2021-09-24

Transcriptomics in the Diagnosis of Genetic Myopathies

Joel, Matthew M.

UNIVERSITY OF CALGARY

Transcriptomics in the Diagnosis of Genetic Myopathies

by

Matthew M. Joel

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN NEUROSCIENCE

CALGARY, ALBERTA

SEPTEMBER, 2021

Abstract

The myopathies are a diverse group of primary muscle disorders that arise for a variety of reasons, including both acquired disease (*i.e.* autoimmune disorders), or from genetic variation (the genetic myopathies). RNA sequencing is the application of next-generation sequencing technologies to sequence the transcriptomes of cells and tissues, yielding a functional, and regulatory snapshot of a sample. Comparing the transcriptomes of the autoimmune disorder inclusion body myositis, and a variety of genetic myopathies, including samples with mitochondrial, myofibrillar, dystrophic, or otherwise nonspecific pathology, showed an extensive immunological influence on those with myositis. There are more nuanced differences in the transcriptomes of the histologically grouped conditions among this cohort, including the previously described *FGF21* upregulation in mitochondrial myopathies. Long non-coding RNAs are a neglected species of RNA with myriad regulatory roles. Several non-coding transcripts were identified among the studied groups, that will serve as candidates for testing their biomarker potential for muscle diseases. We tested the utility of RNAseq at diagnosing the genetic myopathy participants of this cohort, identifying four cases where potentially pathogenic variants were detected by accounting for transcript isoform abundance. Genes with transcriptional findings, and potentially pathogenic variants included *FLNC*, *MYOT*, *NEB*, and *SELENON*. The approach may not be optimal for diagnosing individuals with presumed mitochondrial disease, where minimal differences were observed in mitochondrial transcripts. Ultimately, RNAseq provides another tool for clinicians to investigate genetic disorders, and assist with differential diagnosis.

Keywords: myopathy, transcriptomics, lncRNA, variant prioritization, clinical genetics

Preface

This thesis constitutes the original, unpublished, & independent work of the author M. Joel. Experiments reported herein were approved by the Conjoint Health Research Ethics Board at the University of Calgary (ID: REB16-2196_REN4), in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.

Acknowledgements

This work would not have been completed without the mentorship and support of my supervisors Dr. Gerald Pfeffer and Dr. Jason de Koning, and my committee members Dr. Paul Arnold and Dr. Quan Long. I'd like to thank Dr. Adrienne Benediktsson, my undergraduate mentor who instilled in me a love of research. I'd also like to thank the members, past and present, of the Pfeffer and de Koning labs. Kristy Martens and Carly Pontifex, for cataloging and preparing my samples for sequencing and for so warmly welcoming me to the lab. Dr. Ivan Krukov for his assistance with all things technical and for his amenability to bounce ideas off of, relevant to our work or otherwise. Tyler Soule for making our early morning lab meetings exponentially more cheerful. And Zachary Nurcombe, Robyn Wells-Sutherland, and Brooke Belanger for the camaraderie throughout this program over these past years. I'd also like to thank the estate of Katharine Sarah Melinda Mei-Ling Thomas for the generous contribution to my studies in the form of a research scholarship for rare diseases and biomedical engineering.

To Mom & Dad, I need to thank you for support for everything I've done throughout these last 27 years, with minimal but completely justified complaint; I could not have done this without you, literally. And to my dearest friend McKinley Wiens, for helping me enjoy the present while paving the future, and for always being there for me, thank you.

For Mom & Dad, turns out your basement goblin can write

Table of Contents

List of Tables

List of Figures

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
ACMG	American College of Molecular Genetics
ALS	Amyotrophic lateral sclerosis
AMP	Association for Molecular Pathology
BH	Benjamini-Hochberg
BMD	Becker muscular dystrophy
CADD	Combined Annotation Dependent Depletion
CI	Confidence interval
CK	Creatine kinase
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DE	Differential expression
DM	Myotonic dystrophy
DMD	Duchenne muscular dystrophy; dystrophin
eQTL	Expressed quantitative trait locus
FC	Fold change
FSHD	Facioscapulohumeral dystrophy
FTD	Frontotemporal dementia
GATK	Genome Analysis Toolkit
gnomAD	Genome Aggregation Database
GO	Gene ontology
GTF	Gene transfer format
GVCF	genome variant call formatted file
GWAS	Genome-wide association study
hIBM	Hereditary inclusion body myopathy
HLA	Human leukocyte antigen
IBM	Inclusion body myositis
IG	Immunoglobulin
IPA	Ingenuity Pathway Analysis
kNN	k-Nearest neighbours
lncRNA	Long non-coding RNA
MFM	Myofibrillar myopathy
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MtM	Mitochondrial myopathy
ncRNA	Non-coding RNA
NGS	Next generation sequencing
NMD	Nonsense mediated decay
OMIM	Online Mendelian Inheritence of Man
ONT	Oxford Nanopore Techonologies

PC	Principal component
PCA	Principal component analysis
RNA	Ribonucleic acid
RNAseq	RNA sequencing
RPKM	Reads per kilobase per million mapped reads
rRNA	ribosomal RNA
scRNAseq	Single-cell RNAseq
SMRTseq	Single-molecule real-time sequencing
SNP	Single-nucleotide polymorphism
snRNAseq	Single-nuclear RNAseq
STAR	Spliced Transcripts Alignment to a Reference
TPM	Transcripts per million
VEP	Variant Effect Predictor
VUS	Variant of unknown significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
μRNA	Micro-RNA

Die Philosophen haben die Welt nur verschieden interpretiert;

es kommt aber darauf an, sie zu verändern

The philosophers have only interpreted the world, in various ways.

The point, however, is to change it.

Karl Marx, *Theses on Feuerbach*

Chapter 1 – Introduction

1.1 Myopathy

Myopathy is a broad class of disorders of the musculoskeletal system. Weakness is the common symptom of myopathy, frequently accompanied by muscular pain, involuntary muscle activity, or disability. We can further break down the myopathies into two separate groups; genetic cases, where inherited or *de novo* mutation is responsible for muscular dysfunction, or acquired, where the myopathy is an adverse effect from the affected individual's environment. The most common genetic myopathies, and consequentially the most studied, are the dystrophinopathies: Duchenne & Becker's muscular dystrophies (DMD, & BMD respectively), myotonic dystrophy (DM; *dystrophia myotonia*), and facioscapulohumeral dystrophy (FSHD). However, there are rarer myopathies with genetic causes, including myofibrillar myopathy (MFM), and mitochondrial myopathy (MtM). These, along with the acquired myopathy inclusion body myositis (IBM), will be the focus of discussion for the purposes of this thesis.

1.1.1 Dystrophinopathies

Duchenne, and Becker's muscular dystrophy are hereditary degenerative diseases caused by genetic variation in dystrophin, affecting an estimated 1:3500 male births[1-6]. DMD, and BMD tend to follow an X-linked recessive pattern of inheritance, due to dystrophin's presence on the X-chromosome[1-3,6]. Therefore, females tend to be asymptomatic carriers for the disease, unless there is a skewed X-inactivation, prioritizing gene expression from the affected chromosome[3,4]. DMD, and BMD are typically thought to fall on a disease spectrum. Where DMD is recognized as being a severe manifestation, with early-onset disability[2,6], while BMD

is typically regarded as a less severe, but more variable disease, with disability expected anywhere between a participant's early-twenties, to beyond their sixties[2,6]. Dystrophin is regarded as the largest gene in the human genome, and is transcribed to produce three tissue-specific protein-coding isoforms[4]. Due to its size, dystrophin is prone to harbouring many genetic variants. Nearly 3000 are attributed to a muscular dystrophy phenotype[3]. Generally, the variation impacts the reading frame of the mature messenger RNA (mRNA) product. In DMD you find premature stop gains from a shifted reading frame[2,4-6]. Whereas, in BMD the mature mRNA has a preserved reading frame, that can code for a truncated, but relatively functional protein[2,4-6]. DMD is a rare case where genetic modifiers have been identified. For example, a reduction in expression of the *CUGBP1* isoform Celf2a, was identified by Martone *et al*.[7] to rescue dystrophin function, resulting in a phenotype more similar to BMD.

1.1.2 Myotonic Dystrophy

Myotonic dystrophy is distinct from the other myopathies by occurrence of myotonia, cardiac conduction disorders, insulin resistance, cataracts, and cognitive impairment[1,8]. Unlike DMD, and BMD, DM is an autosomal dominant disease, affecting about 1:12500 people[1,9]. The molecular cause being gain-of-function tri- and tetranucleotide repeat expansions[8]. Specifically, in type 1 DM, we find a CTG repeat expansion in *DMPK*[8]. Type 2 DM instead has a CCTG repeat expansion in the zinc finger protein-coding gene *ZNF9*[8]. While other trinucleotide expansion disorders, such as Huntington's, or spinocerebellar ataxia are caused by translation of the expanded repeats, resulting in protein aggregation, DM pathology is a product of RNA interference, where the repeats interfere with alternative splicing[10]. Like the other trinucleotide

repeat disorders, genetic anticipation can be observed in successive generations, where the unstable repeat expands on the germline, correlating with earlier onset, or increased severity[11].

1.1.3 Facioscapulohumeral Dystrophy

Facioscapulohumeral dystrophy is another superficially autosomal dominant muscle disorder that affects approximately 1:25000 people[1,4,12]. FSHD is distinct from other myopathies by presenting with a descending pattern of atrophy and weakness originating in the muscles of the face and upper body[1]. More severe cases may be associated with extra-muscular comorbidities, including intellectual disability, epilepsy, retinal disease, or neurological forms of hearing-loss[1]. While considered autosomal dominant in the literature, FSHD is not completely penetrant[13]. There is significant variation in the severity of symptoms between generations of affected families, and even twins[13]. The first source of FSHD was found to be the microsatellite array, and CpG island *D4Z4*[9]. Unlike the repeat expansions observed in DM, and other diseases, the structural variant with FSHD is a reduction in the number of repeats[12]. However, his reduction may be insufficient to cause disease; approximately 3% of the population has a number of repeats (4-8) within the pathological range (< 10)[12]. The next cause of FSHD identified was variation in *SMCHD1*, that appears to affect the epigenetic regulation of the *D4Z4* region[1,12].

1.1.4 Myofibrillar Myopathy

Myofibrillar myopathy is a clinically heterogeneous myopathy that tends to present in distal muscle groups, with the occasional coincidence of cardiomyopathy, or peripheral neuropathy[14,15]. A MFM diagnosis is typically made by histological findings, where it is characterized by atrophic and disorganized muscle fibres[14,15]. The pathology is due to Z-disk

dysfunction from the abnormal aggregation of associated proteins[14,15]. Genes that are frequently implicated are associated with Z-disk organization, including: desmin (*DES*), α-B crystallin (*CRYAB*), myotilin (*MYOT*), filamin C (*FLNC*), *ZASP*, Bcl2-associated athanogene (*BAG3*), or titin (*TTN*)[15]. However, most individuals do not have a genetic diagnosis[15].

1.1.5 Mitochondrial Myopathy

Given that mitochondria have been an integral part in the history of Eukarya, it is not surprising that dysfunction can result in a variety of diseases in humans. Among those, are the primary mitochondrial myopathies. Skeletal muscle is heavily dependent on mitochondria for chemical potential energy from oxidative phosphorylation, calcium homeostasis, synthesis of several macromolecules, and redox signalling, and are consequently packed with the organelle[16,17]. Variation in the mitochondrial genome, containing 13 respiration-related protein coding genes, and 2 ribosomal and 22 transfer RNAs, or any of the nuclear chromosomes that contain the remaining several hundreds of proteins critical to mitochondrial function, and interaction with the rest of the cell, can result in disease[18]. Mitochondrial diseases, including MtM, tend to be clinically heterogeneous, often believed to be a result of heteroplasmy, where a proportion of a cell's mitochondria are mutated, resulting in multiple subpopulations[18].

1.1.6 Inclusion Body Myositis

Inclusion body myositis is one such case. IBM is a late-onset myopathy that primarily impacts the flexors of the hands and feet, and the extensors of the knee[19,20]. Rarely, the weakness can impact respiratory muscles, resulting in dysphagia, and increasing the risk of aspiration pneumonia[19]. IBM is a member of the presumed autoimmune myopathies[19]. This

presumption is based on several findings: auto-reactive antibodies in patient sera, elevated cytokines, cytotoxic T-cell lymphocytosis, and infiltration in affected muscle, and increased risk in some human leukocyte antigen (HLA) haplotypes[19,20].

1.1.7 Diagnosis of Myopathy

The diagnostic battery for myopathies involves a series of blood draws, electromyography studies, muscle biopsy, magnetic resonance imaging (MRI), and genetic sequencing[13,19,21]. Other tests may be considered to evaluate comorbidities or rule out differential diagnoses[13]. Creatine kinase (CK) is a nonspecific marker for muscular damage, and is often elevated in certain myopathies[13]. In IBM, the lymphocytosis will be observed in a complete blood count[21]. There are several histological features that can be found in the myopathies[13,19]. Centralized nuclei are indicators of muscle disease, and can be found in the aptly named centronuclear myopathies, as well as IBM[13,19]. Rimmed vacuoles are another feature found in many late onset *ZASP*- or titinopathies, early-onset distal myopathies, limb-girdle muscular dystrophy, and IBM[13,19]. Inflammatory infiltrates are characteristic of IBM and the other autoimmune myopathies, but can also be observed in cases such as FSHD where necrotic muscle must be cleared away[19]. Electron microscopy may be employed to identify ultra-structural anomalies[13]. Today it is common to confirm findings by molecular analyses[19].

1.1.8 Sources of Variability in Disease Presentation

Beyond the nuances caused by differences in pathogenic variants, genetic anticipation, inheritance patterns, or alternative splicing, there are influences on disease presentation. An example is biological sex. In muscles, there are metabolic differences between male, and female

muscles, where type I fibres dominate muscles in females, in contrast to the type II in males[22,23,24]. This has been found in both the vastus lateralis, and the biceps brachii[22,24]. Aging also impacts the sexes differently in terms of muscle health, which can be seen in the sex-specific prevalence of age-related sarcopenia, obesity, or insulin resistance[23]. Sex differences have also been observed in *ANO5*-based limb girdle muscular dystrophies, where males tend to have higher prevalence, and severity[25,26]. A similar trend was found in FSHD males, when compared to females[27]. Addressing confounding variables in studies will be increasingly important, as we try to understand the relationship between genetics, and myopathic changes.

1.2 Molecular Diagnosis

1.2.1 Molecular Diagnostic Tools

In a clinical setting, genetic sequencing technologies are typically employed to identify the molecular etiology for a participant's disease. Targeted sequencing may be utilized to capture a participant's variation within genes associated with their disease, including structural variation like the deletions in DMD, or microsatellite copy number variation in FSHD. Alternatively, next-generation sequencing (NGS) such as whole exome or genome sequencing (WES, WGS, respectively) would identify single-nucleotide, and small indel variation within the entirety of the participant's coding regions, or genome, respectively. NGS would increase our sensitivity of finding a pathogenic variant, though at the expense of sensitivity to large structural variants.

1.2.2 ACMG Criteria for Molecular Diagnosis

A problem arises when a participant does not possess variants that are well characterized in terms of the disease in question, or where the severity of their phenotype is inconsistent with a

plausible variant. For the former, the American College of Medical Genetics (ACMG) with the Association for Molecular Pathology (AMP) have a published set of guidelines for classifying variants[28]. Where variants are categorised on their degree of pathogenicity based on allele frequency in the population, coding sequence consequence, segregation with disease, computational predictions, and experimentally verified biological impacts.[28]. A variant's designation is decided by the combinations of evidence available, where several lines of evidence of pathogenicity are required to declare a variant pathogenic[28]. When conflicting, or insufficient evidence of pathogenicity, or benignity is available, the ACMG guidelines recommend a designation of "variant of uncertain significance" (VUS)[28]. There have been several attempts to further systematize the ACMG guidelines, including formulating their criteria as a Bayesian classification framework, where the combinations of evidence were found to be more or less internally consistent[29], or modifying the guidelines to accommodate continuous variables, over their current logical values (i.e. using an allele frequency value, rather than checking if it exceeds a threshold)[30]. There have also been attempts to translate the guidelines to accommodate copy number variants (CNVs), such as exon deletions, or tandem repeats[31]

1.2.3 Limitations to Molecular Diagnosis

While the ACMG criteria improve the reproducibility of clinical genetics, they are confounded by various sources of incomplete penetrance. For example, if a SNP modifies splicing such that a critical exon is included, or there is an improper intron retention event, the degree of biological impact would be missed by the criteria, without experimentation. Genetic modifiers could also impact splicing, or the penetrance of a potentially pathogenic variant, and this too could cause a pathogenic variant to slip under the radar. ACMG does allow experimental

evidence to influence variant interpretation, however, thoroughly testing individual variants can easily become overwhelming with the sheer scale of WES, or WGS. The guidelines also make several assumptions that are not always defensible, particularly for late-onset conditions. This can be illustrated with one of the most influential lines of evidence for ACMG, allele frequency. When applying allele frequency thresholds to variant prioritization, we are inherently assuming that the general population sampled consists of strictly healthy individuals, at all time points in their life. This completely neglects the possibility of late-onset conditions, where an individual in a population study can present as healthy at the time of sequencing, only to develop disease later in life. The same can be said for computational impact predictors designed to interpret variants based on evolutionary conservation, where a variant may not be selected against if disease manifests beyond an individual's reproductive prime. Further, the later a disease presents, the more difficult segregation analyses become, as parents are more frequently deceased, and children have yet to develop a phenotype. Ideally, it would be possible to extract participant-specific information that would improve our power in prioritizing plausibly pathogenic variants, to improve diagnosis, treatment, and genetic counseling.

1.3 RNA Biology

The transcriptome is the body of all products translated from the genome. It is a dynamic entity, specific to each cell, and is influenced by its interactions with its environment. The transcriptome contains, but is not limited to: the protein-coding mRNAs, ribosomal rRNAs that catalyze translation, transfer tRNAs that carry amino acids to the ribosome and match them to their codons during translation, and various other species of non-coding RNAs (ncRNA), that act in a regulatory capacity.

1.3.1 Alternative Splicing

The diversity of the transcriptome, and proteome is due to alternative splicing. Combinations of exon inclusion, intron retention, splice site usage, *etc.* allow for fine control over cellular processes, without the genetic bulk. The regulation of alternative splicing is complex, comprised of an ensemble of protein splicing enhancers and silencers, the ribonucleoproteins that comprise the machinery of the spliceosome, and the sequences recognized by these molecules[5,32]. These processes are spatially, and temporally, constrained[32]. With the efficacy of exon definition dependent on the kinetics of splicing factor binding, immature RNA secondary structure, signalling events within the cell, and even the kinetics of RNA polymerase during transcription[32]. Exon enhancers and silencers, splice sites and intronic splicing motifs, RNA secondary structure, and transcription kinetics are all sequence sensitive. Even SNPs could plausibly tip the balance of a gene's splicing into aberrant territory.

Variation could impact splicing at the level of exon definition, introduction of cryptic splice sites, loss of efficiency with recursively spliced introns, or otherwise non-canonical intron retention[33,34]. Exons themselves are defined by the binding of exonic splicing enhancers to regulatory elements within the RNA[33,34]. These factors assist in the recruitment of spliceosome components, specifically the small-nuclear ribonucleoproteins (snRNPs) U1, and U2, that bring the ends of two exons into proximity[33]. Both SNPs and structural variants could impact the efficiency of splicing factors recognizing and binding to their enhancers. Likewise, variation at canonically defined splice-sites could similarly impact the efficiency of snRNP binding, and exon definition. The opposite is also true; cryptic splicing, for example, is the aberrant splicing of inefficient splice sites[33]. Many of these splice sites are found in the human genome, and

utilized to some extent[33]. Even common cryptic splicing events often result in intron retention[33]. Introns of course being relatively enriched with stop codons, means these transcripts are often downregulated by non-sense mediated decay (NMD)[33]. Another mechanism of intron retention is a loss of efficiency in recursive splicing. In longer introns, a single pair of splice sites may be insufficient to efficiently splice out an intron[33]. The solution has been recursive splicing, where portions of the intron are iteratively removed in succession during transcript maturation[33]. This is, of course, dependent on the efficiency of the intronic splice site, and a reduction in that efficiency could result in intron retention. Lastly, outside of these processes, introns can be retained due to stalling of RNA polymerase during transcription[33]. GC-rich sequences can reduce the transcription velocity of RNA polymerase, the stalling of which can prevent the time-sensitive binding of splicing factors meant to remove the intron[33].

1.3.2 RNA Sequencing

RNA sequencing (RNAseq) is the application of NGS methods towards the sequencing of each RNA molecule contained with a tissue, or cell. For our purposes, RNAseq may make it possible to prioritize a participant's variation with by quantifying the expression of individual exons, transcript isoforms, or entire genes. Effectively providing investigators with a sense of the dysregulation underlying the pathology. There are multiple ways variation could impact a participant's transcriptome. Frameshifts, and stop gains may result in a reduction in mature mRNA isoforms due to NMD[33]. The significance of splice variants would be associated with the degree to which an exon is differentially expressed. If a variant impacts transcription regulation, it could be detectable by quantifying RNA reads. Additionally, when 1/3 of described

pathogenic variants are predicted to impact splicing[33], transcriptomics could provide insight into the pathogenesis of a participant's disease.

1.3.3 Skeletal Muscle Transcriptome

Splicing is critical to the differentiation, and physiology of every cell-type. Muscles in particular are dependent on changes in the abundance of transcript and protein isoforms to execute their developmental program, and to fit their functional niche[5]. In fact, muscle is among the tissues with the highest degree of alternative splicing, and tissue-specific exons[5,35]. For example, both the sarcomeric protein titin, and microfilament-associated regulator α-tropomyosin have not just muscle-specific, but skeletal, cardiac, or smooth muscle-specific exons[5]. This specificity is not static, and even different stages in myogenesis will have their own specific exons[5]. As another example, active muscle has a significantly higher energy demand than, say myoblasts, and a part of maturation is the transition between exons of a mitochondrial calcium channel, that increases the sensitivity of energy production to calcium[5]. Entire protein functions could be reversed during development. As is the case of *MEF2*, that undergoes a shift between isoforms containing mutually exclusive exons, where one exon causes *MEF2* to act as a transcriptional repressor, while the other causes it to act as an activator.

1.3.4 Muscle Transcriptome in Disease

The complexity of alternative splicing in skeletal muscle is particularly noticeable in muscle disease. The low-complexity repeat-expansions in *DMPK*, and *ZNF9* observed in DM have been identified to sequester the muscleblind-like splicing regulator (*MBNL*) into nuclear inclusions[5,8,32]. The resulting loss of available *MBNL* causes a reduction in the expression of the

insulin receptor (*INSR*), a muscle specific chloride channel (*CLCN1*), a cardiac isoform of troponin, and even an exon of dystrophin (*DMD*) causing the characteristic symptoms of DM[5,8, 36]. The remaining transcripts are more commonly found during development, and a similar observation can be made in the dystrophinopathies, and in FSHD[5,8, 9]. In FSHD this is caused by untimely *DUX4* expression, either by a reduction of copy-number or methylation of *D4Z4*, that inhibits the cell's ability to transition to a mature phenotype[9,12]. The variants that can cause DMD, or BMD include splicing variants, that cause the premature termination, or protein truncation[2,4-6].

1.3.5 Long Non-coding RNAs

So far, the discussion on cellular physiology, and disease etiology has been described by changes to protein function. However, proper cell function is not restricted to translation of mRNAs. Long non-coding RNAs (lncRNAs) are an increasingly studied RNA species; distinct from mRNA in that they do not contain a significant open reading frame, and therefore not conventionally translated, but different from other ncRNAs by their length (>200 bp). Expression of different lncRNAs varies even more widely among different cell-types and developmental stages than mRNAs[3]. They have been identified interacting with both nucleic acids, and proteins[3]. They can modulate expression and function of genes by recruitment of chromatin remodelling factors, impacting transcription initiation, influencing splicing, regulation of the nuclear architecture, polysome kinetics, including influence over internal ribosome entry sites (IRES motifs), or the sequestration of complementary micro RNAs (μRNAs)[3]. LncRNAs have been implicated, either functionally, or as biomarkers in a variety of diseases, including, but not limited to, multiple sclerosis[37], various cancers[38,39], and cardiovascular disease[40].

1.3.6 Skeletal Muscle & lncRNAs

LncRNAs too are heavily implicated in muscle physiology. For example: an enhancer-derived RNA, essentially a transcribed transcription enhancer, from myoblast determination protein 1 (MyoD) regulates both its parent gene, and myogenin (MyoG), critical transcriptional regulators throughout muscle development[3]. The steroid receptor RNA activator 1 (*SRA1*), is a gene that is unique in that the protein product's function is to bind a lncRNA transcript of the same gene, and in muscle, another of the transcriptional regulators of MyoD is this *SRA1* lncRNA[3].

1.3.7 Myopathy & lncRNAs

Myopathies also have several associations to lncRNAs, Dystrophin codes for multiple lncRNAs[3,4]. Some of these isoforms have been found to be differentially expressed in female carriers of disease-associated variants, regardless of symptom presentation[3,4]. Also in DMD and BMD, is the dysregulation of the lncRNA lincMD1 in myoblasts[3,36]. LincMD1 is a muscle specific transcript that is necessary for the latest stages of muscle differentiation[3,36]. It appears to function as a μRNA sponge, sequestering μRNAs that would otherwise interfere with the progression of the cell's developmental program[36]. This effect has been rescued by supplementing DMD participant-derived myoblasts with exogenous lincMD1[3,36]. FSHD too has associations with a lncRNA. Within the *D4Z4* microsatellite array implicated in FSHD pathogenesis, is a lncRNA *DBET* (D4Z4 binding element transcript)[41]. This transcript is able to recruit a chromatin methyltransferase (*ASH1L*), further contributing to the disinhibition of the locus[41]. Clearly there is precedence for dysregulation of lncRNAs in myopathies.

1.4 Purpose

Genetic variation does not just have consequences at the protein level. Particularly in muscle disease, where splicing, aberrant expression, and lncRNAs are implicated. While many of the diagnostic criteria are fairly specific to many of the myopathies. There are always the cases that do not fit perfectly into any one diagnosis. There is a vast network of interactions between proteins, nucleic acids, etc. where dysfunction at multiple nodes could manifest in similar ways. However, trying to accommodate the interconnectedness of cellular processes can quickly overwhelm an individual when identifying a molecular diagnosis. And the consequences of any one variant can be difficult to assess. Such as when it impacts a non-coding region, occurs at a splice site, or causes a synonymous change at the codon level. There are programs that can predict the efficiency of splice sites, but splicing is dependent on cellular context. There is a way, however, to identify the effect of variation on alternative splicing, and transcriptional regulation, by means of RNAseq. It is also possible to identify the impact of variation on the expression of non-coding RNAs, who's biological functions are increasingly investigated.

1.5 Aims

1.5.1 Aim 1 – Differential Expression Between Genetic & Acquired Myopathies

The plethora of data generated from RNAseq lends well to discovery-based studies to better understand the cellular effects, and markers of different biological conditions. In the context of this project, this entails contrasting acquired, and genetic myopathy cases, and their various pathologies, to better understand the cellular mechanisms of disease, and to identify RNA-based biosignatures, that could be used as an additional diagnostic tool.

1.5.2 Aim 2 – Application of RNAseq to Clinical Molecular Diagnosis

For rare diseases, conventionally available information for variant prioritization may be inadequate for identifying the most plausible causal variant(s) for an individual's disease. For a variant to impact biology, it plausibly needs to impact protein or RNA structure, or gene regulation. Given that RNAseq can provide both quantitative data in the form of read counts, and differential expression, and a platform for variant calling, it is plausible that it can be utilized for molecular diagnostics, with the benefit of providing some biological context with which to prioritize variants.

Chapter 2 – Methods

2.1 Ethics Approval

Ethical considerations for this project were reviewed, and approved by the Conjoint Health Research Ethics Board at the University of Calgary (ID: REB16-2196_REN4), in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.

2.2 Participant Cohort & Sample Processing

Participants were recruited by the principal investigator, Dr. Gerald Pfeffer, under his clinical practice. Sixteen participants were included in the present study, including 3 presenting with myositis, 3 with mitochondrial, 4 with myofibrillar, 2 with dystrophic, and 4 with non-specific findings reported in their histopathology results (Table 1), with a mean age of onset of 49 (± 4.26; standard error of the mean), 10 samples with elevated serum CK concentrations (cohort mean of 609 ± 125 U/L; standard error of the mean). Clinically ordered targeted sequencing was available for many of the genetic myopathy samples.

Samples were collected at the time of clinical biopsy, from the vastus lateralis, deltoid, gastrocnemius, quadriceps femoris, or tibialis anterior, flash frozen in liquid nitrogen, and stored at -80°C until processing. Tissue was homogenized in a Precellys 24 Tissue Homogenizer (Bertin Instruments, Paris, France), with 2.8 mm ceramic beads in 2.0 mL microcentrifuge tubes as per manufacturer's instructions. RNA was isolated from tissue homogenates with RNeasy Fibrous Tissue Kits (Qiagen, Hilden, Germany), according to manufacturer's instructions.

Table 2.1: Participant cohort

Participant ID	Histopathology Presentation	Sex	Age at Onset	Highest Recorded Serum CK (U/L)	Biopsy Source
1330	Myositis	Male	64	384	Vastus lateralis
1328	Myositis	Male	59	685	Vastus lateralis
1133	Myositis	Female	60	70	Deltoid
1308	Myofibrillar	Male	46	685	Gastrocnemius
1255	Myofibrillar	Male	1	71	Vastus lateralis
1373	Dystrophic	Male	35	700	Quadriceps femoris
1021	Dystrophic	Male	65	1120	Tibialis anterior
1105	Myofibrillar	Male	64	787	Deltoid
1099	Mitochondrial	Male	53	242	Deltoid
1312	Mitochondrial	Female	62	976	Deltoid
1514	Nonspecific	Male	52	1867	Vastus lateralis
1515	Nonspecific	Female	50	1207	Quadriceps femoris
1517	Mitochondrial	Male	45	106	Deltoid
1519	Myofibrillar	Male	29	672	Vastus lateralis
1523	Nonspecific	Female	62	45	Deltoid
1524	Nonspecific	Female	37	133	Deltoid

2.3 RNA Library Preparation & Sequencing

RNA sequencing was performed at the Centre for Genomics & Informatics at the University of Calgary, with the following procedure. Libraries were prepared from 500 ng of isolated RNA with TruSeq™ Standard Total RNA Library Preparation Kits (Illumina, Albany, NY, USA) in accordance with the manufacturer's "Low-input Sample" protocol, including depletion of ribosomal RNA (rRNA) with RiboZero™ magnetic beads (avoiding transcript type biases as may be seen with other library preparations, i.e. poly-A selection for mRNAs), indexed with Illumina's TruSeq i7 adapter indices. Samples underwent 15 cycles of PCR enrichment, and clean-up, followed by quality validation by TapeStation™ (Agilent, Santa Clara, CA, USA) for verification of rRNA depletion, and qPCR with KAPA™ qPCR Library Quant Kit for Illumina (Roche Sequencing, Pleasanton, CA, USA) for library quantification before pooling. Paired-end 50 bp fragment sequencing was performed on an Illumina NovaSeq™ 6000 (Illumina, Albany, NY, USA) with the appropriate kit for an average of 100 million reads per-sample, to ensure accuracy for both variant calling, and detection of low-abundance transcripts (i.e., lncRNAs).

2.4 Variant Calling Pipeline

Alignment, and variant calling of raw sequencing reads was performed with the Broad Institute's Genome Analysis Toolkit (GATK), in accordance with their best practices pipeline for calling germline variants from RNAseq reads[42-44]. This entails alignment of the reads to an indexed human GRCh38 genome with a splice-junction aware aligner, STAR[45]. include Base Quality Score Recalibration (BQSR) to address systemic errors in base call, HaplotypeCaller for variant calling, and variant recalibration[43]. HaplotypeCaller is advantageous over other variant calling tools due to its fidelity in detecting small indels, and calling reference states, or

insufficient read-depth across samples, in generating a genomic variant call formatted file (GVCF)[43]. The developers openly acknowledge HaplotypeCaller deliberately calls more false positive variants than other tools, in order to maximize sensitivity[43]. This is remedied by implementing their recommendation of variant recalibration and filtering[43]. Processed participant variants will be annotated with VEP (Variant Effect Predictor; Ensembl)[46] for allele frequency as documented by gnomAD, CADD scores[47], and OMIM phenotypes, in addition to default parameters, for clinical variant prioritization.

2.5 RNA Quantification & Differential Expression Analysis

Transcript abundance was estimated using the program Salmon, mapping sequenced reads to a transcriptome containing both coding, and non-coding transcripts from Ensembl v100, the latest version at the beginning of this study, with decoys generated from the reference genome to reduce false mappings[48]. The desirable features of Salmon include its speed, comparable to another popular mapper Kallisto (both of which display superior speed when compared to traditional aligners), and its accounting for GC-content, 5'-sequence, and 3'-sequence biases in estimating transcript abundances, which are prone to introducing false positive results[48]. While Salmon can quantify previously aligned reads, the developers have found there is actually a marginal increase in error when doing so, when compared to its baseline quantification algorithm[48]. Differential expression (DE) of both transcripts and genes was quantified with the R package DESeq2[49]. DESeq2 controls for library size, and fits the data to a negative binomial generalized linear model, utilizing a Wald test for hypothesis testing by comparing the parameters of the null and alternative models, and adjusts for multiple hypothesis testing with the Benjamini-Hochberg correction (BH)[49]. A similar program developed by the

same group, DEXSeq, was utilized to quantify reads mapping to document exons (contained in the GRCh38 GTF provided by Ensembl v100), and differential expression analysis, using an algorithm similar to DESeq2[50]. However, the variability in exon usage proved difficult to interpret for variant prioritization purposes, and the decision was made to focus on transcript-, or gene-level analyses.

2.6 Gene Ontology & Ingenuity Pathway Analysis

Enrichment analysis was performed with DAVID's (Database for Annotation, Visualization and Integrated Discovery) functional clustering tool[51], focusing on the gene ontology categories of biological function, cellular compartments, and molecular function, on the list of genes from the differential expression analysis between acquired, and genetic myopathy samples, where the genes selected fulfilled the following criteria: overexpression in the acquired cases (log_2FC [fold change] < -1), and statistically significant increase in expression (BH corrected p-value < 0.05). Statistically significantly (BH corrected p-value < 0.05) differentially expressed (|log2FC| > 1) genes were forwarded for Ingenuity Pathway Analysis© (IPA; Qiagen, Hilden, Germany)[52] testing for canonical pathway enrichment, in order to complement the GO findings, and to compare to recent literature[53]. IPA was repeated to illustrate the differences between deltoid-, and vastus lateralis-derived samples, to illustrate the degree of biological heterogeneity between muscle groups.

2.7 Clinical Variant Prioritization

Variant call formatted files were annotated with VEP (Variant Effect Predictor; Ensembl)[46] for allele frequency as documented by gnomAD, and OMIM phenotypes, in

addition to default parameters, for clinical variant prioritization. For consistency with clinical approaches, prioritization was based on the lines of evidence considered under the AMCG-AMP's guidelines[28]: gnomAD allele frequency for criteria PM2, BA1, and BS1, variants unique to each specimen for PP4, and PM6, and functional impact predictions, and OMIM annotations for PVS1, PS1, PM4, PM5, and PP2 (Table 2.2)[28]. Transcript count statistics were employed to investigate their application in prioritization, similar to the ACMG criteria PS3, and BS3, which allow for biological impacts to inform prioritization. Specifically, an allele frequency threshold of 0.001 was applied to reduce the false negative rate for possible recessive causal variants. We focused on variants unique to each participant on the grounds that their phenotypes diverged sufficiently, and a believed lack of relatedness, meant that shared variants would be uncommon. Our search was limited to moderate to severe impacts, as predicted by VEP (including missense, splice acceptor or donor, stop gain or loss, start gain or loss, in-frame indels, or frameshifts). Variants in genes associated with neuromuscular disorders were prioritized (Appendix Table A; compiled from http://www.musclegenetable.fr). Transcript counts and their Z-scores were employed as a measure to compare each sample to the remainder of the cohort for each transcript, or exon. Transcripts with Z-scores exceeding the 95% confidence interval (CI) for each transcript were selected for filtering, assuming a normal distribution ($|Z| \geq 1.96$).

2.8 Other Packages & Libraries Used

General data manipulation, and plotting was performed with Python (v3.8.10), with the numpy (v1.20.3), pandas (v1.3.0), and matplotlib (v3.4.2) with seaborn (v0.11.1), scipy (v1.6.2), and statmodels (v0.12.2) packages. The statistics-oriented programming language R (v4.1.0) was used for operations involving packages from the bioinformatics tool repository Bioconductor

Table 2.2: Summary of applicable ACMG-AMP guideline criteria.

ACMG Criterion	Description
PVS1	Loss of function variant, where loss of function common in disease
PS1	Identical missense variant for disease
PM2	Absence from gnomAD, otherwise extremely rare for presumed recessive disorders
PM4	inframe, stop-loss variants
PM5	Novel missense variant for disease
PP2	Missense variant, where missense is common in disease
PP3	Ensemble of computational evidence for pathogenicity
PP5	Reputable source classified variant as pathogenic
BS1	Allele frequency greater than expected
BP1	Missense, where loss of function typical for disease
BP4	Computational evidence for benignity
BP6	Reputable source classified variant as benign

(v1.30.16)[54], including: tximport (v1.20.0)[55] to read count files generated by Salmon, ensembldb (v2.16.3)[56] and biomaRt (v2.48.2)[57] for accessing genomic data from Ensembl, GenomicRanges (v1.44.0)[58] for data manipulation involving genomic positions, Gviz (v1.36.2)[59] for plotting genomic loci, and class (v.7.3.19) for k-nearest neighbours (kNN) cross validation of the PCA results.

Chapter 3 – Transcriptomic differences between acquired & genetic myopathies

3.1 Results

3.1.1 Differential Expression Between Acquired & Genetic Myopathy

Testing for DE between etiologies reveals most statistically significant, differentially expressed genes are expressed in the acquired cases (Figure 3.1). Gene ontology analysis with DAVID identifies multiple immune response-related terms that are enriched with genes expressed in these samples (Figure 3.2), with the cell-type specific terms predominantly represented by T-cell related activities, over other immune cells. The top canonical pathways identified by IPA further demonstrate this pattern (Figure 3.3). Performing the DE analysis on individual transcripts shows a similar pattern, where most statistically significant transcript isoforms are expressed by the acquired samples (Figure 3.4A). Focusing on lncRNA transcripts brings the number of significantly, differentially expressed isoforms closer to parity with the genetic samples (Figure 3.4B).

3.1.2 Differential Expression Analysis Across Histological Phenotypes

Since the cohort is composed of samples representing several types of myopathy, DE was tested for every pair of myopathy phenotypes, categorized by histopathological findings (myositis, mitochondrial, myofibrillar, dystrophic, or nonspecific findings). PCA was performed on the samples' raw read counts to identify the similarity between each sample, and each condition (Figure 3.5). The first principal component (explaining σ^2 = 88.486%) appeared sufficient to delineate between the acquired and genetic samples. The phenotypes within the genetic group showed visual clustering within the first three principal components (PCs; collecti-

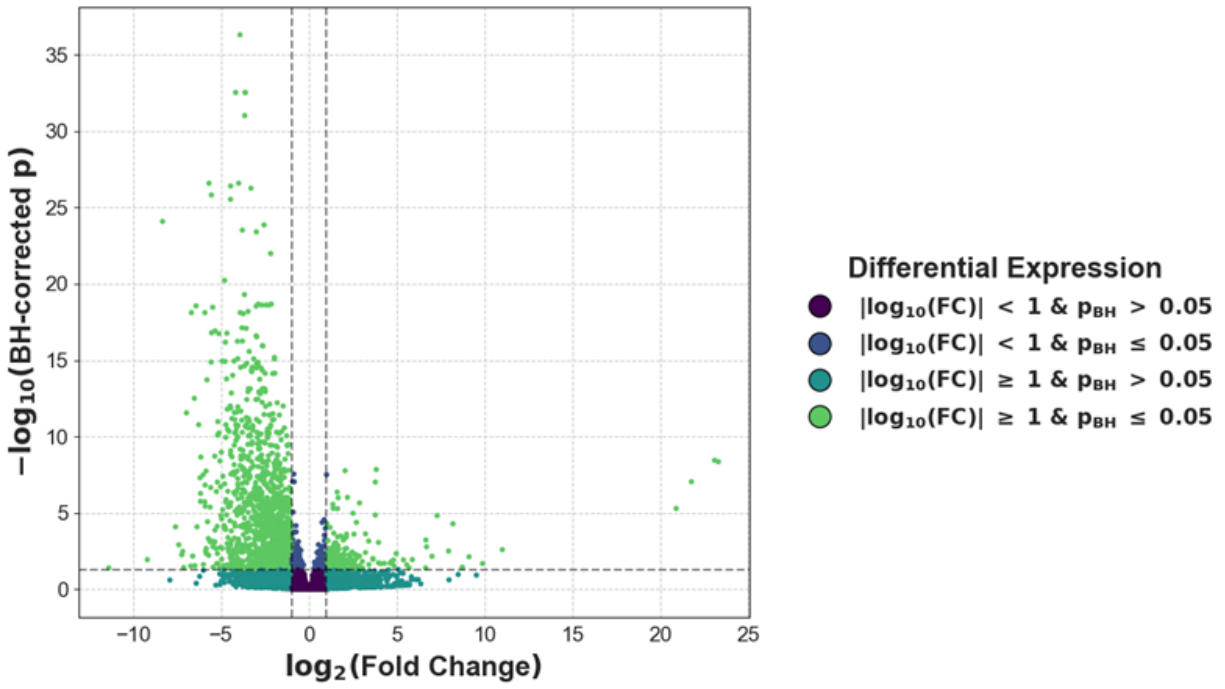


Figure 3.1: Differential expression between genetic and acquired myopathy cases. Comparison of the DE, and statistical significance for each detected gene between genetic ($log_2FC > 0$) and acquired ($log_2FC < 0$) myopathy RNAseq samples, revealing most differentially expressed genes are upregulated in the IBM cases.

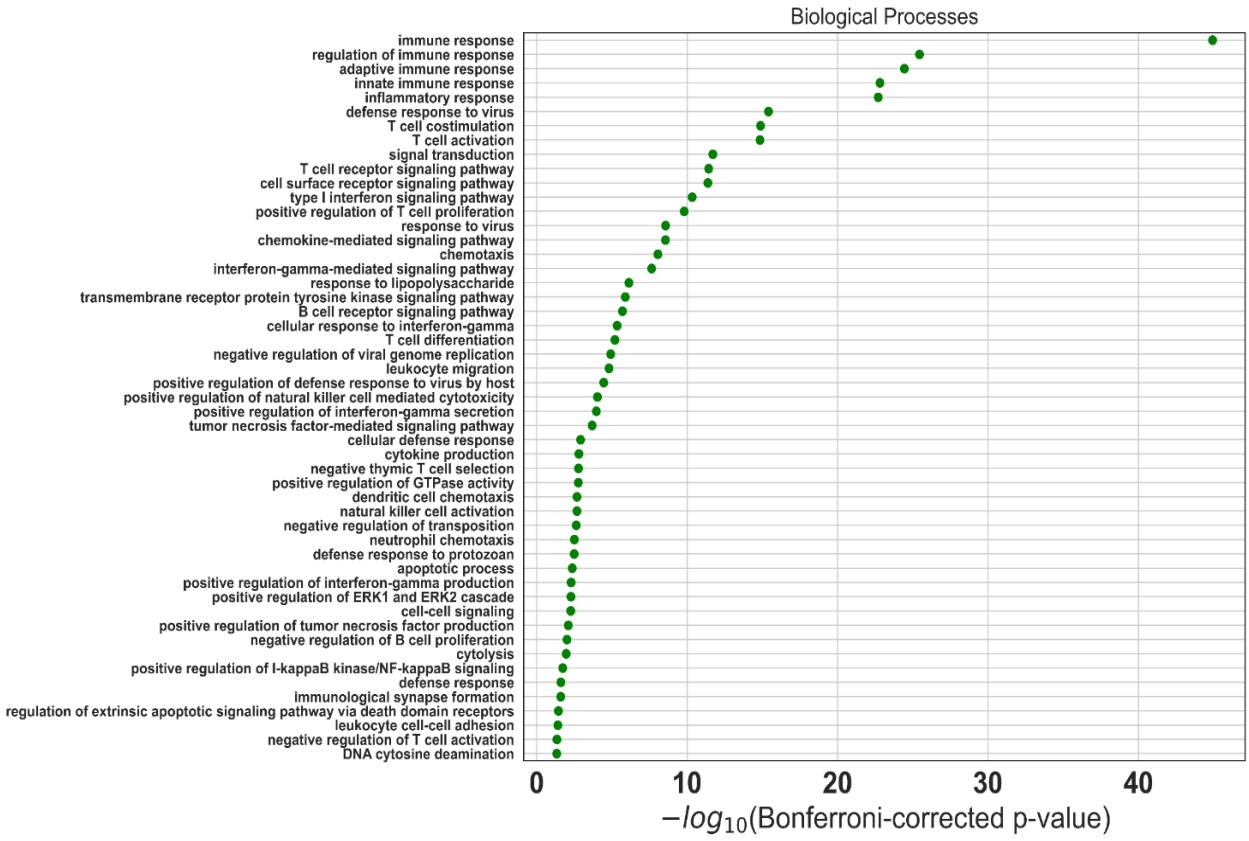


Figure 3.2: Differential expression in IBM cases dominated by immunity. Genes that were significantly upregulated in the acquired cases ($\log_2FC < 1$ & BH-adjusted $p < 0.05$) were analysed for GO enrichment by DAVID, demonstrating a statistically significant immunological influence.

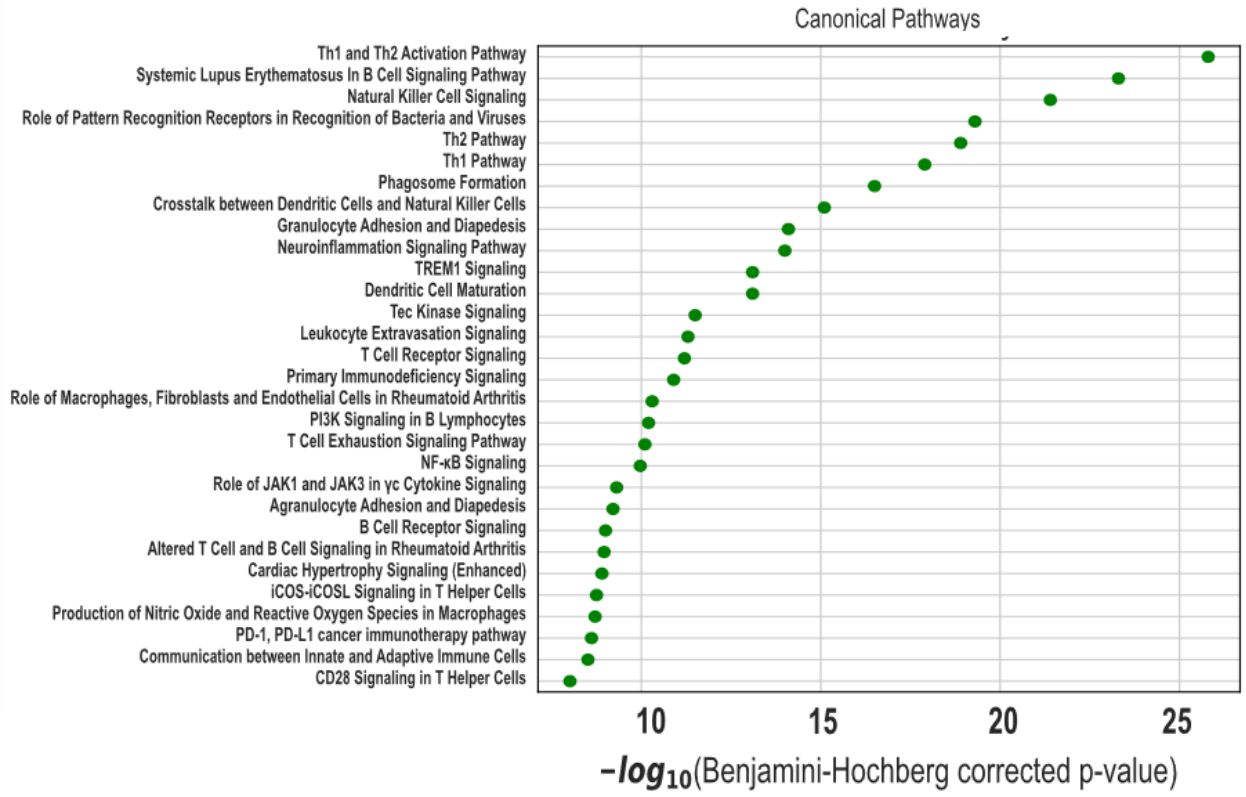


Figure 3.3: Differential expression between genetic myopathies and IBM dominated by immune pathways. Ingenuity Pathway Analysis was implemented to complement the DAVID findings, showing a similar enrichment for immunological pathways**.**

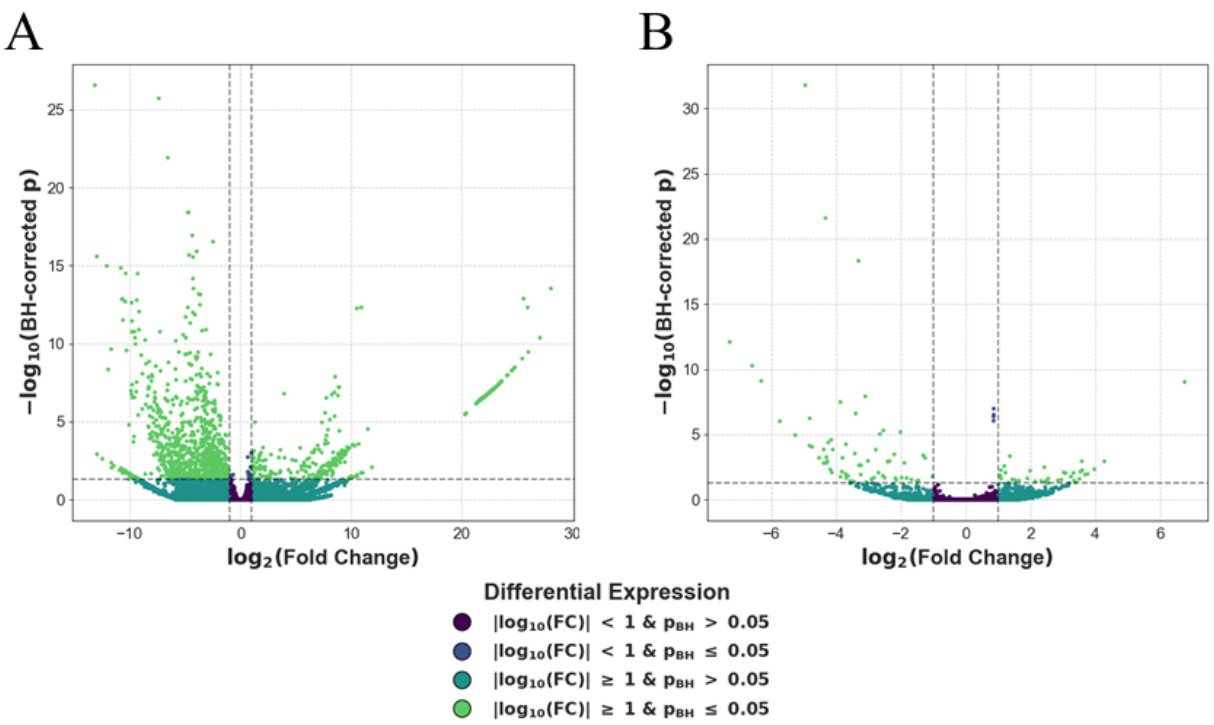


Figure 3.4: LncRNAs detected as differentially expressed between genetic cases and IBM. Comparison of the DE, and statistical significance for each detected transcript, between genetic ($log_2FC > 0$) and acquired ($log_2FC < 0$) myopathy RNAseq samples, revealing that more unique transcripts too, are upregulated in the IBM participants **(A)**. Focusing on transcripts with a non-coding designation by Ensembl, reveals the presence of statistically significant upregulation of various lncRNAs upregulated in both conditions **(B)**.

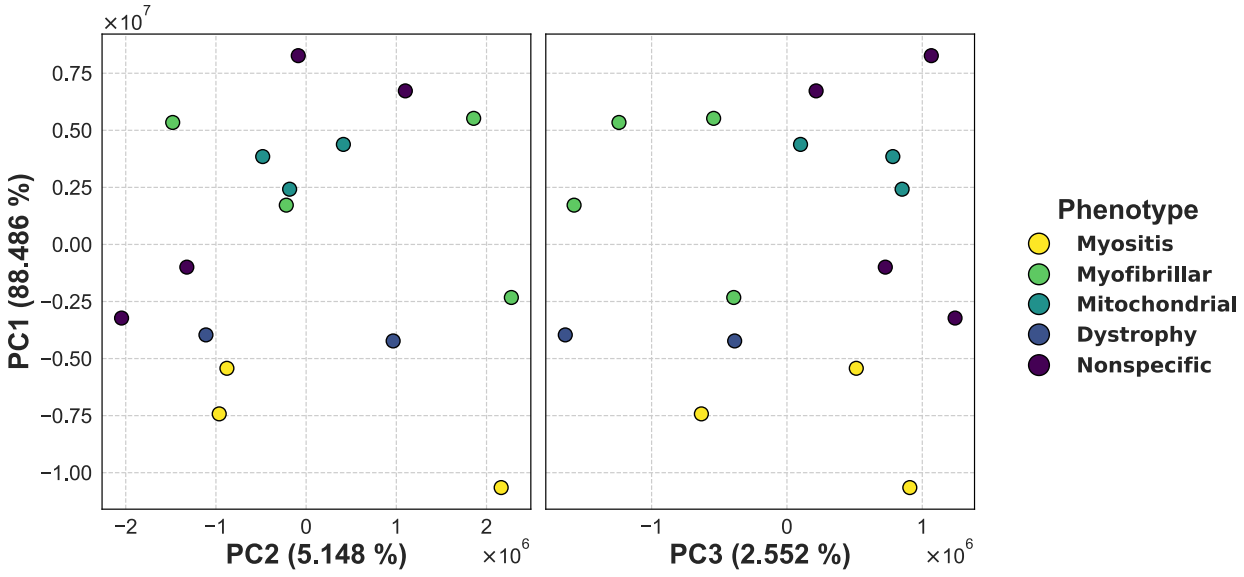


Figure 3.5: Clustering of pathology observed in PCA. The first three principal components (accounting for 96.17% of observed variance) appear to delineate between groups of samples with distinct pathological findings. HLA, and IG gene counts were omitted to prevent their polymorphic nature from interfering with clustering.

-viely σ^2 = 96.186%), with the exception of the samples with nonspecific histological findings, who clustered around the myofibrillar, and mitochondrial samples. Cross-validation across all PCs with kNN results in a maximum accuracy of 100% (k = 1) when classifying between genetic and acquired cases and 68.75% (k = 2) when classifying between the specific phenotypes.

Statistically significant differentially expressed genes were more diffuse across most of the comparisons, than in the grouped analysis (Figure 3.6), though the myositis cases demonstrated the most statistically, differentially expressed genes in each of its four comparisons. The genes that were consistently differentially expressed in at least three comparisons for each phenotype were highlighted to identify whether there were histology-specific findings (Figure 3.6). Increasing this criterion to genes upregulated in all four comparisons excluded all genes for the non-myositis samples. A similar analysis was performed focusing on lncRNAs (Figure 3.7), where each condition had at least two lncRNAs significantly differentially expressed in three of its four comparisons. Figures 3.8-3.12 summarize the genes, and lncRNAs meeting these criteria for each of the conditions, where the top 10 genes, after excluding T-cell receptor, and immunoglobulin (IG) genes due to their variable nature, and lncRNAs are shown for the myositis group (789 genes, and 46 lncRNAs were found to be consistently differentially expressed). With the exception of the myositis samples, the gene sets were not large enough to perform a GO enrichment analysis, or IPA. The mitochondrial myopathy group appears to have the most consistently upregulated genes, including *FGF21*, and *MTND6P4*, that are increased in all four comparisons. The lncRNAs show a different story, with only two, where the highest (a *JPX* transcript) is shared with the myofibrillar cases. The myofibrillar cases had the fewest number of genes and lncRNAs consistently upregulated, inclu-

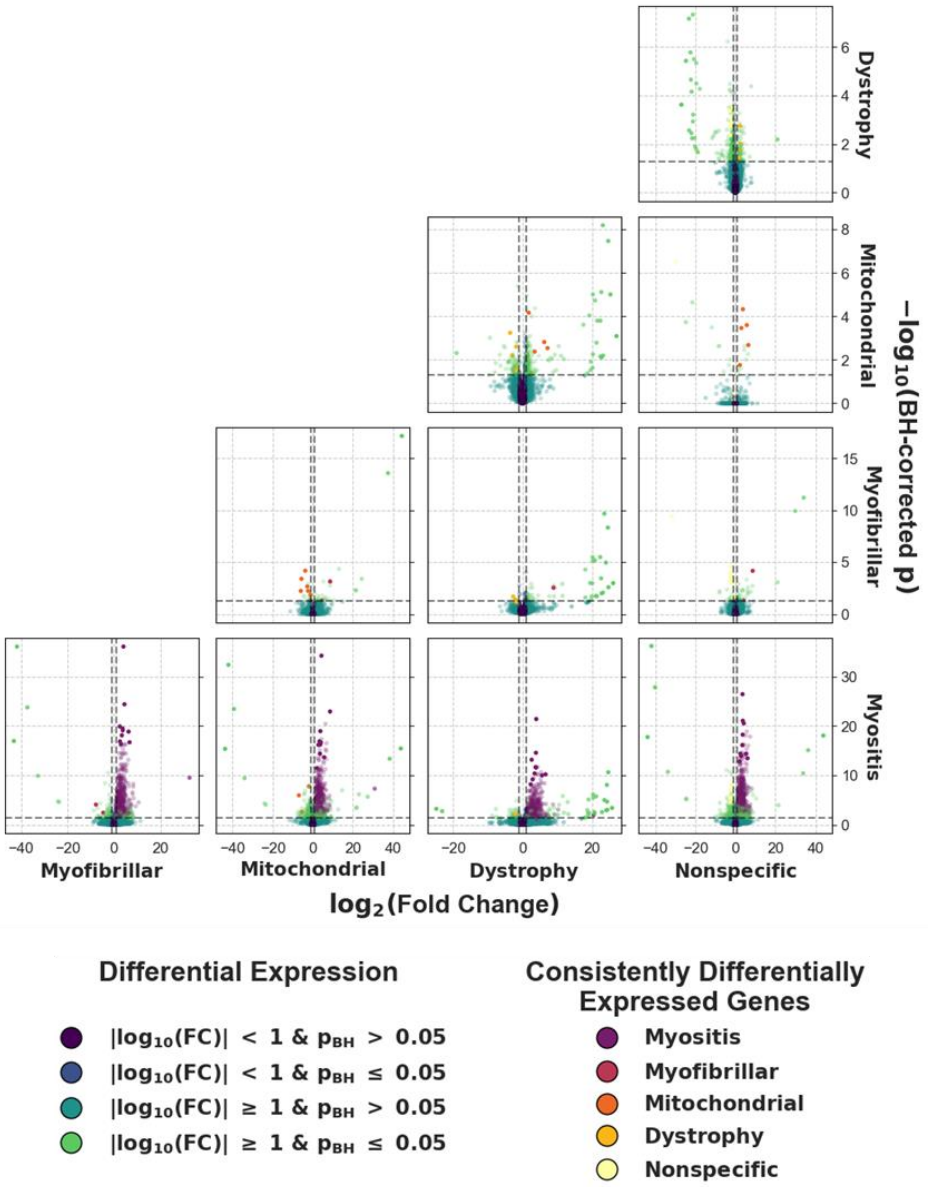


Figure 3.6: Distinct myopathies have differentially expressed genes in most comparisons. Genes that were statistically, and differentially expressed in at least 3 comparisons for a single group are highlighted.

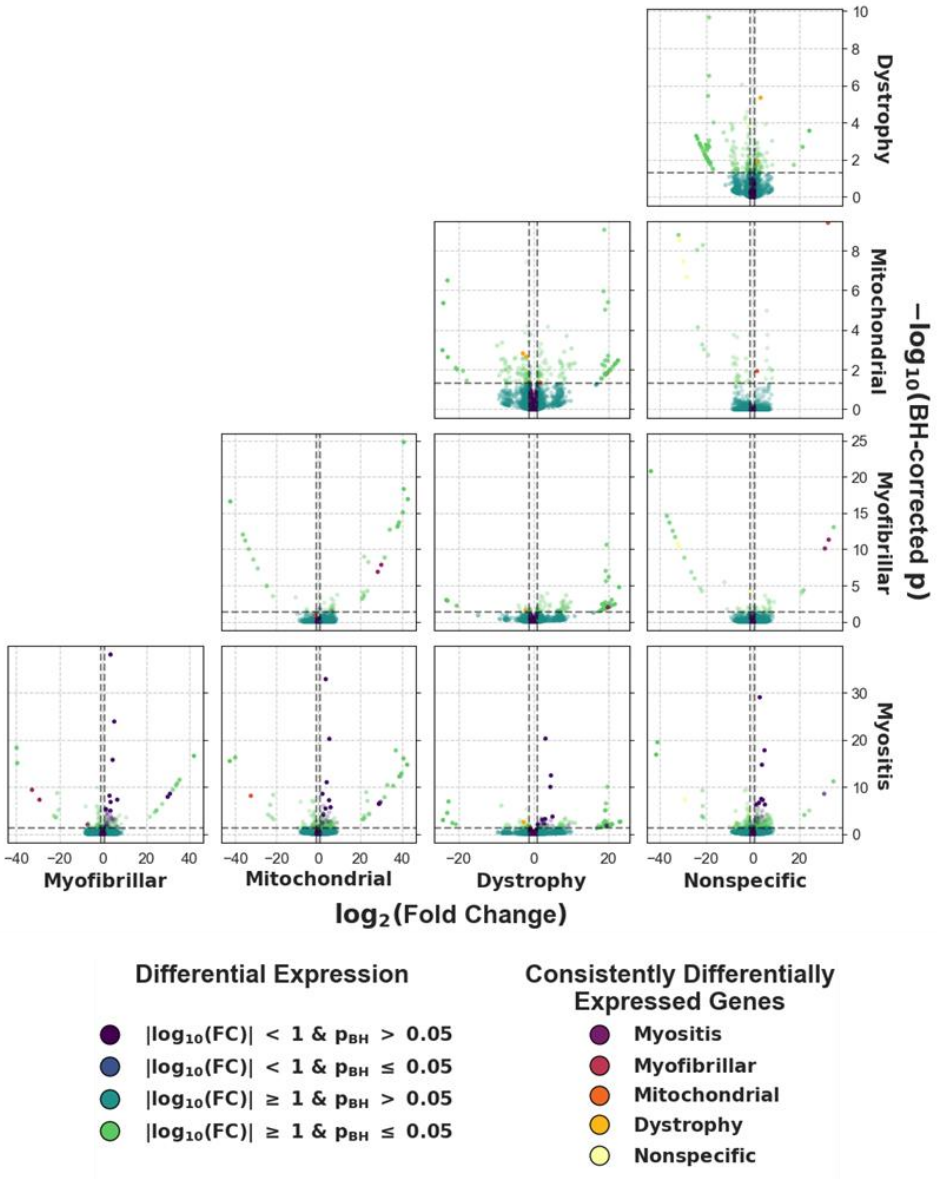
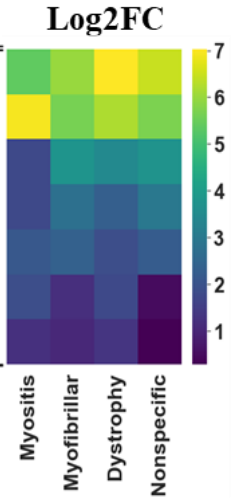


Figure 3.7: Distinct myopathies have differentially expressed lncRNAs in most comparisons. An identical comparison was made between the histologically divided groups to focus on lncRNAs.

A

Upregulated Genes
(≥3 comparisons)

Gene Symbol	Description
FGF21	fibroblast growth factor 21
MTND6P4	MT-ND6 pseudogene 4
TRIB3	tribbles pseudokinase 3
ASNS	asparagine synthetase (glutamine-hydrolyzing)
SLC13A4	solute carrier family 13 member 4
ADAMTSL5	ADAMTS like 5
MPST	mercaptopyruvate sulfurtransferase



B

Upregulated lncRNAs
(≥3 comparisons)

Transcript ID	Gene Symbol	Description
ENST00000660856	JPX	JPX transcript, XIST activator
ENST00000450589	GAS5	growth arrest specific 5

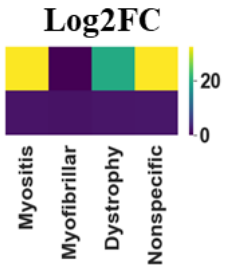
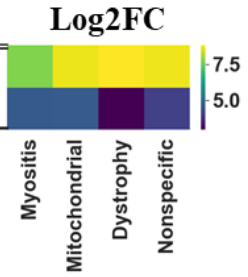


Figure 3.8: Genes and lncRNAs consistently upregulated in mitochondrial myopathy samples. Genes that were differentially, and statistically significantly expressed in the mitochondrial samples (n = 3) in at least three of the four comparisons between phenotypes **(A)**. Similarly presenting lncRNA transcripts **(B)**.

A

Upregulated Genes
(≥3 comparisons)

Gene Symbol	Description
RPL10P9	ribosomal protein L10 pseudogene 9
COMP	cartilage oligomeric matrix protein



B

Upregulated lncRNAs
(≥3 comparisons)

Transcript ID	Gene Symbol	Description
ENST00000660856	JPX	JPX transcript, XIST activator
ENST00000423477	BOLA3-AS1	BOLA3 divergent transcript
ENST00000419485	ZNRD1ASP	zinc ribbon domain containing 1 antisense, pseudogene
ENST00000427705		

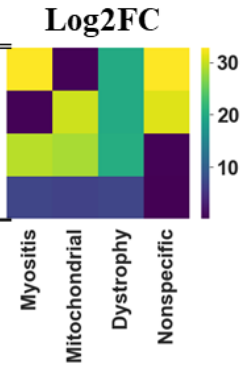
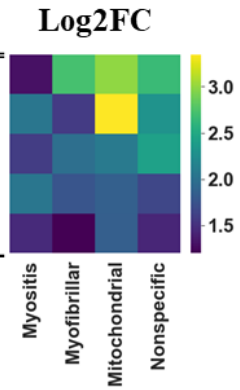


Figure 3.9: Genes and lncRNAs consistently upregulated in myofibrillar myopathy samples. Genes that were differentially, and statistically significantly expressed in the myofibrillar samples (n = 4) in at least three of the four comparisons between phenotypes **(A)**. Similarly presenting lncRNA transcripts **(B)**.

A

Upregulated Genes
(≥3 comparisons)

Gene Symbol	Description
KIAA1324	KIAA1324
-	novel transcript
ANGPT4	angiopoietin 4
-	vimentin 2, pseudogene
-	novel transcript, antisense to STEAP4



B

Upregulated lncRNAs
(≥3 comparisons)

Transcript ID	Gene Symbol	Description
ENST00000455531	MEG3	maternally expressed 3
ENST00000652827	-	novel transcript
ENST00000631117	-	novel transcript, antisense to STEAP4

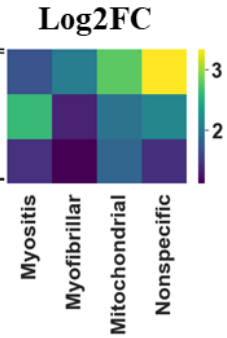
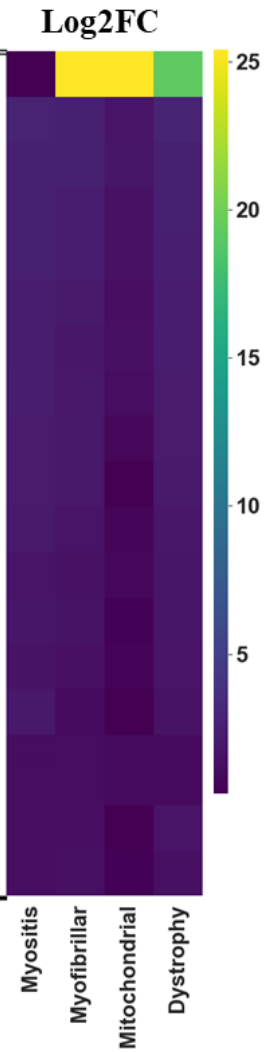


Figure 3.10: Genes and lncRNAs consistently upregulated in dystrophic samples. Genes that were differentially, and statistically significantly expressed in the dystrophic samples (n = 2) in at least three of the four comparisons between phenotypes **(A)**. Similarly presenting lncRNA transcripts **(B)**.

A

Upregulated Genes
(≥3 comparisons)

Gene Symbol	Description
RNU6-33P	RNA, U6 small nuclear 33, pseudogene
MT-ND3	NADH:ubiquinone oxidoreductase core subunit 3
MT-ATP8	ATP synthase membrane subunit 8
MTATP6P1	MT-ATP6 pseudogene 1
MT-ND4L	NADH:ubiquinone oxidoreductase core subunit 4L
MT-ATP6	ATP synthase membrane subunit 6
MT-ND4	NADH:ubiquinone oxidoreductase core subunit 4
MT-ND5	NADH:ubiquinone oxidoreductase core subunit 5
MT-ND2	NADH:ubiquinone oxidoreductase core subunit 2
MT-TL1	tRNA-Leu (UUA/G) 1
MT-ND1	NADH:ubiquinone oxidoreductase core subunit 1
MT-CO2	cytochrome c oxidase II
-	novel transcript, antisense to FUZ
MT-CYB	cytochrome b
SCARNA7	small Cajal body-specific RNA 7
WHAMMP2	WAS homolog associated with actin, golgi, and microtubules pseudogene 2
-	novel transcript, antisense to ST8SIA5
FZD9	frizzled class receptor 9



B

Upregulated lncRNAs
(≥3 comparisons)

Transcript ID	Gene Symbol	Description
ENST00000615804	-	uncharacterized LOC102724843
ENST00000666863	-	novel transcript
ENST00000419485	ZNRD1ASP	zinc ribbon domain containing 1 antisense, pseudogene
ENST00000427705		
ENST00000600669	-	novel transcript, antisense to FUZ

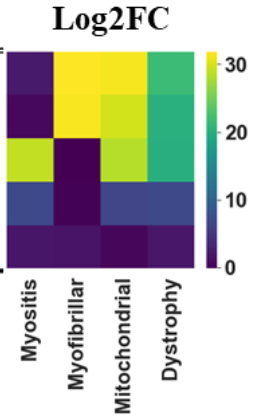
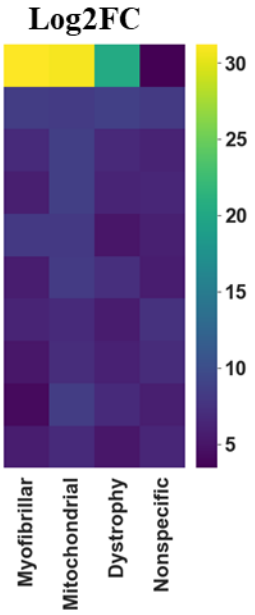


Figure 3.11: Genes and lncRNAs consistently upregulated in myopathy samples without a clear histological phenotype. Genes that were differentially, and statistically significantly expressed in the non-specific samples (n = 4) in at least three of the four comparisons between phenotypes **(A)**. Similarly presenting lncRNA transcripts **(B)**.

A

Upregulated Genes
(≥3 comparisons)

Gene Symbol	Gene Description
RNU6-33P	RNA, U6 small nuclear 33, pseudogene
NPM1P7	nucleophosmin 1 pseudogene 7
CXCL9	C-X-C motif chemokine ligand 9
UBD	ubiquitin D
CXCL13	C-X-C motif chemokine ligand 13
CD1B	CD1b molecule
CCL13	C-C motif chemokine ligand 13
ZBED2	zinc finger BED-type containing 2
ADAMDEC1	ADAM like decysin 1
GBP6	guanylate binding protein family member 6



B

Upregulated lncRNAs
(≥3 comparisons)

Transcript ID	Gene Symbol	Description
ENST00000423477	BOLA3-AS1	BOLA3 divergent transcript
ENST00000666863	-	novel transcript
ENST00000615804	-	uncharacterized LOC102724843
ENST00000467369	HCP5	HLA complex P5
ENST00000616926	-	novel transcript, antisense CCL3
ENST00000478808	USP30-AS1	USP30 antisense RNA 1
ENST00000643127	LINC02694	long non-protein coding RNA 2694
ENST00000417079	LINC00426	long non-protein coding RNA 426
ENST00000645416	LINC02694	long non-protein coding RNA 2694
ENST00000666495	HCP5	HLA complex P5

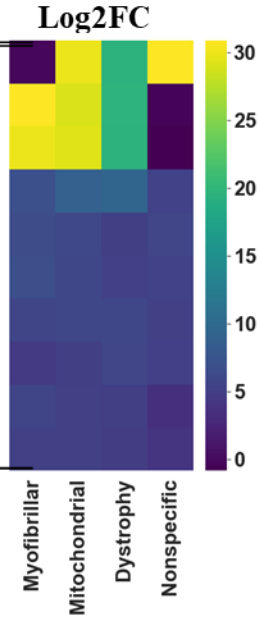


Figure 3.12: Top 10 genes and lncRNAs consistently upregulated in myositis samples. Genes that were differentially, and statistically significantly expressed in the myositis samples (n = 3) in at least three of the four comparisons between phenotypes, filtered to top 10 ranked by mean $\log_2$FC across all comparisons **(A)**. Similarly presenting lncRNA transcripts **(B)**.

-ding a ribosomal pseudogene, and *COMP*, that wasn't DE'd when compared to the dystrophic samples. In addition to the *JPX* transcript shared with the mitochondrial samples, an antisense *BOLA3* transcript shared with the myositis samples, and two *ZNRD1ASP* transcripts shared with the non-specific samples, were upregulated. The dystrophic cases had the most undescribed genes and lncRNAs consistently upregulated. It is interesting to note that many of the consistently upregulated genes are non-coding by nature, including many pseudogenes. For the non-specific pathology samples, all but one of the genes upregulated are marginally increased, and include many mitochondrial genes, that are not DE'd when compared to the mitochondrial samples. The gene with the highest average DE, is shared with the myositis cases, a small nuclear RNA pseudogene. In the myositis samples, the most exaggerated expression belonged to the snRNA pseudogene shared with the nonspecific cases, the antisense *BOLA3* transcript, mentioned with the myofibrillar cases, and two uncharacterized lncRNAs shared with the nonspecific cases. Many of the identified genes, and lncRNA's in the myositis' top 10 DE list are heavily implicated in the immune system, such as the chemokines *CKCL9*, *CXCL13*, and *CCK13*, or two HLA complex P5 lncRNAs.

3.1.3 Other Differential Expression Comparisons

Given the reality of achieving an age-, sex-matched, identically sampled cohort for analysis, possible other contributors to the variance between the samples were analysed, including sex (Figure 3.13), serum CK (Figure 3.14), and biopsied muscle (Figure 3.15).

The distribution of differentially expressed ($|\log_2 FC| > 1$) genes between the males and females of the cohort appears uniform across the genome, with the exception of the Y chromoso-

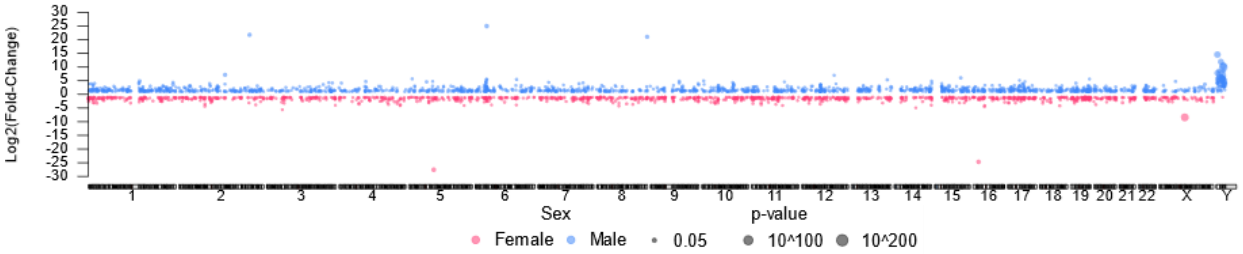


Figure 3.13: ChrY responsible for most significantly expressed genes between sexes. Manhattan plot of the log_2FC of genes observed between the males, and females of the group. The distribution of genes is relatively uniform across the genome, with the exception of the Y chromosome, accounting for the most significantly differentially expressed genes. The highly statistically significant X-chromosomal gene upregulated in females corresponds to *Xist*.

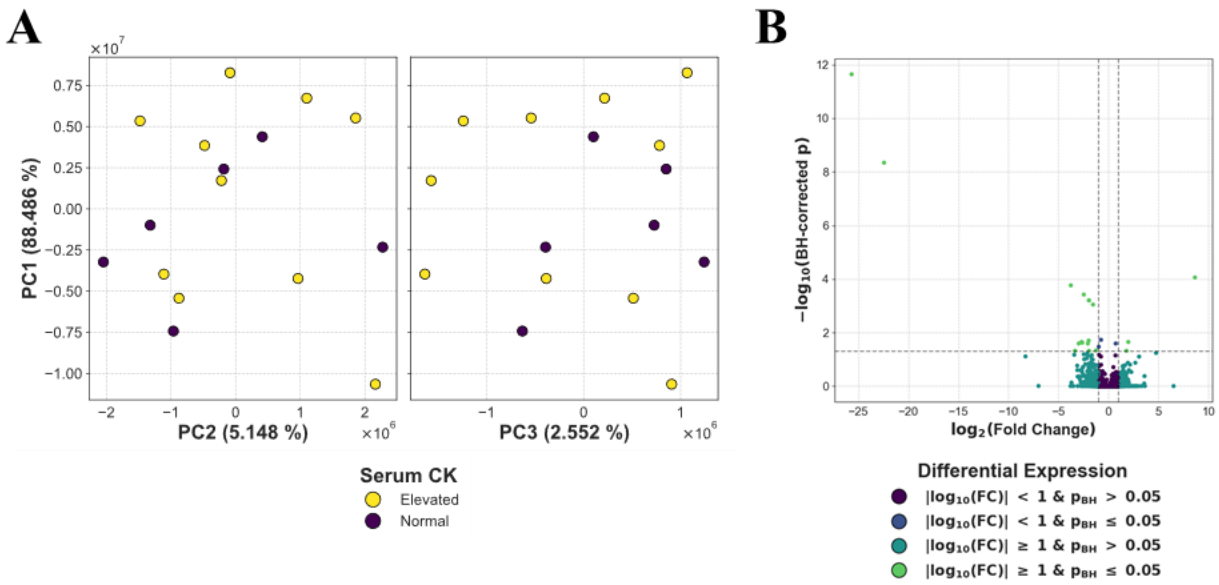


Figure 3.14: Differential expression by samples' serum CK status. Overlaying serum CK status on the first three principal components, showing minimal clustering between the normo- and hyperCKemic participants **(A).** Most genes that were differentially expressed between the normal, and elevated serum CK samples, fail to meet statistical significance **(B).**

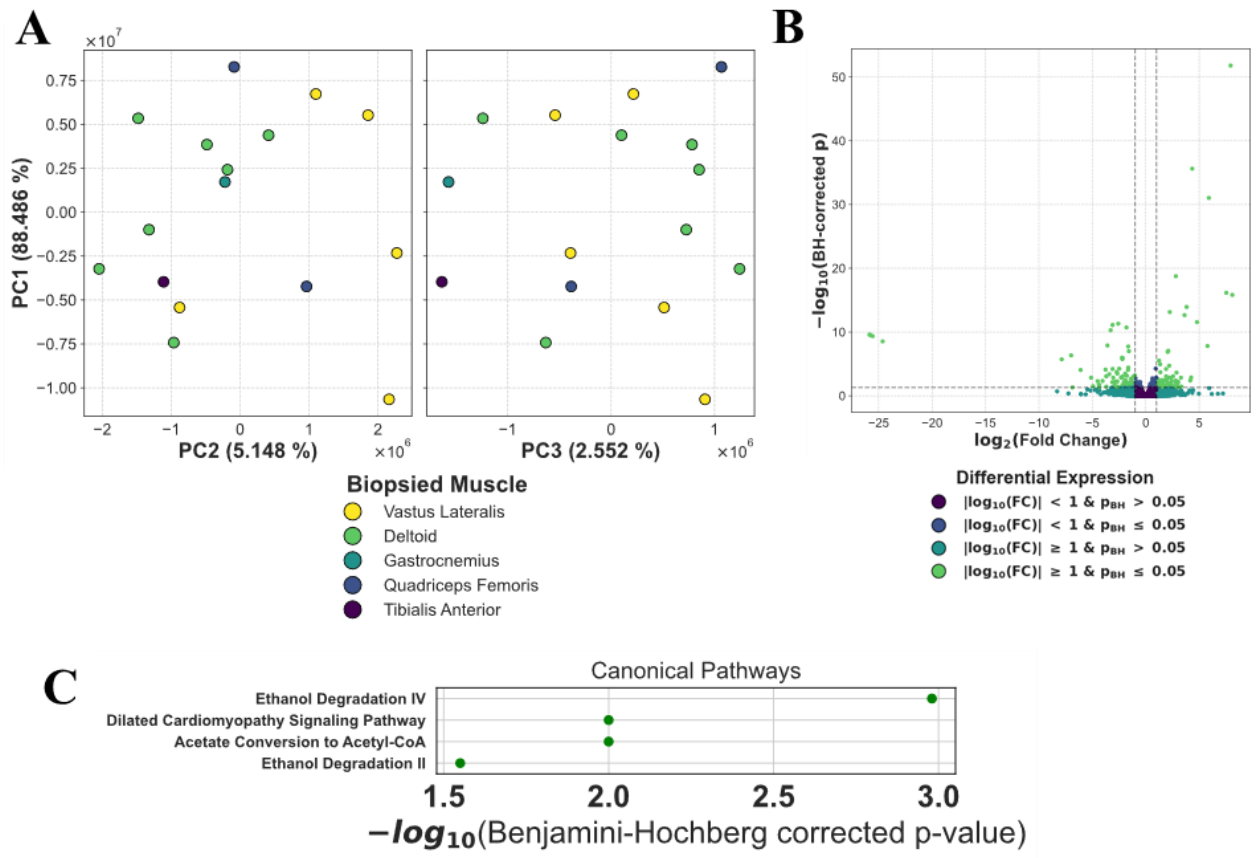


Figure 3.15: Differential expression between the biopsied muscle groups. Principal component analysis overlaid with each sample's biopsy location, showing minimal clustering between the five muscles sampled **(A)**. DE between the two most common muscles sampled in the cohort, the vastus lateralis, and deltoid, increases in the positive direction indicate increased expression in the vastus lateralis **(B)**. The distribution of differentially expressed genes appears symmetrical. IPA canonical pathway analysis reveals four pathways, enriched with statistically significantly (BH-corrected $p < 0.05$) differentially expressed ($|\log_2 FC| > 1$) genes between the vastus lateralis, and deltoid, that pass multiple hypothesis testing **(C).** Note, only "Ethanol Degradation IV" displayed a directionality (Z = -1.0), indicating genes in this pathway were predominantly expressed in the deltoid.

-me, that also accounts for the statistically significant extremes. 76 DE'd genes met statistical significance (BH p-adjusted < 0.05), however there was insufficient pathway enrichment to meet statistical significance in IPA. Accuracy of kNN cross-validation on all PCs did not exceed 62.50% (at k = 3)

Participants with elevated CK do not appear to cluster when their status is overlaid on the first three PC and had a maximum accuracy from cross-validation with kNN of 50% (at k = 3). Additionally, few genes were found to be significantly DE'd between the two conditions. There were an insufficient number of genes statistically DE'd to perform DAVID, or IPA.

Similarly, when overlaying biopsied muscle over the first three principal components, visually, there appears to be minimal clustering, and kNN cross-validation accuracy of 50% (at k = 1). However, the DE analysis between the two most frequently sampled muscles, the vastus lateralis, and deltoid, reveals many genes are statistically significantly differentially expressed. To identify whether there are any overall differences between the two muscle groups, a canonical pathway enrichment analysis was performed on the statistically significantly (BH-corrected p-value < 0.05), differentially expressed ($|\log_2 FC| > 1$) genes, with IPA, revealing four pathways passed BH multiple hypothesis testing correction: two pathways pertaining to ethanol metabolism, acetate esterification with coenzyme A, and a pathway related to dilated cardiomyopathy; only "Ethanol Degradation IV", (including the genes *ACSL1, ACSS1, ALDH1A2, TYRP1*) displayed any directionality (Z = -1.0), indicating the genes in this pathway are predominantly expressed in the deltoid.

3.2 Discussion

3.2.1 Inclusion Body Myositis Transcriptome Dominated by Immune Infiltration

Perhaps unsurprisingly, the majority of GO terms, and IPA canonical pathways enriched with genes DE'd in the acquired, myositis samples are immunological in nature. A similar result has been observed recently, as seen in a bioRxiv preprint from Johari *et al.* (2021), where the top IPA pathways enriched in their comparison between IBM, a cohort of tibial muscular dystrophy participants, and histologically normal controls, included terms found in the present study: dendritic cell maturation, T-cell receptor signalling, T-cell exhaustion signalling, and iCOS, CD28 signalling [53]. However, the enrichment of apoptotic pathways, and explicit mention of calcium ion mobilization were not observed in the present study, and we observed signals from more B-cell oriented pathways. There are differences between the approaches. Here transcript, and gene abundances were estimated with a pseudoaligner, Salmon, whereas Johari *et al.* devoted effort to physically aligning their reads to a genome with STAR, before simply counting with featureCounts[53], where the lack of accounting for 5'-, 3'-, and GC-biases from sequencing and library prep, or mis-mapping of repetitive reads may contribute to some of the variance. Alternatively, where IBM samples were compared strictly to other myopathic cases in this study, Johari *et al.* included histologically normal muscle from amputees, where the muscle may not contain increased amounts of apoptosis-related transcripts that could artificially enrich IPA's T-cell specific apoptosis pathway. This also highlights a concern with gene set enrichment-like analyses, where genes can be implicated in a variety of different pathways, the enrichment of pathways may not necessarily mean the gene is fulfilling its role in that pathway alone, or at all. It is promising, however, that there are still some pathways that survived independent replication.

In the vein of verifying the GO, and IPA results, there were attempts to perform a cell-type deconvolution on the samples, to confirm whether there was an increased immune cell influence on the IBM transcriptomes. However, there appears to be a lack of a comprehensive single cell (scRNAseq)/single nuclear RNAseq (snRNAseq), or even cell-type specific bulk tissue RNAseq datasets publicly available for muscle. An adequate dataset for cell-type deconvolution would contain all cells that could be present in a tissue sample[59], for a muscle biopsy that could include type I & type II fibres, satellite cells, endothelial cells, adipose, fibroblasts, pericytes & smooth muscle from the local vasculature, and of course both myeloid, and lymphoid lineage immune cells. When it comes to scRNAseq, the issue with skeletal muscle is mature myofibrils being syncytia, meaning single muscle cells cannot be isolated. There is work being done to remedy this problem with snRNAseq, where individual nuclei are instead sequenced[60,61]. Indeed, there is already evidence of nuclear heterogeneity within the muscle, and murine dystrophic models[61]. As snRNAseq datasets are released publicly, muscular cell-type deconvolution will become feasible for experimental purposes, and perhaps even as a clinical tool to reinforce histopathological findings, as a substitute for immunohistochemical studies for identifying present cell types, or even understanding the nuances of the many compartments within muscle . A day may come when diseases like IBM are subdivided, and treated based on the heterogeneity of the cells present.

3.2.2 Genes Differentially Expressed Across the Myopathies

While most of the histological groups had few upregulated genes, many pseudogeneous or undescribed, there are some promising results. First, among the top 10 genes upregulated in the myositis cases, are several chemokine ligands that are implicated in immune function.

Including *CXCL9*, which has been previously associated with IBM[62]. Second, is *FGF21* found consistently, and relatively highly, in the mitochondrial myopathy group. *FGF21*, or fibroblast growth factor 21, is traditionally known to influence lipid and sugar metabolism, encouraging ketogenesis, mobilizing lipids, conversion of white to brown adipose, and influencing appetite[63]. In a pathological setting, FGF21 may be released following metabolic stress, causing muscle atrophy, or osteopenia[64]. While typically released by the liver during fasting conditions, it appears MtM muscle recapitulates a starving-like response during disease progression, explaining the secretion of FGF21[63,65]. Interestingly, in a similar RNAseq study, comparing healthy controls, and progressive external opthalmoplegia (PEO; a manifestation of MtM), Forsstrom *et al.* (2019) also saw increases of asparagine synthetase (*ASNS*), and Tribbles pseudokinase 3 (*TRIB3*), both of which were found consistently upregulated in our MtM cohort[63]. While the differences in methodology (healthy vs. PEO, rather than between myopathies, and abundance estimation with HTseq, rather than salmon) prevents direct comparison of the FC of these genes, the effect follows a similar pattern, where *FGF21* has the highest degree of expression, followed by *ASNS*, and *TRIB3*[63]. This effect was observed in disease caused by heteroplasmic mtDNA deletions[63], which could help diagnose the mitochondrial cohort. At the very least, these findings suggest we are in fact capturing truly pathognomonic signals in these DE analyses.

3.2.3 LncRNAs Differentially Expressed in Different Myopathies

While sparse, each myopathy type present in this study had lncRNAs DE'd when compared with each other. It is interesting to note that both the mitochondrial, and myofibrillar cohort have a *JPX* transcript upregulated, when its function is related to X-chromosome

inactivation in females[66,67]. Particularly when there was only a single female with mitochondrial disease in this cohort, and none with myofibrillar pathology. Though, there are situations beyond X inactivation where *JPX* transcripts was found to be functional[68]. However, Ensembl lists several dozen *JPX* transcripts annotated by Havana, where the transcript found upregulated in the mitochondrial, and myofibrillar cases (ENST000000660856), does not have any RNAseq-level supporting its existence; it is possible this finding is simply an artefact from the transcriptome used for quantification. The other lncRNA found upregulated in the mitochondrial samples is ENST00000450589 in *GAS5*. While it was found with lower $\log_2$FC, in each comparison, it was found upregulated compared to each of the four other groups. The function of *GAS5* does not appear to be fully known, where roles were found in macrophage mortality in atherosclerosis[69], colorectal cancer progression[70], and modulating the citric acid cycle[71], among others. The latter is an interesting coincidence, given this transcript was found upregulated in the samples with mitochondrial disease. However, given the diversity of functions this one gene is associated with, it is plausible its functions are transcript dependent, and without knowing which transcripts were implicated in which function, it is possible this finding is little more than a coincidence. The last upregulated lncRNA found in the literature is *MEG3*, found upregulated in the dystrophic cases. *MEG3* transcripts have been found to have conserved motifs that can stimulate p53[72]. There has been a relationship observed between p53, and activating *DUX4* in FSHD[73], so there may be a functional relationship here, though it requires further investigation.

3.2.4 Impact of Confounding Variables

Given the limitations of a small cohort, it is important to ensure the signal identified in the previous experiments were due to pathological changes and not from other variables. The

three that had sample sizes large enough to test were biological sex, serum CK, and biopsied muscle. The metric adapted to compare similarity between the samples was the accuracy of a cross-validation kNN, trained on all PCs. None of the three covariates could achieve a higher accuracy as the disease classifications, particularly when the genetic cases were grouped together. While there were examples of genes being upregulated in either sex across the genome, only 76 were statistically significant, and those did not enrich any IPA pathways, suggesting sex had a minimal impact on the DE analyses. The significance of *Xist* in females, and the Y-chromosomal genes in males, lends legitimacy to our findings. An interesting observation is that when delineating the samples by histology, the non-specific group had the most females (Fisher's exact $p = 0.0632$, when compared with the rest of the cohort with specific pathology). Whether this is more than a coincidence however is unknown, as the sample size is limited. While there have been transcriptomic studies comparing female, and male muscles[23,24], we were unable to replicate any findings. This could be due to several factors, including their disease state muddying their transcriptome, or the differences in age across the cohort (54.2 ± 10.8 for females, and 46.6 ± 19.2 for males, mean age $\pm \sigma$).

Serum CK activity is considered a highly non-specific indicator of skeletal, and cardiac muscle pathology, however, we thought it might be interesting to identify whether there were any transcriptomic differences between those who were hyperCKemic, and not. There are several CK isoforms found in humans, distinguished by subunit composition (muscle- or brain- type CK)[74]. We hypothesized at first that those with elevated CK activity would be actively transcribing, and translating CK genes (*CKM, CKB*, for the muscle, and brain specific CKs), however neither were elevated. This doesn't deny that scenario as a possibility, as transcription

does not necessarily correlate with translation. There also exists a heterodimer composed from both CKM, and CKB[75], changes to these quaternary protein structures would also not be visible at the transcript level. The insignificance of the CK findings is not totally unexpected given its variability based on severity, and disease. It is also possible, given CK levels oscillate, that the highest recording did not capture the true CK peak for some of the participants. What can be said, however, is the minimal impact, if any, serum CK status had on the overall DE analysis.

The last covariate tested in this cohort was the source of biopsied muscle. Two groups had enough samples to reliably perform a DE analysis, the deltoid, and the vastus lateralis of the quadricep. This comparison had the largest distribution of genes differentially expressed, among the confounding variable tests. However, it seems the impact muscle group had on the DE between pathologies may still be minimal. First, samples did not cluster near similarly sampled muscles when overlaid on the PCs. Second, few pathways were found to be enriched with the genes DE'd between the two, with only one actually showing a directionality. Further, 3/4 pathways found to be enriched by IPA, were all associated with short-chain carbon metabolism (either ethanol, or acetate), suggesting the differences between the muscle groups may simply be nuances to their metabolism. Much like sex, however, there has been minimal investigation into the transcriptomic differences between muscle groups. Considering with most of the myopathies described the muscles impacted vary by etiology, if muscle groups were in fact identical, should they not experience the same pathology? Or perhaps more accurately, should pathology not correlate with environmental influences like weight-bearing or use?

Chapter 4 – Transcriptomics in variant prioritization of genetic myopathy

4.1 Results

4.1.1 Variant Prioritization & Reclassification of Clinically Identified Variants

For the 13 presumed genetic myopathy participants in the cohort, 8 had candidate variants detected by clinically ordered targeted sequencing (Table 4.1). The participants with mitochondrial histopathology were not referred for clinical sequencing, and one myofibrillar, and one nonspecific participant, did not have any variants reported. To summarize the collective findings, the 8 samples had between 1, and 8 variants considered pathogenic or of unknown significance. Two of the variants have not been reported in NCBI's dbSNP. All variants, except for one (rs781353247), were found to be heterozygous. All but two (rs1370166904 & rs762500701) had predicted protein-coding consequences, mainly missense, though two of which were predicted to cause frameshifts (again rs781353247 in 1255, and also rs368104077 in 1523), and an inframe deletion (rs794727697). Three are listed as pathogenic in ClinVar (rs764698870, for AR diseases including: Nonaka myopathy, sialuria, or GNE myopathy, rs794727697, for AR limb-girdle muscular dystrophy 2A, and rs368104077, for Eichsfeld congenital muscular dystrophy, again presumed to be AR), with the remainder considered uncertain, conflicting, or are otherwise absent. The most frequent variant (rs139576982, in 1021) had an AF of 189/282304 in gnomAD's exomes, and genomes.

Given that the participant's families have not been sequenced, applying systematic criteria like ACMG, that heavily prioritize pedigrees, and variant segregation, can be difficult.

Table 4.1: Variants identified by clinically ordered targeted sequencing.

Participant ID	Histology	Variant dbSNP ID (If available)	Genotype	HGNC Symbol	HGVS	gnomAD	ClinVar
1308	Myofibrillar	rs764698870	Het.	*GNE*	p.A555V	12/251492	pathogenic
1255	Myofibrillar	rs781353247	Hom.	*MYOT*	p.A125Lfs*5	2/282802	uncertain
		rs571902899	Het.	*CACNA1S*	p.V923M	5/250190	uncertain
		rs148865136	Het.	*LRP4*	p.I1728V	99/282352	uncertain
		rs782308478	Het.	*PLEC*	p.H4192L	0/246720	uncertain
1373	Dystrophic	rs1370166904	Het.	*TNPO3*	c.2711+5G>A	1/251150	-
		rs777919963	Het.	*DYNC1H1*	p.D464N	5/251326	uncertain
		rs774035582	Het.	*COL12A1*	p.L1262S	7/248900	-
		rs780100460	Het.	*PLEKHG5*	p.G963W	4/181528	-
1021	Dystrophic	rs139576982	Het.	*PNPLA2*	p.R79Q	189/282304	conflicting
1105	Myofibrillar	rs769744438	Het.	*RYR1*	p.E1175K	7/282856	uncertain
		rs368686970	Het.	*COL12A1*	p.R2314W	6/241190	uncertain
		rs952020807	Het.	*DES*	p.E434K	2/201122	uncertain
		rs772640415	Het.	*GFPT1*	p.K640R	2/251428	uncertain
		rs762500701	Het.	*MUSK*	c.754-6T>A	7/247522	uncertain
		rs201291446	Het.	*NEB*	p.R8252H	28/237930	uncertain
		3-14130877-C-G	Het.	*TMEM43*	p.S73C	0/152228	uncertain
1515	Nonspecific	rs794727697	Het.	*CAPN3*	p.K254Δ	10/251418	pathogenic
1524	Nonspecific	11-22274612-T-A	Het.	*ANO5*	p.V760D	0/250838	-
		rs767777879	Het.	*GNE*	p.K26N	11/250546	-
1523	Nonspecific	rs368104077	Het.	*SELENON*	p.N238Kfs*?	45/280576	pathogenic

ACMG does ascribe significance to experimentally confirmed biological impacts, the question is whether a high through-put experiment could contribute sufficient information to utilize this line of evidence. To that end, the clinically identified variants were classified under the ACMG guidelines, with the information available (including gnomAD allele frequencies, computational variant impact predictions, courtesy of CADD, and review of OMIM, and ClinVar). The impact of RNAseq on the variants' priority was determined by a list of criteria: first, is the variant detected? Specifically for predicted splice variants, the absence of the variant was considered sufficient to assume splicing efficiency was not impacted, and the variant was likely benign. Second, were there any transcript isoforms where the participant in question deviated from the remainder of the cohort? The designation changes for clinically identified variants are summarized in Table 4.2. For missense variants, insignificant deviations were not considered sufficient to warrant changing the variants designation. Cases where this second criterion increased our certainty in variant pathogenicity include a homozygous *MYOT* frameshift variant, where the sample's expression is dramatically reduced from the rest of the cohort, and a heterozygous missense *NEB* variant, where a large transcript is increased relative to the cohort. These two cases will be elaborated on in the following section. In total, two variants had their pathogenicity designation elevated, and four were lowered, one likely pathogenic variant under ACMG, to VUS, and four VUS to benign. The likely pathogenic variant was downgraded on account of the ambiguity of the coverage of the area (Figure 4.1). Variant rs794727697 in participant 1515 is an inframe deletion of three nucleotides from *CAPN3*. However, this variant was not called by HaplotypeCaller from RNAseq. Further investigation of the exon the variant was supposed to be found in, shows the deletion is one half of a GAAGAA, where a reduction in coverage for the second GAA is observed but well within the range of coverage observed throug-

Table 4.2: Clinical variants with ACMG-designation change

Patient ID	Variant (dbSNP ID)	Designation Change[a]	Detected by RNAseq
1255	rs781353247	L. Pathogenic → Pathogenic	Yes
1373	rs1370166904	VUS → Benign	No
	rs780100460	VUS → Benign	No
1105	rs762500701	VUS → Benign	No
	rs201291446	VUS → L. Pathogenic	Yes
1515	rs794727697	L. Pathogenic → VUS	No

[a] Pathogenicity for participant, not necessarily in general

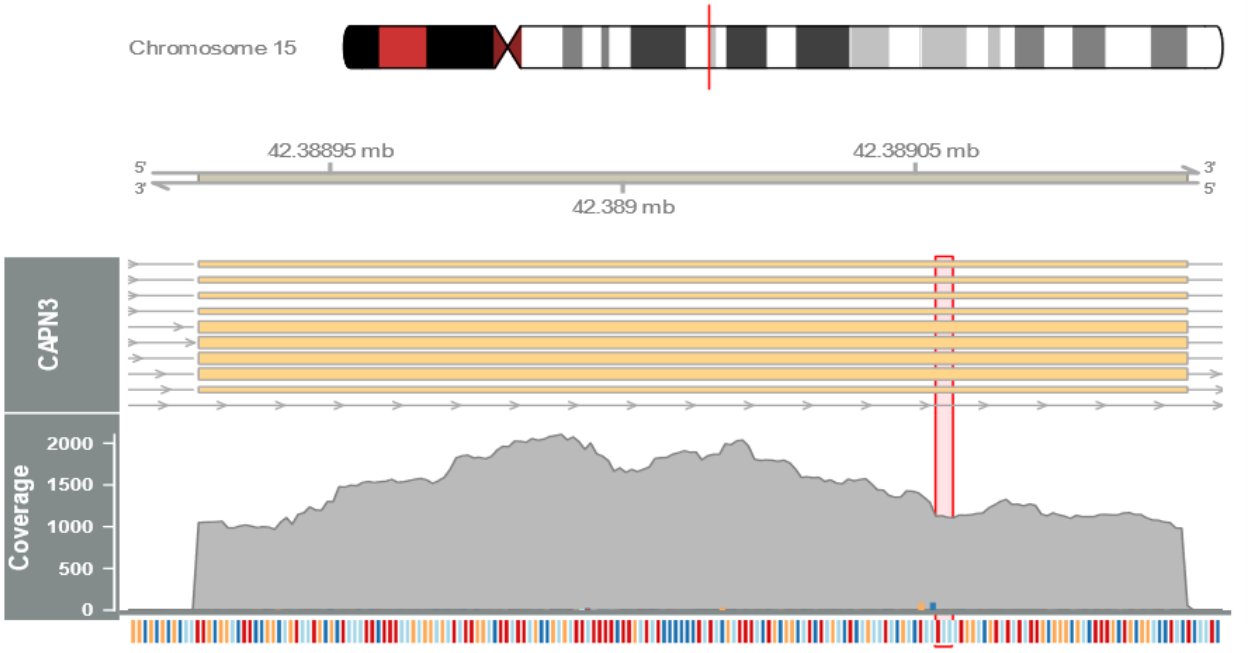


Figure 4.1: Ambiguous coverage of *CAPN3* variant. Clinical sequencing reported a variant in *CAPN3* for participant sample 1515, CAPN3:p.K254Δ, that was not called from RNAseq. Coverage of the region, with the reported variant highlighted shown.

-hout the exon. No other variants in *CAPN3* were identified in sample 1515.

It is important to note that most VUS would remain such under ACMG criteria, as the information like CADD scores, or population allele frequencies are static facts, and won't change by transcript level findings in an individual, and there are fewer supports for benignity, than for pathogenicity. As such, VUS where a benign support criterion was assigned after the RNAseq, are listed as benign for the purposes of illustration.

While the ACMG guidelines provide a framework for variant prioritization, with large genomic datasets, it can prove impractical for clinicians, and geneticists to classify each variant accordingly. Instead, the principals described by ACMG may be used to narrow down a list of variants to the most plausible candidates, ideally one that is more digestible. Criteria which can be used to quickly parse through a list of variants include allele frequency, where a threshold can be applied, or ordered to provide some degree of priority, coding sequence consequences, where frameshifts, or missense variants will be easier to interpret than untranslated region variants for example, and lastly, focusing on genes implicated in similar conditions. The combinations of these criteria were tested on the variants identified from sequencing this cohort. Including a criterion pertaining to transcript expression, where variants in genes containing an isoform that exceeded the 95%-ile CI of the cohort ($|Z| > 1.96$), were retained, to test whether the additional information available to RNAseq can be useful in variant prioritization (Figure 4.2). Filtering on variants either absent from gnomAD, or present at a frequency less than 1/1000, proved redundant for this dataset, where most variants were retained. Particularly as another population-related criterion was included, where variants were only considered if they were unique to the

individual, compared to the others in the cohort. Other strategies were where the gnomAD AF criterion was replaced with pathogenic CADD scores ($CADD_{PHRED} \geq 12.0$), or one narrowing our search to variants that are predicted to cause protein-coding changes. Narrowing the search space down to the 153 genes previously associated with myopathies, naturally provided the smallest list of variants when focusing on a single filtering criterion. Lists with fewer than 10 variants on average were achievable when using all 4 of the filtering criteria, for either CADD scores, or coding variants, though some samples had empty lists.

4.1.2 Case Studies

Ultimately, the goal is to find genetic diagnoses for the samples described here, and others. Of the thirteen samples that are presumed to have a genetic etiology for their disease, four cases will be described here, illustrating their disease, and the most plausible variant(s) identified.

First, participant 1308 presented with an asymmetrical foot drop at 46, ultimately progressing until orthoses were necessary for walking by age 53. Upper body, axial, and bulbar musculature was spared from weakness. Strength deficits were observed for dorsiflexion, but not plantarflexion, at the ankle, and for extension, and flexion at the toes. Most recent serum CK was elevated, at 632 U/L. A muscle biopsy was obtained from their right gastrocnemius, with histology showing neurogenic atrophy, with αB-crystallin-(+), dystrophin-(+), and desmin-(+) inclusions, and myopathic features, suggestive of a myofibrillar myopathy, with eosiniphillic aggregates, rimmed vacuoles, the aforementioned accumulations, and myofibrillary disarray, Z-

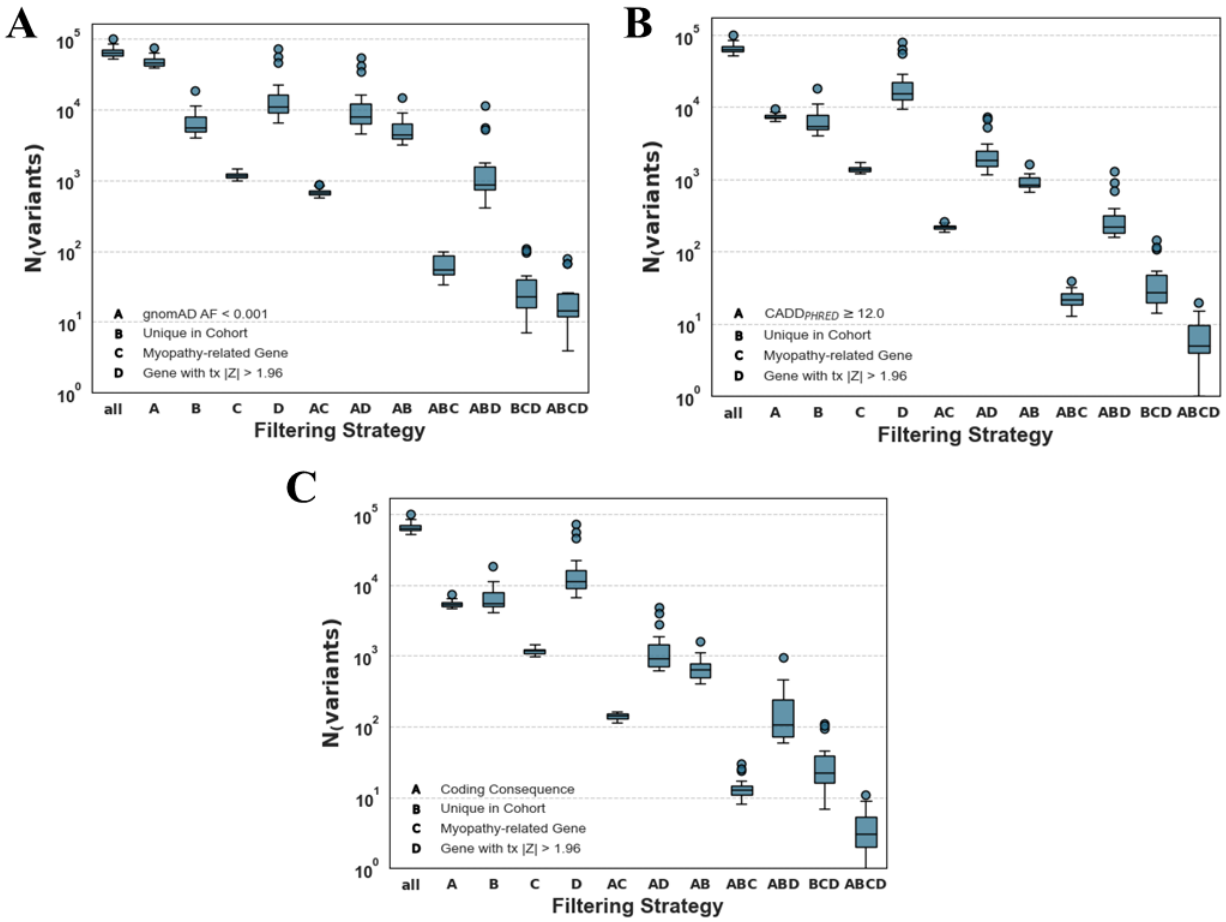


Figure 4.2: Use of transcript counts as a feature for manual variant prioritization. The resulting search space for variant prioritization from various combinations of ACMG-consistent filtering strategies and for inclusion with a metric to identify variants within genes with a dysregulated transcript, determined by exceeding the 95% confidence interval for each individual transcript, calculated from all samples (*n = 16*; **A**). As gnomAD allele frequencies were not informative for this dataset, alternative strategies were employed. Including a focus on variants that are predicted pathogenic by CADD **(B),** or predicted to impact protein coding sequences **(C)**.

line streaming, and accumulations of granular material in the sarcoplasm. Muscle MRI shows fatty infiltration of the semimembranosus (SM), semitendinosus (ST), and biceps femoris (BF), a pattern reminiscent of *FLNC*-based myofibrillar myopathies[27]. Clinical genetic testing identified a single heterozygous variant, rs764698870 (GNE:p.A555V), in *GNE*, that is reported as pathogenic in ClinVar, though with several cases suggesting a recessive, or compound heterozygous inheritance pattern[76-78]. This variant was also identified by RNAseq, however, another plausible variant, rs201905890 (FLNC:p.E534K), in *FLNC* was considered, accompanied with increased expression of protein-coding isoforms, that passed the stringent variant prioritization criteria discussed previously (Figure 4.3).

Second, participant 1255 presented with weakness in the legs originating in childhood, before progressing to requiring a cane to talk by age 34. Distal muscle wasting and weakness was observed in both legs upon examination. Serum CK was normal (<195 U/L). Vastus lateralis biopsy revealed myopathic changes, with variation in fibre size, with occasional small, rounded type 1 fibres, rare optically clear, or granular sarcolemmal vacuoles, and rare fibres with desmin-(+), and dystrophin-(+) inclusions. A muscle MRI is unavailable for this individual. Clinical sequencing identified a homozygous single nucleotide deletion, rs781353247 (*MYOT*:c.372delA. MYOT:p.A125Lfs*5) in *MYOT*, among three other heterozygous missense VUS in *CACNA1S*. *LRP4*, and *PLEC*. The *MYOT* variant was also detected in the RNAseq, and two of the three transcript isoforms of *MYOT* were found to be significantly reduced in this individual (Figure 4.4). The variant is found at the beginning of an early exon, within the coding region of the two deficient isoforms, but prior to the coding region of the third transcript, suggesting the frameshift may be promoting nonsense-mediated decay.

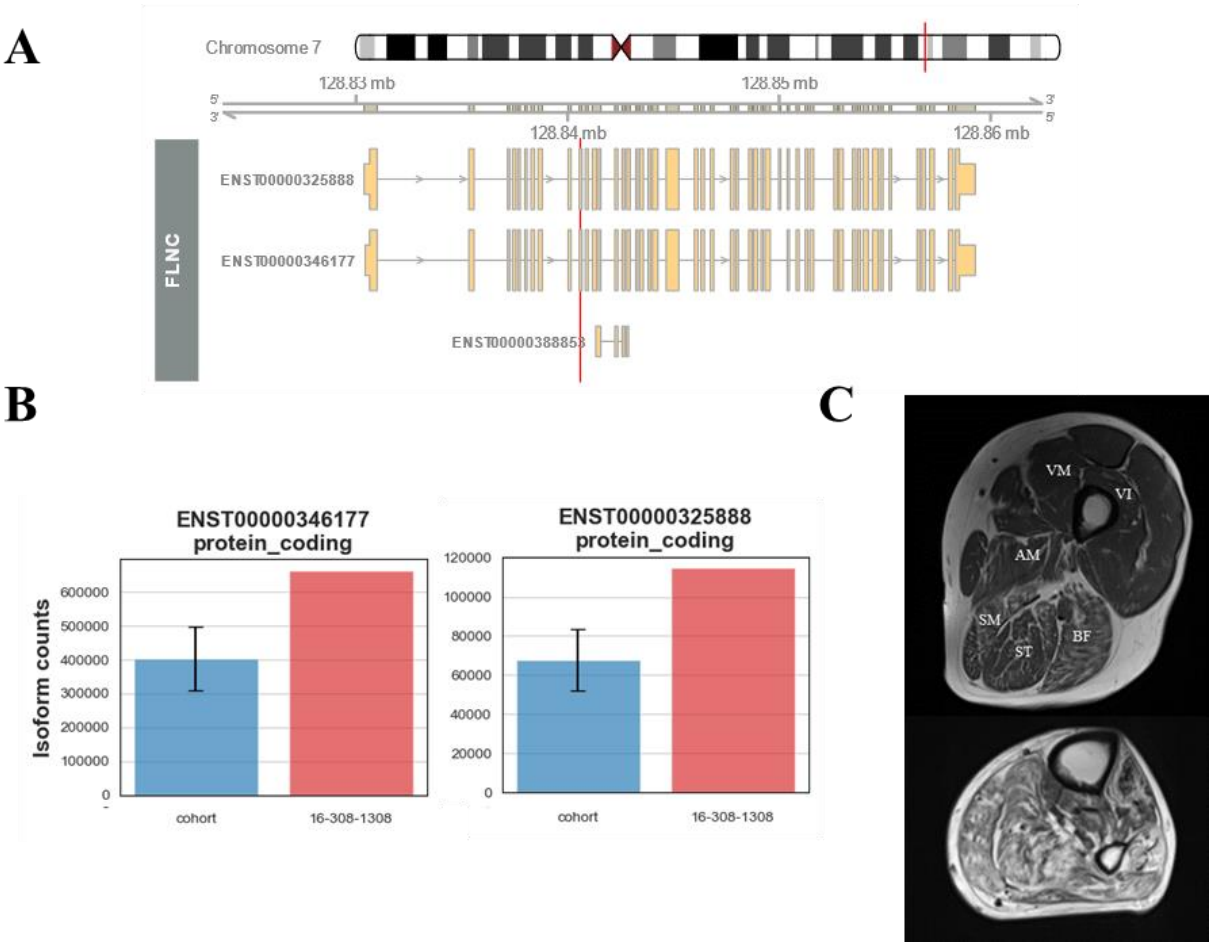


Figure 4.3: *FLNC* transcript count abnormalities in FLNC:p.E534K. Filamin C transcript isoform exon usage, with highlighted region indicating single nucleotide variant observed in sample 1308 **(A)**. The same sample has increased read counts for the two protein-coding isoforms, compared with the remainder of the cohort **(B)**. MRI for participant 1308 showing fatty infiltration of the SM, ST, & BF, and much of the muscles of the lower leg **(C)**. Vastus intermedius (VI), vastus medialis (VM), adductor magnus (AM), semimembranosus (SM), semitendinosus (ST), biceps femoris (BF).

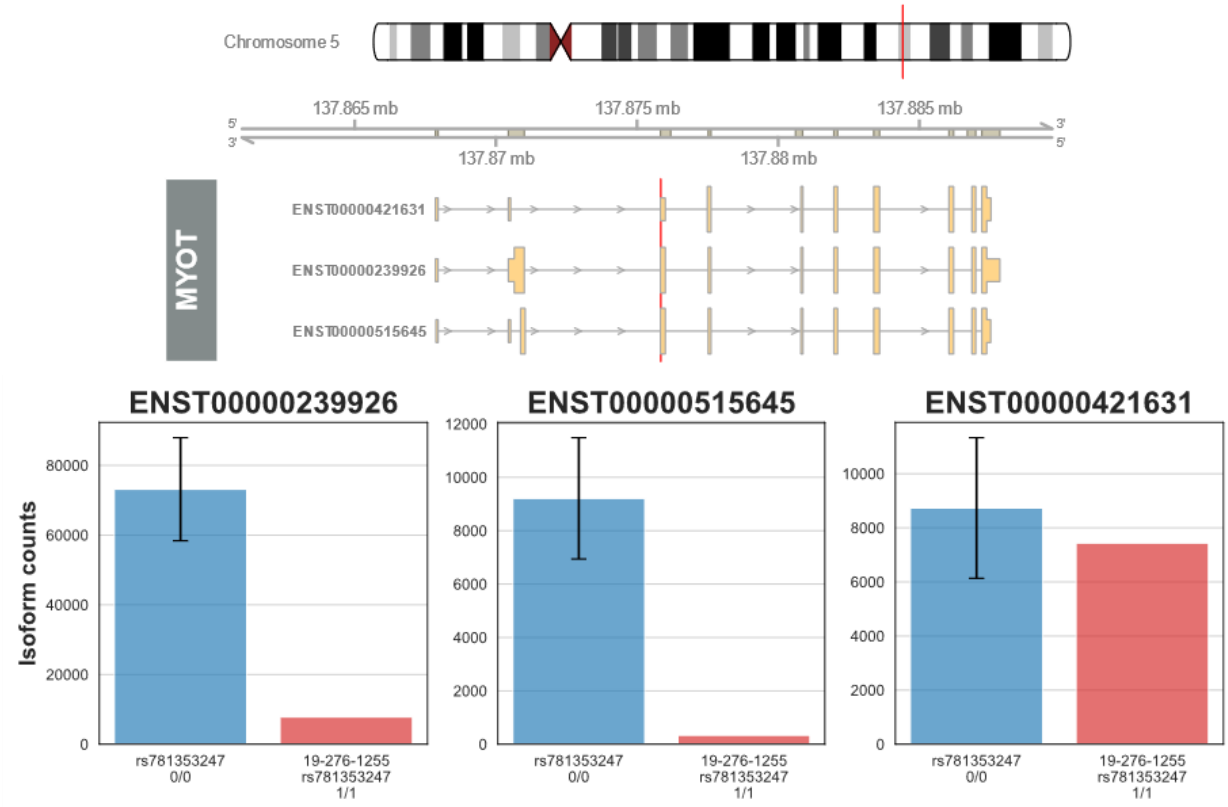


Figure 4.4: *MYOT* transcript count abnormalities in MYOT:p.A125Lfs*5. Myotilin protein-coding transcript isoform exon usage, with highlighted region indicating single nucleotide deletion observed in sample 1255 **(A)**. The same sample has reduced read counts for the two isoforms with the longest open reading frame, compared with the remainder of the cohort **(B)**. The isoform that does not contain the frameshifting variant in its ORF has a read count comparable to the cohort.

Third, participant 1105 was diagnosed with a distal myopathy, with early respiratory muscle impairment at 64. Deltoid biopsy revealed a chronic, moderate myofibrillar-like myopathy, including internalized nuclei, endomysial fibrosis, vacuoles with eosinophilic material, and "wiped-out" pallor and myofibrillar disarray on mitochondrial enzyme immunohistochemistry, αB-crystallin-(+) foci, and Z-band streaming. Eight heterozygous missense variants, and one heterozygous intronic variant, were found on clinical sequencing. Two of the missense variants were declared as pseudodeficiency alleles in GAA. A variant in *RYR1* was declared as likely pathogenic by the sequencing provider; insufficient information was available to recapitulate a similar classification with ACMG criteria, and was subsequently listed as a VUS in Table 4.2. A variant that was found on clinical sequencing, and in RNAseq, that passed filtering criteria with the isoform expression consideration, was rs201291446 (NEB:p.R8252H) in *NEB*, among other *NEB* variants unique to the participant in this cohort. An additional finding in *NEB* is a splice donor variant 2-151565043-C-T, that is absent from gnomAD, and dbSNP, but immediately adjacent to rs1203384703, with a frequency of 1/140280 in gnomAD's genomes, and predicted by SpliceAI to impact splicing. Neither the participant's, nor the adjacent splice variant is reported in ClinVar, and the missense variant is listed with uncertain significance for nemaline myopathy 2. At the transcript level, a highly similar isoform to that expressed most in the cohort was found overexpressed ($Z > 1.96$) in this participant's sample, among another isoform similar to the canonical, and overexpressed transcript (but where $|Z| < 1.96$; Figure 4.5). Since the first two isoforms differ by a single exon, coverage of the two adjacent exons was confirmed to follow the pattern observed with the appropriate isoforms. The third transcript has been discontinued from ensemble since performing this analysis, but its reported protein sequence is identical to that of the overexpressed transcript in this participant.

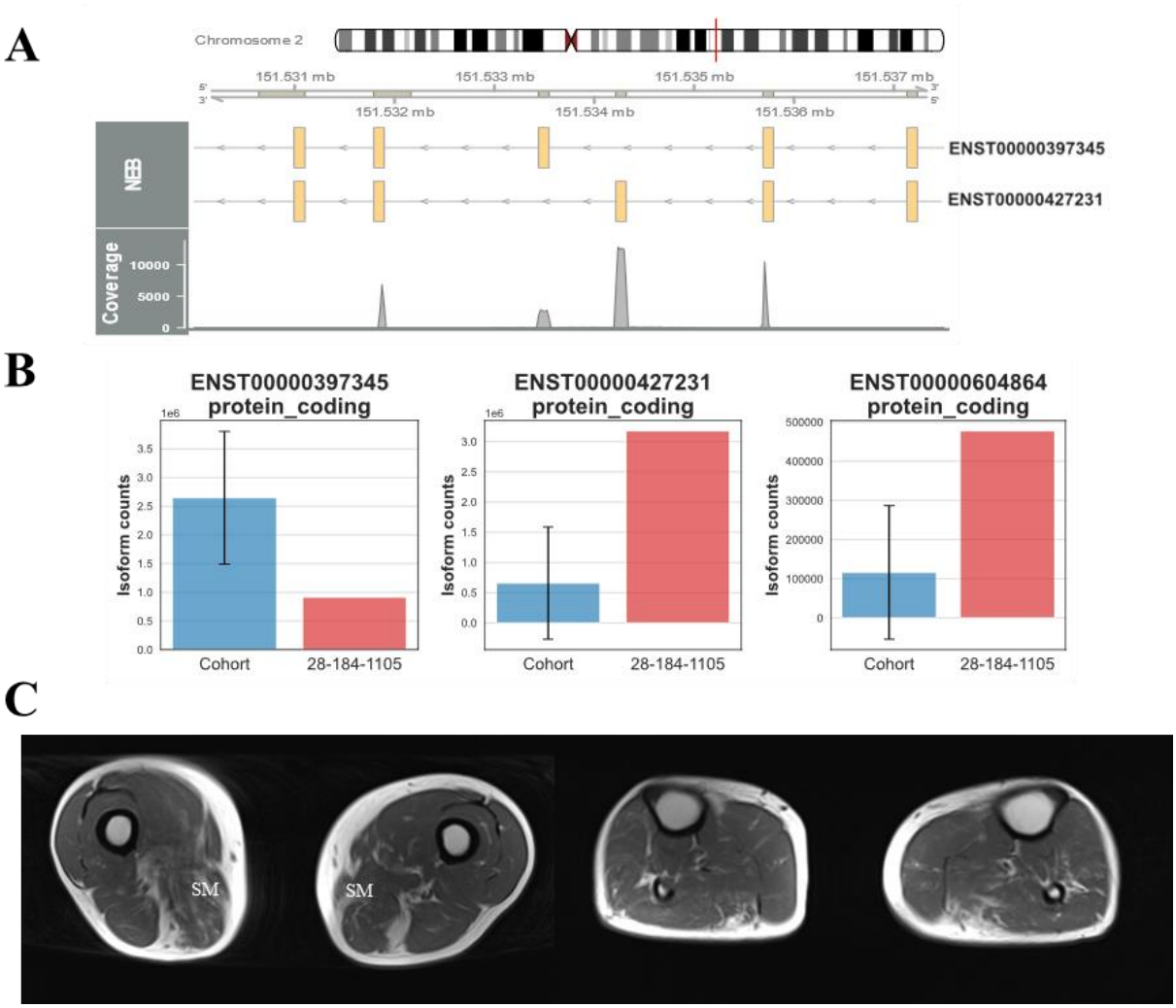


Figure 4.5: *NEB* transcript abnormalities in an individual with several unique variants. Sequencing coverage of the exclusive exons between ENST00000397345, and ENST00000427231 in *NEB* **(A)**. Expression of three *NEB* protein-coding transcript isoforms for sample 1105, showing decreased expression of the predominant isoform ENST00000397345, and increased expression of ENST00000427231, and ENST00000604864, where the former passed the filtering criteria (|Z| > 1.96; **B)**. Participant thigh and leg MRI showing isolated semimembranosus involvement **(C).**

Their MRI shows isolated changes to semimembranosus, where *NEB* myopathies vary in posterior compartment presentation, either sparing it entirely, or causing signal hyperintensities within the muscles of the hamstrings[79]. It is unknown which variant would be responsible for these findings.

The last clinical vignette that will be presented here is participant 1523, who presented with proximal leg weakness at 61, with diffuse weakness on examination, and normal CK (45 U/L). Deltoid biopsy demonstrated mild, non-specific changes, including myofibre size variation, scattered nuclear knots, type 2, COX-(-) fibres, perimysial thickening with fatty replacement, no inflammation, vacuoles, or dystrophic processes. Clinical sequencing found a heterozygous frameshift variant in *SELENON*, rs368104077 (SELENON:p.N238Kfs*?), declared as pathogenic in ClinVar, found in gnomAD at 45/280576. The variant was called from the RNAseq, and identified after passing the transcript expression criterion, where the two most common isoforms are reduced in this sample, and the third, less common, isoform is dramatically increased (Figure 4.6). The variant occurs in exon 4/12, or 5/13 for the two reduced isoforms respectively, and exon 4/13 for the increased transcript. The latter differs from the first two by inclusion of an exon distal to the variant. MRI showed generalised sarcopenia, which could be considered consistent with more severe manifestations of *SELENON* myopathies[80]. No other rare variants in *SELENON* were detected for this individual.

It is interesting to note, that despite three participants presumed to have mitochondrial disease, very few participants had transcripts that deviated sufficiently from the cohort for those variants to be considered. To give these participants a chance at a molecular diagnosis, that crite-

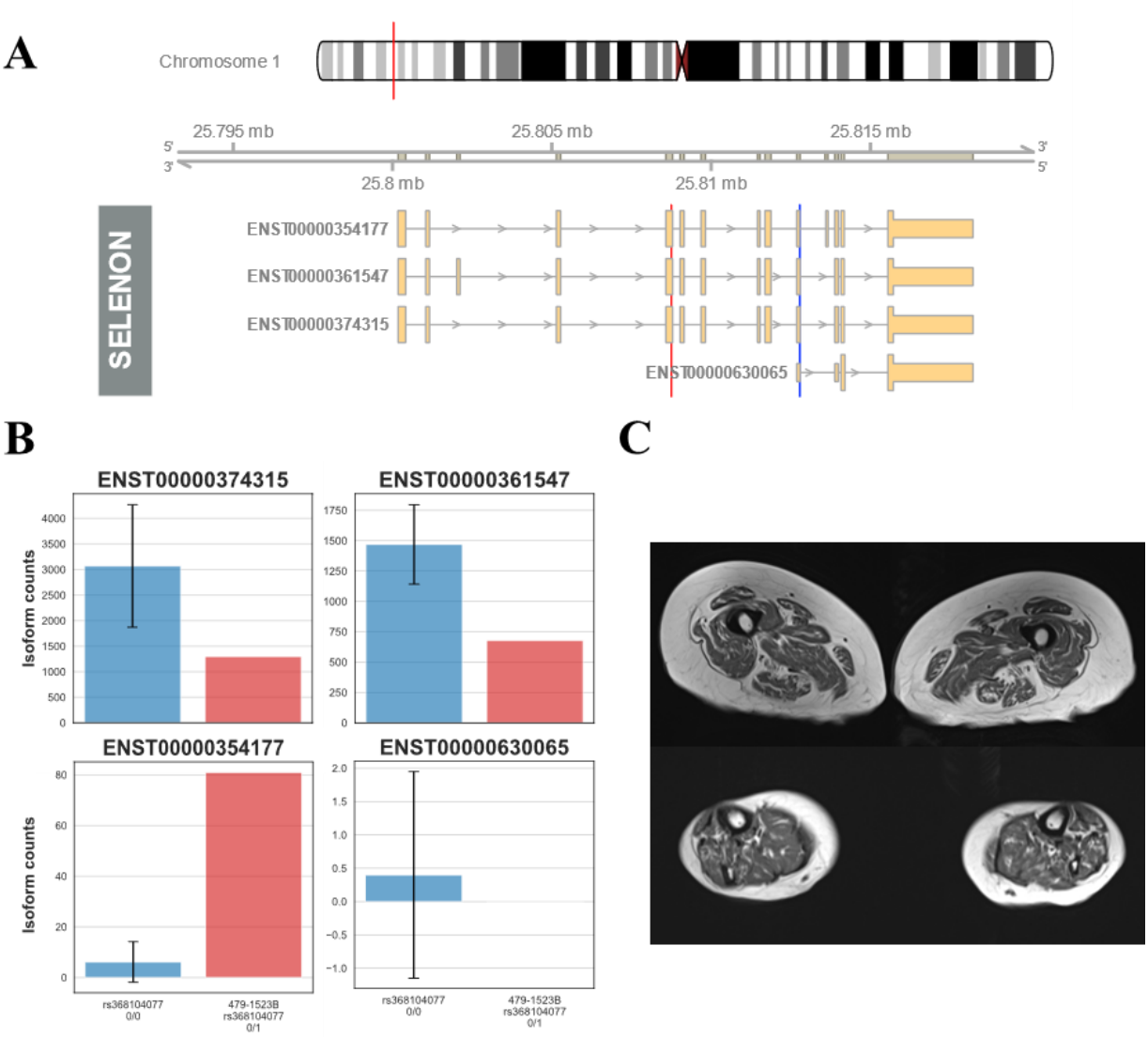


Figure 4.5: *SELENON* transcript count abnormalities in SELENON:p.N238Kfs*?. *SELENON* protein-coding transcript isoform exon usage, with highlighted region indicated single nucleotide deletion observed in sample 1523 (red), and the next selenocysteine redefinition element (blue) **(A)**. This sample has fewer reads mapped to the two most common isoforms in the rest of the cohort, and a dramatic increase in reads mapped to the third **(B)**. Participant MRI showing generalized sarcopenia **(C).**

-rion was lifted for mitochondrial genes, effectively treating them as if they instead had whole exome sequencing, rather than RNAseq, and a list of the most significant variants, predicted to have protein-coding impacts, or deletions, was constructed (Table 4.3). Participant 1099 had five missense variants across *MT-CO3, MT-ND5*, and *MT-CYB*. Only rs200855215 in *MT-MD5* had any indications of causing disease reported in ClinVar, however other investigators reported it as benign. Participant 1312 had two variants in *MT-CYB*, both with conflicting interpretations on ClinVar. And lastly, 1517 had five missense variants, one that was absent from dbSNP entirely, two absent from, and three listed as benign in ClinVar. This participant did, however, contain a unique homoplasmic deletion, far upstream of several mitochondrial genes.

4.2 Discussion

4.2.1 Transcript Expression can be Informative in Variant Prioritization

For a variant to influence any particular trait, there should be some biologically quantifiable difference from otherwise normal conditions. Ideally, as a cost saving measure, the system to detect variants, and the quantification method would be identical, such as RNAseq. Given the splicing impacts of most variants is still difficult to estimate, variants were considered if any transcript within their resident gene were unusually expressed in order to avoid excessive false negatives. This resulted in about 1/10 of a reduction in search space for prioritizing variants, which necessitated combining it with other parameters. While it hinders gene discovery attempts, focusing on variants within documented myopathy genes provide the most digestible variant lists, when used with the other parameters. Using either CADD scores, or focusing on coding consequences resulted in similar length datasets. Coding consequences allows an investigator to focus on the more intuitively damaging variants. However, most of those variants

Table 4.3: mtDNA missense variants, and deletions found in mitochondrial samples.

Participant ID	Variant (dbSNP ID)	HGNC Symbol	HGVS	ClinVar Designation	ALT Reads (%)[a]
1099B	rs2853825	*MT-CO3*	p.V91I	Benign	98.4
	rs1556424136	*MT-ND5*	p.I100V	Benign	99.9
	rs1556424302	*MT-ND5*	p.T432A	Benign	100
	rs200855215	*MT-ND5*	p.Q434R	Pathogenic & Benign	100
	rs199951903	*MT-CYB*	p.G251S	L. Benign & Risk Factor & Benign	99.8
1312	rs41518645	*MT-CYB*	p.D171N	Conflicting & Benign	100
	rs200336777	*MT-CYB*	p.V356M	Benign & Pathogenic	99.6
1517A	rs377569791	-	MT:494delC	-	99.0
	rs193303045	*MT-ATP6*	p.A177T	Benign	99.9
	rs2853492	*MT-ND4*	p.V230M	-	10.2
	MT-12074-A-C	*MT-ND4*	p.M439L	-	99.9
	rs386829198	*MT-ND5*	p.T556A	Likely Benign & Benign	99.7
	rs28357681	*MT-CYB*	p.F18L	Benign	99.9

[a] Fraction of sequenced reads with ALT allele, compared to total reads sequenced for that location.

would be missense, and therefore more difficult to associate with the transcript counts. CADD scores on the other hand, lack the innate intuition over the meaning of their scores, but allow prioritization of intronic and regulatory variants that would manifest effects at the transcript level, in addition to a means of rank prioritizing variants, rather than filtering.

One limitation in the application of transcript counts to variant prioritization is the use of data from a disease cohort, where each sample could have transcripts that would differ from a healthy control. A similar study as been performed in the past where a myopathy cohort was instead compared to the Genotype-Tissue Expression database (GTEx), focusing primarily on detecting splicing abnormalities[81]. Publicly available expression data will be preferable to sequencing entire cohorts for clinical investigation. However, experimental differences may reduce translatability. Cummings *et al.* managed to achieve a molecular diagnosis for about 60% of their cohort, when candidates had already been found by WES, or WGS[81]. While remarkable, the reliance on another NGS technique makes the practicality of RNAseq less feasible for a clinical setting, where costs are a concern. While this study includes plausible variant candidates for several participants that were identified by clinical sequencing, the methodology described herein produced them independently of the clinical results. Additionally, the aforementioned study benefitted from a cohort of similarly phenotyped cases, while the present study evaluated participants of divergent phenotypes, and consequently molecular etiologies.

4.2.2 Genetic Myopathy Variant Prioritization

Four of thirteen genetic myopathy participants have plausible molecular etiologies based on transcriptome-level information. Two had the offending variants detected by clinical

sequencing. The other two had candidates discovered by targeted sequencing, but were either supplanted by an alternative variant that more consistently explained the clinical features, or another variant in the same gene as one detected previously, where which, or both, are pathogenic is unknown.

FLNC variation has been implicated in several diseases, including various forms of MFM and hypertrophic cardiomyopathy[82-85]. In terms of myopathy, *FLNC* variants have been found to cause AD MFM[82], and recently, an AR congenital myopathy[83]. Support for a myofibrillar disease includes the midlife age of onset, predominant presentation in legs, elevated serum CK activity, histopathology results including the rare vacuole, and desmin-(+) and dystrophin-(+) aggregates, and posterior compartment of the thigh involvement as seen by MRI[82]. However, the missense variants associated with *FLNC* MFM are closer to either the N-, or C-terminals of the protein. Whereas FLNC:p.E534K sits deeper in the protein, closer to the variant discovered in the congenital myopathy (FLNC:p.P442R)[83]. Furthermore, carriers of the congenital myopathy variant appeared normal[83]. It also is not known how a missense variant might result in an increase of transcripts, as was observed in this participant. While there is salient information here to consider a *FLNC* diagnosis, clinical sequencing did detect a previously reported variant in *GNE*.

Two participants had clinical candidates in the gene *GNE*, an enzyme that takes part in sialic acid metabolism. In both cases, the sample's *GNE* isoform expression was similar to the rest of the cohort, resulting in the variants being neglected by the utilized prioritization criteria. However, it is likely that even in a pathogenic state, *GNE* transcripts would remain relatively

normal. Previous studies have found that across normal, and hereditary inclusion body myopathy (hIBM; not to be confused with IBM) GNE protein expression remains stable[86,87]. Pathogenic variants in *GNE* tend to impact the epimerase, or kinase activity of the protein, where functional defects are detected by measuring sialic acid residues on glycoproteins[86-88]. While GNE activity is regulated at the protein level, by a negative feedback loop[87], in the absence of transcriptional feedback, it is unlikely simple missense variants would impact expression, and therefore still be plausible candidates of pathogenic variants. There is however, one more obstacle for considering these *GNE* variants as pathogenic: hIBM tends to be autosomal recessive, where identified participants have either been homozygous for a single variant, or compound heterozygous for two variants, that could impact either of GNE's functional domains[86,88]. In the absence of a second variant detected for samples 1308, and 1524, the variant in *FLNC* remains more plausible for 1308, particularly given their MRI results, and 1524 remains without a plausible pathogenic candidate.

Most *MYOT* myopathies display AD inheritance, and have been associated with missense variants in the 2nd exon[89]. Where the profile matches that of participant 1255 includes the variation in muscle fibre size, rare vacuoles, and desmin-(+) and dystrophin-(+) inclusions found by histology[89]. Unusually however, where the youngest observed onset in a *MYOT* myopathy was in the young adults of a family with a particular variant[89], it is believed participant 1255 may have congenital disease, with weakness reported in early childhood, before reported with diminished ambulation at age 34. Additionally, it begs the question that if simple missense mutations can cause relatively severe disease, how is this participant no worse off for having almost no expression of the two largest, and most frequent *MYOT* isoforms? Intriguingly, *MYOT*

knockout mice have been generated previously, and presented with no muscular defects[90]. The similarity between human and murine *MYOT* has already been commented on by investigators characterizing novel *MYOT* variants[89,90], but does this result indicate mice and humans may have differences in their dependence on *MYOT*? Alternatively, while purely speculative at the moment, could there be some compensatory mechanism for congenital defects in *MYOT*?

Another gene with hard to predict variant consequences is *NEB*. Participant 1105 had a missense variant identified clinically, in addition to a splice site variant detected during this study in *NEB*, however, neither occur near the exclusive exon between the two predominant isoforms detected. Additionally, *NEB* products typically lay in the 600-900 kDa range[91,92], while the two predominant transcripts both have 987 kDa protein products. The apparent increase in expression of these two isoforms may be attributable to the quantification methods used in RNAseq. The pseudomapper used here, Salmon, breaks up the transcriptome into 31-mer (by default) fragments, and aligns those to the reads to quickly estimate transcript abundance, where the proportion of an isoform within the transcriptome will be estimated by its unique splice sites. Other methods, like *de novo* assembly, would assemble the reads into the longest transcripts possible, to achieve a parsimonious result. Since the smaller isoforms of *NEB* are going to contain many of the exons contained in the largest isoforms, unless there is a dramatic difference among the expression of the small isoforms, the most parsimonious choice would be to assume the reads originated from the largest transcripts. This remains speculative at the moment, and requires further investigation to understand the *NEB* transcriptome of this participant. That said, the true difference between the normal isoform, and the highest isoform in this participant is a single exon. Further, even GTEx finds limited expression of this exon (exon

144 in their records) in skeletal muscle. Therefore, the question remains, is this exon somehow problematic in muscle?

Selenoprotein N, the protein product of *SELENON*, is one of the rare selenoproteins in the human genome, that incorporate selenocysteine into their amino acid sequence by a complex mechanism that redefines the stop codon UGA[93,94]. Its function is still unknown, though it has been found to assist in sarcoplasmic calcium concentration maintenance[93], and can be found coupled to the ryanodine receptors, voltage-gated calcium channels (*RYR1-3*) in muscle[95]. Selenocysteine residues are believed to typically reside at active sites, where SELENON contains up to two[93,94]. One of these two sites would be disrupted by the frameshift observed in sample 1523. *SELENON* (or formerly *SEPN1*) related myopathies include the minicore myopathies, which are named for the small lesions observed by histology[96,97]. Other features include fibre size variation, and absence of dystrophic processes, both of which were reported in participant 1523. However, these myopathies are typically associated with missense variants, and follow an AR inheritance patter[96,97]. A similar single-base insertion has been observed previously, though the disease was attributed to a compound heterozygous condition with an out-of-phase missense variant[97]. The question then is whether this individual's disease is a product of the one variant, possibly explaining the later onset, or if there is another variant in *SELENON* that went undetected by targeted sequencing, and RNAseq.

The last point to discuss is the application of these techniques to diagnosing participants with presumed mitochondrial disease. As mentioned in the previous sections, mitochondrial diseases, and myopathies in particular, tend to occur as heteroplasmies[18], frequently by mtDNA

deletions[63]. However, most of the mtDNA variants detected in this cohort were both missense, and virtually homoplasmic. Among the candidate mtDNA variants, only one appears plausibly heteroplasmic, with about 10% of reads at the variant site containing the alternative allele. It is possible HaplotypeCaller could not calculate sufficient confidence to call low abundance heteroplasmic variants. There have been several studies concerning NGS approaches to heteroplasmy[98-100]. Even with extensive coverage up to 1000X, only a heteroplasmic allele frequency of 3% is attainable with complete certainty[100]. Achieving that level of certainty for low-grade heteroplasmy also requires additional library enrichment steps, far beyond what was covered in the present study[99]. As such, complementary analyses may be required to complement RNAseq, or other NGS, when mitochondrial diseases are under consideration.

Chapter 5 – Discussion & Conclusions

5.1 Thesis Overview

The purpose of this thesis was two-fold. First, to compare various genetic myopathies, and an acquired myopathy at the transcriptomic level, and explore the lncRNAs differentially expressed between them for further study. And second, to look at the clinical utility of RNAseq, in identifying plausibly pathogenic variants among genetic myopathy participants. We found that most of the difference between IBM, and the genetic myopathy participants, can be explained by inflammatory processes. Additionally, candidate lncRNAs that are specific for one or two myopathies were identified for future biomarker analysis. Sex, serum CK activity, and muscle group, did not appear to dramatically influence these findings. Next, we found that transcript count statistics from RNAseq may be informative during variant prioritization, particularly when accompanying additional criteria. Finally, we describe four cases where RNAseq yielded salient results in the search for plausibly pathogenic molecular etiologies.

5.2 RNAseq for Disease Gene & Transcript Discovery

These days, RNAseq is the quintessential gene discovery technique. And yet, ironically, it seems most RNAseq based studies with genetic myopathies are focused on diagnosis, rather than understanding the pathology[81,101]. The unique splicing landscape of the muscle transcriptome has been researched in mice[102]. From a disease perspective, human myosites have been compared at the transcriptome level, not only characterizing their differences, but training a classifier on the data to assist in diagnosis[103]. However, it appears the present study is the first of its kind to compare the transcriptomes of genetic myopathies. The benefits of these types of

analyses extend beyond improving diagnostics. It is plausible that investigating the transcription landscape throughout the disease course would improve our understanding of its pathogenesis, and help identify potential points of intervention. Granted, consecutively harvesting muscle from several participants may not be feasible. However, participant primary, and mutation induced fibroblasts have been used previously to understand the pathogenicity of variants in the past[104]. With induction of myoblast differentiation from fibroblasts, it may be more feasible to model these diseases *in vitro*[105].

One limitation of this study is the dependence on a reference transcriptome. This introduces problems in two ways. First, by introducing biases in quantifying transcript abundance. Second, by preventing the discovery, and quantification of novel isoforms. One of the first developments with transcriptomics was *de novo* transcriptome assembly, through tools like Cufflinks, and later Trinity[102,103,106,107]. These tools have different approaches to assembly, Cufflinks constructs the shortest path available based on fragments overlapping splice sites[102,103], and Trinity constructs de Bruijn graphs for each transcript group[106,107]. However, abundance estimation for both ultimately depends on the abundance of reads unique to each assembled transcript[102,103,106,107], where variation in exon coverage could introduce a bias. One way to prevent this bias would be sequencing the entirety of each transcript, preventing any ambiguity in the relative abundance between isoforms. This is particularly true for many muscle-relevant genes, like *TTN*, *DMD*, *SYNE1-2*, or *NEB* that are relatively large, and contain myriad potential transcripts. Or, from a variant calling perspective, copy number repeat variants, like the polyglutamine repeat expansions seen in the spinocerebellar ataxias[108], and in the androgen receptor with spinobulbar muscular atrophy[109], the elusive hexanucleotide repeat in *c9orf72*

associated with the amyotrophic lateral sclerosis (ALS) – frontotemporal dementia (FTD) disease spectrum[110,111], or the copy number variants seen in DM and FSHD[8,9, 12], can be difficult to accurately measure from NGS techniques. Long read sequencing, such as Oxford nanopore sequencing (ONT), or Pacific Bioscience's single-molecule real-time sequencing (SMRTseq) technology, both considered the third generation of sequencing technology, could be the solution to the limitations seen with NGS. The advent of ONT, and SMRTseq have provided novel opportunities to explore sequences. ONT has been used to detect individual modifications of nucleobases in sequences[112,113], that would otherwise elude the unspecialized NGS techniques. More pertinent to the current study, ONT provides a means of RNAseq without the need of reverse transcription, or library amplification, that could introduce biases into the workflow[114]. However, where ONT, and SMRTseq fall behind NGS, is in base calling fidelity, and consequently, reduced certainty in variant calling, though they are constantly improving, and there are developments available to close this gap[115,116].

A common theme in trying to explain our findings with the upregulated lncRNAs, is the lack of literature on their function. Despite this gap of knowledge, we know many lncRNAs are functional, from genome-wide association studies (GWAS), and expression quantitative trait loci (eQTL) studies[117]. Dissecting the function of all lncRNAs is no easy task either, they are typically less conserved than mRNAs, have higher regulatory complexity, and variable cell-type specificity[117]. There are efforts now, however, to understand the non-coding transcriptome across cell-types[118], and in disease states like ALS[119]. For example, the lncRNA *NEAT1* is associated with paraspeckle formation in nuclei, as a type of post-transcriptional modification centre, similar to Cajal bodies[120]. In ALS however, there is aberrant expression in motor

neurons, that associate with the ALS-related genes *TDP-43*, and *FUS*[119]. Similarly in FTD, antisense transcripts of the notorious *c9orf72* locus have been found in inclusions, sequestering RNA binding proteins, demonstrating a disease-specific lncRNA function[121]. While annotating lncRNAs is relatively straightforward (absence of an open reading frame, length >200bp), high-throughput functional evaluation requires a more creative approach[118]. FANTOM6 released their proof of concept, where lncRNA function will be determined through RNA interference studies in cell culture, where growth and viability parameters are automatically quantified[118]. While a full understanding of the noncoding transcriptome is still many years out, there has been interest in identifying the diagnostic, and prognostic potential of serum, tissue, and exosomal lncRNAs, including in various cancers[122,123], cardiovascular disease[124], and multiple sclerosis[37]. Pursuing these markers now may provide a foothold to investigate the nuances of various pathologies as the functions of lncRNAs become understood.

5.3 Clinical Utility of RNAseq

In the present study, four of thirteen genetic myopathy participants had candidate pathogenic variants associated with RNAseq findings, for a yield of 30%. Many of these were also found by clinical sequencing, though the investigator was blind to these findings at the time of variant prioritization. Reasons that clinically relevant variation may be missed includes efficient splicing, or *cis*- and *trans*-regulatory region variants that would not normally be transcribed. Additionally, even if a transcript-level abnormality may be detected, the offending variant itself may be missed from its effects on exon retention, or splicing efficiency.

While it is too soon to confidently consider these molecular diagnoses, particularly in the absence of causal associations or previous documentation, it appears we have achieved a yield similar to what has been demonstrated for WES of singleton participants[81,125-128]. Though a recent study suggests higher diagnostic yields may be skewed to younger cohorts/diseases of earlier onset[128] which may suggest the present yield should exceed expectations. Naturally, WES of trios confers the highest diagnostic yields[125-127], where disease segregation, carrier status, and *de novo* mutations can be isolated. However, the opportunities to sequence parents are increasingly limited the later the onset of a disease in question is. It is important to note that these diagnoses could not be made independently of clinical findings, as MRI and histology was considered for each. That said, there is still promise for RNAseq in the clinic, particularly for accessible tissues, such as muscle, skin, or serum.

Beyond the sequencing of trios, there are other means that could improve yields. First, if platform-, and program-based biases can be addressed, it would be possible to compare transcript abundances to publicly available databases, similar to the previously mentioned (see section 4.2.1) Cummings *et al.*, and Gonorazky *et al.* approaches with GTEx data[81,101]. As it stands, investigating these biases prove to be a barrier in broader reproducibility, and comparability, particularly for clinical needs. These studies controlled for these biases by analysing GTEx sequences with their respective pipelines[81,101]. However, this requires significantly higher demands for storage, and processing power, consequently reducing the plausibility for a clinical approach, unless predetermined control datasets are made available. Additionally, the present, and both of the previous studies, have all taken three distinct approaches to quantitation. Cummings *et al.* used a pipeline similar to GTEx, with an alignment with STAR, and abundance

estimation with RNA-SeQC, summarizing their counts in terms of reads per kilobase per million mapped reads (RPKM)[81]. Gonorazky *et al*. aligned reads with STAR, and quantified transcripts with featureCounts, also as RPKM, analysing them with edgeR (which has marginially less precision than DESeq2[48]). The present study, aligned with STAR as the previous two, however, alignments were only used for variant calling. Instead, transcript abundance was estimated with Salmon, reducing processing time, and accounting for some sequencing biases[48]. Here, statistics were performed on raw counts (though transcripts per million [TPM] is an option with salmon), as recommended by the group that developed DESeq2[49], as RPKM is relative to RNA concentration, making comparisons between samples more difficult[180]. While the diversity of tools available for analysing RNAseq is promising, as it indicates active development and improvement in the field, it may be time for specialists to develop a consensus on RNAseq pipelines, particularly with regard to abundance estimation, where these three studies differed the most.

Second, the most selective variant prioritization criterion was our focus on myopathy-associated genes, however, without a more gene-inclusive means of prioritizing variants, the genome-wide potential of RNAseq cannot be fulfilled. Omitting this filter however provided variant lists too extensive for a clinician to routinely prioritize. Luckily, there are efforts to develop variant prioritization techniques that consider evolutionary, biochemical, and phenotypic information such as TAIGA2, a variant classifier trained on an ensemble of variant impact prediction tools, with additional weighting for variants found in genes associated with human phenotypes[130].

5.5 Future Directions

Part of the motivation behind this work was to identify potential dysregulated lncRNAs that could be tested as disease biomarkers. As such, there are several ways to move forward with this data. First, these results should be verified by qPCR, or digital droplet PCR. Second, with the lncRNAs that passed verification, a replication cohort will be needed to test their biomarker potential.

In terms of using lncRNAs as biomarkers, there may yet be hesitation in their use if muscle biopsy is required. However, there is increasing interest in liquid biopsies, by means of testing exosomal RNAs. Exosomes are circulating vesicles, that may contain most biomolecules, including nucleic acids, and they have been found being produced by skeletal muscle, and other relevant tissues, such as adipose, or immune cells[131-135]. There is already research into exosomal RNAs, including lncRNAs, on cardiovascular disease[134], metabolic disorders[135], and exosomal μRNAs in muscular dystrophies, congenital and inflammatory myopathies[132], and neuromuscular disease[133]. Muscle health seems uniquely well suited for exosomal analyses, given 40% of body mass, on average, can be attributed to the musculoskeletal system.

There is room to verify the candidate pathogenic variants illustrated in the current study, to substantiate the claims of pathogenicity. Some example experiments that may be conducted include culturing participant primary fibroblasts. There is also the issue of most participants evading a diagnosis. These may be forwarded for WGS to increase their changes at achieving a diagnosis. There is still promise for these participants, particularly given the success of Cummings *et al*. when combining WGS or WES with their RNAseq[81].

5.6 Conclusions

This discovery-oriented study was devised first-and-foremost to identify potential molecular biomarkers to delineate between the various genetic, and acquired myopathies, that could be further explored. We found several lncRNAs, and gene-level signals that were specific to the histological groups. We also identified the main difference between the transcriptomes of IBM, and genetic myopathies is most likely from the immune infiltration characteristic of IBM. Given the potential wealth of information provided by RNAseq, we also attempted to identify pathogenic variation in our sample cohort. This appears to be the first study of its kind to focus on late-onset myopathies, in difficult-to-diagnose individuals with divergent manifestations of disease. Four of thirteen participants had salient transcript-level findings that could be associated with local variants. The data gathered may yet prove useful for the remaining ten, should they be forwarded for WGS. As sequencing becomes an increasingly popular clinical tool, the necessity of personalize, and precision approaches to medicine will become apparent.

References

1 Mah, J. K., Korngut, L., Fiest, K. M., Dykeman, J., Day, L. J., Pringsheim, T., & Jette, N. (2016). A Systematic Review and Meta-analysis on the Epidemiology of the Muscular Dystrophies. *Canadian Journal of Neurological Sciences / Journal Canadien Des Sciences Neurologiques, 43*(1), 163–177. https://doi.org/10.1017/cjn.2015.311

2 Koenig, M., Beggs, A. H., Moyer, M., & Bettecken, K. H. T. (1989). The Molecular Basis for Duchenne versus Becker Muscular Dystrophy: Correlation of Severity with Type of Deletion. *American Journal of Human Genetics, 45*,498–506.

3 Neguembor, M., Jothi, M., & Gabellini, D. (2014). Long noncoding RNAs, emerging players in muscle differentiation and disease. *Skeletal Muscle, 4*(1), 8. https://doi.org/10.1186 /2044-5040-4-8

4 Bovolenta, M., Erriquez, D., Valli, E., Brioschi, S., Scotton, C., Neri, M., Falzarano, M. S., ,,, Ferlini, A. (2012). The *DMD* Locus Harbours Multiple Long Non-Coding RNAs Which Orchestrate and Control Transcription of Muscle Dystrophin mRNA Isoforms. *PLoS ONE, 7*(9), e45328. https://doi .org/10.1371/journal.pone.0045328

5 Nakka, K., Ghigna, C., Gabellini, D., & Dilworth, F. J. (2018). Diversification of the muscle proteome through alternative splicing. *Skeletal Muscle, 8*(1), 8. https://doi.org/10.1186 /s13395-018-0152-3

6 Malhotra, S., Hart, K., Klamut, H., Thomas, N., Bodrug, S., Burghes, A., Bobrow, M., … Wortonm, R. (1988). Frame-shift deletions in participants with Duchenne and Becker muscular dystrophy. *Science, 242*(4879), 755–759. https://doi.org/10.1126/science .3055295

7 Martone, J., Briganti, F., Legnini, I., Morlando, M., Picillo, E., Sthandier, O., Politano, L., & Bozzoni, I. (2016). The lack of the Celf2a splicing factor converts a Duchenne genotype into a Becker phenotype. *Nature Communications, 7*(1), 10488. https://doi.org/10.1038/ncomms10488

8 Lin, X., Miller, J. W., Mankodi, A., Kanadia, R. N., Yuan, Y., Moxley, R. T., Swanson, M. S., & Thornton, C. A. (2006). Failure of *MBNL1*-dependent post-natal splicing transitions in myotonic dystrophy. *Human Molecular Genetics, 15*(13), 2087–2097. https://doi.org/10.1093/hmg/ddl132

9 van den Heuvel, A., Mahfouz, A., Kloet, S. L., Balog, J., van Engelen, B. G. M., Tawil, R., Tapscott, S. J., & van der Maarel, S. M. (2019). Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. *Human Molecular Genetics, 28*(7), 1064–1075. https://doi.org/10.1093/hmg/ddy400

10 Hu, N., Antoury, L., Baran, T. M., Mitra, S., Bennett, C. F., Rigo, F., Foster, T. H., & Wheeler, T. M. (2018). Non-invasive monitoring of alternative splicing outcomes to identify candidate therapies for myotonic dystrophy type 1. *Nature Communications, 9*(1), 5227. https://doi.org/10.1038/s41467-018-07517-y

11 Kamsteeg, E.-J., Kress, W., Catalli, C., Hertz, J., Witsch-Baumgartner, M., Buckley, M., ... & Scheffer, H. (2012). Best practice guidelines and recommendations on the molecular diagnosis of myotonic dystrophy types 1 and 2. *European Journal of Human Genetics, 20,*1203–1209.

12 Ruggiero, L., Mele, F., Manganelli, F., Bruzzese, D., Ricci, G., Vercelli, L., Govi, M., … Tupler, R. (2020). Phenotypic Variability Among Participants With *D4Z4* Reduced

Allele Facioscapulohumeral Muscular Dystrophy. *JAMA Network Open*, *3*(5), e204040. https://doi.org/10.1001 /jamanetworkopen.2020.4040

13 Gilbreath, H. R., Castro, D., & Iannaccone, S. T. (2014). Congenital Myopathies and Muscular Dystrophies. *Neurologic Clinics*, *32*(3), 689–703. https://doi.org/10.1016/j.ncl .2014.04.006

14 Selcen, D. (2011). Myofibrillar myopathies. *Neuromuscular Disorders*, *21*(3), 161–171. https://doi.org/10.1016/j.nmd.2010.12.007

15 Vincent, A. E., Grady, J. P., Rocha, M. C., Alston, C. L., Rygiel, K. A., Barresi, R., Taylor, R. W., & Turnbull, D. M. (2016). Mitochondrial dysfunction in myofibrillar myopathy. *Neuromuscular Disorders*, *26*(10), 691–701. https://doi.org/10.1016/j.nmd.2016.08.004

16 Friedman, J. R., & Nunnari, J. (2014). Mitochondrial form and function. *Nature*, *505*(7483), 335–343. https://doi.org/10.1038/nature12985

17 Vincent, A. E., Ng, Y. S., White, K., Davey, T., Mannella, C., Falkous, G., ... Picard, M. (2016). The Spectrum of Mitochondrial Ultrastructural Defects in Mitochondrial Myopathy. *Scientific Reports*, *6*(1), 30610. https://doi.org/10.1038/srep30610

18 Eisner, V., Lenaers, G., & Hajnóczky, G. (2014). Mitochondrial fusion is frequent in skeletal muscle and supports excitation–contraction coupling. *Journal of Cell Biology*, *205*(2), 179–195. https://doi.org/10.1083/jcb.201312066

19 Greenberg, S. A. (2019). Inclusion body myositis: Clinical features and pathogenesis. *Nature Reviews Rheumatology*, *15*(5), 257–272. https://doi.org/10.1038/s41584-019-0186-x

20 Greenberg, S. A., Pinkus, J. L., Kong, S. W., Baecher-Allan, C., Amato, A. A., & Dorfman, D. M. (2019). Highly differentiated cytotoxic T cells in inclusion body myositis. *Brain*, *142*(9), 2590–2604. https://doi.org/10.1093/brain/awz207

21 Dimachkie, M. M., & Barohn, R. J. (2014). Distal Myopathies. *Neurologic Clinics, 32*(3), 817–842. https://doi.org/10.1016/j.ncl.2014.04.004

22 Lindholm, M. E., Huss, M., Solnestam, B. W., Kjellqvist, S., Lundeberg, J., & Sundberg, C. J. (2014). The human skeletal muscle transcriptome: Sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *The FASEB Journal, 28*(10), 4571–4581. https://doi.org/10.1096/fj.14-255000

23 Gheller, B. J. F., Riddle, E. S., Lem, M. R., & Thalacker-Mercer, A. E. (2016). Understanding Age-Related Changes in Skeletal Muscle Metabolism: Differences Between Females and Males. *Annual Review of Nutrition, 36*(1), 129–156. https://doi. org/10.1146/annurev-nutr-071715-050901

24 Liu, D., Sartor, M. A., Nader, G. A., Gutmann, L., Treutelaar, M. K., Pistilli, E. E., IglayReger, H. B., ... Gordon, P. M. (2010). Skeletal muscle gene expression in response to resistance exercise: Sex specific regulation. *BMC Genomics, 11*(1), 659. https://doi.org/10.1186/1471-2164-11-659

25 Sarkozy, A., Hicks, D., Hudson, J., Laval, S. H., Barresi, R., Hilton-Jones, D., Deschauer, M., … Lochmüller, H. (2013). *ANO5* Gene Analysis in a Large Cohort of Participants with Anoctaminopathy: Confirmation of Male Prevalence and High Occurrence of the Common Exon 5 Gene Mutation. *Human Mutation, 34*(8), 1111–1118. https://doi.org/10.1002/humu.22342

26 Penttila, S., Palmio, J., Suominen, T., Raheem, O., Evila, A., & Gomez, N. M. (2012). Eight new mutations and the expanding phenotype variability in muscular dystrophy caused by *ANO5*. *Neurology, 78*,897-903.

27 Kley, R. A., Leber, Y., Schrank, B., Zhuge, H., Orfanos, Z., Kostan, J., ... Vorgerd, M. (2021). *FLNC*-Associated Myofibrillar Myopathy: New Clinical, Functional, and Proteomic Data. *Neurology Genetics, 7*(3), e590. https://doi.org/10.1212/NXG.0000000000000590

28 Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine, 17*(5), 405–423. https ://doi.org/10.1038/gim.2015.30

29 Amendola, L. M., Muenzen, K., Biesecker, L. G., Bowling, K. M., Cooper, G. M., Dorschner, M. O., Driscoll. C., … Jarvik, G. P. (2020). Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies. *The American Journal of Human Genetics, 107*(5), 932–941. https://doi.org/10.1016/j.ajhg.2020.09.011

30 Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y.-Y., Kobayashi, Y. ... Topper, S. (2017). Sherloc: A comprehensive refinement of the ACMG–AMP variant classification criteria. *Genetics in Medicine, 19*(10), 1105–1117. https://doi.org/10.1038/gim.2017.37

31 Brandt, T., Sack, L. M., Arjona, D., Tan, D., Mei, H., Cui, H., Gao, H., ... Meck, J. M. (2020). Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. *Genetics in Medicine, 22*(2), 336–344. https://doi.org/10.1038/s41436-019-0655-2

32 Fu, X.-D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics, 15*(10), 689–701. https://doi.org/10.1038 /nrg3778

33 Sibley, C. R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews Genetics, 17*(7), 407–421. https://doi.org/10.1038/nrg.2016.46

34 Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics, 3*(4), 285–298. https://doi.org/10.1038/nrg775

35 Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Kalsotra, A., Cooper, T. A., & Johnson, J. M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics, 40*(12), 1416–1425. https://doi. org/10.1038/ng.264

36 Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., & Bozzoni, I. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. *Cell*, *147*(2), 358–369. https://doi.org/10 .1016/j.cell.2011.09.028

37 Gupta, M., Martens, K., Metz, L. M., de Koning, A. J., & Pfeffer, G. (2019). Long noncoding RNAs associated with phenotypic severity in multiple sclerosis. *Multiple Sclerosis and Related Disorders, 36*,101407. https://doi.org/10.1016/j.msard.2019.101407

38 Hua, J. T., Chen, S., & He, H. H. (2019). Landscape of Noncoding RNA in Prostate Cancer. *Trends in Genetics, 35*(11), 840–851. https://doi.org/10.1016/j.tig.2019.08.004

39 Schulte, C., Barwari, T., Joshi, A., Zeller, T., & Mayr, M. (2020). Noncoding RNAs versus Protein Biomarkers in Cardiovascular Disease. *Trends in Molecular Medicine, 26*(6), 583–596. https://doi.org/10.1016/j.molmed.2020.02.001

40 Fattahi, S., Kosari-Monfared, M., Golpour, M., Emami, Z., Ghasemiyan, M., Nouri, M., & Akhavan-Niaki, H. (2020). LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: A novel approach to personalized medicine. *Journal of Cellular Physiology, 235*(4), 3189–3206. https://doi.org/10.1002/jcp.29260

41 Cabianca, D. S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., & Gabellini, D. (2012). A Long ncRNA Links Copy Number Variation to a Polycomb/Trithorax Epigenetic Switch in FSHD Muscular Dystrophy. *Cell, 149*(4), 819–831. https://doi.org/ 10.1016/j.cell .2012.03.035

42 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491–498. https://doi.org/10 .1038/ng.806

43 Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples [Preprint]. *Genomics*. https://doi.org/10. 1101/201178

44 Van der Auwera, G. A. (March 2014) The GATK Best Practices for variant calling on RNAseq, in full detail. https://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail (Retrieved June 24,2020).

45 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

46 McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology, 17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

47 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44–57.

48 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods, 14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

49 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org /10.1186/s13059-014-0550-8

50 Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. https://doi.org/10.1101/gr.133744 .111

51 Liewluck, T., Pho-Iam, T., Limwongse, C., Thongnoppakhun, W., Boonyapisit, K., Raksadawan, N., ... Sangruchi, T. (2006). Mutation analysis of the GNE gene in distal myopathy with rimmed vacuoles (DMRV) participants in Thailand. *Muscle & Nerve*, *34*(6), 775–778. https://doi.org/10.1002/mus.20583

52 Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics, 30*(4), 523–530. https://doi.org/10.1093/ bioinformatics/btt703

53 Johari, M., Vihola, A., Palmio, J., Jokela, M., Jonson, P. H., Sarparanta, J., ... Udd, B. (2021). Comprehensive transcriptomic analysis shows disturbed calcium homeostasis and deregulation of T lymphocyte apoptosis in inclusion body myositis [Preprint]. *BioArxiv.* https:// www.biorxiv.org/content/10.1101/2021.06.30 .450477v1.full.pdf

54 Morgan, M. (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16. https://CRAN.R-project.org/package-BiocManager

55 Soneson, C., Love, M. I., Robinson, M. D. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research, 4*,1521. https:// dx.doi.org/1012688/f1000research,7563.2

56 Rainer, J., Gatto, L., Weichenberger C. X. (2019) ensembledb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics, 35*(17), 3151-3153. https:// dx.doi.org/10.1093/bioinformatics/btz031

57 Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., de Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics, 21*,3439-3440

58 Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., Carey, V. J. (2013). Software for Computing and Annotating Genomic Range. *PLOS Computational Biology, 9(*8), e1993118. https://dx.doi.org/10.1371/journal.pcbi.1003118

59 Hahne, F., Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. *Methods in Molecular Biology, 1418*,335=351

60 Zeng, W., Jiang, S., Kong, X., El-Ali, N., Ball, A. R., Ma, C. I.-H., ... Mortazavi, A. (2016). Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Research, 44*(21), e158. https://doi.org/10.1093/nar/gkw739

61 Kim, M., Franke, V., Brandt, B., Lowenstein, E. D., Schöwel, V., Spuler, S., Akalin, A., & Birchmeier, C. (2020). Single-nucleus transcriptomics reveals functional compartmentalization in syncytial skeletal muscle cells. *Nature Communications, 11*(1), 6375. https://doi.org/10.1038/s41467-020-20064-9

62 Raju, R., Vasconcelos, O., Granger, R., & Dalakas, M. C. (2003). Expression of IFN-γ-inducible chemokines in inclusion body myositis. *Journal of Neuroimmunology, 141*(1–2), 125–131. https://doi.org/10.1016/S0165-5728(03)00218-2

63 Forsström, S., Jackson, C. B., Carroll, C. J., Kuronen, M., Pirinen, E., Pradhan, S., Marmyleva, A., ... Suomalainen, A. (2019). Fibroblast Growth Factor 21 Drives Dynamics of Local and Systemic Stress Responses in Mitochondrial Myopathy with mtDNA Deletions. *Cell Metabolism, 30*(6), 1040-1054.e7. https://doi.org/10.1016/j.cmet.2019.08.019

64 Tezze, C., Romanello, V., & Sandri, M. (2019). *FGF21* as Modulator of Metabolism in Health and Disease. *Frontiers in Physiology, 10*,419. https://doi.org/10.3389/fphys.2019.00419

65 Tyynismaa, H., Carroll, C. J., Raimundo, N., Ahola-Erkkilä, S., Wenz, T., Ruhanen, H., Guse, K., ... Suomalainen, A. (2010). Mitochondrial myopathy induces a starvation-like response. *Human Molecular Genetics, 19*(20), 3948–3958. https://doi.org/10.1093/hmg/ddq310

66 Tian, D., Sun, S., & Lee, J. T. (2010). The Long Noncoding RNA, Jpx, Is a Molecular Switch for X Chromosome Inactivation. *Cell, 143*(3), 390–403. https://doi.org/10.1016/j.cell.2010.09.049

67 Carmona, S., Lin, B., Chou, T., Arroyo, K., & Sun, S. (2018). LncRNA Jpx induces *Xist* expression in mice using both trans and cis mechanisms. *PLOS Genetics, 14*(5), e1007378. https://doi.org/10.1371/journal.pgen.1007378

68 Pan, J., Fang, S., Tian, H., Zhou, C., Zhao, X., Tian, H., ... Gong, Z. (2020). LncRNA *JPX/miR-33a-5p/Twist1* axis regulates tumorigenesis and metastasis of lung cancer by activating Wnt/β-catenin signaling. *Molecular Cancer, 19*(1), 9. https://doi.org/10.1186/s12943-020-1135-9

69 Chen, L., Yang, W., Guo, Y., Chen, W., Zheng, P., Zeng, J., & Tong, W. (2017). Exosomal lncRNA *GAS5* regulates the apoptosis of macrophages and vascular endothelial cells in atherosclerosis. *PLOS ONE, 12*(9), e0185406. https://doi.org/10.1371/journal.pone.0185406

70 Ni, W., Yao, S., Zhou, Y., Liu, Y., Huang, P., Zhou, A., ... & Li, J. (2019). Long noncoding RNA GAS5 inhibits progression of colorectal cancer by interacting with and triggering YAP phosphorylation and degradation and is negatively regulated by the m6A reader YTHDF3. *Molecular Cancer, 18*(1), 143. https://doi.org/10.1186/s12943-019-1079-y

71 Sang, L., Ju, H., Yang, Z., Ge, Q., Zhang, Z., Liu, F., ... Lin, A. (2021) Mitochondrial long non-coding RNA GAS5 tunes TCA metabolism in response to nutrient stress. *Nature Metabolism, 3*,90-106.

72 Uroda, T., Anastasakou, E., Rossi, A., Teulon, J.-M., Pellequer, J.-L., Annibale, P., ... Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 *Pathway. Molecular Cell, 75*(5), 982-995.e9. https://doi.org/10.1016/j.molcel.2019.07.025

73 Grow, E. J., Weaver, B. D., Smith, C. M., Guo, J., Stein, P., Shadle, S. C., Hendrickson, P. ... Cairns, B. R. (2021). P53 convergently activates *Dux*/*DUX4* in embryonic stem cells and in facioscapulohumeral muscular dystrophy cell models. *Nature Genetics, 53*(8), 1207–1220. https://doi.org/10.1038/s41588-021-00893-0

74 Panteghini, M. (1988). Serum isoforms of creatine kinase isoenzymes. *Clinical Biochemistry, 21*(4), 211–218. https://doi.org/10.1016/S0009-9120(88)80003-1

75 Munsat, T. L., Baloh, R., Pearson, C. M., & Fowler Jr., W. (1973). Serum Enzyme Alterations in Neuromuscular Disorders. *JAMA, 226*(13), 1536–1543.

76 Darvish, D., Vahedifar, P., & Huo, Y. (2002). Four novel mutations associated with autosomal recessive inclusion body myopathy (MIM: 600737). *Molecular Genetics and Metabolism, 77*(3), 252–256. https://doi.org/10.1016/S1096-7192(02)00141-5

77 Noguchi, S., Keira, Y., Murayama, K., Ogawa, M., Fujita, M., Kawahara, G., ... Nishino, I. (2004). Reduction of UDP-N-acetylglucosamine 2-Epimerase/N-Acetylmannosamine Kinase Activity and Sialylation in Distal Myopathy with Rimmed Vacuoles. *Journal of Biological Chemistry, 279*(12), 11402–11407. https://doi.org/10.1074/jbc.M313171200

79 Jungbluth, H., Sewry, C. A., Counsell, S., Allsop, J., Chattopadhyay, A., Mercuri, E., North, K., ... Muntoni, F. (2004). Magnetic resonance imaging of muscle in nemaline myopathy. *Neuromuscular Disorders, 14*,779–784. https://doi.org/10.1016/j.nmd.2004.08.005

80 Hankiewicz, K., Carlier, R. Y., Lazaro, L., Linzoain, J., Barnerias, C., Gómez-Andrés, D., Avila-Smirnow, D., ... Quijano-Roy, S. (2015). Whole-Body Muscle Magnetic Resonance Imaging in *SEPN1*-Related Myopathy Shows a Homogeneous and Recognizable Pattern. *Muscle & Nerve, 52*(5), 728–735. https://doi.org/10.1002/mus.24634

81 Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., Bolduc, V., … MacArthur, D. G. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine, 9,* eaal5209.

82 Vorgerd, M., van der Ven, P. F. M., Bruchertseifer, V., Löwe, T., Kley, R. A., Schröder, R., Lochmüller, H., ... Huebner, A. (2005). A Mutation in the Dimerization Domain of Filamin C Causes a Novel Type of Autosomal Dominant Myofibrillar Myopathy. *The American Journal of Human Genetics, 77*(2), 297–304. https://doi.org/10.1086/431959

83 Kölbel, H., Roos, A., Evangelista, T., Nolte, K., Johnson, K., Töpf, A., Wilson, M., ... Schara, U. (2020). First clinical and myopathological description of a myofibrillar myopathy with congenital onset and homozygous mutation in *FLNC*. *Human Mutation, 41*,1600–1614.

84 Brodehl, A., Ferrier, R. A., Hamilton, S. J., Greenway, S. C., Brundler, M.-A., Yu, W., Gibson, W. T., ... Gerull, B. (2016). Mutations in *FLNC* are Associated with Familial Restrictive Cardiomyopathy. *Human Mutation, 37*(3), 269–279. https://doi.org/10.1002/humu.22942

85 Valdés-Mas, R., Gutiérrez-Fernández, A., Gómez, J., Coto, E., Astudillo, A., Puente, D. A., Reguero, J. R., ... López-Otín, C. (2014). Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. *Nature Communications, 5*(1), 5326. https://doi.org/10.1038/ncomms6326

86 Krause, S., Aleo, A., Hinderlich, S., Merlini, L., Tournev, I., Walter, M. C., Argov, Z., Mitrani-Rosenbaum, S., & Lochmuller, H. (2007). GNE protein expression and subcellular distribution are unaltered in HIBM. *Neurology, 69*(7), 655–659. https://doi.org/10.1212/01.wnl.0000267426.97138.fd

87 Milman Krentsis, I., Sela, I., Eiges, R., Blanchard, V., Berger, M., Becker Cohen, M., & Mitrani-Rosenbaum, S. (2011). GNE Is Involved in the Early Development of Skeletal and Cardiac Muscle. *PLoS ONE, 6(*6), e21389. https://doi.org/10.1371/journal.pone.0021389

88 Pogoryelova, O., Wilson, I. J., Mansbach, H., Argov, Z., Nishino, I., & Lochmüller, H. (2019). *GNE* genotype explains 20% of phenotypic variability in *GNE* myopathy. *Neurology Genetics, 5*(1), e308. https://doi.org/10.1212/NXG.0000000000000308

89 Hauser, M. A. (2000). Myotilin is mutated in limb girdle muscular dystrophy 1A. *Human Molecular Genetics, 9*(14), 2141–2147. https://doi.org/10.1093/hmg/9.14.2141

90 Moza, M., Mologni, L., Trokovic, R., Faulkner, G., Partanen, J., & Carpén, O. (2007). Targeted Deletion of the Muscular Dystrophy Gene Myotilin Does Not Perturb Muscle Structure or Function in Mice. *Molecular and Cellular Biology, 27*(1), 244–252. https://doi.org/10.1128/MCB.00561-06

91 Lam, L. T., Holt, I., Laitila, J., Hanif, M., Pelin, K., Wallgren-Pettersson, C., Sewry, C. A., & Morris, G. E. (2018). Two alternatively-spliced human nebulin isoforms with either exon 143 or exon 144 and their developmental regulation. *Scientific Reports, 8*(1), 15728. https:/doi.org/10.1038/s41598-018-33281-6

92 Donner, K., Sandbacka, M., Lehtokari, V.-L., Wallgren-Pettersson, C., & Pelin, K. (2004). Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. *European Journal of Human Genetics, 12*(9), 744–751. https://doi.org/10.1038/sj.ejhg.5201242

93 Pitts, M. W., & Hoffmann, P. R. (2018). Endoplasmic reticulum-resident selenoproteins as regulators of calcium signaling and homeostasis. *Cell Calcium, 70,*76–86. https://doi.org/10.1016/j.ceca.2017.05.001

94 Lescure, A., Rederstorff, M., Krol, A., Guicheney, P., & Allamand, V. (2009). Selenoprotein function and muscle disease. *Biochimica et Biophysica Acta, 1790(*11), 1569–1574. https://doi.org/10.1016/j.bbagen.2009.03.002

95 Jurynec, M. J., Xia, R., Mackrill, J. J., Gunther, D., Crawford, T., Flanigan, K. M., Abramson, J. J., ... Grunwald, D. J. (2008). Selenoprotein N is required for ryanodine receptor calcium release channel activity in human and zebrafish muscle. *Proceedings of the National Academy of Sciences, 105*(34), 12485–12490. https://doi.org/10.1073/pnas.0806015105

96 Moghadaszadeh, B., Petit, N., Jaillard, C., Brockington, M., Roy, S. Q., Merlini, L., Romero, N., ... Guicheney, P. (2001). Mutations in *SEPN1* cause congenital muscular dystrophy with spinal rigidity and restrictive respiratory syndrome. *Nature Genetics, 29*(1), 17–18. https://doi.org/10.1038/ng713

97 Ferreiro, A., Quijano-Roy, S., Pichereau, C., Moghadaszadeh, B., Goemans, N., Bönnemann, C., Jungbluth, H., ... Guicheney, P. (2002). Mutations of the Selenoprotein N Gene, Which Is Implicated in Rigid Spine Muscular Dystrophy, Cause the Classical Phenotype of Multiminicore Disease: Reassessing the Nosology of Early-Onset Myopathies. *The American Journal of Human Genetics, 71*(4), 739–749. https://doi.org/ 10.1086/342719

98 Kaneva, K., Merkurjev, D., Ostrow, D., Ryutov, A., Triska, P., Stachelek, K., ... Gai, X. (2020). Detection of mitochondrial DNA variants at low level heteroplasmy in pediatric

CNS and extra-CNS solid tumors with three different enrichment methods. *Mitochondrion, 51*,97–103. https://doi.org/10.1016/j.mito.2020.01.006

99 Rensch, T., Villar, D., Horvath, J., Odom, D. T., & Flicek, P. (2016). Mitochondrial heteroplasmy in vertebrates using ChIP-sequencing data. *Genome Biology*, *17*(1), 139. https://doi.org/10.1186/s13059-016-0996-y

100 González, M. del M., Ramos, A., Aluja, M. P., & Santos, C. (2020). Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing. *Mitochondrion, 50*,88–93. https://doi.org/10.1016/j.mito.2019.10.006

101 Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., ... Dowling, J. J. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics*, *104*(3), 466–483. https://doi.org/10.1016/j.ajhg.2019.01.012

102 Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., ... Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

103 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., ... Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

104 Bucelli, R. C., Arhzaouy, K., Pestronk, A., Pittman, S. K., Rojas, L., Sue, C. M., Evilä, A., ... Weihl, C. C. (2015). *SQSTM1* splice site mutation in distal myopathy with rimmed vacuoles. *Neurology*, *85*(8), 665–674. https://doi.org/10.1212/WNL.0000000000001864

105 Lattanzi, L., Salvatori, G., Coletta, M., Sonnino, C., Cusella De Angelis, M. G., Gioglio, L., ... Cossu, G. (1998). High efficiency myogenic conversion of human fibroblasts by adenoviral vector-mediated *MyoD* gene transfer. An alternative strategy for ex vivo gene therapy of primary myopathies. *Journal of Clinical Investigation*, *101*(10), 2119–2128. https://doi.org/10.1172/JCI1505

106 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

107 Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

108 Sullivan, R., Yau, W. Y., O'Connor, E., & Houlden, H. (2019). Spinocerebellar ataxia: An update. *Journal of Neurology*, *266*(2), 533–544. https://doi.org/10.1007/s00415-018-9076-4

109 Leckie, J. N., Joel, M. M., Martens, K., King, A., King, M., Korngut, L. W., ... Schellenberg, K. L. (2021). Highly Elevated Prevalence of Spinobulbar Muscular Atrophy in Indigenous Communities in Canada Due to a Founder Effect. *Neurology Genetics*, *7*(4), e607. https://doi.org/10.1212/NXG.0000000000000607

110 DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., … Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of *C9ORF72* Causes Chromosome 9p-

Linked FTD and ALS. *Neuron, 72*(2), 245–256. https://doi.org/10.1016/j.neuron.2011.09.011

111 Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., … Traynor, B. J. (2011). A Hexanucleotide Repeat Expansion in *C9ORF72* Is the Cause of Chromosome 936-Linked ALS-FTD. *Neuron, 72*(2), 257–268. https://doi.org/10.1016/j.neuron.2011.09.010

112 Liu, Q., Georgieva, D. C., Egli, D., & Wang, K. (2019). NanoMod: A computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics, 20*(S1), 78. https://doi.org/10.1186/s12864-018-5372-8

113 Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.-L., & Wang, K. (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications, 10*(1), 2449. https://doi.org/10.1038/s41467-019-10168-2

114 Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., … Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods, 16*(12), 1297–1305. https://doi.org/10.1038/s41592-019-0617-2

115 Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods, 18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

116 Edge, P., & Bansal, V. (2019). Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications, 10*(1), 4660. https://doi.org/10.1038/s41467-019-12493-y

117 Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., Denisenko, E., … Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature, 543*(7644), 199–204. https://doi.org/10.1038/nature21374

118 Ramilowski, J. A., Yip, C. W., Agrawal, S., Chang, J.-C., Ciani, Y., Kulakovskiy, I. V., Mendez, M., … Carninci, P. (2020). Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Research, 30*(7), 1060–1072. https://doi.org/10.1101/gr.254219.119

119 Nishimoto, Y., Nakagawa, S., Hirose, T., Okano, H., Takao, M., Shibata, S., Suyama, S., ... Okano, H. (2013). The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis. *Molecular Brain, 6*(1), 31. https://doi.org/10.1186/1756-6606-6-31

120 Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An Architectural Role for a Nuclear Noncoding RNA: *NEAT1* RNA Is Essential for the Structure of Paraspeckles. *Molecular Cell, 33*(6), 717–726. https://doi.org/10.1016/j.molcel.2009.01.026

121 Mizielinska, S., Lashley, T., Norona, F. E., Clayton, E. L., Ridler, C. E., Fratta, P., & Isaacs, A. M. (2013). C9orf72 frontotemporal lobar degeneration is characterised by frequent neuronal sense and antisense RNA foci. *Acta Neuropathologica, 126*(6), 845–857. https://doi.org/10.1007/s00401-013-1200-z

122 Fattahi, S., Kosari-Monfared, M., Golpour, M., Emami, Z., Ghasemiyan, M., Nouri, M., & Akhavan-Niaki, H. (2020). LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: A novel approach to personalized medicine. *Journal of Cellular Physiology, 235*(4), 3189–3206. https://doi.org/10.1002/jcp.29260

123 Hua, J. T., Chen, S., & He, H. H. (2019). Landscape of Noncoding RNA in Prostate Cancer. *Trends in Genetics, 35*(11), 840–851. https://doi.org/10.1016/j.tig.2019.08.004

124 Schulte, C., Barwari, T., Joshi, A., Zeller, T., & Mayr, M. (2020). Noncoding RNAs versus Protein Biomarkers in Cardiovascular Disease. *Trends in Molecular Medicine, 26*(6), 583–596. https://doi.org/10.1016/j.molmed.2020.02.001

125 Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., … Marshall, C. R. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine, 20*(4), 435–443. https://doi.org/10.1038/gim.2017.119

125 Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., … Bale, S. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine, 18*(7), 696–704. https://doi.org/10.1038/gim.2015.148

127 Ghaoui, R., Cooper, S. T., Lek, M., Jones, K., Corbett, A., Reddel, S. W., Needham, M., ... Clarke, N. F. (2015). Use of Whole-Exome Sequencing for Diagnosis of Limb-Girdle Muscular Dystrophy: Outcomes and Lessons Learned. *JAMA Neurology, 72*(12), 1424. https://doi.org/10.1001/jamaneurol.2015.2274

128 Thuriot, F., Gravel, E., Buote, C., Doyon, M., Lapointe, E., Marcoux, L., Larue, S., ... Lévesque, S. (2020). Molecular diagnosis of muscular diseases in outpatient clinics: A Canadian perspective. *Neurology Genetics, 6*(2), e408. https://doi.org/10.1212/NXG.0000000000000408

129 Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences, 131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3

130 Saha Mandal, A. (2019) Predicting the Evolutionary and Medical Significance of Human Genetic Variants with Machine Learning [Unpublished doctoral thesis]. University of Calgary, Calgary, AB.

131 Madison, R. D., & Robinson, G. A. (2019). Muscle-Derived Extracellular Vesicles Influence Motor Neuron Regeneration Accuracy. Neuroscience, 419, 46–59.

132 Lam, N. T., Gartz, M., Thomas, L., Haberman, M., & Strande, J. L. (2020). Influence of microRNAs and exosomes in muscle health and diseases. *Journal of Muscle Research and Cell Motility, 41*(4), 269–284. https://doi.org/10.1007/s10974-019-09555-5

133 Yang, C., Yang, W., Wong, Y., Wang, K., Teng, Y., Chang, M., Liao, K., ...Kung, H. (2020). Muscle atrophy-related myotube-derived exosomal microRNA in neuronal dysfunction: Targeting both coding and long noncoding RNAs. *Aging Cell, 19*(5). e13107 https://doi.org/10.1111/acel.13107

134 Yuan, Z., & Huang, W. (2021). New Developments in Exosomal lncRNAs in Cardiovascular Diseases. *Frontiers in Cardiovascular Medicine, 8,* 709169. https://doi.org/10.3389/fcvm.2021.709169

135 Huang, Z., & Xu, A. (2021). Adipose Extracellular Vesicles in Intercellular and Inter-Organ Crosstalk in Metabolic Health and Diseases. *Frontiers in Immunology, 12*, 608680. https://doi.org/10.3389/fimmu.2021.608680

Appendix

Table A: Myopathy associated genes. Accessed from http://www.musclegenetable.fr

HGNC Symbol				
ABHD5	*DAG1*	*HNRNPDL*	*MYL1*	*SCN4A*
ACAD9	*DES*	*HRAS*	*MYL2*	*SELENON*
ACADVL	*DMD*	*HSPB8*	*MYMK*	*SGCA*
ACTA1	*DNAJB6*	*INPP5K*	*MYO18B*	*SGCB*
ACTN2	*DNM2*	*ISCU*	*MYOT*	*SGCD*
ACVR1	*DPM1*	*ITGA7*	*MYPN*	*SGCG*
ADSS1	*DPM2*	*KBTBD13*	*NEB*	*SLC16A1*
AGL	*DPM3*	*KLHL40*	*ORAI1*	*SLC22A5*
ANO5	*DUX4*	*KLHL41*	*PABPN1*	*SLC25A20*
B3GALNT2	*DYSF*	*KLHL9*	*PAX7*	*SMCHD1*
B4GAT1	*EMD*	*KY*	*PFKM*	*SPEG*
BAG3	*ENO3*	*LAMA2*	*PGAM2*	*SPTBN4*
BIN1	*ETFA*	*LAMP2*	*PGK1*	*STAC3*
BVES	*ETFB*	*LARGE1*	*PGM1*	*STIM1*
CACNA1H	*ETFDH*	*LDB3*	*PHKA1*	*SYNE1*
CACNA1S	*FHL1*	*LDHA*	*PLEC*	*SYNE2*
CAPN3	*FKRP*	*LIMS2*	*PNPLA2*	*TCAP*
CASQ1	*FKTN*	*LMNA*	*PNPLA8*	*TMEM43*
CAV3	*FLAD1*	*LMOD3*	*POGLUT1*	*TNNT1*
CAVIN1	*FLNC*	*LPIN1*	*POMGNT1*	*TNPO3*
CCDC78	*FXR1*	*LRP12*	*POMGNT2*	*TOR1AIP1*
CFL2	*GAA*	*MAP3K20*	*POMK*	*TPM2*
CHKB	*GBE1*	*MATR3*	*POMT1*	*TPM3*
CLN3	*GMPPB*	*MB*	*POMT2*	*TRAPPC11*
CNTN1	*GNE*	*MEGF10*	*PRKAG2*	*TRIM32*
COL12A1	*GOLGA2*	*MSTN*	*PYGM*	*TRIM54*
COL6A1	*GOSR2*	*MSTO1*	*PYROXD1*	*TRIM63*
COL6A2	*GYG1*	*MTM1*	*RBCK1*	*TRIP4*
COL6A3	*GYS1*	*MYBPC3*	*RXYLT1*	*TTN*
CPT2	*HACD1*	*MYH2*	*RYR1*	*VCP*
CRPPA	*HNRNPA1*	*MYH7*	*RYR3*	*VMA21*
CRYAB	*HNRNPA2B1*			