed at 4dpf in the severe *gmppb* mutants according to an unadjusted p value < 0.05 were also subset. These two lists were merged to look for differentially expressed miRNAs with differentially expressed mRNA targets. Finally, since miRNAs typically induce degradation of their mRNA targets, miRNA and mRNA targets were subset based on an opposite pattern of differential gene expression (i.e. miRNA was downregulated, and mRNA was upregulated OR miRNA was upregulated, and mRNA was downregulated). This data was used to generate networks in Cytoscape version 3.7.2[72].

2.10 Comparison of gene expression to cardiac regeneration

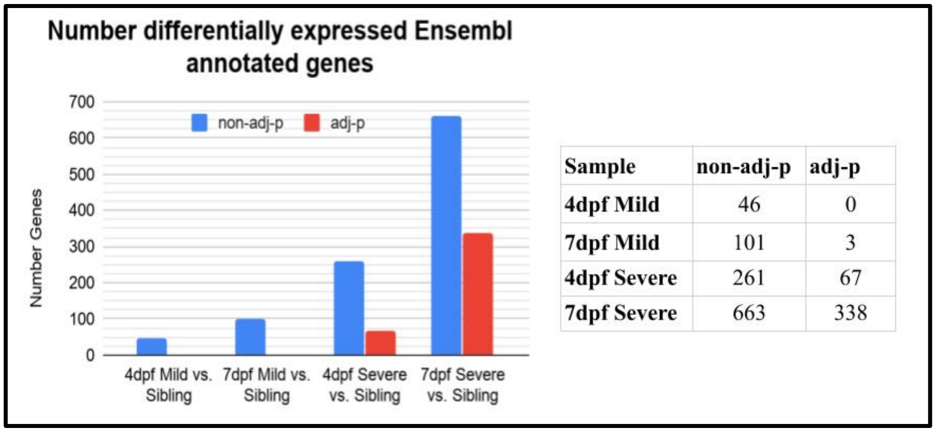
To highlight genes that might be involved in muscle-related phenotypes, protein-coding and miRNA gene expression was compared to that of a previous study exploring differentially expressed genes involved in cardiac regeneration in zebrafish[73]. Venny 2.0 was used for comparison and gene ontologies for the protein-coding genes were determined as described in section 2.7.

3. RESULTS

3.1 mRNA sequencing of *gmppb* mutants

3.1.1 Identifying differentially expressed Ensembl annotated genes.

To identify potential gene candidates involved in modulating phenotypic severity, differentially expressed gene transcripts were determined by comparing expression between pairs of sample groups (Figure 6). The gene transcripts analyzed were those annotated by Ensembl. As expected, using adjusted p-values, a more stringent significance threshold decreases the number of genes defined as differentially expressed. At both timepoints, severe mutants compared to sibling controls had a greater number of differentially expressed genes relative to the mild mutants compared to sibling controls. Next, a comparison of sets of differentially expressed genes from the four pairwise comparisons was performed to determine the temporal and sample-specific expression of the Ensembl annotated genes (Figure 7).



Sample	non-adj-p	adj-p
4dpf Mild	46	0
7dpf Mild	101	3
4dpf Severe	261	67
7dpf Severe	663	338

Figure 6. From mRNA sequencing, a list of differentially expressed gene transcripts defined by a threshold of $p<0.05$, or adjusted $p<0.05$ were subset. Those with Ensembl annotations are included in this figure. The table shows the number of differentially expressed genes in each sample according to the threshold definition of differentially expressed.

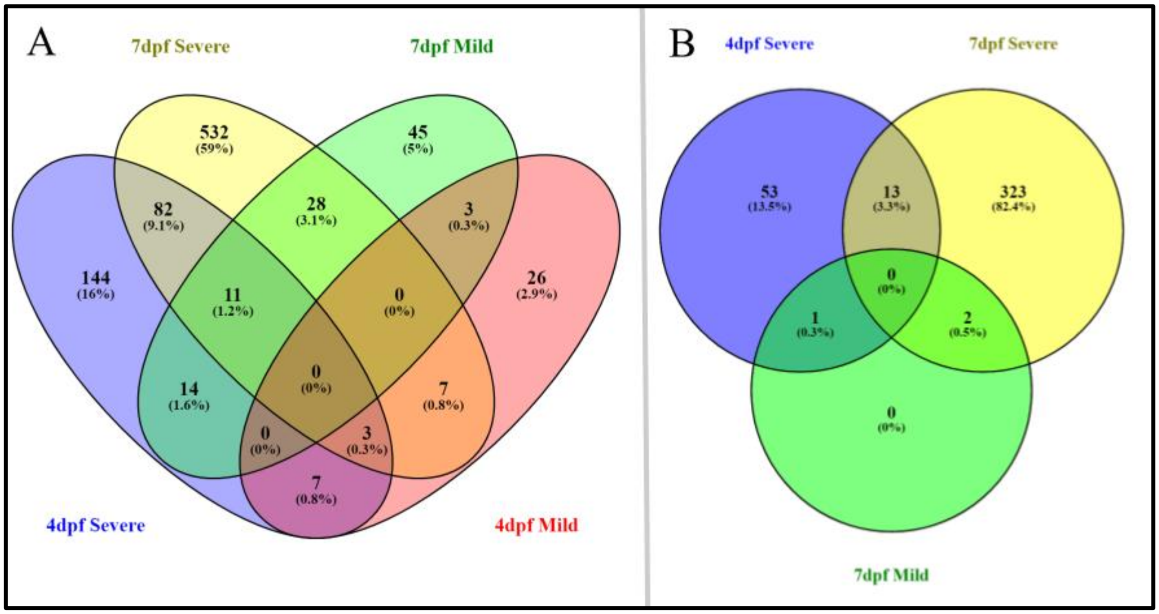


Figure 7. Venn diagrams of differentially expressed Ensembl annotated genes. Panel A includes genes with **unadjusted** p-values < 0.05 while panel B includes genes with **adjusted** p-values < 0.05. Regions of overlap indicates genes that are differentially expressed at the multiple specified timepoints or phenotypes. Percentages are included to indicate the number of genes in each category over the total number of differentially expressed genes.

Next, to characterize genes differentially expressed in the severe mutants at both 4dpf and 7dpf, Gene Ontology annotations were performed using PANTHER and DAVID. Of the 82 differentially expressed genes (p-value < 0.05) common between 4dpf and 7dpf, but not shared with 4dpf or 7dpf mild, only 47 were mapped to genes represented in PANTHER or DAVID. Gene Ontology terms of interest included muscle organization, extracellular matrix, cell adhesion, mitochondrial function, the immune system, and transcriptional regulators (Tables 3 and 4).

The ten most differentially expressed Ensembl annotated genes were subset as indicated by the largest positive or negative fold change ($\log_2$ FC) and an adjusted p-value < 0.05. Eight of these genes came from the 7dpf severe mutants and two came from the

4dpf severe mutants (Table 2). Three of these genes lack gene descriptions, and two of them were not annotated as protein-coding genes.

It is important to note that in these analyses, all Ensembl genes, regardless of whether they were protein-coding or not, were included. Therefore, non-coding genes, including miRNA precursors and annotated lncRNAs, were present. Of the 881 total genes differentially expressed across all pairwise comparisons, 22 were annotated as lncRNAs and two as miRNA precursors. Of just the 82 genes differentially expressed in both the severe mutants, one was annotated as a lncRNA and there were no annotated miRNA precursors.

Table 2. Ten most differentially expressed Ensembl annotated genes.

Gene ID	Name	Gene Description	4dpf Severe			7dpf Severe		
			Base	$\log_2$FC	Adj P	Base	$\log_2$FC	Adj P
ENSDARG 00000008840	hmbsa	hydroxymethylbilane synthase a	202	-0.49	0.528	164	2.93	2E-08
ENSDARG 00000011533	sema6dl	sema, transmembrane, and cytoplasmic domain, semaphorin 6D, like	124	4.99	0.000	4	-0.97	3E-01
ENSDARG 00000011921	txnl1	thioredoxin-like 1	1	0.30	0.000	4	-2.82	2E-04
ENSDARG 00000044484	cdc23	CDC23 (cell division cycle 23, yeast, homolog)	84	-0.18	0.820	55	2.77	4E-13
ENSDARG 00000056590	calca	calcitonin/calcitonin-related polypeptide, alpha	31	-2.02	0.000	32	3.78	8E-13
ENSDARG 00000074676	rprd2b	regulation of nuclear pre-mRNA domain containing 2b	4	0.57	0.000	11	-2.67	4E-06
ENSDARG 00000076320	ano9a	anoctamin 9a	873	-0.12	0.926	766	2.67	2E-08
ENSDARG 00000093922	CR354542.1	Processed transcript	55	0.10	0.919	38	2.79	8E-08
ENSDARG 00000104365	BX323064.2	Antisense	32	0.75	0.332	113	-3.39	1E-36
ENSDARG 00000104919	si:ch211-153b23.3	Protein coding	61	2.81	0.000	13	-1.23	2E-01

Table 3. Differentially expressed (**unadj p <0.05**) genes with selected Gene Ontology annotations at 4 and 7dpf in the severe mutants.

Gene ID	Name	Gene Description	4dpf Severe			7dpf Severe		
Muscle Organization			Base	log$_2$FC	P Val	Base	log$_2$FC	P Val
ENSDARG00000000563	ttn.1	titin, tandem duplicate 1	52	0.83	0.039	46	-0.98	0.042
ENSDARG00000028213	ttn.2	titin, tandem duplicate 2	47	0.73	0.047	42	-0.92	0.030
ENSDARG00000045302	smpx	small muscle protein X-linked	124	-0.70	0.001	131	-0.41	0.027
Extracellular Matrix								
ENSDARG00000042816	mmp9	matrix metallopeptidase 9	43	1.00	0.001	36	-0.85	0.008
ENSDARG00000061904	fhod3b	formin homology 2 domain containing 3b	24	-0.97	0.012	90	-2.56	0.000
Cell Adhesion								
ENSDARG00000093008	adgrf3b	adhesion G protein-coupled receptor F3b	13	-1.36	0.005	30	-0.87	0.024
Immune System								
ENSDARG00000042816	mmp9	matrix metallopeptidase 9	43	1.00	0.001	36	-0.85	0.008
Mitochondrial Function								
ENSDARG00000038643	alas2	aminolevulinate, delta-, synthase 2	258	-1.62	0.001	733	-1.31	0.002
ENSDARG00000063922	mt-nd6	NADH dehydrogenase 6, mitochondrial	581	0.87	0.000	374	-0.55	0.041
ENSDARG00000069852	lipt2	lipoyl(octanoyl) transferase 2	75	0.58	0.013	70	-0.64	0.006
Developmental Pathways								
ENSDARG00000074148	**RAS** rbpjl	recombination signal binding protein for ig kappa J region	1073	0.53	0.001	587	1.81	0.000
ENSDARG00000040959	**NOTCH** rabl3	RAB, member of RAS oncogene family-like 3	126	0.48	0.033	65	1.34	0.000

Table 3 Continued

Gene ID	Name	Gene Description	4dpf Severe			7dpf Severe		
Transcriptional Regulation			Base	log$_2$FC	P Val	Base	log$_2$FC	P Val
ENSDARG00000034300	sema3c	sema domain, immunoglobulin domain (Ig), (semaphorin) 3C	93	-1.69	0.000	207	-0.87	0.000
ENSDARG00000035187	abl1	c-abl oncogene 1, non-receptor tyrosine kinase	111	0.39	0.042	102	-0.53	0.019
ENSDARG00000036036	mdka	midkine a	83	-0.57	0.015	98	-0.45	0.045
ENSDARG00000038859	rgs20	regulator of G protein signaling 20	191	-0.60	0.010	202	0.60	0.032
ENSDARG00000040959	rabl3	RAB, member of RAS oncogene family-like 3	126	0.48	0.033	65	1.34	0.000
ENSDARG00000053370	eif3jb	eukaryotic translation initiation factor 3, subunit Jb	154	0.57	0.026	98	0.67	0.004
ENSDARG00000055792	foxo4	forkhead box O4	11	1.50	0.000	8	-0.89	0.047
ENSDARG00000056079	l3mbtl2	L3MBTL histone methyl-lysine binding protein 2	38	0.99	0.001	16	1.17	0.002
ENSDARG00000056590	calca	calcitonin/calcitonin-related polypeptide, alpha	31	-2.02	0.000	32	3.78	0.000
ENSDARG00000071727	si:dkey-37o8.1	si:dkey-37o8.1	196	-0.53	0.003	230	-0.34	0.041
ENSDARG00000074148	rbpjl	recombination signal binding protein for ig kappa J region	1073	0.53	0.001	587	1.81	0.000
ENSDARG00000075670	rereb	arginine-glutamic acid dipeptide (RE) repeats b	34	0.93	0.005	36	-1.26	0.000
ENSDARG00000095332	si:dkey-14d8.1	si:dkey-14d8.1	28	0.81	0.023	22	-0.95	0.010
ENSDARG00000100536	nkrf	NFKB repressing factor	109	0.57	0.029	83	0.46	0.029

Table 4. Differentially expressed (**adjusted p <0.05**) genes with selected Gene Ontology annotations at 4 and 7dpf in the severe mutants.

Gene ID	Name	Gene Description	4dpf Severe			7dpf Severe		
			Base	log$_2$FC	Adj P	Base	log$_2$FC	Adj P
ENSDARG00000005941	clul1	clusterin-like 1 (retinal)	30	-2.13	0.000	121	0.25	0.000
ENSDARG00000074148	**RAS** rbpjl	recombination signal binding protein for ig kappa J region-like	1073	0.53	0.020	587	0.23	0.000
ENSDARG00000038643	alas2	aminolevulinate, delta-, synthase 2	258	-1.62	0.028	733	0.41	0.020
ENSDARG00000034300	sema3c	sema domain, Ig, secreted, semaphorin 3C	93	-1.69	0.000	207	0.25	0.008
ENSDARG00000056079	l3mbtl2	L3MBTL histone methyl-lysine binding protein 2	38	0.99	0.018	16	0.38	0.026
ENSDARG00000056590	calca	calcitonin/calcitonin-related polypeptide, alpha	31	-2.02	0.000	32	0.48	0.000
ENSDARG00000074148	rbpjl	recombination signal binding protein for ig kappa J region-like	1073	0.53	0.020	587	0.23	0.000

3.1.2 Identifying novel transcripts for lncRNA analysis.

The ensembl annotated gene transcripts represented only a fraction of the total differentially expressed genes in the *gmppb* mutants. It is likely that within this subset of differentially expressed genes, previously unidentified lncRNAs exist. To identify potential lncRNAs, a previously described workflow was used to subset novel transcripts with low coding potential and low repetitiveness (Figure 8).

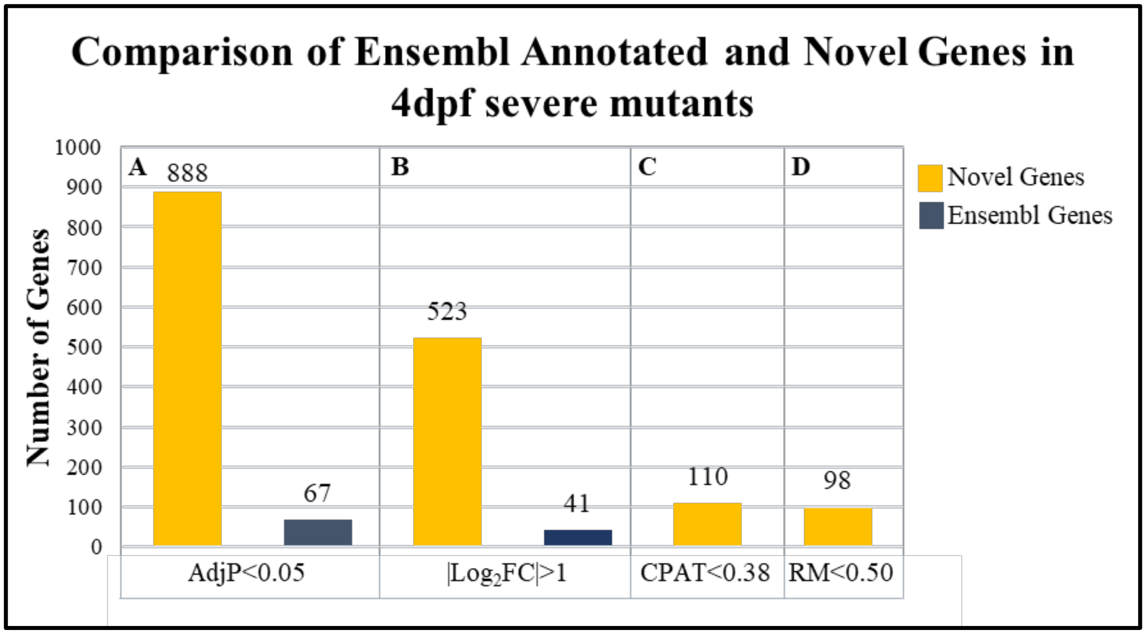
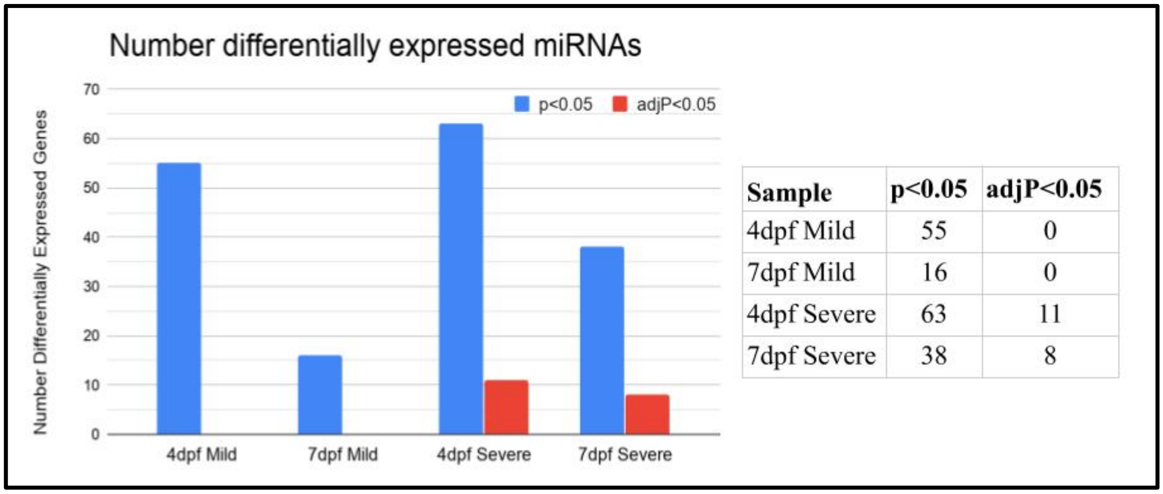


Figure 8. Identification of potential lncRNAs. A. Total genes (blue = ensembl annotated, yellow = unannotated) with adjusted p-values less than 0.05 in 4dpf severe mutants. B. Genes from panel A with $Log_2FC > 1$ or < -1. C. Novel genes from panel B with Coding Potential < 0.38 as determined by CPAT. D. Novel Genes from panel C with < 50% repetitive bases.

3.2 Small RNA sequencing of *gmppb* mutants to identify differentially expressed miRNAs

Since standard mRNA sequencing selects for poly-adenylated transcripts, and processed miRNA transcripts are not poly-adenylated, small RNA sequencing was performed to measure processed miRNA expression[74]. In the dataset, there were a total of 265 miRNAs expressed across all samples. Figure 9 shows the number of differentially expressed miRNAs in each sample; Figure 10 shows the overlap in differentially expressed miRNAs between samples.



Sample	p<0.05	adjP<0.05
4dpf Mild	55	0
7dpf Mild	16	0
4dpf Severe	63	11
7dpf Severe	38	8

Figure 9. Number of differentially expressed miRNA as indicated by adjusted p-value < 0.05 and unadjusted p-values < 0.05.

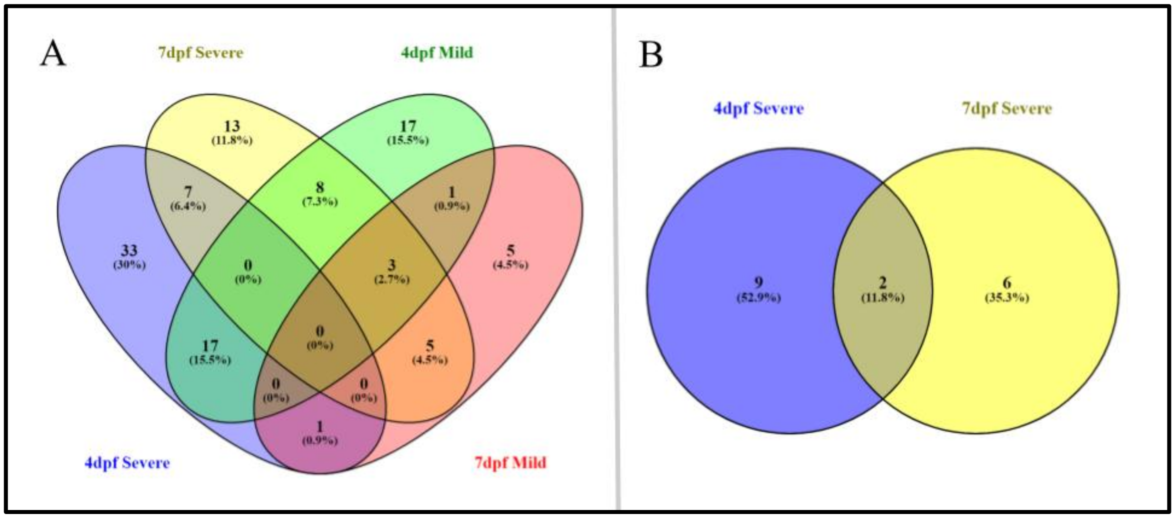


Figure 10. Overlap of differentially expressed miRNAs. A: miRNAs that are differentially expressed according to an unadjusted p-value < 0.05. B: miRNAs that are differentially expressed according to an adjusted p-value < 0.05. Venn diagrams made with Venny 2.0.

3.3 MiRNA co-expression networks

Co-expression networks can be used to identify sets of miRNAs that have correlated expression patterns that can be used to infer common function. The topology of these networks can be analyzed to determine nodes and edges that contribute towards network stability. Number of miRNA nodes in each of the three co-expression networks is listed in Table 5. Both characteristic path length and resultant network size are two parameters used to identify important nodes (Figures M1 and M2). MiRNAs that, upon removal, cause relatively large changes in characteristic path length are listed in Table 6.

Table 5. Number of nodes in miRNA co-expression networks.

Sample:	**Number of Nodes**
Sibling 4dpf and 7dpf	221
Mild 4dpf and 7dpf	255
Severe 4dpf and 7dpf	264

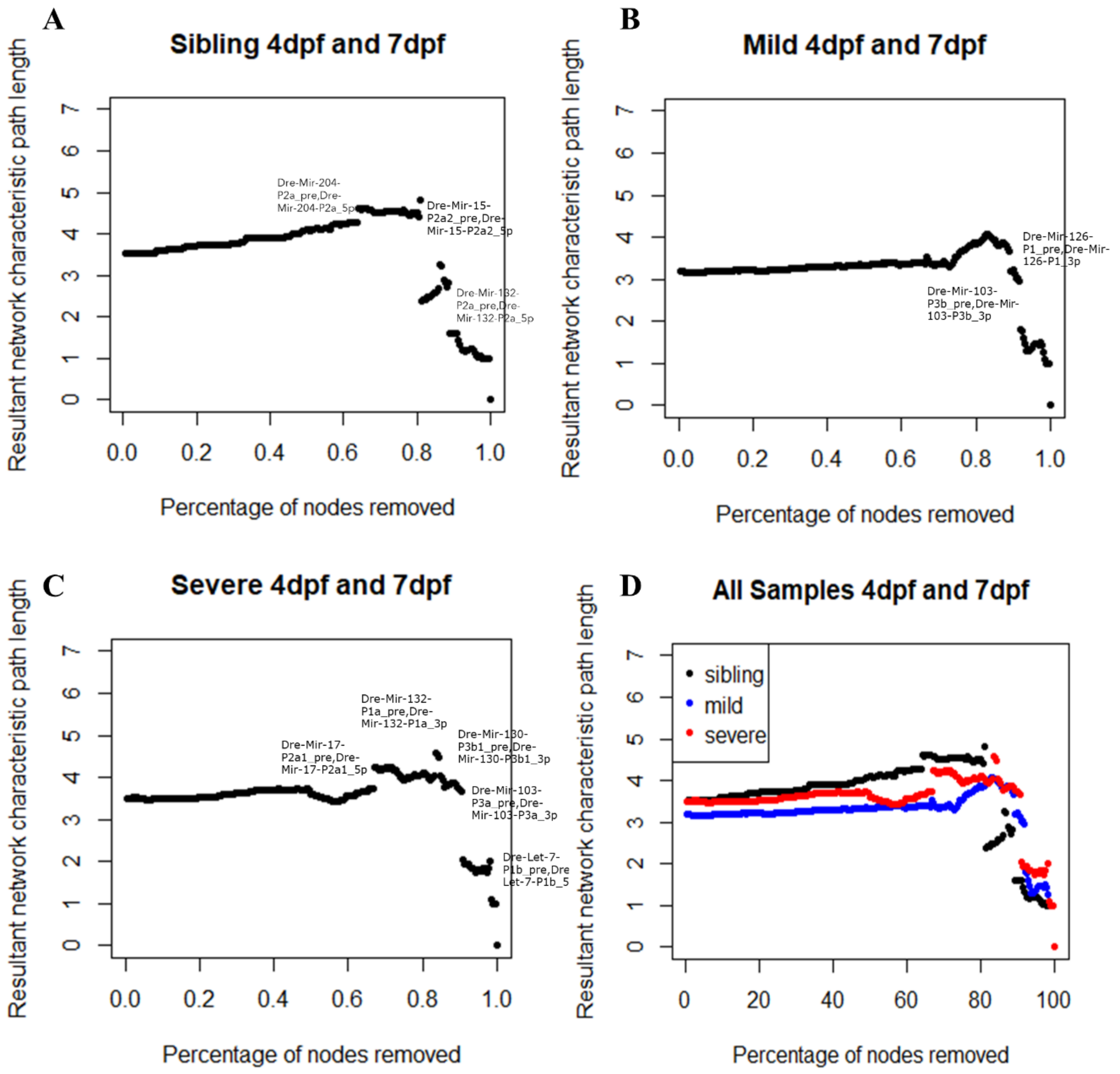


Figure 11. Characteristic path length upon removal of individual nodes in miRNA expression. miRNAs in each plot that upon removal cause relatively large changes in path length are labeled. Resultant characteristic path length upon removal of miRNA nodes in sibling 4dpf and 7dpf co-expression network (A), mild 4dpf and 7dpf co-expression network (B), and severe 4dpf and 7dpf co-expression network (C). D has all three networks overlaid into one graph.

Table 6. miRNAs that upon removal cause gaps in characteristic path length

4dpf 7 dpf Sibling	**4dpf 7dpf Mild**	**4dpf 7dpf Severe**
Dre-Mir-204-P2a-5p	Dre-Mir_103-P3b-3p	Dre-Mir-17-P2a1-5p
Dre-Mir-15-P2a2-5p	Dre-Mir-126-P1-3p	Dre-Mir-132-P1a-3p
Dre-Mir-132-p2a-5p		Dre-Mir-130-P3b1-3p
		Dre-Mir-103-P3a-3p
		Dre-Let-7-P1b-5

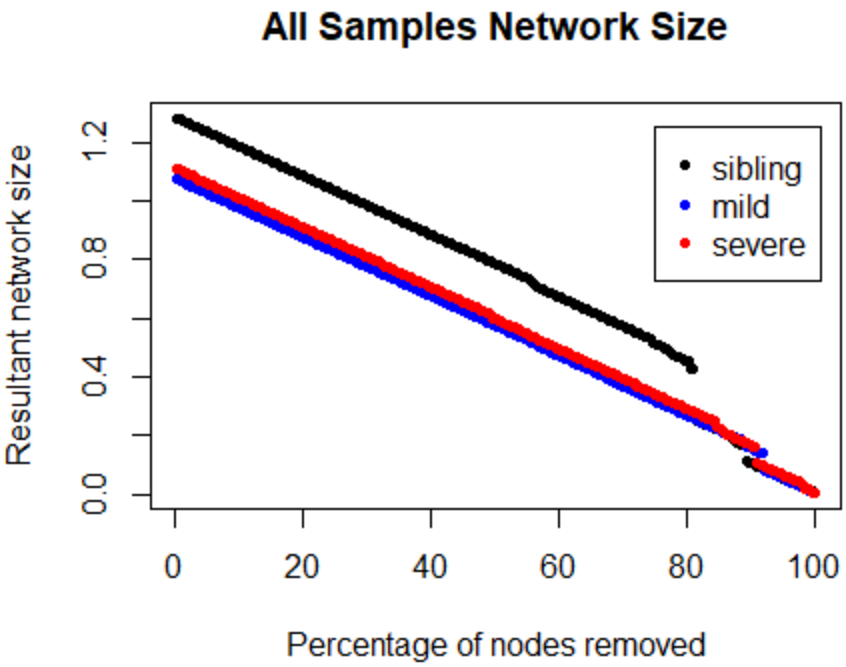


Figure 12. Network size upon removal of nodes from sibling (black), mild (blue), and severe (red) 4dpf and 7dpf co-expression networks.

3.4 miRNA target prediction at 4dpf in the *gmppb* severe mutants

TargetScanFish was used to predict mRNA targets of miRNAs, and analyses were performed to identify miRNA and mRNA targets with opposite expression at 4dpf in the severe mutants. Only the predominantly expressed miRNAs were included in this analysis. The first network with miRNAs upregulated and Ensembl annotated genes consists of a total of 8 upregulated miRNAs and 13 downregulated Ensembl genes, for a total of 37 interactions (Figure 13, Panel A). The second network has 7 downregulated miRNAs and 9 upregulated Ensembl genes for a total of 18 interactions (Figure 14, Panel B). Expression of each of the targeted genes is included in Tables 7 and 8.

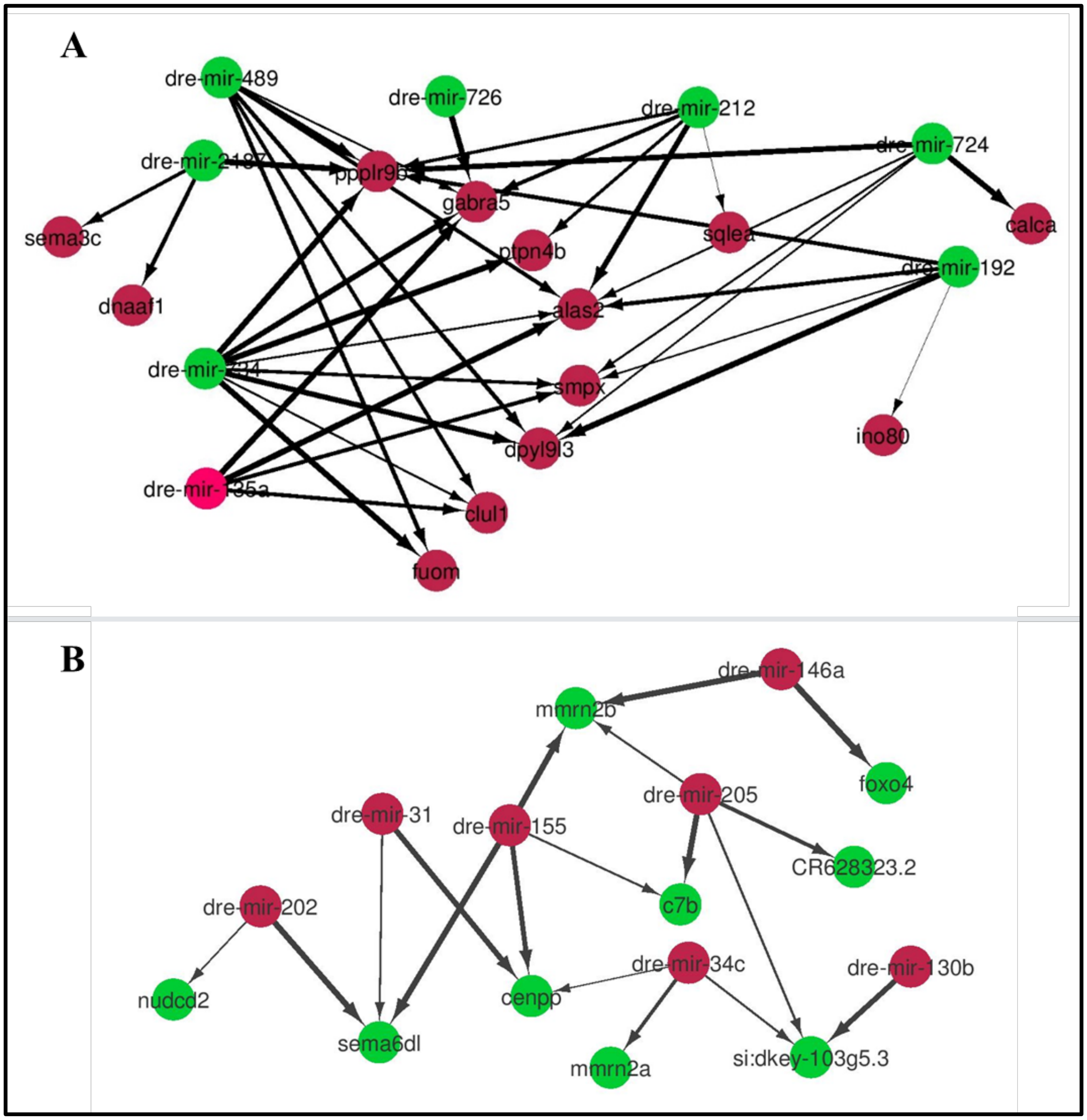


Figure 13. MiRNA and mRNA target interactions at 4dpf in *gmppb* severe mutants. A: Each of the green nodes is an upregulated miRNA (unadjusted p <0.05); each of the red nodes are downregulated ensembl annotated genes (adjusted p < 0.05). B: Green nodes are upregulated ensembl annotated genes, red nodes are downregulated miRNAs. Edges are denoted by arrows and show miRNAs targeting mRNAs based on TargetScanFish data. The width of the edges is based on the context score of the interaction between the miRNA and mRNA.

Table 7. Upregulated mRNA targets of downregulated miRNAs in 4dpf Severe mutants.

Gene ID	Name	Gene Description	4dpf Severe		
			Base	log$_2$FC	Adj P
ENSDARG00000011533	sema6dl	Sema, transmembrane, cytoplasmic domain semaphorin 6D like	124	4.99	5.3E-41
ENSDARG00000044135	cenpp	centromere protein P	134	0.56	4.0E-02
ENSDARG00000053062	CR628323.2	gap junction epsilon-1 protein-like	84	0.73	2.0E-02
ENSDARG00000055792	foxo4	forkhead box O4	11	1.50	8.3E-03
ENSDARG00000057121	c7b	complement component 7b	23	1.63	9.6E-03
ENSDARG00000062319	si:dkey-103g5.3	si:dkey-103g5.3	56	2.12	8.2E-08
ENSDARG00000073711	mmrn2b	multimerin 2b	261	0.82	4.8E-04
ENSDARG00000076135	mmrn2a	multimerin 2a	54	0.78	3.0E-02
ENSDARG00000078059	nudcd2	NudC domain containing 2	207	1.53	3.9E-10

Table 8. Downregulated mRNA targets of upregulated miRNAs in 4dpf Severe mutants.

Gene ID	Name	Gene Description	4dpf Severe		
			Base	$\log_2$FC	Adj P
ENSDARG00000000183	ptpn4b	protein tyrosine phosphatase non-receptor type 4b	461	-1.49	4.5E-07
ENSDARG00000005941	clul1	clusterin-like 1 (retinal)	30	-2.13	3.0E-04
ENSDARG00000012030	dnaaf1	dynein, axonemal, assembly factor 1	18	-1.21	1.3E-02
ENSDARG00000034300	sema3c	Sema and ig domain, secreted, (semaphorin) 3C	93	-1.69	2.2E-05
ENSDARG00000038643	alas2	aminolevulinate, delta-, synthase 2	258	-1.62	2.8E-02
ENSDARG00000039422	fuom	fucose mutarotase	30	-1.15	2.0E-02
ENSDARG00000045302	smpx	small muscle protein X-linked	124	-0.70	3.1E-02
ENSDARG00000056590	calca	calcitonin/calcitonin-related polypeptide, alpha	31	-2.02	3.1E-05
ENSDARG00000070432	ino80	INO80 complex ATPase subunit	459	-1.01	1.2E-02
ENSDARG00000070730	gabra5	gamma-aminobutyric acid (GABA) A receptor, alpha 5	41	-0.93	1.6E-02
ENSDARG00000079013	dpy19l3	dpy-19 like C-mannosyltransferase 3	313	-0.99	2.2E-05
ENSDARG00000079366	ppp1r9ba	protein phosphatase 1, regulatory subunit 9Ba	135	-0.71	5.9E-03
ENSDARG00000079946	sqlea	squalene epoxidase a	386	-0.71	9.2E-03

3.5 Splicing analysis of *gmppb* severe mutants

To produce a truncated *gmppb* protein product, a stop cassette was designed by the Henry lab for incorporation into the early 5' end of *gmppb*. The stop cassette had a total length of 75 nucleotides with homology domains on either end, each 20 nucleotides long; thus, the stop codons were from position 20-55. The left homology region was homologous to intron 2 of *gmppb* and the right homology region was homologous to the end of intron 2 and the beginning of exon 3 (Figure 14).

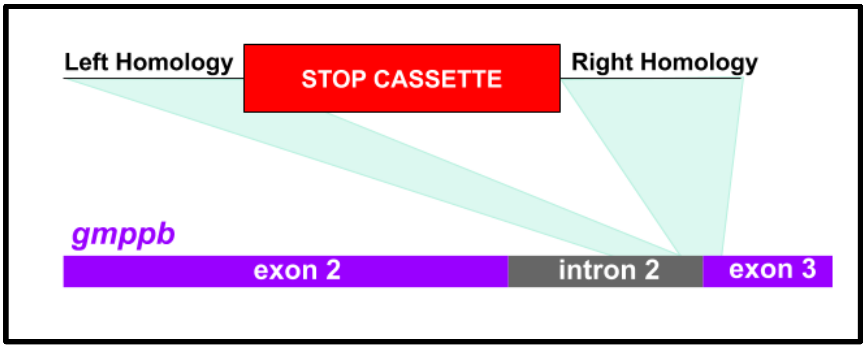


Figure 14. Stop cassette structure. The stop cassette consists of a left *gmppb* homology domain and a right *gmppb* homology domain, with the stop codon region containing stop codons in every reading frame. Shown in light blue is the where the regions of homology share sequence with *gmppb*.

To verify incorporation of the stop cassette into the *gmppb* mutants, PCR and sequencing of the gene *gmppb* was performed in 18 severe mutants. The mutants were categorized based on the sequencing results. Two categories of mutations were found, with one outlier. The first mutation, henceforth named the "partial double insertion", was present in 3 of the mutants. It consisted of insertion of one *full* stop region and one *partial* stop region. In between the two stop cassettes was a right homology domain, and the left homology region was truncated, with only 6 of the 20 original nucleotides (Figure 15). Notably, incorporation of this type of mutation resulted in a stop codon in every frame which would be expected to yield a truncated *gmppb* protein product[75].

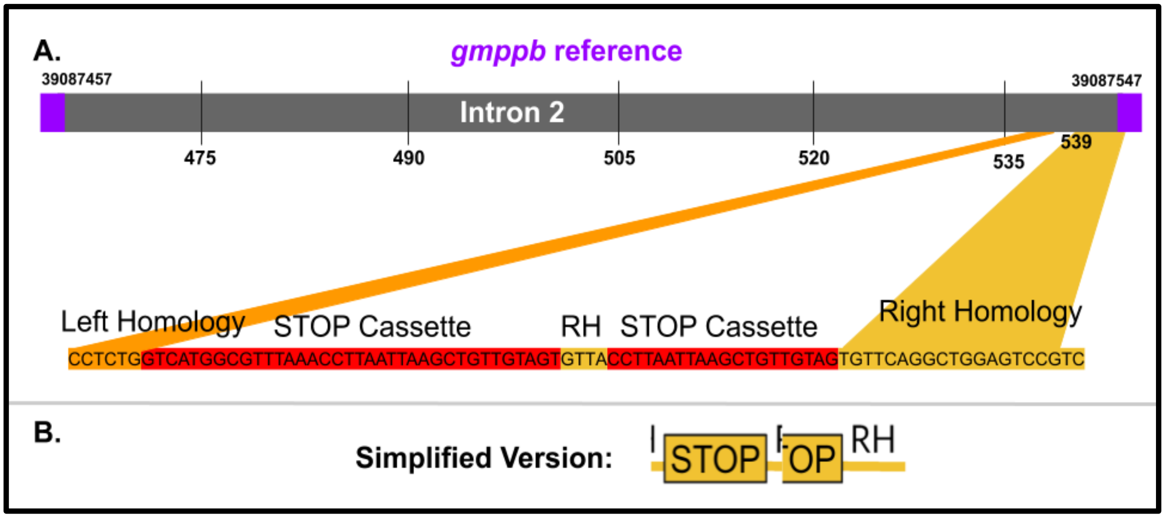


Figure 15. Partial double insertion mutation structure. A. In three of the 18 sequenced mutants, the stop cassette insertion consisted of incorporation of a partial stop cassette followed by a full stop cassette. B. Shown is a simple pictorial representation of the mutation with a subset of the left homology, a full stop cassette, a portion of the right homology, a partial stop cassette, and then the right homology region.

A second category of mutations was characterized by a “TG Gap” and was present in 14 of the 18 sequenced severe mutants (Figure 16). It was characterized by not only a lack of stop cassette insertion, but a truncated left homology region where 8 nucleotides were missing resulting in a truncated intron. ExpRESy indicated that stop codons were present in five of the six possible reading frames.

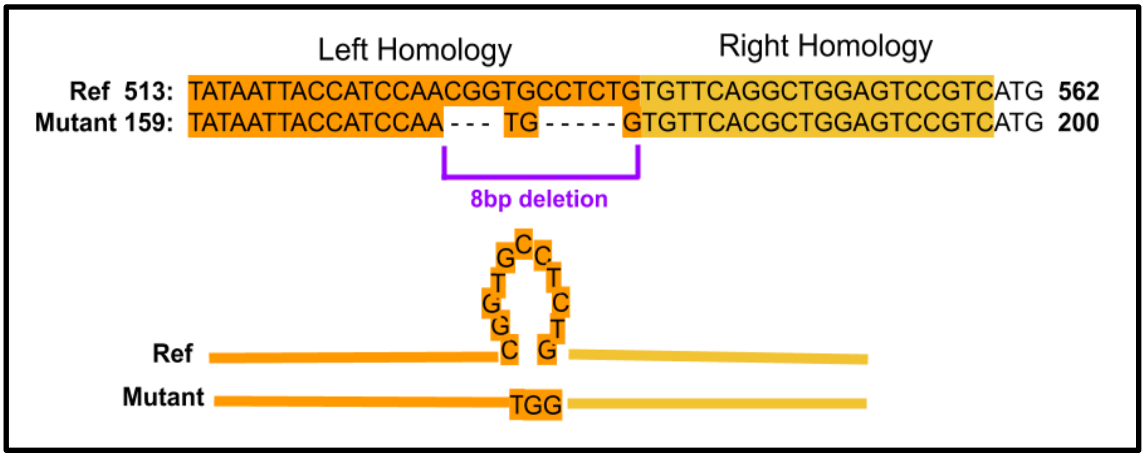

Left Homology
Right Homology
Ref 513: TATAATTACCATCCAACGGTGCCTCTGTGTTCAGGCTGGAGTCCGTCATG 562
Mutant 159: TATAATTACCATCCAA - - - TG - - - - - GTGTTCACGCTGGAGTCCGTCATG 200
8bp deletion
G C C T T G C G T C G
Ref
Mutant
TGG

Figure 16. “TG Gap” mutation consists of an 8 base pair deletion between the left and right homology domains of *gmppb*.

The findings from these analyses were next incorporated retroactively into the previously collected RNA Sequencing data by searching for specific subsequences of characteristic mutant regions in the FASTQ files. The queries “TAGTCTTACCTT” and “AAGGTAACACTA” were used to identify reads that represented the double stop cassette insertion mutation type in each of the samples (Figure QT) for both the forward and reverse reads. To verify that these were from the correct region of the genome (i.e. to confirm the were reads from *gmppb*), a secondary search was performed to determine the subset of these reads that contained six nucleotides of exonic *gmppb* that based on a read length of ~150nts should be present adjacent to the stop cassette (GGAGGC (e2) and GTCCGT (e3)) To find reads that represented the “TG Gap”, the subsequences “TGAACACCATTGGA” and “ACTTGTGGTAACCT” were searched for. In a similar fashion as previously described, the reads were subsequently searched for sequences from exon 2 and 3 of *gmppb* to verify that the reads were from the correct region (Figure 17).

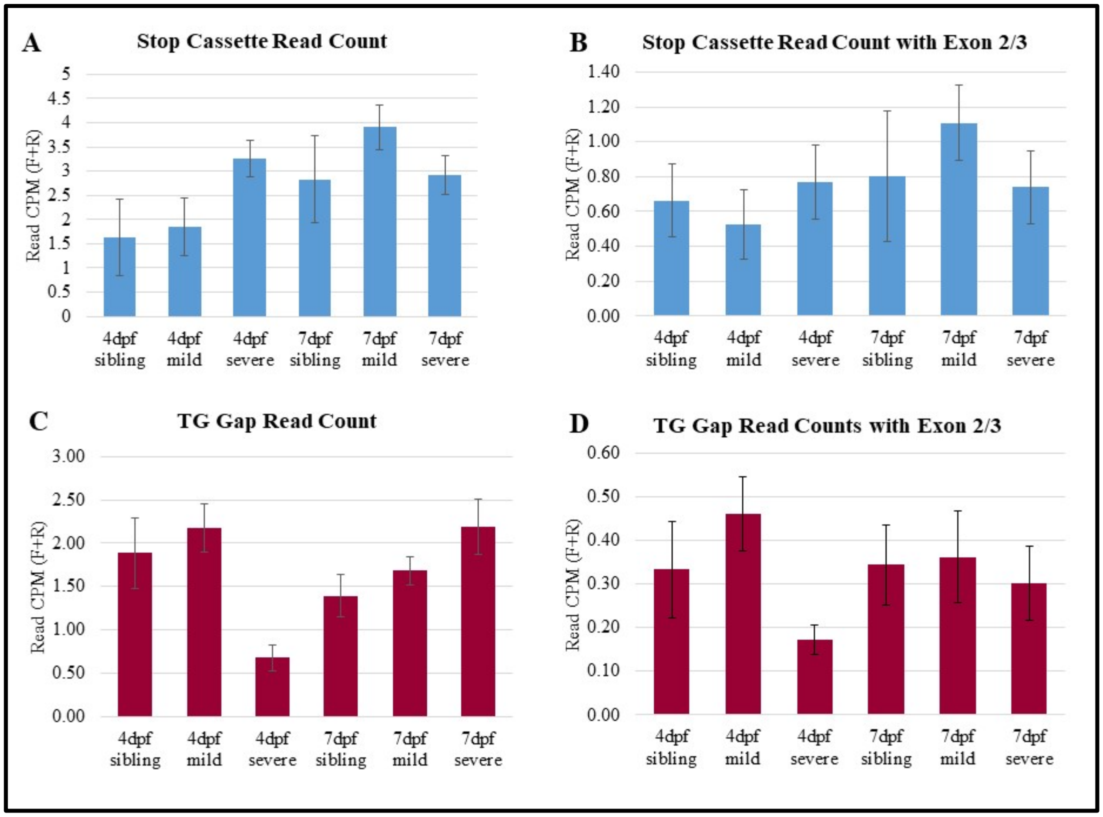


Figure 17. Characterization of reads from RNA Sequencing. The number of reads were expressed as counts per million (CPM). A: Results from searching for "TAGTGTTACCTT" and "AAGGTAACACTA", sequences representing four nucleotides of the right homology of the stop cassette (SC) and 4nts of the stop codons of the SC to either end. B: Results from searching for an additional sequence in the resulting reads from A, "GGAGGC" a sequence present in exon 2 of *gmppb* and "GTCCGT", a sequence present in exon 3 of *gmppb*. This step ensures that the reads come from *gmppb*. C: In a similar fashion to graph A, a sequence, this time representing the TG Gap mutation identified from the PCR of *gmppb* mutants, was searched for. D: Number reads with exons 2 and 3 subsequences.

4. DISCUSSION & FUTURE DIRECTIONS

4.1 mRNA Sequencing of *gmppb* mutants

4.1.1 Differences in number of differentially expressed genes confirms differences in *gmppb* mild and severe mutants.

Differentially expressed genes in the mild and severe mutants were identified using adjusted and non-adjusted p-values to identify genes that contributed to phenotypic severity. The shear difference in the number of differentially expressed genes correlates with the Henry's Lab method of distinguishing between the mild and severe mutants (Figure 6). According to the unadjusted p-values, there are ~6 times more differentially expressed genes in the severe mutants as opposed to the mild mutants at 7dpf; according to the adjusted p-value, there are over 100 times more differentially expressed genes. Another interesting trend is that there are ~6 times more differentially expressed genes in the severe mutants at 7dpf as compared to the severe mutants at 4dpf. This could be due to compounding effects from irregulated pathways at early time points, leading to cellular responses aiming to restore the mutants to "normal" or "healthy" conditions. For example, multiple pathways must be activated to restore the tissue damage, and likely many of these are immune mediators. To better understand the gene pathways that contributed towards the severe phenotype, genes differentially expressed in the severe 4dpf and 7dpf mutants, but not the mild 4dpf and 7dpf mutants, were identified and further characterized.

4.1.2 Gene Ontology characterization reveals phenotype-relevant genes.

Gene Ontology annotations for the genes differentially expressed in *only* the 4dpf and 7dpf severe mutants were determined using PANTHER and DAVID. These databases use manually curated and electronic annotations of Gene Ontology terms to characterize genes from the zebrafish model organism database, ZFIN. When examining functional annotation of genes, it is important to recognize that the zebrafish research community is smaller than the mouse and human research communities and the number of annotations are thereby smaller. Our entry list of 82 zebrafish genes resulted in a list of 47 genes with annotations. PANTHER and DAVID list Gene Ontology annotations for the genes that provide information about the biological processes, molecular function and cellular components associated with the protein products of the genes. Because of the relatively small number of input genes (82 for unadjusted p-values and 13 for adjusted p-values), the only term that was enriched was "nucleus" with a p-value of 0.005. A list of MD relevant terms that appeared in the PANTHER and DAVID annotations was selected and genes were highlighted for further characterized based on these annotations. Terms included muscle organization, extracellular matrix associated, cellular adhesion, immune function, mitochondrial function, transcriptional regulators, and development pathways.

The immune response to skeletal muscle disorganization and destruction is an important component of Muscular Dystrophy. Typically, following necrosis, the cellular debris is removed by macrophages and muscle satellite cells (MuSCs) migrate to the site of injury and proliferate to replace the lost tissue and restore muscle function. However, in most forms of MD, the MuSCs are unable to properly restore muscle function, and instead fibrotic tissue is deposited. The molecular mechanisms that modulate the immune response are best characterized in Duchenne's Muscular Dystrophy (DMD)[76]. In DMD, dystrophic muscle is invaded by CD4+ and CD8+ T cells, macrophages, eosinophils, and natural killer cells[77,78]. Depletion of myeloid and lymphoid populations decreases myonecrosis [79–81]; depletion of B and T cells reduced fibrosis and TGF-β1(15), and ablation of inflammatory signals like IFNγ reduced the severity of muscle pathology[82]. Immunosuppressive medications, like the glucocorticoid prednisone, has also been shown to improve muscle strength in DMD patients and decrease myofiber injury[83]. However, the immune response in MD is not strictly deleterious. For example, M2 macrophages are induced by Il-4 and IL-13 to inhibit damaging inflammation and M1 macrophage mediated cytotoxicity[84]. Thus, the immune response is an important and delicate component of the pathology of MD and its role in dystroglycanopathy warrants further research.

Multiple immune relevant genes were differentially expressed in the severe mutants. Of the 81 genes differentially expressed in both the 4dpf and 7dpf severe and not the mild mutants, only one was related to the immune system: matrix metalloproteinase-9 (*mmp9*). MMPs have been shown to play an important role in myofiber functionality and skeletal muscle cell migration, differentiation, and regeneration. One way they do so is by degrading the extracellular matrix to allow for MuSC migration and differentiation to replace lost tissue. Inhibition of MMPs suppresses migration of MuSCs to the site of injury and impedes regeneration[85,86]. MMP-9 upregulation has been shown as a clinical biomarker for Duchenne's Muscular dystrophy[87]. In the severe mutants, at 4dpf, *mmp9* was upregulated, suggesting that the mutants were responding to the muscle damage by promoting satellite cell migration and differentiation, but at 7dpf, it was downregulated, suggesting a lack of response to damage (Table 3).

Another interesting candidate that was differentially expressed in the 4dpf and 7dpf with an adjusted p-value < 0.05 was clusterin-like-1 (*clul1*). The clusterin protein is a molecular chaperone that inhibits apoptosis through stabilization of the Ku70-Bax protein complex[88]. In the 4dpf severe mutants, *clul1* was significantly downregulated, corresponding to an increase in apoptosis according to the aforementioned pathway (Table 4). This could be a possible mechanism that results in myocyte death. However, apoptosis is often a response, a last resort to other cellular damage – thus, we suggest that this is a downstream effect of cellular damage in the MD mutants.

Genes were also ranked based on the ratio of expression as indicated by the fold change (Log_2 FC). The top 10 largest fold changes in expression were all in the severe mutants, eight of these were from the 7dpf severe mutants and two were from the 4dpf severe mutants. This is roughly the proportion we would expect based on the genes in each category that were differentially expressed according to the adjusted p-value (67 genes in the 4dpf severe and 338 genes in the 7dpf severe). Of the 10 most differentially expressed genes, functions were related to cell growth, nervous system development, and muscle contraction and relaxation.

Antomacin 9a (*ano9a*) is part of the antomacin family which encodes calcium-chloride channels and it was differentially expressed in the 7dpf severe *gmppb* mutants. There is little research characterizing *ano9*, however another member of this family is implicated in MD. Antomacin 5 (*ANO5*) mutations are one of the causes of Limb-Girdle MD. The protein product of this gene encodes a calcium-activated chloride channel that is most abundant in the endoplasmic reticulum (ER) of skeletal muscle[89]. It is predicted to be involved in regulating muscle contraction and relaxation. ANO5 also maintains calcium homeostasis which promotes plasma membrane repair in damaged myofibrils. Upon deletion of *ANO5* or pharmacological inhibition of injury-triggered calcium flow form the ER to the cytoplasm, enabled injured patient myocytes to repair[90]. Although the specific role of *ano9* is unknown, based on the role of *ANO5*, it is an interesting candidate for future research.

Cell division cycle 23 (*cdc23*) was upregulated in the 7dpf severe mutants. *Cdc23* is a mitotic regulator that allows for cell cycle progression through the anaphase promoting complex (APC). We predict that the upregulation of this gene is likely related to an upregulation in cell division in an attempt to replace the lost myocytes. Interestingly, *cdc23* was the only cell division cycle gene that was upregulated in the 4dpf or 7dpf mutants.
Finally, the sema family proteins were differentially expressed. *Sema6dl* was one of the genes with the largest fold change. It was upregulated with a Log_2FC of nearly 5.0. *Sema3* was downregulated in both the 4 and 7dpf severe mutants. The sema receptors have been implicated in multiple signaling pathways with roles in regulating innervation. Specifically, *sema3* has been shown to be involved in the innervation of skeletal muscle in the diaphragm[91] and *sema2* loss-of-function mutants have ectopic innervation in the muscle[92]. The loss of motor control in MD may be related to a lack of reinnervation or improper reinnervation after myocyte damage. In MD in general, the Neuronal Muscular Junctions (NMJ), the peripheral synapses that induce muscle contraction, are disorganized – a phenotype shared in the *gmppb* mutants.

Interestingly, two of the most differentially expressed genes were not protein coding – one was an antisense transcript and the other was a processed transcript, in the future we would like to look at these genes in more detail, as they could be unannotated lncRNAs.

Of course, there are many different lenses to view the gene expression data and identify candidates for future research. Another approach would be looking at genes that are differentially expressed in both the mild and severe mutants - but have a higher proportional change in expression in the severe as compared to the mild as indicated by the Log_2 FC.

Of the 881 annotated differentially expressed ensembl genes, 22 were annotated as lncRNAs and two as miRNA precursors. In the past, we have found that the Ensembl annotated lncRNAs tend to be highly repetitive – often exceeding 80% repetitive bases. Thus, we would like to analyze these lncRNAs further and search for any functional annotations in the literature in the future.

4.1.4 Identifying novel transcripts for lncRNA analysis.

To identify potential lncRNAs, novel transcripts were identified with low coding potential and less than 50% of repetitive bases according to a previously described lncRNA annotation workflow[93]. One thing that stands out in the 4dpf severe *gmppb* mutants is the high proportion of differentially expressed unannotated, novel genes as compared to the differentially expressed ensembl-annotated genes. In the past, this might have been considered "transcriptional noise"[94], but difficulty still persists in identifying transcripts that are most likely to be related to the phenotype – and doing so requires extensive manual analysis. To characterize these 98 potential lncRNAs, ORFFinder will be used to look for open reading frames and BLAST will be used to identify transcripts with alignment to known protein coding genes. Furthermore, these transcripts will be aligned to novel lncRNAs that were differentially expressed in caudal and cardiac regeneration in the zebrafish to look for commonly differentially expressed transcripts. Additionally, the genes to either side of the potential lncRNAs will be determined to indicate possible regulatory targets since some lncRNAs act in a "cis" mechanism. Homologous lncRNAs in humans will be identified. These analyses will be combined to generate a list of lncRNAs for qPCR validation.

4.2 Small RNA sequencing of *gmppb* mutants to identify differentially expressed miRNAs

To identify differentially expressed miRNAs, regulators of mRNA degradation, small RNA Sequencing was performed. The number of differentially expressed miRNAs according to an unadjusted p-value < 0.05 is similar in the 4dpf mild (55) and severe mutants (63) (Figure 9). However, when you consider the differentially expressed miRNAs according to the adjusted p-value, there were no differentially expressed miRNAs in the mild mutants. We suspect that there were not any miRNAs with an adjusted p-value < 0.05 because we had only three biological replicates per sample group. Additional biological replicates would need to be characterized in a future experiment to more accurately characterize the biological variation in these sample groups. Since we only have three biological replicates, an unadjusted p-value can be investigated further to reveal differences in the phenotypic differences in the mild and severe mutants. In the analysis of the miRNAs using unadjusted p-values, another trend was a lower number of differentially expressed miRNAs in the 7dpf samples than in the 4dpf samples. This is contrasting to the trend observed in the Ensembl annotated genes where there were more differentially expressed genes in the 7dpf samples than in the 4dpf samples. Again, we suspect that this is a result of only 3 biological replicates characterized by small RNA Sequencing compared to 4 biological replicates for standard RNA sequencing.

4.3 MiRNA co-expression networks reveal miRNAs that contribute towards network structural integrity.

Three gene co-expression networks were produced using 4dpf and 7dpf siblings, 4dpf and 7dpf mild *gmppb* mutants, and 4dpf and 7dpf severe *gmppb* mutants. Because we filtered out lowly expressed miRNAs, the number of co-expressed miRNAs (nodes) in each network was different. The number of nodes was lowest in the sibling network, intermediate in the mild network, and highest in the severe network (Table 5). The sibling showed the longest overall characteristic path length, the severe showed an intermediate characteristic path length, and the mild displayed the shortest characteristic path length (Figure M1). Network node attack analysis revealed multiple miRNAs that resulted in large changes in the characteristic path length after they were removed from the network (Table 6). These miRNAs are candidates for further investigation. A shorter characteristic path length implies biological networks working in conjunction and synergy[74]. Perhaps, the additional stressor (the *gmppb* mutation), causes coordinated changes in miRNA expression in an attempt to combat the dysregulated pathways and processes. It is possible that the severe *gmppb* mutants have an intermediate characteristic path length because of an inability to respond as effectively to the physiological changes induced by the *gmppb* mutation. However, using this type of analysis in gene networks is a novel, thus, we need to perform further research to determine how changes in network topology correlate with genetic function.

4.4 MiRNA target prediction at 4dpf in the *gmppb* severe mutants reveals an overlapping network of miRNAs regulating a cohort of Ensembl annotated genes

By determining differentially expressed miRNAs with differentially expressed Ensembl annotated mRNA targets with opposite expression, miRNA/mRNA interactions can be predicted. In the 4dpf samples, a total of 55 such interactions were identified. The gene targets of these interactions were further characterized in relation to functions and processes implicated in the pathological progression of MD.

Four of the mRNA targets were previously identified based on the 10 most differentially expressed analysis (*calca*) and differential expression (adj-p < 0.05) in the severe mutants at 4dpf and 7dpf (*clu1 (yes1)*, *alas2*, and *sema3c*). This analysis therefore identified potential upstream regulators of potentially MD-relevant genes and pathways.

To identify mRNA/miRNA interactions of interest, Gene Ontology annotations of the targeted mRNAs were analyzed, and a literature search was performed to identify mRNAs with MD-relevant function. From this analysis, 8 genes of interest were identified with functions related to cell growth regulation, immune activation, and neuron transmission and function.

Cenpp and *foxo4* are modulators of cell growth. *Cenpp* is a subunit of the CENPI-associated centromeric complex and is required for kinetochore function and miotic progression[96]. In the severe 4dpf mutants, *cenpp* was upregulated - suggesting an increase in cell division – likely in an attempt to replace the damaged tissue. Multiple downregulated miRNAs are predicted to target *cenpp,* including dre-mir-31, dre-mir-155, and dre-mir-34c. Forkhead box O transcription factors (foxo) are involved in cellular proliferation, stress resistance, and apoptosis[97]. Overexpression of FOXO3 is linked to skeletal muscle atrophy through induction of atorign-1, a ubiquitin ligase[98]. *Foxo4* is not well characterized, but based on its inclusion in the foxo family, it could play a similar role in inducing musclar atrophy. FOXO3 was significantly downregulated in the 4dpf severe mutants – suggesting an increase in muscle atrophy. This gene was predicted to be targeted by dre-mir-146a.

The protein complement 7b (*c7b*) is a component of the complement pathway of the innate immune response that recruits the MAC attack complex to induce apoptosis of target cells. *C7b* was upregulated, suggesting an increase in the innate immune response in 4dpf severe mutants that might contribute towards the loss of muscle mass in the mutants. *C7b* was predicted to be targeted by dre-mir-205 and dre-mir-155.

The genes Protein Tyrosine Phosphatase Non-Receptor Type 4 (*ptpn4b*), Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha5 (*gabra5*), and Protein phosphatase 1, regulatory subunit 9Ba (*ppp1r9ba*) are involved in neuronal function and protection. In the gmppb mutants, the neural muscular junctions (NMJs) are improperly formed and are likely related to loss of motor control in the mutant zebrafish (Figure 3). *Ptpn4b* has been shown to be involved in neural circuit formation in the brain of drosophila as it aids in establishing and stabilizing axonal projection patterns[99]. Thus, *ptpn4* inhibition causes an increase in neuronal apoptosis[100]. In the 4dpf *gmppb* severe mutants at 4dpf severe, *ptpn4b* was downregulated – suggesting a decrease in this process. *Gabra5* encodes a subunit of the GABA receptor, which is a ligand-gated chloride channel found in the brain. This receptor's ligand is GABA, an inhibitory neurotransmitter. Another symptom related to MD is epilepsy which occurs in an estimated 6-7% of MD pediatric patients, as compared to 0.5-1% in the general population[101,102]. Thus, this gene's decreased expression in the

4dpf severe mutants could be related to the loss of motor control *inhibition* in MD patients. *Ppp1r9b* is a factor involved in promoting the formation of filopodia outgrows that can be further remodeled to form dendritic spines that allow for excitation of neurons in the brain[103]. *Ppp1r9b* was downregulated in the mutants, suggesting a decrease in formation of filopodia outgrowths and a decrease in neural development and function. *Ptpn4b*, *gabra5*, and *ppp1r9ba* were each predicted to be targeted by dre-mir-734. *Ptpn4b* and *gabra5* were predicted to be targeted by dre-mir-212, dre-mir-135, and dre-mir-489; *Ptpn4b* was additionally predicted to be targeted by dre-mir-2187 and dre-mir-192.

Overall, this analysis was able to identify upstream regulators of protein-coding genes with functions that appear to be related to the pathology of MD. To understand the role of miRNAs in targeting mRNAs more thoroughly, these same interactions should be identified in the other samples: 7dpf severe mutants, 4dpf mild mutants, and 7dpf mutants, and the results should be compared.

4.5 Splicing analysis: *gmppb* severe mutants are characterized by two categories of mutations

The splicing analysis revealed multiple types of mutations present in the severe mutants, emphasizing the need to untangle how this might contribute towards the phenotype severity. Interestingly, both types of mutations apparently resulted in the same general phenotype, since all eighteen of the sequenced fish has severe phenotypes. This is perhaps based on the observation that even with the “TG” gap mutation type, a substantial deletion and frameshift mutation is induced that results in stop codons in five of the six reading frames. However, for this to have any major impact on the length of the resultant protein, the intronic region must be translated since the mutation occurs in the 3’ region of the intron. Perhaps, the mutation is disrupting the normal splicing of the mRNA which leads to inclusion of the intron either resulting in a truncated protein via the included stop codon or a dysfunctional protein based on the inclusion of the intron. Spliceosomes are small nuclear ribonucleoproteins that carry out splicing. The vast majority of eukaryotic introns are U2-type introns which are marked by a “GT” at the 5’ end and a “AG” at the 3’ end[104]. However, since the STOP cassette sequence contains multiple “GT” and “AT” dinucleotides, predicting exactly how the mutations affect the splicing is difficult and would be more accurately determined via functional studies, such as determining the structure of the resulting protein.

Moreover, the presence of the different types of mutations in the fish is curious. Why did some of them receive the double insertion while others have the “TG” gap? Some of this might best be explained by scientific variability, whether it be in the fish themselves or the handling of them. In respect to the “TG” gap, it is possible that homology directed repair just never happened. After the cut was induced via the CRISPR/Cas9 system, perhaps the stop cassette oligonucleotide was not proximal to the site so cut sequence was eventually ligated back together after a few nucleotides were chewed off on either end.

Finally, of course when using a CRISPR/Cas9 system, the guide RNA must be carefully and precisely designed to ensure the lowest possible chance of off-target affects. ChopChop[105], the tool that was used to generate the oligonucleotide sequences uses an internal algorithm that rates the specificity of the sgRNA. Furthermore, to verify this, we aligned the sgRNA sequence “GGACTCCAGCCTGAACACAG” against the GRCz11 zebrafish assembly using BLAT and only found one match – suggesting proper design of the sgRNA. Thus, off-target affects are minimal.

After all this, the characteristic mutation search results were our attempt to incorporate our findings from the PCR into the previously collected RNA Sequencing data. The goal was to determine if mutation types could explain the variability in phenotype in the fish. What we found is that there only seems to be a small association with the stop cassette mutation type in the 4dpf severe mutants. They appear to have a higher percentage of “double partial insertions” and a lower percentage of “TG” gaps – suggesting that perhaps the increase in severity (as compared to the mild mutants) is partially explained by the different types of mutations. However, after searching for a short fragment of exon 2 or 3 in the reads matching the mutation types, this trend seems to disappear, making us question whether or not the reads are from *gmppb*, or if they are in fact reads from a different region of the genome. In order to investigate this, a more robust method of validating the location of the reads would need to be used, likely with a method other than grep. Furthermore, to really understand the effect of the DNA changes on the processed mRNA transcript and subsequent translated amino acid sequence, a tool that can identify intron/exon boundaries of zebrafish transcripts is needed. This would allow for prediction of protein products which could be confirmed via protein isolation from mutants and amino acid identification.

BIBLIOGRAPHY

1. Duchenne Muscular Dystrophy - NORD (National Organization for Rare Disorders). Available at: https://rarediseases.org/rare-diseases/duchenne-muscular-dystrophy/. (Accessed: 16th January 2020)

2. Astrea, G. *et al.* Broad phenotypic spectrum and genotype-phenotype correlations in GMPPB-related dystroglycanopathies: an Italian cross-sectional study. *Orphanet J. Rare Dis.* **13**, 170 (2018).

3. Durbeej, M. *et al.* Disruption of the β-sarcoglycan gene reveals pathogenetic complexity of limb-girdle muscular dystrophy type 2E. *Mol. Cell* **5**, 141–151 (2000).

4. Den Dunnen, J. T. *et al.* Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. *Am. J. Hum. Genet.* **45**, 835–47 (1989).

5. Hu, X. Y. *et al.* Partial gene duplication in Duchenne and Becker muscular dystrophies. *J. Med. Genet.* **25**, 369–76 (1988).

6. Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509–17 (1987).

7. Ibraghimov-Beskrovnaya, O. *et al.* Primary structure of dystrophin-associated glycoproteins linking dystrophin to the extracellular matrix. *Nature* **355**, 696–702 (1992).

8. Langenbach, K. J. & Rando, T. A. Inhibition of dystroglycan binding to laminin disrupts the PI3K/AKT pathway and survival signaling in muscle cells. *Muscle Nerve* **26**, 644–653 (2002).

9. Moresi, V., Adamo, S. & Berghella, L. The JAK/STAT pathway in skeletal muscle pathophysiology. *Frontiers in Physiology* **10**, (2019).

10. Bouchet-Séraphin, C., Vuillaumier-Barrot, S. & Seta, N. Dystroglycanopathies: About Numerous Genes Involved in Glycosylation of One Single Glycoprotein. *J. Neuromuscul. Dis.* **2**, 27–38 (2015).

11. Structure of Skeletal Muscle | SEER Training. Available at: https://training.seer.cancer.gov/anatomy/muscular/structure.html. (Accessed: 20th January 2020)

12. 10.2 Skeletal Muscle – Anatomy and Physiology. Available at: https://opentextbc.ca/anatomyandphysiology/chapter/10-2-skeletal-muscle/. (Accessed: 20th January 2020)

13. 10.3 Muscle Fiber Contraction and Relaxation – Anatomy and Physiology. Available at: https://opentextbc.ca/anatomyandphysiology/chapter/10-3-muscle-fiber-contraction-and-relaxation/. (Accessed: 20th January 2020)

14. 10.6 Exercise and Muscle Performance – Anatomy and Physiology. Available at: https://opentextbc.ca/anatomyandphysiology/chapter/10-6-exercise-and-muscle-performance/. (Accessed: 20th January 2020)

15. MAURO, A. Satellite cell of skeletal muscle fibers. *J. Biophys. Biochem. Cytol.* **9**, 493–495 (1961).

16. Tedesco, F. S., Dellavalle, A., Diaz-Manera, J., Messina, G. & Cossu, G. Repairing skeletal muscle: Regenerative potential of skeletal muscle stem cells. *Journal of Clinical Investigation* **120**, 11–19 (2010).

17. Conboy, I. M. & Rando, T. A. The regulation of Notch signaling controls satellite cell activation and cell fate determination in postnatal myogenesis. *Dev. Cell* **3**, 397–409 (2002).

18. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

19. Devoto, S. H., Melançon, E., Eisen, J. S. & Westerfield, M. Identification of separate slow and fast muscle precursor cells in vivo, prior to somite formation. *Development* **122**, 3371–3380 (1996).

20. Berger, J. & Currie, P. D. Zebrafish models flex their muscles to shed light on muscular dystrophies. *DMM Disease Models and Mechanisms* **5**, 726–732 (2012).

21. Mazelet, L., Parker, M. O., Li, M., Arner, A. & Ashworth, R. Role of Active Contraction and Tropomodulins in Regulating Actin Filament Length and Sarcomere Structure in Developing Zebrafish Skeletal Muscle. *Front. Physiol.* **7**, (2016).

22. Kanagawa, M. *et al.* Molecular recognition by LARGE is essential for expression of functional dystroglycan. *Cell* **117**, 953–964 (2004).

23. Willer, T. *et al. Targeted disruption of the Walker-Warburg syndrome gene Pomt1 in mouse results in embryonic lethality*. (2004).

24. Hara, Y. *et al.* A Dystroglycan Mutation Associated with Limb-Girdle Muscular Dystrophy. *N. Engl. J. Med.* **364**, 939–946 (2011).

25. Roscioli, T. *et al.* Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of α-dystroglycan. *Nat. Genet.* **44**, 581–585 (2012).

26. Carss, K. J. *et al.* Mutations in GDP-mannose pyrophosphorylase B cause congenital and limb-girdle muscular dystrophies associated with hypoglycosylation of α-dystroglycan. *Am. J. Hum. Genet.* **93**, 29–41 (2013).

27. Belaya, K. *et al.* Editor's Choice: Mutations in GMPPB cause congenital myasthenic syndrome and bridge myasthenic disorders with dystroglycanopathies. *Brain* **138**, 2493 (2015).

28. Cabrera-Serrano, M. *et al.* Expanding the phenotype of GMPPB mutations. *Brain* **138**, 836–844 (2015).

29. Gagnon, J. A. *et al.* Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs. *PLoS One* **9**, e98186 (2014).

30. Digitalcommons@umaine, D. & Bailey, E. *Neuromuscular Development and Phenotypic Variation in Neuromuscular Development and Phenotypic Variation in Zebrafish Models of Dystroglycanopathy Zebrafish Models of Dystroglycanopathy*. (2019).

31. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172**, 393–407 (2018).

32. Somarowthu, S. *et al.* HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell* **58**, 353–361 (2015).

33. Rossignol, F., Vaché, C. & Clottes, E. Natural antisense transcripts of hypoxia-inducible factor 1alpha are detected in different normal and tumour human tissues. *Gene* **299**, 135–140 (2002).

34. Derrien, T., Guigó, R. & Johnson, R. The Long Non-Coding RNAs: A New (P)layer in the &quot;Dark Matter&quot;. *Front. Genet.* **2**, 107 (2011).

35. Neguembor, M. V., Jothi, M. & Gabellini, D. Long noncoding RNAs, emerging players in muscle differentiation and disease. *Skelet. Muscle* **4**, 8 (2014).

36. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).

37. Shukla, G. C., Singh, J. & Barik, S. MicroRNAs: Processing, maturation, target recognition and regulatory functions. *Mol. Cell. Pharmacol.* **3**, 83–92 (2011).

38. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology* **9**, (2018).

39. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics* **16**, 421–433 (2015).

40. Shibasaki, H. *et al.* Characterization of a novel microRNA, miR-188, elevated in serum of muscular dystrophy dog model. *PLoS One* **14**, e0211597 (2019).

41. Raz, T. *et al.* Protocol Dependence of Sequencing-Based Gene Expression Measurements. *PLoS One* **6**, e19287 (2011).

42. KAPA RNA HyperPrep Kits with RiboErase - Roche Sequencing Solutions. Available at: https://sequencing.roche.com/en-us/products-solutions/by-category/library-preparation/rna-library-preparation/kapa-rna-hyperprep-kit-with-riboerasehmr.html. (Accessed: 21st January 2020)

43. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. (2010).

44. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**,

907–915 (2019).

45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

46. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).

47. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

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

49. Qiagen. QIAseq miRNA Library Kit Handbook available online at https://www.qiagen.com/us/products/discovery-and-translational-research/next-generation-sequencing/rna-sequencing/mirna-small-rnaseq/qiaseq-mirna-ngs/. in (2020).

50. Holme, P., Kim, B. J., Yoon, C. N. & Han, S. K. Attack vulnerability of complex networks. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **65**, 14 (2002).

51. Berlusconi, G., Calderoni, F., Parolini, N., Verani, M. & Piccardi, C. Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLoS One* **11**, (2016).

52. Wu, X. *et al.* Structural vulnerability of complex networks under multiple edge-based attacks. in *Proceedings - 2018 IEEE 3rd International Conference on Data Science in Cyberspace, DSC 2018* 405–409 (Institute of Electrical and Electronics Engineers Inc., 2018). doi:10.1109/DSC.2018.00065

53. Lin, C. C., Mitra, R., Cheng, F. & Zhao, Z. A cross-cancer differential co-expression network reveals microRNA-regulated oncogenic functional modules. *Mol. Biosyst.* **11**, 3244–3252 (2015).

54. Weirauch, M. T. Gene Coexpression Networks for the Analysis of DNA Microarray Data. in *Applied Statistics for Network Biology: Methods in Systems Biology* **1**, 215–250 (Wiley-VCH, 2011).

55. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

57. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

58. Wang, W.-C. *et al.* miRExpress: Analyzing high-throughput sequencing data for

profiling microRNA expression. *BMC Bioinformatics* **10**, 328 (2009).

59. Fromm, B. *et al.* MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* **48**, D132–D141 (2020).

60. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

61. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74–e74 (2013).

62. Smit, AFA & Green, P. RepeatMasker at ftp.genome.washington.edu/RM/RepeatMasker.html.

63. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–6 (2011).

64. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford).* **2010**, baq020 (2010).

65. Belinky, F. *et al.* PathCards: multi-source consolidation of human biological pathways. *Database* **2015**, (2015).

66. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).

67. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

68. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

69. Thomas, P. D. *et al.* PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

70. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

71. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

72. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

73. King, B. L. & Yin, V. P. A Conserved MicroRNA Regulatory Circuit Is Differentially Controlled during Limb/Appendage Regeneration. *PLoS One* **11**, e0157106 (2016).

74. Cai, X., Hagedorn, C. H. & Cullen, B. R. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**, 1957–1966 (2004).

75. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).

76. Villalta, S. A., Rosenberg, A. S. & Bluestone, J. A. The immune system in Duchenne muscular dystrophy: Friend or foe. *Rare Dis.* **3**, e1010966 (2015).

77. Vetrone, S. A. *et al.* Osteopontin promotes fibrosis in dystrophic mouse muscle by modulating immune cell subsets and intramuscular TGF-β. *J. Clin. Invest.* **119**, 1583–1594 (2009).

78. Arahata, K. & Engel, A. G. Monoclonal antibody analysis of mononuclear cells in myopathies. I: Quantitation of subsets according to diagnosis and sites of accumulation and demonstration and counts of muscle fibers invaded by T cells. *Ann. Neurol.* **16**, 193–208 (1984).

79. Wehling, M., Spencer, M. J. & Tidball, J. G. A nitric oxide synthase transgene ameliorates muscular dystrophy in mdx mice. *J. Cell Biol.* **155**, 123–131 (2001).

80. Wehling-henricks, M. *et al.* Major basic protein-1 promotes fibrosis of dystrophic muscle and attenuates the cellular immune response in muscular dystrophy. *Hum. Mol. Genet.* **17**, 2280–2292 (2008).

81. Spencer, M. J., Montecino-Rodriguez, E., Dorshkind, K. & Tidball, J. G. Helper (CD4+) and cytotoxic (CD8+) T cells promote the pathology of dystrophin-deficient muscle. *Clin. Immunol.* **98**, 235–243 (2001).

82. Villalta, S. A., Deng, B., Rinaldi, C., Wehling-Henricks, M. & Tidball, J. G. IFN-γ Promotes Muscle Damage in the mdx Mouse Model of Duchenne Muscular Dystrophy by Suppressing M2 Macrophage Activation and Inhibiting Muscle Cell Proliferation . *J. Immunol.* **187**, 5419–5428 (2011).

83. Manzur, A. Y., Kuntzer, T., Pike, M. & Swan, A. V. Glucocorticoid corticosteroids for Duchenne muscular dystrophy. in *Cochrane Database of Systematic Reviews* (ed. Manzur, A. Y.) (John Wiley & Sons, Ltd, 2008). doi:10.1002/14651858.CD003725.pub3

84. Farini, A. *et al.* T and B lymphocyte depletion has a marked effect on the fibrosis of dystrophic skeletal muscles in thescid/mdx mouse. *J. Pathol.* **213**, 229–238 (2007).

85. Nishimura, T. *et al.* Inhibition of matrix metalloproteinases suppresses the migration of skeletal muscle cells. *J. Muscle Res. Cell Motil.* **29**, 37–44 (2008).

86. Lei, H., Leong, D., Smith, L. R. & Barton, E. R. Matrix metalloproteinase 13 is a new contributor to skeletal muscle regeneration and critical for myoblast migration. *Am. J. Physiol. Physiol.* **305**, C529–C538 (2013).

87. Ogura, Y., Tajrishi, M. M., Sato, S., Hindi, S. M. & Kumar, A. Therapeutic potential of matrix metalloproteinases in Duchenne muscular dystrophy. *Frontiers in Cell and Developmental Biology* **2**, (2014).

88. Trougakos, I. P. *et al.* Intracellular clusterin inhibits mitochondrial apoptosis by suppressing p53-activating stress signals and stabilizing the cytosolic Ku70-bax protein complex. *Clin. Cancer Res.* **15**, 48–59 (2009).

89. Bolduc, V. *et al.* Recessive Mutations in the Putative Calcium-Activated Chloride

Channel Anoctamin 5 Cause Proximal LGMD2L and Distal MMD3 Muscular Dystrophies. *Am. J. Hum. Genet.* **86**, 213–221

90. Chandra, G. *et al.* Dysregulated calcium homeostasis prevents plasma membrane repair in Anoctamin 5/TMEM16E-deficient patient muscle cells. *Cell Death Discov.* **5**, (2019).

91. Saller, M. M. *et al.* The role of Sema3-Npn-1 signaling during diaphragm innervation and muscle development. *J. Cell Sci.* **129**, 3295–3308 (2016).

92. Winberg, M. L., Mitchell, K. J. & Goodman, C. S. Genetic analysis of the mechanisms controlling target selection: Complementary and combinatorial functions of netrins, semaphorins, and IgCAMs. *Cell* **93**, 581–591 (1998).

93. King, B. L. *et al.* RegenDbase: a comparative database of noncoding RNA regulation of tissue regeneration circuits across multiple taxa. *npj Regen. Med.* **3**, 10 (2018).

94. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends in Genetics* **22**, 1–5 (2006).

95. Zhang, J. *et al.* Identifying miRNA synergism using multiple-intervention causal inference. doi:10.1101/652180

96. Okada, M. *et al.* The cenp-H–I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. *Nat. Cell Biol.* **8**, 446–457 (2006).

97. Wang, Y., Zhou, Y. & Graves, D. T. FOXO transcription factors: Their clinical significance and regulation. *BioMed Research International* **2014**, (2014).

98. Sandri, M. *et al.* Foxo transcription factors induce the atrophy-related ubiquitin ligase atrogin-1 and cause skeletal muscle atrophy. *Cell* **117**, 399–412 (2004).

99. Whited, J. L., Robichaux, M. B., Yang, J. C. & Garrity, P. A. Ptpmeg is required for the proper establisment and maintenance of axon projections in the central brain of Drosophila. *Development* **134**, 43–53 (2007).

100. Préhaud, C. *et al.* Attenuation of rabies virulence: Takeover by the cytoplasmic domain of its envelope protein. *Sci. Signal.* **3**, (2010).

101. Goodwin, F., Muntoni, F. & Dubowitz, V. Epilepsy in Duchenne and Becker muscular dystrophies. *Eur. J. Paediatr. Neurol.* **1**, 115–119 (1997).

102. Pane, M. *et al.* Duchenne muscular dystrophy and epilepsy. *Neuromuscul. Disord.* **23**, 313–315 (2013).

103. Terry-Lorenzo, R. T. *et al.* Neurabin/Protein Phosphatase-1 Complex Regulates Dendritic Spine Morphogenesis and Maturation. *Mol. Biol. Cell* **16**, 2349–2362 (2005).

104. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).

105. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).

106. Muscular Dystrophy: Options for Complication Management. Available at: https://www.uspharmacist.com/article/muscular-dystrophy-options-for-complication-management. (Accessed: 20th January 2020)

APPENDICES

APPENDIX A: COMPARISON OF MUSCULAR DYSTROPHY TYPES

Table A1. The nine main forms of MD[106].

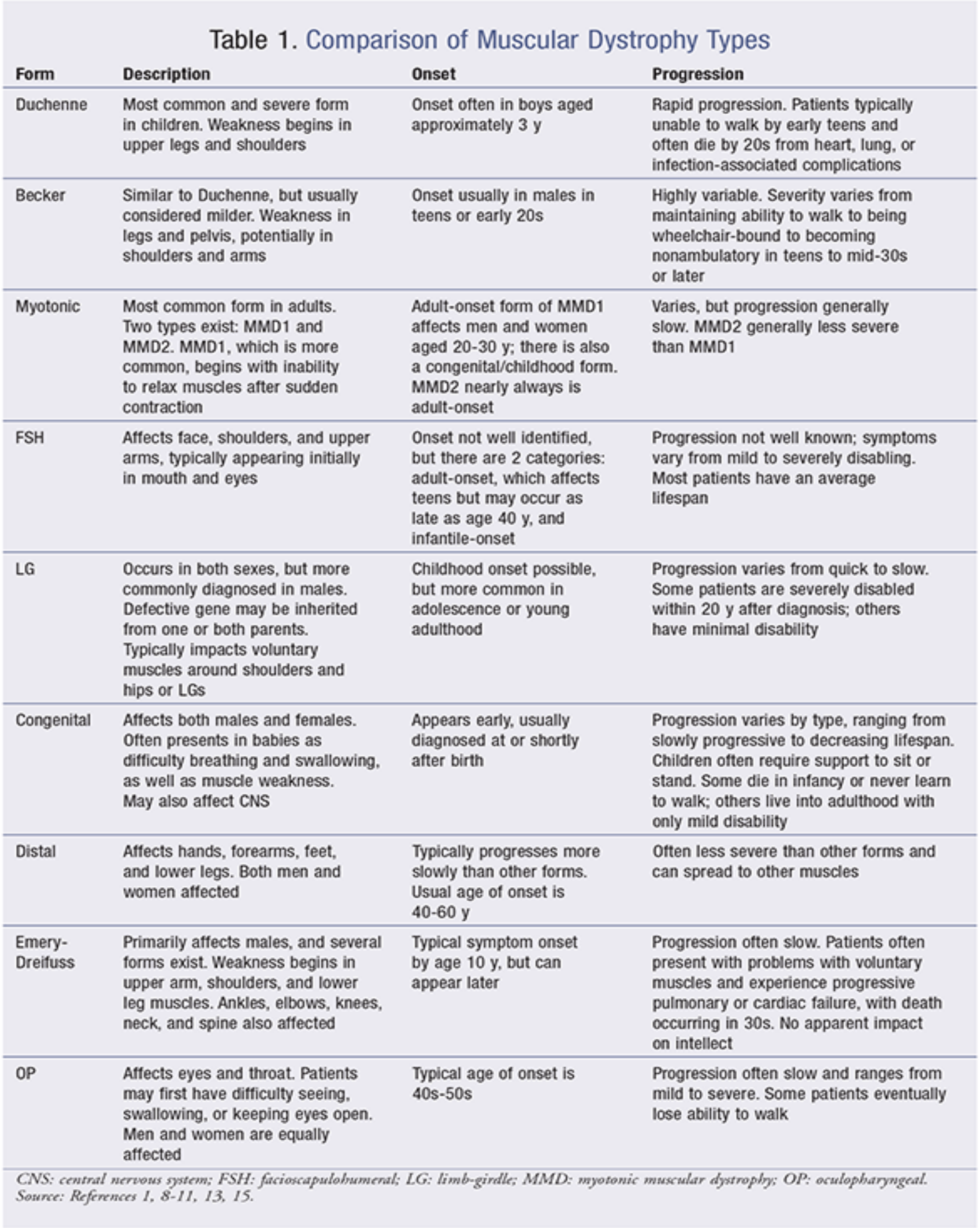


Table 1. Comparison of Muscular Dystrophy Types

Form	Description	Onset	Progression
Duchenne	Most common and severe form in children. Weakness begins in upper legs and shoulders	Onset often in boys aged approximately 3 y	Rapid progression. Patients typically unable to walk by early teens and often die by 20s from heart, lung, or infection-associated complications
Becker	Similar to Duchenne, but usually considered milder. Weakness in legs and pelvis, potentially in shoulders and arms	Onset usually in males in teens or early 20s	Highly variable. Severity varies from maintaining ability to walk to being wheelchair-bound to becoming nonambulatory in teens to mid-30s or later
Myotonic	Most common form in adults. Two types exist: MMD1 and MMD2. MMD1, which is more common, begins with inability to relax muscles after sudden contraction	Adult-onset form of MMD1 affects men and women aged 20-30 y; there is also a congenital/childhood form. MMD2 nearly always is adult-onset	Varies, but progression generally slow. MMD2 generally less severe than MMD1
FSH	Affects face, shoulders, and upper arms, typically appearing initially in mouth and eyes	Onset not well identified, but there are 2 categories: adult-onset, which affects teens but may occur as late as age 40 y, and infantile-onset	Progression not well known; symptoms vary from mild to severely disabling. Most patients have an average lifespan
LG	Occurs in both sexes, but more commonly diagnosed in males. Defective gene may be inherited from one or both parents. Typically impacts voluntary muscles around shoulders and hips or LGs	Childhood onset possible, but more common in adolescence or young adulthood	Progression varies from quick to slow. Some patients are severely disabled within 20 y after diagnosis; others have minimal disability
Congenital	Affects both males and females. Often presents in babies as difficulty breathing and swallowing, as well as muscle weakness. May also affect CNS	Appears early, usually diagnosed at or shortly after birth	Progression varies by type, ranging from slowly progressive to decreasing lifespan. Children often require support to sit or stand. Some die in infancy or never learn to walk; others live into adulthood with only mild disability
Distal	Affects hands, forearms, feet, and lower legs. Both men and women affected	Typically progresses more slowly than other forms. Usual age of onset is 40-60 y	Often less severe than other forms and can spread to other muscles
Emery-Dreifuss	Primarily affects males, and several forms exist. Weakness begins in upper arm, shoulders, and lower leg muscles. Ankles, elbows, knees, neck, and spine also affected	Typical symptom onset by age 10 y, but can appear later	Progression often slow. Patients often present with problems with voluntary muscles and experience progressive pulmonary or cardiac failure, with death occurring in 30s. No apparent impact on intellect
OP	Affects eyes and throat. Patients may first have difficulty seeing, swallowing, or keeping eyes open. Men and women are equally affected	Typical age of onset is 40s-50s	Progression often slow and ranges from mild to severe. Some patients eventually lose ability to walk

CNS: central nervous system; FSH: facioscapulohumeral; LG: limb-girdle; MMD: myotonic muscular dystrophy; OP: oculopharyngeal.
Source: References 1, 8-11, 13, 15.

APPENDIX B – SCRIPTS

B1 Identifying predominantly expressed miRNA

```
#Grace Smith. 1/16. This code will determine which miRNA end (5' or 3') is more highly
expressed
setwd("File/Path")

# Read in miRNA expression data
counts <- read.table("all_counts_normalized.txt",sep="\t",header=T)
counts[is.na(counts)] <- 0

#make lists of each of the arms
counts_5p <- counts[1,]
counts_3p <- counts[2,]

#add the rows based on the 2nd column containing either 3p or 5p
for (i in 3:550)
 if (any(grepl("3p", counts[i,2])))
 { counts_3p <- rbind(counts_3p, counts[i,])
 } else {
    counts_5p <- rbind(counts_5p, counts[i,]) }

#write the tables
write.table(counts_3p, "3p_counts.txt", sep="/t")
write.table(counts_5p, "5p_counts.txt", sep="/t")

#change the column names of each
colnames(counts_5p) <- c("miRNA", "arm", "SIB4_5p", "MIL4_5p", "SEV4_5p",
"SIB7_5p", "MIL7_5p", "SEV7_5p")
colnames(counts_3p) <- c("miRNA", "arm", "SIB4_3p", "MIL4_3p", "SEV4_3p",
"SIB7_3p", "MIL7_3p", "SEV7_3p")

#now I want to compare the expression of the two arms
merged <- merge(counts_3p, counts_5p, by.x=1, by.y=1, all.x=TRUE, all.y=TRUE)
data <- merged[3:8]
data <- cbind(data, merged[10:15])
row.names(data) <- merged[,1]
data[is.na(data)] <- 0

avg <- data.frame(rowSums(data[1:6]))
avg <-  cbind(avg, rowSums(data[7:12]))
row.names(avg) <- merged[,1]
colnames(avg) <- c("3p", "5p")
```

```
avg <- cbind(avg, (avg$`3p`-avg$`5p`))
#now, if 5p is higher, col3 = -; if 3p is higher col3 = +!!!!!
sum(avg[3]>0) #number 3p miRNAs that are expressed strand
sum(avg[3]<0) #number 5p miRNAs that are expressed strand


#now I want to write a file that only contains the predominantly expressed miRNAs and
not the passengers
predom_miRNA <- data.frame(merged[,1], stringsAsFactors=FALSE)

for (i in 1:287)
 if (avg[i,3]>0) {
  predom_miRNA[i,2] <- "3p"
  } else {
  predom_miRNA[i,2] <- "5p" }

colnames(predom_miRNA) <- c("miRNA", "arm")
p <- data.frame(predom_miRNA[,2])
row.names(p) <- predom_miRNA[,1]
write.table(p, "predominantly_expressed_miRNA_arm.txt", sep="")

#in excel I made a file that has the correct labels and then miRNAs with and without a *
at the end
a <- read.table("predominantly_expressed_miRNA_arm_passenger.txt", header=T)
b <- merge(a, counts[,2], by.x=1, by.y=1)
write.table(b, "labeled_predominantly_expressed_miRNA_arm.txt", row.names=F)

sum(grepl("\\*", b$predom_miRNA))
#tells you that 22 of the miRNAs that are predominantly expressed were labeled as
passengers!
```

B2 Repeat Masker analysis to identify candidate lncRNAs

```
#Grace Smith 1/23: Script reads in repeat masked files, calculates % repetitive bases

library(stringr)
setwd("Folder")
filenames <- list.files(pattern="masked.txt")  #repeat masked files to read in

all_transcripts <- data.frame("ID", "total_bases", "repetitive_bases", "ratio_rep_bases")
colnames(all_transcripts) <- c("ID", "total_bases", "repetitive_bases", "ratio_rep_bases")

for (f in 1:length(filenames))
{
  b <- read.table(filenames[f], sep="\t", header=F)
  A=N=0
  for (i in 1:nrow(b))
   {
    if (grepl(">", b[i,1]))     #means new transcript encountered
     {
      if (i==1 & f==1)
      {
       tran <- grep("MSTRG.\\d*.\\d*", b[i,1], value=T)   #if 1st row & 1stfile, do
nothing
      } else {
        new <- data.frame(tran, A, N, (N/A))
        colnames(new) <- c("ID", "total_bases", "repetitive_bases", "ratio_rep_bases")
        all_transcripts <- rbind(new, all_transcripts)
        tran <- grep("MSTRG.\\d*.\\d*", b[i,1], value=T)   #get next id
        A = N = 0    #reset parameters
      }
     } else {
      A = A+str_count(b[i,1])
      N = N+str_count(b[i,1], "N")
     }
  }
  ###if the end of file is reached, add the final row
  new <- data.frame(tran, A, N, (N/A))
  colnames(new) <- c("ID", "total_bases", "repetitive_bases", "ratio_rep_bases")
  all_transcripts <- rbind(new, all_transcripts)
  A=N=0
}

write.table(all_transcripts, "4dpf_Severe_ratio_repeat_masked_bases_adj_P.txt",
sep="\t")

Threshold <- 0.6
threshold_transcript <- subset(all_transcripts,
(all_transcripts$ratio_rep_bases<Threshold))
name <- paste(Threshold,
"Threshold_4dpf_Severe_ratio_repeat_masked_bases_adj_P.txt", sep="_")
write.table(threshold_transcript, name, sep="\t")
```

AUTHOR'S BIOGRAPHY

Grace Smith was born in Rockhill, South Carolina. Her family moved to Maine when she was eight, and since then, she has developed into a true Mainer – skiing, hiking, and rocking t-shirts and Bean Boots in the snow – she's got it down pat.

She majored in Molecular & Cellular Biology and Biochemistry with a minor in Computer Science. During her undergraduate, she was a member of Track and Field Club, a tutor for TRIO support services, a Maine Learning Assistant, and a member of the senior honor society, All Maine Women. She became a member of the King Laboratory at the end of her freshman year which quickly became one of her favorite undergraduate experiences. She was fortunate enough to participate in the Novartis Summer Scholar internship the summer after her sophomore year and the Amgen Scholars program at Washington University School of Medicine in St. Louis the summer after her junior year. She was awarded the Barry Goldwater Scholarship her senior year.

She hopes to enroll in a dual MD/PhD program in the fall of 2022, and until then, will be at the National Cancer Institute via the ICRC post-baccalaureate program.