Predicting Becker and Duchenne Muscular Dystrophy Progression through the neuroMuscular ObserVational Research Datahub

A Major Qualifying Project submitted to the Faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfillment of the requirements for the degree of Bachelor Of Science

Justin Krittaponrattana Moy

Submitted to:

Dr. Elizabeth Ryder, Advisor

April 28, 2022

Abstract

Duchenne Muscular Dystrophy (DMD) and Becker Muscular Dystrophy (BMD) are two muscular dystrophies characterized by progressive muscle loss and weakness. Both diseases are caused by mutations in the *dmd* gene which produces dystrophin, a key structural protein in muscle cells; however, the prognosis of each disease differs drastically with monetary and logistical implications for families. This MQP aims to elucidate the uncertainty of disease severity by leveraging BMD and DMD data from the neuroMuscular ObserVational Research (MOVR) datahub and the Genotype-Tissue Expression (GTEx) dataset. For patients with whole exon deletions or duplications (n=563), the Brooke Upper Extremity (BUE) scale progression and Vignos Lower Extremity (VLE) scale progression was predicted separately using three different predictor sets: exons, isoforms, and tissues. It was found that predictive exons for BUE progression were clustered around exon 25 which corresponds to an actin binding site on dystrophin. Predictive exons for VLE were clustered around exon 60 which corresponds to the dystroglycan binding site on dystrophin.

Acknowledgements

I am grateful for the support and willingness of the Muscular Dystrophy Association throughout this project in providing and understanding the data. I would also like to thank my advisor Professor Elizabeth Ryder for helping me see this project through its many recalibrations and pivots. Lastly, I would like to thank Professor Dmitry Korkin and his lab, Professor Zheyang Wu, and Professor Gonzalo Andres Contador Revetria for helping with their domain knowledge and giving me feedback. I would also like to thank WPI for providing me the opportunity to carry out this work.

Table of Contents

List of Figures

List of Tables

Introduction

Imagine paying $51,000 more in health care expenses for your child every year (Larkindale et al, 2013). Imagine burying your child before they even turn 30. Imagine watching your child learn to run and walk only to be confined to a wheelchair by the age of 10. Your doctor tells you your child could end up anywhere on a spectrum between the above or living with relatively mild muscle weakness that persists through their entirely normal lifespan. For 1 in 3500 males globally being confined to a wheelchair by 10 and dying by 30 is not a nightmare, but the reality of living with Duchenne muscular dystrophy (DMD) (Emery, 1991).

Duchenne muscular dystrophy, which is caused by a mutation of the *dmd* gene on the X chromosome, is the most common type of muscular dystrophy. The *dmd gene* produces the dystrophin protein, a key muscle protein that maintains structural integrity during movement. DMD and its milder form Becker muscular dystrophy (BMD) are dystrophinopathies in a group of progressive muscle diseases called muscular dystrophy. As of 2020, over 110 genes are known to be responsible for a muscular dystrophy (Benarocch et al, 2019). Although the United States Food and Drug Administration considers DMD a rare disease, for those living with the disease it can feel omnipresent (NIH, 2021). In both dystrophinopathies, there is increased muscle necrosis, sarcolemmal weakness, and lack of muscle regeneration (Duan et al, 2021). Both of these diseases are debilitating, but disease progression can differ drastically between patients.

DMD manifests in early childhood with motor delays and general weakness beginning at the lower limbs and eventually progressing to the upper limbs. Generally DMD patients lose their ambulation by their early teens and die by the age of 28 (Broomfield et al, 2021). Other symptoms include cardiomyopathy, decreased lung function, calf hypertrophy and waddling gait (Yiu and Kornberg, 2015). About one third of those living with DMD also have cognitive deficiency (Thangaranjh et al, 2019). In contrast, BMD symptoms are much milder, with an age of onset around 12 years of age and a normal life expectancy (Emery, 2002; Capitanio et al, 2020). BMD occurs at a rate of 1.6 per 100,000 people (Salari et al, 2022).

When measuring the severity of BMD and DMD two scales can be used: the Brooke Upper Extremity (BUE) Scale and the Vignos Lower Extremity (VLE) Scale. The Brooke Scale was developed in 1981 to measure clinical trial effectiveness. The scale progresses from 1-6 with 1 being "patient can abduct the arms in a full circle until they touch above the head" and 6 being

"patient cannot raise hands to the mouth and has no useful function of the hands" (Brooke et al, 1981). The Vignos Scale was developed in 1963 to monitor pediatric patients. The scale progresses from 1-10 with 1 being "walks and climbs stairs without assistance" and 10 being "is confined to bed" (Vignos et al, 1963).

For those living with dystrophinopathies, improvements in quality of care have led to a longer lifespan, but without an underlying cure, those unprepared for the longer lifespan may face unexpected issues. As DMD and BMD patients live longer, caregivers also have to deal with growing older and planning for the care of their children (Yamaguchi and Suzuki, 2015).

Research into dystrophinopathies has been slowed by the complexity of the *dmd* gene. The *dmd gene* is the largest gene in the human genome with a size of about 2,200 Kb and contains 79 exons located on the X-chromosome (NCBI, 2021). The large size of the gene combined with its complex splicing makes a myriad of targets for pathogenesis. A mutation of any of the exons or intronic regulatory elements could lead to disease anywhere on the spectrum between severe DMD and mild BMD. It is of no surprise that it took 28 years from the discovery of dystrophin in 1988 to develop a drug to treat the disease, Exondys 51, and even then, it only helps patients with a mutation in exon 51 (Sarepta, 2016).

With the understanding that mutations in dystrophin can lead to a wide range of phenotypes and the reality of slow pharmaceutical development, the focus shifts from curative measures to preparative ones. The ability to predict disease progression based on noninvasive measures would help prepare families for the life ahead of them. Using the longitudinal patient data and genetic information of DMD and BMD patients from the Muscular Dystrophy Association's (MDA) neuroMuscular ObserVational Research (MOVR) datahub, this MQP aims to use the genetic data to predict disease severity.

Background

Normal Skeletal Muscle Function

Skeletal muscles' purpose is to generate contractile force which is used in our everyday tasks like walking or eating. Skeletal muscle itself is composed of bundles of muscle fibers which themselves contain many myofibrils. For a muscle to contract, a nerve signal is sent to the neuromuscular junction where it triggers the contractile unit called the sarcomere to contract. A sarcomere is primarily made up of actin, myosin, and the Z-disk (Figure 1). When myosin is bound by ATP, it pulls on the actin filament pulling the Z-disks closer together and thus contracting the muscle. Surrounding these sarcomeres is the sarcolemma, a membrane that connects to the extracellular matrix (ECM). In order to properly contract and prevent damage, the movement of sarcomeres must be coordinated with the sarcolemma so that force is transmitted equally (Mukund and Subramanian, 2019). This is the role of the costamere. The costamere is composed of two complexes, the dystrophin associated protein complex (DAPC) and the focal adhesion complex (FAC) (Figure 1). The focal adhesion complex binds the Z-Disk while the DAPC complex binds to f-actin. Both of these complexes bind to the sarcolemma and the ECM (Feher, 2017).

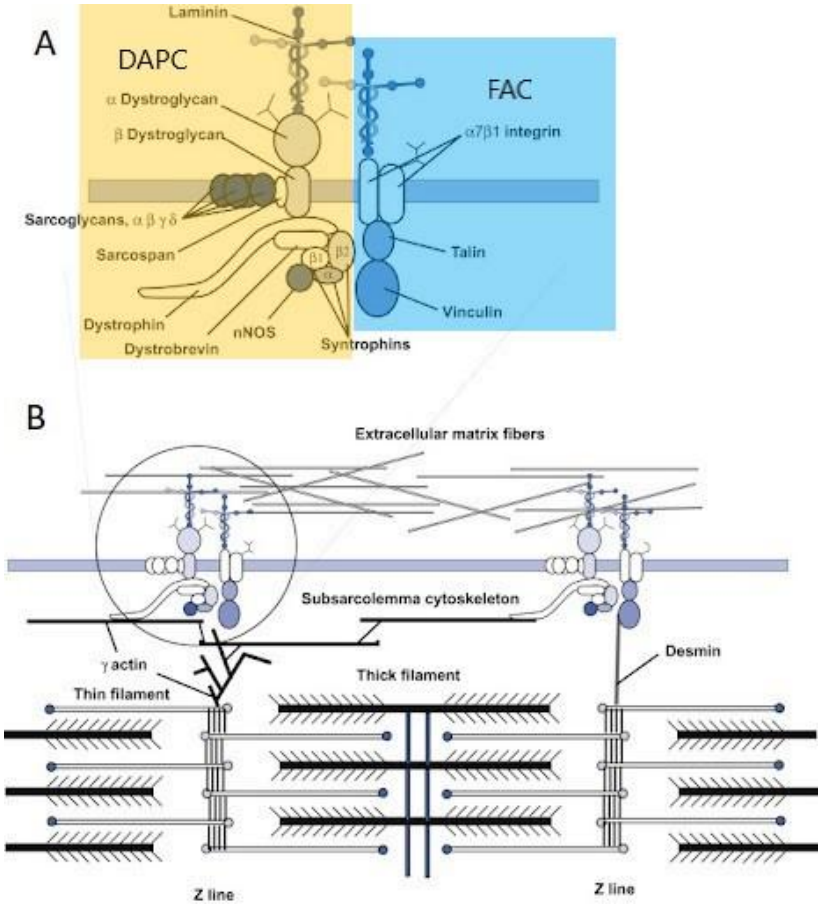

Figure 1: ***Structural components of muscle fibers***. *Figure 1 A) Protein components of the dystrophin associated protein complex (DAPC) and the focal adhesion complex (FAC). B) The DAPC and FAC together form the costamere, which transmits mechanical force between the sarcomere and the extracellular matrix (ECM). The sarcomere is made up of the Z-line, f-actin (thick filament), and myosin (thin filament). The DAPC complex binds to actin while the focal adhesion complex while FAC binds to the Z-line. Image adapted from (Feher, 2017)*

Dystrophinopathic Skeletal Muscle

Seeing the important role that the DAPC complex has in transmitting mechanical stress, it is of no surprise that disruptions to both the costamere and connection to the ECM leads to muscle inflammation, necrosis, and fibrosis. We can see this damage directly by comparing the images of healthy muscle tissue to a human with DMD (Figure 2). In the patient with DMD, the lack of structure in the hematoxylin and eosin staining indicates hemorrhaging (Figure 2a, e). In the stain of connective tissue and muscle fiber, the increased blue indicates fibrosis, which is the replacement of muscle fibers with connective tissue (Figure 2b, f). In normal muscle, dystrophin

colocalizes with laminin (Figure 2c, d). As shown in *Figure 1A*, dystrophin and laminin associate in the same complex; in the absence of dystrophin laminin loses its ability to anchor the muscle fibers to the ECM (Figure 2g, h).

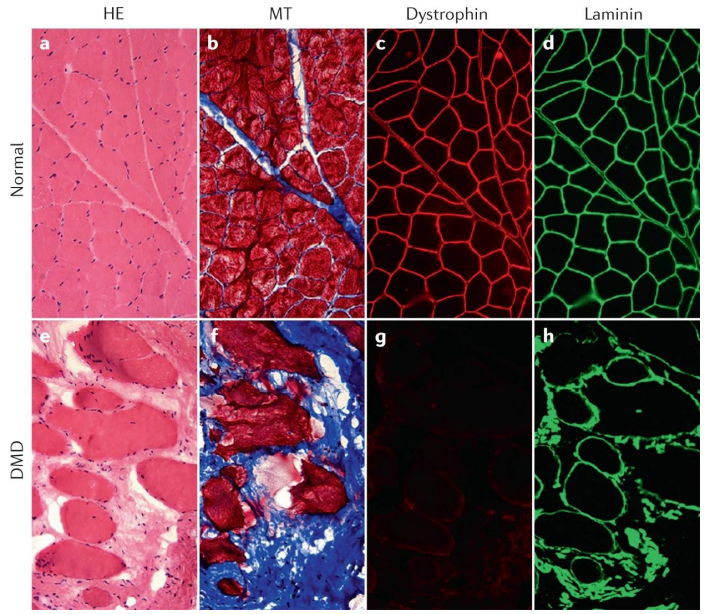

Figure 2: Comparison of normal and DMD muscle cells*. Panel a and e are stained with hematoxylin and eosin. Nuclei are stained purple, and the ECM and cytoplasm are stained pink. Panel b and f are stained with masson trichrome. Connective tissue is stained blue and muscle fibers are stained red. Panel c and g stain dystrophin in red. Panel d and h show laminin stained in green. Both the dystrophin and laminin pictures are captured using immunofluorescence Images from Duan et al, 2021*

From a mechanical perspective, this inherent weakness is exacerbated by muscle contraction which makes dystrophinopathies progressive. In a study on the *mdx* mouse, a mouse model missing dystrophin, mechanical contractile stress was placed on the cells. The authors found that although in both wildtype and *mdx* mice muscle damage occurred after muscle contraction, the sarcolemma was much weaker and led to a higher percentage of damage in *mdx* mice (Petroff et al, 1993). Without functioning dystrophin, this stress leads to progressive muscle damage. In a study on *mdx* mice, mice hindlimbs were restrained for 14 days. The mice with restrained hind limbs showed decreased degeneration compared to unrestrained mice (Mokhtarian et al, 1999). This combined with a finding that dystrophin helps organize the costamere points to its structural role (Rybakova et al, 2000).

DMD Gene

Dystrophin is encoded by the *DMD* gene located on the X chromosome. The gene itself contains 7 promoters, three upstream and four internal (Figure 3). The three upstream promoters produce full length dystrophins Dp427m in muscles, Dp427b and Dp427p in the brain with Dp427p being produced specifically in the purkinje cells. Additionally the internal promoters produce smaller dystrophin isoforms with a functional cysteine rich and C-terminal domain (Muntoni et al, 2003). Although Dp stands for dystrophin protein and the following number refers to the molecular weight, in the literature that notation can refer to either the protein or the transcript. DMD isoforms that occur in skeletal muscle and brain tissue are of particular interest due to the muscle degeneration and cognitive symptoms associated with dystrophinopathies. Muscle cells produce full length dystrophin which includes all 79 exons; however, some alternative splicing events have been reported. In a recent RNA-Seq study, the researchers found 12 alternative splicing events (ASE). Of note 66% of the 12 ASEs preserved the reading frame (Bougé et al, 2017). Although not much is known about the function of these ASEs in muscle cells, the deletion of exon 71 and 78 is of note. This transcript, also called Dp71 is expressed in embryonic skeletal muscle in multiple species, but not expressed in adult skeletal muscle. Interestingly this transcript is also expressed in adult brain cells (Tennyson et al, 1996). Recently studies have been performed on dystrophin isoforms in the brain that have overturned previous research. A 2017 study found that Dp427p was not expressed in purkinje cells in the human brain, but rather only in mice brains. This same study found that not only were isoforms Dp140, Dp71, and Dp40 heavily expressed in the brain, but were also significantly coexpressed with genes responsible for cognitive disorders (Doorenweerd et al, 2017). Not only has coexpression been found, but another study found correlation between mutated exons and Full Scale IQ score (Taylor et al, 2010). Introns constitute 99.5% of the gene. Even mutations in intronic regions can lead to pathogenic mutations highlighting the importance of correct splicing in protein function. Even in healthy individuals, complicated splicing activity occurs like exon skipping and intron retention (Tuffery-Giraud et al, 2017).

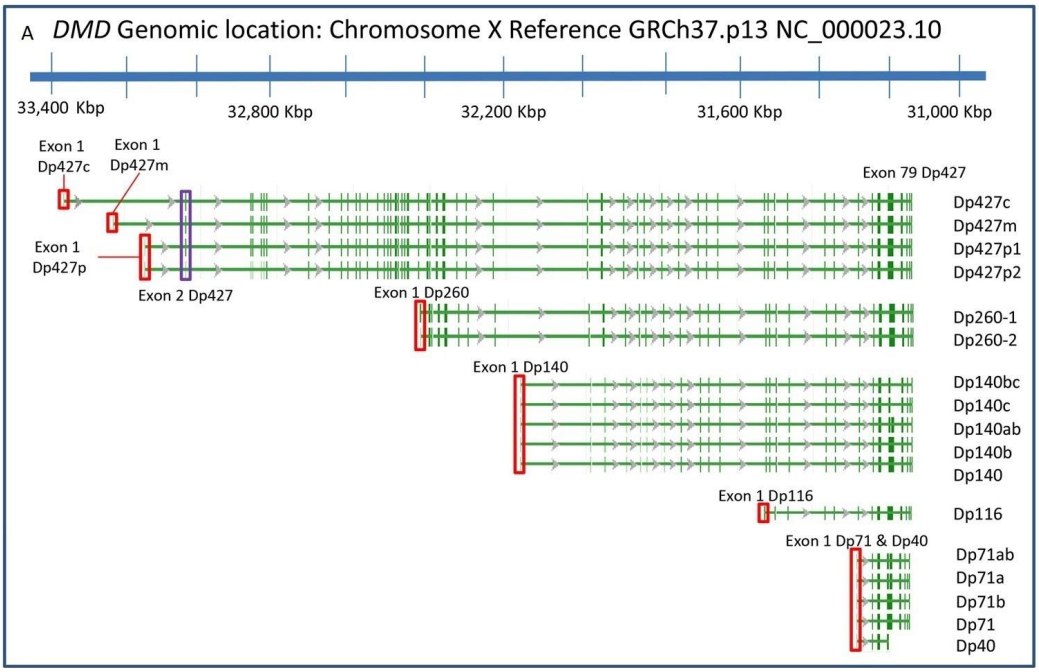

Figure 3: DMD isoforms present in humans. *Although each isoform is labeled starting at exon 1, each exon 1 for each category is a different exon. DMD Isoforms fall into 5 major categories: Dp427, Dp260, Dp140, Dp116, and Dp71. Image from [Doorenweerd et al, 2017]*

Dystrophin Protein

Due to alternative splicing, the dystrophin protein has many isoforms of varying lengths (Figure 4). The full length protein found in muscle is 3685 amino acids long and contains 4 domains: the N-terminal actin-binding domain, the rod domain, the cysteine-rich domain, and the C-terminal domain (Koenig et al, 1988).

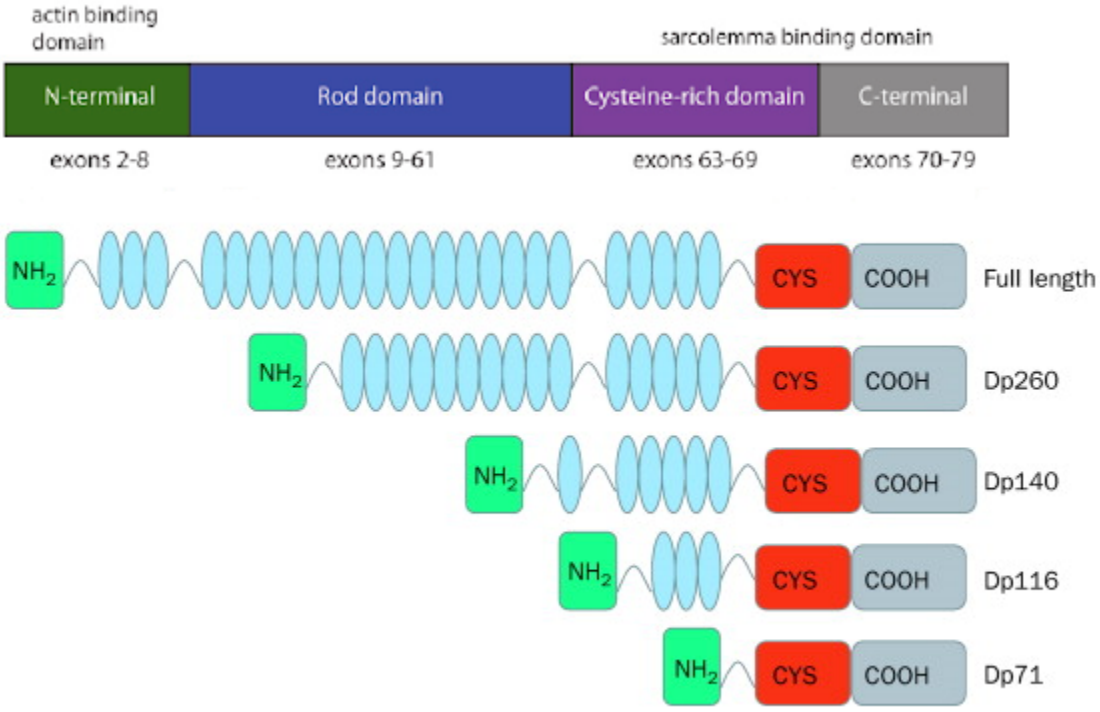

Figure 4: Domains of dystrophin. *This figure shows the four domains of dystrophin along with the exons that each domain contains. The N-terminal domain contains an actin binding domain. The rod domain is the longest, but also the most redundant. The cysteine rich and c-terminal domain provide important scaffolding for signaling proteins and sarcolemmal binding sites. The blue ovals represent spectrin repeats in the rod domain. Adapted from Muntoni et al, 2003 and Annuar et al 2010*

N-terminal Actin Binding Domain

The N-terminal actin binding domain (NABD) of dystrophin binds with f-actin in a one to one ratio (Rybakova et al, 2006). F-actin is the fibrous polymerized form of gamma-actin that, in muscle cells, forms both the sarcomere and cytoskeleton (Dominguez and Holmes, 2011). The NABD contains two important calponin homology (CH) domains, CH1 and CH2 which directly

bind to f-actin (Figure 5). This domain is named such because it was first discovered in the protein calponin. CH1 directly binds to f-actin while CH2 increases binding affinity of CH1 by guiding CH1 to the correct location (Yin et al, 2020). Despite knowing the general function of the NABD, there is still controversy regarding its exact structure *in vivo* with one model proposing dimerization in solution and another proposing a monomer in solution (Norwood et al, 2000; Singh and Mallela, 2012).

Rod Domain

The next domain of the dystrophin protein is the rod domain which is composed of 24 spectrin repeats and 4 hinge regions (Figure 5). The spectrin repeats primarily bind to the sarcolemma, but also bind to lipids and neuronal nitric oxide synthase. The spectrin repeats also contain a second actin binding domain (Legardinier et al, 2009; Lai et al, 2013). Each hinge area is rich in proline which allows for more flexibility during muscle contraction. These features are consistent with dystrophin's function as a scaffolding protein (Koenig and Kunkel, 1990).

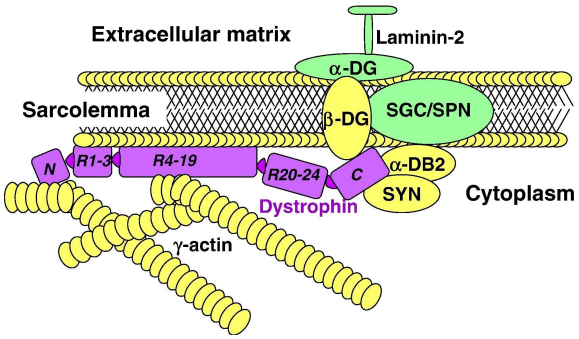

Figure 5: Dystrophin associated protein complex. *Notice that along the rod domain, dystrophin binds to both actin and the sarcolemma. Hinges in the rod domains provide flexibility. C, C-terminal domain. DG, dystroglycan. SGC, sarcoglycan. SPN, sarcospan. db, dystrobrevin. SYN, syntrophin. Adapted from Legardinier et al 2009*

Cysteine Rich Domain

Following the rod domain is the cysteine rich domain. This domain actually contains four domains, a WW domain, two EF hands, and a ZZ domain (Ervasti, 2007; Fletcher et al, 2010). Taken together these domains bind to and stabilize the interaction of dystrophin and *β*-dystroglycan. The WW domain, named for its two conserved tryptophans, bind to

β-dystroglycan while the two EF hands alter the conformation of the WW domain to allow binding (Rentschler et al, 1999). Lastly, the ZZ domain stabilizes the interaction by binding zinc at the region where dystrophin binds β-dystroglycan (Hnia et al, 2007).

C-terminal Domain

The last domain is the C-terminal domain which contains the binding site for syntrophins and dystrobrevin. This domain contains two coiled coil regions, the second of which binds to dystrobrevin (Morris et al, 1998). Additionally, the region that contains the syntrophin binding site is contained in a region of alternative splicing which may point to a lack of necessity of syntrophin (Suzuki et al, 1995).

Dystrophin Associated Protein Complex

Dystrophin is a key component of the dystrophin associated protein complex (DAPC) which contains several of the proteins mentioned in previous sections. Proteins of the DAPC include dystroglycan, dystrobrevin, syntrophin, sarcoglycan, sarcospan, and laminin. Within the DAPC are two dystroglycans α and β which come from a single polypeptide that is then cleaved (Holt et al, 2000). Dystroglycans connect dystrophin to laminin. the Transmembrane β-dystroglycan binds directly to dystrophin as described previously while extracellular α-dystroglycan binds to both laminin and β-dystroglycan. α-dystroglycan is also thought to help facilitate acetylcholine receptor formation. Dystrobrevins are intracellular proteins that are implicated in acetylcholine receptor stability and neuromuscular junction formation (Gawor and Proszynski, 2017). Syntrophins act as adaptor proteins that organize the cytoskeleton and recruit signaling molecules (Iwata et al, 2004).

Both sarcoglycans and sarcospans are transmembrane proteins. There are four types of sarcoglycans that pass through the sarcolemma, α-δ. The sarcoglycans only function correctly if all four of the sarcoglycans are produced simultaneously and correctly. These proteins help relieve mechanical stress and connect the sarcolemma to α-dystroglycan. Mutations in each of the sarcoglycans lead to different types of limb girdle muscular dystrophy (Holt and Campbell, 1998). Connected to the sarcoglycans is sarcospan. Sarcospan is another transmembrane protein that forms a subcomplex with sarcoglycans. Sarcospan plays a role in the glycosylation of α-dystroglycan and the regeneration of muscles through the Akt pathway (Marshall et al, 2012).

Lastly, laminin is an extracellular protein that makes up a large part of the basement membrane. Laminin binds to α-dystroglycan and plays a role in PI3K activation which promotes cell survival (Langenbach and Rando, 2002). Laminin also increases cell membrane integrity. In an experiment by Han et al, mutations in muscles that prevented the binding of laminin to α-dystroglycan had an absolute force 200mN less than the wild type (Han et al, 2009). Mutations in laminin-2 lead to the muscular dystrophy MDC1A (Helbling-Leclerc et al, 1995).

Predicting Phenotype

Despite the large possibility for mutations to occur in the lengthy intronic regions, a majority of pathogenic *dmd* mutations are whole exon duplications or deletions. In a study that used the TREAT-NMD DMD Global Database, which contains 7149 mutations, 86% were either deletions or duplications with most deletions on either exon 45 through 55 or 2 through 20. Only 22 mutations were in mid intronic regions. Point mutations accounted for 10.57% of mutations (Bladen et al, 2015). Although the study described above only details DMD, dystrophin mutations can cause both DMD and BMD. In most cases the phenotype of a mutation is predicted by the reading frame rule. Reading frame refers to how DNA is transcribed into RNA. During transcription, the DNA is read in groups of three base pairs; however, there can be three different frames that can be used (Figure 6).

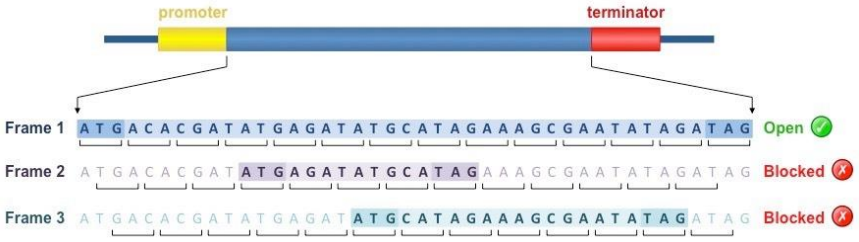

Figure 6: Reading frames on a gene. *On a sequence of DNA there are three reading frames. Each reading frame is shifted over by one base pair, hence if the starting base was shifted over 3 times, the reading frame would return to the first one. By changing the reading frame, the groups of three base pairs called codons change which may produce completely different proteins. Additionally changes in reading frame may lead to premature stop codons which will create truncated proteins. Image from Cornell, 2016*

The reading frame rule for BMD and DMD states that out of frame mutations will cause DMD while in frame mutations will lead to the less severe BMD phenotype (Monaco et al,

1988). In addition to the reading frame rule, the location of the mutation also changes the phenotype. BMD is seen in indel mutations of the actin binding domain, deletions in the central rod domain, and truncations after exon 74. On the other hand, DMD phenotypes are associated with mutations in the cysteine rich domain and C-terminal domain (Aartsma-Rus et al, 2006).

Recent studies have also pointed to some BMD cases being caused by changes in splicing regulation. In one instance a splice enhancer was converted into a splice silencer due to a nonsense mutation. Although the mutation created a premature stop codon in exon 25 which would predict DMD, the patient showed a BMD phenotype. Further measurements found that both a truncated transcript and a transcript missing exon 25 were present. The study constructed a minigene in both HeLa and mouse cells to understand how this mutation was leading to the BMD phenotype despite creating a stop codon. The authors found the mutation prevented the binding of a splice enhancer which prevented the premature stop codon from being integrated into the final transcript. Thus the reading frame was maintained and a full length protein was produced sans exon 25 (Zhu et al, 2019).

Considering the range of genotypes and phenotypes, one would think that machine learning could be leveraged for predictive analysis; however, focus has been given genotype-phenotype correlation after diagnosis rather than prediction during diagnosis. A 2011 study found that mutation region and type primarily affected age of onset in BMD and cognitive deficiencies in both BMD and DMD; however, this correlation was found only after symptoms manifested themselves (Magri et al, 2011). In a study in India, phenotype and genotype were correlated, but only at a specific time point with patients already showing symptoms (Vengalil et al, 2017). In these studies, a formal predictive model was not built; however, machine learning has been applied in discriminating DMD and BMD. In one study 2536 MRI images were used to train a model to detect whether a patient had dystrophinopathy. The model outscored three human judges with an accuracy of 91%. Upon this success, the researchers then tried to classify DMD and BMD with a 94% and 82% accuracy respectively (Yang et al, 2021).

The lack of predictive modeling on non-image clinical data presents an opportunity to create a useful predictor for the severity of a disease rather than just a disease class or probable correlations. By using clustering, and classification algorithms on the MDA MOVR dataset a more robust model can be built.

Methods

Datasets

neuroMuscular ObserVational Research (MOVR)

MOVR is an observational data hub maintained by the Muscular Dystrophy Association. The database was started in 2018 and collected its initial data from the US Neuromuscular Disease Registry. It currently collects data from 50 locations across the country and tracks 7 diseases: ALS, BMD, DMD, FSHD, LGMD, Pompe, and SMA. In the context of this MQP the BMD and DMD datasets will be used since the difference between the diseases is only phenotypical and the diseases are caused by the mutations on the *dmd* gene.

DEMOGRAPHICS FORM	DIAGNOSIS FORM	ENCOUNTER FORM	Spinal Conditions, Neuroimaging
Disease Type	Date/Age of Diagnosis	Encounter Date	Nutritional & GI Therapies
Enrollment Date	Clinical Diagnosis	Height/Weight	Pulmonary, Cardiology
Gender/DOB/Race/Eth	Muscle Biopsy	Clinical Trial Participation	Multidisciplinary Care Info
Insurance Information	Body Region Affected	Surgical History	
Education Information	Family History	Falls/Hospitalizations	DISCONTINUATION FORM
Employment Info	Molecular/DNA Results	Medications	Cause/Date
	Gross/Developmental Motor Milestones	Mobility & Mental Status	Reason for Study Withdrawal
		Assistive Devices	Date of Death
		Disease Progression	Cause of Death

Figure 7: Datasets in MOVR. *The MOVR dataset is broken into 4 Domains: Demographics, Diagnosis, Discontinued, and Encounters. Each domain came as a Microsoft Excel file. Figure from MOVR Slides*

Genotype-Tissue Expression Project (GTEx)

GTEx is a resource run by Broad Institute that has high quality transcriptome data for a multitude of genes including the *dmd* gene. The data was collected from over 1000 healthy individuals in 54 tissues. It contains tissue specific gene expression levels, eQTLs, and expression data of different transcripts. The RNA sequence data itself is inaccessible through GTEx; however, it does exist in a visualized format on the website. For the purposes of this project the tissue specific transcript expression data was downloaded.

Preprocessing

To preprocess the MOVR data, each domain was converted into a dataframe in R. Several boilerplate columns were dropped and for the Demographic, Diagnosis, and Discontinuation domain, the patient and facility ID were concatenated to form the primary key. For the Encounter domain, patient, facility, and case ID were concatenated to form the primary key.

To distill each patient's diagnosis into a computer usable format only patients that had indel mutations were selected (n=563). For each patient the start and end of their mutation was used to create a list of affected exons. Next for each of these patients a row was created with each column being an exon. The exons were then marked as zero for an unaffected exon, and one for an affected exon. In addition to marking affected exons, this table was then used to create a vector of affected isoforms with zero representing unaffected and one representing affected exons. This matrix was created by using the affected exons for patients. If an affected exon was contained in the isoform, then the isoform was marked with a one.

For the encounter dataset several metrics were dropped initially leaving only 34 columns remaining. Binary categories were converted into zeros and one, and scales were converted into numbers.

To create a dataset for the prediction phase, two columns were used: Brooke Arms and Shoulders, and Brooke Hips and Legs. These columns correspond to the Brooke Upper Extremity Scale and the Vignos Lower Extremity Scale respectively. For each of these metrics, the dates of birth from the Demographics dataframe and encounter dates were used to determine the age at each encounter. Next patients were selected that had more than one measurement of each metric over time. This led to a dataset with n=297 for the Brooke Scale and n=301 for the Vignos Scale.

Even though the GTEx data contained only transcript expression data for healthy individuals, an attempt was made to combine the data to account for how tissues would be affected by various exon mutations. To do this, 16 transcripts were chosen from the GTEx dataset that corresponded to known *dmd* transcripts in humans. Then using the table of affected isoforms for each patient, the dot product was taken with the matrix of transcript expression for each tissue. This created a table for each patient of how affected each tissue was with zero being the least affected and the sum of each transcript's Transcripts Per Kilobase Million (TPM) in a tissue being the most affected.

Classification

The datasets used by the machine learning algorithms were the affected isoforms and the encounter dataset. For the encounters the algorithms were run with and without in frame or out of frame as a feature. For classification algorithms a stratified 20%-80% test train split was used. When labels were necessary, BMD and DMD were used as labels.

Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is used to project multidimensional data into a two-dimensional visualization that shows clustering. UMAP represents the data as an N-dimensional graph with nodes being values and the number of neighbors being the edge. This graph is then projected into a lower dimension (McInnes et al, 2018). The UMAP python library was used to create a UMAP visualization. The UMAP projection was used twice, once with the affected isoform data and then again with the encounter data. In both graphs, the points were colored by whether they were BMD or DMD.

Support Vector Machine

Support Vector Machines work by creating the largest margin between clusters of classes. SVM creates linear separability by projecting the data into a higher dimension if not immediately separable through a kernel function (Boser, Guyon, and Vapnik, 1992). The standard scikit-learn implementation using the RBF kernel was used on the encounter data to discriminate between DMD and BMD.

Logistic Regression

Logistic regression uses a sigmoid function to create a separation between classes. The standard scikit-learn implementation of logistic regression with and without principal component analysis was used on the encounter data to discriminate between DMD and BMD.

Decision Tree and Random Forest

A decision tree works by splitting classes based on individual columns until all points are classified. In scikit-learn's default implementation, the gini index is used to choose which column to split on first. It works by choosing columns that will be the most discriminatory. Random forest works by creating multiple decision trees on different features. The class chosen is the class most trees choose (Ho, 1995). Scikit learn's default implementation was used to classify the encounter dataset.

Prediction

For the prediction phase, two metrics were used as the response variable for forward stepwise linear regression: the linear slope *m* and the logistic growth rate *k* (Table 1).

Line	Logistic Curve
$y = mx + b$	$y = \frac{L}{1+e^{k(x-x_0)}}$

Table 1: Equations containing response variables for slope and logistic growth rate .

These two values were chosen to account for the different progressions that appeared in patients. Some patients remained constant while others only increased or decreased, and some would remain constant, progress, and then plateau again. To generate both *k* and m, the Brooke Upper Extremity Scale and Vignos Lower Extremity scale were used because of their high coverage among patients with whole exon indel mutations. For each of these patients, their time of encounter was converted to their age and the score was plotted against age. Then either a linear or logistic curve was fitted to those points. *k* and *m* were then extracted and used as a response variable for the exon, tissue, and isoform dataset. For each of these regressions, four metrics were used as a measure of fit: Adjusted R^2, Mallows Cp (Cp), Schwarz Information Criterion (SBC), and Akaike Information Criterion (AIC) (Table 2). They were chosen because they all incorporate the number of features used in the regression.

Adjusted R^2	$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1}\right]$
Cp	$C_p = \frac{SSE_p}{MSE_p} - (n - 2p)$
SBC	$SBC = nlog\left(SSE_p\right) - (n - p)log(n)$
AIC	$AIC = nlog\left(SSE_p\right) - nlog(n) + 2p$

Table 2: Formulae for regression metrics. *This table shows the formula for each metric used to evaluate regressions. n is the number of rows and p is the number of features.*

Results

Data Summarization

As of 2022, this is the first report to specifically explore the genetic data contained within the BMD and DMD datasets. Therefore exploratory summarizations were first executed to understand what could be analyzed. First, to understand the types and distributions of mutations, two visualizations were created. The first visualization shows the number of mutations in the dataset by frame type and type of mutation. As shown in the figure below, most mutations are either deletion or duplication mutations. Notice that the most mutations are out of frame mutations, which corresponds to the large number of DMD patients in the dataset. Additionally, this dataset contains a lot of missing data in regard to frame type since only deletion and duplication mutations contain information on frame type.

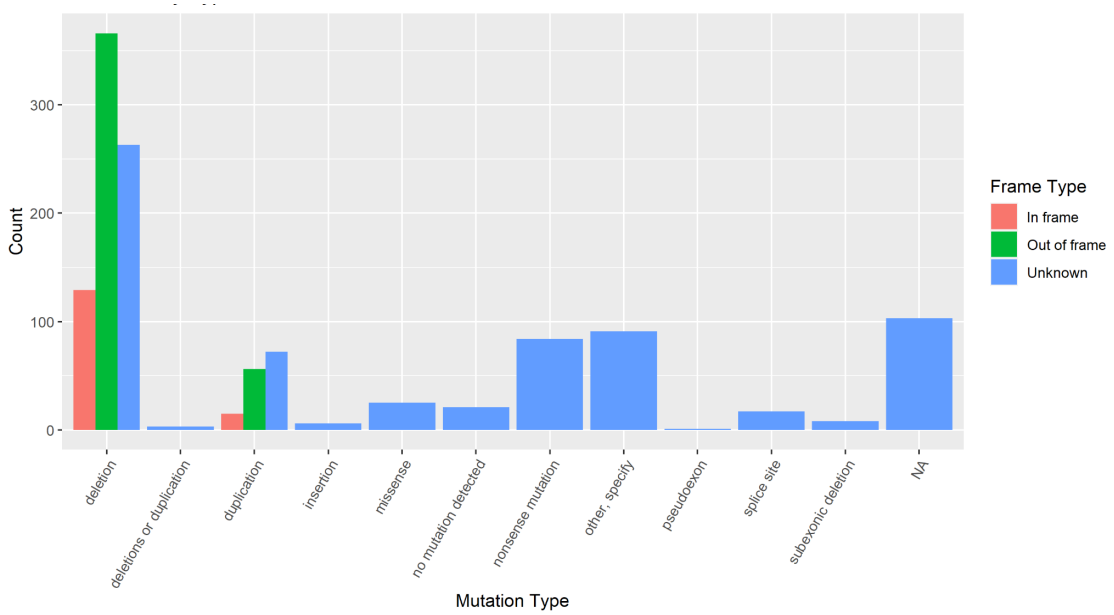

Figure 8: Mutations by type and frame. *Types of mutations are shown on the x-axis and the number of patients with that mutation type on the y-axis. For insertion and deletion mutations, the mutations are further divided by frame type.*

Figure 9 shows the number of mutations across all exons. Most BMD mutations are in frame while most DMD mutations are out of frame. Additionally, there is a large number of mutations between exon 45 and exon 60.

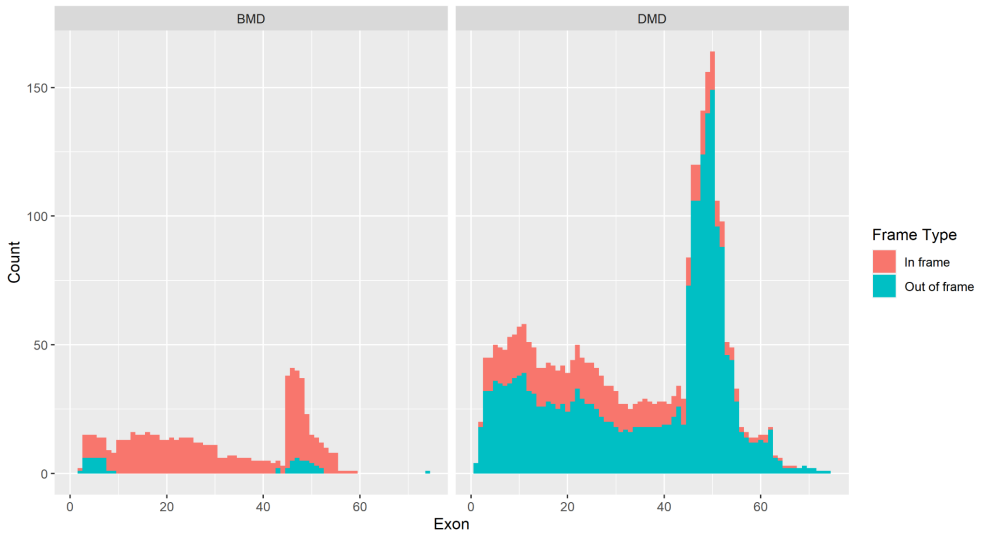

Figure 9: Distribution of affected exons. *The x-axis is the exon number and the y-axis is the number of individuals with mutations at that location. The figure is further divided by DMD and BMD.*

In addition to understanding the location and types of mutations. It was also important to understand the number of encounters that each patient had with the medical system. Figure 10 shows the number of visits each person has, classified by disease type. As expected, the number of patients with each visit number decreases as the number of visits decreases.

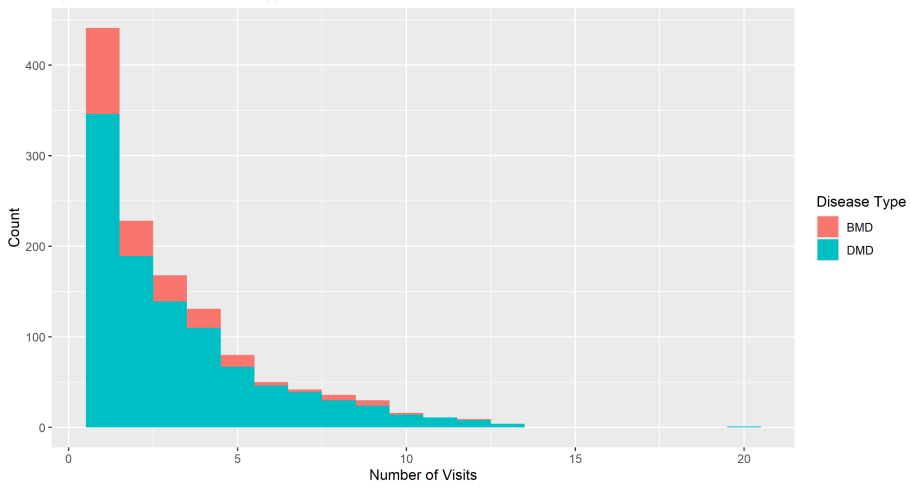

Figure 10: Distribution of visit count by individual and disease type. *The x-axis shows the number of visits, and the y-axis shows the number of patients on the y-axis.*

Next the GTEx TPM data was summarized to find out an ideal cutoff for tissue expressions. However, no obvious cutoff was observed, and because even genes with low

expression may be significant, a cutoff was ultimately not used. Figure 11 shows that for each of the 288 TPMs (18 tissues and 16 transcripts), most values were between zero and one.

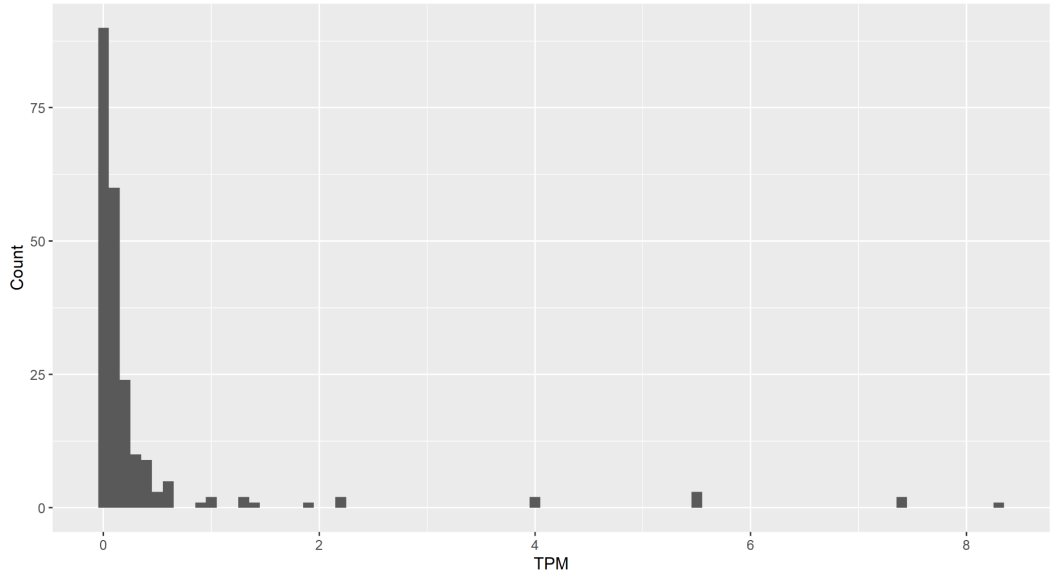

Figure 11: Distribution of TPM values. *In the above figure TPM is represented on the x-axis while the count of transcripts that fell in that TPM is on the y-axis. The bin size for TPM is 0.1.*

Classification

UMAP

UMAP was used on the encounter, exon, isoform, and tissue dataset to first explore if there was any inherent structure to the data. It was expected that individuals with similar features including disease type would cluster together. This was based on the assumption that people with similar mutations or disease symptoms may have the same disease; however, projection of both the encounter and exon dataset showed no clear structure.

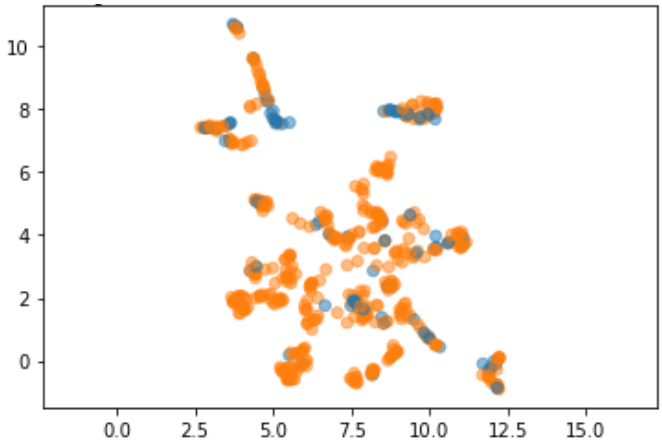

Figure 12: UMAP projection of encounter dataset. *This figure shows each patient as a point. There is no clear clustering. The points are colored based on which disease they have: DMD or BMD.*

The UMAP projections of the isoform and tissue datasets were identical. The clusters were very tight, but did not align with specific disease types. Figure 13 shows UMAP with frame type included as a feature with eight clusters apparent; however UMAP without frame type as a feature produced five clusters which aligns with the general number of transcript groups present in humans. Considering that when frame type, either in or out of frame, is included as a feature, the number of clusters would almost double. Additionally it makes sense that the tissue and isoform data would produce identical clusters since each patient was multiplied by the identical matrix of tissue TPMs.

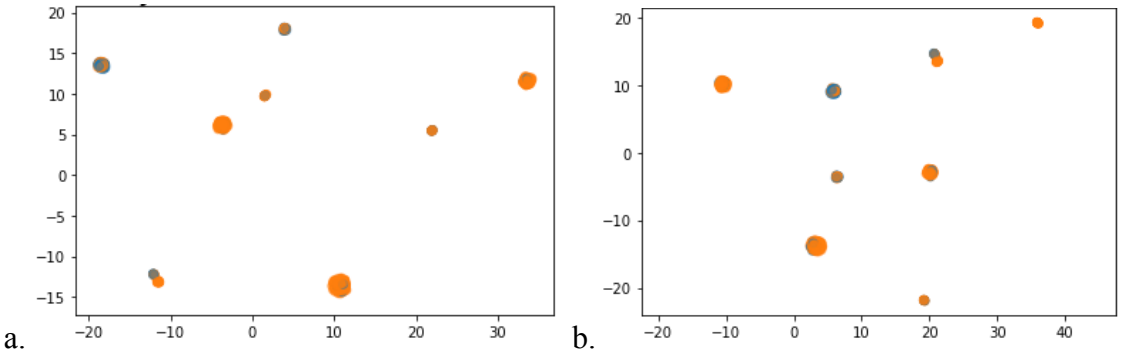

Figure 13: UMAP projections of isoform and tissue dataset. *a) isoform dataset b) tissue dataset. These panels show each patient as a point. The points are colored based on which disease they have: DMD or BMD.*

Machine Learning Algorithms

In order to determine if the data was useful for doing simple classification into disease types, SVM, logistic regression, decision tree, and random forest were all run at the encounter, exon, isoform, and tissue level. Table 3 shows the f-1 macro scores generated by each algorithm for each predictor type. f-1 macro is calculated by taking the unweighted average f-score for each class of data where the f score is $2\frac{precision \bullet recall}{precision+recall}$.

	Encounter	Exons	Isoforms	Tissues
SVM	0.4555	0.7554	0.7288	0.4555
Logistic Regression	0.6271	0.7358	0.7555	0.4555
Decision Tree	0.6346	0.7852	0.7480	0.4555
Random Forest	0.6673	0.7916	0.7497	0.4555

Table 3: f-1 macro scores for different predictors and algorithms.

Over the course of multiple stratified 5-fold cross validations on the models, exons were the best dataset to classify disease type. Decision trees and random forest performed about the best for the encounter and exon data, with only very small differences among algorithms for the isoforms data. Tissue data performed poorly as a predictor regardless of algorithm.

Prediction

Following the use of exon, isoform, and tissue data to classify between BMD and DMD, it was determined that exons performed the best, so exon data was used to predict the progression of the disease. On average each patient had 4.5 encounters spanning over 3 years. For the purposes of measuring progression, two metrics were used: the Brooke Upper Extremity Scale (BUE) and the Vignos Lower Extremity Scale (VLE) because of their high coverage. Graphing age against scales shows 3 cases: Linear change, constancy, or logistic change.

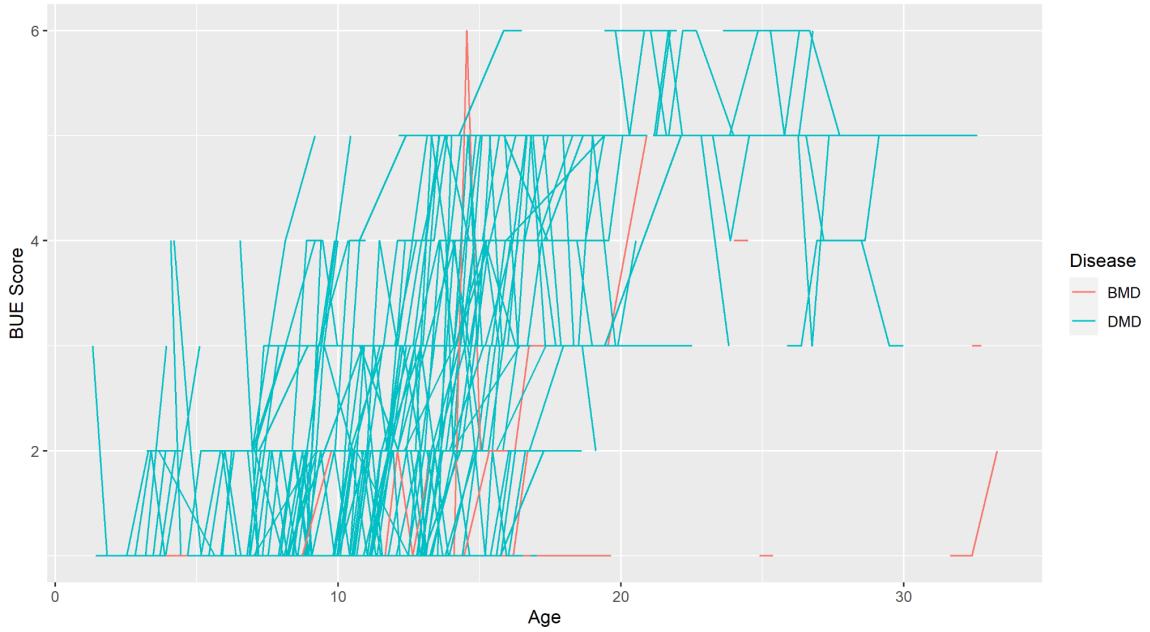

Figure 14: BUE score vs. age by patient and disease type. *This figure shows the progression of each patient on the Brooke Upper Extremity Scale as they increase in age. In this graph each patient is colored based on their disease rather than by individual.*

Although it is difficult to see when all patients are graphed together, several different patterns are observed: some patients remain constant, others increase in disease severity over time and then plateau, while some patients oscillate in severity. It is because of these differences in progression that both slope *m* and logistic growth *k* were used to characterize disease progression. To develop the best model, four scores were used in forward stepwise regression. Adjusted R^2, Fowlkes-Mallows Cp, SBC and AIC scores were used for both *m* and *k* obtained from the BUE and VLE scales to find the optimal number of parameters.

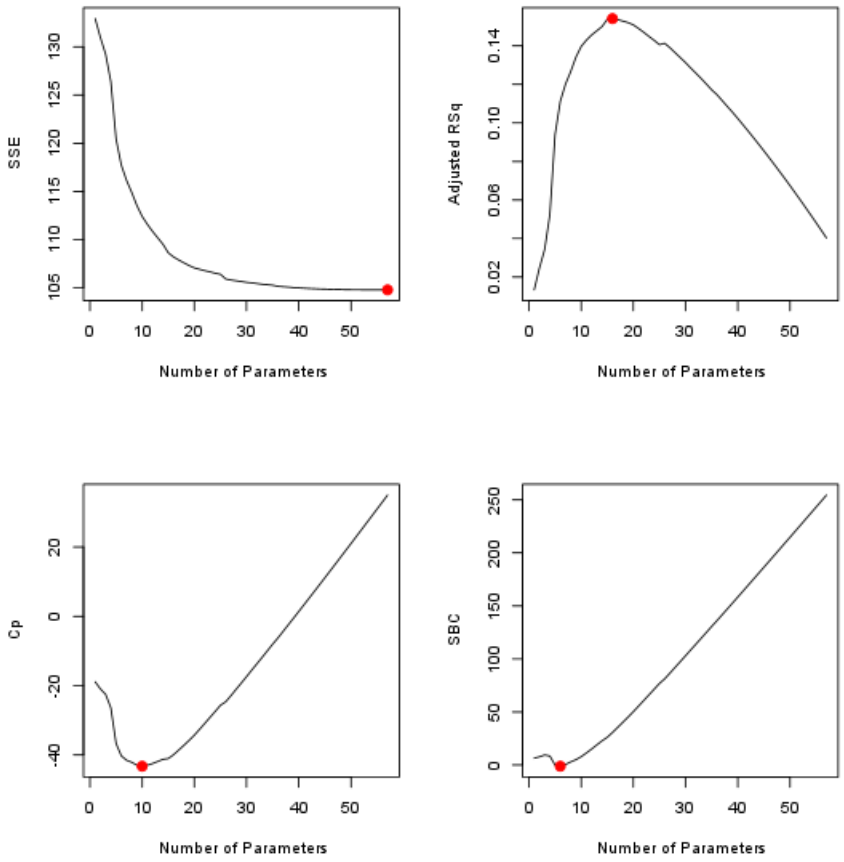

Figure 15: Best number of parameters by evaluation metric. *The figure shows the SSE, Adjusted R^2, Cp, and SBC score on the y-axis for predicting the BUE score derived slope. The red dot highlights the number of parameters needed for the best score.*

For the Brooke Upper Extremity slope, Adjusted R2 had 16 predictive exons, Cp had 10 predictive exons, SBC had 6 predictive exons, and AIC had 10 predictive exons. For the logistic growth rate Adjusted R2 had 5 predictive exons, Cp had 4 predictive exons, SBC had 1 predictive exons, and AIC had 4 predictive exons.

	Slope (*m*)	Logistic Growth Rate (*k*)
Adjusted R^2	17, 22, 23, 24, 25, 26, 28, 30, 32, 33, 34, 42, 58, 60, 63, 79	55, 58, 60, 61, 79
Cp	17, 22, 23, 25, 26, 28, 30, 34, 63, 79	55, 58, 60, 61
SBC	17, 23, 25, 26, 30, 34	58
AIC	13, 18, 19, 20, 22, 23, 27, 51, 63, 64	45, 47, 48, 49

Table 4: Brooke Upper Extremity best model exons by predictor and evaluation metric.

For the Vignos Lower Extremity slope Adjusted R2 had 13 predictive exons, Cp had 10 predictive exons, SBC had 1 predictive exons, and AIC had 7 predictive exons. For the logistic growth rate Adjusted R2 had 5 predictive exons, Cp had 4 predictive exons, SBC had 1 predictive exons, and AIC had 0 predictive exons. In the case of AIC, having zero predictive exons means that a horizontal line is better at predicting the logistic growth rate than any line.

	Slope (*m*)	Logistic Growth Rate (*k*)
Adjusted R^2	4, 11, 15, 32, 33, 34, 35, 55, 62, 63, 64, 65, 68	55, 57, 58, 60, 79
Cp	4, 11, 15, 32, 35, 55, 62, 64, 65, 68	55, 57, 58, 60
SBC	55	57
AIC	45, 51, 53, 55, 60, 61, 62	

Table 5: Vignos Lower Extremity best model exons by predictor and evaluation metric.

Because these best models are derived from different evaluation metrics, they cannot be directly compared to each other. For instance, a model that minimizes Cp, may not minimize SBC. Additionally, each exon chosen by the models has a significance value; however, in minimizing each metric, insignificant terms are included. Because there is much ambiguity in the best model, rather than choosing one model, showing which exons the models chose as most predictive highlights two clusters of mutations as shown in Figure 16: one from exon 17-35 and

another from exon 50-65. Through this visualization, the best exon can be chosen as the most frequent exons found in the models. When compared between upper and lower extremities the exons predictive of the upper extremities are clustered around exon 25, and exons predictive of the lower extremities are clustered around exon 60.

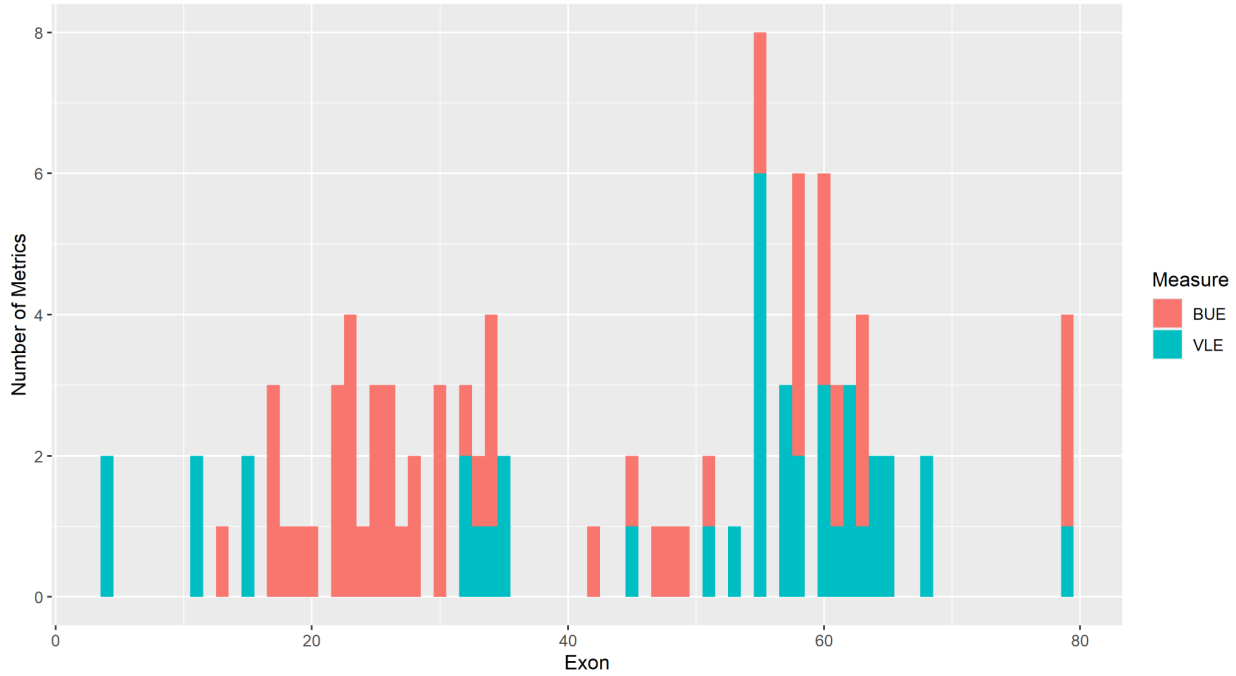

Figure 16: Distribution of best model exons. *There are two clusters of best model exons: Brooke Upper Extremity predictors centered around exon 25 and Vignos Lower Extremity predictors centered around exon 60.*

Discussion and Conclusions

This MQP sought to utilize the novel neuroMuscular ObserVational Research datahub to develop disease classification models for Becker and Duchenne Muscular Dystrophy, as well as predict disease progression over time. Through the preprocessing of the data it was found that the data aligns with the reading frame rule previously found in the literature. Duchenne muscular dystrophy results from out of frame mutations while Becker muscular dystrophy results from in frame mutations. Despite the wealth of data present, it was unfortunate that for many patients, a lot of the specific mutation data was missing including the location and frame type. This cut the usable data in half as a large part of the project revolved around using the location of mutations. Despite this missing data, the GTEx transcript expression data and isoforms was used to attempt to create better classifiers; however, both classification and using those datasets as predictors failed to perform well and provide meaningful results. In the future, a better method may be employed to incorporate the transcript expression data with the patient's genetic data.

In contrast to the isoform and tissue data, using exons with machine learning algorithms showed consistency in classifying disease type. In addition, using exons as a predictor for disease progression identified two clusters of exons that were selected by multiple metrics in stepwise regression analysis. The cluster around exon 25 aligned with an actin binding site and was primarily important in predicting the Brooke Upper Extremity scale. On the other hand, the second cluster around exon 60 aligned with the dystroglycan binding site and corresponds to Vignos Lower Extremity predictors. Both of these locations are in known mutation hotspots on the rod domain. It is suspected that these hotspots are caused by the extremely repetitive spectrin repeats causing template slippage (Ankala et al, 2012); however no study has shown a correlation specifically between these regions and the upper and lower extremities. Further studies could biologically validate these correlations; however more robust data collection would need to be applied. Additionally having more computational power would allow for a more exhaustive regression to be carried out rather than only using a forward algorithm.

Code and Data Availability

All code used in this project is available at https://github.com/crelicthecleric/MD-ML-MQP. For any inquiries please contact Dr. Elizabeth Ryder at ryder@wpi.edu

This study was conducted using data from the Muscular Dystrophy Association's neuroMuscular ObserVational Research (MOVR) DataHub®. MOVR is operated through participating MDA Care Centers with the support of participants, site PIs, coordinators and staff. MOVR data is available through the Muscular Dystrophy Association. Email mdamovr@mdausa.org for details.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 02/01/22. For more information go to https://gtexportal.org/home/.

Bibliography

Aartsma-Rus, A., Van Deutekom, J. C., Fokkema, I. F., Van Ommen, G. B., & Den Dunnen, J. T. (2006). Entries in the Leiden duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle & Nerve*, *34*(2), 135-144. https://doi.org/10.1002/mus.20586

Ankala, A., Kohn, J. N., Hegde, A., Meka, A., Ephrem, C. L., Askree, S. H., Bhide, S., & Hegde, M. R. (2012). Aberrant firing of replication origins potentially explains intragenic nonrecurrent rearrangements within genes, including the human *DMD* gene. *Genome Research*, *22*(1), 25-34. https://doi.org/10.1101/gr.123463.111

Annuar, A. A., Wong, K. T., Ching, A. S., Thong, M. K., Wong, S. W., Alsiddiq, F., Ong, L. C., & Goh, K. J. (2010). Exercise induced cramps and myoglobinuria in dystrophinopathy – a report of three Malaysian patients. *Neurology Asia*, *15*(2), 125-131. https://www.neurology-asia.org/index.php

Benarroch, L., Bonne, G., Rivier, F., & Hamroun, D. (2019). The 2020 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscular Disorders*, *29*(12), 980-1018. https://doi.org/10.1016/j.nmd.2019.10.010

Bladen, C. L., Salgado, D., Monges, S., Foncuberta, M. E., Kekou, K., Kosma, K., Dawkins, H., Lamont, L., Roy, A. J., Chamova, T., Guergueltcheva, V., Chan, S., Korngut, L., Campbell, C., Dai, Y., Wang, J., Barišić, N., Brabec, P., Lahdetie, J., … Lochmüller, H. (2015). The TREAT-NMD DMD Global Database: Analysis of more than 7,000 Duchenne Muscular Dystrophy mutations. *Human Mutation*, *36*, 395-402. https://doi.org/10.1002/humu.22758

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. https://doi.org/10.1145/130385.130401

Bougé, A., Murauer, E., Beyne, E., Miro, J., Varilh, J., Taulan, M., Koenig, M., Claustres, M., & Tuffery-Giraud, S. (2017). Targeted RNA-SEQ profiling of splicing pattern in the DMD gene: Exons are mostly constitutively spliced in human skeletal muscle. Scientific Reports, 7(1). https://doi.org/10.1038/srep39094

Brooke, M. H., Griggs, R. C., Mendell, J. R., Fenichel, G. M., Shumate, J. B., & Pellegrino, R. J. (1981). Clinical trial in duchenne dystrophy. I. The design of the protocol. *Muscle & Nerve*, *4*(3), 186-197. https://doi.org/10.1002/mus.880040304

Broomfield, J., Hill, M., Guglieri, M., Crowther, M., & Abrams, K. (2021). Life expectancy in duchenne muscular dystrophy. Neurology, 97(23), e2304-e2314. https://doi.org/10.1212/wnl.0000000000012910

Capitanio, D., Moriggi, M., Torretta, E., Barbacini, P., De Palma, S., Viganò, A., Lochmüller, H., Muntoni, F., Ferlini, A., Mora, M., & Gelfi, C. (2020). Comparative proteomic analyses of duchenne muscular dystrophy and Becker muscular dystrophy muscles: Changes contributing to preserve muscle function in Becker muscular dystrophy patients. Journal of Cachexia, Sarcopenia and Muscle, 11(2), 547-563. https://doi.org/10.1002/jcsm.12527

Clerk, A., Morris, G. E., Dubowitz, V., Davies, K. E., & Sewry, C. A. (1993). Dystrophin-related protein, utrophin, in normal and dystrophic human fetal skeletal muscle. The Histochemical Journal, 25(8), 554-561. https://doi.org/10.1007/bf02388063

Cornell, B. (2016). *Gene identification*. BioNinja. https://ib.bioninja.com.au/options/untitled/b2-biotechnology-in-agricul/gene-identificatio n.html

Dominguez, R., & Holmes, K. C. (2011). Actin Structure and Function. *Annual Review of Biophysics*, *40*, 169-186. https://doi.org/10.1146/annurev-biophys-042910-155359

Doorenweerd, N., Mahfouz, A., Van Putten, M., Kaliyaperumal, R., 't Hoen, P., Hendriksen, J., Aartsma-Rus, A., Verschuuren, J., Niks, E., Reinders, M., Kan, H., & Lelieveldt, B. (2018). Timing and localization of human dystrophin isoform expression provide insights into the cognitive phenotype of duchenne muscular dystrophy. Neuromuscular Disorders, 28, S10. https://doi.org/10.1016/s0960-8966(18)30318-3

Duan, D., Goemans, N., Takeda, S., Mercuri, E., & Aartsma-Rus, A. (2021). Duchenne muscular dystrophy. *Nature Reviews Disease Primers*, *7*(13). https://doi.org/10.1038/s41572-021-00255-4

Emery, A. E. (1991). Population frequencies of inherited neuromuscular diseases—A world survey. *Neuromuscular Disorders*, *1*(1), 19-29. https://doi.org/10.1016/0960-8966(91)90039-u

Emery, A. E. (2002). The muscular dystrophies. *The Lancet*, *359*(9307), 687-695. https://doi.org/10.1016/s0140-6736(02)07815-7

Ervasti, J. M. (2007). Dystrophin, its interactions with other proteins, and implications for muscular dystrophy. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1772*(2), 108-117. https://doi.org/10.1016/j.bbadis.2006.05.010

Feher, J. J. (2017). Quantitative human physiology: An introduction (2nd ed.). Academic Press.

Fletcher, S., Adams, A. M., Johnsen, R. D., Greer, K., Moulton, H. M., & Wilton, S. D. (2010). Dystrophin Isoform induction in vivo by antisense-mediated alternative splicing. *Molecular Therapy*, *18*(6), 1218-1223. https://doi.org/10.1038/mt.2010.45

Fokkema, I. F., Taschner, P. E., Schaafsma, G. C., Celli, J., Laros, J. F., & Den Dunnen, J. T. (2011). LOVD v.2.0: The next generation in gene variant databases. Human Mutation, 32(5), 557-563. https://doi.org/10.1002/humu.21438

Gawor, M., & Prószyński, T. J. (2017). The molecular cross talk of the dystrophin-glycoprotein complex. *Annals of the New York Academy of Sciences*, *1412*(1), 62-72. https://doi.org/10.1111/nyas.13500

Han, R., Kanagawa, M., Yoshida-Moriguchi, T., Rader, E. P., Ng, R. A., Michele, D. E., Muirhead, D. E., Kunz, S., Moore, S. A., Iannaccone, S. T., Miyake, K., McNeil, P. L., Mayer, U., Oldstone, M. B., Faulkner, J. A., & Campbell, K. P. (2009). Basal lamina strengthens cell membrane integrity via the laminin G domain-binding motif of a-dystroglycan. *Proceedings of the National Academy of Sciences*, *106*(31), 12573-12579. https://doi.org/10.1073/pnas.0906545106

Helbling-Leclerc, A., Zhang, X., Topaloglu, H., Cruaud, C., Tesson, F., Weissenbach, J., Tomé, F., Schwartz, K., Fardeau, M., Tryggvason, K., & Guicheney, P. (1995). Mutations in the laminin α2–chain gene (LAMA2) cause merosin–deficient congenital muscular dystrophy. *Nature Genetics*, *11*(2), 216-218. https://doi.org/10.1038/ng1095-216

Hnia, K., Zouiten, D., Cantel, S., Chazalette, D., Hugon, G., Fehrentz, J., Masmoudi, A., Diment, A., Bramham, J., Mornet, D., & Winder, S. (2007). ZZ domain of dystrophin and utrophin: Topology and mapping of a β-dystroglycan interaction site. *Biochemical Journal*, *401*(3), 667-677. https://doi.org/10.1042/bj20061051

Holt, K. H., & Campbell, K. P. (1998). Assembly of the sarcoglycan complex: Insights for muscular dystrophy. *Journal of Biological chemistry*, *273*(52), 34667-34670. https://doi.org/10.1074/jbc.273.52.34667

Holt, K. H., Crosbie, R. H., Venzke, D. P., & Campbell, K. P. (2000). Biosynthesis of dystroglycan: Processing of a precursor propeptide. *FEBS Letters*, *468*(1), 79-83. https://doi.org/10.1016/s0014-5793(00)01195-9

Iwata, Y., Sampaolesi, M., Shigekawa, M., & Wakabayashi, S. (2004). Syntrophin is an actin-binding protein the cellular localization of which is regulated through cytoskeletal reorganization in skeletal muscle cells. *European Journal of Cell Biology*, *83*(10), 555-565. https://doi.org/10.1078/0171-9335-00415

Koenig, M., & Kunkel, L. M. (1990). Detailed analysis of the repeat domain of dystrophin reveals four potential hinge segments that may confer flexibility. *Journal of Biological Chemistry*, *265*(8), 4560-4566. https://doi.org/10.1016/s0021-9258(19)39599-7

Koenig, M., Monaco, A., & Kunkel, L. (1988). The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. *Cell*, *53*(2), 219-228. https://doi.org/10.1016/0092-8674(88)90383-2

Lai, Y., Zhao, J., Yue, Y., & Duan, D. (2012). a2 and a3 helices of dystrophin R16 and R17 frame a microdomain in the 1 helix of dystrophin R17 for neuronal NOS binding. *Proceedings of the National Academy of Sciences*, *110*(2), 525-530. https://doi.org/10.1073/pnas.1211431109

Langenbach, K., & Rando, T. (2002). Inhibition of dystroglycan binding to laminin disrupts the PI3K/AKT pathway and survival signaling in muscle cells. *Muscle & Nerve*, *26*(5), 644-653. https://doi.org/10.1002/mus.10258

Larkindale, J., Yang, W., Hogan, P. F., Simon, C. J., Zhang, Y., Jain, A., Habeeb-Louks, E. M., Kennedy, A., & Cwik, V. A. (2013). Cost of illness for neuromuscular diseases in the United States. *Muscle & Nerve*, *49*(3), 431-438. https://doi.org/10.1002/mus.23942

Legardinier, S., Raguénès-Nicol, C., Tascon, C., Rocher, C., Hardy, S., Hubert, J., & Le Rumeur, E. (2009). Mapping of the lipid-binding and stability properties of the central rod domain of human dystrophin. *Journal of Molecular Biology*, *389*(3), 546-558. https://doi.org/10.1016/j.jmb.2009.04.025

Luxner, L. (2019, February 14). *Defying the odds of living with duchenne, decade after decade*. Muscular Dystrophy News. https://musculardystrophynews.com/2019/02/08/defying-the-odds-of-living-with-duchenne-decade-after-decade/

Magri, F., Govoni, A., D'Angelo, M. G., Del Bo, R., Ghezzi, S., Sandra, G., Turconi, A. C., Sciacco, M., Ciscato, P., Bordoni, A., Tedeschi, S., Fortunato, F., Lucchini, V., Bonato, S., Lamperti, C., Coviello, D., Torrente, Y., Corti, S., Moggio, M., … Comi, G. P. (2011). Genotype and phenotype characterization in a large dystrophinopathic cohort with extended follow-up. *Journal of Neurology*, *258*(9), 1610-1623. https://doi.org/10.1007/s00415-011-5979-z

Marshall, J. L., Holmberg, J., Chou, E., Ocampo, A. C., Oh, J., Lee, J., Peter, A. K., Martin, P. T., & Crosbie-Watson, R. H. (2012). Sarcospan-dependent Akt activation is required for utrophin expression and muscle regeneration. *Journal of Cell Biology*, *197*(7), 1009-1027. https://doi.org/10.1083/jcb.201110032

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Mokhtarian, A., Lefaucheur, J. P., Even, P. C., & Sebille, A. (1999). Hindlimb immobilization applied to 21-day-old mdx mice prevents the occurrence of muscle degeneration. Journal of Applied Physiology, 86(3), 924-931. https://doi.org/10.1152/jappl.1999.86.3.924

Monaco, A. P., Bertelson, C. J., Liechti-Gallati, S., Moser, H., & Kunkel, L. M. (1988). An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics*, *2*(1), 90-95. https://doi.org/10.1016/0888-7543(88)90113-9

Morris, G. E., Sedgwick, S. G., Ellis, J. M., Pereboev, A., Chamberlain, J. S., & Nguyen thi Man. (1998). An Epitope structure for the C-terminal domain of dystrophin and Utrophin. *Biochemistry*, *37*(31), 11117-11127. https://doi.org/10.1021/bi9805137

Mukund, K., & Subramanian, S. (2019). Skeletal muscle: A review of molecular structure and function, in health and disease. *Mechanisms of Disease*, *12*(1), e1462. https://doi.org/10.1002/wsbm.1462

Muntoni, F., Torelli, S., & Ferlini, A. (2003). Dystrophin and mutations: One gene, several proteins, multiple phenotypes. *The Lancet Neurology*, *2*(12), 731-740. https://doi.org/10.1016/s1474-4422(03)00585-4

National Center for Biotechnology Information. (2021, November 14). *DMD dystrophin [Homo sapiens (human)] - Gene - NCBI*. Retrieved November 15, 2021, from https://www.ncbi.nlm.nih.gov/gene/1756?_ga=2.241646470.505676344.1636740782-1986090702.1633564186

National Institute of Health. (n.d.). *List of FDA orphan drugs*. Genetic and Rare Diseases Information Center (GARD). https://rarediseases.info.nih.gov/diseases/fda-orphan-drugs/D

Norwood, F. L., Sutherland-Smith, A. J., Keep, N. H., & Kendrick-Jones, J. (2000). The structure of the N-terminal actin-binding domain of human dystrophin and how mutations in this domain may cause duchenne or Becker muscular dystrophy. *Structure*, *8*(5), 481-491. https://doi.org/10.1016/s0969-2126(00)00132-5

Petrof, B. J., Schrager, J. B., Stedman, H. H., & Sweeney, H. L. (1993). Dystrophin protects the sarcolemma from stresses developed during muscle contraction. *Proceedings of the National Academy of Sciences*, *90*(3), 3710-3714. https://doi.org/10.1073/pnas.90.8.3710

Rentschler, S., Linn, H., Deininger, K., Bedford, M., Espanel, X., & Sudol, M. (1999). The WW domain of dystrophin requires EF-hands region to interact with β-dystroglycan. *Biological Chemistry*, *380*(4). https://doi.org/10.1515/bc.1999.057

Rybakova, I. N., Humston, J. L., Sonnemann, K. J., & Ervasti, J. M. (2006). Dystrophin and utrophin bind actin through distinct modes of contact. *Journal of Biological Chemistry*, *281*(15), 9996-10001. https://doi.org/10.1074/jbc.M513121200

Rybakova, I. N., Patel, J. R., & Ervasti, J. M. (2000). The dystrophin complex forms a mechanically strong link between the sarcolemma and Costameric actin. *Journal of Cell Biology*, *150*(5), 1209-1214. https://doi.org/10.1083/jcb.150.5.1209

Salari, N., Fatahi, B., Valipour, E., Kazeminia, M., Fatahian, R., Kiaei, A., Shohaimi, S., & Mohammadi, M. (2022). Global prevalence of duchenne and Becker muscular

dystrophy: A systematic review and meta-analysis. *Journal of Orthopaedic Surgery and Research*, *17*(1). https://doi.org/10.1186/s13018-022-02996-8

Sarepta. (2018). *Access the mechanism of action with exon skipping | Exondys 51 (eteplirsen) injection*. EXONDYS 51 (eteplirsen) injection for Healthcare Professionals. https://www.exondys51hcp.com/about-exondys-51/mechanism-of-action

Singh, S., & Mallela, K. (2012). The N-terminal actin-binding tandem calponin-homology (CH) domain of dystrophin is in a closed conformation in solution and when bound to F-actin. *Biophysical Journal*, *103*(9), 1970-1978. https://doi.org/10.1016/j.bpj.2012.08.066

Suzuki, A., Yoshida, M., & Ozawa, E. (1995). Mammalian a1-syntrophin and b1-syntrophin bind to the alternative splice-prone region of the dystrophin COOH terminus. *Journal of Cell Biology*, *128*(3), 373-381. https://doi.org/10.1083/jcb.128.3.373

Taylor, P. J., Betts, G. A., Maroulis, S., Gilissen, C., Pedersen, R. L., Mowat, D. R., Johnston, H. M., & Buckley, M. F. (2010). Dystrophin gene mutation location and the risk of cognitive impairment in duchenne muscular dystrophy. *PLoS ONE*, *5*(1), e8803. https://doi.org/10.1371/journal.pone.0008803

Tennyson, C., Dally, G. Y., Ray, P. N., & Worton, R. G. (1996). Expression of the dystrophin isoform Dp71 in differentiating human fetal myogenic cultures. Human Molecular Genetics, 5(10), 1559-1566. https://doi.org/10.1093/hmg/5.10.1559

Thangarajh, M., Hendriksen, J., McDermott, M., Martens, W., Hart, K., & Griggs, R. (2019). Relationships between DMD mutations and neurodevelopment in dystrophinopathy. *Neurology*, *93*(17), 1597-1604. https://doi.org/10.1212/WNL.0000000000008363

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*. https://doi.org/10.1109/icdar.1995.598994

Tuffery-Giraud, S., Miro, J., Koenig, M., & Claustres, M. (2017). Normal and altered pre-mrna processing in the DMD gene. *Human Genetics*, *136*(9), 1155-1172. https://doi.org/10.1007/s00439-017-1820-9

Vengalil, S., Preethish-Kumar, V., Polavarapu, K., Mahadevappa, M., Sekar, D., Purushottam, M., Thomas, P. T., Nashi, S., & Nalini, A. (2017). Duchenne muscular dystrophy and Becker muscular dystrophy confirmed by multiplex ligation-dependent probe amplification: Genotype-phenotype correlation in a large cohort. *Journal of Clinical Neurology*, *13*(1), 91. https://doi.org/10.3988/jcn.2017.13.1.91

Vignos, P. J., Spencer, G. E., & Archibald, K. C. (1963). Management of progressive muscular dystrophy of childhood. *JAMA: The Journal of the American Medical Association*, *184*(2), 89. https://doi.org/10.1001/jama.1963.03700150043007

Yang, M., Zheng, Y., Xie, Z., Wang, Z., Xiao, J., Zhang, J., & Yuan, Y. (2021). A deep learning model for diagnosing dystrophinopathies on thigh muscle MRI images. *BMC Neurology*, *21*(1). https://doi.org/10.1186/s12883-020-02036-0

Yamaguchi, M., & Suzuki, M. (2015). Becoming a back-up carer: Parenting sons with duchenne muscular dystrophy transitioning into adulthood. Neuromuscular Disorders, 25(1), 85-93. https://doi.org/10.1016/j.nmd.2014.09.001

Yin, L., Schnoor, M., & Jun, C. (2020). Structural characteristics, binding partners and related diseases of the Calponin homology (CH) domain. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.00342

Yiu, E. M., & Kornberg, A. J. (2015). Duchenne muscular dystrophy. *Journal of Paediatrics and Child Health*, *51*(8), 759-764. https://doi.org/10.1111/jpc.12868

Zhu, Y., Deng, H., Chen, X., Li, H., Yang, C., Li, S., Pan, X., Tian, S., Feng, S., Tan, X., Matsuo, M., & Zhang, Z. (2019). Skipping of an exon with a nonsense mutation in the DMD gene is induced by the conversion of a splicing enhancer to a splicing silencer. *Human Genetics*, *138*(7), 771-785. https://doi.org/10.1007/s00439-019-02036-2