



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학석사 학위논문

**Integrated Analysis of Genome and  
Transcriptome to Enhance  
Understanding of Rare  
Neuromuscular Disorders**

희귀 신경근 질환의 유전체, 전사체  
통합 분석 연구

2019년 8월

서울대학교 대학원  
의과학과 의과학전공  
Jana Kneissl

# Integrated Analysis of Genome and Transcriptome to Enhance Understanding of Rare Neuromuscular Disorders

지도 교수 최무림

이논문은 의학석사 학위논문으로 제출함

2019 년 4 월

서울대학교 대학원  
의과학과 의과학전공

Jana Kneissl

Jana Kneissl 의 의학석사 학위논문을 인준함

2019 년 06 월

이승재 \_\_\_\_\_ (인)

최무림 \_\_\_\_\_ (인)

채종희 \_\_\_\_\_ (인)

# Integrated Analysis of Genome and Transcriptome to Enhance Understanding of Rare Neuromuscular Disorders

Thesis supervisor: Murim Choi (PhD)  
Thesis submission: April, 2019

Seoul National University  
College of Medicine  
Department of Biomedical Sciences  
Functional Genomics  
Jana Kneissl

Approval by thesis committee  
June, 2019

Seung-Jae Lee (PhD) \_\_\_\_\_

Murim Choi (PhD) \_\_\_\_\_

Jong Hee Chae (MD, PhD) \_\_\_\_\_

Abstract

# Integrated Analysis of Genome and Transcriptome to Enhance Understanding of Rare Neuromuscular Disorders

Jana Kneissl

Department of Biomedical Sciences

College of Medicine

Seoul National University

**Introduction.** Whole exome sequencing has become a robust and standard tool for rare diseases diagnosis thanks to advantages in cost and data handling. However, whole exome sequencing-based diagnosis rates typically do not exceed 50%, which can be attributed to the difficulty of interpreting variants of uncertain significance, as well as to the disregard of non-coding variants, including variants in intronic and regulatory regions in the genome. Therefore, I explored the utility of transcriptome sequencing as a compensatory approach in rare neuromuscular disorders diagnosis.

**Methods.** Whole exome sequencing of 94 patients with undiagnosed neuromuscular disorders was collected from Seoul National University Children's Hospital and analyzed for variants in known neuromuscular disease genes. Additional transcriptome sequencing was performed for 63 of the whole exome sequenced patients and for ten patients without genome data. Transcriptome data were utilized for cryptic damaging variants, differentially expression, aberrant splicing and allele specific

expression analysis. Furthermore, non-negative matrix factorization was applied to identify expression-based clustering and cluster-specific gene ontology was derived.

**Results.** Whole exome sequencing analysis identified candidate variants in 49% of patients, with 83% of them located within known disease genes. Structural variants with questionable pathogenicity were discovered in twelve cases. RNA-Sequencing based variant calling lead to further discovery of heterozygous candidate variants in nine samples, five of which did not undergo whole exome sequencing. Allele specific expression identified two likely candidate genes and differential gene expression analysis lead to the prioritization of sets of genes in an additional four samples. Lastly, aberrant splicing of *DMD*, *TTN* and *MICU1* was detected in each of four samples. Non-negative matrix factorization-based clustering resulted in the identification of six clusters with distinct gene expression profiles.

**Discussion.** Firstly, I aimed to evaluate whether transcriptome sequencing can provide additional evidence for the interpretation of whole exome sequencing variants. Overall, transcriptome sequencing was able to detect abnormalities associated with the previously identified mutation in less than 30% of positive whole exome sequencing cases. For samples without whole exome sequencing result, I successfully used transcriptome sequencing to identify potential pathogenic causes in 18 cases. In conclusion, transcriptome sequencing proved to be a useful tool for the diagnosis of whole exome sequencing negative samples, but did not prove to have great utility for the interpretation of pathogenic whole exome sequencing variants.

---

**Keywords:** whole exome sequencing, transcriptome sequencing, neuromuscular disorders, multiomics, diagnosis, Mendelian disorder, variant discovery, expression-based clustering

**Student Number:** 2017-21922

# Table of Contents

1. INTRODUCTION.....	1
1.1. Advancement through next generation sequencing.....	1
1.2. Genetics of neuromuscular disorders (NMD).....	3
1.3. Transcriptome sequencing-based NMD diagnosis.....	8
2. METHODS.....	12
2.1. Data collection.....	12
2.2. Whole exome sequencing data analysis.....	13
2.3. Transcriptome sequencing analysis.....	15
2.4. Non-negative matrix factorization based clustering.....	19
3. RESULTS.....	22
3.1. Data collection.....	22
3.2. Phenotype information.....	23
3.3. Whole exome sequencing results.....	25
3.4. Transcriptome sequencing quality control.....	28
3.5. Transcriptome-based clustering.....	31
3.6. Exome variants in transcriptome sequencing.....	35
3.7. Transcriptome-sequencing based diagnosis.....	39
4. DISCUSSION.....	48
5. REFERENCES.....	57
6. APPENDIX.....	63
6.1. Supplementary Figures.....	63
6.2. Supplementary Tables.....	67
7. 국문초록.....	71

# Table of Figures

FIGURE 1.....	4
FIGURE 2.....	6
FIGURE 3.....	11
FIGURE 4.....	22
FIGURE 5.....	24
FIGURE 6.....	25
FIGURE 7.....	26
FIGURE 8.....	29
FIGURE 9.....	30
FIGURE 10.....	31
FIGURE 11.....	32
FIGURE 12.....	33
FIGURE 13.....	34
FIGURE 14.....	34
FIGURE 15.....	35
FIGURE 16.....	37
FIGURE 17.....	38
FIGURE 18.....	39
FIGURE 19.....	41
FIGURE 20.....	44
FIGURE 21.....	44
FIGURE 22.....	47
FIGURE 23.....	63
FIGURE 24.....	64
FIGURE 25.....	64
FIGURE 26.....	65
FIGURE 27.....	66

## Index of Tables

TABLE 1.....	9
TABLE 2.....	28
TABLE 3.....	40
TABLE 4.....	40
TABLE 5.....	42
TABLE 6.....	42
TABLE 7.....	45
TABLE 8.....	70

# 1. Introduction

## 1.1. Advancement through next generation sequencing

### 1.1.1. Identification of genotype-phenotype associations

The technological innovation of next generation sequencing technologies and the associated drop of sequencing cost have revolutionized the study of disease pathobiology within the last decade. Especially, whole exome sequencing has been integrated as a routine diagnostic tool in clinical practice. Whole exome sequencing captures the estimated 2-3% of the human genome encoding proteins, making it a cost-effective alternative to whole genome sequencing. Moreover, compared to targeted gene sequencing or gene-panel sequencing, whole exome sequencing captures most genes at once, allowing for a later reanalysis with easy incorporation of database updates or novel published disease-phenotype information [1]. This has resulted in an identification of candidate genes for approximately half of all hitherto known rare diseases and has led to the understanding of the underlying genetics in many cancers [2].

However, for an overwhelming number of disorders a genetic cause has not yet been identified. Moreover, even for diseases with established genetic cause, not all cases show mutations within the known genes. In fact, it has been established, that mutations in different genes can cause clinically indistinguishable phenotypes (locus heterogeneity), and mutations within the same gene can cause a variety of phenotypes (phenotypic heterogeneity). Additionally, even carriers of a known disease allele do not necessarily display the associated phenotype (incomplete penetrance), further complicating sequencing based diagnosis [2].

### 1.1.2. Integration of genome and transcriptome

Even though genome sequencing has been successfully employed to gain insight in the genetic background of somatic and germline disorders, it generally requires functional follow-up studies to assign disease-risk or pathogenicity to a genomic loci with certainty. A possible way to bridge the gap between genomic data and clinical data is the application of transcriptome sequencing. Transcriptome sequencing directly provides information on how a certain genotype affects expression levels in the healthy and diseased and can therefore be used as a means to evaluate the functional impact of the genotype.

For many complex disorders, large-scale genome-wide association studies (GWAS) have allowed the identification of disease susceptibility loci. These are single nucleotide variants significantly more frequently encountered in affected individuals than the healthy population. However, the interpretation of these loci has remained challenging, mainly owing to the limited knowledge on the functional impact of such single nucleotide polymorphisms. An example of how transcriptome data can aid in the assessment of such variants was given by Ferreira *et al.*, who utilized RNA-sequencing data to uncover genes in the vicinity of previously identified GWAS risk loci, which showed abnormal expression levels in breast cancer [3]. Similarly, Lamontagne *et al.* integrated genomic data collected for GWAS studies and transcriptome sequencing from a large lung expression-quantitative trait loci (eQTL) study to prioritize the risk loci for chronic obstructive pulmonary disorder [4].

Furthermore, transcriptome sequencing has been successfully implemented for the discovery and interpretation of non-coding causative variants in Mendelian disorders. For instance, it has been

shown that individuals affected by X-linked dystonia Parkinsonism share a haplotype of seven single nucleotide variants and a retrotransposon located within a 449kb region on chromosome X. Integration of this genomic information with RNA-sequencing of cultured neurons revealed that indeed the retrotransposon not only affects the expression of *TAF1*, a gene located within the 449kb region, but also leads to intron retention and aberrant splicing of the gene. Thus, the retrotransposon presents a likely genetic cause for X-linked dystonia Parkinsonism [5].

In conclusion, previous studies have successfully demonstrated that the combination of genomic and transcriptomic information greatly aids in the interpretation of genomic data and can therefore lead to the prioritization and discovery of novel disease genes.

## **1.2. Genetics of neuromuscular disorders (NMD)**

“Neuromuscular disorders” is a collective term including a variety of diseases which affect the normal functioning of the muscle either by directly impairing muscle structure and metabolism or by indirectly affecting the signal transfer from neurons to muscles. Previous efforts have led to the identification of over five-hundred genes linked to over nine-hundred distinct NMDs [6]. Here, I provide a short overview over some of the most important insights into the genetics of neuromuscular disorders.

### **1.2.1. Congenital myopathies: from pathology to genetics**

Congenital myopathies are neuromuscular disorders that usually present with muscle weakness and hypotonia at the time of birth and have an estimated prevalence of 1:20,000 [6]. Historically, they have been classified by subtype-specific muscle biopsy changes into four

subcategories: nemaline myopathy, core myopathy, centronuclear myopathy and congenital fiber-type disproportion (Figure 1).

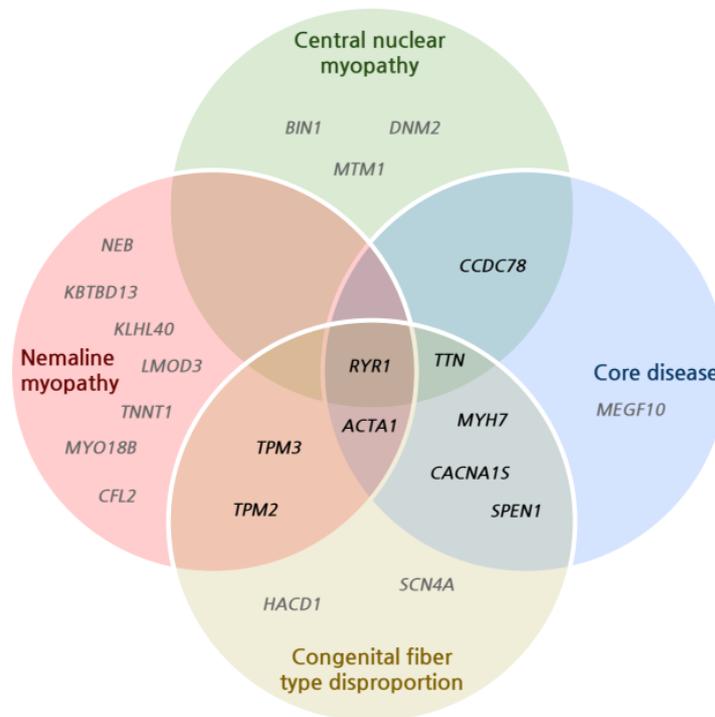


Figure 1: Venn diagram of known genes causing different subtypes of congenital myopathies (adapted from Gonorazky, et al., 2018)

The histopathological feature that distinguishes nemaline myopathies is the presence of nemaline bodies, also called rods, in the muscle biopsy. Rod body myopathies are caused by mutations in genes encoding components of the actin thin filament in the muscle (*ACTA1*, *NEB*, *TPM2*, *TPM3*, *TNNT1*, *LMOD3*), which lead to either a faulty formation of the actin filament or disturb the interaction between the thin filament and the adjacent thick myosin filament [7]. Other genes associated with nemaline myopathies include *KLHL40*, *KLHL41* and *KBTBD13*, genes whose protein products have been shown to regulate thick filament breakdown and protein turnover [9].

Centronuclear myopathies are defined as myopathies for which

central nuclei can be observed in over 25% of myofibers. A severe subtype called myotubular myopathy is caused by X-linked recessive mutations in *MTM1* and frequently leads to death in early childhood. *MTM1* plays a role in the regulation of the neuromuscular junction structure and function and hence, affects the formation and maintenance of the excitation-contraction coupling apparatus. Mutations in other genes, which also play a role in the excitation-contraction coupling such as *DNM2*, *MTMR14*, *SPEG* and *BIN* have also been linked to centronuclear myopathies [7].

Mutations in *RYR1* can likewise induce a centronuclear myopathy phenotype, however, they more frequently cause central core myopathies. Central cores are muscle fibers lacking reactivity to oxidative stains, most commonly due to a lack of mitochondria. It has been proposed that the disease phenotype is caused by a reduced release of calcium ions, either caused by mutations directly in *RYR1*, or by mutations in *SEPN1*, which modulates the calcium ion re-uptake by *RYR1* [9].

Commonly, congenital fiber type disproportions are described as a fourth subtype, but it has also been proposed that they represent a preliminary stage before the development of any of the above mentioned subtype-specific pathology features. The latter hypothesis is supported by the fact that mutations in the associated genes (*SEPN1*, *RYR1*, *TPM3*) have been described to frequently lead to other subtypes [7](Figure 1).

Even though the four described categories constitute the majority of congenital myopathies, a variety of other genes have been described to likewise cause a congenital myopathy phenotype. For instance, mutations in *SCN4A*, *STAC3* and *SPTBN4* have been shown to

result in myosin-storage myopathies presenting with a highly variable clinical presentation [8].

### 1.2.2. Muscular dystrophies and the dystrophin complex

Muscular dystrophies present with progressive muscle weakness of skeletal muscles, but differ in the distribution of affected body regions and involvement of other organ systems. Common phenotypes shared among patients include muscle atrophy, joint contractures, hypertrophy and myotonia.

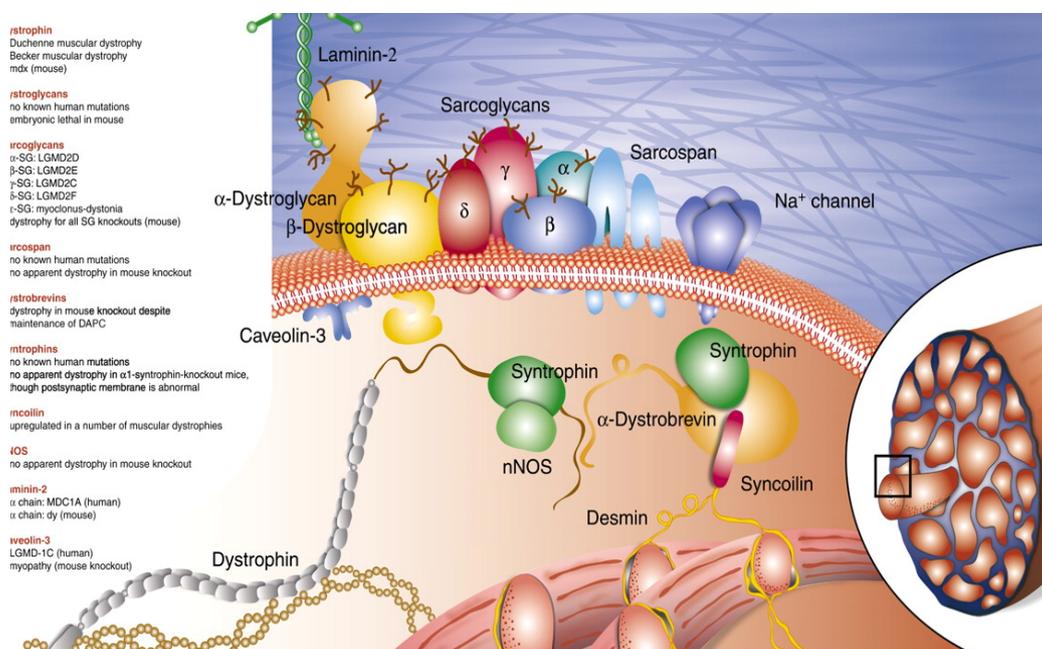


Figure 2: Muscular disorders caused by mutations in genes encoding the dystrophin associated protein complex (DAPC) (Ehmsen, Poon and Davies, 2002)

The most prevalent muscular dystrophy subtypes are Duchenne muscular dystrophy (prevalence 8.3 in 100,000 boys) and Becker's muscular dystrophy (prevalence 7.3 in 100,000 boys) caused by mutations in *DMD*, the largest gene in the human genome [9]. Over 90% of the phenotypic difference between the two disorders can be explained by whether *DMD*'s open-reading frame is disrupted by the mutation. In the less severe Becker's muscular dystrophy, a shortened

but partly functional protein is expressed, while in the severe Duchenne muscular dystrophy the reading frame is disrupted and the protein is completely non-functional [10].

Dystrophin, the protein encoded by *DMD*, is part of a complex responsible for the connection between the extracellular matrix and the intracellular cytoskeleton. Abnormalities in a multitude of other genes encoding proteins related to the dystrophin complex have similarly been described to cause muscular dystrophies (Figure 2). For instance, mutations in genes encoding sarcoglycans (*CAPN3*, *CAV3*, *DYSF*) lead to the development of limb-girdle muscular dystrophies and abnormal glycosylation of  $\alpha$ -Dystroglycan (*DAG1*) induces Fukuyama congenital muscular dystrophy, muscle-eye-brain disease and Walker-Warburg syndrome [12]. Similarly, mutations in Laminin-2 (*LMNA2*), a protein interacting with  $\alpha$ -Dystroglycan, can result in the development of Emery-Dreifuss muscular dystrophy, another subtype of congenital muscular dystrophies [13].

### **1.2.3. Other neuromuscular disorders**

Congenital myopathies and congenital muscular dystrophies show strong phenotypic and genetic overlap. Commonly, they present with disease onset in early childhood or even at birth. In contrary, *GNE*-related myopathy, caused by mutations in N-acetylglucosamine epimerase, and myofibrillar myopathies present with onset later in life. N-acetylglucosamine epimerase malfunctioning in *GNE*-myopathies results in a reduced addition of sialic acid to glycoproteins and glycolipids, impairing their function and hence affecting a variety of cellular processes [14]. Myofibrillar myopathies are characterized by protein aggregates and disorganized myofibers. Genes encoding proteins that have been shown to abnormally aggregate include *DES*,

*MYOT, CRYAB, LBD3, ZASP, FLNC* and *BAG3*. All these genes are associated with structure and function of Z-discs or chaperone-associated autophagy [15].

Another important subgroup of neuromuscular disorders are metabolic disorders, which can be caused by a dysregulation of the cellular energy household. The different cellular energy production pathways define the metabolic disorder subtypes. Glycolytic enzyme defects lead to an impaired use of glycogen as energy source and result in weakness, cramps and myonecrosis upon exertion. A similar phenotype can be caused by mutations in genes coding for proteins important for the lysosomal storage of glycogen. On the other hand, mutations of genes involved in lipid metabolism impair the use of lipids for energy [16]. Moreover, mitochondrial disorders present an essential subtype of metabolic disorders.

In conclusion, neuromuscular disorders make up an extremely diverse disease group with strong symptomatic overlap, making pure phenotype-based diagnosis difficult even for the most experienced clinicians. Whole exome sequencing has already identified many causative genes and has greatly contributed to the understanding of the underlying pathomechanisms. However, with more than half of all samples not resulting in a definite diagnosis by exome sequencing alternative methods have to be explored.

### **1.3. Transcriptome sequencing-based NMD diagnosis**

Recent approaches have aimed to evaluate transcriptome sequencing as a tool in the diagnosis of neuromuscular and metabolic disorders. Table 1 shows a summary of the three major studies published to this day.

	Cummings et al, 2017	Gonorazky et al, 2019	Kremer et al, 2017
<b>Samples</b>			
# of samples total	63	65	105
RNA-source	muscle biopsy	muscle biopsy	fibroblasts
<b>RNA results</b>			
# diagnosed by alternative splicing	17	8	5
# diagnosed by allele specific expression	0	3	5
# diagnosed by variant calling	0	10	0
# diagnosed by gene expression	0	10	3
<b>Diagnosis rate</b>			
total	35%	36%	10%

Table 1: Overview over previous studies using RNA-sequencing based diagnosis for neuromuscular and metabolic disorders

In their pioneering study of 63 patients affected from undiagnosed rare muscle disorders, Cummings *et al.* were able to achieve an overall RNA-sequencing based diagnosis rate of 35%. All their findings were based on the detection of aberrant splicing. To do so they used an in-house script comparing the 63 patients to 184 control samples derived from the Genotype-Tissue Expression project (GTEx) [17]. Their method led to an identification of a median of five abnormal splicing events in 190 neuromuscular disease-associated genes. The detection of abnormal splicing events allowed them to assign pathogenicity to a missense variant in *TTN*, synonymous variants in *RYR1* and *POMGNT1* and to a highly recurrent de novo intronic mutation in *COL6A1*. Even though they performed outlier gene expression and allele specific expression analysis, they did not yield any significant results and the analyses merely played a supportive role [18].

In a follow-up study Gonorazky *et al.* studied a cohort of 65 samples without whole exome sequencing results. Even though their final diagnosis rate of 36% was similar to the previous report, only 26.7% of the diagnosed samples showed aberrant splicing. Additionally, they were able to identify candidate variants by RNA-

sequencing based variant calling in ten cases and prioritized further thirteen variants via outlier gene expression and allele specific expression analysis. Similar to Cummings, *et al.*, they utilized control muscle samples derived from the GTEx project during their analyses [7].

A slightly different approach was taken by Kremer *et al.*, when they utilized RNA-sequencing for the diagnosis of 105 mitochondriopathy patients. Following three analysis approaches, they identified gene expression outliers, monoallelic rare variants and aberrant splicing events in all samples. In comparison to the other published studies, they did not use any external controls, but instead compared each sample to the remaining samples of the cohort. For forty of the included samples previous whole exome sequencing diagnosis was available. However, only for 14 (35%) diagnosed samples, they were able to discover any kind of evidence in RNA-sequencing. Notably, they did not find any abnormalities in samples with heterozygous missense variants. Purely RNA-sequencing based diagnosis was achieved in 10% of cases with their most remarkable finding being a recurrent intronic mutation in *TIMMDC1* found in three unrelated families included in this study [19].

These three previously published studies explored a variety of different analyses and have successfully shown, how transcriptome sequencing can be applied to diagnose patients affected by Mendelian disorders.

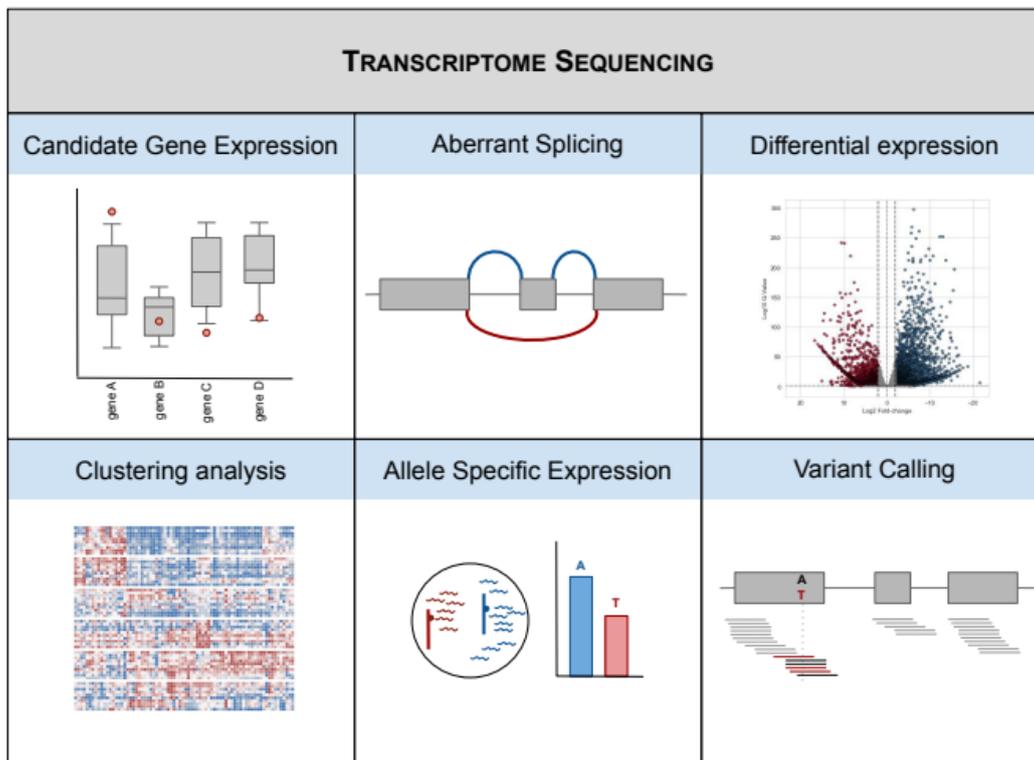


Figure 3: Analysis performed using RNA-sequencing data that can aid in the diagnosis of whole exome sequencing negative samples and evaluate whole exome sequencing samples with variants of uncertain significance

Here, I integrated genome and transcriptome data generated from patients with various types of neuromuscular disorders. Whereas previous studies have focused on alternative splicing analysis and sample-specific changes, I aimed to construct a comprehensive transcriptomic profile driven by the heterogeneity of the cohort. Therefore, I not only performed a variety of transcriptomic analysis that aimed to diagnose each sample individually (Figure 3), but also applied a clustering-algorithm to identify hidden cohort subgroups with distinct transcriptome profiles. Thus, I provide a genetic overview of the NMD patients by genome and transcriptome and example cases that were resolved by various analytic approaches.

## **2. Methods**

### **2.1. Data collection**

The study was approved by Seoul National University's Institutional review board (IRB#1707-126-872, IRB#1101-110-353).

#### **2.1.1. Whole exome sequencing**

Patient blood samples were obtained by clinicians and DNA extraction was performed by a hospital-associated laboratory. Extracted DNA was then sent to Theragen sequencing, where Illumina paired end sequencing with Agilent SureSelect Human All Exome v5 exome capture kit was performed. Raw data was downloaded from the company server in form of FASTQ files. All patients or their guardians gave informed consent.

#### **2.1.2. Transcriptome sequencing**

Muscle biopsy samples were obtained for patient and controls and biopsies were sent to Theragen sequencing. The company performed RNA extraction and provided a quality report including the RNA integrity number, volume and concentration for each sample. After excluding low quality samples, Theragen performed library preparation using TruSeq Stranded total RNA kit and Illumina paired end sequencing. FASTQ files were provided for download upon completion of the sequencing process from the company server.

Out of 718 available GTEx skeletal muscle RNA-Sequencing samples, 136 samples were selected, downloaded and converted to FASTQ files. The sample selection excluded samples affected by diseases which could influence muscle gene expression levels, such as “amyotrophic lateral sclerosis” or “unexplained weakness”. The data access was approved and samples were obtained from the database of

genotypes and phenotypes (dbGAP) [20].

## **2.2. Whole exome sequencing data analysis**

### **2.2.1. Whole exome sequencing alignment**

FASTQ files were aligned to reference genome Hg19 (UCSC, 2009) using Bwa Mem (v.0.7.16a) and mate coordinates were added to the BAM files using Samtools Fixmate (v.1.4) [21, 22]. After sorting the BAM file by coordinates, duplicate reads were marked by Picard MarkDuplicates [23]. The Picard software package (v.2.9.4) was further utilized to add read-groups, sample-IDs and sequencing platform information to the BAM files. Insertions and deletions were realigned using the Genome Analysis Toolkit (GATK, v.3.8.0) software RealignerTargetCreator and IndelRealigner to assure high-quality indel identification [24]. Indel realignment locally realigns indels to minimize the number of mismatching bases. In order to correct against systematic bias introduced into quality scores assigned by the sequencing machines, the quality scores were re-calibrated using the Genome Analysis Toolkit's BaseRecalibrator package. Finally, the BAM file quality was assessed using Samtools Flagstat and Qualimap Bamqc (v.2.2.1) [25]. The quality reports for all samples were combined using MultiQC (v.1.7) [26].

### **2.2.2. Whole exome sequencing candidate variant discovery**

Variants were called from the resulting BAM file using Samtools Pileup [22] and normalized with Bcftools (v.1.3) [27] to ensure correct annotation by SnpEff (v.4.3i) [28] in the following step. After SnpEff annotated the variant calling file with predicted variant effect and predicted variant impact, the related software SnpSift [29] was used to add public database information. Each VCF was annotated with the

following databases: dbSNP [30], 1000Genomes [31], ExAC [32], UK10K [33], KOVA [34], Deciphering developmental disorders project [35] and an in-house database.

Final VCF files were converted into a table format and filters were applied for variant prioritization. First, I excluded all variants with a genotype quality (GQ) value under eighty to minimize the amount of false-positive variants. Due to difficulty in interpretation of variant impact, I also excluded any variant located in an intronic region and variants with predicted low impact. Moreover, single nucleotide variants and small indels with a recorded allele frequency  $> 0.01$  in any of the above mentioned public databases were considered likely benign. Additionally, I filtered out all variants that were found within more than 10% of the cohort.

For trio samples, I explored three possible inheritance hypotheses: compound heterozygous variants, homozygous variants and de novo variants. If not indicated differently in the clinical information, I assumed that parents were unaffected and therefore filtered out any homozygous variants inherited from the parents. Where clinical information indicated a family history, I adjusted the filtering hypotheses to match the observed inheritance pattern. For singleton samples, I investigated homozygous variants, potentially compound heterozygous (two variants within the same gene) and heterozygous variants located in known disorder genes recorded in the Online Mendelian Inheritance in Man (OMIM) database [36]. Candidate variants were presented to clinicians to discuss the accordance between sample phenotype and previously described phenotypes associated with each candidate gene. In collaboration with the respective clinicians and expert geneticists, I reached a final result for each sample.

### **2.2.3. Whole exome sequencing structural variant detection**

To assess, whether a sample carried copy number variations, I extracted the average coverage for each target region provided in the Agilent BED file. After normalizing the coverage information using the average total coverage depth for the respective sample, I calculated the per-base coverage for each target region. Subsequently, I compared the normalized average-per-base-coverage for each sample to a control. For trio samples both parents were used as a control, while for singleton samples an average of all trio parent samples served as a control. The ratio of the per-base average coverage information between sample and control was calculated and plotted for each chromosome. Regions showing abnormal coverage ratios, were investigated closer and compared to structural variants described in gnomAD-SV [37].

Loss of heterozygosity was assessed using an in-house script. The pileup output file produced during the variant calling process was used to calculate the minor allele frequency at each covered genomic position [22]. The minor allele frequency was then plotted against the genomic position for each chromosome, samples and regions without detected minor allele frequency around 0.5 were selected. For each detected region, I searched for overlapping potentially damaging homozygous variants in the available whole exome sequencing data.

## **2.3. Transcriptome sequencing analysis**

### **2.3.1. Variant calling from RNA-Sequencing**

Variant calling from RNA-Sequencing was performed following a slightly modified version of GATK's best practices published online [38]. Using STAR aligner (v.2.5.4b), raw RNA sequencing FASTQ files were aligned to the Hg19 reference genome (UCSC, 2009) [39]. To increase the

accuracy of splice junction alignment, I opted to use the STAR two-pass mode available as a command line option. Sample, library and sequencing platform information was added to the sorted BAM files and duplicate reads were marked using Picard tools (v.2.9.0) [23]. To reduce false positive variant calls in splice regions, GATK's software (v.3.8.0) SplitNCigarReads was used to extract exon segments and hard clip intron regions. Moreover, the tool simultaneously reassigned mapping qualities to each read to convert the quality score assigned by STAR to the quality scores used by downstream tools. Finally, I applied GATK's base-quality score recalibration software to correct for systematic errors made in the sequencing process.

Variants for each sample were called using GATK's Haplotype caller in GVCF mode and subsequently, all sample GVCF files were combined into a cohort-GVCF file. GATK's GenotypeGVCF software was then called to perform joint genotyping on all samples [38]. Following the recommendations from the best practices, the variants from the resulting VCF file were filtered if the assigned phred-scale confidence score fell below 20. Moreover clusters of at least three variants within a 35bp window were flagged. Before performing annotation with SnpEff (v.4.3r) [28], the VCF file was normalized with Bcftools (v.1.3) [27]. Finally, SnpSift [29] was utilized to add information from 1000Genomes [31], gnomAD [40], ExAC [32], KOVA [34], UK10K [33], Deciphering Developmental Disorders [35], Clinvar [41] and an in-house database. The VCF file was then split and converted into sample-specific table files using Scikit-allel (v.1.2.0) [42].

Similarly to the process for variant prioritization from whole exome sequencing, I excluded variants with GQ value below 80 and flagged variants with a cohort allele count over ten percent of the

cohort size. Moreover, I prioritized any variant detected in known muscle or disease genes and variants with low public allele frequency.

### **2.3.2. Alternative Splicing analysis**

For discovery of aberrant splicing, I used the Leafcutter software (v.0.2.8) and its associated visualization tool LeafViz [43]. First, I realigned all FASTQ files to the Hg19 reference assembly using the intronMotif option available for STAR aligner (v.2.5.4b). Subsequently, BAM files were converted to junction files for all samples and controls. The junction files were then combined into cluster files, which carry per cluster read counts for each individual sample. The splicing analysis was performed separately for each sample, using all other samples and the four available control samples as an object of comparison. To allow for a comparison with less than two samples per group, I set the available flags for minimum amount of samples per group and minimum number of samples per intron to one.

For each sample, I then extracted all clusters with an adjusted p-value below 0.05 and annotated the genes with information from the OMIM database [36]. For samples with less than twenty significant splicing events, I evaluated each splicing event separately for potential pathogenicity. First, I checked whether the reported phenotype in any of the abnormally spliced OMIM genes matched the samples phenotype. For genes without a known disease-association, I checked for any publications mentioning a connection between the gene and muscle development and maintenance. In case the analysis returned more than twenty significant splicing events recorded for one sample, I used Toppgene [44] to check whether the genes are related with respect to gene ontology, pathway, genomic location or transcription factors. Where trio whole exome sequencing data was available, I further

searched for any de novo intronic variants within the alternatively spliced gene.

### **2.3.3. Gene counts and differential gene expression**

BAM files aligned for alternative splicing analysis were used as an input for Subread Featurecount package (v.1.6.1) [45] and raw count files were retrieved. Count quality control was performed using the Qualimap Countqc software (v.2.2.1) [25] and by plotting and investigating the logarithmic count distribution. In-between sample correlation was used to check for outlier samples. Due to poor quality, sample CDC\_NM16.1 was excluded from any further RNA-Sequencing analysis.

Raw counts were combined into one matrix and used as an input file for the DESeq2 package (v.1.22.1)[46] in R. Similarly to the above described alternative splicing analysis, I compared each sample to all other samples and controls using the sequencing batch as a potential confounding variable. For each sample differentially expressed genes with adjusted p-value below 0.05 were extracted and annotated with information from the OMIM database. Furthermore, for samples with less than twenty differentially expressed genes, I investigated each gene separately, searching for a connection between gene and muscle disorder in OMIM and Pubmed. Alternatively, if more than twenty differentially expressed genes were detected, I performed gene ontology analysis using Toppgene [44] focusing on gene ontology and pathway.

Additionally, I performed gene expression outlier analysis. Assuming a normal distribution of the expression levels across samples for each gene, I calculated the probability of each minimum and maximum occurring [47]. Subsequently, I applied Bonferroni correction for multiple testing using the Statsmodel module for Python [48].

#### **2.3.4. Allele Specific Expression**

In order to perform allele specific expression analysis, I adapted the python script described in “Tools and best practices for data processing in allelic expression analysis” by Castel *et al.* [49].

First, all heterozygous variants with a coverage of over twenty and a minor allele frequency between 0.3 and 0.7 were extracted from the normalized unfiltered whole exome sequencing VCF file produced by Samtools. The VCF file consisting of the selected heterozygous variants was annotated with SnpEff and public databases using SnpSift. VCFanno (v.0.2.9) [50] was used to add the CADD score [51] for each variant to the VCF file. The positions of these variants were subsequently converted into a BED file. Next, I used Samtools mpileup on the final BAM files produced by the RNA-Sequencing variant calling pipeline with the BED file as a regions input. Finally, the information from the exome VCF file and the information from the RNA-sequencing mpileup file were combined into a final output file carrying the expression information for all heterozygous exome variants covered in RNA-sequencing.

#### **2.4. Non-negative matrix factorization based clustering**

This analysis was based on the clustering method described in “Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer” by Robertson *et al.* [52].

As an input for non-negative matrix factorization, I used the FPKM matrix calculated based on the raw count outputs obtained by running Featurecounts in the stranded mode. FPKMs were objected to upper quartile normalization per sample and were median centered per gene. Furthermore, the normalized FPKM values were logarithmically

transformed resulting in the matrix  $R$ .

In order to increase speed and efficiency of this analysis, I selected a subset of the top 25% most variable genes as measured by their standard deviation across all samples, and calculated the sample-to-sample Spearman correlation coefficient distance matrix ( $R_{dis}$ ).

Aiming to identify the optimal amount of clusters, I used the resulting distance matrix ( $R_{dis}$ ) as an input for standard hierarchical clustering linkage analysis using SciPy [47] for a random subset of 80% of samples selected for each iteration. For each potential number of cluster  $K_{2-21}$ , I performed linkage analysis  $K*500$  times and retrieved the association between sample and cluster for each iteration. Next, I calculated, how often each two samples were assigned to the same cluster in  $K*500$  iterations for each potential  $K$  ( $M_k$ ). The matrices ( $M_{k1-21}$ ) were summed up and normalized by the total number of iterations leading to the matrix  $M_{norm}$ .

Subsequently, I iteratively performed non-negative matrix factorization on  $M_{norm}$  with potential  $K_{2-21}$  as input and observed the behavior of the cophenetic correlation coefficient across different  $K$  values and noted the optimum  $K_{opt}$  as the last  $K$  value before the cophenetic correlation coefficient dropped (measure of cluster stability). This analysis was repeated twenty times and the most frequently resulting  $K_{opt}$  was chosen as the optimal cluster number  $K^*=6$  [53].

To assign each sample to a cluster, I again performed non-negative matrix factorization on  $M_{norm}$  with a fixed number of clusters  $K^*=6$  as input and retrieved the factorized matrix  $H$ . This matrix contained the association values between each sample and each cluster. I considered each sample to be part of the cluster to which it had the highest association score. Subsequently, I performed non-negative

matrix factorization using the logarithmic count matrix  $R$ , the number of clusters  $K^*=6$  and the normalized matrix  $H$  as input.

Lastly, in order to determine representative genes for each cluster, I calculated the mean expression difference for all genes between samples within a cluster and outside a cluster and selected the top 1% of genes that showed the greatest difference as cluster-specific genes.

## 3. Results

### 3.1. Data collection

#### 3.1.1. Genetic and Transcriptome data collection for SNUH - samples

The Seoul National University Hospital (SNUH) - neuromuscular disorder patient cohort consisted of 117 samples with diverse disease phenotypes. The cohort was selected from a larger SNUH neuromuscular cohort (> 1,000) after filtering for diagnostic single gene or targeted panel sequencing. During this study, whole exome sequencing data was collected for 96 cases, the majority of which (68 samples, 70.83%) were sequenced as singletons. For a subset of twenty-eight patients (29.17%), supplementary parent whole exome sequencing was collected.

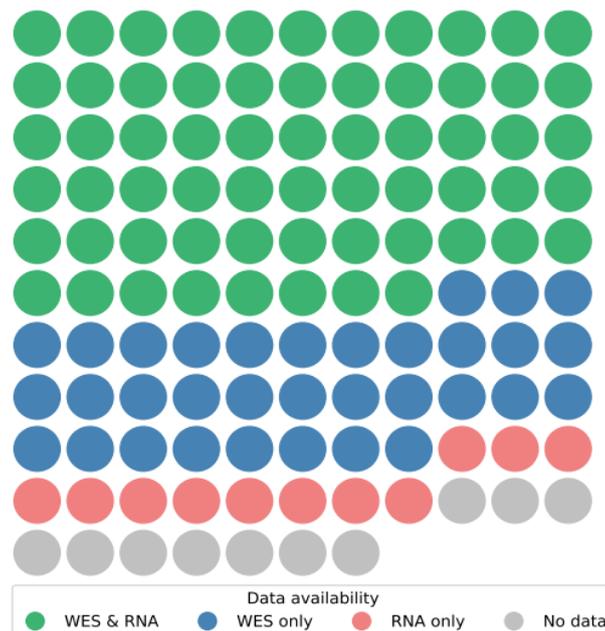


Figure 4: Data availability for each sample included in the SNUH-cohort. Each circle depicts one sample with the color representing the type of the collected sequencing data

Transcriptome sequencing was obtained for 63 of the above described whole exome sequencing samples, and for additional eleven samples without genetic data, leading to a final number of 74 transcriptome-sequenced samples. Note, that the study cohort included ten samples for which no sequencing data had been available until the point of writing (Figure 4).

### **3.1.2. Identification and Correction of SNUH Sample Mix-Ups**

In order to prevent sample mix-ups from falsifying the results, I used the NGSCheckmate software package [54] to retrieve correlation coefficients between raw FASTQ files from whole exome and whole transcriptome sequencing. Due to this analysis two samples were excluded from further analysis: whole exome sequencing of CDC\_NM43.1, which matched whole exome and transcriptome sequencing of CDC\_NM44.1 and transcriptome sequencing of CDC\_NM97.1, which matched whole exome and transcriptome sequencing of CDC\_NM31.1.

### **3.1.3. Transcriptome control data collection from healthy muscle**

Initially, transcriptome sequencing data of 136 healthy skeletal muscle samples was obtained from the Genotype-Tissue Expression atlas (GTEx) [17]. However, due to heavy differences in the expression profiles between SNUH cohort data and GTEx control data, I did not use these samples for downstream analysis (see 3.4.2).

As a compensation, my collaborators provided transcriptome sequencing for four healthy muscle biopsies retrieved during plastic surgery procedures.

## **3.2. Phenotype information**

Clinical information was provided by the responsible clinicians in form of

a spreadsheet containing basic information such as age and gender, suspected diagnosis, presented symptoms, as well as results for an assortment of clinical tests. The obtained data, assigned male gender to 79 and female gender to 36 of the samples included in the SNUH-cohort. Out of the 117 samples, 21 samples (17.9%) had a positive family history, with at least one other relative presenting with a

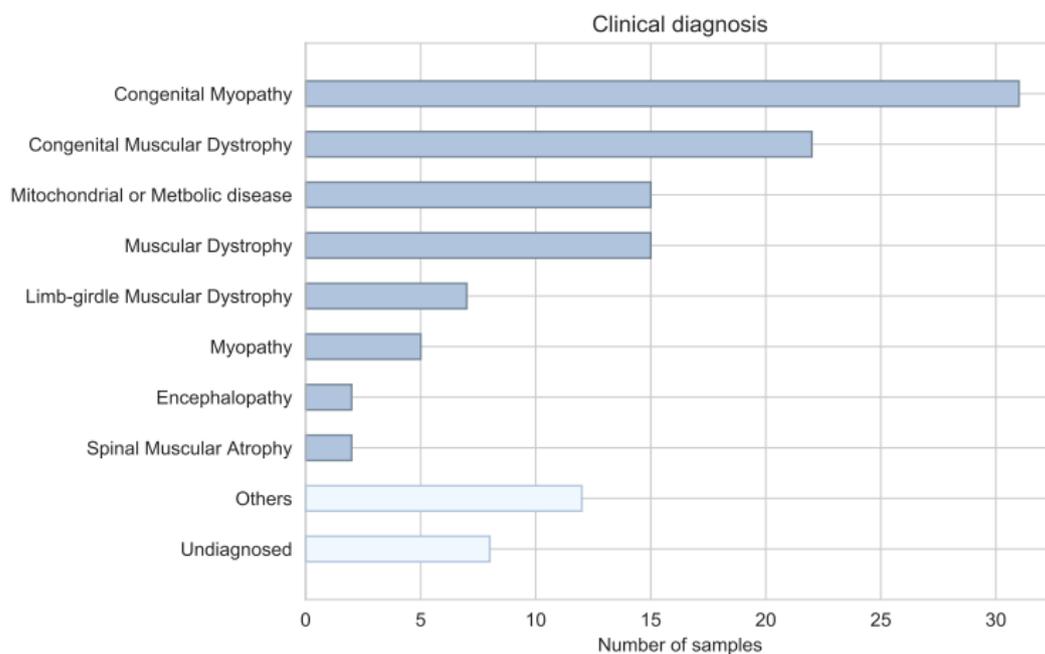


Figure 5: Number of samples per different clinical diagnoses. “Others” include Charcot-Marie-Tooth disease, myotonia, motor neuron disease and congenital myasthenic syndrome.

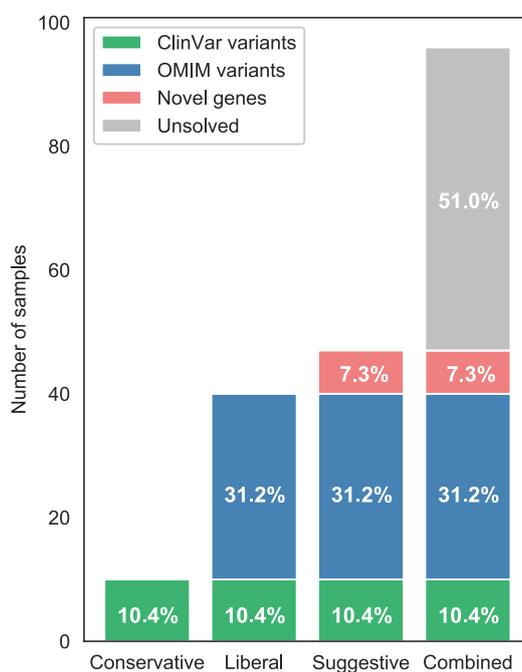
comparable disease phenotype. Affected family members included siblings, parents, grandparents and distant relatives.

The clinical diagnoses included congenital myopathies, congenital muscular dystrophies, as well as metabolic and mitochondrial disorders and are summarized in Figure 5. Common muscle symptoms shared among several samples included muscle weakness (53), muscle atrophy (8), motor developmental delay (18), hypotonia (24), and muscle hypertrophy (15). Additionally, some samples displayed heart (6) or respiratory involvement (10). Neurological symptoms and

developmental delay manifested in ten and eight samples, respectively. For 44 samples genetic tests had been performed previously, including multiplex ligation-dependent probe amplification for *SMN* and *DMD*, gene panel sequencing, target gene sequencing, but had not yielded convincing results.

### 3.3. Whole exome sequencing results

#### 3.3.1. Detection of pathogenic SNVs and small indels



*Figure 6: Results from WES divided into three categories: conservative (known pathogenic variants), liberal (+ known disease genes), suggestive (+ rare variants in novel genes)*

Whole exome sequencing small variant discovery was performed in 94 samples and lead to the identification of a candidate gene in 47 (49.0%) of them. The full list of candidate genes and variants can be seen in supplementary Table 8. In total, 10.4% of samples carried variants described in the ClinVar [41] database as “Pathogenic” or “Likely

pathogenic”, while in an additional 31.2% of samples, variants located in known muscle disease genes were discovered, leading to a liberal diagnosis rate of 41.6%. For three samples, a heterozygous variant in a known gene associated with a clinically comparable disease was detected, but a second variant to fulfill the previously published recessive inheritance remained undiscovered (CDC\_NM1.1: *TACO1*, CDC\_NM9.1: *NEB*, CDC\_NM60.1: *MYO9A*). Furthermore, I suggest variants in four genes not previously associated with neuromuscular disorders as potential causative in four samples (*SYTL2*, *RRMB2*, *TRAPPC1*, *ARRDC4*). Including these variants a final suggestive diagnosis rate of 49.0% (Figure 6) was achieved.



Figure 7: Composition of variant types prioritized in WES. Homozygous (HOM), heterozygous (HET) and compound heterozygous (CMPHET) variants were investigated.

The composition of the candidate variants is depicted in Figure 7. In summary, I identified compound heterozygous variants in 25 samples,

homozygous variants in seven and heterozygous variants in 15 cases. All detected homozygous variants were missense variants, while for heterozygous variants, loss of function (frameshift, stopgain) variants were discovered in four cases. Likewise the majority of detected compound heterozygous variants consisted of two missense variants, with all but one sample carrying a missense variant in at least one allele. Two compound heterozygous loss of function variants were only detected in one sample (CDC\_NM87.1, *TTM*) with a clinically consistent presentation of titinopathy.

Sanger sequencing confirmation and segregation was successfully performed for the prioritized candidate variants in fourteen samples.

### **3.3.2. Structural variant discovery**

Using coverage information from whole exome sequencing data, I searched for sample-specific copy number variations and regions showing loss of heterozygosity. This analysis resulted in the discovery of a region of loss of heterozygosity on chromosome 22 in sample CDC\_NM60.1 which included a homozygous variant in *APOL4*. However, due to the lack of evidence for an association of *APOL4* with the neuromuscular disorders, the region was not further investigated.

Copy number deletions were detected in three and copy number duplications in nine samples, resulting in a copy number discovery rate of 12.77% (Table 2). In order to exclude non-pathogenic copy number variations, I searched the gnomAD-SV database for any structural variants overlapping the regions highlighted in the patients. In ten out of the twelve cases, at least one overlapping structural variant was reported in gnomAD-SV. Moreover, in three of those cases (CDC\_NM25.1, CDC\_NM52.1, CDC\_NM70.1) the gnomAD variation

was of the same type and completely contained the copy number variations detected during this analysis.

<b>Sample</b>	<b>Estimated location of copy number variation</b>
CDC_NM10.1	Duplication chr11:134,100,000-134,300,000
CDC_NM16.1	Deletion chr5:40,800,000-41,100,000
CDC_NM25.1	Deletion chr8:39,250,00-39,450,000
CDC_NM27.1	Duplication chr6:7,100,000-7,500,000
CDC_NM29.1	Duplication chr3:170,500,000-171,000,000
CDC_NM52.1	Duplication chr3:75,800,000-75,900,000
CDC_NM53.1	Duplication chr2: 236,500,000-237,500,000
CDC_NM66.1	Duplication chr2:89,400,000-90,500,000
CDC_NM70.1	Duplication chr14:71,400,000-71,600,00
CDC_NM78.1	Deletion chr1:146,000,000-148,000,000
CDC_NM80.1	Duplication chr9:135,900,000-136,000,000
CDC_NM90.1	Duplication chr16:56,800,000-56,900,000

*Table 2: List of copy number variation per sample with the approximate region*

To allow for further assessment of clinical relevance, all genes overlapping the copy number regions were extracted and researched for previous muscle disease associations using OMIM and Pubmed. However, no gene with previous neuromuscular disease association was contained in the discovered copy number variations.

### **3.4. Transcriptome sequencing quality control**

#### **3.4.1. Muscle biopsy RNA quality assessment**

Using the GTEx-project as a pointer, I originally considered using a RNA-integrity number (RIN)-value cutoff of six for allowing samples to proceed from RNA isolation to whole transcriptome sequencing. However, with only 20% of the SNUH samples in the first sequencing batch passing this requirement, I decided to manually evaluate the RNA integrity profile for each sample. The validity of this approach was further supported by the observation, that no correlation between RIN-value and the number of high-quality aligned reads (Figure 8) could be observed.

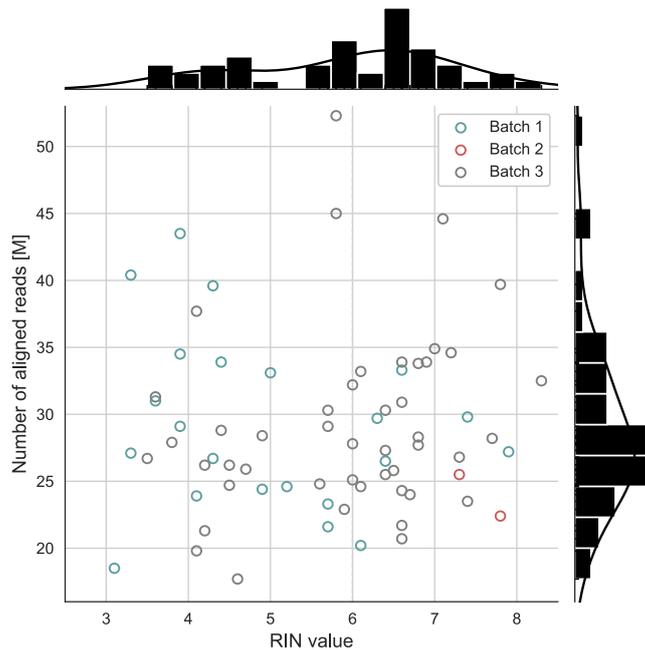


Figure 8: Scatterplot showing RIN value and number of aligned reads in million colored by sequencing batch.

### 3.4.2. Integrating GTEx control samples with SNUH transcriptome data

Following previous related studies [18,50], I intended to utilize skeletal muscle RNA-sequencing data provided by the GTEx-project as control for transcriptome analysis. To assure sufficient similarity for such usage, I compared count density profiles between GTEx and SNUH samples and performed t-distributed stochastic neighbor embedding (tSNE) based clustering.

Comparison of the count densities of GTEx realigned samples, original GTEx counts (publicly available on their website) and SNUH counts including all detected genes showed a distinct distribution for all four sample types with SNUH cases and GTEx realigned muscle showing the greatest resemblance (Figure 9, top). In comparison, the density distributions based on protein-coding genes showed a clear convergence of the four types into two groups: GTEx samples (realigned

and original) and SNUH samples (cases and controls) (Figure 9, middle). This grouping, although reduced in severity, was also reproduced when basing the analysis on muscle disorder related genes only (Figure 9, bottom).

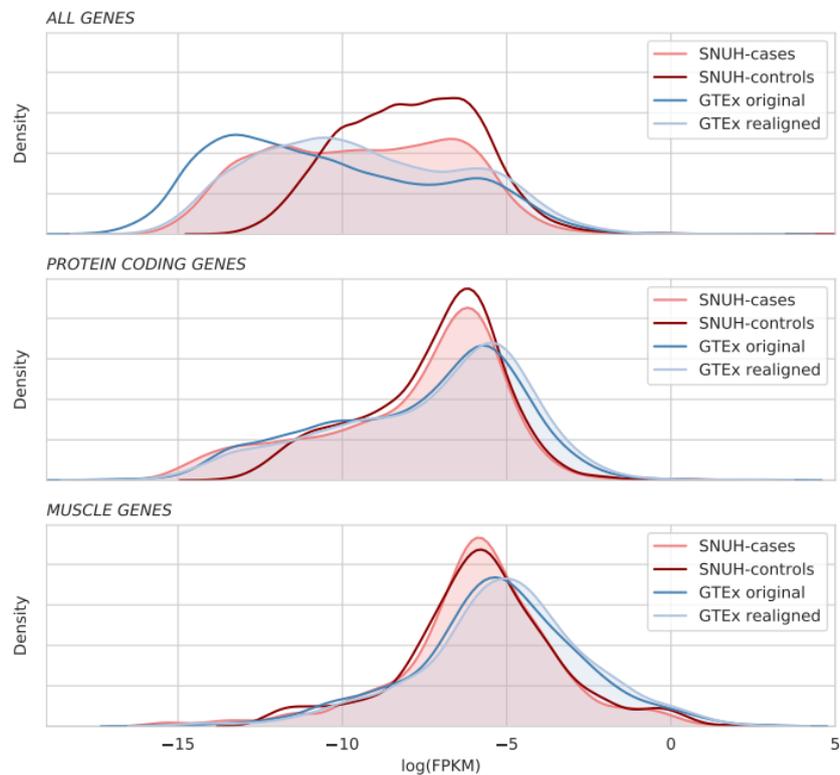


Figure 9: FPKM density distribution of FPKMs of skeletal muscle from different sources. The distributions are based on all detected genes, protein-coding genes and muscle genes.

Subsequently, I performed tSNE based clustering including count data for other tissues from GTEx to investigate whether the difference between GTEx samples and the SNUH cohort might be due to sample contamination with other tissues. As shown by Cummings *et al.* such contamination should be observable as clustering of samples with non-muscle tissues [18]. However, the tSNE analysis based on protein-coding genes showed a distinct cluster formed by the SNUH cohort and control samples instead of co-clustering with GTEx muscle or other tissues (Figure 10, left). In contrary, clustering based on muscle-disorder related

gene expression only, showed co-clustering between GTEx muscle and SNUH samples (Figure 10, right). Interestingly, sample CDC\_NM16.1 clustered with GTEx adipose tissue samples when basing clustering on muscle gene expression.

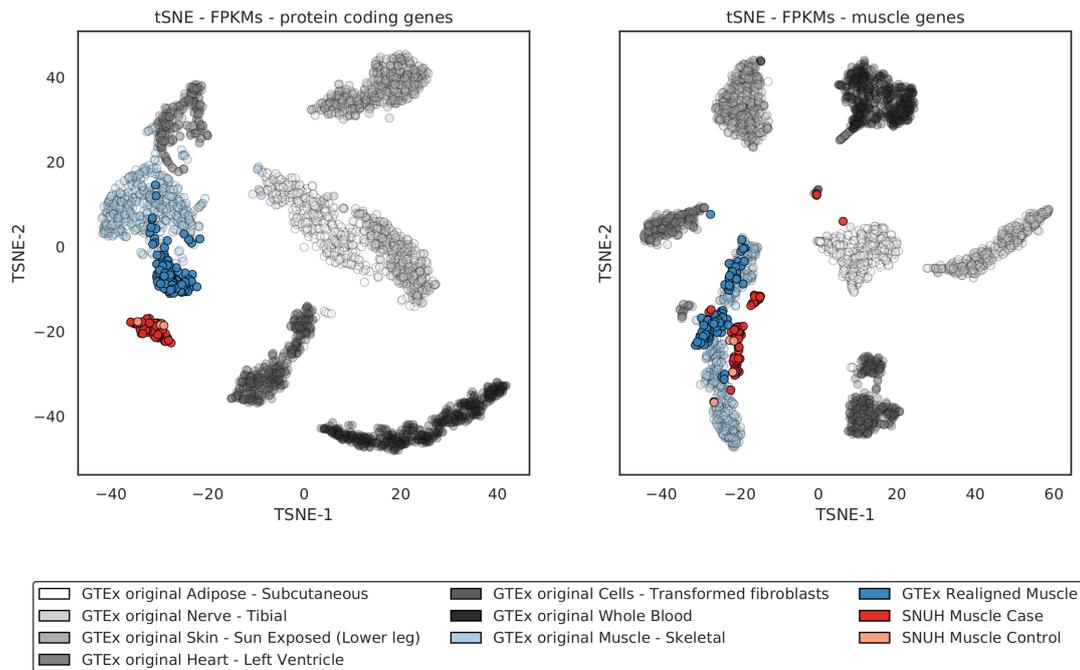


Figure 10: tSNE of GTEx original FPKMs (gray, light blue), realigned GTEx skeletal muscle FPKMs (blue), SNUH skeletal muscle cases (red) and controls (orange). Left side shows tSNE based on all protein coding genes, right side shows tSNE based on muscle genes.

Overall, the analysis showed significant differences in the transcription data sets between SNUH and GTEx that could not be sufficiently reduced by realignment of GTEx data or the exclusion of non-coding genes.

### 3.5. Transcriptome-based clustering

In order to discover underlying expression patterns in neuromuscular disorders, I performed transcriptome based clustering using non-negative matrix factorization. The analysis included all 74 muscle disorder samples with available transcriptome data, as well as the four

healthy control muscle samples.

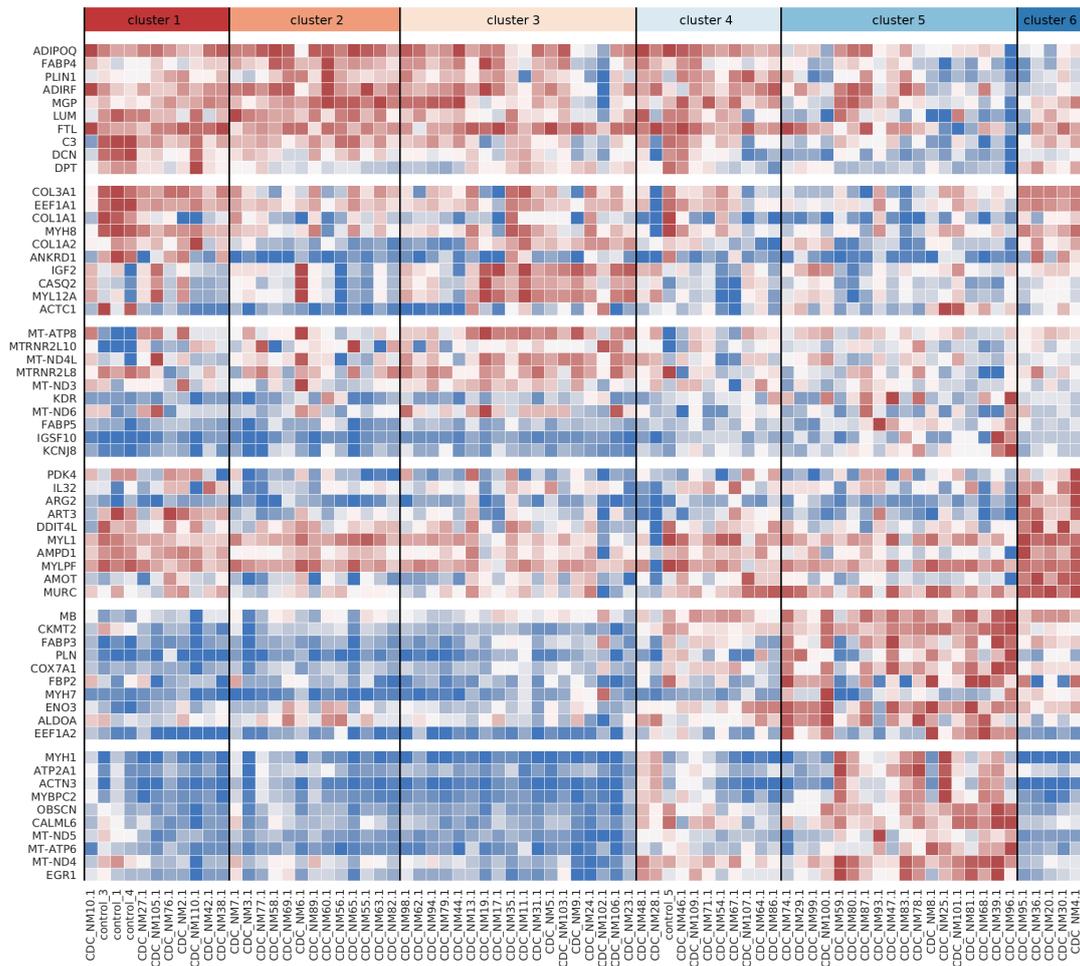


Figure 11: Cluster gene expression z-score derived from logarithmic FPKM values for all samples and the top 10 genes identified for each cluster

The unsupervised clustering of protein-coding gene expression values using non-negative matrix factorization resulted in the identification of six clusters with an average of 12.6 samples. Three out of four control samples were assigned to cluster-1, while the remaining control sample was sorted into cluster-4. Subsequently, I extracted cluster-specific gene expression patterns by selecting cluster-associated genes based on their mean difference in expression between a specific cluster and all other clusters. Figure 11 displays the expression values for the top ten genes with the highest cluster association scores for all

sample.

Next, I performed gene ontology based gene enrichment analysis using the top one-hundred cluster-associated genes (Figure 12). Cluster-1 associated genes were found to be enriched for circulatory and blood vessel genes. Commonly with cluster-6, cellular components enrichment analysis returned nucleosome and DNA packaging complex. Cluster-2 and -5 were associated with extracellular matrix space and components and shared a common association with collagen. In comparison, cluster-3's genes were associated with oxidoreductase activity, cytochrome-c oxidase activity and energy metabolism. Cluster-4 showed relation to transcription and cAMP and purine-containing compound response.

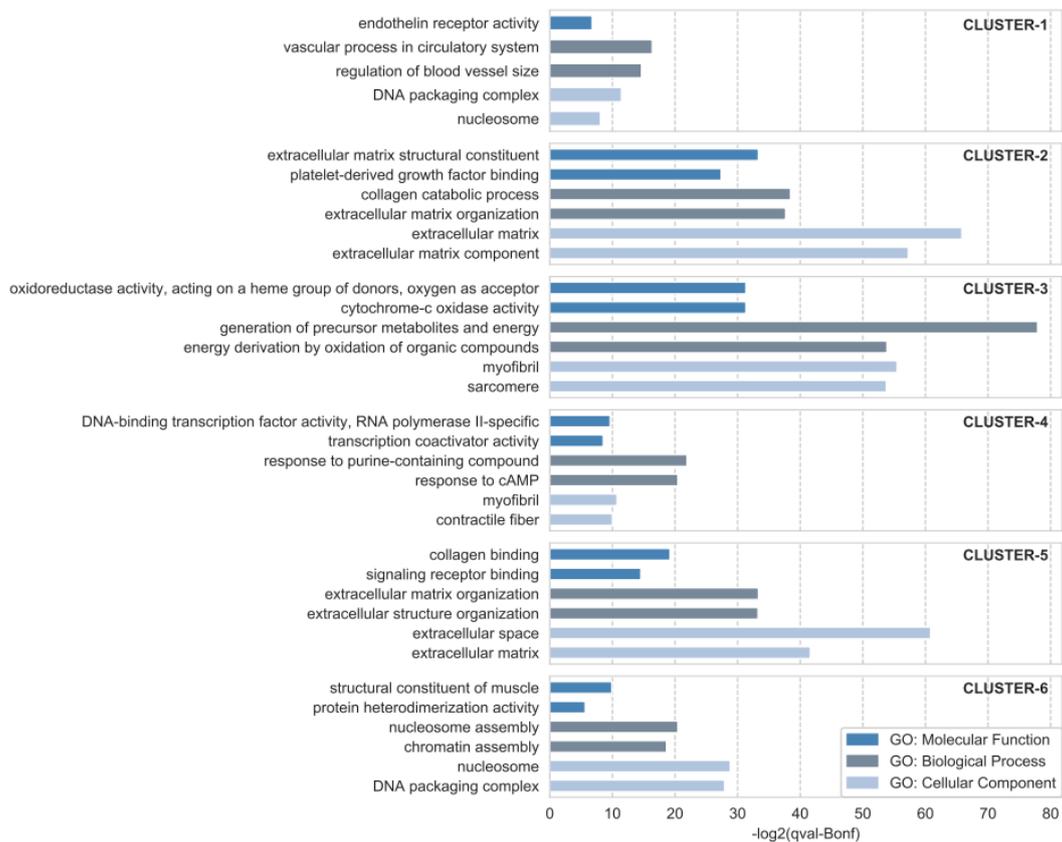


Figure 12: Gene ontology based on the top 100 genes associated with each cluster

Moreover, I explored potential correlations between the clusters and genotype, sample parameters and clinical phenotype. Comparison

of clusters and age revealed a low average age of samples within cluster-6, the smallest cluster with just five samples (Figure 13).

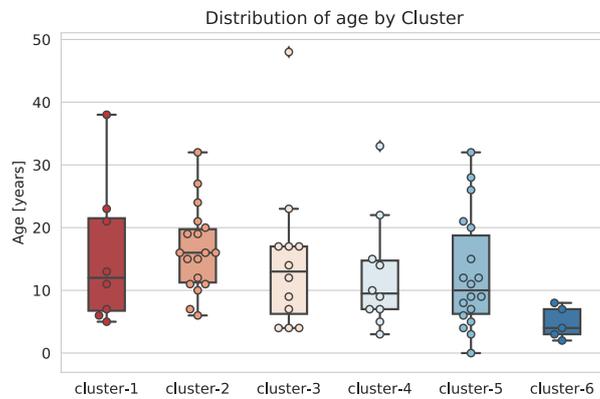


Figure 13: Box plot of sample age by cluster shows low average age in cluster-6

Concurrently, I investigated the relationship between disease category and associated cluster. However, different disease categories showed to be distributed evenly across all clusters (Figure 14).

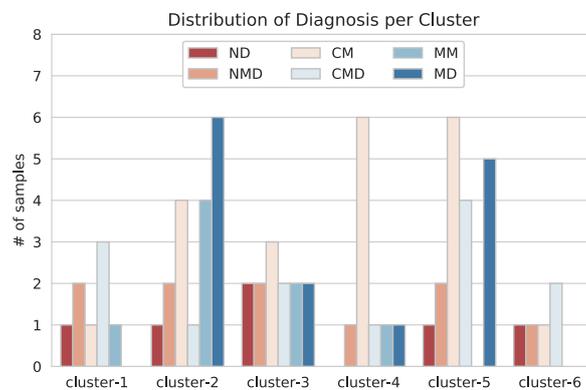


Figure 14: Number of samples for each clinical diagnosis in each cluster

Because the different disease-subtypes show overlapping symptoms, I further questioned, whether specific symptoms could explain the cluster assignment, but no obvious association could be observed (supplementary Figure 27).

In conclusion, utilization of non-negative matrix factorization

resulted in the discovery of six clusters with representative genes associated with distinct gene functions, processes and cellular components. The cluster assignments of the samples could not be explained by clinical phenotypes and genotypes, so the determining factor is yet to be discovered.

### **3.6. Exome variants in transcriptome sequencing**

Out of the 47 samples for which whole exome sequencing identified a candidate variant, transcriptome sequencing was available for 33. Approximately half of those samples (17 samples, 51.51%) carried compound heterozygous variants, six samples homozygous (18.18%), and ten samples heterozygous variants (30.30%).

#### **3.6.1. Effect of variants on gene expression levels**

First, to evaluate the effect of variants on gene expression levels, I performed differential gene expression analysis for all 33 samples. Surprisingly, in none of the samples, a significant differential expression of the mutated gene could be detected. In order to explore non-significant outliers, I visually compared gene expression levels for each gene between samples with and without a detected variant (Figure 15).

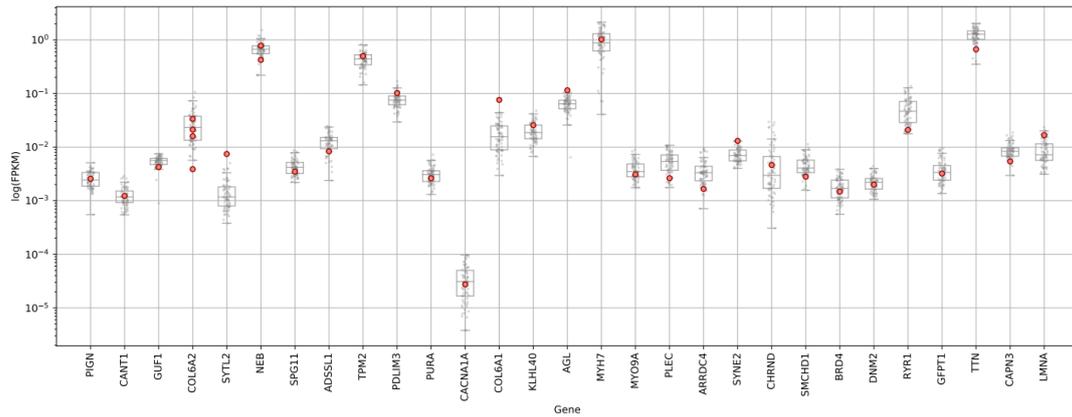


Figure 15: Box plots of logarithmic FPKMs for genes with whole exome sequencing candidate variants. Samples without candidate variant for each gene are displayed in light gray, while samples with candidate variant in respective gene are displayed in red.

While most samples showed expression levels within the interquartile range, some samples showed comparably high or low expression levels of their respective candidate gene. The minimum *COL6A2* FPKM value was detected for CDC\_NM30.1, a sample carrying a heterozygous stopgain mutation within the gene, while other samples with detected *COL6A2* mutations showed average expression levels. In contrary, CDC\_NM39.1 showed high expression levels for *COL6A1* in which whole exome sequencing identified a ClinVar known pathogenic missense variant. Other samples showing relatively high expression levels in their respective candidate genes were CDC\_NM8.1 for *SYTL2*, CDC\_NM55.1 for *AGL* and CDC\_NM68.1 for *SYNE2*.

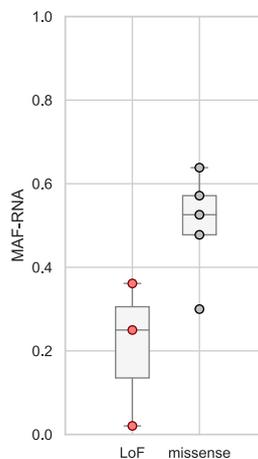
For no sample significant under- or over-expression of the candidate gene could be detected using differential gene expression and outlier gene expression analysis. However, when visually comparing FPKM levels I was able to observe slightly increased or decreased expression of the candidate gene in five samples.

### 3.6.2. Allele specific expression of candidate variants

Next, I was interested in whether heterozygous and compound heterozygous variants discovered in whole exome sequencing would

show allele specific expression favoring the intact or less damaging allele.

In total, transcriptome sequencing data was available for ten samples with heterozygous candidate variants. Out of them six samples carried a missense variant, two samples a stopgain variant and one sample each showed a frameshift variant and an inframe deletion. Due to insufficient coverage by RNA-sequencing the missense variant in CDC\_NM38.1 *CACNA1A* was excluded from allele specific expression analysis. Comparing heterozygous loss of function (LoF) and missense variants, I noticed that loss of function variants tended to show a lower minor allele frequency (MAF) in transcriptome sequencing (Figure 16). Interestingly, the missense variant with the lowest minor allele frequency displayed in this plot (CDC\_NM9.1, *NEB*) was suspected to be compound heterozygous.



*Figure 16: Comparison of MAF for LoF and missense variants in RNA-sequencing*

Moreover, I gathered transcriptome data for 17 compound heterozygous samples. Out of them, no coverage could be detected for either variant in CDC\_NM36.1 (*PURA*) and for one variant each in

CDC\_NM63.1 (*PLEC*) and CDC\_NM65.1 (*ARRDC4*). Figure 17 depicts the minor allele frequencies detected in transcriptome sequencing for the two compound heterozygous variants in each sample. For a majority of samples both variants were located within the green area, representing evidence for biallelic expression (CDC\_NM6.1, CDC\_NM11.1, CDC\_NM55.1, CDC\_NM68.1, CDC\_NM76.1, CDC\_NM83.1, CDC\_NM86.1, CDC\_NM87.1). In comparison, focusing on CDC\_NM59.1 I could observe a high minor allele frequency for the missense variant, while the loss of function allele showed a lower minor allele frequency. Other samples did not show a clear pattern of allele specific expression.

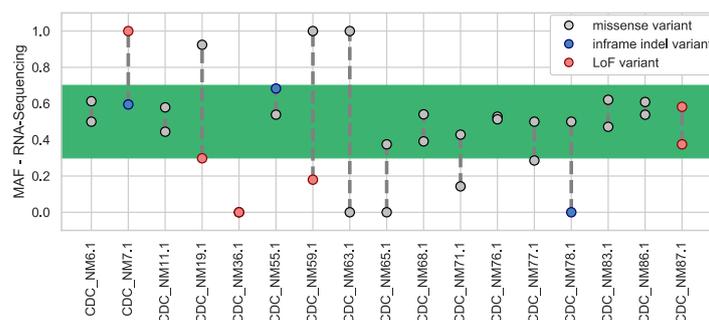


Figure 17: MAF of compound heterozygous WES variants per sample. The variants are colored by variant type. The green range shows approximate range of biallelic expression.

### 3.6.3. Alternative splicing

The candidate variants identified in whole exome sequencing included a predicted splice-acceptor variant in CDC\_NM7.1 located in *COL6A2*. However, splicing analysis using leafcutter did not detect any significant abnormal splicing within this gene.

### 3.6.4. Gene expression of genes within copy number variations

Lastly, I investigated whether samples with copy number variation would show a higher or lower expression level of the genes located within the

copy number variation (CNV). For samples CDC\_NM78.1, I detected a deletion including the genes *CLDN11*, *PRPL22L1*, *EIF5A2*, *TNKK1K*, *PRKAB2*, and *FMO5*. For all genes besides *EIF5A2*, the sample showed expression levels within the lower 25% quartile compared to the other samples. Similarly, the FPKM values for sample CDC\_NM29.1 laid within the top 25% of all samples for three out of the four genes located within the duplicated region (Figure 18).

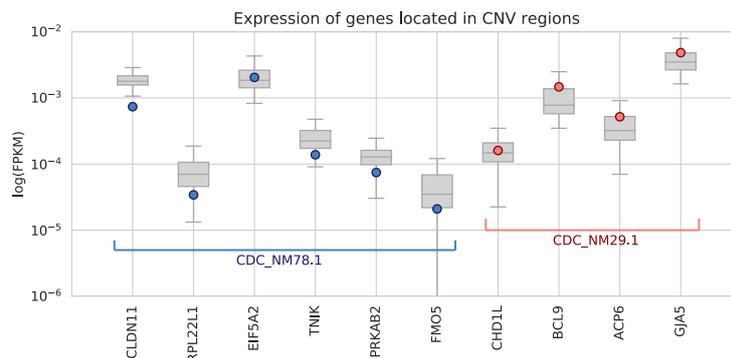


Figure 18: Logarithmic FPKMs for samples with CNVs in WES. FPKMs for all other samples are displayed as box plots. CDC\_NM78.1 showed a deletion, while CDC\_NM29.1 showed a duplication in whole exome sequencing

### 3.7. Transcriptome-sequencing based diagnosis

Out of 74 samples with available transcriptome sequencing, 40 samples remained without a exome sequencing based diagnosis. For such samples I focused on finding abnormalities in the expression profile that could provide indications of the genetic cause of the disorder.

#### 3.7.1. RNA-sequencing based variant calling

First, I performed RNA-sequencing based variant calling to discover expressed variants in intronic, untranslated regions (UTRs), and exonic regions. During the analysis I prioritized rare variants in known muscle disease genes and genes with other OMIM associations. For samples with available exome sequencing, I moreover compared called variant

genotypes between genetic and transcriptomic data and excluded all variants rejected as causative by the responsible clinicians during whole exome sequencing analysis. This analysis led to the identification of potential candidate variants in nine samples (Table 3). Out of the nine samples, whole exome sequencing based diagnosis had previously failed to identify a candidate variant in five samples. For the remaining four samples whole exome sequencing data was not available. However, it is to be noted that the variants have not yet been discussed with the clinicians and have not been confirmed by Sanger-Sequencing.

Sample	Gene	GT	Variant
CDC_NM10.1	<i>MYH6</i>	het	chr14:23863506:c.2456A>T:p.N819I
CDC_NM31.1	<i>LMNB1</i>	het	chr5:126161784:c.1596A>G
CDC_NM35.1	<i>SLC2A1</i>	het	chr1:43395389:c.742A>G:p.N819I
CDC_NM74.1	<i>ACTA1</i>	het	chr1:229568530:c.227G>A:p.G76D
CDC_NM93.1	<i>COL6A3</i>	het	chr2:238257255:c.6930dupA:p.G2311fs
CDC_NM96.1	<i>SYNE1</i>	het	chr6:152552537:c.21028G>T:p.V7010L
CDC_NM100.1	<i>ACTA1</i>	het	chr1:229568098:c.535C>T:p.R179C
CDC_NM103.1	<i>SYNE2</i>	het	chr14:64681095:c.19240T>C:p.S6414P
CDC_NM106.1	<i>AFF4</i>	het	chr5:132240060:c.1087A>C:p.K363Q

Table 3: List of candidate variants prioritized during RNA-Sequencing variant calling

### 3.7.2. Allele specific expression (ASE) of rare variants

For the 63 samples with transcriptome and exome sequencing data available, I extracted heterozygous variants from the whole exome sequencing data and investigated their minor allele frequency in RNA-sequencing.

An average of 1,868 heterozygous exome variants were sufficiently covered in RNA-sequencing for each sample. Overall, I could see that the majority of heterozygous variants (mean 1470 variants, 78.7%) were biallelic expressed (minor allele frequency between 0.3 and 0.7). Additionally, a mean of 203 (10.9%) heterozygous variants showed a bias towards the reference allele, while a mean of 36 variants (1.9%) of variants showed a bias towards the non-reference allele. This

analysis, moreover, included a mean 221.6 heterozygous whole exome variants located in known muscle disease genes of which up to five variants per sample favorable expressed the non-reference allele (Table 4).

	Mean number of variants for all genes	Mean number of variants for muscle genes
0.0 < RNA-MAF < 0.1	86.2 ± 29.7	32.5 ± 18.8
0.1 < RNA-MAF < 0.3	117.1 ± 33.7	19.5 ± 13.5
0.3 < RNA-MAF < 0.5	865.8 ± 266.6	96.7 ± 24.3
0.5 < RNA-MAF < 0.7	604.2 ± 186.4	65.5 ± 17.7
0.7 < RNA-MAF < 0.9	25.7 ± 10.5	1.5 ± 1.5
0.9 < RNA-MAF < 1.0	10.7 ± 5.9	0.1 ± 0.2
<b>Total</b>	<b>1868.4 ± 553.8</b>	<b>221.6 ± 50.2</b>

Table 4: Number of variants included in ASE analysis listed by RNA-MAF

Exploratory analysis showed a cluster strongly enriched for muscle genes with allelic imbalance towards the reference allele (Figure 19, left). However, upon closer investigation, I could show, that the majority of these variants originated in either *NEB* or *TTN* (Figure 19, right).

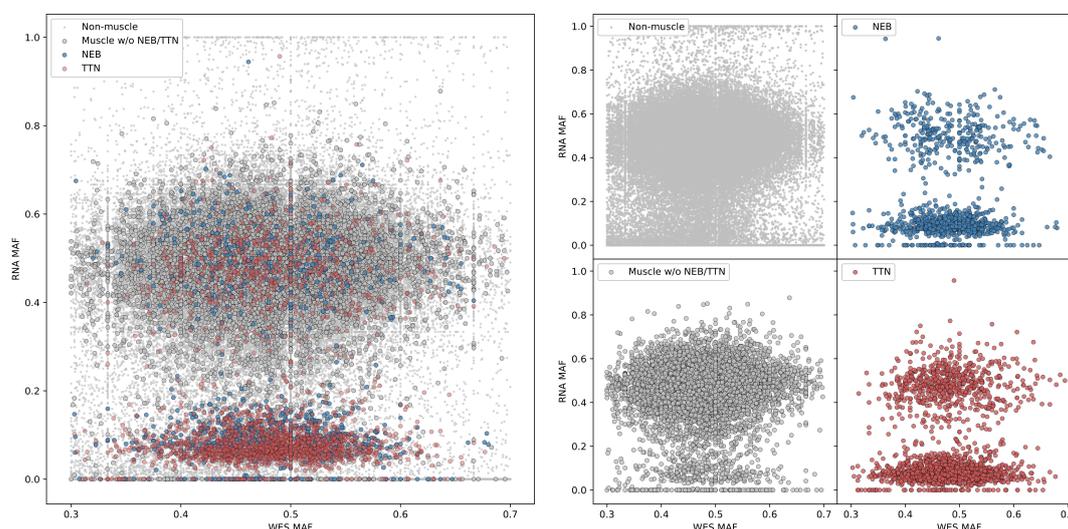


Figure 19: ASE visualized as relationship between WES-MAF and RNA-MAF. Overall plot on the left side shows a cluster at RNA-MAF between 0.0 and 0.3. The right side shows non-muscle genes (light-gray), muscle genes without *TTN* and *NEB* (gray), *TTN* (blue) and *NEB* (right) separately.

Moreover, I annotated the allele specific expression variants with

CADD score, gnomAD allele frequency and predicted variant impact and used the annotation to identify rare monoallelic expressed variants. For each detected variant I then searched for previous reports about monoallelic expression or known imprinting. In eight samples I identified a total of nine genes carrying interesting monoallelic expressed rare variants of which three variants were further prioritized (Table 5). Notably, I discovered a rare splice site variant in *NEB* for CDC\_NM9.1, a sample carrying a heterozygous stop-gain variant within the same gene. Moreover, CDC\_NM4.1 showed a monoallelic splice variant in *MYL3*, a gene associated with hypertrophic cardiomyopathy and CDC\_NM67.1 showed two monoallelic rare expressed variants in *TNXB*, an extracellular matrix protein.

Sample	Gene	Loc	NT-Change	OMIM
CDC_NM4.1	<i>MYL3</i>	chr3:46899716	c.*13+5G>C	Hypertrophic cardiomyopathy
CDC_NM9.1	<i>NEB</i>	chr2:152544262	c.24160-15A>G	Nemaline myopathy
CDC_NM67.1	<i>TNXB</i>	chr6:32035469	c.6513C>T	Ehlers Danlos syndrome
		chr6:32036450	c.5937C>T	

Table 5: Monoallelic rare variants prioritized during allele specific expression analysis. Location (Loc) and nucleotide change (NT-Change)

### 3.7.3. Differential gene expression analysis

Differential gene expression analysis was performed for all forty samples individually using all remaining samples as a control. In cases differential gene expression analysis returned less than fifteen significantly (adjusted p-value < 0.05) anomalous expressed genes, I researched each of the genes for known disease associations and previously described connection to muscular homeostasis. This method led to the discovery of muscle related abnormal genes in four samples (Table 6).

Sample	Gene	log2FC	qvalue	OMIM
CDC_NM81.1	<i>MYLK3</i>	-5.03	0.004	.
	<i>TECRL</i>	-4.94	0.006	Ventricular tachycardia, catecholaminergic
	<i>TNNI1</i>	-4.60	7.6E-05	.
CDC_NM95.1	<i>TNNC1</i>	-4.35	0.0099	Cardiomyopathy hypertrophic/distal
	<i>TPM3</i>	-3.84	0.03	CAP myopathy, CFTP, Nemaline myopathy
	<i>HK2</i>	-4.67	0.02	3MC syndrome
CDC_NM100.1	<i>AQP4</i>	-7.16	0.01	.
	<i>MYH7</i>	-5.16	0.0001	Cardiomyopathy, Laing distal myopathy, etc.
CDC_NM101.1	<i>MYL3</i>	-5.43	0.0008	Cardiomyopathy hypertrophic
	<i>MYL2</i>	-4.37	0.02	Cardiomyopathy hypertrophic
	<i>TNNC1</i>	-3.54	0.008	Cardiomyopathy hypertrophic/distal

Table 6: Differentially expressed muscle genes detected in four samples

Alternatively, for samples with more than fifteen significant differentially expressed genes, I performed gene enrichment analysis for up- and downregulated genes and looked for significant gene ontology associations. For eight samples more than fifteen genes were significantly up- or downregulated. CDC\_NM25.1 and CDC\_NM28.1 showed significant enrichment of downregulated genes for cytoskeletal protein binding, actin binding, muscle filament sliding and actin-myosin sliding, while CDC\_NM93.1 showed an up-regulation of genes involved in proton-transporting ATPase activity, transmembrane ion movement, leukocyte activation, lysosome and vacuole (Supplementary Figure 23). CDC\_NM46.1 displayed enrichment for RNA-binding and processing, as well as nucleolar part and ncRNA metabolic processes among its up-regulated genes, while it showed a down-regulation of genes involved in ubiquitin transferase, protein modifications and nucleoplasm. In comparison CDC\_NM96.1 presented with up-regulation of GTPase motor activity, pattern recognition and immune response, while cytoskeletal protein binding, muscle contraction, myofibril and sarcomer were enriched for down-regulated genes. Lastly, CDC\_NM102.1 gene enrichment analysis displayed up-regulation of enzyme binding, chromatin binding and GTPase activity and a down-regulation of RNA

binding, ribosome constituent, catabolic processes, mitochondria and ribonucleoproteins (Figure 24).

#### 3.7.4. Outlier gene expression analysis

Furthermore, I investigated, whether the minima and maxima for each gene represent significant outliers. The analysis resulted in the identification of 3,674 significant maxima, but no significant minima. In Figure 20, it can be observed, that most minima were located within two standard deviations of the mean and all minima fell within the significant p-value cutoff 0.05. In contrary maxima peaked around 3 standard deviations from the mean and showed a secondary peak at 8-9 standard deviations from the mean.

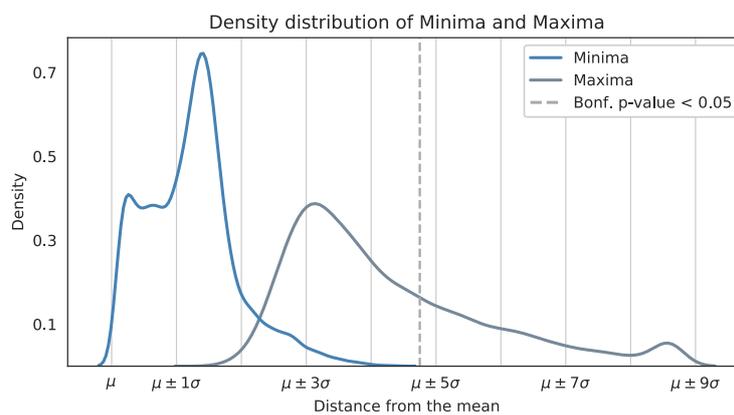


Figure 20: Density distribution of minima and maxima visualized as distance from the mean in standard deviations

#### 3.7.5. Aberrant splicing in transcriptome sequencing

Intronic variants can cause aberrant splicing events, such as exon-skipping and intron-retention, which can adversely affect protein function. Therefore, I contemplated to discover sample-specific splice junctions by comparing splicing events in each sample to all other samples and controls using leafcutter. This software first collects splice junctions from each sample and then tests each sample for each detected splice junction.

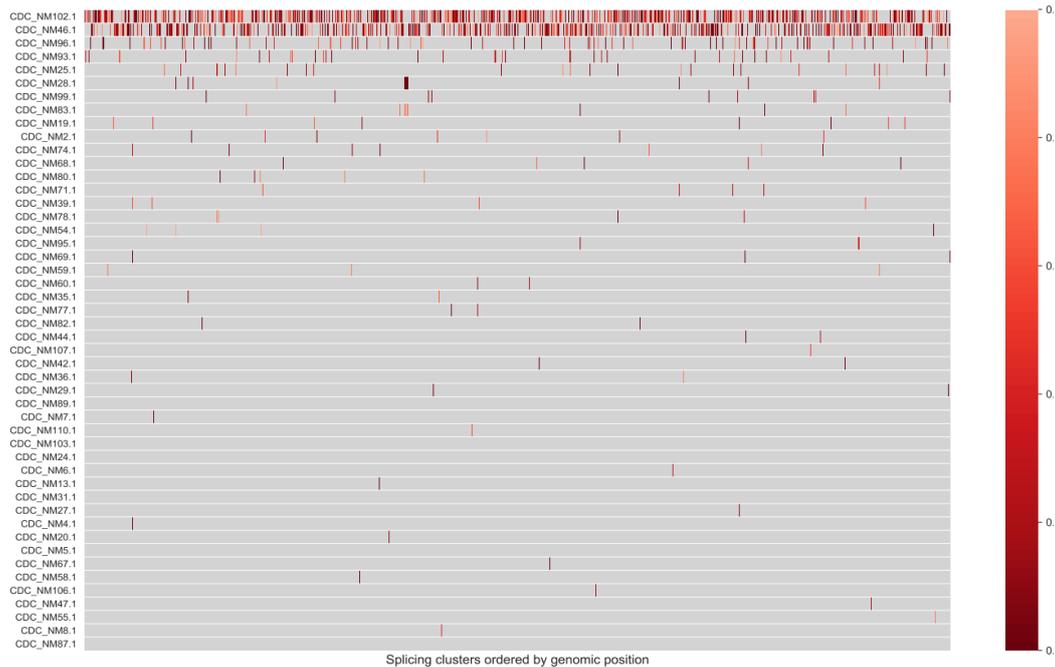


Figure 21: Heatmap showing the adjusted p-value for each sample for each analyzed splicing cluster ordered by genomic position. Gray areas show clusters in samples which did not show a adjusted p-value under 0.05.

Firstly, investigating the overall distribution of the number of significant splicing events (adjusted p-value < 0.05), it could be observed that most samples showed less than ten significant splicing events (Figure 21, supplementary Figure 25). However, for three samples the analysis resulted in the detection of over fifty significant splice junction each (CDC\_NM96.1: 87, CDC\_NM46.1: 426, CDC\_NM102.1: 537 splice junctions). In order to investigate whether the abnormally spliced genes could be related to one-another, I performed gene enrichment analysis for the three samples. Notably, all three samples showed an enrichment of genes related to RNA binding, cytoskeletal binding or enzyme binding. While CDC\_NM102.1 and CDC\_NM96.1 both further showed enrichment of muscle cell development and contractile fiber genes, CDC\_NM46.1's abnormally spliced genes were enriched for macromolecule metabolism and mitochondria.

Sample	Gene	OMIM	Adjusted p-value
CDC_NM54.1	<i>DMD</i>	Duchenne Muscular Dystrophy	6.02E-16
CDC_NM55.1	<i>DMD</i>	Duchenne Muscular Dystrophy	0.04
		Cardiomyopathy, dilated, 1G	
		Cardiomyopathy, familial hypertrophic, 9	
CDC_NM67.1	<i>TTN</i>	Muscular dystrophy, limb-girdle 10	4.09E-44
		Myopathy, myofibrillar 9, with early respiratory failure	
		Salih myopathy	
		Tibial muscular dystrophy, tardive	
CDC_NM82.1	<i>MICU1</i>	Myopathy with extrapyramidal signs	2.96E-22

Table 7: Significant aberrant splicing events of muscle genes detected by leafcutter

Identical to the differential gene expression analysis, I individually researched each aberrant spliced gene for samples with less than fifteen significant detected splice junctions. The Individual analysis for each sample revealed muscle gene abnormal splicing events in four samples (Table 7). For sample CDC\_NM67.1 alternative splicing in *TTN*, a gene linked to titinopathy, was shown. Upon closer investigation, I found a de novo inframe [CTT] deletion located at the exon border causing a three basepair shortened exon. Samples CDC\_NM54.1 and CDC\_NM55.1 were clinically described to show a typical *DMD*- related phenotype with dystrophin 3 loss visible in the pathology. However, neither whole exome sequencing nor RNA-sequencing variant calling was able to identify *DMD* variants for either samples. Using this alternative splicing approach, I was able to identify cryptic *DMD* splice junctions in both samples. Lastly, in sample CDC\_NM82.1 I discovered a homozygous exon skipping event in *MICU1*, a gene previously described for myopathy with extrapyramidal signs. The detected abnormal splicing junctions are visualized in Figure 22.

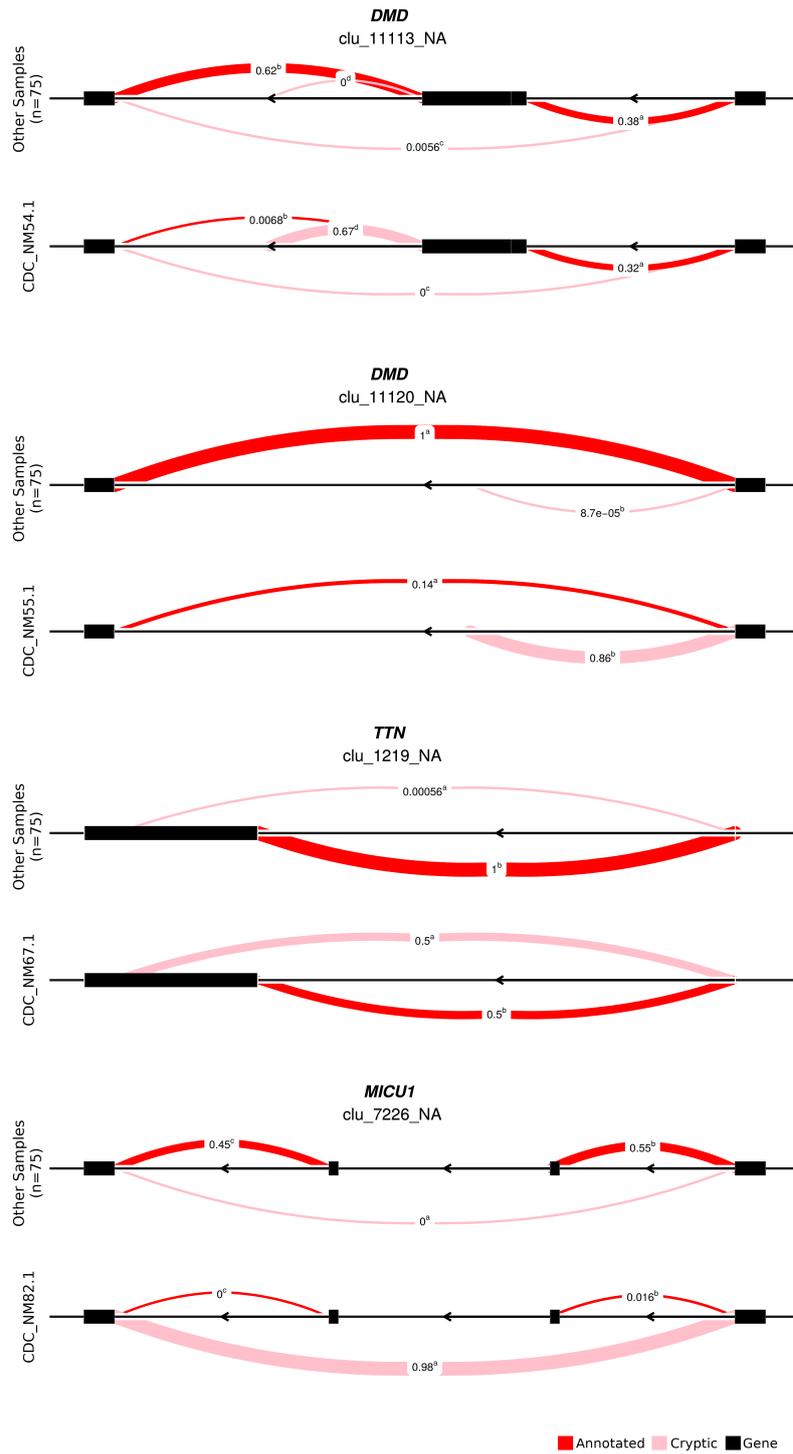


Figure 22: Aberrant splicing detected in four samples for DMD, TTN and MICU1

## 4. Discussion

Here, I present a comprehensive analysis of exome and transcriptome for a diverse cohort of 117 neuromuscular disorder patients. Analysis of 96 patient exomes yielded a diagnosis rate of 49% with the majority of variants located in known muscle disease genes. Subsequently, I investigated whether the influence of the detected variants could be observed in transcriptome sequencing. For five samples slightly abnormal expression levels could be detected. Moreover, allele specific expression analysis revealed allelic imbalance of heterozygous loss of function variants and monoallelic expression of a missense variant in a sample with a missense and a loss of function compound heterozygous variant in *COL6A2*. For samples with negative whole exome sequencing, transcriptome sequencing further helped identify heterozygous missense variants in nine samples and led to the discovery of abnormal *DMD* splicing in two samples. While allele specific expression analysis guided me in the discovery of two likely disease related monoallelic rare expressed variants, differential gene expression highlighted abnormal thin and thick filament expression levels in two samples. Lastly, I utilized transcriptome data to identify distinct expression-based clusters. These clusters were not related to phenotype or genotype of the samples, opening up the interesting discussion on what might be the driving force underneath the clusters.

Based on previous research of muscle transcriptome, I aimed to use GTEx healthy muscle samples as control. However, comparison of counts distribution and tSNE-based clustering revealed striking difference between SNUH transcriptome sequencing results and those obtained from GTEx. These differences are unlikely to have originated from the changes between healthy and diseased samples, because the

internal SNUH control samples clustered well with the SNUH cases. Most likely, the differences originated from the different sequencing processes (stranded vs. non-stranded, whole RNA vs. polyA-enriched RNA) and downstream analyses. Interestingly, when examining muscle genes only, SNUH and GTEx samples co-clustered showing high similarity of tissue-specific gene expression. However, one sample of the SNUH cohort co-clustered with GTEx adipose tissue samples, suggesting either contamination of the biopsy with adipose tissue or extreme changes in expression of the sample.

Whole exome sequencing resulted in an identification of a candidate gene in 49% of samples. However, the majority of variants has yet to be confirmed by Sanger-sequencing. Heterozygous variants inherited from unaffected parents, as well as compound heterozygous variants that show to be located on the same allele, will have to be excluded from the final diagnosis rate. Previous studies applying whole exome sequencing for the diagnosis of neuromuscular disorders achieved overall diagnosis rates of 69% [56], 79% [57] and 57% [58]. In comparison the diagnosis rate of 49% described here is relatively small. This could be due to the fact, that a considerable amount of the samples had previously undergone other genetic studies, such as muscle gene panel sequencing. It might even be considered surprising, that for such samples whole exome sequencing revealed pathogenic mutations in muscle disorder genes that had not been discovered in previous tests.

Moreover, I aimed to evaluate whether functional information gained from transcriptome sequencing could help in the interpretation of whole exome sequencing candidate variants. In total I investigated the transcriptomic data for 33 samples with positive whole exome sequencing. Variants in four samples were not sufficiently covered in

transcriptome sequencing. Two of those genes (*PURA*, *CACNA1A*) are associated with primarily neurological disorders and expression in muscle was hence not expected. Overall expression values for the respective candidate genes were observed to be cohort lowest or highest in five samples. However, significant differential expression could not be shown in any. Missense variants frequently show no effect on expression levels, however it could have been possible, that they influence RNA degradation, stability or transcription factor binding. Moreover, investigation of allele specific expression revealed allelic imbalance towards the reference in two heterozygous loss of function variants. Samples with compound heterozygous variants carry a different variant in each allele. It is therefore interesting to study, whether both alleles are expressed equally in a sample. Here, I could show that the majority of samples with compound heterozygous variants expressed both variants simultaneously. However, one sample showed monoallelic expression of the allele carrying a missense allele, while the loss of function allele was not expressed. However, more data will be necessary to provide an evidence-based interpretation of these results: Firstly, sequencing bias and alignment bias could have falsified the minor allele frequency and secondly, the majority of the variants have not been confirmed as compound heterozygous by Sanger sequencing and might thus be located on the same allele.

Whole exome sequencing analysis further resulted in the identification of twelve copy number variations. However, all but two variations showed overlapping structural variants in gnomAD-SV and none of them included a known muscle disease gene. Copy number variations are predicted to impact gene dosage, however when comparing the expression levels of affected genes between samples

with copy number variations and controls, only small differences could be observed. Therefore, I was not able to assign pathogenicity to any of them.

Overall transcriptome sequencing provided additional value for functional interpretation in nine out of 33 samples (27%). However, the evidence is arguably very weak and additional functional studies will be necessary.

For forty samples without whole exome sequencing results, I aimed to utilize transcriptome sequencing as a means to identify pathogenic events leading to neuromuscular disorders. First, I performed RNA-sequencing based variant calling and were able to identify candidate genes for nine samples. Four of those samples had previously undergone exome sequencing without successful variant prioritization. All variants were found in heterozygous state, so segregation analysis and Sanger confirmation will still be necessary to assign pathogenicity to any of the variants.

An average of 221 heterozygous muscle gene variants from whole exome sequencing were sufficiently covered in transcriptome sequencing to assess allele specific expression. The analysis showed allelic imbalance towards the non-reference allele for *TTN* and *NEB* across multiple samples. This is to my knowledge the first report of allelic imbalanced expression of *TTN* and *NEB*. Moreover, I discovered four rare monoallelic expressed variants. In CDC\_NM9.1 I was able to uncover a rare splice site variant showing monoallelic expression. The sample further showed an additional stop-gain variant in *NEB*, leading me to the hypothesis that the sample expresses only the less damaging allele. The splice site variant is extremely rare showing an allele frequency of 0.000001 in gnomAD and no record of homozygous allele

counts. Even though alternative splicing analysis did not reveal any evidence for aberrant splicing in *NEB*, I suggest the two compound heterozygous variants as likely pathogenic and recommend segregation analysis to confirm them. Additionally, I discovered a rare monoallelic expressed variant in *MYL3*. The gene has been described to cause autosomal dominant or recessive cardiomyopathy. However, due to CDC\_NM4.1's clinical information suggesting Ulrich's muscular dystrophy or congenital myopathy, I expect the *MYL3* variant to be non-pathogenic. Lastly, monoallelic rare variant analysis revealed two protein coding heterozygous missense variants in *TNXB* for sample CDC\_NM67.1. The gene has previously been described to cause Ehlers Danlos syndrome as well as autosomal recessive primary myopathy [59]. In contrary to congenital disease onset in patient CDC\_NM67.1, the patient in the previous report stayed asymptomatic until age 30. However, due to the lack of further clinical reports of recessive *TNXB* related myopathy, I could not exclude the variant as potential cause.

Aberrant splicing analysis showed significant splicing events in *MICU1*, *TTN* and *DMD* for four samples. I suggest to accept abnormal *DMD* splicing as pathogenic for CDC\_NM54.1 and CDC\_NM55.1, because both samples displayed loss of dystrophin 3 in their muscle biopsies. Whole exome sequencing had formerly led to the identification of a compound heterozygous *AGL* variant in CDC\_NM55.1. However, one of the two variants has been described as 'Likely benign' in the ClinVar database. Hence, I came to the conclusion that the cryptic splice site in *DMD* presents a more likely disease cause than the compound heterozygous *AGL* variant. It is noteworthy that both samples underwent previous *DMD* target gene sequencing without convincing results. Indeed, if possible, I recommend reanalysis of that data in search

of intronic variants in proximity of the cryptic splice sites. This analysis further revealed an abnormal splicing in *TTN* for sample CDC\_NM67.1, that I could show to actually be a three basepair deletion. Mutations in *TTN* can cause autosomal dominant myopathy and muscular dystrophy, but, owing to the outright size of the gene, interpretation of variants of uncertain significance remains challenging. For CDC\_NM82.1, a patient suffering from limb-girdle muscular dystrophy and congenital myopathy, I discovered a homozygous exon skipping event in *MICU1*, a gene encoding the mitochondrial  $\text{Ca}^{2+}$  uptake uniporter related to myopathy with extrapyramidal signs and learning disabilities [60]. Considering that there was no notion of any extra-muscular symptoms included in the clinical information, I conclude that abnormal splicing of *MICU1* is unlikely disease causing.

Differential gene expression revealed a significant down-regulation of *MYH7*, *MYL2* and *MYL3* in CDC\_NM101.1. All three genes encode thick filament proteins and have been related to cardiac myosin and cardiomyopathy. *MYH7* has further been described in autosomal dominant and recessive myopathies. The under-expression of the three myosin genes might suggest a thick filament dysfunction, possibly due to a mutation in *MYH7*. In contrary, the sample CDC\_NM95.1 showed a down-regulation of *TNNI1*, *TNNC1*, *TPM3*, three genes related to the thin filament. This could be interpreted as an indication for a thin filament dysfunction with potential genetic mutations in actin, troponin or tropomyosin genes.

Aberrant splicing analysis revealed a high number of abnormally spliced genes for CDC\_NM96.1, CDC\_NM46.1 and CDC\_NM102.1. Thus, I suspected an upstream mutation that affects splicing patterns. However, whole exome sequencing data did not reveal any rare variants

located within known splicing-associated genes. The two latter samples further showed a high number of significantly differentially expressed genes. While RNA binding, RNA processing, ribonucleoar protein complex biogenesis related gene expression levels were found to be increased in CDC\_NM46.1, protein modifications showed to be downregulated. This could indeed indicate a splicing defect, but could also be a mere expression of early stage muscular dystrophy during which ribosome biogenesis is increased [61]. Similarly, CDC\_NM102.1 displayed an abnormal RNA household with down-regulation of RNA-binding, ribosome constituents and ribonucleoprotein and up-regulation of transcriptional regulation, chromatin binding and nucleoplasm. Whole exome sequencing identified a homozygous mutation in *LMNA* in the sample, which could strongly affect nucleus and chromatin structure [62].

Expression based clustering revealed six clusters in the SNUH cohort. I assigned representative genes to each cluster and performed gene ontology analysis. Out of four controls three were assigned to the same cluster. It remains debatable as to why one control was assigned to a different cluster, but possible reasons include age-specific expression profiles, muscle biopsy contamination or athletic status of the sample. Cluster-specific genes were selected as genes that show greatest difference in expression between the specific cluster and the other clusters. Hence, gene enrichment analysis of these genes could not reveal, whether a certain process is up- or downregulated, but merely signifies that a certain process seems to be different within one cluster compared to the others. In the case of cluster-1 and cluster-2 gene enrichment analysis returned strong signals for extracellular matrix-related process and collagen. As a matter of fact, difference in

gene expression levels of genes affecting extracellular matrix have been described. The over- and under-expression of a variety of matrix metalloproteases have been related to muscle fibrosis and suggested as biomarkers, previously [62, 63]. Genes selected for cluster-1 revealed to be enriched for endothelial receptors and other vascular processes. Using electromicroscopy, it has been discovered that indeed vascular structures in affected muscle differs greatly from control samples: vascular endothelial cells displayed blister like swelling and capillary diameter was found to be greatly increased [63]. Further studies have additionally described increased vascular-endothelial growth factor expression in affected samples [56]. Moreover, cluster-1 shared an association of cluster gene enrichment for DNA packaging and nucleosome with cluster-6. Epigenetic changes have been described to play an important role in skeletal muscle hypertrophy and ribosome production during early stages of muscular dystrophy [57]. Due to cluster-1 containing three control samples, it could be hypothesized, that the samples within the cluster were still in an early disease stage or had a less severe muscle phenotype at the time of the biopsy. Interestingly, cluster-3 specific genes were related to oxidoreductase activity and cytochrome-c oxidase as well as energy metabolism, giving rise to the hypothesis that samples within cluster-3 showed abnormal mitochondrial activity. Finally, cluster-6 showed a distinct expression profile of cAMP-response genes. Cyclic AMP is an important second messenger, that is involved in the regulation of a variety of cellular process including sarcoplasmic calcium dynamics, contractility and muscle regeneration [65]. I furthermore, tested whether any of the clinical parameters showed a correlation with the cluster assignment, but could not find any significant associations. Hence, I suggest that the

clusters might represent different stages of disease progression. The composition of the muscle can change greatly for progressive disorders such as Duchenne's muscular dystrophy with an increased fibrosis. Therefore, further investigation of cluster and pathology results could be performed to test this hypothesis.

In conclusion, I present a comprehensive study on the value of integration of whole exome and transcriptome sequencing for the diagnosis of rare neuromuscular disorders. RNA-sequencing successfully helped in the interpretation of whole exome sequencing variants and revealed interesting expression patterns that open new areas for further investigation.

## 5. References

- [1] G. Bertier and Y. Joly, "Clinical exome sequencing in France and Quebec: what are the challenges? What does the future hold?," *Life Sci. Soc. Policy*, vol. 14, no. 1, 2018.
- [2] C. Bacchelli and H. J. Williams, "Opportunities and technical challenges in next-generation sequencing for diagnosis of rare pediatric diseases," *Expert Rev.*, 2016.
- [3] M. A. Ferreira *et al.*, "Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer," *Nat. Commun.*, vol. 10, no. 1, p. 1741, 2019.
- [4] M. Lamontagne *et al.*, "Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations," *Hum. Mol. Genet.*, vol. 27, no. 10, pp. 1819-1829, 2018.
- [5] T. Aneichyk *et al.*, "Dissecting the Causal Mechanism of X-Linked Dystonia- Parkinsonism by Integrating Genome and Transcriptome Assembly," *Cell*, vol. 172, no. 5, pp. 897-909, 2018.
- [6] G. Bonne, F. Rivier, and D. Hamroun, "GeneTable of Neuromuscular Disorders," *World Muscle Society*, 2019. [Online]. Available: <http://www.musclegenetable.fr/index.html>.
- [7] H. D. Gonorazky, C. G. Bönnemann, and J. J. Dowling, "The genetics of congenital myopathies," *Handb. Clin. Neurol.*, vol. 148, pp. 549-564, 2018.
- [8] D. Cassandrini *et al.*, "Congenital myopathies: Clinical phenotypes and new diagnostic tools," *Ital. J. Pediatr.*, vol. 43, no. 1, pp. 1-16, 2017.
- [9] H. Jungbluth *et al.*, "Congenital myopathies: Disorders of excitation-contraction coupling and muscle contraction," *Nat. Rev. Neurol.*, vol. 14, no. 3, pp. 151-167, 2018.
- [10] J. C. Carter, D. W. Sheehan, A. Prochoroff, and D. J. Birnkrant, "Muscular Dystrophies," *Clin. Chest Med.*, vol. 39, no. 2, pp. 377-389, 2018.
- [11] K. Zhang, X. Yang, G. Lin, Y. Han, and J. Li, "Molecular genetic testing and diagnosis strategies for dystrophinopathies in the era of next generation sequencing," *Clin. Chim. Acta*, vol. 491, no. January, pp. 66-73, 2019.
- [12] J. Ehmsen, E. Poon, and K. Davies, "The Dystrophin-Associated Protein Complex," *J. Cell Sci.*, vol. 2002, pp. 2801-2803, 2002.

- [13] S. mi Kang, M. H. Yoon, and B. J. Park, "Laminopathies; Mutations on single gene and various human genetic diseases," *BMB Rep.*, vol. 51, no. 7, pp. 327-337, 2018.
- [14] P. O., G. C. J.A., N. N., H. L. H., and R. A., "GNE myopathy: From clinics and genetics to pathology and research strategies," *Orphanet J. Rare Dis.*, vol. 13, no. 1, pp. 1-15, 2018.
- [15] J. P. Fichna, A. Maruszak, and C. Żekanowski, "Myofibrillar myopathy in the genomic context," *J. Appl. Genet.*, vol. 59, no. 4, pp. 431-439, 2018.
- [16] A. El-Gharbawy and J. Vockley, "Inborn Errors of Metabolism with Myopathy," *Pediatr. Clin. North Am.*, vol. 65, no. 2, pp. 317-335, 2017.
- [17] L. J. Carithers and H. M. Moore, "The Genotype-Tissue Expression (GTEx) Project," *Biopreserv. Biobank.*, vol. 13, no. 5, pp. 307-308, 2015.
- [18] B. B. Cummings *et al.*, "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing," *Sci. Transl. Med.*, vol. 9, no. 386, p. eaal5209, 2017.
- [19] L. S. Kremer *et al.*, "Genetic diagnosis of Mendelian disorders via RNA sequencing," *Nat. Commun.*, vol. 8, pp. 1-11, 2017.
- [20] "Database of Genotypes and Phenotypes (dbGAP)," *Information, National Center for Biotechnology*, 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/gap>.
- [21] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589-595, 2010.
- [22] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078-2079, 2009.
- [23] Broad Institute, "Picard toolkit," *Broad Institute, GitHub repository*, 2019. [Online]. Available: <http://broadinstitute.github.io/picard/>.
- [24] M. D. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel *et al.*, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 254-260, 2010.
- [25] K. Okonechnikov, A. Conesa, and F. García-Alcalde, "Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data," *Bioinformatics*, vol. 32, no. 2, pp. 292-294, 2016.
- [26] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize

- analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047-3048, 2016.
- [27] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin, “BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data,” *Bioinformatics*, vol. 32, no. 11, pp. 1749-1751, 2016.
- [28] P. Cingolani *et al.*, “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3,” *Landes Biosci.*, vol. 5, no. 1, pp. 29-30, 2013.
- [29] P. Cingolani *et al.*, “Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift,” *Front. Genet.*, vol. 3, no. MAR, pp. 1-9, 2012.
- [30] P. Bolme, B. Borgstrom, and K. Carlstrom, “Longitudinal study of adrenocortical function following allogeneic bone marrow transplantation in children,” *Horm. Res.*, vol. 43, no. 6, pp. 279-285, 1995.
- [31] T. 1000 G. P. Consortium, “A global reference for human genetic variation,” *Nature*, vol. 20, no. 2, pp. 163-178, 2015.
- [32] M. Lek *et al.*, “Analysis of protein-coding genetic variation in 60,706 humans,” *Nature*, vol. 536, no. 7616, pp. 285-91, 2016.
- [33] D. Muddyman, C. Smee, H. Griffin, and J. Kaye, “Implementing a successful data-management framework: The UK10K managed access model,” *Genome Med.*, vol. 5, no. 11, p. 1, 2013.
- [34] S. Lee *et al.*, “Korean Variant Archive (KOVA): A reference database of genetic variations in the Korean population,” *Sci. Rep.*, vol. 7, no. 1, pp. 1-9, 2017.
- [35] D. D. Disorders, “Large-scale discovery of novel genetic causes of developmental disorders,” *Nature*, vol. 519, no. 7542, pp. 223-228, 2015.
- [36] J. H. U. McKusick-Nathans Institute of Genetic Medicine, “Online Mendelian Inheritance in Man, OMIM,” 2019. [Online]. Available: <https://omim.org/>.
- [37] R. L. Collins *et al.*, “An open resource of structural variation for medical and population genetics,” *bioRxiv*, p. 578674, 2019.
- [38] Broad Institute, “GATK Best Practices for Variant Calling from RNA-Sequencing,” 2019. [Online]. Available: <https://software.broadinstitute.org/gatk/documentation/article.php?>

id=3891.

- [39] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15-21, 2013.
- [40] K. J. Karczewski *et al.*, “Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes,” *bioRxiv*, p. 531210, 2019.
- [41] M. J. Landrum *et al.*, “ClinVar: Improving access to variant interpretations and supporting evidence,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062-D1067, 2018.
- [42] A. Miles, S. Rae, P. Ralph, and R. Pisupati, “scikit-allel: A Python package for exploring and analysing genetic variation data,” 2019. [Online]. Available: <https://github.com/cggh/scikit-allel>.
- [43] Y. I. Li *et al.*, “Annotation-free quantification of RNA splicing using LeafCutter,” *Nat. Genet.*, vol. 50, no. 1, pp. 151-158, 2018.
- [44] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, “ToppGene Suite for gene list enrichment analysis and candidate gene prioritization,” *Nucleic Acids Res.*, vol. 37, no. SUPPL. 2, pp. 305-311, 2009.
- [45] Y. Liao, G. K. Smyth, and W. Shi, “FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features,” *Bioinformatics*, vol. 30, no. 7, pp. 923-930, 2014.
- [46] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.
- [47] E. Jones, T. Oliphant, and P. Peterson, “SciPy: Open source scientific tools for Python,” *Online accessed*, 2019. [Online]. Available: <http://www.scipy.org/>.
- [48] S. Seabold, J. Perktold, C. Fulton, and K. Shedden, “StatsModels - Statistics in Python,” *Github repository*, 2019. [Online]. Available: <https://github.com/statsmodels>.
- [49] S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen, “Tools and best practices for data processing in allelic expression analysis,” *Genome Biol.*, vol. 16, no. 1, pp. 1-12, 2015.
- [50] B. S. Pedersen, R. M. Layer, and A. R. Quinlan, “Vcfanno: Fast, flexible annotation of genetic variants,” *Genome Biol.*, vol. 17, no. 1, pp. 1-9, 2016.
- [51] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, “CADD: Predicting the deleteriousness of variants throughout the human

- genome,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886-D894, 2019.
- [52] A. G. Robertson *et al.*, “Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer,” *Cell*, vol. 171, no. 3, p. 540-556.e25, 2017.
- [53] M. Zitnik and B. Zupan, “NIMFA: A Python Library for Nonnegative Matrix Factorization,” vol. 13, pp. 849-853, 2018.
- [54] S. Lee, S. Lee, S. Ouellette, W. Y. Park, E. A. Lee, and P. J. Park, “NGSCheckMate: Software for validating sample identity in Next-generation sequencing studies within and across data types,” *Nucleic Acids Res.*, vol. 45, no. 11, p. e103, 2017.
- [55] H. D. Gonorazky *et al.*, “Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease,” *Am. J. Hum. Genet.*, vol. 104, no. 3, pp. 466-483, 2019.
- [56] J. P. Fichna *et al.*, “Whole-exome sequencing identifies novel pathogenic mutations and putative phenotype-influencing variants in Polish limb-girdle muscular dystrophy patients,” *Hum. Genomics*, vol. 12, no. 1, pp. 1-12, 2018.
- [57] D. Schofield *et al.*, “Cost-effectiveness of massively parallel sequencing for diagnosis of paediatric muscle diseases,” *npj Genomic Med.*, vol. 2, no. 1, 2017.
- [58] S. Puusepp *et al.*, “Effectiveness of whole exome sequencing in unsolved patients with a clinical suspicion of a mitochondrial disorder in Estonia,” *Mol. Genet. Metab. Reports*, vol. 15, no. March, pp. 80-89, 2018.
- [59] I. Pénisson-Besnier *et al.*, “Compound heterozygous mutations of the TNXB gene cause primary myopathy,” *Neuromuscul. Disord.*, vol. 23, no. 8, pp. 664-669, 2013.
- [60] S. Musa *et al.*, “A Middle Eastern Founder Mutation Expands the Genotypic and Phenotypic Spectrum of Mitochondrial MICU1 Deficiency: A Report of 13 Patients,” *JIMD Rep.*, 2010.
- [61] F. von Walden, C. Liu, N. Aurigemma, and G. A. Nader, “mTOR signaling regulates myotube hypertrophy by modulating protein synthesis, rDNA transcription, and chromatin remodeling,” *Am. J. Physiol.*, vol. 311, pp. 663-672, 2016.
- [62] M. Mio *et al.*, “Structural instability of lamin A tail domain modulates its assembly and higher order function in Emery-Dreifuss muscular dystrophy,” *Biochem. Biophys. Res. Commun.*, vol. 512, no. 1, pp. 22-28, 2019.
- [63] S. Sugino, M. Miyatake, Y. Ohtani, K. Yoshioka, T. Miike, and M. Uchino,

“Vascular alterations in Fukuyama type congenital muscular dystrophy,”  
*Brain Dev.*, vol. 13, no. 2, pp. 77-81, 1991.

[64] T. Saito, Y. Yamamoto, T. Matsumura, H. Fujimura, and S. Shinno,  
“Serum levels of vascular endothelial growth factor elevated in patients  
with muscular dystrophy,” *Brain Dev.*, vol. 31, no. 8, pp. 612-617, 2009.

[65] R. Berdeaux and R. Stewart, “cAMP signaling in skeletal muscle  
adaptation: hypertrophy, metabolism, and regeneration,” *Am. J. Physiol.*, vol.  
303, 2012.

## 6. Appendix

### 6.1. Supplementary Figures

#### 6.1.1. Differential gene expression analysis

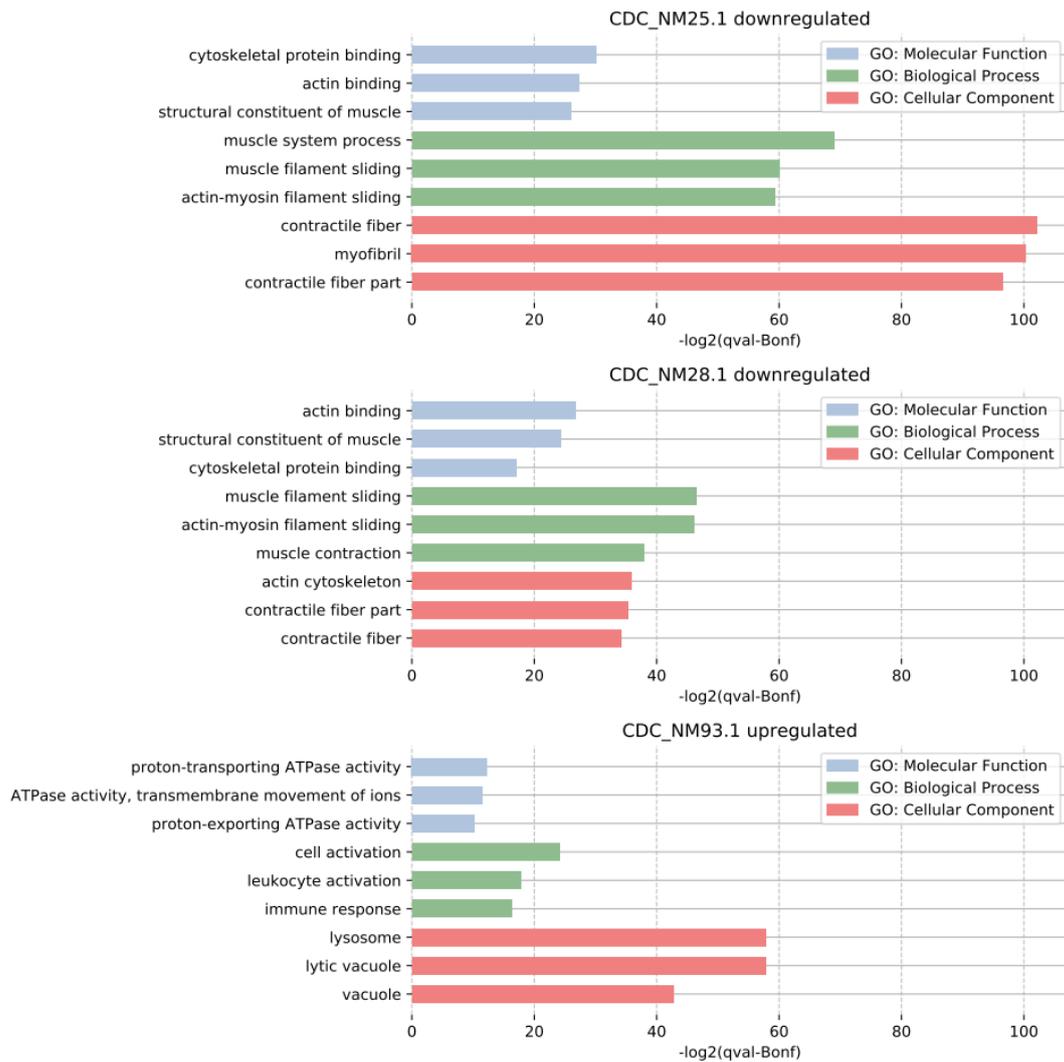


Figure 23: Gene enrichment analysis for CDC\_NM25.1, CDC\_NM28.1 and CDC\_NM93.1. All samples only showed significant enrichment for either up- or downregulated genes.

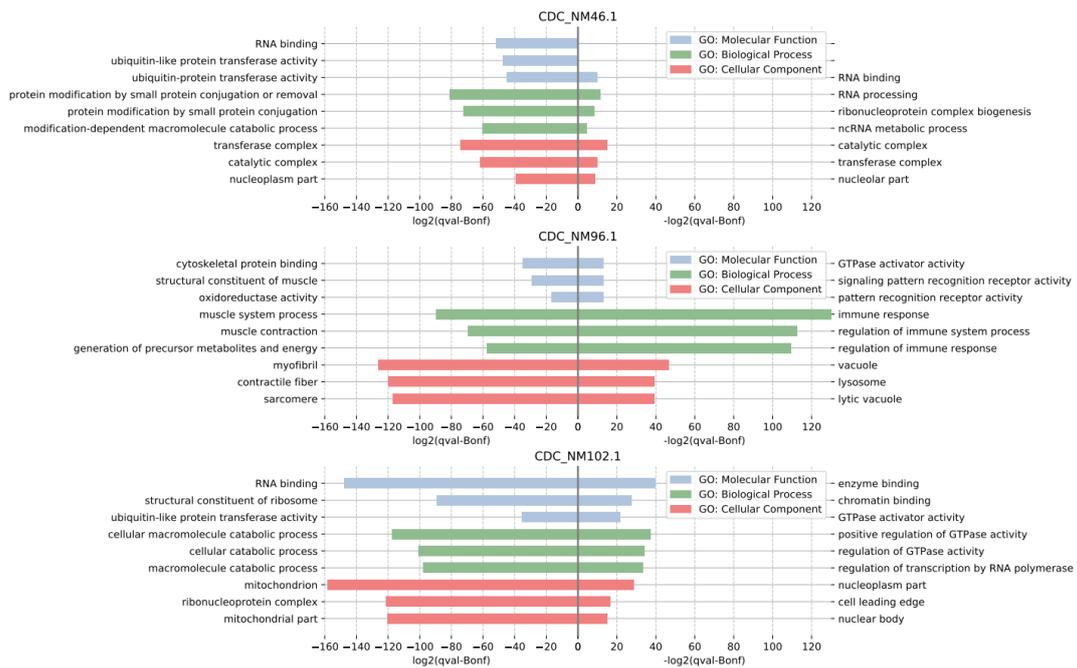


Figure 24: Gene ontology analysis results for CDC\_NM46.1, CDC\_NM96.1 and CDC\_NM102.1 differentially expressed genes

### 6.1.2. Aberrant splicing analysis

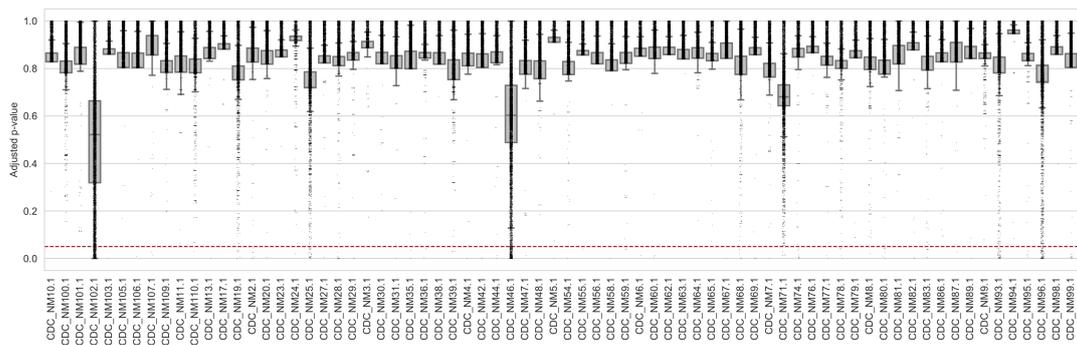


Figure 25: Distribution of p-values for each sample shows that more significant splicing events were detected in CDC\_NM102.1, CDC\_NM46.1 and CDC\_NM96.1. While most other samples show very few significant splicing clusters, these three samples show an even distribution of splicing junctions across all p-values

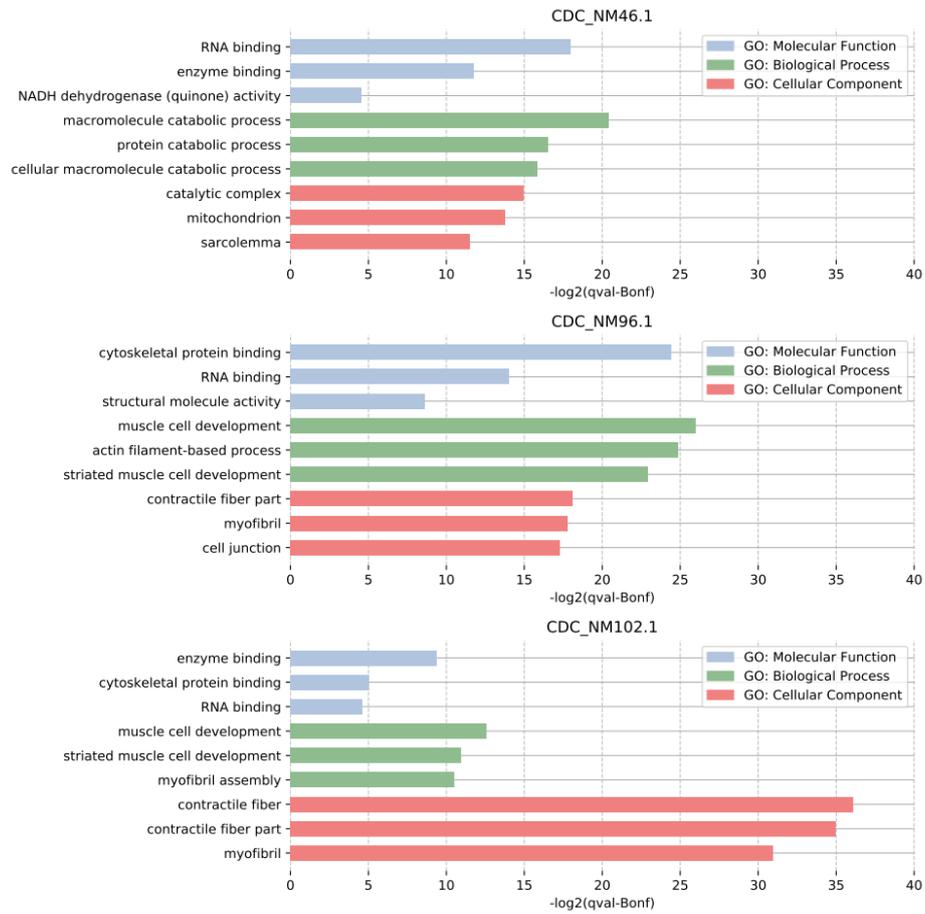


Figure 26: Gene ontology analysis for samples with more than twenty significant splicing events. All three samples showed enrichment for RNA-binding. CDC\_NM46.1 further showed association with catabolic processes and mitochondria, while CDC\_NM96.1 showed an enrichment for structural muscle cell development and cell function. CDC\_NM102.1 shared several associations with CDC\_NM96.1: muscle cell development, striated muscle cell development, contractile fiber and myofibril.

### 6.1.3. Non-negative matrix factorization clustering

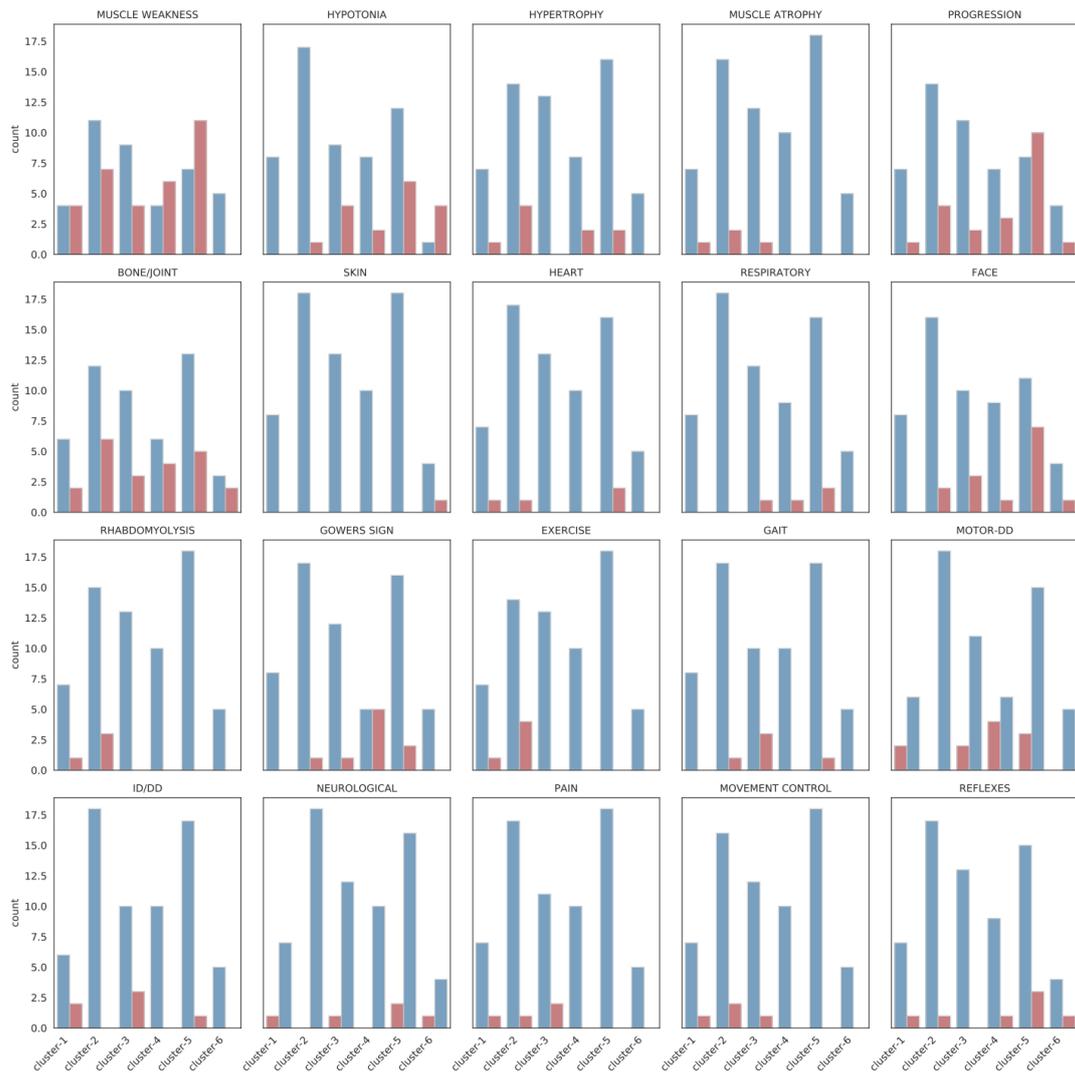


Figure 27: Number of samples with a specific symptom (red) and without the symptom (blue) for each cluster shows no specific accumulation of samples with the same symptom into the same cluster

## 6.2. Supplementary Tables

### 6.2.1. Whole exome sequencing variants

Sample	Gene	GT	Variant	GnomAD-AF	Conser- vation	OMIM	Note
CDC_NM1.1	<i>TACO1</i>	het	chr17:61678654:c.G212C:p.G71A	0.0000040 49	1	Mitochondrial complex IV deficiency (AR, Mi)	second variant
CDC_NM2.1	<i>PIGN</i>	het	chr18:59757713:c.A2279G:p.Q760R	0.0000109 3	14	Multiple congenital anomalies-hypotonia-seizures syndrome 1 (AR)	
CDC_NM5.1	<i>CANT1</i>	hom	chr17:76993221:c.G484T:p.G162W	0	0	Desbuquois dysplasia 1 (AR) Epiphyseal dysplasia, multiple, 7 (AR)	
CDC_NM6.1	<i>GUF1</i>	cmp het	chr4:44684357:c.C514G:p.Q172E chr4:44691907:c.T1250A:p.L417Q	0 0.00027	2 0	Epileptic encephalopathy, early infantile, 40 (AR)	
CDC_NM7.1	<i>COL6A2</i>	cmp het	chr21:47538526:c.A1117-2G chr21:47552175:c.2776_2784dupAATG CCATC:p.N926_1928dup	0 0.0000211 7	0 24	Myosclerosis, congenital (AR) Bethlem myopathy 1 (AR, AD) Ulrich congenital muscular dystrophy (AR, AD)	
CDC_NM8.1	<i>SYTL2</i>	hom	chr11:85431930:c.G1535A:p.S512N	0	2		novel
CDC_NM9.1	<i>NEB</i>	het	chr2:152396932:c.A20818T:p.K6940*	0	0	Nemaline myopathy 2 (AR)	second variant
CDC_NM11.1	<i>SPG11</i>	cmp het	chr15:44914955:c.A2287G:p.T763A chr15:44864940:c.T6284C:p.L2095S	0.0000039 82 0.0006929	1 2	Amyotrophic lateral sclerosis 5, juvenile (AR) Charcot-Marie-Tooth disease, axonal, type 2X (AR) Spastic paraplegia 11 (AR)	
CDC_NM16.1	<i>BICD2</i>	cmp het	chr9:95480972:c.C1955T:p.S652L chr9:95480997:c.G1930GT:p.A644S	0 0	1 0	Spinal muscular atrophy, lower extremity-predominant, 2A (AD) Spinal muscular atrophy, lower extremity-presominant , 2B (AD)	
CDC_NM19.1	<i>ADSSL1</i>	cmp het	chr14:105207568:c.G910A:p.D304N chr14:105208309:c.1048delA:p.I350fs	0.0000340 4 0.0000814 9	0 0	Myopathy, distal, 5 (AR)	
CDC_NM20.1	<i>TPM2</i>	het	chr9:35689262:c.G121A:p.E41K	0	1	Arthrogryposis multiplex congenita, distal, type 1 (AD) Arthrogryposis, distal, type 2B (AD) CAP myopathy 2 (AD) Nemaline myopathy 4 (AD)	
CDC_NM22.1	<i>ADSSL1</i>	cmp het	chr14:105207568:c.G910A:p.D304N chr14:105208309:c.1048delA:p.I350fs	0.00003 0.00008	0 0	Myopathy, distal, 5 (AR)	
CDC_NM23.1	<i>PDLIM3</i>	hom	chr4:186456560:c.C29T:p.P10L	0.0005	1		
CDC_NM26.1	<i>RYR1</i>	cmp het	chr19:39075661:c.G14725A:p.E4909K chr19:39075603:c.C14667G:p.Y4889*	0.000008 0	13 0	Central core disease (AR, AD) King-Denborough syndrome (AD) Minicore myopathy with external ophthalmoplegia (AR) Neuromuscular disease, congenital, with uniform type 1 fiber (AR, AD) Malignant hyperthermia susceptibility 1 (AD) Myosclerosis, congenital (AR)	
CDC_NM28.1	<i>COL6A2</i>	het	chr21:47535950:c.G883A:p.G295R	0	1	Bethlem myopathy 1 (AR, AD) Ulrich congenital muscular dystrophy (AR, AD)	

Sample	Gene	GT	Variant	GnomAD-AF	Conser- vation	OMIM	Note
CDC_NM30.1	<i>COL6A2</i>	het	chr21:47542795:c.C1615T:p.R539*	0.00001	1	Myosclerosis, congenital (AR) Bethlem myopathy 1 (AR, AD) Ulrich congenital muscular dystrophy (AR, AD)	
CDC_NM36.1	<i>PURA</i>	cmp het	chr5:139493769:c.G3T:p.M1? chr5:139493861:c.G95T:p.G32V	0 0.000009	18 5	Mental retardation, autosomal dominant 31 (AD)	
CDC_NM37.1	<i>HADHB</i>	cmp het	chr2:26496604:c.A340G:p.N114D chr2:26496604:c.G64+5A	0.00001 0	0	Trifunctional protein deficiency (AR) Epileptic encephalopathy, early infantile, 42 (AD) Episodic ataxia, type 2 (AD)	
CDC_NM38.1	<i>CACNA1A</i>	het	chr19:13373585:c.G4064A:p.R1355Q	0	0	Migraine, familial hemiplegic, 1 (AD) Migraine, familial hemiplegic, 1, with progressive cerebellar ataxia (AD) Spinocerebellar ataxia 6 (AD)	
CDC_NM39.1	<i>COL6A1</i>	het	chr21:47409043:c.G850A:p.G284R	0	0	Bethlem myopathy 1 (AR, AD) Ulrich congenital muscular dystrophy 1 (AR, AD)	
CDC_NM40.1	<i>RRMB2</i>	het	chr8:103244471:c.A326G:p.K109R		20		novel
CDC_NM45.1	<i>GNE</i>	cmp het	chr9:36219937:c.G1807C:p.V603L chr9:36249315:c.G131C:p.C44S	0.00002 0	0 0	Nonaka myopathy (AR)	
CDC_NM47.1	<i>KLHL40</i>	hom	chr3:42730521:c.G1582A:p.E528K	0.00006	0	Nemaline myopathy 8 (AR) Central core disease (AR, AD) King-Denborough syndrome (AD) Minicore myopathy with external ophthalmoplegia (AR)	
CDC_NM50.1	<i>RYR1</i>	cmp het	chr19:39008165:c.G9852C:p.W3284C chr19:38945873:c.A1441-2G	0 0	0 0	Neuromuscular disease, congenital, with uniform type 1 fiber (AR, AD) Malignant hyperthermia susceptibility 1 (AD)	
CDC_NM55.1	<i>AGL</i>	cmp het	chr1:100327823:c.309_311delTGG:p.G104del chr1:100379160:c.G4027A:p.E1343K	0 0.0005	17 9	Glycogen storage disease IIIa (AR) Glycogen storage disease IIIb (AR) Cardiomyopathy, dilated, 1S (AD) Cardiomyopathy, hypertrophic, 1 (AD) Laing distal myopathy (AD)	
CDC_NM56.1	<i>MYH7</i>	het	chr14:23882989:c.5754_5768delCAAGCTGCGGGCCAA:p.N1918_A1922del	0	1	Left ventricular noncompaction 5 (AD) Myopathy, myosin storage (AR, AD) Scapulooperoneal syndrome, myopathic type (AD)	novel
CDC_NM57.1	<i>TRAPPC1</i>	het	chr17:7833983:c.378dupT:p.R127fs		14		
CDC_NM59.1	<i>COL6A2</i>	cmp het	chr21:47538952:c.1189_1196delGGCAACAG:p.G397fs chr21:47552249:c.C2843G:p.T948R	0 0.000004	25 1	Myosclerosis, congenital (AR) Bethlem myopathy 1 (AR, AD) Ulrich congenital muscular dystrophy (AR, AD)	
CDC_NM60.1	<i>MYO9A</i>	het	chr15:72191053:c.A3791C:p.Q1264	0.0006	3	Myasthenic syndrome, congenital, 24, presynaptic (AR) Epidermolysis bullosa simplex with nail dystrophy (AR) Epidermolysis bullosa simplex with muscular dystrophy (AR)	
CDC_NM63.1	<i>PLEC</i>	cmp het	chr8:144999068:c.G5440A:p.G1814S chr8:144998444:c.G6064A:p.A2022T	0.0002 0.004	40 12	Epidermolysis bullosa simplex with pyloric atresia (AR) Epidermolysis bullosa, Ogna type (AD) Muscular dystrophy, limb-girdle 17 (AR)	

Sample	Gene	GT	Variant	GnomAD-AF	Conser- vation	OMIM	Note
CDC_NM65.1	<i>ARRDC4</i>	cmp het	chr15:98504164:c.G73C:p.E25Q chr15:98512491:c.G764A:p.R255Q	0 0.00001	9 0		
CDC_NM66.1	<i>BAG3</i>	hom	chr10:121436725:c.A1659T:p.E553D	0.0002	10	Cardiomyopathy, dilated, 1HH (AD)	
CDC_NM68.1	<i>SYNE2</i>	cmp het	chr14:64519978:c.A9347G:p.K3116R chr14:64461863:c.T2883G:p.N961K	0.002 0.00003	2 19	Myopathy, myofibrillar, 6 (AD) Emery-Dreifuss muscular dystrophy 5 (AD)	
CDC_NM70.1	<i>ITGA3</i>	cmp het	chr17:48158692:c.C2839T:p.R947* chr17:48156216:c.G2326T:p.G776W	0 0.0007	6 19	Interstitial lung disease, nephrotic syndrome, and epidermolysis bullosa, congenital (AR)	
CDC_NM71.1	<i>CHRND</i>	cmp het	chr2:233391303:c.C117G:p.N39K chr2:233396133:c.C892T:p.R298C	0.006 0.00003	19 8	Myasthenic syndrome, congenital, 3A, slow-channel (AD)	
CDC_NM72.1	<i>FAT1</i>	het	chr4:187557999:c.G3712A:p.E1238K	0	1	Myasthenic syndrome, congenital, 3C, associated with acetylcholine receptor deficiency (AR)	
CDC_NM76.1	<i>NEB</i>	cmp het	chr2:152423951:c.G17887A:p.V5963I chr2:152543980:c.G2590A:p.D864N	0.002 0.00002	18 0	Multiple pterygium syndrome, lethal type (AR)	
CDC_NM77.1	<i>SMCHD1</i>	cmp het	chr18:2778215:c.C5525T:p.A1842V chr18:2705733:c.A1884G:p.I628M	0.000004 0	3 7	Myasthenic syndrome, congenital, 3B, fast-channel (AR)	
CDC_NM78.1	<i>BRD4</i>	cmp het	chr19:15350584:c.3328_3330delGTG:p.V1110del chr19:15376322:c.A692C:p.D231A	0.0002 0.00002	0 3	Nemaline myopathy 2 (AR)	
CDC_NM80.1	<i>DNM2</i>	het	Chr19:1090405:c.G1102A:p.E368K	0	0	Bosma arhinia microphthalmia syndrome (AD)	
CDC_NM83.1	<i>RYR1</i>	cmp het	chr19:39003064:c.C9413T:p.P3138L chr19:38959747:c.G3523A:p.E1175K	0.00002 0.00002	0 1	Fascioscapulohumeral muscular dystrophy 2, digenic	
CDC_NM86.1	<i>GFPT1</i>	cmp het	chr2:69569331:c.C1156T:p.R386C chr2:69565658:c.A1243C:p.M415L	0 0	0 0	Centronuclear myopathy 1 (AD)	
CDC_NM87.1	<i>TTN</i>	cmh et	chr2:179542427:c.T34212G:p.Y11404* chr2:179395588:c.C105754T:p.R35252*	0 0.000008	50 50	Charcot-Marie-Tooth disease, axonal type 2M (AD)	
CDC_NM88.1	<i>DMD</i>	het	chrX:31792192:c.A7427G:p.N2476S	0.00009	1	Charcot-Marie-Tooth disease, dominant intermediate B (AD)	
CDC_NM89.1	<i>CAPN3</i>	het	chr15:42676684:c.316dupT:p.C106fs	0	0	Lethal congenital contracture syndrome 5 (AR)	

Sample	Gene	GT	Variant	GnomAD- AF	Conser- vation	OMIM	Note
CDC_NM102.1	<i>LMNA</i>	hom	chr1:156105884:c.1129C>T:p.R377C	0	9	Muscular dystrophy, congenital (AD) Central core disease (AR, AD) King-Denborough syndrome (AD)	
CDC_NM108.1	<i>RYR1</i>	cm ph et	chr19:39013724:c.10316G>A:p.G3439D chr19:39071088:c.14595_14597delC AA:p.N4865del	0.000004 0	20 0	Minicore myopathy with external ophthalmoplegia (AR) Neuromuscular disease, congenital, with uniform type 1 fiber (AR, AD) Malignant hyperthermia susceptibility 1 (AD)	

*Table 8: Whole exome sequencing variants per sample*

## 7. 국문초록

# 희귀 신경근 질환의 유전체, 전사체 통합 분석 연구

자나  
의과학과 의과학전공  
서울대학교대학원

Whole exome sequencing (WES)은 비용 및 데이터 처리의 용이성으로 인하여 희귀질환 진단등에 매우 효과적인 방법이 되었다. 그러나 variant of unknown significances (VUS)를 해석하는 어려움과 non-coding 변이형을 확인할 수 없다는 점 등의 이유로 WES 기반의 희귀질환 진단률은 대부분 50%를 넘지 못한다. 따라서, 본 연구에서는 희귀질환 진단의 보완적인 접근법으로 새로이 전사체 분석법을 도입할 것을 제시하고자 한다. 이를 위하여 서울대학교 어린이병원 소아신경과에서 임상적으로 진단되지 못한 근신경질환 환자 94명을 대상으로 WES 분석을 실시하고, 이미 알려진 근신경질환의 원인 유전자 변이들을 분석하였다. 추가적으로, 기존에 WES 분석이 수행된 63명의 환자군과 이 외의 10명의 환자군을 추가하여 전사체 분석을 수행하였다. 전사체 데이터를 이용하여 damaging 변이 분석, allele-specific expression 분석, 환자군과 정상군에서 다르게 발현하는 유전자 (DEG) 및 비정상적인 splicing 양상을 탐색하는 분석을 수행하였다. 또한, non-negative matrix factorization 분석 기법을 통해 유전자 발현 프로파일을 기반으로 한 군집화를 수행하고, 각 군집을 특징 짓는 유전자 그룹을 도출하였다. 그 결과, WES 분석을 통하여 49%의 환자에서 후보 원인 변이를 확인하였으며, 그 중 83%의 환자에서는 알려진 근신경질환 원인 유전자의 변이를 확인하였다. 12명의 환자에서는 그 가능성이 확실하지 않은 구조 변이를 확인하였다. 전사체 데이터

기반의 변이 분석을 통하여, WES 을 수행하지 않은 5 명의 환자를 포함한 총 9 명의 환자에서 heterozygous 변이를 추가로 발견하였다. Allele-specific expression 분석을 통하여 2 개의 후보 원인유전자를 발견하였고, DEG 분석 결과, 4 명의 환자에서 잠재적인 원인 유전자 그룹을 선별할 수 있었다. 또한, 4 명의 환자에서 *DMD*, *TTN*, *MICU1* 유전자들의 비정상적인 splicing 이 확인되었다. non-negative matrix factorization 기반 군집화 분석 결과, 유전자 발현 양상을 기반으로 한 6 개의 군집을 확인할 수 있었다. 본 연구를 통하여 전사체 분석법이 기존의 WES 기법 기반 분석의 효과적인 보완 기법이 될지의 여부를 확인하고자 하였다. 전사체 분석 결과, WES 기법을 통해 원인 유전자 변이가 확인된 환자들 중 9 명에게서 같은 맥락의 전사체 이상을 확인할 수 있었으며, WES 을 수행하지 않은 환자들 중 18 명에게서도 잠재적인 원인 유전자 변이를 확인하였다. 따라서 전사체 분석법은 기존의 분석기법으로 원인 유전자 변이를 발견할 수 없는 증례의 진단에 유용한 도구로 사용될 수 있음을 시사한다.

---

**주요어:** 유전체, 전사체, 신경근 질환, 멀티오믹스, 희귀질환 진단, 멘델유전질환, 변이 발굴, 전사체-기반 군집화

**학 번:** 2017-21922

# Acknowledgements

I would like to express my deepest appreciation to all those who have assisted me in pursuing my Masters degree at SNU. For providing me with the opportunity and the incitement to learn and grow and for creating an atmosphere in which scientific discoveries can thrive, I would like to offer my special thanks to my supervisor Dr. Murim Choi. Without his advice and guidance, this work would not have been possible. Furthermore, I wish to express my gratitude to my thesis committee, who have offered me constructive suggestions to improve my research in content and structure and to my collaborators at SNU Children's hospital, who assisted with the collection and interpretation of the data presented in this work.

Moreover, I would like to thank all the members of the SNU Functional Genomics laboratory for being incredibly supportive and always offering a helping hand. Thanks to them, my graduate student life was full of joy and laughter.

Lastly, I would like to thank Hooram and his family for providing me with a home away from home and my loving mother, grandparents and brother for their heartfelt encouragement and emotional support throughout my life.