CLINICAL UTILITY OF TARGETED RNA-SEQ IN NEUROMUSCULAR AND IMMUNE DISORDERS

A Dissertation
Presented to
The Academic Faculty

by

Kiera R Berger

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
May 2022

CLINICAL UTILITY OF TARGETED RNA-SEQ IN NEUROMUSCULAR AND IMMUNE DISORDERS

Approved by:

Dr. Gregory Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Patrick McGrath
School of Biological Sciences
Georgia Institute of Technology

Dr. Madhuri Hegde
Global Lab Services
PerkinElmer, Inc

Dr. Subra Kugathasan
Departments of Pediatrics and Human Genetics
Emory University School of Medicine

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Date Approved: April 19, 2022

ACKNOWLEDGEMENTS

I would like to extend my immense gratitude and appreciation for my advisor, Dr. Greg Gibson. Greg has been both patient and pushy, each at exactly the right moment. Throughout my time here, he has been a wonderful mentor and advocate for me. Greg taught me to be a more confident scientist and person and I would not have been able to successfully complete this program without his guidance and support.

I am also grateful for the guidance and insights I received from my committee members: Dr. I. King Jordan, Dr. Madhuri Hegde, Dr. Patrick McGrath, and Dr. Subra Kugathasan. I would particularly like to thank Dr. Madhuri Hegde, whose excitement and confidence in my work helped make me more self-assured in my conclusions.

I also thank Lisa Redding for all her assistance on the administrative side of the program. She is an exceptionally kind and encouraging person and has always promptly answered all of my questions, which significantly reduced my anxiety over the past six years.

I am extremely appreciative of the friends I made at the start of this journey, Mary Beth McWhirt and Emily Norris. You each have provided wonderful but completely different friendships, and I needed both of you to maintain my spirits throughout my time here. Along the same lines, I must thank my lifelong friend Stephanie Hubbard, who always finds the time to catch up despite the physical distance between us.

The love and support of my family has been critical to achieving this goal. My parents, Mary and Don Berger, instilled in me a thirst for knowledge and a belief I can do anything and have provided me with support and guidance beyond what any child would

reasonably expect. My brother, Andrew, without whom I likely would not have found my way to Georgia Tech, and Willi Rechler, the best sister-in-law I could possibly hope for, kept me feeling close to home and I always know the two of you are there for me. Finally, Riley Flanagan has been a constant source of encouragement and reminders to take a break and go outside, and I could not have done this without him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ACMG	American College of Medical Genetics
AEI	Allele Expression Imbalance
ASE	Allele Specific Expression
B	Benign
CD	Crohn's Disease
CNV	Copy Number Variation
IBD	Inflammatory Bowel Disease
IGV	Integrative Genomics Viewer
LB	Likely Benign
LGMD	Limb Girdle Muscular Dystrophy
LoF	Loss of Function
LP	Likely Pathogenic
NGS	Next Generation Sequencing
NMD	Neuromuscular Disorders
nmd	nonsense mediated decay
OMIM	Online Mendelian Inheritance in Man
P	Pathogenic
PBMC	Peripheral Blood Mononuclear Cells
PCA	Principal Component Analysis
PID	Primary Immune Deficiencies
PSI	Percent *of exons* Spliced In
PTV	Protein Truncating Variant
RIN	RNA Integrity Number

SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
UC	Ulcerative Colitis
VEOIBD	Very Early Onset Inflammatory Bowel Disease
VUS	Variant of Uncertain Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

SUMMARY

This PhD thesis explores the analysis of RNA-seq data and the value it can provide to the diagnosis of rare genetic disorders. While individually rare, these incurable disorders collectively affect as many as 1 in 10 Americans and are mostly without effective treatment. Recent developments in personalized medicine and targeted gene-therapies indicate the dawn of a new era for rare disease treatment is on the horizon, but the advancement of therapeutics is dependent on increasing our knowledge of the causative molecular mechanisms underlying each specific disease. Despite the meteoric rise of whole genome and whole exome sequencing in the past decade, diagnostic yield for the majority of rare disorders hovers around 35%. RNA-seq is a less commonly used technology, but holds the potential to show us the intermediary step between DNA and protein and the direct effects variants may have on it. My research focuses on developing a multi-faceted approach to analysis of RNA for rare disease diagnostics, using targeted gene panels to achieve higher read depth and the correlation of multiple data types to maximize identification of pathogenic events.

Chapter 2: In this case study of an individual affected by neuromuscular disease, we explored the use of targeted RNA-seq for the affected tissue. Based on the unusual clinical presentation, the patient and both parents underwent whole exome sequencing. A single heterozygous missense variant was found in *GNE*, which had been previously characterized as pathogenic. However, *GNE*-myopathies are inherited in an autosomal recessive manner, and no second variant was identified. RNA-seq of the muscle showed mono-allelic expression of the pathogenic allele, and subsequent aCGH identified an upstream deletion of 7.08kb. *GNE* transcript abundance in RNA indicated a 50% reduction

in the patient, which along with the monoallelic expression of the pathogenic variant confirmed that the promoter region deletion leads to loss of expression of the *GNE* transcript.

Chapter 3: My second study explored transcriptome-wide splicing patterns in the ileal and rectal tissue of patients with inflammatory bowel disease. Exon inclusion was calculated for each paired duplicate sample and analysis of principal components found splicing to be less variable than gene expression overall, though both showed clear separation by tissue type. Eight individuals exhibited aberrant splicing and gene expression profiles that are at least partially explained by altered ratios of specific cell types suggestive of active inflammation. The ileal samples of two individuals with Crohn's disease had significantly divergent splicing patterns that resembled the splicing profile of rectal tissue, indicating that differential splicing contributes to the pathology of inflammatory bowel disease.

Chapter 4: This extensive study of a sizeable cohort of patients clinically suspected of *Dysferlinopathy* exemplified the diagnostic power of RNA-seq for a single gene. Genomic sequencing of *DYSF* and a protein assay of *Dysferlin* were performed for the entire cohort and a subset of the full cohort underwent targeted RNA-seq of whole blood. Prior to RNA-seq, the diagnostic yield of the cohort was 33%. This study relied on a novel pipeline for the comprehensive analysis of RNA including splicing, allele specific expression, exon usage, and transcript abundance. Emphasis was placed on concordance of clinical and laboratory data for a complete diagnosis. RNA-seq identified exonic variants, aberrant splicing events, and allele specific expression. Even with only around a fifth of the full cohort undergoing RNA-seq, the total diagnostic yield was raised

to 44%. In addition, we found that allele specific expression due to nonsense-mediated decay can be used to phase variants.

Chapter 5: In this final study, we explored how the pipeline developed for RNA-seq analysis in neuromuscular disorders would fare in the diagnosis of disorders with a less clear-cut causative gene or inheritance pattern. RNA-seq of PBMCs in a small cohort of patients suspected of a primary immunodeficiency or very-early onset inflammatory bowel disease who had previously undergone whole exome sequencing was performed for a panel of genes associated with immune disease. Although only a few instances of aberrant splicing were found, results were frequently suggestive of di- or multi-genic inheritance and brought into question the traditional model of variant classification that is prone to dismiss potentially disease contributive variants that are only semi-rare.

CHAPTER 1. INTRODUCTION

In order to understand why RNA-sequencing is relevant to clinical diagnostics, it is first necessary to review the recent history of rare human congenital disease and the rise of clinical genetic testing.

1.1 Rare diseases

An overview of the rare disease information presented here is provided in Figure 1.1. There are over 7000 defined rare diseases, with 80% or more considered to be genetic in origin.[1,2] While each individual disease is exceedingly rare, collectively around 30 million Americans are estimated to be affected by a rare disease. Despite being a relatively small portion of the population, rare disease patients make up roughly 25% of pediatric hospitalizations and place a sizeable burden on the healthcare system with lifetime estimated costs often exceeding $5 million[3-6]. Addressing rare diseases is an important issue for healthcare. Over half of rare diseases are associated with a reduced lifespan and years of life lost due to rare disease is 2X that of diabetes and 4X that of infectious disease[7].[1] Because each of these conditions affects only a handful of individuals, there has historically been a lack of research and resources surrounding the diagnosis and treatment of rare disease. The Orphan Drug Act of 1983 aimed to improve this situation by incentivizing the development of therapeutics for rare diseases. While this did lead to the approval of hundreds of orphan drugs, to date just 3% (226) of all rare diseases have an approved FDA treatment[8]. That is not to say that all hope is lost,

[1] Almost certainly no longer accurate.

however. From 2015-2020 orphan drugs made up between 39 and 53% of all new FDA approvals, and 11% of registered clinical trials correspond to a rare disease[9].

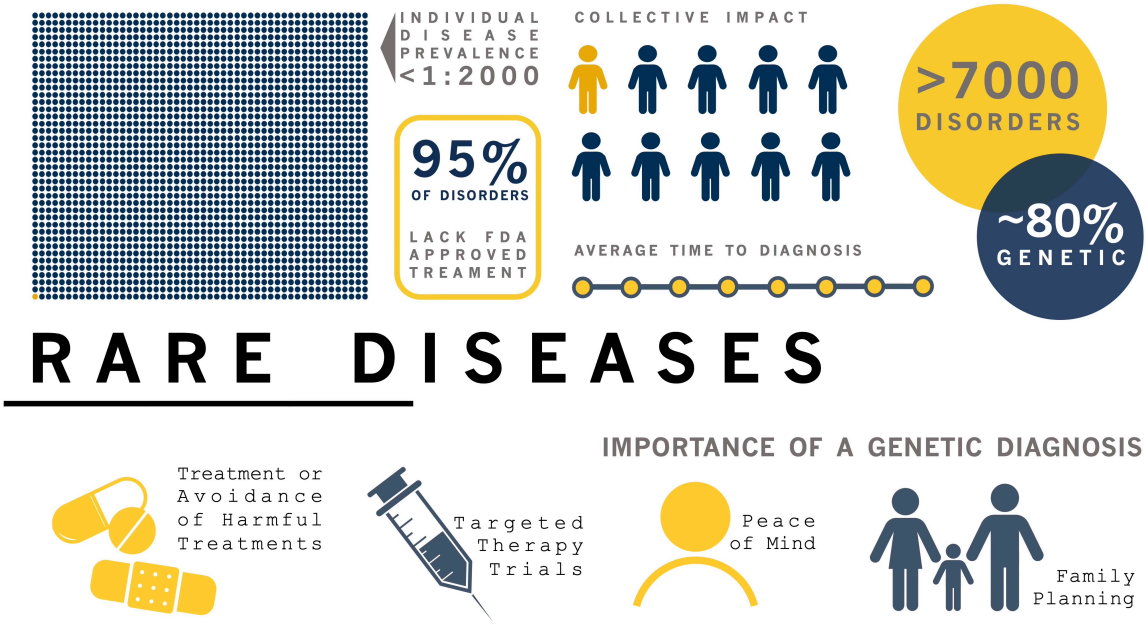

Figure 1.1 An overview of rare disease.

It is incredibly important that rare disease patients receive a diagnosis. Because of overlapping and heterogeneous phenotypes along with confounding factors like common disorders masking typical symptoms, receiving a diagnosis increasingly means identification of the underlying genetic cause, often referred to as a molecular or genetic diagnosis[10,11]. In addition, therapies available or in development commonly target specific affected regions or variants, making a genetic diagnosis essential for treatment or enrollment in clinical trials. Managers of clinical trials may need to know the specific diagnosis of people with a suspected rare disease before enrolling them, because the

therapies are designed to be effective only for specific patient subsets.[12] Such clinical trials will also benefit from better information relating to the causative variants for a specific disorder so treatments can be targeted to the mutation, as for example with *CFTR*-genotype-guided therapy for cystic fibrosis.[13] The Jain Foundation has funded a clinical trial for treatment of a subtype of neuromuscular disease, dysferlinopathy, where enrolment depends on a molecular diagnosis of a pathogenic *DYSF* (HGNC ID: 3097) genotype. One aim of this thesis is to enhance such diagnosis.

Even when there is no existing drug therapy, a diagnosis can allow patients to receive symptom-alleviating treatment[14]. Studies indicate that between 20-50% of genetic diagnoses lead to a recommendation for change in disease management or treatment[15,16]. It remains true, however, that a majority of patients with a rare disease have no treatment options available or hope for a treatment option in their lifetime. It has been speculated that a diagnosis, for these patients, would be detrimental to their mental health. Recent studies have conclusively shown that in nearly all cases this is false[17], and furthermore there are tangible benefits to receiving a diagnosis[18], beyond the psychological (Figure 1.1). A conclusive diagnosis is an end to the diagnostic odyssey that takes, on average, 8 years, as well as an end to the constant laboratory tests (some of which, like biopsies, are invasive and painful)[19]. A diagnosis also prevents unnecessary care like unhelpful or even detrimental therapies[20]. Despite over half of rare diseases presenting at birth or in childhood[11], undiagnosed adults experience more difficulty navigating the healthcare system[19], making early diagnosis even more important. In some countries, patients are unable to access social and ancillary services without a diagnosis[14], increasing their burden, both financial and otherwise. Then there are the social and psychological benefits like peace of mind, family and life planning, and

availability of specific support groups. Finally, even if all the previously mentioned benefits of a definitive diagnosis do not apply, rare disease patients report post-diagnosis that they are finally taken seriously by health care providers, even when the providers have no knowledge about their specific condition[18]. This is particularly common amongst women and people of color[21,22], whose unaffected counterparts are already known to experience an uphill battle in receiving adequate and necessary care in the face of medical disavowal.[2] The validation provided by a definitive diagnosis leads to empowerment, self-confidence, and improved health-related quality of life[23].

1.2 Rare disease genetics

Research has come a long way in the diagnosis of rare disorders of genetic origin. Since the shotgun sequencing of the draft human genome two decades ago, sequencing technologies have advanced almost exponentially. As costs have decreased at nearly the same rate, early diagnosis through genetic sequencing has become more popular.[24,25] Here, I will discuss the types of rare disease genetic testing, the standards for sequence variant interpretation, and the major challenges and debates present in the field.

1.2.1 Genomic sequencing

There are three main test types ordered in the clinical laboratory setting for the diagnosis of rare disease: targeted gene panels, whole exome sequencing (WES), and whole genome sequencing (WGS). As sequencing costs decreased, single gene tests were largely replaced with disease or tissue specific panels of genes. These panels

[2] For rare diseases involving fatigue and weakness, women are often told their symptoms are psychosomatic or related to menopause or their menstrual cycle, which harkens back to the good old days of female hysteria.

generally depend on a backbone of gene-disease association research to gather groups of genes most likely to contain the underlying cause for a particular range of phenotypes. This remains the most common test type ordered by clinicians, despite the push towards WES and WGS in the research community[11]. Targeted panels are inexpensive and offer a high read depth for increased confidence. However, this sequencing method covers only exonic regions (+/- about 5bp), meaning that intronic variants will not be detected. The biggest downside to targeted panels is that they will not be able to identify the causal variant if it is located in a gene that is not a part of the sequenced panel. WES improves upon this aspect of targeted sequencing in that it covers all known protein coding genes. Like targeted panels, however, it covers only exonic regions. Just as its name suggests, WGS covers the entire genome including exons, introns, and intergenic sequences. The pros and cons surrounding the different test options are discussed further later in this introduction, in the context of diagnostic yields and variant interpretation.

1.2.2 Variant interpretation

The early days of identifying a causative variant primarily involved linkage analysis and required animal models to prove pathogenicity. In the past 20 years or so, the technology to sequence the human genome has improved far faster than our understanding of it. The research community is still recovering from the errors that arose from faulty assumptions in early rare disease research. Multiplex pedigrees, at one time the standard for diagnostics, overestimated disease penetrance[15]. As a result, the penetrance and prevalence of most rare diseases are considered uncertain[8]. Next Generation Sequencing (NGS) allowed for causal inference of variants as well as in-silico predictions of pathogenicity. Large population sequence databases of presumed healthy individuals like ExAC[26] and gnomAD[27] provided more accurate metrics for variant

frequency. Variants that had been defined as pathogenic due to their presence in families with rare disease were found to be more common than previously thought[28]. In the first decade of the millennium, new companies cropped up to bring NGS technology to clinical diagnostics. By the early teens, it was apparent to the rare disease diagnostic community that the lack of consensus surrounding variant pathogenicity had become a problem both for researchers and the patients they aimed to help[29,30], with one study finding that up to a third of pathogenic variants in the Human Gene Mutation Database (HGMD) were in fact likely benign[31]. In 2015, the American College of Medical Genetics (ACMG) released a set of guidelines for variant interpretation.[32] These were produced by a large collaborative process to establish standards by which to determine the significance of a variant and are the current foundation for clinical diagnostics. These standards have widely been adopted by both researchers and clinical genetic testing laboratories, but they are often not applied consistently due to the complexity of the evidence and the fact that they were written in broad terms to apply to a number of types of genetic testing.[33] Variants are classified as pathogenic (P), likely pathogenic (LP), of unknown significance (VUS), likely benign (LB), or benign (B) using a combination of criteria from population data, computational data, functional data, and segregation patterns.[32] The biggest obstacles to implementation of these criteria include inconsistent standards of evidence, inability to phase pairs of alleles, and failure to incorporate emerging lines of evidence.

The presence of clinical testing laboratories distinct from rare disease research laboratories means that many rare variants have been seen in patients but are not published in the literature. Because patients with a specific rare disease are few and far between and the variants are, by definition, not often found in the general population, it is important to share this information across the diagnostic community. The public database

ClinVar was created so that researchers and laboratories could submit variants alongside the evidence for their interpretation[34].

As of June 2021, OMIM[35] lists 6,091 human phenotypes whose molecular basis is unknown, while ClinVar[36] lists almost 1.5 million variant records claiming some evidence for causation of a condition. Whole genome sequencing (WGS) is dramatically expanding clinical diagnosis, but there are multiple challenges to be overcome[10,37].

The process of identifying genetic causality in rare disease is incredibly difficult. There are 3 billion base pairs in the human genome, of which just 1% code for proteins (referred to as the exome)[38]. Changes to the coding sequence, particularly changes that lead to premature truncation of the protein like nonsense variants and frameshift deletions or insertions, are the easiest to link to rare disease[29]. We know now that the regions between exons and between genes are important for regulatory purposes[39], but there is still a dearth of knowledge about the exact mechanics. Furthermore, it is challenging and often not possible to determine the effect of variants that occur in these regions[40].

1.2.3 Diagnostic yield and the challenges to interpretation

Our ability to interpret sequence data was long ago eclipsed by our ability to generate it. The use of trios in WES and WGS, where a proband and two closely related relatives (ideally parents) are sequenced, helps sort through the large number of potential rare variants by allowing for phasing and identification of *de novo* variants. Although diagnostic yields for some disorders, like cystic fibrosis (CF; OMIM #219700), have reached upwards of 90% with the use of DNA sequencing[41], most others hover around 30%.[42-44]

The introduction of Whole Exome Sequencing (WES) allowed for gene/disease association discoveries to accelerate. WES is now routinely used in research settings for gene discovery and variant identification for rare diseases. While Whole Genome Sequencing provides more information than WES, WES is often chosen for its lower cost both in sequencing and in data storage/interpretation (though neither will be a meaningful barrier a decade from now). There is a push for WES/WGS to be used as a primary tool for clinical rare disease diagnostics, evidenced by the test offerings at clinical genetic testing laboratories and research initiatives focused on newborn screenings and database building using the technologies. While this work is undoubtedly incredibly important to the implementation of precision medicine strategies, it is unfortunately not very helpful to patients currently suffering from an undiagnosed rare disease. Previous studies have indicated that for many diseases, WES/WGS offer an improvement in diagnostic yield relative to candidate gene screening of only 5-15%.[42-44] There are a number of underlying reasons for this, but chief among them is the burden of variants of uncertain significance (VUS). Another barrier to WES/WGS becoming a common tool in clinical diagnostics is cost. Though it can be argued that WGS/WES as the initial diagnostic test saves money in the long run,[12] it is more expensive than individual single-gene or multi-gene panel tests and a recent survey indicated that most insurance is far less likely to cover WES or WGS.[45-48]

A belief commonly held by researchers focused on clinical diagnostics is that WGS is the future – in the future every baby will be sequenced immediately, and this will capture issues early on. WGS is future-proof, in the sense that you can always reanalyze the data after strides are made in variant interpretation and/or identification of new genes associated with a disease. Studies focus on building huge cohorts of WGS sequences,

believing that if we sequence enough we will identify the causative variants mostly by statistical enrichment.[10,24,25,49,50] I would like to challenge this notion. Because of variability in penetrance[1], rarity of disease, and the incredibly high burden of VUS, WGS is simply not a sufficient option. By all means we should continue to grow these databases, as they undeniably have improved diagnostics and will likely eventually serve the desired purpose. But they alone do not help the majority of people with a rare disease and won't within their lifetimes. The critical need is for VUS to be evaluated more systematically. Although it is time-consuming and expensive to conduct individual experiments that definitively prove a VUS is a causative pathogenic variant, the effort pays off in diagnostic accuracy and hence patient health.

Against this background, I note that RNA seq is cheaper, identifies more clinically relevant variants and importantly, directly observes the effects of certain variants. It can resolve some of the VUS and also identify variants not seen in DNA sequencing. It also adds to the knowledge base – once a variant is seen to be pathogenic in RNA it can be used to go back and reanalyze WGS.

1.3 Splicing

Alternative splicing is a normal and necessary form of isoform and gene expression regulation, though the mechanisms by which it happens are only partially understood[51]. Types of alternative splicing events include exon skipping, alternative donor/acceptor sites (exon extension, partial exon exclusion), retained introns, and mutually exclusive exon usage (Figure 1.2). Exon skipping is the most prevalent type of alternative splicing in humans.[52] A splicing event is aberrant if it does not occur naturally in a known, annotated transcript or in the population.

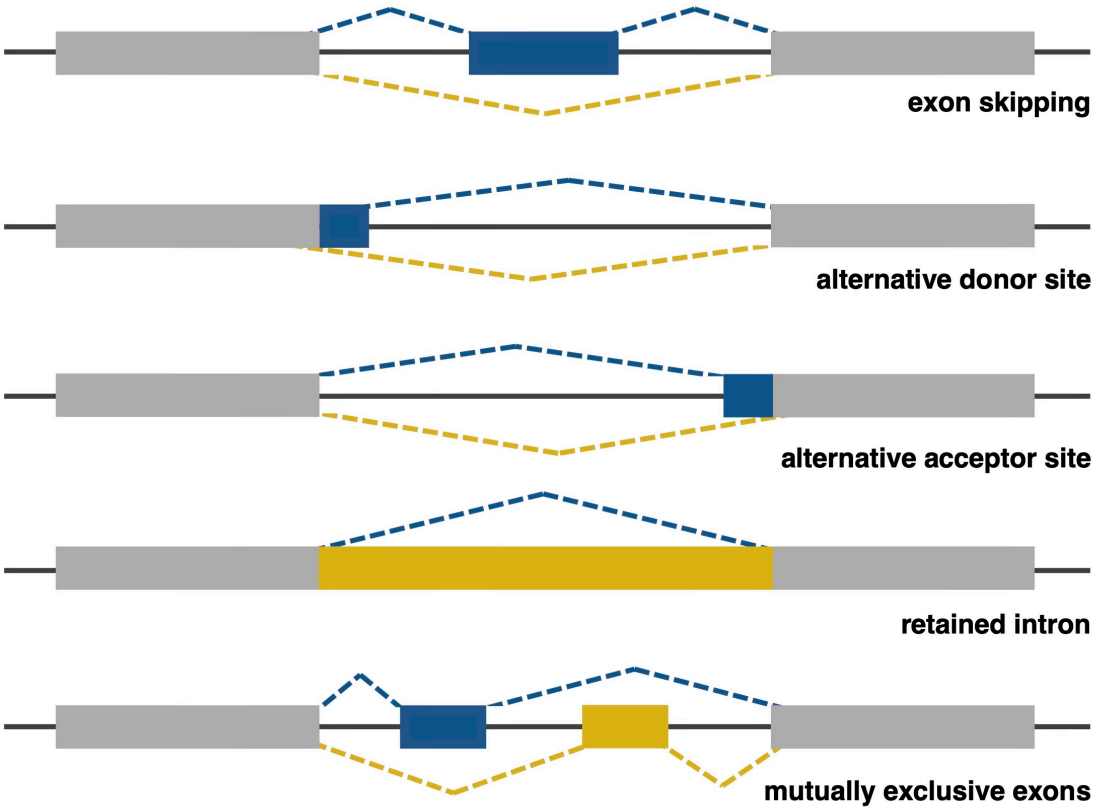

Figure 1.2 Types of alternative splicing.

Apart from variants that disrupt the canonical splice sequence (the two base pairs adjacent to the beginning and end of every exon), variant effects on splicing have largely been ignored. We cannot even say for certain what percentage of variants in the human genome affect splicing. Estimates range from as little as 9% to more than 50%[51,53-55]. Splice variants make up around 15% of reported variants in ClinVar, and 75% of those disrupt the canonical splice site[56-58]. It is clear that variants affecting splicing are understudied and as a result are more likely to be missed when testing for rare disease. While in-silico tools for predicting the effect of missense variants on protein function are

quite accurate and well-tested, similar tools to predict splice variants are mostly out of date and unreliable given our historic lack of understanding of non-exonic sequences and regulatory mechanisms[51,53]. More recently, the Transcript-inferred Pathogenicity score (TraP)[59] and the machine-learning based SpliceAI[53] have shown marked improvement in predicting whether a variant will impact splicing. It is questionable, though, whether these tools have been incorporated into analysis pipelines at clinical testing laboratories. The majority of these labs use the Alamut software, which uses only the older, less accurate splice predictors[51]. Moreover, it is important to note that the predictors, even when they accurately identify variants that affect splicing, are unable to determine the penetrance of the alternative splicing and in many cases cannot even say what the new splicing coordinates will be. Determining the precise effects of specific splice variants is essential to advancing both our knowledge of splicing mechanics and our ability to accurately interpret the pathogenicity of variants. In one example, a *BRCA1* splicing variant long classified as pathogenic was found to only partially disrupt splicing in the affected individuals, potentially leaving enough wild-type transcript to retain normal protein function[60,61]. There are currently three methods commonly used to show the effect of splicing variants. Minigene constructs are beneficial when the patient sample is not available, as any variant can be inserted and splicing effects observed. However, these constructs focus on just the small affected region and thus lack the context to view the larger effect on the gene, not to mention the inability to observe how penetrant the event is in the patient's tissue. RT-PCR methods directly observe the effect of the variant using the patient's tissue, but require developing specific primers for each individual variant that is suspected of impacting splicing. More recently, RNA-sequencing has been used to directly observe aberrant splicing in the affected patient without any prior knowledge of their genetic variation[43,62,63].

1.4 Approaches to RNA-based diagnostics

RNA sequencing provides a way to potentially classify some VUS, but also allows for identification of functional defects that may not have been seen using DNA alone, such as aberrant splicing.

There are a number of factors that make splicing analysis challenging and may provide some explanation for why it is not a more common clinical tool for those working to diagnose rare disease. Most RNA-seq performed in research labs is short read, but long-reads are optimal for isoform analysis/alternative splicing analysis. However, the high error rate and low throughput of long-read sequencing means results would need to be confirmed with a low error rate method like short read sequencing.[64-66] As the cost and accuracy of long-read sequencing continues to improve and robust tools are developed for analysis, it will likely become the method of choice for identification of structural variants. In the meantime, a handful of researchers have developed methods to identify alternate splice events in short reads, depending on the higher read depth and statistical comparisons of splice junction reads between sample groups.

One of the first studies to advocate for RNA-seq in clinical genetics was Cummings et al., who reported a 35% increase in diagnostic yield when using whole mRNA sequencing of skeletal muscle for neuromuscular disorders.[62] Their method of identifying aberrant splicing, which they called Mendelian RNA-seq, was novel and improved upon many aspects of computational analysis that were lacking in previous methods. It enabled the identification of splice events that were occurring in just one or a few patients and were not necessarily captured in all available reads. Criticisms of the Cummings approach are that the cut-off parameters were arbitrary and there was no test

for statistical significance.[63,67,68] However, the Cummings method is the only one that acknowledges the reality that rare disease variant identification and interpretation often depends on manual review and curation of the evidence. Indeed, even the ACMG guidelines for variant classification state that some of their defined cut-offs are arbitrary and recommend geneticists use their professional judgement.[32]

Review of the literature on identification of alternative splicing events confirms that bioinformatic tools developed for this purpose tend to place undue emphasis on statistical significance. This leads to many tools performing well only on group comparisons and often completely ignores the nuance of rare genetic variant identification and interpretation. Each individual rare pathogenic splicing event identified is likely to only be present in one or a few diseases cases in a cohort. This makes it difficult to compare groups of samples. One of the earliest developed research tools is MISO[69], which is essentially PSI (percent *of exons* spliced in) calculations with some additional statistics. It only compares individual samples, leading to many false positives and negatives. Another tool, DEXSeq,[70] is dependent on exon counts and does not actually examine splice sites. It only detects large changes (such as exon skipping, but only when fully penetrant) and only works well with sample groups. In 2019, Fresard et al introduced a method that they propose identifies the causal gene and variants using blood RNA-seq, which makes it easier to analyze a large cohort and large number of genes.[63] However, successful application depends entirely on whether the causal variant alters expression of the gene in the studied tissue, which often limits it to just loss-of-fuction (LoF) variants since the extreme effect of nonsense-mediated decay on gene expression level in RNA is easily detectable by statistical methods. Adoption of strict filters and tests for significance means that many splicing events that are not fully penetrant and/or occur in just one individual

slip through the cracks. In just the past two years, a number of tools have been developed including LeafcutterMD, SPOT, and FRASER that purport to identify rare splicing variants using robust statistical or machine-learning methods.[67,71,72] It is likely that one or more of these tools will be an incredibly helpful addition to a rare disease RNA-seq analysis framework, but it is unlikely that one tool alone will be sufficient to perform a complete analysis[73].

1.5 Disease relevant tissue

One of the drawbacks of RNA sequencing in comparison to DNA is the need to use tissue that expresses the gene or genes of interest, ideally the tissue affected by the disorder. A handful of studies have claimed that blood is not useful for diagnostics of rare disease when it is not the target tissue, for example in generalized neuromuscular disorders.[51,62,74] These conclusions depend primarily on comparisons of read abundance from whole mRNA in skeletal muscle and blood, finding that only around half of disease-associated genes are adequately expressed in blood. Two flaws are readily apparent in this argument.

The first is the assumption that expression of more disease-associated genes necessarily leads to more diagnoses, and hence that analysis should target the tissue with the most relevant expression. If the results of mRNA sequencing were the only information available, this would be valid. However, most patients have undergone a diagnostic odyssey involving clinical evaluations and other laboratory tests, providing a plethora of information which could narrow down the list of potential genes even in a disease group as heterogeneous as neuromuscular disorders. If these genes are expressed in an accessible tissue such as blood, then the targeted RNA-seq approach can be well

justified.[63] In addition, landmark RNA-seq diagnostic studies have repeatedly found that the number of aberrant events that are potentially causative identified in each individual is prohibitively high to proceed with the manual analysis[62,63,67,72]. Each of these studies deployed methods to narrow the list to just a handful of events per sample, in many cases prioritizing candidate genes using the patients clinical or genomic data. By using a targeted panel, one would simply be performing this prioritization prior to sequencing. A more valid argument against blood RNA-seq, when blood is not the disease associated tissue, is the presence of tissue-specific splicing profiles. In a gene with many isoforms, it is important to ensure both that variants can be identified in the exons expressed in the disease-associated tissue and that any variants identified actually affect the disease-relevant transcript(s).

The second flawed assumption is that whole mRNA sequencing is the only option. While total mRNA provides a global transcriptomic profile, it is dominated by a handful of the most highly expressed genes in the given tissue.[75,76] In many cases, these genes are not relevant to the disease and many of the genes of interest are left with low read depth despite being expressed in the tissue. Performing targeted RNA-sequencing on a panel of curated genes known to be associated with the disease allows for higher read abundance per gene, even with a lower overall sequencing depth. Importantly, it also raises the representation of even low-abundance and short exons. Simply put, 10X coverage of 20,000 transcripts can be replaced by 100X coverage of 200 highly relevant transcripts for one tenth the sequencing cost.

While all of these points support the conclusion that disease-specific tissue is ideal for RNA-seq analysis, it must be noted that obtaining that tissue is a significant barrier to RNA-seq becoming a more widespread tool. Tissue biopsies are painful, invasive, and not

regularly collected in patients for diagnostic or research use, particularly when other options are available[51]. One of the main reasons that previous RNA-seq studies have focused on neuromuscular disorders is that muscle biopsies are frequently part of the diagnostic process already. There is more difficulty in obtaining control muscle biopsies for these research cohorts as well (particularly in pediatric cases), leading many to choose total mRNA seq and follow the sequencing protocols of GTEX samples in order to take advantage of the large control cohort it can offer.[62,65,68] However, at least one study found that the transcriptome of pediatric individuals tends to differ from adults prior to the age of 10, suggesting that GTEX is not a good control cohort for pediatric RNA-seq research.[77] The use of targeted RNA sequencing panels on blood samples, when the panel and cohort are properly selected, can be just as, or more, successful in diagnosing rare disease cases. As part of this thesis I show that whole blood, which is often more easily obtainable than affected tissue and is regularly taken for other routine tests, can be used in lieu of the disease tissue in instances where the gene of interest is also expressed in blood.

1.6 Specific rare diseases

The specific rare diseases and groups of disorders studied as a part of this thesis are introduced here. A more thorough discussion of each takes place in the associated chapter.

1.6.1 Dysferlinopathies

Neuromuscular Disorders (NMDs) are a group of over 200 conditions that involve the peripheral nervous and muscular systems, many of them genetic in origin.[42,78] Although individually each disorder is very rare, as a whole NMDs affect more than 1:3000

people with onset ranging from birth to late in life, and symptoms vary from mild weakness to severe disability and shortened lifespan. There is no cure or even treatment for most NMDs. In addition to this phenotypic heterogeneity, NMDs also exhibit heterogeneity at the genetic and allelic level.

Dysferlinopathies are an autosomal recessive subset of NMDs that are caused by mutations in the *Dysferlin* (*DYSF*; HGNC ID: 3097) gene primarily consisting of Limb Girdle Muscular Dystrophy Type 2B (LGMD2B; OMIM #253601) and Miyoshi Myopathy Type 1 (MM; OMIM #254130) but also including HyperCKemia and Distal Myopathy with anterior tibial onset (DMAT; OMIM #606768).[79] Because phenotypically similar forms of Limb Girdle Muscular Dystrophy and Miyoshi Myopathy are caused by mutations in other genes, identification of the causative variants in the *Dysferlin* gene is required for a diagnosis of a dysferlinopathy. Receiving a specific genetic diagnosis is important for patients, not only for peace of mind and family planning, but also for avoidance of harmful treatments. Furthermore, enrollment in current and future clinical trials for therapies will likely target specific genes and mutation types, as is the case for clinical trials related to a treatment that was recently approved for Duchenne Muscular Dystrophy.[80]

1.6.2 Primary Immune Deficiencies

Primary immune deficiencies (PID) describe a group of more than 300 disorders caused by genetic variants affecting the immune system.[81] Children are more commonly affected than adults, and estimates of prevalence range from 1:1200 births to 1:10,000.[82] National surveys have shown that more than half of PID cases are not diagnosed until adulthood. PIDs are often marked by chronic infections, and 70% of surveyed patients reported at least one previous hospitalization.[83-85] In some cases,

early intervention and treatment can allow the patient to live a relatively normal life, making early diagnosis one of the top priorities. In the same survey, nearly half of those diagnosed with a PID reported no hospitalizations after being diagnosed, underscoring the importance of early diagnosis.

1.6.3 Very Early Onset Inflammatory Bowel Disease

The genetic contribution to inflammatory bowel disease (IBD) is known to be complex, with GWAS studies implicating multiple genes and variants with a range of effect sizes.[86-88] However, some evidence suggests that in the small percentage of cases occurring in infants and children < 5 years of age, the cause may in some cases be monogenic in nature.[89,90] These cases are known as very early-onset inflammatory bowel disease (VEO-IBD). Registries at the Children's Hospital of Philadelphia and at Boston Children's Hospital, among others, are slowly expanding the recognition of this condition, and already over 50 genes have been causally linked to the condition.[91-93] Mutations affecting the protein structure encoded by genes such as *CTLA4* (HGNC ID: 2505), *FOXP3* (HGNC ID: 6106), and *STAT3* (HGNC ID: 11364) implicate altered regulation of immune function, and hence imply that transcriptome analysis of blood from patients with VEO-IBD, similar to PID, might well improve the clinical diagnostic yield for these conditions.

1.7 Thesis outline

In this thesis, I will show the clinical utility of RNA-seq for neuromuscular and immune disorders through four studies that differ in approach and/or disease. In Chapter 1, a GNE-myopathy case study validated the targeted RNA-seq NMD gene panel and showed how allele expression can inform further functional assays. In Chapter 2, an

analysis of transcriptome-wide splicing in adolescent IBD patients revealed widespread differential isoform usage in two patients with Crohn's Disease causing their ileal tissue to resemble rectal tissue, implicating splicing as a contributor to IBD disease pathology. Chapter 4 involves the development and refinement of my bioinformatic pipeline and analysis approach for targeted RNA-seq. In addition to validating the use of blood as a substitute for disease tissue in certain NMDs, I show the power of RNA-seq to evaluate VUS, identify aberrant splicing, and phase variants without parental samples. Finally, in Chapter 5 I apply the analysis approach from Chapter 4 to a small cohort of PID and VEOIBD patients, where results indicate even rare diseases without simple Mendelian inheritance patterns benefit from RNA-seq.

CHAPTER 2. CLINICAL UTILITY OF RNA SEQUENCING TO RESOLVE AN UNUSUAL GNE MYOPATHY

This chapter is adapted from the published case report [94] (citation below) to focus on and emphasize my contribution, namely the analysis of RNA-seq data.

Chakravorty S, **Berger K**, Arafat D, Nallamilli BRR, Subramanian HP, Joseph S, Anderson ME, Campbell KP, Glass J, Gibson G, Hegde M. Clinical utility of RNA sequencing to resolve unusual GNE myopathy with a novel promoter deletion. Muscle & nerve. 2019;60(1):98-103.

2.1 Abstract

Mutations in the UDP N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase (*GNE*) gene can cause GNE-related myopathies that are mostly autosomal recessive diseases associated with the juvenile-onset neuromuscular disorder known as hereditary inclusion body myopathy (HIBM). In this report, we describe a patient showing an unusual manifestation and progression of HIBM with skeletal muscle weakness and some inflammation, and harboring compound heterozygous mutations in *GNE* gene. We used exome sequencing, cDNA sanger confirmation, NGS-based whole mRNA-sequencing, and microarray-based comparative genomic hybridization (aCGH) to detect the novel combination of compound heterozygosity. We found monoallelic expression of the *GNE* gene harboring the V727M missense variant suggesting the deletion of an upstream promoter for the other normal allele reducing *GNE* gene expression by ~50% in muscle. We performed gene set enrichment analysis (GSEA) on significantly ($p<0.05$) differentially regulated muscle genes which showed pathways such as that in dermatomyositis, inflammatory pathways, extracellular matrix organization and cell adhesion, myogenesis, mitochondrial oxidative energy metabolism, slow twitch fiber

generation are mostly affected that correlated well with affected cellular components and genes' molecular functions, reminiscent of an abnormal HIBM. Our study shows the importance of considering a-CGH and functional assays such as transcriptome sequencing in the clinic for faster and definitive molecular diagnosis for GNE-related myopathies.

2.2 Background

In order to enhance molecular diagnostic yield in neuromuscular disorders (NMDs), functional assays downstream of genomic DNA-level has been recommended by the American College of Medical Genetics and Genomics and College of American Pathologists (ACMG-AMP) committee guidelines[32]. Integrating transcriptome sequencing and other functional assays with genome sequencing has been previously shown to increase efficiency of annotating functional variants [95,96], which enhances our understanding of gene-variant association to disease etiology [97] and furthers clinical diagnostics of NMDs. Here, we describe using functional genomic approaches for definitive molecular diagnosis of NMDs in an abnormal progressive GNE myopathy case.

GNE-related myopathy or GNE myopathy (OMIM #605820) is a rare, recessive inherited, degenerative NMD involving skeletal muscle disorder caused by variants in the human *GNE* gene (HGNC: 23657; NC_000009.12) with early adult onset [98-101], also known as vacuolar myopathy sparing the quadriceps, or hereditary inclusion body myopathy (HIBM), or inclusion body myopathy or myositis 2 (IBM-2), or Nonaka myopathy. The disease is progressive and leads to marked disability within 10–20 years of initial symptoms [101]. Although all leg and all hip muscles get affected, relative sparing of the quadriceps occurs. But these muscles become affected at advanced stages of the disease

along with neck muscles and respiratory muscles [102-104]. Moreover, dilated cardiomyopathy and cardiac conduction abnormalities have been reported in patients, some of whom suffered sudden cardiac death [105-107]. If inflammation occurs, the disease is generally diagnosed as inclusion body myositis or polymyositis which is one of the inflammatory myopathies. The *GNE* gene encodes a bifunctional enzyme, uridine diphosphate (UDP)-N-acetylglucosamine (GlcNAc) 2-epimerase/acetylmannosamine (ManNAc) kinase (GNE) [99,108]. The GNE enzyme catalyzes the first two rate-limiting steps in the biosynthesis of 5-N-acetylneuraminic acid (Neu5Ac) or sialic acid [109,110] that are found as the non-reducing terminal glycans on various glycoproteins and glycolipids functioning in a variety of cellular signaling pathways [111], especially the sarcolemmal and extracellular matrix (ECM) glycoproteins, such as the sarcoglycans and dystroglycan, suggesting their importance in cell-cell adhesion and cell-extra/intracellular matrix interactions.

The molecular pathology of GNE myopathy is still not clear. Most GNE myopathy patients harbor bi-allelic (compound heterozygote) missense variants in *GNE* which reduce GNE epimerase and kinase enzymatic activities impairing sialic acid production which appears to be the main cause of pathology but is not clearly understood [112-115].

Here we report a novel compound heterozygote variant combination consisting of a novel deletion variant and a common missense variant in *GNE* in an early adult patient that shows abnormal progression of the disease. We emphasize the functional assays we used in this case to definitively diagnose and explain the clinical utility of RNA-sequencing for diagnostics.

2.3 Materials and Methods

Methods are described below based on their chronological order of testing for the proband and parents. Due to the clinical manifestation of the patient, exome sequencing was performed which resulted no definite diagnosis. After this, RNA-sequencing was performed which gave clues to test deletions/duplication using aCGH that together resulted in definitive genetic diagnosis. To characterize further using target muscle biopsy tissue and to understand the connection between patient clinical features, pathophysiology and molecular diagnosis, we performed gene expression analysis and gene ontology-gene set enrichment analysis on the RNA-seq muscle transcript abundance data and performed western blot analysis on the biopsy to understand glycosylation defects in the patient muscle.

2.3.1 Exome sequencing and analysis

Peripheral blood was collected into EDTA tubes from the patient proband and his parents. Genomic DNA was extracted using GenElute™ Blood Genomic DNA (Sigma-Aldrich NA2020) according to the manufacturer's protocol. Exome Sequencing (ES) was performed on genomic DNA using the NimbleGen (Madison, WI) v3.0 targeted sequence capture method to enrich for the exome. These targeted regions were then sequenced using the Illumina (San Diego, CA) HiSeq 2000 sequencing system with 100 base pair (bp) paired-end reads at an average coverage of 100X in the target region. The targeted regions included the exon and 10 bp of flanking intronic sequence. In general, ES assays performed at Emory Genetics Laboratory (EGL) have an overall coverage of 92.9%, with as high as 94.8% coverage in the coding region.

2.3.2 Bioinformatics analysis and variant classification

Bioinformatic analysis was performed using NextGENe software from SoftGenetics (State College, PA). The NextGENe output was customized to mine variants from EGL's internal variant database EmVClass [116] other public databases, and variant prediction tools, such as SIFT [117] and PolyPhen [118]. All variants detected were classified using population frequency data available from the Exome Variant Server (http://evs.gs.washington.edu/EVS/) and previous reports of disease association and pathogenicity available through the Human Gene Mutation Database (HGMD), National Center for Biotechnology Information PubMed, and Google. Variant classification was performed based on American College of Medical Genetics and Genomics- Association for Molecular Pathology (ACMG-AMP) guidelines [119]. A detailed overview of the bioinformatics pipeline and variant annotation protocol is described elsewhere [120]. All variants were curated and reviewed by board-certified laboratory directors and maintained as an in-house variant database; they were made publicly accessible via EGL's online tool, EmVClass[116].

2.3.3 Sanger confirmation of genomic DNA

PCR-amplified genomic DNA and RT-PCR amplified cDNA products of the *GNE* gene were Sanger sequenced for confirmation of the variants. An automated primer design script, developed and validated in-house, was used to design primers[121].

2.3.4 Next Generation Sequencing (NGS)-based RNA-sequencing

High quality (RNA Integrity Number; RIN>7) RNA was extracted from right thigh muscle biopsies of the patient and six control individuals using Aurum™ Total RNA Fatty

and Fibrous Tissue Kit (cat# 7326870) following the manufacturer's protocol. Library preparation was performed using Agilent Sure Select XT RNA Reagent kit (cat# G9692B) following manufacturer's protocol. Patient samples in replicate and six control normal individual muscle biopsy mRNA samples were sequenced in the Georgia Tech Molecular Evolution Core on an Illumina Next Seq instrument to obtain high output paired-end by 150bp reads at a depth of more than 150 million reads. From the whole transcriptome sequence data, we focused our analysis on 273 genes that are known to be NMD-associated and are known to have skeletal muscle expression as retrieved from GTEX portal[122]. These 273 genes were curated initially based on the associations to NMDs listed in http://www.muscle.ca/wp-content/uploads/2012/10/Disorder_List_ENG_May2017.pdf.

2.3.5 Bioinformatics workflow

Raw sequencing reads were checked for quality using FastQC[123]. Reads were not trimmed beyond removal of adapter sequences [124]. Human reference genome GRCh38 (NCBI) and NCBI *Homo sapiens* Annotation Release 106 were obtained from Illumina iGenomes and sequenced reads were aligned using the splice-aware alignment program STAR version 2.5.2b in 2-pass mode [125].

Quality metrics for all samples were obtained by running Picard RNASeqMetrics (*Available online at: http://broadinstitute.github.io/picard)* and principal component analysis (PCA) on gene expression was performed to check for outlier status based on tissue composition or contamination. Initial unsupervised clustering of samples by relative gene expression levels clearly differentiated the patient sample from six controls.

For analysis of the variant of interest, alignment files were loaded into Integrated Genomics Viewer (IGV) [126] to confirm positive identification of variants found in ES and analyze the transcript structure of the candidate gene for potential variant effects such as abnormal splicing or allele specific expression.

To identify possible downstream effects, differential gene expression analysis, including both normalization to account for sequencing depth and RNA composition as well as identification of surrogate variables to correct for un-modeled variation, was performed using the R packages DESeq2 version 1.16.1 [127] and SVA version 3.24.3 [128]. Read counts mapped per gene were obtained using HTSeq-Count [129] and the analysis was conducted following the Bioconductor DESeq2 vignette and the RNA-Seq Gene Expression Analysis workflow [130]. To maintain consistency, the same annotation file used for alignment was used for all downstream analysis. The result file was filtered to include only significantly up or down regulated genes using Benjamini-Hochberg adjusted[131] $p<0.05$, from our curated panel of 273 NMD-associated genes.

2.3.6 Array CGH

The targeted gene high-resolution oligonucleotide CGH array was custom designed on Oxford Gene Technologies (OGT) 180 K platform to detect deletions and duplications. Long oligonucleotides (~45–60 mer) were used to design the array, with repeat sequence masking implemented to ensure greater sensitivity and specificity. The *GNE* gene and its upstream was covered by 720,000 probes including covering the *GNE* 13 exons at an average spacing of 15 bp between probes and the intronic region was covered at an average spacing of 25 bp between probes. Use of intronic

oligonucleotide probes allows detection of dosage changes within the entire genomic region of the gene and determination on the approximate breakpoints [132].

2.3.7 Gene Ontology-Pathway Analysis

We performed a tiered approach to investigate the genetic and pathway networks affected by genes that are statistically significantly differentially regulated in the proband muscle biopsy compared to the six control biopsies. We performed a gene-ontology-based analysis with MSigDB (Molecular Signature Database) [133,134] that curates gene ontologies (GOs), biological processes or pathways, molecular functions, and cellular compartments separately that are significantly associated with the gene clustering ($p<0.01$; $FDR<0.05$) based on greater than 1325 biologically defined gene sets, similar to our recent study on infantile spasm[135]. We retrieved the top 100 enriched gene set pathways/cellular compartment/molecular functions that are affected in the patient muscle. The individual gene sets were then manually collapsed with biological evidence. The criteria behind manual compiling are based on a) established hierarchical superfamily of the GO functions [136] and external links with ontology and hierarchy for non-GO gene sets from MSigDB database, and b) biological similarity of the individual functions and pathways (eg. “HALLMARK_MYOGENESIS” and “GO_MUSCLE_CONTRACTION” were compiled into “Muscle Development and Contraction"). The compilation criterion was consistently followed and performed. The value k is the number of significant genes from the patient muscle that are differentially regulated and found in a particular enriched gene set, and value K is the total number of significant genes in a gene set. The ratio k/K is the proportion of significant genes in the patient muscle found in an enriched gene set.

2.4 Results

2.4.1 Patient Clinical Presentation and Family History

The patient is a twenty one year old male with Indian ancestry with generalized muscle weakness and an abnormal progressive neuromuscular disorder (NMD) and a clinical suspicion of Polymyositis or HIBM.

2.4.2 Exome sequencing revealed a single monoallelic V727M missense variant

To identify variants of interest, whole-exome sequencing was performed on genomic DNA extracted from samples submitted from the proband, biological mother, and biological father; 97% of the exome-capture region was covered at a read depth of 20X or greater. A known "likely pathogenic" missense variant [c.2179C>T (p.V727M)] was identified in *GNE*. This variant was reported to be "pathogenic" or "likely pathogenic" by number of studies and is quite prevalent (rs121908627; allele frequency of 0.0141) among South-East Asian population (Indian and Thai descent) [99,137-144] but confirmatory functional evidence is still lacking as to its clear functional implication in the GNE protein. The V727M location in kinase domain (C-terminal end) of GNE potentially suggests a possible, yet not definite, role in pathogenesis towards its kinase activity to phosphorylate ManNAc which is known to be a key process in sialic acid biosynthesis for glycoprotein glycosylation [145]. But, recently, due to the high prevalence of V727M in south-east Asian population, [144] the pathogenicity of V727M variant has been questioned. In fact, it is not clearly known if the V727M variant is pathogenic due to its effect in reducing allelic expression or reducing the kinase activity. Most GNE-related myopathy cases harboring a single V727M variant are prevalent as compound heterozygotes with the other allele harboring exonic/intronic deletions or missense variants, but our exome sequencing did

not reveal any other pathogenic variants in *GNE* or other genes. Therefore, we were not able to achieve a definite molecular diagnosis at this stage.

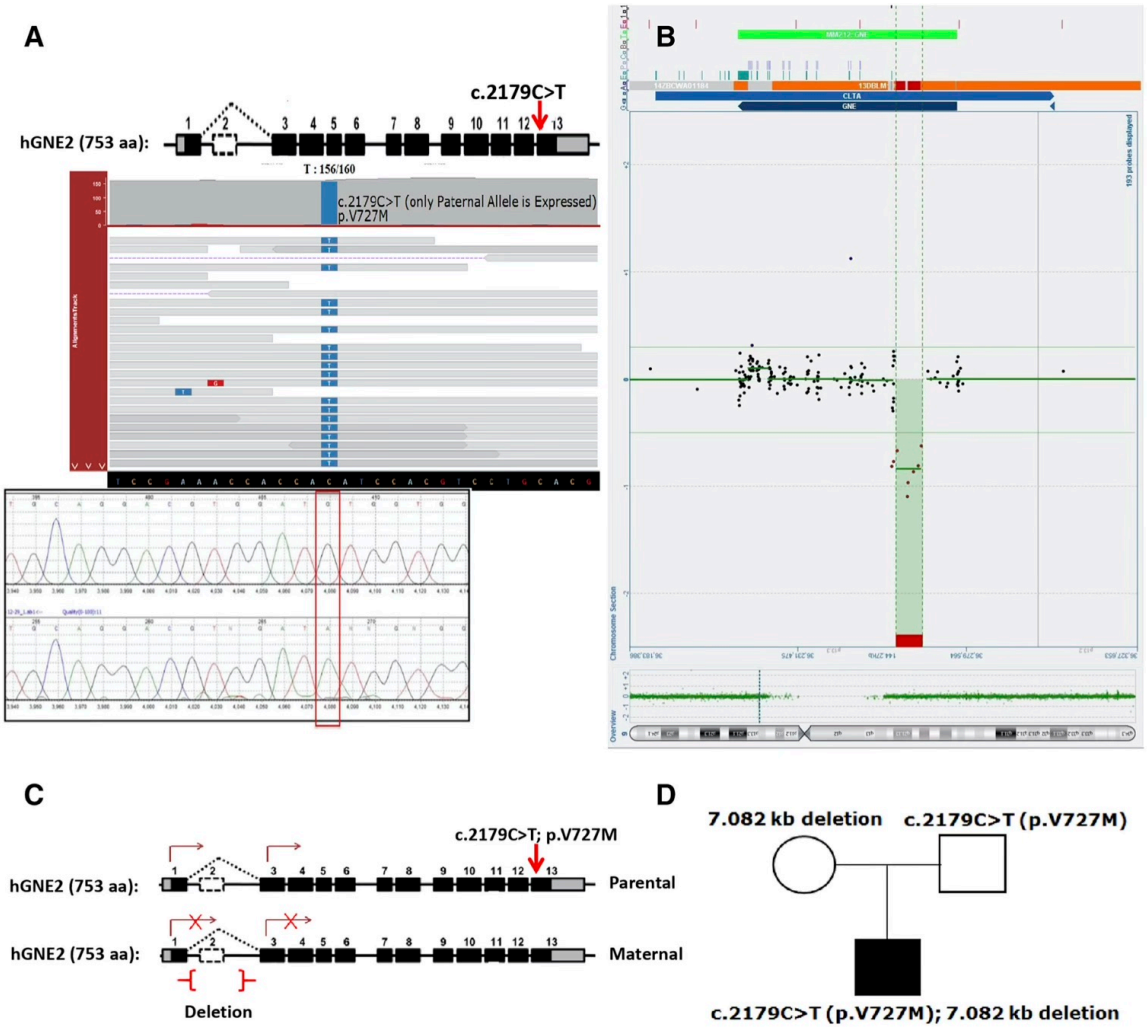

Figure 2.1 A novel *GNE* promoter deletion. (A) Integrated Genomics Viewer (IGV) pile up of RNA-sequencing showing monoallelic expression of GNE gene with only the allele harboring c.2179C>T:G>A (p.V727M) missense "likely pathogenic" variant expressed. The red arrow indicates the position of the V727M variant in exon 13 of the GNE gene. Sanger sequencing confirmation was performed on cDNA showing monoallelic expression as shown below. (B) aCGH signal showing a deletion upstream of the GNE gene with genomic breakpoints at nucleotide positions g.36,259,402 and 36,266,483 was detected in this individual (SCV000599234). This deletion is 7.08 kb in size and encompasses the untranslated exon 2 of the hGNE2 transcript but upstream of the hGNE1 transcript of the GNE gene. (C,D) Exome sequencing and later aCGH of trios reveal that monoallelic expression was due to expression of only the paternal allele of GNE in the proband.

2.4.3 Transcriptome sequencing revealed monoallelic expression of V727M allele

NGS-based transcriptome sequencing (RNA-seq) reads of whole mRNA using target muscle biopsy revealed presence of only the V727M allele of the *GNE* gene (Figure 2.1A). The depth of the alternate allele is very high giving us confidence in our variant call. Exome sequencing revealed heterozygous V727M *GNE* variant and the absence of the normal V727 GNE allele from RNA-seq reads suggest monoallelic expression of the V727M mutant GNE allele. This result suggested a possible deletion or duplication event that can block expression of the normal allele for which we performed array comparative genomic hybridization (aCGH) using patient genomic DNA.

2.4.4 aCGH reveal an large deletion upstream of untranslated exon 1 of GNE

Recently, a novel ~11.3 kb deletion encompassing exon 2 was found along with a single V727M variant [146]. In our study, aCGH deletion/duplication analysis of the *GNE* gene, revealed a novel ~7-9 kb deletion (g.36,259,402 and 36,266,483) upstream of the untranslated exon 1 of the *GNE* gene (Figure 2.1B-D) which was not previously reported nor found in genomic databases. To our knowledge, this is the first report of a large deletion upstream of all exons (non-exonic/intronic) of the *GNE* gene in any GNE myopathy patient. This result shows a novel compound heterozygous variant combination of a large deletion upstream of *GNE* gene along with previously reported a single missense variant V727M. Since the RNA-seq result clearly shows monoallelic expression of the mutant V727M allele, this suggests that the upstream deletion is heterozygous encompassing promoter region of the normal GNE allele that abolishes its expression.

2.4.5 Gene Expression show 50% reduction in GNE gene expression

Cluster analysis of transcript abundance of 273 NMD-associated genes from RNA-seq data in the patient sample in replicate compared to data from 6 control normal biopsies show separate clustering of the patient samples and the controls suggesting the transcriptome profile of the patient muscle is clearly different from normal control. Out of 273 genes, 89 muscle genes are significantly differentially expressed between the patient and control samples. Importantly, *GNE* gene expression is reduced by ~50% compared to 6 controls (Figure 2.2) further suggesting only the monoallelic expression of V727M mutant allele.

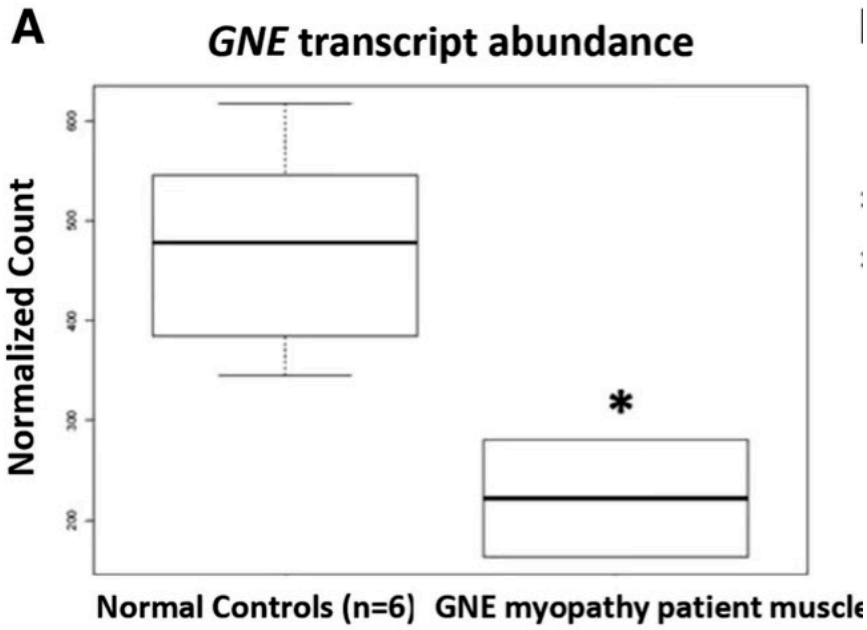

Figure 2.2 Reduced *GNE* expression. Approximately 50% lower-expression ($P < 0.05$) of GNE in GNE myopathy patient muscle compared with that in 6 control normal muscle biopsies.

2.4.6 Gene Ontology and Gene Set Enrichment Analysis

Gene ontology-based gene set enrichment (GO-Pathway) analysis on the differentially-regulated 89 genes identified major enriched “pathways,” “cellular compartments,” and “molecular functions.” The common biology is a predicted effect on of protein and lipid glycosylation affecting the cytoskeleton-intracellular matrix and ECM cross-talk through sarcolemmal proteins, important for the sarcomere integrity.

2.5 Discussion

Our study provides important insights for molecular diagnostic approaches to understand the pathological and molecular nature of unusual myopathies. We report here a family having a patient with a novel upstream promoter-region large deletion in the *GNE* gene, which abolishes expression of the respective allele. Previous reports showed that patients with compound heterozygous variants in both epimerase and kinase GNE domains manifest more severe phenotypes than those with both variants in 1 domain[147], suggesting that mild pathogenicity of missense variants in each domain needed for more disease severity. Although V727M pathogenicity is uncertain given its relatively high prevalence in South Asians, the most parsimonious conclusion given many other similar reports is that this compound heterozygous state contributes to the pathology.

The second causal variant was inferred from the combination of aCGH and RNA-seq that definitively diagnosed the case as GNE-related myopathy and led to identification of multiple gene expression perturbations. Previously, Zhu et al.[148] showed that large promoter region deletions in *GNE* are common in already clinically diagnosed GNE-myopathy patients, and Garland et al.[146] showed that a combination of such deletions

and a V727M missense variant causes a more severe reduction in *GNE* expression than the combination of V727M and another missense variant. Here, we show that such variant combinations are associated with unique GNE-related myopathy pathology and the clinical/molecular diagnostic hurdles faced. Consequently, it is likely that the combination of reduced transcription due to promoter region deletion and possible V727M-induced subtle altered kinase activity is required for the unique HIBM-like symptoms. Further functional studies are needed to classify the pathogenicity of V727M.

As per ACMG guidelines[32], because the deletion variant causes a 50% reduction in *GNE* gene expression, we clinically classify the variant as “pathogenic.” This potentially results in a significant reduction in key sarcolemmal protein α-DG glycosylation and aberrant expression of core α-DG and β-DG, which along with altered expression of genes and pathways found in GO-pathway analysis could explain the muscle wasting and weakness.

Importantly, this study shows the power of using aCGH, RNA-seq and focused functional assays on target muscle tissue following clinical/pathological clues for improving diagnostic efficiency and timeliness in the evaluation of undiagnosed myopathies. We believe that this approach will be broadly applicable to the diagnosis of NMDs and will thus harness the advances in clinical genomics and developing precision therapies.

CHAPTER 3. ALTERED SPLICING ASSOCIATED WITH THE PATHOLOGY OF INFLAMMATORY BOWEL DISEASE

Berger K, Somineni H, Prince J, Kugathasan S, Gibson G. Altered splicing associated with the pathology of inflammatory bowel disease. Human genomics. 2021;15(1):1-10.

3.1 Abstract

3.1.1 Background

Aberrant splicing of individual genes is a well-known mechanism promoting pathology for a wide range of conditions, but disease is less commonly attributed to global disruption of exon usage. To explore the possible association of aberrant splicing with inflammatory bowel disease, we developed a pipeline for quantifying transcript abundance and exon inclusion transcriptome-wide and applied it to a dataset of ileal and rectal biopsies, both obtained in duplicate from 34 pediatric or young adult cases of ulcerative colitis and Crohn's disease.

3.1.2 Results

Expression and splicing covary to some extent, and eight individuals exhibited aberrant profiles that can be explained by altered ratios of epithelial to stromal and immune cells. Ancestry-related biases in alternative splicing accounting for 5% of the variance were also observed, in part also related to cell-type proportions. In addition, two individuals were identified who had 284 exons with significantly divergent percent spliced-in exons, including in the established IBD risk gene *CEACAM1*, which caused their ileal samples to resemble rectum.

3.1.3 Conclusions

These results imply that quantitative differences in splice usage contribute to the pathology of inflammatory bowel disease in a previously unrecognized manner.

3.2 Background

Defective RNA splicing contributes to the etiology of a wide variety of diseases.[149] Single gene defects that weaken or abolish splice sites, or activate cryptic ones, have been associated with over 200 human diseases, including progeria, cystic fibrosis, muscular dystrophies, and some cancers.[150-153] Computational analyses have further identified variants in over 80,000 splicing regulatory motifs [154], and scores such as TraP (TRanscript inferred Pathogenicity Score) provide pre-computed predictions of likely splice defects for polymorphisms affecting all human genes.[59] Just as importantly, global mis-regulation of the splicing of hundreds of genes due to aberrant activity of components of the spliceosome, is known to contribute to pathology for a variety of conditions, notably myelodysplastic syndrome, myotonic dystrophy, several neurological disorders, and cancer metastasis.[155-159]

The inflammatory bowel diseases (IBD) ulcerative colitis (UC) and Crohn's disease (CD) afflict approaching 1% of adults in developed countries and have been rising in prevalence globally for several decades.[160] They are well known to involve aberrant gene expression in the gut [161,162] as well as peripheral immune system [163,164], and signatures of severe disease at diagnosis have been associated with progression to complicated disease or remission [165-167] and are being developed as biomarkers of therapeutic response.[168] There is also some indication that gene expression is to some extent ancestry-dependent, resulting in mis-regulation of pathways related to cytokine

signaling, extracellular matrix function, and mitochondrial activity that is biased toward more adverse outcomes in African Americans.[169] To date, to our knowledge, there have not been any reports of splicing defects in IBD, so we asked whether transcriptome profiles assessed by RNAseq of bulk ileal and rectal tissues might provide evidence for unusual splice isoforms associated with IBD in a dataset of paired ileal and rectal biopsies from a cohort of 34 young individuals with CD or UC.

3.3 Methods

We analyzed whole mRNA sequencing profiles of 124 samples obtained from 34 young donors with IBD (age range 8-20 years). Duplicate biopsies of both the ileum and rectum were analyzed, in general 4 samples per donor, although 6 donors were represented by only 3 samples and three by a single biopsy from each location. Individuals were closely matched for ancestry (18 European, 16 African American), sex (16 male, 18 female), disease type (20 Crohn's disease, CD; 14 ulcerative colitis, UC), and disease status at time of sampling (20 established cases, 14 cases at diagnosis). All donors were tumor free at biopsy. Following quality control, total transcript abundance was measured for 18,929 genes, and percent-spliced-in (PSI) estimates [170] were obtained for 7,001 variable exon bins.

3.3.1 Sequencing

RNA was extracted and library preparation was performed using the Illumina TruSeq Stranded mRNA kit. Paired-end 100bp stranded sequencing was performed for all samples on an Illumina HiSeq at a median read depth of 22.7 million (range: 10.2-106 million) read pairs.

3.3.2 Preprocessing

FastQC was run on raw fastq files to ensure mean phred scores per sequence and per base were above 27, to check consistency among samples in per sequence GC content, per base N content, sequence length distribution, and sequence duplication levels, and to check for the presence of adapters.[171] Samples were trimmed up to but not beyond the adapter using trimmomatic.[172] Samples were aligned with the STAR splice aware aligner to hg38 using the Gencode v29 primary assembly sequence and annotation.[125,173] Default parameters were used with the following exceptions: to increase accuracy of splice site mapping and discovery, two-pass mode was invoked; novel splice junctions were required to have a minimum overhang of 8bp, and a minimum of 5 unique reads was required for a splice site to be included in the splice junction output. To ensure each read used in downstream analysis was accurately mapped and results were not affected by high homology regions such as pseudogenes, all multimapping reads (which map equally well to two locations in the genome) were filtered out. We further confirmed that all reads aligning to the CEACAM1 alternative splice bin did not align to the duplicate pseudogene [174], which possesses sufficiently divergent nucleotide sequence to prevent multi-mapping.

3.3.3 QC

In order to remove samples exhibiting extreme 5' or 3' bias or mapping issues that could affect splice calculations, sample quality was assessed using the Quality of RNA-Seq Tool-set (QoRTs) which evaluates cumulative gene diversity, gene-body coverage, and number of observed splice junction loci.[175] One rectal sample from individual 6 and one from individual 26 were observed to be extreme outliers in 3' bias and were removed,

leaving both individuals with two ileal and one rectal sample. To confirm that each sample from an individual was indeed the same individual, variant calling was performed at Purcell's 5k sites following GATK best practices and identity-by-descent was compared using output from PLINK.[176-178] A PI_HAT minimum threshold of 0.7 was used to confirm a match between two samples. The single rectal sample from individual 18 failed the identity-by-descent QC measure, leading to the removal of all samples from individual 18.

3.3.4 Gene Expression Analysis

Overall differential gene expression was performed with DESeq2, using the STAR raw read counts per gene output and including ancestry, disease, location, and the interaction of disease and location in the design formula.[127] Prior to analysis, genes were filtered for mean coverage >5 reads and both Principal Component Analysis (PCA) and Principal Component Variance Analysis (PCVA) were performed on the final set of 18,929 genes, with results listed in Table 3.1. PCA captures the major components of covariance of gene expression, and PVCA sums the amount of variance in each PC that is associated with the influencing factor, weighted by the variance in gene expression explained by the PC. We only analyzed the first 10PC of both RNA abundance and PSI since smaller PC explained less than 1% of the variance each and tend to capture differences among individuals or noise.

Gene expression was also used to estimate abundance of specific cell types. Lists of genes expressed in immune, epithelial, and fibroblast cells were created from single cell RNAseq data for these cell types obtained from the colonic mucosa of ulcerative colitis patients [179], using thresholds of >5 counts per million (CPM) in one cell type and <1

CPM in the other two cell types. We then generated PC1 for each list and estimated the correlation with location and ancestry in order to evaluate the contribution of cell type abundance to these effects.

3.3.5 Splicing Analysis

Splicing patterns between individuals were compared using the Percent Spliced In (PSI) metric, which was calculated per exonic bin for each sample. PSI is independent of library size and yields a score between 0 and 1 representing the proportion of isoforms that include a particular exonic bin. Inclusion (IR) and exclusion (ER) read counts were obtained following the protocol outlined in Schafer et. al.[170] using the splice junction output from the STAR alignment, and the recommendation of requiring >10 ER to identify alternatively spliced exon bins was used to inform the following filtering steps. For each sample, if a site had <10 ER the PSI score was rounded up to the nearest tenth (IR >10) to lessen the impact of low exclusion counts or NA (IR <10) to indicate no coverage, allowing more exonic bins to be evaluated across all samples without low ER counts dominating the analysis. To limit analysis to genes expressed in both tissues, rows where one or more samples had no coverage were excluded (511,191 exon bins, leaving 108,091). Rows where all samples had the exact same PSI (0 or 1) were removed (64,129), reducing analysis to only those sites where one or more samples had variability in level of exon exclusion. Subsequently, to focus on splice bins with potential group-wise differences, further filtering was performed to exclude splice bins where 40% of samples were close to constitutively included or excluded (>95% and <5% PSI, respectively). This filtering reduced the original 619,282 potential splice sites to 7,001 in the final analysis, with a mean PSI score of 0.45. PCA and PCVA were performed on the PSI estimates before and after the final stage of filtering, yielding very similar results presented in Table

3.1. To identify differentially used exonic bins, linear mixed models were performed on the final set of PSI scores with the lme4 R package including fixed effects of disease, location, and ancestry and the interaction between disease and location, and a random effect of individual.[180]

3.3.6 Characterization of Differential Splicing in the Spliceopathy samples

A heatmap of 50 exon bins Figure 3.2a was created to visualize differences in the splicing patterns of sample groups. Because the largest contributor of variance in PSI was tissue location, the 50 most significant exonic bins by location obtained from the previously described analysis were used. Hierarchical clustering of samples for the 50 exonic bins was performed using the Euclidian distance and complete linkage method.

The difference between two PSI averages for each site was again used to observe the extent of variation between groups (Figure 3.2d). Samples were split into ileum, rectum, intermediate ileum, intermediate rectum, and spliceopathy, and the difference in average PSI for each comparison was categorized at every exonic bin in the filtered 7001 exonic bins used for analysis.

For the identification of PSI sites in the spliceopathy samples that were significantly different from the differentiated ileum, the differentiated rectum, or both, exonic bin filtration criteria were relaxed slightly. To limit analysis to genes expressed in all three groups, rows where more than five ileal samples, five rectal samples, or one spliceopathy sample had no coverage were removed. The missing values in rows where 1-5 ileal or rectal samples had no coverage were replaced with the tissue location average at that site, allowing these additional 27,514 exonic bins to potentially be included in this analysis. The remainder of the filtration steps were carried out as before, this time reducing the

original 619,282 potential splice sites to 9,499 in the final analysis, with a mean PSI score of 0.43. To identify differentially used exonic bins, linear mixed models were performed on the final set of PSI scores including fixed effects of group (ileum, rectum, or spliceopathy) and ancestry and a random effect of individual.

3.4 Results

3.4.1 Effects of disease, location, and ancestry on splicing and gene expression

In order to quantify the influences of disease, location and ancestry on splicing and gene expression, we first computed the principal components (PC) for both the transcript abundance and PSI (percent spliced in) counts from the RNAseq dataset, and then generated a weighted sum of the influences on these measures. This principal component variance analysis revealed that three quarters (75%) of the expression variability and one third (33%) of the splicing variability was captured by the first ten principal components of the respective measures, indicating that gene level expression is far more variable than exon usage between rectal and ileal tissue. For gene expression, 40.3% of the variance was between locations (ileum and rectum), 2.3% between ancestry groups (European and African), 0.2% between disease subtypes (UC and CD), and 0.8% captured by the interaction between Location and Disease. Corresponding percentages for the splicing variance were 20.7% between locations, 5.8% ancestry groups, 0.7% disease and 0.8% the interaction effect. These proportions and the contributions to each PC are provided in Table 3.1, which also shows that the variance contributions to PSI are relatively unaffected by the threshold of inclusion, being similar for datasets with 108,091 or 7,001 exon bins. In both cases, then, consistent with previous studies, by far the largest effect is between ileum and rectum [181], a meaningful ancestry component is observed [169], and twice as

much variance is due to differences in the effect of disease on the two tissues than to disease across both tissues.

Table 3.1 Principal Component Variance Analysis (PCVA) decomposition of sources of variance. The table shows the percent variance in gene expression or splicing (PSI at two inclusion thresholds) explained by the first ten PCs and their sum, as well as the weighted contribution of each variance component term (ancestry, disease, location, and the interaction of disease and location) to these 10 PC. Gene expression is for 18929 genes, and number of PSI bins is before and after the final two stages of filtering (n=119 samples for all).

Principal component / Variance component	Gene Expression	PSI (108091 Exon bins)	PSI (7001 Exon bins)
PC1	28.9%	7.6%	9.3%
PC2	25.1%	5.5%	6.9%
PC3	5.2%	4.1%	4.1%
PC4	4.4%	3.0%	3.2%
PC5	3.5%	2.6%	2.9%
PC6	2.3%	2.4%	2.7%
PC7	2.0%	2.3%	2.3%
PC8	1.6%	1.8%	2.0%
PC9	1.6%	1.7%	1.9%
PC10	1.3%	1.7%	1.6%
SUM of PC1-PC10	75.9%	33.0%	27.9%
Ancestry	2.3%	5.8%	5.3%
Location	0.2%	0.7%	0.5%
Disease	40.3%	20.7%	27.9%
Location*Disease	0.8%	0.8%	1.0%

At the 5% False Discovery Rate, there were 9,569 differentially expressed genes by location, 1,847 by ancestry, and just 556 by disease, although 1,570 showed an interaction effect, implying that most disease effects, as expected, are specific to the ileum (in CD) or rectum (in UC). Correspondingly, there were 1,885 significant PSI by location, 90 by ancestry, and none by disease or showing an interaction effect, implying that disease has a much smaller impact on splicing in each tissue than it does on overall expression.

The first principal component (PC1) of gene expression and the first two principal components (PC1 and PC2) of splicing provide particularly strong separation by location as seen in Figure 3.1a,b respectively, with the exception of samples from 8 individuals highlighted by the solid squares which are also extreme for PC2. We provide evidence in Figure 3.1c,d that these major components of variation reflect the proportions of the three major tissue compartments [179,182], specifically with elevated epithelial contribution to the ileum relative to rectum, and immune and fibroblast contributions to the intermediate samples. Analysis of the genes altered in the intermediate-type ileal and rectal cases showed indicate that the differentiation of these samples is likely driven by an amplified immune response. Of note, the IBD-associated genes *TLR2*, *TLR4*, and *NOD2* exhibit elevated expression in intermediate ileal and intermediate rectal samples.

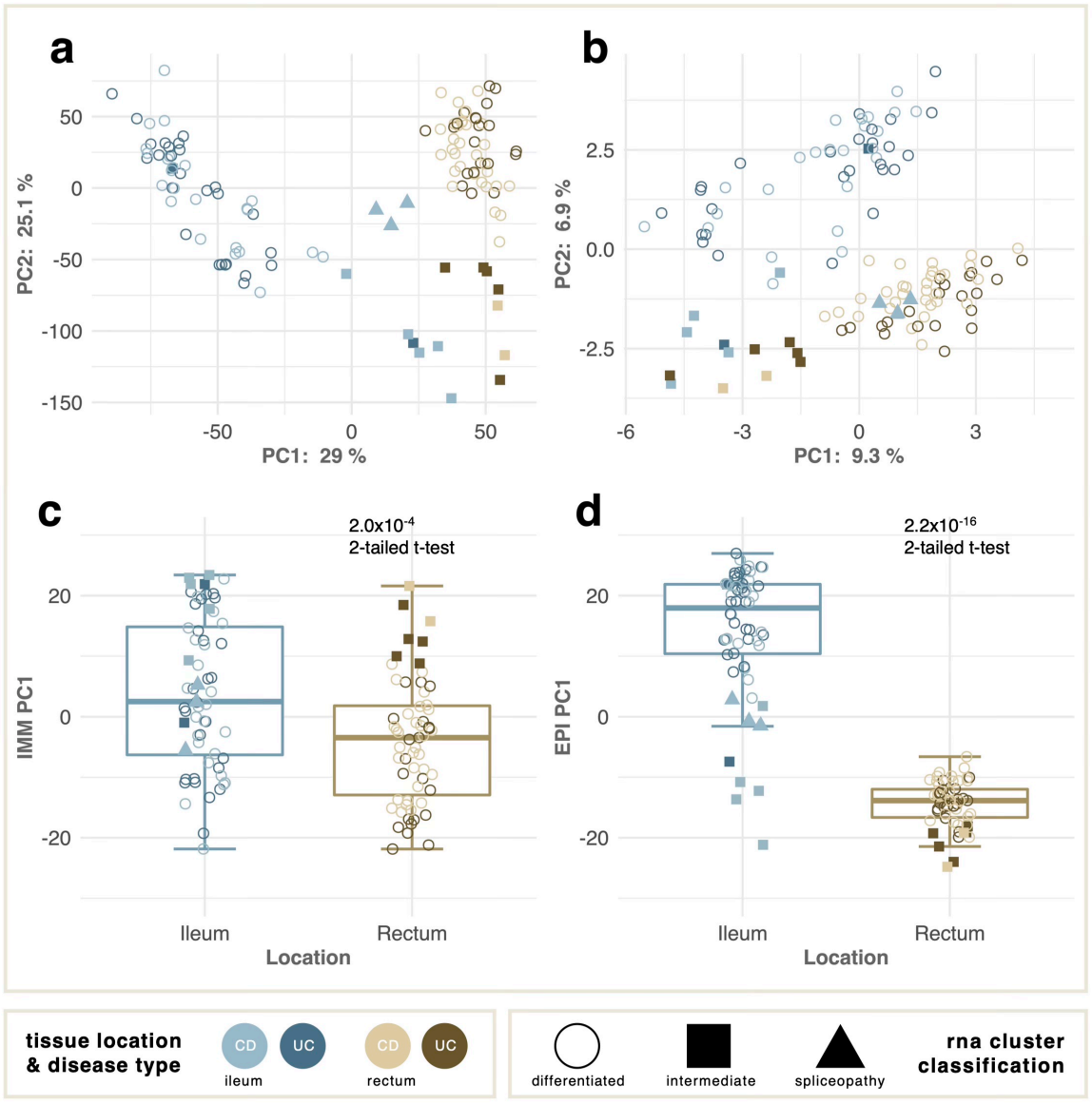

Figure 3.1 Principal components of transcript variation. (a) PC2 vs PC1 of transcript abundance showing separation of ileal (blue, n=60) and rectal (brown, n=59) samples along PC1, and of intermediate samples (solid squares) along PC2. 6 of 8 intermediate individuals are represented by two samples each; different individuals are intermediate in the two tissues. (b) PC2 vs PC1 of exon usage (PSI) showing similar separation by tissue, but with three ileal samples (blue triangles) clustering with the rectal set. Percentages refer to variance explained, shading to disease status. (c,d) Differential abundance of immune (c) and epithelial (d) cell contributions summarized by PC1 of compartment-specific gene expression differentiate ileum and rectum

3.4.2 Aberrant Profiles define "Spliceopathies"

Three ileal samples highlighted by the solid blue triangles in Figure 3.1a,b (2 from one donor, 1 from another who did not have a paired ileal biopsy) have rectal-like splicing yet gene expression intermediate between ileum and rectum. These two Crohn's disease cases thus have particularly altered splicing, suggesting that their disease is due to a "spliceopathy". Analysis of variance detected 284 differentially used splice sites in the three samples compared to ileum, and the heat map in Figure 3.2a highlights how these ileal samples globally more resemble rectum in terms of exon usage. These differentially used splice sites come from 246 genes, of which only 104 were found to be differentially expressed at the gene level, further supporting the suggestion of a "spliceopathy". A representative example, *CEACAM1,* itself an established IBD risk gene [183] whose product regulates mucosal inflammation via T-cells [184], is shown in Figure 3.2b where exon bin 12 has low, rectal-like PSI in ileum, whereas the other intermediate samples are more ileal-like. Overall expression of the gene is normal (Figure 3.2c).

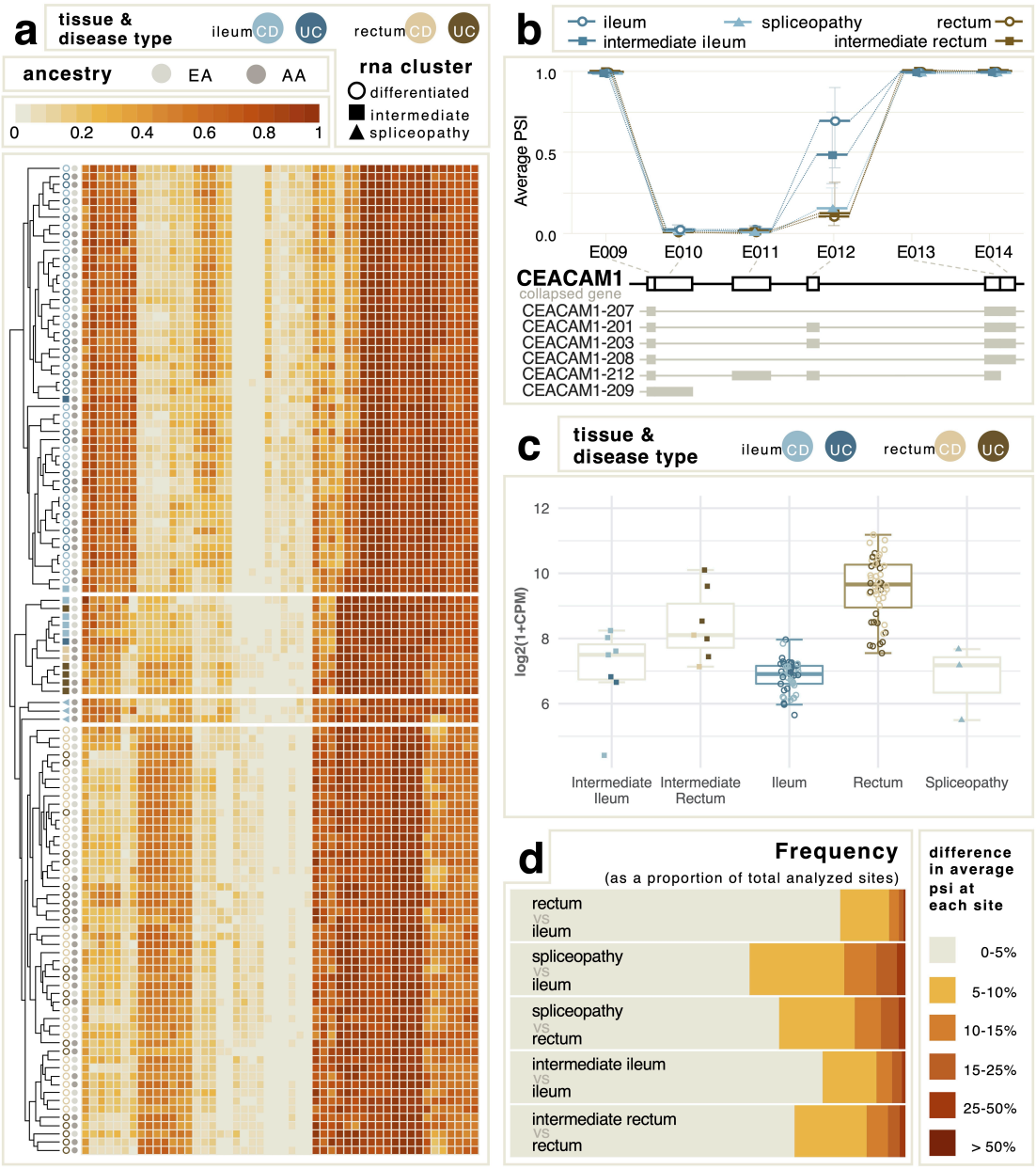

Figure 3.2 Characteristics of the spliceopathy samples. **(a)** Heat map of the top 50 most differentially abundant exons showing broad clustering by tissue (rectum to the left) but not disease status. **(b)** Average PSI of exon bins 9 through 14 of *CEACAM1*, showing average levels of E012 (corresponding to *CEACAM1* exon 7) differ by tissue and state. **(c)** Gene expression of *CEACAM1* by cluster. Intermediate ileum and spliceopathy samples are not significantly different from differentiated ileum, whereas intermediate rectum and differentiated rectum are both significantly elevated relative to ileum. **(d)** The proportion of sites with indicated difference in average PSI for comparisons of ileum (n=50) to rectum (n=52), spliceopathy ileum (n=3) to both ileum and to rectum, and intermediate ileum (n=7)or rectum (n=7) to corresponding differentiated tissue. The most differential splicing is observed in each bin above 5% for the spliceopathies

There are two main isoform types of *CEACAM1* that differ in the length of the cytoplasmic tail. The inhibitory functions of the long cytoplasmic tail isoform (*CEACAM1-L*) are well studied and *CEACAM1-L* is known to be the predominantly expressed isoform in human lymphocytes. Though *CEACAM1-S* functions are less well characterized, it has been linked to mucosal immune regulation and recent studies show that intestinal T-cells primarily express this isoform.[185] Exon bin 12 corresponds to exon 7 of the *CEACAM1* gene, which is included only in the *CEACAM1-L* isoform and also contains regions involved in the alternative splicing of this gene.[183,186] Analysis of this region did not identify any SNPs that may lead to the differential isoform ratio observed in these samples. While no other CEACAM family members exhibited altered splicing profiles we did observe elevated gene expression of *CEACAM5,* known to be a marker of Crohn's disease, in the spliceopathy samples.[187] However, expression was consistent with the level seen among rectal samples, further supporting the hypothesis of a transcriptome-wide defect causing these ileal samples to resemble rectal tissue.

Underscoring that the defective splicing is transcriptome-wide, Figure 3.2d shows the fraction of exons in bins of differential usage for various contrasts, with the greatest deviations seen for the spliceopathy samples. Despite the widespread nature of the defective splicing, separation of samples by tissue-type using PCA (principal component analysis) could also be performed reliably using just 96 of the 284 splice sites that were also differentially used in the rectal samples compared to the ileum, but did not distinguish the spliceopathy samples and rectal samples, making these exonic bins "rectal-like". Gene ontology analysis of the genes encompassing these exons identified an enrichment of genes involved in fructose catabolism. Three out of five genes in the pathway (*KHK*, *TKFC*, and *GLYCTK*) had exons exhibiting rectal-like splicing in the ileal spliceopathy samples.

In addition to differential isoform usage, overall transcript abundance of ketohexokinase (*KHK*) and triokinase (*TKFC*) was also reduced in the spliceopathy samples, to a level intermediate between rectum and ileum. Breath testing has been used to demonstrate that fructose malabsorption is quite common in individuals with ileal Crohn's disease [188], consistent with the hypothesis that an excess of short chain carbohydrates may be a trigger for pathogenesis.

Expansion of the gene ontology analysis to genes encompassing all 284 exons found to be differentially used in spliceopathy compared to ileum also identified enrichment of RNA splicing and spliceosome processes, suggesting that the rectal-like splicing observed in the spliceopathies is driven by an unknown aberration in the mRNA processing mechanisms of these patient's ileal tissue. However, there was no evidence from splicing, expression, or genotype data for the involvement of any of the three RNA-binding proteins known to influence alternative splicing of *CEACAM1*.[186] Two-way hierarchical clustering of PSI for the top 24 most spliceopathy-affected exons from 17 splicing-related genes showed that the two samples from one individual are clear outliers, while the single aberrant biopsy from the second individual falls within a small cluster of rectal-like ileal samples. These two cases thus likely have different genetic etiologies. It is not possible from this dataset to discern whether a single mutation is responsible for the profiles, or whether a combination of genetic and environmental factors lead to disruption of the splicing of these gene products, which then mediates the broader set of aberrant splice events.

3.5 Discussion

Our results establish that altered splicing is a relevant feature of the IBD gut. Since splicing is to some extent co-regulated with transcription [149,189], covariation of both aspects of gene expression is observed, for example in similarity of the principal components. An appreciable fraction of individuals have more rectal-like ileal expression and splicing also because of alterations in the proportions of epithelial, stromal and immune cells. These differences are to some extent ancestry-biased, notably with elevated stromal (fibroblast) expression in European relative to African Americans (Figure 3.3a,b). This observation extends our recent demonstration of ancestry-related differences in ileal gene expression involving pathways that also associate with disease severity [169]. Our ability to determine the cause and observe the downstream effects of the "spliceopathy" is limited by both the low number of individuals it was observed in and the design of our study. Future research could shed more light on the frequency, effects, and possible cause of this type of aberration by analyzing single cell RNAseq and variants from whole genome sequencing in addition to bulk mRNAseq in the rectal and ileal tissue. It will be important to define the molecular mechanisms responsible for the coordinated splicing defects, and to evaluate whether they suggest personalized therapeutic interventions.

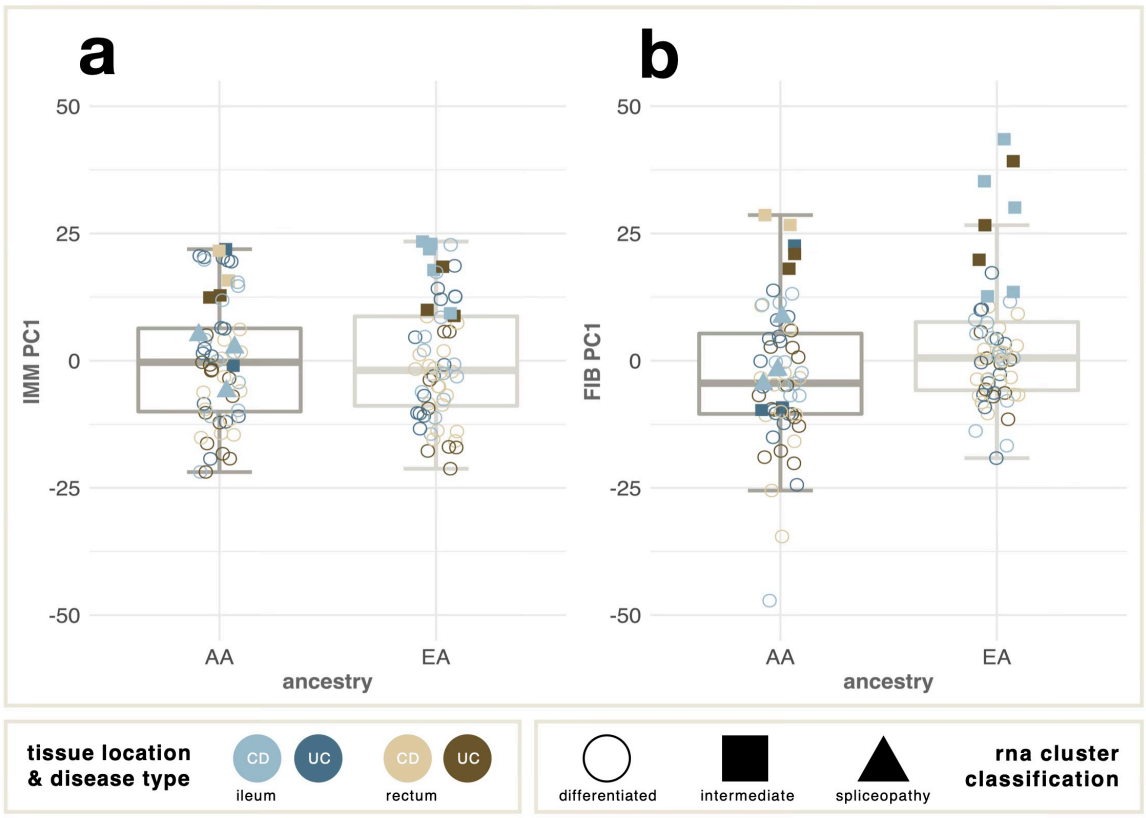

Figure 3.3 Association of ancestry with tissue proportion in biopsies. (a) PC1 of immune specific expression is similar between the two ancestry groups. (b) PC1 of fibroblast specific expression is significantly different between the two ancestry groups, implying a reduced proportion of fibroblasts in the African-ancestry biopsies ($p = 0.005$, 2-tailed t-test). Note that the aberrant intermediate samples have particularly elevated fibroblast expression in both groups, whereas the two "spliceopathy" cases, both African American, have relatively normal fibroblast proportions

A further noteworthy aspect of this study is the development of a pipeline for quantitative analysis of splicing data from RNAseq. The popular MISO (Mixture of Isoforms) protocol [69] incorporates fragment length distributions and exon-level abundance estimates into probabilistic estimation of altered isoform usage, but is intended for single sample comparisons. Several other existing approaches to detection of aberrant exon usage are incorporated into standard RNAseq analysis tools [70,190,191], while an

approach based on identification of novel or cryptic splice junctions in cases compared with controls, led to identification of the molecular basis for 17 of 48 previously undiagnosed neuromuscular disease cases .[62] Here we combine attributes of each of these algorithms along with quantitative evaluation of exon usage to identify suites of concurrent aberrant splicing in outlier individuals. PSI filtering (see Methods) allows for a focus on exons that are not constitutively expressed and therefore contribute to differential isoform usage, without being limited to annotated isoforms. Similar conclusions were observed at a variety of thresholds of inclusion, but careful filtering to rule out artefacts of low expression or exon coverage allowed us to focus on a core set of a few hundred genes perturbed in two cases of spliceopathy.

This study was performed using whole mRNA, which has long been the standard for gene expression analysis and, by extension, exon level and splice site analysis. However, several aspects of whole mRNA sequencing are not ideal for analysis of these smaller features, and a case can be made for targeted RNA-seq when at all possible. Our samples, following Illumina and ENCODE recommendations [192], had a median sequencing depth of 45 million reads. While this provides robust analysis at the gene level, it is quite limiting in terms of how many splice sites can be evaluated with accuracy. After removing exonic bins in genes with low or no coverage (60% of sites), the mean number of informative reads per bin for any given sample is 169 while the average maximum number of reads at a single site is over 750,000. At over half of the remaining sites, the median PSI score across all samples is 0.99, rendering those sites uninformative for differential usage analysis. A careful review of literature and public RNA-seq databases such as GTEX could identify genes that, though highly expressed in the target tissue, are not relevant for the proposed analysis. By using this information to create a targeted RNA-

seq panel, researchers can achieve a higher read depth for a more robust analysis of splicing without needing to increase overall sequencing depth or sacrifice gene level analysis.

3.6 Conclusions

Consistent with previous studies, we found tissue location to be the largest contributor to variability in gene expression and splicing. Though gene expression differences between tissues are often accompanied by changes in splicing, as one might expect since different cell types may utilize different isoforms, neither analysis shows the whole picture on its own. The observation of the ileal samples in two CD patients exhibiting intermediate gene expression but clear rectal-like splicing indicates that differential splicing is a previously unrecognized contributor to IBD disease pathology. Because the aberrations are seen in the full splicing profile rather than a specific aberrant splicing event, we refer to these cases as a “spliceopathy”. Our results indicate that inclusion of splicing analysis when performing RNA-Seq experiments for the study of human disease could play an important role in identifying additional contributions to the pathology of not only IBD, but also other complex diseases.

CHAPTER 4. INTEGRATIVE ANALYSIS OF TARGETED RNA-SEQ IN WHOLE BLOOD INCREASES DIAGNOSTIC YIELD FOR DYSFERLINOPATHY

This chapter has been adapted from a version available on MedRxiv[193] (citation below). Substantial changes were made for clarity and to focus on the RNA-seq analysis and interpretations that I performed. This chapter marks the development and refinement of my approach to rare variant analysis in RNA, and the methods have been expanded accordingly to better detail and explain the rationale and purpose behind each part.

Chakravorty S, **Berger K**, Rufibach L, Gloster L, Emmons S, Shenoy S, Hegde M, Dinasarapu AR, Gibson G. Combinatorial clinically driven blood biomarker functional genomics significantly enhances genotype-phenotype resolution and diagnostics in neuromuscular disease. medRxiv. 2021.

4.1 Abstract

Even after extensive genetic testing, 50-60% of neuromuscular-disease patients remain undiagnosed, hindering precision-medicine and clinical-trial-enrollment. This is due to: a) clinical-genetic-heterogeneity; b) high-prevalence of variants-of-uncertain-significance (VUSs); (c) unresolved genotype-phenotype-correlations for patient stratification, and (d) lack of minimally-invasive biomarker-driven-assays. We therefore implemented a combinatorial phenotype-driven blood-biomarker functional-genomics approach to enhance diagnostics and trial-readiness by elucidating disease mechanisms of a neuromuscular-disease patient-cohort clinically suspected of Dysferlinopathy, the second-most-prevalent LGMD in the US.

Using a panel of 273 genes implicated in neuromuscular disease, we performed blood-based targeted RNA-seq on a subset of 69 cohort patients to validate our integrative

approach to analysis and identify patient variants not seen in DNA, such as splicing, to provide additional diagnoses.

Targeted RNA-seq was highly successful at diagnosing Dysferlinopathy, resolving 63% of RNA-seq cases and improving the diagnostic yield of the overall cohort to 46.7%, an 11% increase over DNA-sequencing. We resolved nearly half of VUSs identified in DNA-seq, observing aberrant splicing in 10 variants. Importantly, the high read depth and consistency of nonsense-mediated-decay in the presence of protein truncating variants (PTVs) allowed for reliable phasing of *DYSF* variants without the need for trio sequencing. Our results show that RNA-seq is a powerful tool for the diagnosis of rare disease and that minimally invasive samples such as blood can be used in place of the disease tissue under certain circumstances.

4.2 Introduction

Limb-girdle muscular-dystrophies (LGMDs) are one of the most prevalent and heterogeneous inherited neuromuscular-disorders (NMDs) with >30 monogenic clinically-overlapping subtypes[194]. Among them, Dysferlinopathy (OMIM 254130, 253601, 606768), a recessively-inherited muscular dystrophy caused by variants in the *DYSF* (MIM 603009) gene[195,196], with variability in clinical presentations [197-199] is the second most prevalent LGMD[194,200,201]. Definitive molecular-diagnosis is typically a pre-requirement to enroll patients with such clinico-genetic heterogeneity into clinical-trials. Recently, in a large LGMD 35 gene-panel next-generation-sequencing (NGS) program, we achieved 27% diagnostic yield[194]. However, 72% of all clinically reportable variants were variants of uncertain significance (VUSs) resulting in ~50% of cohort, including at least 90 patients with *DYSF* VUSs or unresolved compound

heterozygosity without known phasing, remaining undiagnosed, hindering trial-enrollment. Importantly, with upcoming trials of gene-therapy (NCT02710500) and others on the horizon, improved understanding of genotype-phenotype correlations by identifying mechanism of pathogenicity of not only VUSs, but also pathogenic or likely pathogenic variants at the molecular level is essential to stratify patients appropriately for better readiness to clinical trials or precision medicine initiatives.

Rigorous VUS-reclassification and resolution of genotype-combinations per American College of Medical Genetics and Genomics (ACMG) guidelines[32] requires understanding disease mechanisms using an integrative approach combining functional-assays with genotype and phenotype correlation[202-204]. Gene-based or other biomarker testing from easily accessible tissue, such as blood or urine, is needed since muscle biopsies or skin-derived transdifferentiated myotubes are invasive, costly, and adipocyte contamination can compromise quality. DYSF protein-expression in blood, although shown by us to be an effective Dysferlinopathy biomarker[205-208], was not able to provide definitive genotype-phenotype correlations, resolve carrier-range detection for patients clearly Dysferlinopathy-suspected, and was unable to reclassify VUSs. Alternatively, transcriptome-sequencing (RNA-Seq) using patient muscles or myotubes or fibroblasts or blood without in-depth focused clinical-correlation increased diagnostic-yield to a maximum of 36% in clinically diverse cohorts[62,63,74,209].

The *Dysferlin* gene presents as an ideal gene for performing RNA-seq from whole blood because it is overexpressed in monocytes and hence blood[210] and the 14 documented, protein-product producing isoforms are relatively large (55-58 exons) yet have only four exons that are naturally alternatively spliced (Figure 4.1), referred to here as 1, 1a, 5a, 17, and 40a[211,212]. In the disease tissue, skeletal muscle, the main isoform

is NM_003494 which utilizes alternative exons 1 and 17. The primary isoforms expressed in whole blood NM_001130980 (inclusion of exons 1 and 5a) and NM_001130981 (exons 1, 5a, and 40a). Very few pathogenic variants have been identified in any of the alternatively spliced exons, but isoform differences between the analyzed and diseased tissue should be considered carefully when interpreting variants found there.

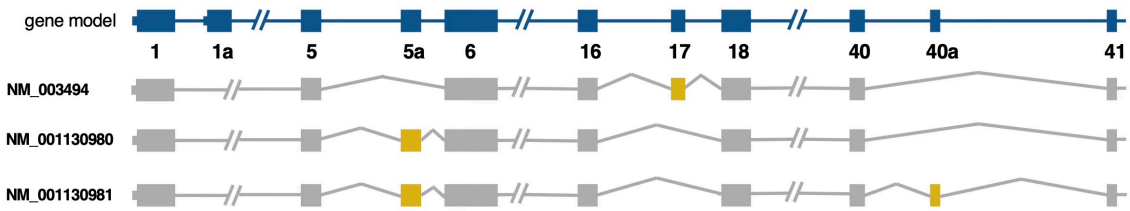

Figure 4.1 The structure of three *DYSF* transcripts. NM_003494 is the primary transcript in skeletal muscle, while NM_001130980 and NM_001130981 are the primary transcripts in whole blood. Only regions with naturally alternatively spliced exons are shown. The gene model represents an aggregate of all possible exons.

We show here that blood-based targeted-RNA-Seq with clinical-correlation does have high resolution when the candidate gene, such as *DYSF*, is adequately expressed as suggested but not shown by two other groups[62,74]. Though blood-based whole transcriptome RNA-Seq on trios (proband and parents) provided 38% diagnostic yield in combination with exome or genome sequencing[213], for adult neuromuscular disorders that are late-onset such as Dysferlinopathy and many other neuromuscular disorders parental DNA/RNA may not be available for segregation studies. Here, using NMD-specific targeted higher depth blood-based RNA-Seq, we show resolution of phases of previously unresolved compound heterozygous variants using allele-expression imbalance (AEI) as was suggested previously[214]. In our large cohort of 364 patients

clinically suspected of Dysferlinopathy or related-LGMD, we used targeted RNA-seq of just 69 cases to raise the diagnostic yield from 34% to 44% and resolve discrepancies between genotype and protein expression in monocytes. We explain how phenotype and functional assays can inform efficient diagnostic testing and illustrate the importance of a combinatorial approach to functional genomics to improve understanding of variant-gene-disease relationships.

4.3 Materials and Methods

4.3.1 Study design

A total of 364 patients of diverse ethnicities (including Americas and Europe) with clinical-suspicion of Dysferlinopathy (LGMD2B/Miyoshi Myopathy/Distal Myopathy with Anterior Tibial Onset/related LGMD) or related-LGMD and 15 normal-control samples from unaffected individuals were recruited between 2016 and 2019 in this study at Emory University based on the following inclusion- and exclusion-criteria. Inclusion-criteria: Patients with Dysferlinopathy or related LGMD clinical-suspicion were selected after comprehensive clinical-evaluation. Exclusion-criteria: Patients with definitive clinical-diagnosis of other unrelated muscular-dystrophy types were excluded in order to target a focused Dysferlinopathy-suspected cohort. Written informed consent was obtained from all participants of the study according to Institutional Review Board approval. Genotype information from prior CLIA-CAP-certified genetic-testing reports for all 364 patients who underwent DNA-testing to identify disease genetic-basis was also collected where available. These genetic tests were heterogeneous, ranging from exome or array-comparative-genomics-hybridization to known variant Sanger-sequencing based on respective physician's discretion. Subsequently, 342 out of 364 patients underwent blood

CD14+ monocyte-assay (MA-assay) for minimally invasive DYSF protein-profile-analysis using immunoblotting.

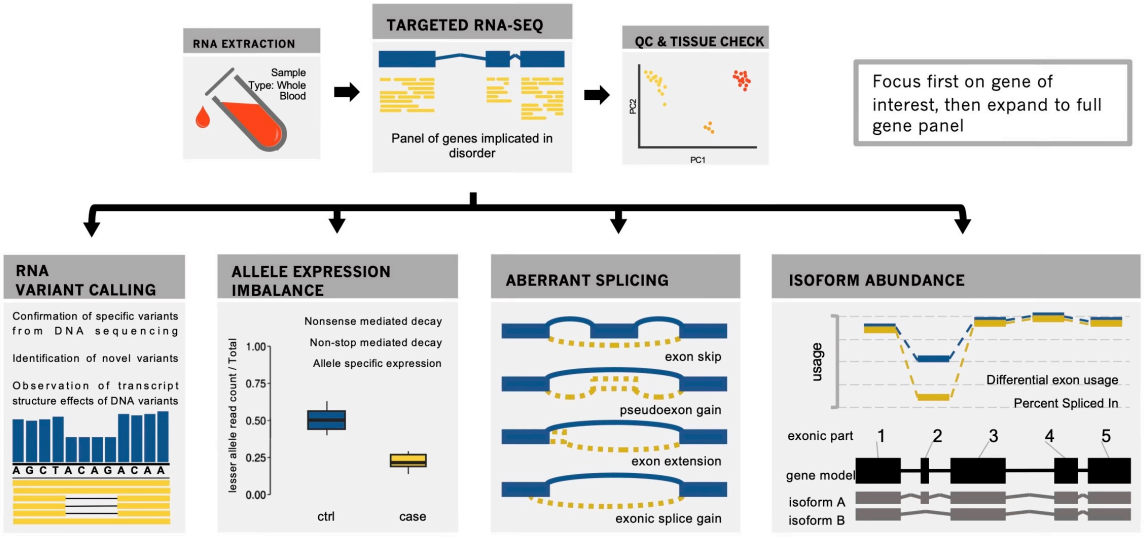

Figure 4.2 Overview of multi-faceted approach to analyzing RNA for the diagnosis of mendelian disease. The information gleaned from each type of analysis is complementary to the others to provide the best support for evidence of pathogenicity.

Thereafter, targeted RNA-Seq using whole blood with an integrative analytical approach (Figure 4.2) was performed on a subset of 69 consenting patients either without complete molecular-diagnosis (48 patients) or without resolution of genotype-phenotype correlation even with confirmed genetic diagnosis in some cases (2 patients). This combinatorial sequential approach enabled understanding the Dysferlinopathy genotype-phenotype landscape. The cohort samples were categorized into 3 general groups based on the results of molecular genetic testing, with a fourth group of negative controls comprised of 8 volunteers with no history of neuromuscular disease. Group A was comprised of samples having two identified heterozygous pathogenic or likely pathogenic

(P/LP) variants (or a single homozygous P/LP variant) by DNA sequencing, so RNA sequencing for these samples served as the positive control group. Group B samples have a single heterozygous P/LP variant or a homozygous VUS. Identification of a second P/LP variant or reclassification of the homozygous VUS to P/LP would result in a molecular diagnosis for these cases. Group C is made up of cases where no P/LP variants have been observed in DNA, requiring identification or reclassification of two variants in order to achieve a molecular diagnosis. Sample numbers for each group are provided in Table 4.1.

Table 4.1 Sample numbers per group for the suspected dysferlinopathy cohort.

	Full Cohort	**RNA-seq**
Group A 2 heterozygous or 1 homozygous P/LP variants in *DYSF*	130	21
Group B 1 heterozygous P/LP variant in *DYSF*	87	30
Group C No P/LP variants in *DYSF*	147	18
Total	364	69

Previous literature using RNA-Seq to diagnose neuromuscular disorders analyzed muscle biopsies and used the publicly available GTEx data as controls for comparison[62,74]. Although GTEx samples are an important resource for comparison, their use as either proxy tissue or as normal controls in RNA-Seq is debatable due to a) potential differences in the methods pipeline and sequencing platform settings, b) possible sample differences in sex, age, and storage conditions since most GTEx sample are from

individuals >40 years age[215]. Therefore, to make our pipeline more clinically cautious and relevant, and to improve analyses, we used internal normal control blood specimens. These samples came from individuals of different ethnicities without any symptoms or individual/family history of neuromuscular or neurological disease, showed ≥100% DYSF protein expression in CD14+ monocytes, and underwent the same sample collection/storage conditions, sequencing platform and overall methods pipeline as patient samples.

Variants were called and evaluated in RNA-seq and data was analyzed for aberrant splicing, allele expression imbalance (AEI), differential exon usage, and transcript abundance. Taking together the results of RNA-seq analysis, available clinical information, %DYSF protein expression in CD14+ monocytes, and DNA-Seq data, we performed phenotype-genotype correlations. The results were clinically correlated to reclassify VUSs, identify pathogenic events at the mRNA level, and understand the pathogenic nature of the variants as per ACMG-AMP guidelines, in order to submit to public databases such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), Human Genome Mutation Database (HGMD: http://www.hgmd.cf.ac.uk/ac/index.php), Human Genome Variation Society (HGVS: https://www.hgvs.org/) and others. All RNA-seq assay results were reported back to patients and/or respective physicians as research reports according to guidelines of the approved Institutional Review Board protocol.

4.3.2 Whole Blood Targeted RNA-Seq Library Preparation and Sequencing

High quality (RNA Integrity Number; RIN>5) RNA was extracted from whole blood of the patients and control individuals using QIAamp RNA Blood Kit (cat # 52304, Qiagen) following the manufacturer's protocol. Various studies have shown that samples with an

RNA Integrity Number (RIN) above 7 show minimal RNA degradation[216,217]. Samples with a RIN below 7 were in some cases still sequenced but were removed from all gene expression analyses as research has not established that it is possible to correct for RNA degradation through normalization to a degree sufficient to support clinical diagnostics.[75,76,216,217] Only blood specimens in EDTA tubes shipped to us within 24hrs from blood draw time based on time log on top of the EDTA vial were used for RNA extraction to control for any time-based RNA degradation effect. Library preparation was performed using SureSelect^{XT} RNA Target Enrichment for Illumina Multiplexed Sequencing kit (cat# G9691- 9000) following manufacturer's protocol. Targeted RNA-Seq was performed to have a more focused clinically relevant platform for neuromuscular disease (NMD) diagnostics and to achieve greater read depth and coverage of the target NMD genes. We used a custom-designed target library probe to capture 273 genes that are known to be NMD-associated and are known to have skeletal muscle expression (≥1TPM) as retrieved from The Genotype-Tissue Expression (GTEx) portal[210]. These 273 genes were curated initially based on the associations to NMDs (Types of Neuromuscular Diseases, http://muscle.ca/wp-content/uploads/2019/08/Disorder_List_ENG_May2017.pdf) as recently done by us[94]. Using the targeted sequencing here ensured higher read depth in our target gene (*DYSF*) while also providing a preliminary characterization of how a targeted panel of neuromuscular disease genes would fare when sequenced in whole blood. Strand-specific sequencing was performed on an Illumina NextSeq instrument to obtain high output 150bp paired-end reads at a depth of more than 15 million reads per sample.

4.3.3 RNA-Seq Analysis

4.3.3.1 RNA-seq pre-processing pipeline

The pipeline for RNA analysis is shown in Figure 4.3. The outputs highlighted in green squares were used for the following five aspects of RNA analysis: Aberrant Splicing, Variant Calling, Allele Expression Imbalance, Isoform Abundance, and Overall Transcript Abundance. The framework of the pipeline was built following GATK best practices for RNA-seq analysis[218] and further refined after comparison of various guidelines and best practices as well as internal testing of various tools. Custom R-Scripts and Codes used for RNA-Seq data-analysis are made publicly available through the Github repository: https://github.com/kiera-gt/rnaseq-nmd.

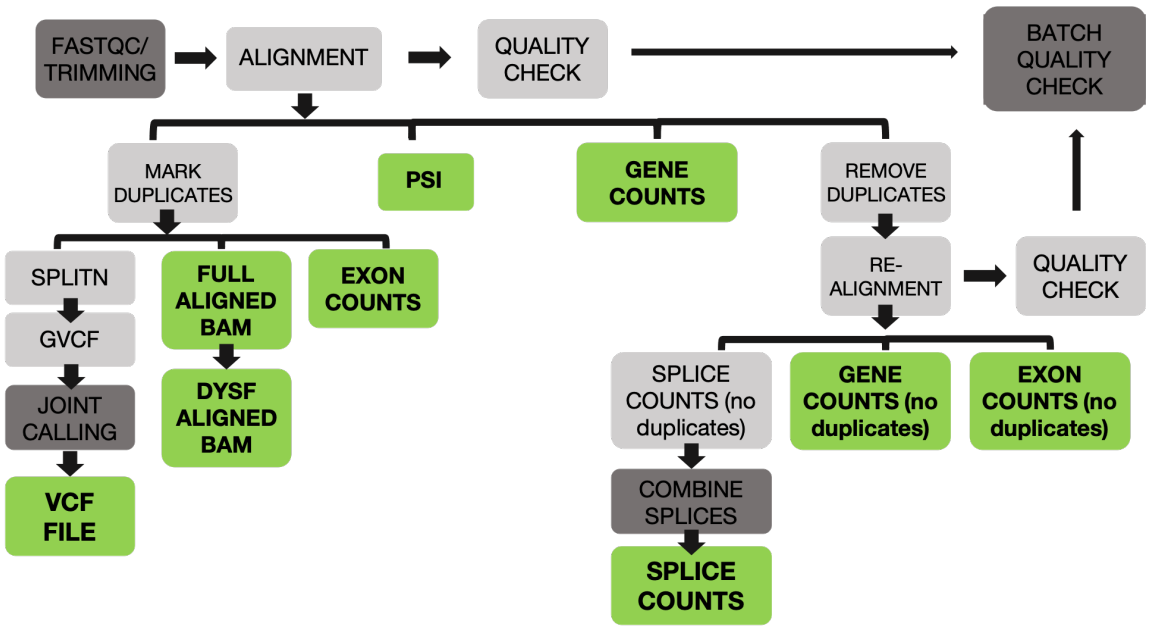

Figure 4.3 Bioinformatics pipeline for the analysis of RNA-seq data. Green squares represent an output file that was analyzed for aberrant splicing, variants, allele expression imbalance, isoform abundance, and/or transcript abundance.

4.3.3.2 Quality Control

Quality Control (QC) measures were implemented at multiple stages of the pipeline. FastQC[171] was run on raw reads and again after adapter trimming. The FastQC tool, like many in this pipeline, was built for the analysis of DNA sequencing reads, but the thorough documentation provides insight into the expectations for performance on whole mRNA sequencing. After alignment, the Quality of RNA-seq Tool-Set (QoRTs)[175] were run on each sample. After all samples completed the individual pipeline, metrics were recorded and QoRTs run for the batch to identify any outliers or possible sequencing issues before running the joint analysis. Alignment metrics tracked for both total reads and only non-duplicate reads include the number of uniquely mapped read pairs, the percent of those reads mapped to the target panel of 273 genes, and the number and percent of reads mapped to *DYSF*. The graphical output of QoRTs produces 36 charts of alignment statistics. While every graph was given a cursory look to check for extreme outliers, the four charts showing cumulative gene assignment diversity and gene-body coverage were reviewed for indications of poor sample quality and 3' or 5' bias, which can influence various aspects of analysis even after applying corrective measures during normalization.[219,220] Finally, a "tissue check" was performed by plotting the first two principal components of gene counts, normalized for library size and regularly log transformed using DESeq2.[221,222] Any outliers or batch effects observed in this and all previous QC steps were noted, and great care taken to account for these issues in downstream analysis steps.

4.3.3.3 Read alignment

Previous research has shown that alignment of RNA performs better and captures more information when reads are not trimmed beyond adapter removal[223].

Trimmomatic[172], a commonly used read trimming tool, was used to remove Illumina universal adapter sequences from reads. Reads were aligned with the STAR[125] splice-aware aligner, using GRCh38 as the reference sequence and NCBI Annotation 106 as the annotation reference. The STAR 2-pass alignment method was used as it consistently achieves high marks in precision and accuracy when tested against other options[224,225]. Aligning to the genome requires split-read alignment which can be more error-prone than transcriptome alignment[65], particularly when a splice event occurs near the beginning or end of an individual read. However, downstream tools perform better and, in some cases, require that RNA reads be mapped to the genome using a splice-aware aligner. Reads were mapped using the same parameters as the ENCODE project,[226,227] with the exception that all multi-mapping reads were removed as the "correct" alignment for these reads cannot be determined reliably.

4.3.3.4 Variant Calling

Duplicate reads were marked in the aligned BAM file, which was then used to call variants in the RNA following GATK Best Practices for Variant Calling in RNA-seq[228] with the alterations described here. The Base Recalibration step in variant calling with GATK is recommended for better calls. However, because we were working with split reads and particularly interested in identifying aberrant splicing, the recalibration and realignment of reads that occurs in this step would alter the original alignment of the reads that was performed with the full (rather than split) reads. Since we examined both the variant calling output and the raw aligned reads, the decision was made to forgo base recalibration to save time and computing energy.

Since many samples were analyzed and variant calling in RNA is especially prone to false positive calls, a GVCF file was produced for joint variant calling across all samples

rather than an individual VCF file for each sample. Joint variant calling using HaplotypeCaller was carried out over all 78 samples undergoing RNA-seq to identify variant calling trends. Variant calling across all of the panel genes was performed with recommended parameters, but genome coordinates encompassing the *DYSF* gene were also called separately with relaxation of parameters that are intended to reduce false positive calls. The rationale behind this is to gain an appreciation, while focusing on a single gene that can quickly be manually checked in depth, of how many and what type of variants tend to be missed by the standard parameters, and then to use this information to inform and alter the analysis framework for expansion to multiple genes in future studies. The joint VCF file was annotated with dbsnp ID, gene location, functional domain information, predicted protein change, population data, conservation and missense *in silico* prediction algorithms including PolyPhen2[229] and SIFT[117], and ClinVar[36] classifications using ANNOVAR[230] to aid in prioritizing variants according to ACMG classification guidelines[32]. Variants were also annotated with the Transcript inferred Pathogenicity Score (TraP Score) to help identify variants predicted to impact splicing[59]. Variants were automatically filtered out if they did not pass a minimum quality threshold of 20 or the alternate allele count was less than 10.

To optimize the efficiency of the manual portions of the analysis, the *DYSF* joint VCF file was checked for variants already identified via DNA sequencing, and the region was visualized in the aligned BAM file using the Integrative Genomics viewer (IGV)[231]. *DYSF* variants called in the VCF file were prioritized for manual evaluation by variant quality and indicators of pathogenicity such as variant type, *in silico* predictor tool results, a previous entry or P/LP/VUS in ClinVar[34], and population frequency in the Genome Aggregation Database (gnomAD)[27]. In addition, the aligned reads were viewed, and

each exon manually checked for any variants not present in the VCF file. Reference and alternate allele counts were recorded for every real variant in every sample for use in AEI analysis described below. High quality loss of function (LOF) variants and variants with an allele frequency of <1% in gnomAD were also pulled for evaluation from the remaining 273 genes in the panel. All potentially causative variants were manually evaluated using the IGV viewer to ensure they were not a result of mis- mapping or noise.

4.3.3.5 RNA-Seq Allele Expression Imbalance (AEI)

To evaluate allele expression across *DYSF*, the allele ratio for each individual high confidence single nucleotide variant (SNV) in each sample was calculated by dividing the read count of the lesser-expressed nucleotide (lesser allele) by the total number of reads at the variant position. Read counts for SNVs passing all filters were obtained from the RNA-Seq variant calling VCF file. Each SNV was grouped by the number of PTVs in the sample it belonged to. Because the data did not pass tests for normality or homogeneity of variance, significant differences between groups were calculated using Wilcoxon rank sum tests and p-values were adjusted using the Benjamini-Hochberg method.

4.3.3.6 Allele Expression Imbalance (AEI) Calculation Method

Only exonic heterozygous SNVs in *DYSF* located in constitutively expressed exons, called at >50X per allele and passing all variant quality filters, were considered in the analysis. Samples that were observed to be outliers by gene expression PCA or did not contain any heterozygous SNVs in *DYSF* were excluded from AEI analysis. Further criteria for sample inclusion was a requirement that the sample contained at least two heterozygous SNVs located more than 150 coding bases apart, to show that the observed AEI is consistent across the entire length of the transcript and that any effect seen is not local to any single variant. A total of 50 samples including 6 controls met the criteria for

inclusion. In each sample, allele ratios were calculated for every SNV meeting the stated criteria by taking the lower number of allele-supporting reads divided by the total available reads at that site for Allele A and the greater number of allele- supporting reads divided by the total site reads for Allele B. In this manner, allele expression is divorced from the concept of “reference” or “alternate” allele. Both AEI and overall gene abundance were correlated with the observation of PTVs in a sample and we attempted to keep the gene abundance observation in the visualization of AEI. Variant call depth is not normalized and varied widely within individuals since depth of coverage is not consistent across all exons, so the calculated allele ratios were instead applied to the overall gene TPM for plotting as a more stable representative of abundance. Each sample is represented by two plotted points showing the average of Allele A and B connected by a line. Error bars represent one standard deviation.

4.3.3.7 RNA-Seq Gene Expression Analysis

Non-duplicate read counts for all genes in the panel were obtained from STAR output files. Transcripts Per Million (TPM) Normalization was performed to control for sequencing depth and to make samples directly comparable. Comparisons of gene expression were performed using Welch’s t-test followed by pairwise t-tests with non-pooled SD. P-values were adjusted using the Benjamini-Hochberg method.

4.3.3.8 Exon usage and splice junction counts

Percent Spliced In (PSI) was calculated for each of the 58 unique exons in *DYSF,* following the methods laid out in Schafer et al[232]. PSI informed the extent of exon skipping, aberrant splice events, as well as usage of the 4 known common variable *DYSF* exons. Raw reads were counted per exon following the DEXSeq[70] vignette and used both to complement PSI analysis and to identify regions of *DYSF* where fewer or more

reads mapped than expected. After removing duplicate reads from the initial alignment, the BAM file was converted back to fastq files and realigned to the genome using STAR with all the same parameters. Counts for splice junctions occurring within the coordinates of the *DYSF* gene were extracted from the STAR splice junction output. These individual counts were combined for all samples at every observed splice site. Raw gene counts for the 273 targeted panel genes were extracted from the complete gene counts output from STAR for use in analysis of transcript abundance.

4.3.3.9 Use of duplicate and non-duplicate reads

There is little consensus in the literature on whether analysis should be performed with duplicate reads or with them removed[233-235]. There are pros and cons to both ways. Reads with duplicates are more heavily influenced by sample quality and sequencing errors are amplified, but they may more accurately represent the true expression of genes or structural events. It was determined that a comprehensive RNA analysis such as this one would benefit from the use of non-duplicate reads in some aspects and a comparison of the two in others. Non-duplicate gene counts and exon counts were obtained in addition to counts with duplicates included, while only non-duplicate counts were used for the identification of novel splice junctions.

4.3.3.10 Identifying aberrant splice events

Viewing an individual sample's PSI calculations across *DYSF* and comparing to both controls and other samples provides very quick identification of exon skipping events. Confirmation of these events as well as identification of other types of splice events is done by observing non-duplicate read counts mapped to novel splice junctions and comparing to the counts in controls and other samples. All possible aberrant splice events were inspected carefully in IGV viewer to ensure they were not a result of mis-mapping or

other sequencing or alignment error. Like the Macarthur group[62], the threshold for considering a novel splice event was that the raw counts are within >10% of the shared splice junction of a known splice event. The criticism of this method is that the threshold of 10% is arbitrary, but the reality is that any statistical test performed here would be similarly subjective. The range of disease severity and symptom variability among dysferlinopathies suggests many factors at play, including the possibility that some causative variants are “more pathogenic” than others.[79,236] Ten percent of reads being affected combined with identification of a nearby variant of interest and axillary clinical information provides multiple lines of evidence and errs on the side of caution for reporting.

Some aberrant splice events and structural changes are not easily seen in PSI calculations and raw splice counts. The STAR alignment algorithm weights the donor and acceptor splice site sequences based on previous research about the likelihood of certain sequences resulting in a splice junction[125]. When a single nucleotide or other variant creates a stronger splice donor or acceptor site than the reference sequence in a region but gets spliced out, STAR is unaware of the DNA change and tries to best-fit the resulting mRNA onto the reference genome. This often results in mapping the reads as an insertion or deletion, since that is statistically more plausible than abnormal splicing, and GATK variant calling may or may not identify the change. Here, again, manually viewing the aligned BAM in IGV viewer is essential. Identifying the mismapping, determining the correct splice junction, and feeding that information to STAR for remapping results in a more accurate final call. Aligning to the transcriptome and using a tool designed to accurately identify insertions and deletions may help find these events easier, but does not remove the manual step of working out the correct splice junction.

4.4 Results and Discussion

All variants observed in DNA sequencing were confirmed in mRNA. We confirmed the 2 *DYSF* variants in all 21 validation cases (Group A) and used the 9 samples containing one protein truncating variant (PTV) and 10 with two PTVs to validate our ability to phase variants using allele expression imbalance (AEI) analysis. In Group B (30 cases), 26 patients received confirmed diagnoses and one patient was brought closer to a diagnosis. In Group C (18 cases), 4 patients received confirmed diagnoses and 5 were brought closer to diagnoses with either *DYSF* (3 cases) or another muscular dystrophy (2 cases with pathogenic variants identified in other genes). Of the 48 cases that were unresolved following DNA sequencing, nearly 63% (30 cases) were diagnosed by RNA-seq. An additional 6 cases (12%) gained information from RNA-seq that brought them closer to a diagnosis. As expected, Group B had a much higher success rate (87% diagnosed) than Group C (22%), given that only one variant had to be newly identified or reclassified for these patients.

In total, 39 VUSs identified in DNA were evaluated in RNA, resulting in the reclassification of 17 (Table 4.2). Ten *DYSF* VUSs were found to alter splicing, thereby reclassifying them as pathogenic (PVS1,PS3-criteria)[32]. Determination that a VUS was in *trans* with a pathogenic variant provided the additional criteria to reclassify 7 variants (1 as benign and 6 as LP).

We note that of 20 P/LP *DYSF* variants solely identified in mRNA, 7 were exonic and 3 located within 10bp of the start or end of an exon but were not reported in DNA-seq results. The fact that prior DNA-testing was performed non-uniformly (Exome/Gene-Panel/Sanger) based on respective physician discretion reflects the real-world heterogeneous scenario of clinical genetics diagnostic requisitions from our experience.

Low cost, high coverage targeted RNA-seq may provide an alternative to WGS that is more likely to be ordered due to its similarity to traditional NGS panels.

Table 4.2 Variants of Uncertain Significance (VUS) reclassified by RNA-seq analysis.

Exon/Intron	Position	Variant Type	New Class-ification	Evidence Type
Exon 5	c.401C>T	missense	B	Phasing
IVS 9	c.907-3C>A	extended splice	P	Aberrant splicing
Exon 19	c.1668_1669insGTT	nonframeshift insertion	LP	Phasing
IVS 25	c.2643+5G>A	extended splice	P	Aberrant splicing
Exon 28	c.3031G>C	cryptic splice	P	Aberrant splicing
Exon 29	c.3113G>C	missense	LP	Phasing
Exon 37-40	r.3904-4410del	gross deletion	P	Aberrant splicing
Exon 37	c.4003G>A	missense	LP	Phasing
IVS 39	c.4334-3C>A	extended splice	P	Aberrant splicing
Exon 43	c.4974G>T	missense/leaky splice	P	Aberrant splicing
IVS 45	c.5057+5G>T	extended splice	P	Aberrant splicing
Exon 47	c.5296G>A	missense	LP	Phasing
Exon 49	c.5503A>G	cryptic splice	P	Aberrant splicing
IVS 49	c.5526-7T>G	extended splice	LP	Aberrant splicing
Exon 53	c.6056G>A	cryptic splice	P	Aberrant splicing
Exon 54	c.6196G>A	missense	LP	Phasing

4.4.1 Identification of aberrant splicing resolving geno-phenotype relationship

Out of 69 patients who underwent RNA-seq, 32 exhibited aberrant splicing, indicating this is a major contributor to pathogenicity for *DYSF*. Three patients (B1, B5, and B6), all with only one previously identified P/LP variant, were diagnosed by the identification of varying-sized large-deletions all spanning entire *DYSF* exon 52 (Figure 4.4),causing primarily a complete exon 52 skip with a small-percentage of transcripts skipping both exons 52 and 53. Nonsense-mediated decay (nmd) was also observed. Though exon 52 deletions have been previously identified[237], this is the first report of the three novel deletions causing similar splicing defects. AEI analyses demonstrated that these deletion variants are *in trans* with the second exonic *DYSF* pathogenic variants found in all three cases.

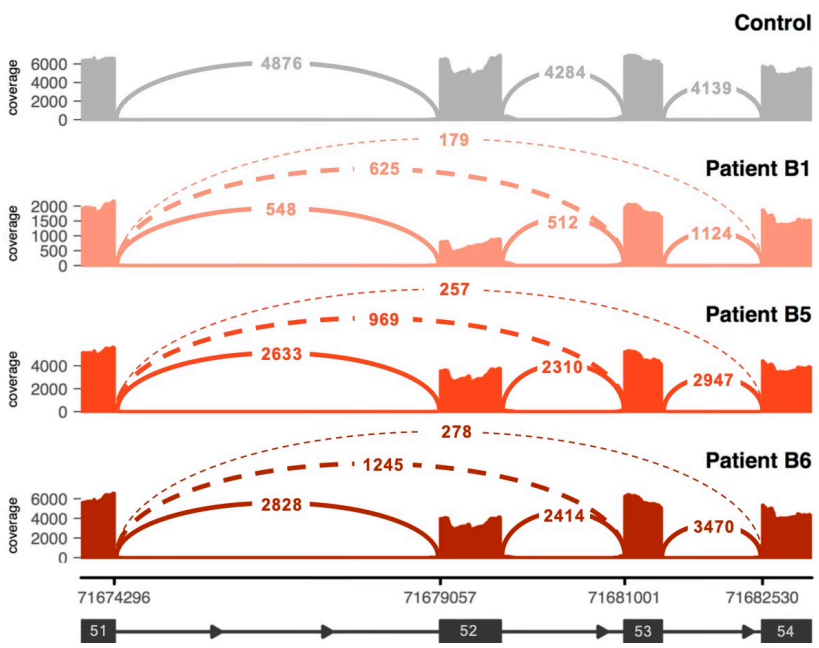

Figure 4.4 Sashimi plot of the exon-skipping event seen in patients B1, B5, and B6. Subsequent genome sequencing (GS) identified the cause as a gross deletion encompassing exon 52. Numbers within continuous and dashed lines indicate number of spliced transcripts. solid lines indicate reference splicing, while dashed lines show observed aberrant splicing.

RNA-Seq can also differentiate between multiple splice events occurring at the same locus. Patient B18 had only one *DYSF*-variant (c.2810+1G>A)[196] reported from DNA-seq and an absence of muscle Dysferlin. RNA-Seq revealed an additional aberrant splicing event using an alternate splice acceptor site in intron 26 that resulted in a 67 bp extension of exon 27 (Figure 4.5). Because these reads use the canonical exon 26 splice donor site, it is unlikely they are a consequence of the c.2810+1G>A variant. Rather, we identified that another intronic variant (c.2811-20T>G) seen only in the aberrantly spliced mRNA reads, disrupts the branch point sequence and leads to preferential use of the alternate splice-acceptor site. This data indicates that the previously identified variant (c.2810+1G>A) and the novel branch-point variant (c.2811-20T>G) are *in trans*. Nonsense-mediated decay due to frameshift and premature protein-truncation of both splice-events is supported by reduced transcript abundance and the observed absence of Dysferlin protein.

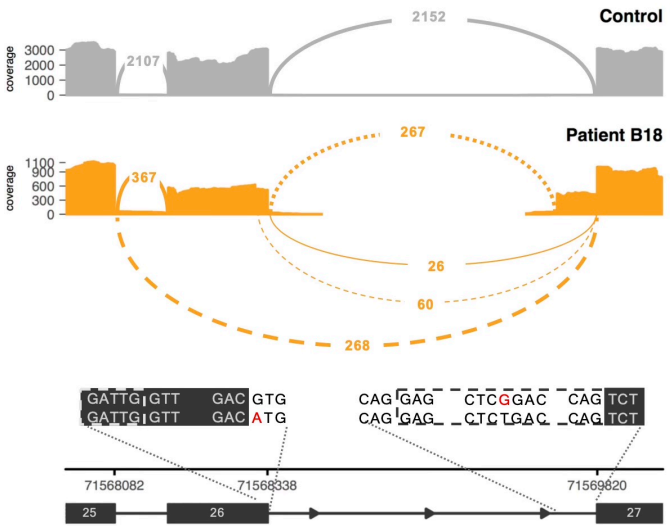

Figure 4.5 Multiple splicing events in a single intron. In patient B18, RNA-Seq identified exon 26 skipping caused by destruction of the splice donor site by the essential splice site IVS26 variant (left inset) and the exon 27 extension caused by a novel branch point variant (right inset). Numbers within continuous and dashed lines indicate number of spliced transcripts. solid lines indicate reference splicing, while dashed lines show observed aberrant splicing.

RNA-Seq allowed us to observe mRNA structural effects that DNA-sequencing alone can miss. Patient B23 RNA-Seq analysis showed that the exon 49 c.5503A>G *DYSF* variant, predicted to be a missense variant, in fact activates a cryptic splice site, resulting in a 23 bp deletion (Figure 4.6).

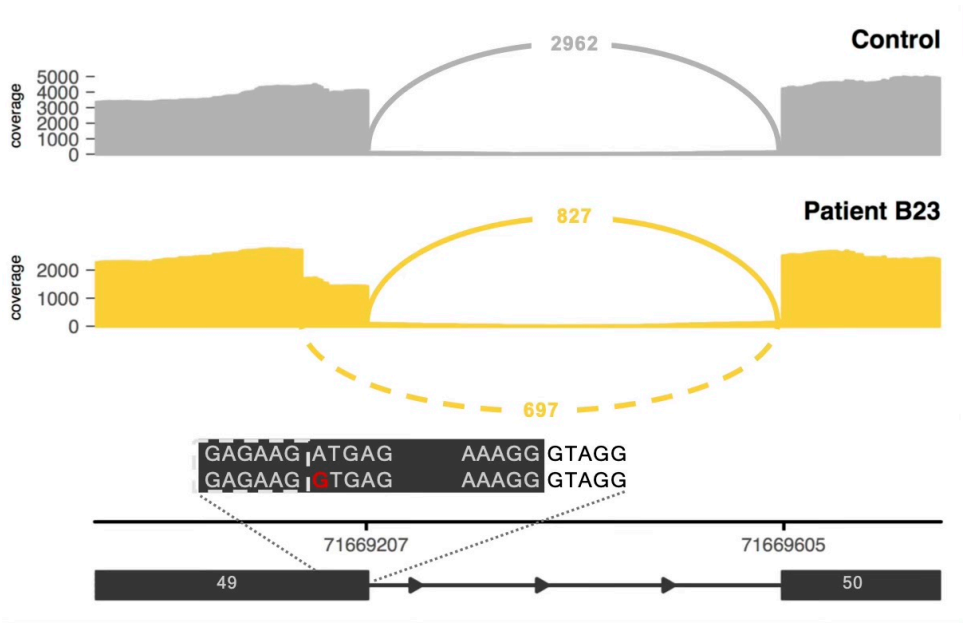

Figure 4.6 Predicted missense variant activates cryptic splicing. Sashimi plot of cryptic splice site variant NM_003494.3:c.5503A>G (inset) in patient B22 which leads to a premature stop codon in the transcript. Numbers within continuous and dashed lines indicate number of spliced transcripts. solid lines indicate reference splicing, while dashed lines show observed aberrant splicing.

4.4.2 Allele expression imbalance (AEI) as a tool to phase variants in adult NMD

In previous studies, nonsense-mediated decay has prevented some protein truncating variants (PTVs) from being called in mRNA[238]. However, the extremely high read-depth by our targeted RNA-Seq panel, even without using any nonsense-mediated-decay inhibitor, allowed us to not only confidently call PTVs, but also use it to

phase *DYSF* variants without parental or offspring sequencing which are not readily available for later onset adult neuromuscular diseases such as Dysferlinopathy. As noted by Cummings et al.,[62] the greater:lesser allele ratio in genes where a sample had one truncating SNV variant tends to be ~75:25. We observed the same trend in SNVs and further noted that deletions/insertions, splice events, and variants in variably expressed exons somewhat follow the same pattern, but with less accuracy due to known issues in reliable mapping/calling these variants. We correlated this ratio with overall transcript abundance of *DYSF*, showing that the differences in transcript abundance among these samples is predominantly a result of nonsense-mediated decay (Figure 4.7). The allele expression imbalance in samples with one truncating variant was so consistent across exons and samples that it could often be reliably used to phase variants. All variant calling tended to show a slight bias towards the reference allele, so we set two levels of confidence (likely and very likely) with ratio thresholds of 60:40 and 70:30.

AEI analysis was able to phase *DYSF* variants in 32 cases, aiding in mechanistic resolution and diagnosis. Within each sample, the allele ratio was found to be consistent for single nucleotide variants (SNVs) across *DYSF* gene. When SNVs were grouped by number of PTVs found in *DYSF*, we found that the lesser-expressed nucleotide expression (lesser allele expression) in patients with one PTV was significantly reduced to ~25% ($p=7\times10^{-13}$) as a result of nmd. For example, in patient A2, the nonsense variant c.331C>T (p.Q111X) in exon 4 is called in 168 out of 993 total reads at that site (~17%). Three other variants were called in the mRNA of patient A2: synonymous SNP c.1827T>C (p.D609D) in exon 20 (1410 of 1680 total reads, 84%), synonymous SNP c.2583A>T (p.S861S) in exon 25 (1465 of 1839 total reads, 80%), and the pathogenic/likely pathogenic missense variant c.6124C>T (p.R2042C) in exon 54 (1535 of 2000 total reads,

77%). Based on the observed AEI, we can be confident that all three of the other variants are in trans with the exon 4 nonsense variant. This method of phasing has the advantage of not requiring trio sequencing but is of course not perfect and should be used cautiously. In patients with biallelic PTVs, both transcripts were subject to nmd and SNV allele-ratio generally returned to 0.5 (Figure 4.7). In these samples we cannot phase each individual SNV in DYSF but can still reliably determine that the PTVs are in trans with one another by extrapolating knowledge of nmd from cases with 1PTV and using transcript abundance as supporting evidence.

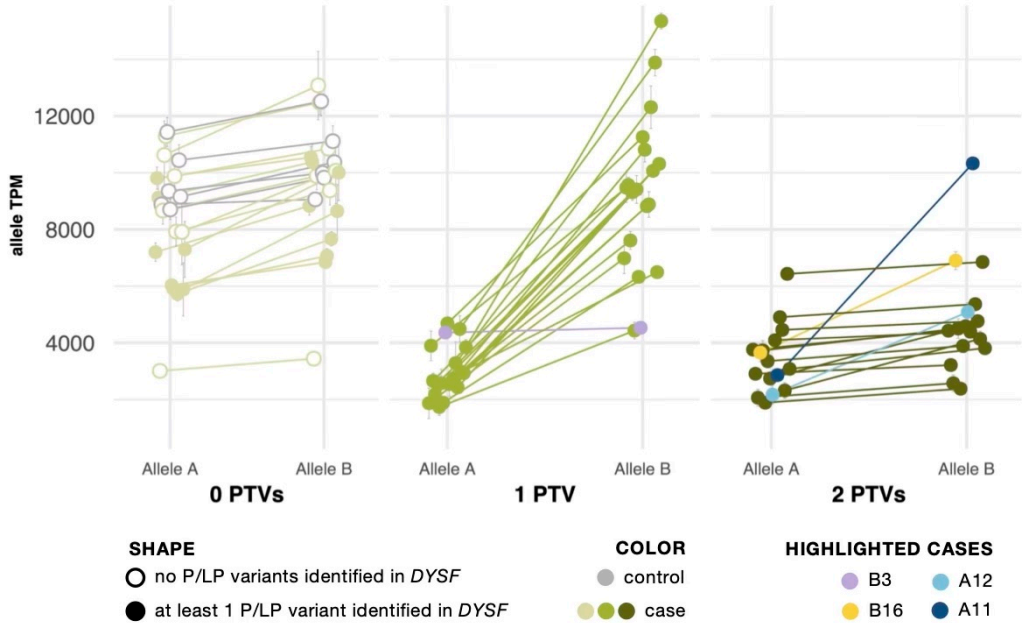

Figure 4.7 Allele expression imbalance caused by nonsense-mediated decay of transcripts containing a protein truncating variant (PTV). Samples are grouped by the number of PTVs observed in DYSF mRNA. See Methods for calculation details. In each sample, the average percent and standard deviation were taken for Allele A and Allele B and mapped onto the sample's overall gene abundance to estimate the abundance of each DYSF allele copy. Cases with one PTV show normal expression of Allele B and reduced expression of Allele A (a result of nonsense-mediated decay acting upon the transcript copy containing the PTV). Cases with bi-allelic PTVs show a reduction in both copies. Highlighted cases (lavender, yellow, teal, navy) are examples of the need to use caution in the interpretation of AEI (see text)

In patient B19, RNA-Seq determined that the c.5296G>A (p.Glu1766Lys) VUS (PM2, PP3, PP4 criteria[32]) is in *trans* with the nonsense pathogenic c.4090C>T (p.Gln136*) variant based on the inverse relationship of AEI ratios, and hence reclassified the c.5296G>A variant as pathogenic (PS3, PM3 criteria). Five other VUSs were able to be reclassified as likely pathogenic in this manner (Table 4.2).

In sample B2, phasing of *DYSF* variants had the opposite result. Reduced Dysferlin staining was seen in muscle. NM_003494.3:c.5022delT and c.401C>T (p.Pro134Leu) *DYSF* variants, previously reported as associated with reduced Dysferlin staining[239] were found to be in *cis*, indicating that c.401C>T is not pathogenic. RNA-Seq identified a different missense *DYSF* VUS [c.6196G>A(p.Ala2066Thr)], previously reported in the homozygous-state in a patient with <5%DYSF[238], in *trans* with the NM_003494.3:c.5022delT deletion, leading us to reclassify c.401C>T as benign, c.6196G>A as pathogenic, and complete Dysferlinopathy diagnosis.

There are of course some caveats to using RNA-seq data for variant phasing. Only one PTV was identified in Patient B3, but the observed AEI more closely matches that of samples with zero or two PTVs. Because the truncating variant seen is located in the last exon of *DYSF*, it is possible that this variant does not lead to nonsense-mediated decay[240]. Another possibility is that there is a second PTV that we were unable to identify, since the overall *DYSF* mRNA abundance is also reduced compared to controls. Subsequent GS found several rare deep intronic *DYSF* variants, most notably an ultra-rare NM_003494.3:c.4509+1586dupG that could possibly cause a regulatory effect, but pathogenicity confirmation could not be achieved. In addition, cases with PTVs affecting alternatively spliced exons (A11, A12) need to be interpreted in that context rather than following the ratio exhibited by other samples. Finally, when a case inexplicably deviates

from the observed pattern (B19), the determination that the variants are *in trans* cannot reliably be made.

4.4.3 Concordance of Monocyte Assay with mRNA Transcript Abundance and Genotype

While muscle Dysferlin protein expression has been found to correlate well with that in blood CD14+ monocytes[205] the same is not true for *DYSF* mRNA-expression from whole-blood, which correlates rather better with the number of *DYSF* PTVs in the sample (Figure 4.8). PTV variants lead to nmd thus decreasing overall gene expression. As a group, samples containing biallelic PTVs exhibited >2-fold decrease in *DYSF* mRNA expression compared to those with no PTVs (Figure 4.8, $p=1\times10^{-10}$). Samples containing just one-PTV also exhibited a decrease in expression, though to a lesser degree (Figure 4.8, $p=1\times10^{-6}$). The discordance between mRNA- and protein-expression is likely explained by pathogenic missense variants and non-frameshift splicing events leading to protein non-functionality or degradation rather than mRNA-decay. For example, patient C13, with <10%DYSF, has a homozygous intronic extended splice-site VUS (IVS49:c.5526-7T>G) which causes an in-frame insertion of two serine residues that does not affect *DYSF* mRNA expression but abolishes DYSF protein expression. This type of resolution to know at what biological level (RNA or protein in this case) the pathogenicity of the variant acts is important to use for patient stratification to reduce variability in responses to clinical trials.

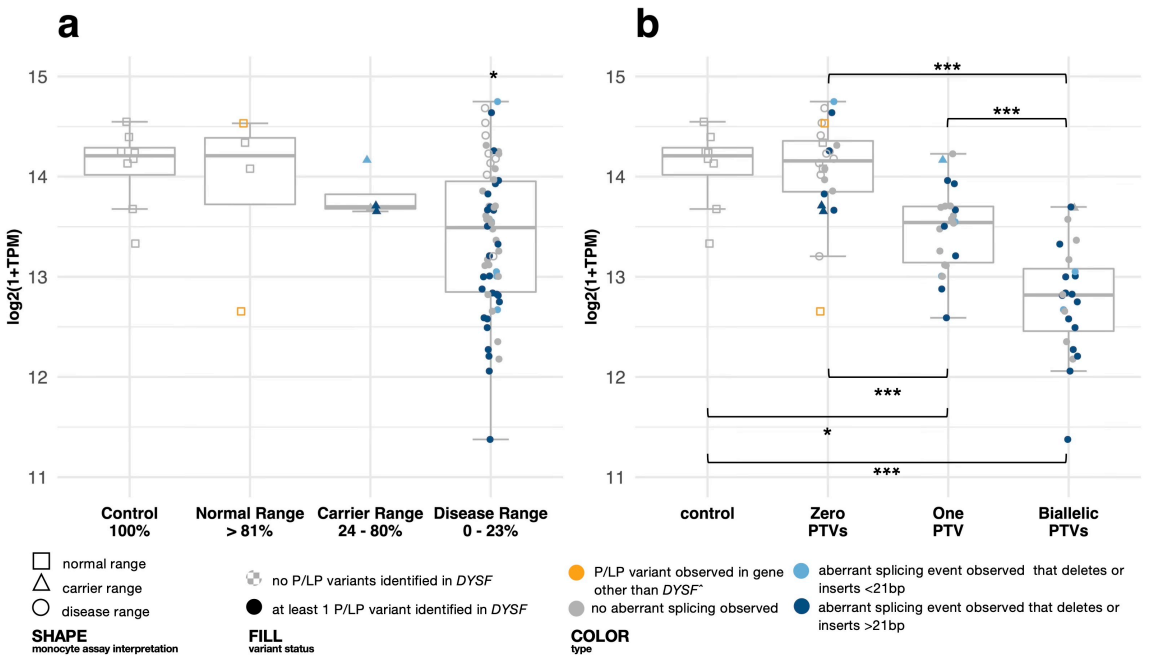

Figure 4.8 *DYSF* mRNA expression. *DYSF* mRNA abundance in Transcripts per Million (TPM) categorized by (a) protein expression in monocytes and (b) number of protein truncating variants (PTVs) observed in the sample. (a) mRNA abundance does not correlate well with protein abundance in monocytes. N=77; control=8, normal range=4, carrier range=4, disease range=61. One-way ANOVA: F=3.3, p=0.03. (b) mRNA abundance correlates well with the number of *DYSF* PTVs observed in a sample, showing that nonsense-mediated decay is the single largest factor acting post transcription and pre-mRNA processing but prior to translation. Of interest, gross insertions or deletions caused by aberrant splicing (marked by navy blue, whether protein truncating or not), do not appear to affect mRNA abundance differently than other variants of a similar type (PTV vs PTV, non-PTV vs non-PTV). N=78; control=8, zero PTVs=24, one PTV=23, biallelic PTVs=23. One-way ANOVA: F=33.3, p=1e-13. Post-hoc comparisons (t-tests with Bonferroni correction for multiple tests): p(control vs 1 PTV)=0.006, p(control vs 2 PTVs)=0.0008, p(0 PTVs vs 1 PTV)=9.2e-05, p(0 PTVs vs 2 PTVs)=1.1e-09, p(1 PTV vs 2 PTVs)=0.0002.

Importantly, RNA-Seq also allowed better patient stratification for those who expressed carrier-range %DYSF. In patients A11 and A12, RNA-Seq helped explain carrier-range expression despite presence of two P/LP variants. The carrier-range %DYSF seen in these patients is due to the natural in-frame skip of exon 17 in the vast majority of blood *DYSF*-transcripts, which results in low expression of their exon 17 or

IVS16 PTVs in blood. However, presence of exon 17 in most muscle-tissue RNA-transcripts results in premature stop and lack of Dysferlin protein expression, leading to Dysferlinopathy. Overall, these results show the utility of RNA-Seq in identifying novel functional variants and aid in our understanding of the nature of their pathogenicity to understand clinical severity.

Though extreme deficits in DYSF protein expression are common in Dysferlinopathies[207], some patients with a clinically-consistent Dysferlinopathy phenotype show only moderate loss of DYSF protein expression[205]. Patients B10 and C6, both with carrier-range %DYSF, were found to share a VUS in *DYSF* exon 43 (c.4794G>T, p.Lys1598His), which has been reported previously in a case with amyloid deposits[241]. RNA-Seq showed that this is a “leaky” splice variant, resulting in a complete skip of exon 43 in approximately half of the affected reads (Figure 4.9), disrupting the DYSF C2F domain and likely causing DYSF mis-folding or deposition in the amyloids seen in this patient muscle. DYSF and amyloid-deposit co-localization is known to be associated with Dysferlinopathy[242,243]. Also, these patients showed milder/slower clinical presentation/progression (muscle-weakness onset after age 45yrs, considerably later than typical onset between 17-20yrs) compared to other Dysferlinopathy patients. Clinical course difference is possibly due to the residual 56-57% of normally spliced *DYSF*-transcripts, allowing mechanistic understanding, patient stratification and better trail-readiness.

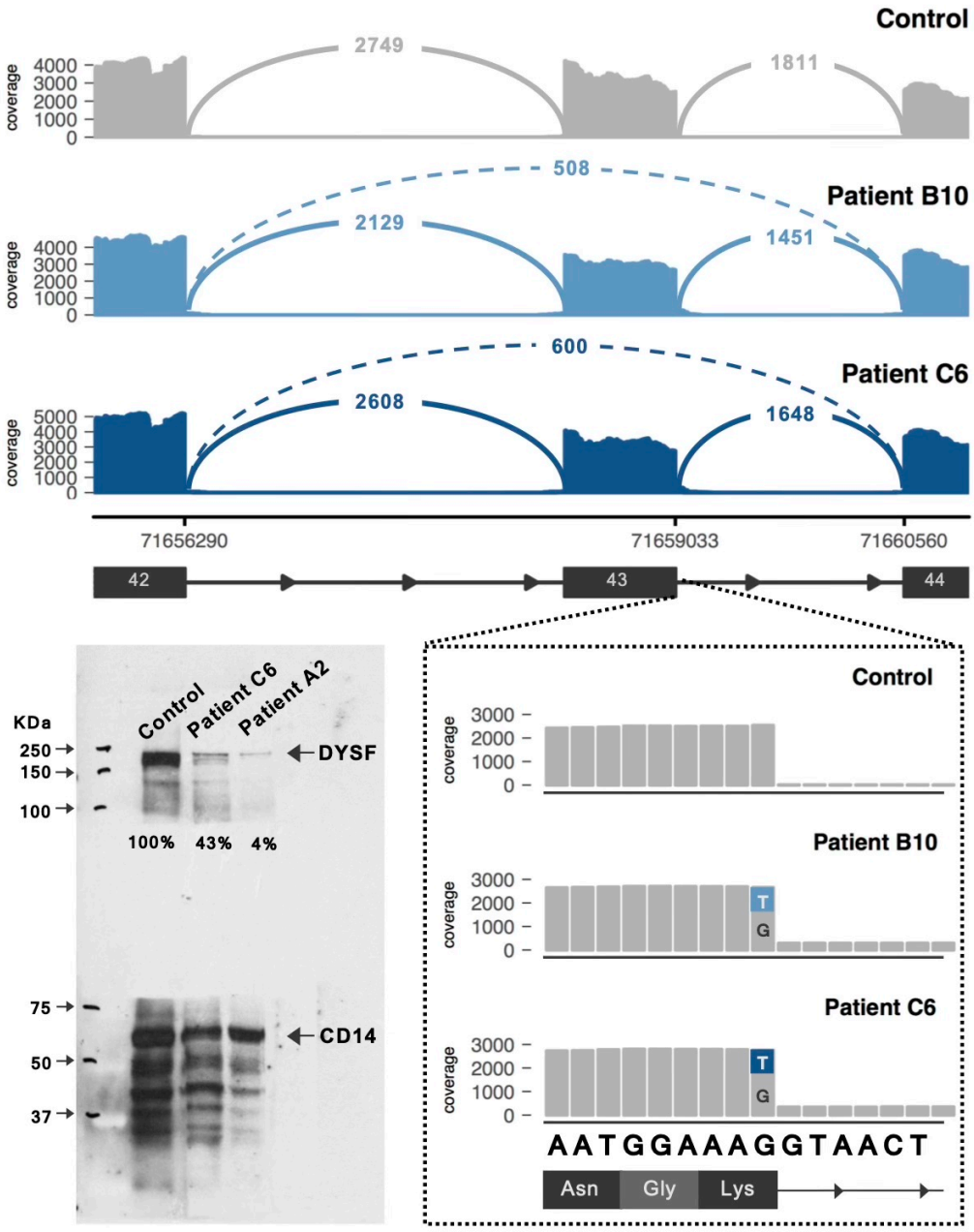

Figure 4.9 Leaky splice variant associated with carrier range *DYSF* protein expression in monocytes. Upper: Sashimi plot of exon-skipping caused by a leaky splice variant in exon 43 (Patients B9 and C6) that is reclassified as pathogenic. Inset: Variant expression in a minority of reads, which show a normal splice pattern. Lower Left: monocyte assay showing reduced DYSF protein expression in patient C6 (patient A2: disease-range reference). Numbers within continuous and dashed lines indicate number of spliced transcripts. solid lines indicate reference splicing, while dashed lines show observed aberrant splicing.

4.4.4 Impact of RNA-seq on Overall Diagnostic Yield

RNA-seq made a significant improvement to diagnostic yield despite being performed for less than a quarter of the cases traditional DNA-seq was unable to resolve (Figure 4.10). Prior to RNA-seq, a diagnosis of dysferlinopathy based on clinical presentation and the presence of two P/LP variants in *DYSF* was made for 130 patients (35.7%). Of the remaining 234 unresolved cases, 48 underwent RNA-seq resulting in an additional 30 diagnoses, bringing the diagnostic yield to 43.9%. The DNA variants for the remainder of the cohort were reanalyzed after the reclassification of 17 VUSs in RNA-seq, providing a diagnosis for 10 more patients for a total diagnostic yield of 46.7%.

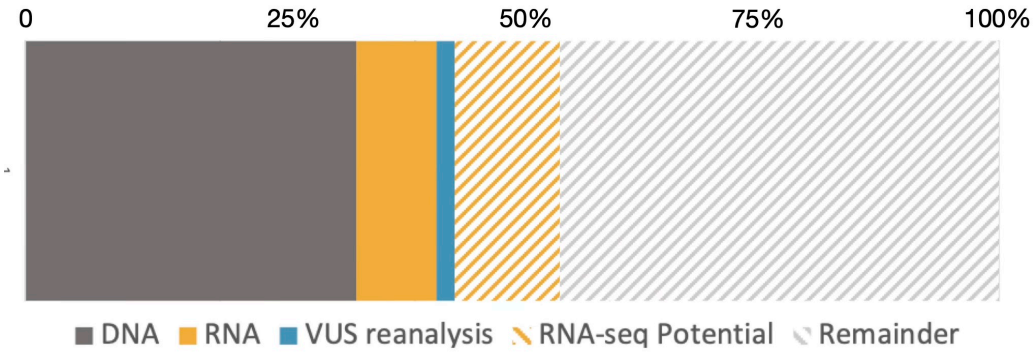

Figure 4.10 Actual and potential diagnostic yield of the full cohort of 364 patients suspected of Dysferlinopathy.

As noted earlier, RNA-seq was more successful at achieving diagnosis when 1 P/LP variant had already been identified in DNA, indicating that patients fitting this description would greatly benefit from RNA-seq. If that rate holds true and we were to sequence the remaining 52 patients with 1 known P/LP variant, the diagnostic yield for the full cohort could increase to 59%. The correlation of disease-range monocyte assay to genotype

suggests it could be used to select patients for RNA-seq even in the absence of DNA sequencing results. Among the 219 disease-range cases, 76% (166/219) had two *DYSF* pathogenic variants (Table A.1) at the conclusion of this study, whereas among 64 carrier-range and 59 normal-range cases, only 8% (5/64) and 2% (1/59) respectively had two *DYSF* pathogenic variants indicating the MA-assay's robustness and that Dysferlin protein absence is highly suggestive of Dysferlinopathy. When aberrant splicing is a substantial contributor to pathogenesis for a disease, such as it is *DYSF* in dysferlinopathy, no form of DNA-sequencing including WES or WGS can provide the diagnostic value that we see with targeted RNA-seq.

4.5 Conclusion

We have clearly demonstrated the clinical utility of blood-based targeted RNA-seq to elucidate variant pathological mechanisms, to understand genotype-phenotype correlations, and significantly enhance diagnostic yield in Dysferlinopathy. Importantly, such analysis enables greater patient stratification which in turn increases readiness for clinical trials and precision medicine initiatives for neuromuscular and other genetically based disorders. In this new genomics era, the intersection of clinical genetics and research genetics based tools needs to be considered to identify more efficient indicators for disease mechanisms, diagnostics, biomarkers and therapy. While muscle is the main target tissue for NMD evaluation, it is possible to use an integrative analysis approach, such as done for this study, for NMD genes that are also adequately expressed in blood or to find other biomarkers that may expedite the diagnostic odyssey NMD patients face.

CHAPTER 5. TARGETED RNASEQ IMPROVES CLINICAL DIAGNOSIS OF VERY EARLY ONSET PEDIATRIC IMMUNE DYSREGULATION

Berger K, Arafat D, Chandrakasan S, Snapper S, Gibson G. Targeted RNA-seq improves clinical diagnosis of very early onset pediatric immune dysregulation. Journal of Personalized Medicine. *Submitted.*

5.1 Abstract

Despite increased use of whole exome sequencing (WES) for the clinical analysis of rare disease, overall diagnostic yield for most disorders hovers around 30%. Previous studies of mRNA have succeeded in increasing diagnoses for clearly defined disorders of monogenic inheritance. We asked if targeted RNA-sequencing could provide similar benefits for primary immunodeficiencies (PIDs) and very early-onset inflammatory bowel disease (VEOIBD), both of which are difficult to diagnose due to high heterogeneity and variable severity. We performed targeted RNA-sequencing of a panel of 260 immune-related genes for a cohort of 13 patients (7 suspected PID cases and 6 VEOIBD) and analyzed variants, splicing, and exon usage. Exonic variants were identified in 7 cases, some of which had been previously prioritized by exome sequencing. For four cases, allele specific expression or lack thereof provided additional insights into possible disease mechanisms. In addition, we identified 5 instances of aberrant splicing associated with 4 variants. Three of these variants had been previously classified as benign in ClinVar based on population frequency. Digenic or oligogenic inheritance is suggested for at least two patients. In addition to validating the use of targeted RNA-sequencing, our results show that rare disease research must find a way to incorporate contributing genetic factors into the diagnostic approach.

5.2 Introduction

Under the general umbrella of personalized medicine, precision genomic medicine refers to investigations designed to diagnose the molecular cause of a clinical condition[244]. Whereas large biobank projects such as the TOPMed Precision Medicine Program[245] focus on gathering genomic and phenotypic data to elucidate patterns in populations that will allow researchers to develop risk predictions, clinicians must be able to design therapeutic interventions tailored to the individual's genetics. As sequencing costs have come down, the focus has shifted from individual diagnostic odysseys involving trains of tests, to whole exome and genome sequencing[15]. Such studies of single individuals or families has typical diagnostic yields between 30% and 50% of patients, in many cases leading to more actionable information than that available from phenotypic considerations alone[10,15,24,25,42,46,49,50,62,246,247]. Precision medicine is also often lauded as the path towards targeted therapies for rare diseases that have previously not received much research funding or attention. Although the vast majority of success stories in precision medicine-based therapies are related to cancer[248,249], inherited immune disorder research is also beginning to yield a number of successful targeted therapies as well[250]. Here we ask whether complementation of exome sequencing with targeted RNA sequencing can increase diagnostic yield in this context.

Primary immunodeficiencies (PIDs), also termed inborn errors of immunity (IEI), encompass over 400 distinct disorders related to immune dysregulation[81,82,84,251]. This includes susceptibility to infection or malignancy, autoimmune and autoinflammatory disorders, and allergies. PIDs have historically been branded as monogenic disorders with traditional Mendelian inheritance[252], but as more PIDs have been identified and we have gained better understanding of immune function and clinical pathogenesis, it has been

recognized that these disorders often display variable penetrance and severity[253]. The term complex immune dysregulation syndrome captures the idea that many cases of aberrant immune activity that share similar presentation nevertheless may have heterogeneous, and sometimes oligogenic, causes[254].

Adult-onset Inflammatory bowel disease (IBD) is known to be a complex disease involving multiple genetic and environmental factors that leads to over-activation of the inflammatory response in the gastrointestinal tract[86,166]. Pediatric IBD makes up a quarter of all diagnosed IBD, and a subset of these cases occur in patients <6 years of age[255]. These are termed very-early onset IBD (VEOIBD) and are generally thought to have a simpler genetic basis[90,92,256]. VEOIBD has also been associated with PIDs, both through symptoms as well as gene involvement[91,92,165,254,257]. Previous exome studies of patients with VEOIBD have identified monogenic causes for a small percentage, but most cases remain without a genetic diagnosis. Monogenic cases of VEOIBD are more likely to have family history of IBD or immunodeficiencies and to be more severe and resistant to conventional treatment[92,256]. Research has also found patients with VEOIBD to have a higher rate of variants in genes associated with PIDs[257], suggesting that some cases for which a monogenic origin is not identified may have a multi-genic etiology.

For both PIDs and VEOIBD, identifying the specific underlying genomic cause is important for treatment[163,165,252,254]. Gene specific targeted therapies enable improved patient management and have been successfully used for several immune conditions[83,84,244]. This group of highly heterogenous disorders often exhibit cascading effects where multiple genes or pathways are pathogenically altered. In some cases, targeting an affected but not causal gene results in worse patient outcomes[254].

Since diagnostic yield from exome sequencing remains well under 50%, the accessibility of immune cells for genomic profiling of peripheral blood samples obtained by standard phlebotomy, raises prospects of RNA-based analyses, specifically RNAseq, that might identify aberrant molecular events such as altered splicing or gene expression[43,51,62,74,149,258]. The relevant mutations might not be observable in exome sequences, or may be of uncertain significance.

Previous cohorts of WES for the diagnosis of PIDs and VEOIBD have prioritized genes known to be associated with immune disorders[252,256,257]. Recent studies have shown that periodic reanalysis of variants results in additional diagnoses[259-263]. This is lauded as a benefit of whole genome (WGS) and whole exome (WES) sequencing, allowing for genes newly associated with a disease to be reconsidered for patients. While most WES and WGS reanalysis reports an increase in diagnostic yield of around 12% with the majority coming from new gene-disease associations, numerous studies have found that for a given disease, patients harbor causative variants in genes that have already been identified for that disorder or group of disorders[264,265]. Novel disease gene discoveries are still happening at a high rate, but there is still a high burden of variants of uncertain significance (VUS) and little incentive to systematically resolve them[266]. Low cost and high read depth motivate the development of targeted panels, which have been used at the DNA level in clinics for several decades now, but are just beginning to be considered for RNA analyses.

RNAseq complements WES and WGS by providing evidence of mRNA effects (or lack thereof) for specific variants as well as identifying alterations in splicing or other structural changes that DNA sequencing methods cannot see[43]. In large cohorts, transcript abundance can also be used to analyze downstream effects of some pathogenic

variants and characterize pathways involved in disease[63,267]. Even where a likely pathogenic variant is identified by DNA analysis, RNAseq can provide supporting evidence that the transcript is affected, and may be used to establish that both alleles are affected in trans by different mutations[193,268]. Although peripheral blood is a mixture of dozens of cell types, so long as the defect is observed in one or more of the major leukocyte or monocyte populations, bulk RNAseq should be a useful source of clinically actionable information. In this study, we show that targeted RNAseq for a set of 13 PID and IBD patients resolves the likely source of immune dysfunction for at least 8 patients, where previous exome analysis had only diagnosed 3 cases.

5.3 Materials and Methods

Targeted RNAseq was performed on samples from 13 patients known to have or suspected of having very early onset inflammatory bowel disease (VEOIBD; 6 samples) primary immune deficiencies (PID; 7 samples). Results of whole exome sequencing (WES), performed previously on all 13 samples, were withheld until initial analysis of RNAseq was completed to compare the success of RNAseq analysis without supplemental genome information.

5.3.1 Sequencing

RNA (median RIN=9.4, range=6.9-9.8) was extracted from peripheral blood mononuclear cells (PBMC) and underwent library preparation using a custom Agilent SureSelect Targeted Enrichment panel of 260 genes linked to VEOIBD or PID, followed by 150bp paired-end sequencing on an Illumina NextSeq platform at an average sequencing depth of 11.5M read pairs.

5.3.2 Alignment and pre-processing

After QC of the raw sequencing files using FASTQC[171], reads were aligned to GRCh38 using GENCODE v29[173] with STAR splice aware aligner version v2.6.1d[125]. In addition to default 2-pass mode parameters, all multimapping reads and splice junctions with <5 supporting reads were filtered out. Post-alignment QC was performed using the Quality of RNAseq Toolset (QoRTs)[175], primarily checking gene-body coverage plots for signs of 5' or 3' bias that may affect downstream analysis.

5.3.3 Variant calling

Variant calls were made on aligned BAM files following GATK Best Practices for Variant Calling in RNAseq[218]. A single VCF file for all samples was created using GenotypeGVCFs and standard filters were applied. Variants were annotated using ANNOVAR[230] and TraP[59] and subsequently prioritized to identify high quality (ALT read depth >= 10 unique non-duplicate reads) exonic variants and variants that may affect splicing. Notably, all three variants (two intronic, one exonic) found to affect splicing had TrAP scores well above the "probably damaging" threshold. The quality of all variants reported were manually confirmed using IGV viewer[231].

5.3.4 Exon usage analysis

To aid in identifying potential aberrant splice events, exon usage was evaluated in two ways. First, percent spliced in (PSI) was calculated following the method laid out in Schafer et al[232]. PSI was used to confirm exon skipping events identified in other forms of data. Second, read counts for collapsed exons were obtained following the DEXSeq protocol[269]. To visualize changes in exon usage while controlling for differences in

overall gene expression, exons were normalized on a per-gene basis using transcripts per million (TPM) in order to factor in exon length. These normalized counts were plotted for each gene and used to visually identify genes and specific exons to prioritize for splicing analysis.

5.3.5 Splicing

Splice counts were obtained for annotated and unannotated junctions following a method adapted from Mendelian RNA-Seq[62]. To ensure that all identified splices were supported by sufficient non-duplicate, uniquely mapping reads, aligned BAM files had duplicate reads removed with Picard Tools and were remapped using STAR after converting back to fastq files. The splice junction output files from STAR were combined and junctions were annotated with the gene name and a list of transcripts that use the junction (for known junctions). To align with established clinical standards for SNP calling, a minimum of 10 non-duplicate uniquely mapped reads were required for an unannotated junction to be further investigated. In addition, these events needed to meet a minimum read support threshold of 10% of the overlapping canonical junction. Events were manually analyzed in IGV viewer to confirm that they were not a result of mis-mapping or a sequencing artefact.

5.3.6 Complementary Analysis

Resulting data files from the above methods were used in a complementary manner to analyze the targeted RNAseq from each patient. After identification of exonic variants, the annotated VCF was used to prioritize genes for manual analysis of the splice counts. Exon usage was used primarily in tandem with spliced read counts to look for aberrant splicing events, but also to identify possible differential isoform usage. Allele

specific expression was calculated with read counts of high quality exonic variants. Gene expression was evaluated with DESeq2 only in conjunction with truncating variants to support conclusions about nonsense mediated decay.

5.4 Results

5.4.1 Development of a Targeted RNAseq panel for Immunodeficiency Analysis

We performed targeted RNA sequencing of 260 genes to an average depth of 11.5M read pairs for 13 patients suspected of, or known to have, an immune-related disorder. A comprehensive analysis of the RNA was completed for each sample, which included variant calling, identification of aberrant splicing, and outlier gene expression. A summary of findings is given in Table 5.1, and the workflow is outlined in Figure 5.1.

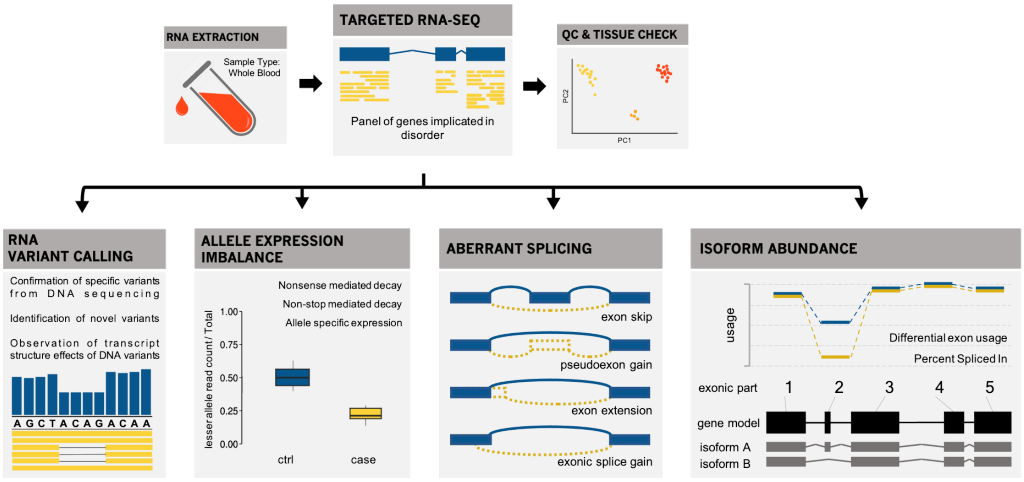

Figure 5.1 An overview to the RNA-seq analysis approach. PBMCs are extracted from whole blood and sequenced using a targeted gene panel. Allele expression imbalance, aberrant splicing, and isoform abundance are analyzed in tandem with called variants to provide a more complete picture of the functional effects of variants on mRNA structure.

Table 5.1 Variants and splicing events of interest

Sample	Gene	HGVS	predicted	Variant Type	Zygosity	splice effect in RNA	SIFT	CADD	Interpro	gnomAD MAF
P25	*MEFV*	c.2040G>C	p.M680I	missense	HET	-	Tolerated	0.002	SPRY	0.0001
	CARD11	c.224G>A	p.R75Q	missense	HET	-	Damaging	33	CARD	-
	TCF25	c.1001_101		nonframeshift	HET	-	-	-	FERM	2.63E-05
P49	*CTLA4*	c.442C>T	p.Q148X	nonsense	HET	-	-	38	Immunoglob	-
P55	*NCF2*	c.812A>G	p.K271R	missense	HET	-	Tolerated	24.4	SH3 domain	-
	NCF1	c.269G>A	p.R90H	missense	HET	-	Damaging	25	Phox	0.001
	WAS	c.689AGA[2]	p.K232del	nonframeshift	HET	-	-	-	-	9.67E-05
P69	*XIAP*			large deletion	HEMI	exon 4-5 skip	-	-		
	MERTK	c.844G>A	p.A282T	missense/	HET	leaky exon skip	Damaging	24.2	Immunoglob	0.0108 (AFR: 0.14)
P89	*NCF1*	c.247G>A	p.G83R	missense	HET	-	Tolerated	20.2	Phox	0.0089
CHB535	*TRAF3*	c.1275C>G	p.Y425X	nonsense	HET	-	-	35	MATH/TRA	-
CHB749	*BTK*	c.1955T>C	p.L652P	missense	HEMI	-	Tolerated	20.6	Protein	-
CHB786	*NOD2*	c.2722G>C	p.G908R	missense	HET	-	Damaging	31	Leucine-rich	0.0113
	ERBIN	c.3704A>C	p.Y1235S	missense	HET	-	Tolerated	19.43	-	1.39E-05
	RBCK1	c.992C>T	p.S331L	missense	HET	-	Damaging	26.3	Zinc finger	9.37E-05
	TYK2	c.2456G>A	p.S819N	missense	HET	-	Tolerated	7.356	Protein	5.03E-05
CHB953	*PIK3CD*	c.1595delG	p.W532X	nonsense	HET	-	-	-	-	-
	CAT	c.903+5G>T	-	splice	HET	intron retention	-	-	-	0.0029 (SAS: 0.01)
	UNC13D	c.154-8T>A	-	splice	HET	intron retention	-	-	-	0.0027 (SAS: 0.01)
	TYK2	c.997_1011	p.V333_E33	nonframeshift	HET	-	-	-	-	-
CHB974	*NOD2*	c.3017dupC	p.A1006fs	frameshift	HET	-	-	-	-	0.0151
	NOD2	c.2722G>C	p.G908R	missense	HET	-	Damaging	31	Leucine-rich	0.0113
CHB1025	*MERTK*	c.844G>A	p.A282T	missense/spli	HET	leaky exon skip	Damaging	24.2	Immunoglob	0.0108 (AFR: 0.14)

Our targeted sequencing panel consisted of 260 genes, 104 of which have been implicated in primary immune deficiencies (PID) and 194 in very early onset inflammatory bowel disease (VEOIBD), with 38 overlapping both disease classes. GC content and gene-body coverage was consistent across all samples. An average of 91% of reads in each sample mapped to the targeted genes, with 196 genes (75%) receiving at least 20 mapped reads in every sample. In order to determine if the coverage extended across the entire gene, however, it is important to look at splice junction coverage. A recent paper developed a calculation for the minimum read sequencing depth (MRSD) needed to express a given gene or genes at a level sufficient for splicing[270]. Because we did not have a PBMC whole mRNA control dataset for direct comparison of MRSD with our targeted panel we used the MRSD web tool and combined the results for whole blood and lymphoblastoid cell lines (LCL), acknowledging that a handful of genes were likely to be specific to a given biotype. MRSD indicated that 88 genes (out of 258 panel genes with splice junctions, 34%) would have greater than or equal to 20 reads mapped to at least 75% of splice junctions in 99% of samples at a sequencing depth <= 50M reads per sample. With targeted RNAseq we found that 140 genes (54%) met the same criteria at an average sequencing depth of ~22M reads (not to mention the fact that we used only non-duplicate reads, which were ~6M per sample). An additional 19 genes have 50% of exons covered, and if we reduce the confidence interval to 75% of samples (since our cohort is disease samples only and we would expect more variation) we end up with a total of 171 genes (66%) with good coverage for a robust splicing analysis.

5.4.2 Application to 7 cases of Primary Immunodeficiency

We used our targeted RNAseq approach to evaluate likely causal mechanisms for 7 cases of primary immunodeficiency being treated at Emory University clinics. Four of

the cases we consider to be resolved on the basis of exonic mutations, two of which were previously noted from WES, whereas the RNA evidence provides additional evidence in support of pathogenicity in two cases. In one of these latter cases, digenic inheritance may be implicated on the basis of a retained intron at a second gene, a phenomenon also seen in a fifth case. This left just two cases completely unresolved. Altered therapeutic intervention is suggested for two cases.

Patient P69 presented with refractory IBD and a history of recurrent fevers. Exome sequencing revealed a hemizygous deletion encompassing exons 4 and 5 of the *XIAP* (HGNC:592) gene, on the basis of which the patient was started on Anakinra[271], an inhibitor of the IL-1 receptor, with the goal of proceeding to a curative bone marrow transplantation (BMT). Analysis of the RNA provided functional evidence that the deletion results in a shift in the reading frame predicted to lead to a premature stop and a protein assay confirmed *XIAP* deficiency. Loss of function (LOF) variants in *XIAP* are known to be causative for X-Linked Lymphoproliferative Syndrome 2 (XLP2, OMIM 300635). To our knowledge, this is the first time this variant has been reported in a patient with a confirmed *XIAP* deficiency.

Symptoms in patient P49 included immune cytopenia, IBD, and eczema, and exome sequencing identified a single heterozygous truncating SNV in *CTLA4* (c.442C>T, p.Gln148*; HGNC:2505). This variant has not previously been reported, but other truncating variants in *CTLA4* are known to be causative for Autoimmune Lymphoproliferative Syndrome Type 5 (ALPS5, OMIM 616100) by way of haploinsufficiency. In addition to a protein assay confirming reduced protein abundance, analysis of RNA revealed allele specific expression occurring in *CTLA4* (Figure 5.2A). The patient was started on Abatacept[272], a CTLA-4 fusion protein that binds to CD80/CD86

and inhibits T-cell activation, as well as the immunosuppressant sirolimus (rapamycin)[273,274], and subsequently underwent bone marrow transplantation. Two other family members were found to harbor the same variant with only mild symptoms, a common finding among families with *CTLA4* LOF variants that suggests other factors may be involved in disease severity[254]. One explanation is compensation by the wild-type allele, which could potentially be observed by targeted RNAseq of relatives in addition to the proband.

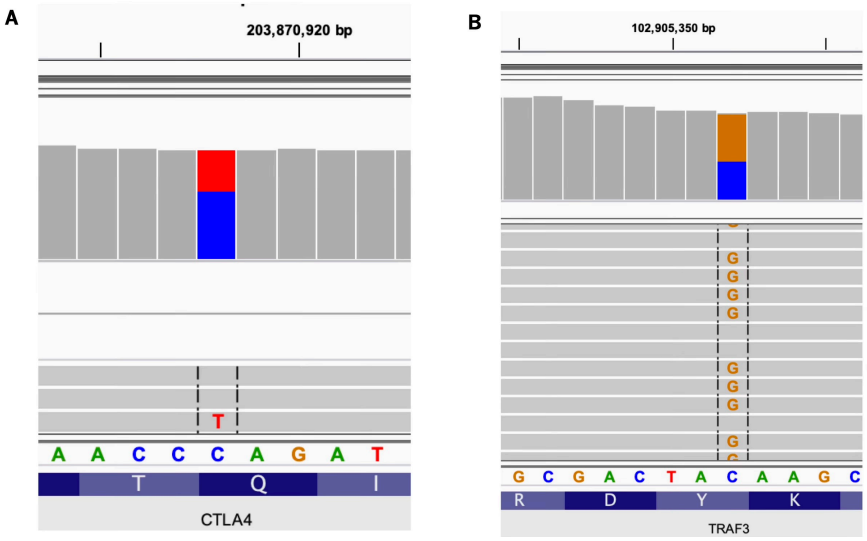

Figure 5.2 Variant allele specific expression. (A) the nonsense variant in *CTLA4* for patient P49 shows unbalanced expression with around 65% of the mapped reads matching the reference allele, suggesting nonsense mediated decay. (B) the nonsense variant in *TRAF3* for patient CHB535 does not exhibit allele specific expression, suggesting it escapes nonsense mediated decay. In fact, the slight bias towards the nonsense allele (~55%) suggests overexpression of the allele given typical reference allele mapping bias.

Patient P25 was previously found to have a dominant negative mutation in *CARD11* (HGNC:16393) causing severe atopic disease (IMD11B, OMIM 617638)[275]. Noting the possibility of other variants contributing to disease heterogeneity and severity,

analysis of RNA continued and identified a pathogenic missense SNV in *MEFV* (HGNC:6998) along with an intron retention event. Although a second variant was not identified in *MEFV*, the highly penetrant M680I mutation has been previously observed in symptomatic carriers of familial Mediterranean fever (FMF, OMIM 249100)[276-278]. Whether or not a carrier with this variant will be symptomatic does not appear consistent within families, suggesting low penetrance and variable expressivity, possibly due to the presence of modifier mutations in other genes. An intron retention event was observed in this patient at the end of *TCF25* (HGNC:29181) intron 9. No causative variant was identified, but due to allele specific expression and increased usage of the remaining *TCF25* exons (Figure 5.3), one possible explanation is a larger duplication. Whole genome sequencing or CNV analysis may shed additional light on this event. *TCF25* is important in transcriptional activity involved in heart development and disease[279]. Disruption of *TCF25* could potentially be an additional susceptibility factor in FMF carriers with a highly penetrant variant.

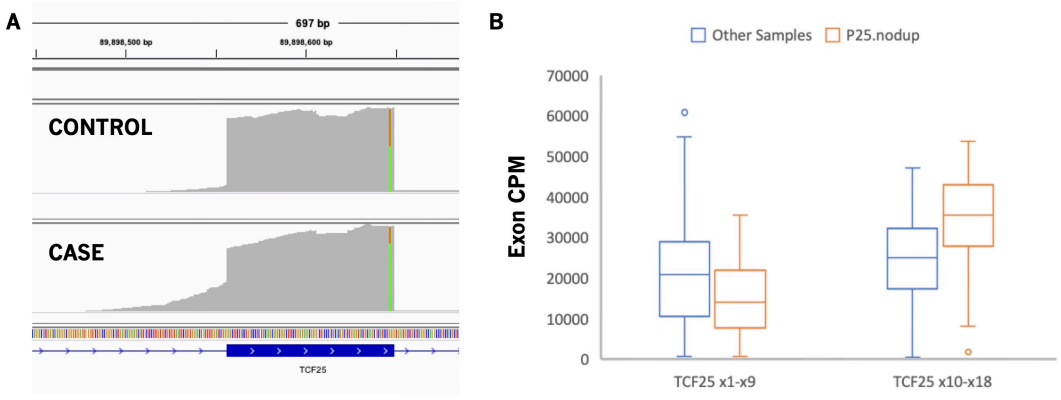

Figure 5.3 Possible aberrant event in *TCF25*. (A) Exon 10 of *TCF25* in an unaffected sample and in P25 showing evidence of intron retention and allele specific expression. The variant seen in this exon is a common polymorphism, but shows unbalanced allele expression in P25. (B) Normalized exon usage in exons prior to the intron retention event compared to the affected exon and following exons.

Patient P55 was found to have rare exonic variants in both *NCF1* (c.269G>A;p.R90H; HGNC:7660) and *NCF2* (c.812A>G;p.Lys271Arg; HGNC:7661), which are primarily linked to chronic granulomatous disease (CGD, OMIM 233700). In the homozygous state, *NCF1* R90H has been associated with a case of pediatric interferonopathy[280]. Splicing and exon usage analysis in this patient suggest the presence of an additional pathogenic event in this gene involving exons 2-3. Unfortunately, the high similarity of pseudogene *NCF1B* (HGNC:32522) make identifying the specifics of this event difficult in RNA alone[281]. Further studies would be needed to confirm whether the mutations in *NCF1* are in trans and hence whether compound heterozygosity explains causation. In any case, the rare events seen in *NCF1* and *NCF2* strongly point towards one or both of these genes being involved in the pathogenesis of this case. In addition, a nonframeshift deletion of a single amino acid was identified in *WAS* (HGNC: 12731), which is associated with Wiskott Aldrich Syndrome (WAS, OMIM 301000). While pathogenicity of this variant has not been determined, we note that carriers of *WAS* pathogenic variants escape disease through non-random X inactivation and that random X-inactivation has been found in symptomatic carriers[282,283], presenting the potential for this to be a second or interacting cause.

Patient P89 also harbors a variant in *NCF1* (p.G83R) which has been previously found to reduce reactive oxygen species and is associated with more severe disease course in pediatric IBD[284]. In addition, though no aberrant splicing was detected, abnormal exon usage was observed in the ubiquitin-modifying enzyme (UME) *USP4* (HGNC:12627). UMEs are involved in the regulation of the IBD disease course[285,286]. Four other rare exonic or UTR variants were identified in other panel genes for this patient (Table 5.1). Exceedingly rare variants have been shown in previous research to be over-

represented in early-onset IBD and primary immune patients[287,288], suggesting a complex multigenic disease origin for this case.

5.4.3 Application to 6 cases of Very Early Onset Inflammatory Bowel Disease

We also used our targeted RNAseq approach to evaluate likely causal mechanisms for 6 cases of very early onset IBD being treated at Boston Children's Hospital. Three of the cases we consider to be resolved on the basis of exonic mutations, two with *NOD2* involvement and one hemizygous for a missense variant that the RNAseq supports elevating to likely pathogenic. Two other cases have variants of interest that influence splicing or transcript abundance, and the final case provides evidence for oligogenic inheritance. Suggestive genetic abnormalities were thus detected in all cases, though follow-up assays would be needed to confirm several of these. Altered therapeutic intervention is suggested for two cases.

NOD2 (HGNC:5331) has been repeatedly associated with IBD[289-291]. A recent study suggested that compound heterozygosity of known *NOD2* risk alleles explains up to 10% of pediatric IBD in European-ancestry cases[292]. In our cohort, patients CHB974 and CHB786 were found to harbor the p.G908R variant. A second *NOD2* risk allele, p.L1007fs, was identified in CHB974, confirming that *NOD2* loss-of-function is likely the causative mechanism in this child. While there were no additional *NOD2* variants found in CHB786, ultra-rare variants were observed in three other genes, including the *NOD2* inhibitor *ERBIN* (HGNC:15842, Table 5.1), indicating this case could be di-genic.

X-Linked Agammaglobulinemia (XLA, OMIM 300755), which is characterized by low B-cell counts and is associated with early-onset colitis, is caused by defects in the *BTK* (HGNC:1133) gene. CHB749 was found to be hemizygous for a missense variant

(c.1955T>C; p.L652P) in the tyrosine kinase domain of *BTK*. This variant has been reported in patients with XLA previously[293,294], but researchers studying *BTK* variant effects on protein have drawn attention to dissimilarities between this variant and other pathogenic *BTK* variants – while the majority of disease-causing *BTK* variants are missense changes in structurally important residues of the tyrosine kinase domain, L652P is not a well conserved location and the residue is exposed in the assembled protein structure[295,296]. However, the change to Proline in this portion of the kinase domain C-lobe breaks the α-helix, making this change more likely to be disruptive to the protein. In the previous studies that reported the L652P variant in a patient with XLA, protein expression was not determined. Assays to assess *BTK* protein expression and B-cell levels should be performed in this patient to confirm a diagnosis of XLA. Since intravenous immunoglobulin to treat infections may not improve inflammation from colitis in patients with *BTK* defects[297], a more creative and personalized treatment plan may be required for this case.

Patient CHB535 was found to contain a nonsense variant in *TRAF3* (c.1275C>G;p.Y425X, HGNC:12033). The extensive functions and interactions of *TRAF3* are still being elucidated, but it is known to be important to inflammatory pathway signaling and gene abnormalities have been associated with many diseases including herpes simplex encephalitis (IIAE5, OMIM 614849), Waldenstrom macroglobulinemia (WM, OMIM 153600), and IBD. The Y425X variant occurs in the highly conserved TRAF-C subunit of the TRAF domain, which is responsible for receptor binding and participates in stabilization of TRAF3 trimerization[298]. Few truncating germline variants have ever been reported in *TRAF3*, and it is not known whether truncating variants are disease-causing. A missense variant was previously found to have a dominant-negative effect on the protein

via destabilization of TRAF3 trimers leading to protein expression of only 17.5% compared to the wild-type[299]. This and other studies have shown that deletion of the TRAF-C domain (the predicted effect of the Y425X truncating variant) does not have the same effect and produces the 30% of protein necessary to maintain normal signaling function (an amount that suggests simple haploinsufficiency is not disease-causing)[300-303]. However, the specific deletion variant created in these studies removed not just the TRAF-C subunit, but also all or part of the TRAF-N subunit shown to be essential for TRAF3 trimerization. Other studies into TRAF3 protein interactions that deleted specifically the TRAF-C subunit in the course of their research unquestionably prove that removal of this domain disrupts inflammatory pathway signaling[304,305], but did not study the wild-type/deletion variant combination. This leaves open the possibility that the Y425X variant observed in CHB535 could act in a dominant-negative manner to cause inflammatory disease. Our RNA analysis provides an important clue to clarifying the pathogenicity of this variant. In order to act in a dominant-negative fashion, the truncated transcript needs to elude degradation by the nonsense-mediated decay (nmd) machinery. This often happens when the truncation occurs in the last coding exon of the transcript[240]. The RNA analysis showed clear, balanced heterozygosity of the variant as well as normal overall expression of *TRAF3* (Figure 5.2B), confirming escape from nmd. In order to show that TRAF3 protein function is sufficiently deficient and declare Y425X likely pathogenic, a protein assay should be performed.

MERTK (HGNC:7027) signaling is important in the negative regulation of inflammation[306]. Two samples, CHB1025 and P69, harbor a missense variant at the end of *MERTK* exon 5 (c.844G>A;p.A282T). This variant has previously been reported in patients with multiple myeloma[307], but in-silico predictors do not agree on whether this

change would be damaging to protein function, and gnomAD[27] frequency in African Americans is 14%. It is no wonder, then, that submissions of this variant to ClinVar have interpreted it as Benign. However, our analysis of mRNA shows this to be a “leaky” variant, where the reduced affinity for the canonical splice site results in a non-frameshift exon skip (Figure 4). Exon 5 of MERTK is part of one of the immunoglobulin-like domains that are important for ligand binding in the inflammatory pathway[308]. Only about 15% of total MERTK transcripts are mis-spliced, making it unlikely to be a disease-causative mutation for retinitis pigmentosa, the rare disease typically associated with the gene, especially given its high frequency in the African population. The possibility remains, though, that this is a risk allele for immune disorders, in combination with XIAP hemizygosity in P69, and an as yet unidentified cofactor in this case of VEOIBD.

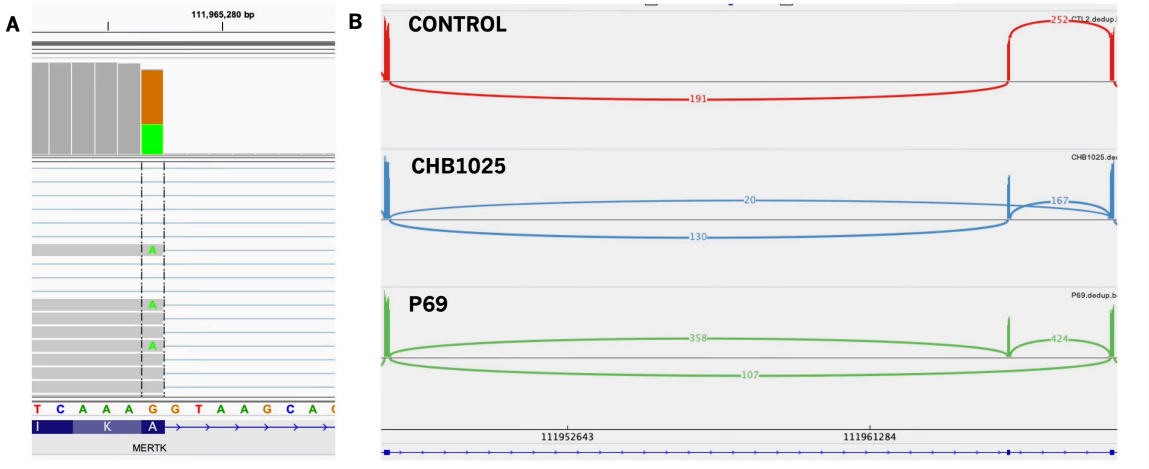

Figure 5.4 “Leaky” exon skip in *MERTK.* (A) The A282T variant is visible at the end of exon 5. About 33% of reads mapping to this location have the G>A change and splice normally. (B) Sashimi plot showing the skip of exon 5 in patients CHB1025 and P69. About 15% of total spliced reads skip exon 5.

Common variable immune deficiency (CVID, OMIM 607594), a primary immune deficiency that has been associated with VEOIBD[257], has been suggested to be polygenic in origin rather than the traditional monogenic mode of congenital disease inheritance. A total of four potentially disease-causing variants were identified in CHB953, three of which are located in genes linked to CVID[309]. A heterozygous nonsense variant was found in *PIK3CD* (HGNC:8977). While PIK3CD-related disease is primarily caused by missense gain of function variants, at least two studies have identified loss of function variants to be disease-causing as well[310,311]. The second variant identified was a 15bp deletion encompassing the exon 7 splice donor site of *TYK2* (HGNC:12440). Analysis of RNA showed the resulting mRNA change to also be a non-frameshift deletion of 15bp, thanks to an alternate splice donor site conveniently located at the beginning of the genomic deletion. Despite this minimal disruption, the deletion removes a portion of the FERM domain, which has been shown to be important to TYK2 protein function[312]. Thirdly, extended splice site variants were found in *UNC13D* (c.154-8T>A; HGNC:23147) and *CAT* (c.903+5G>T; HGNC:1516), both resulting in intron inclusion (Figure 5.5). The *UNC13D* variant has not previously been reported in the literature but has been reported to ClinVar and interpreted as Benign and Likely Benign. The catalase variant was reported as causative for acatalasemia/hypocatalasemia (OMIM 614097) in a study that found a reduction in catalase levels in patients carrying the variant[313]. Again, this variant has been reported to ClinVar, with interpretations of Benign and Uncertain Significance. Both variants have likely been disregarded due to their frequency – while generally rare, they occur in just over 1% of South Asians according to gnomAD. Usage of the nearby canonical splice donor site in *CAT* is roughly 60% of expected, while canonical splicing in *UNC13D* is just 40% of expected for the affected location. No alternative splicing was observed, and a high number of reads mapped to both introns. In both genes, the intron

inclusion is expected to result in premature truncation. *CAT* mRNA levels were not reduced compared to other samples, so a protein assay would be needed to confirm hypocatalasemia. A reduction of *UNC13D* was similarly not seen in the mRNA. However, exonic SNPs across the entirety of the gene were observed in a roughly 40/60 ratio. Coupled with the ~40% use of the canonical splice site, it appears as though rather than nmd this transcript is exhibiting increased expression. While these variants cannot be classified as pathogenic based on the data in this study, we recommend that they be considered VUS until further research can be done on their effects. Construction of mouse strains with combinations of mutants might reveal the oligogenic basis of the pathogenesis in this individual[314].

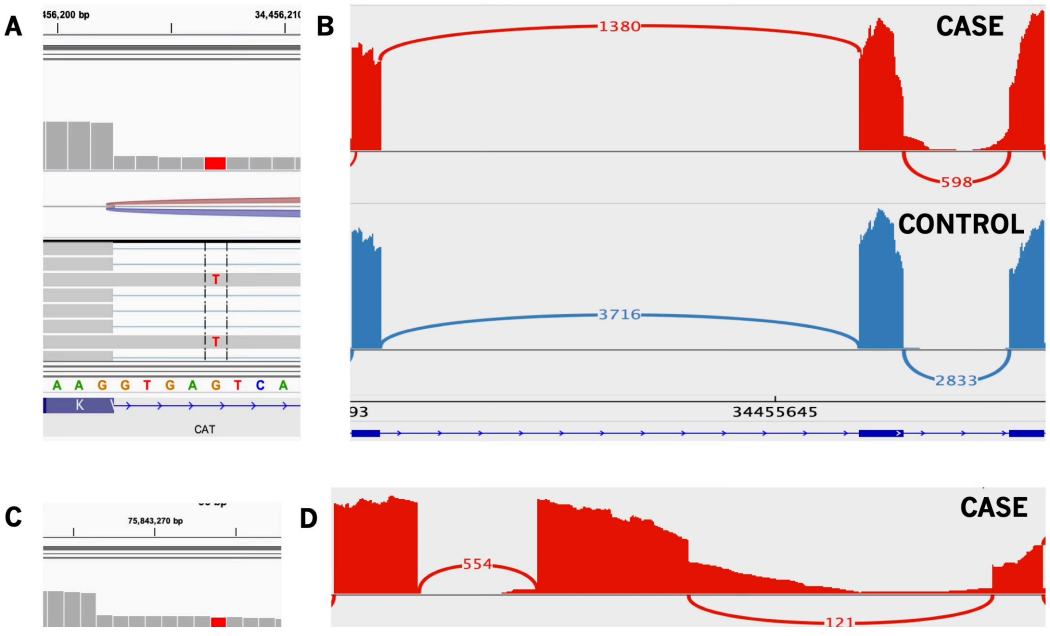

Figure 5.5 Intron retention events in CHB953. (A) The extended splice variant in *CAT* is present in all unspliced reads. (B) Unspliced reads are visible in the *CAT* intron. Normalized spliced read counts from this exon are around 60% of what is seen in other samples. (C) The extended splice variant in *UNC13D* is present in all unspliced reads. (D) Unspliced reads are visible in the *UNC13D* intron. Normalized splice read counts from this exon are around 40% of what is seen in other samples.

5.5 Discussion

This study introduces three innovations with respect to personalized genomic medicine: (i) use of targeted RNAseq to increase the resolution of slicing dysregulation, (ii) development of a modified bioinformatics pipeline for diagnostic evaluation, and (iii) application to a pilot study of 13 cases with two classes of immunodeficiency. Combining these, we provide evidence for previously un-noticed mechanisms of disease for 3 individuals, confirm likely pathology for another 3, and provide suggestive evidence for di- or tri-genic inheritance in two more. The two major limitations of the approach are that it may only be applicable to immune diseases where common blood cell types are involved, and the targeted RNAseq panel may not include the causal gene in some cases.

The decision to use a targeted panel rather than sequencing the entire transcriptome is unusual, but is validated by evidence that it increases the proportion of splice sites with sufficient read depth to evaluate dysregulation. One of the largest limitations of RNAseq for rare disease diagnostics is that the ability to capture a variant is dependent on that gene's expression level in the sequenced tissue type. This generally leads to arguments that the disease-relevant tissue is a necessity for RNAseq and/or that sequencing depth should be at least 50-100M reads per sample[51,62,74]. The recent minimum read sequencing depth (MRSD) study identified whole blood (over LCL, cultured fibroblasts, and skeletal muscle) as the worst option for most gene panels[270]. However, we show that our targeted gene panel outperformed expectations and allowed us to analyze at least 20% more genes of interest than we would have been able to with whole mRNA.

Three variants reported in ClinVar as benign or likely benign were shown to affect splicing in mRNA, drawing the previous interpretation into question. This exemplifies some of the biggest drawbacks to rare variant interpretation in DNA: the dependence on variant

frequency and in-silico predictors. Splice prediction tools, while useful for narrowing in on variants, will never be as accurate as directly assessing the effect through RNA-seq. In addition, assays showing a reduction in functional protein function are interpreted more readily in the context of RNA evidence of the specific change resulting from a splice variant. Variant frequency, while it is (and should) remain a primary way of prioritizing putatively pathogenic variants, should sometimes also be used with some caution where the variant is not necessarily causal, but likely facilitative of dysregulation. Rare disease prevalence is widely thought to be underestimated and is complicated by heterogeneous phenotypes, digenic and polygenic inheritance, and differences between subpopulations[315,316]. For example, acatalasemia/ hypocatalasemia prevalence has been estimated at over 2% in some Asian populations[317,318], making it important to consider individual ancestry when interpreting catalase variants. In a polygenic inheritance model, it is possible that a specific combination of variants that are each individually more common than the disease prevalence together create the disease-causative effect. These nuances are increasingly important as we improve the field of personalized medicine to better understand and treat complex rare disease cases.

The method in Cummings et. al. has been criticized for lacking a statistical basis and arbitrarily choosing cutoff thresholds[63,67,68]. Since publication, multiple tools for a robust statistical analysis for the identification of aberrant splicing have been developed, most notably FRASER[67]. However, the FRASER paper acknowledges that sample size affects the ability of the tool to detect all known splice events, which highlights the important point that rare disease RNAseq analysis tends to involve small sample sizes. Since our work most closely resembles the Cummings et al method, after completing our splicing analysis we also ran FRASER for comparison. Out of seven splice events

manually identified through the original analysis, FRASER detected just two with FDR <0.1. The FRASER paper suggests that z-score and delta PSI be prioritized over p-value, especially in small cohorts like ours. Another three events were successfully identified by FRASER using their suggested delta PSI cutoff of 0.3. When the delta PSI threshold was lowered to 0.1 and the read counts were used to prioritize events (a method quite similar to this study and Cummings et al), FRASER detected 6 of 7 splice events as well as two additional events not found in the original analysis.

For research labs that lack the funds and resources required for whole mRNA sequencing of 100 rare disease samples at a depth of >50M reads per sample, RNAseq analysis is not at all out of reach. With a well-curated panel, targeted RNAseq can nearly double the number of genes that can be analyzed, at a fraction of the total sequencing depth. A smaller number of genes to analyze means that each individual patient sample can undergo a more thorough analysis that combines variant calling, exon usage information, and identification of splice events using both FRASER and the manual methods used in this study and Cummings et al to achieve the highest resolution.

We show the potential of the approach to increase diagnostic yield, but much work needs to be done to incorporate findings of this nature into the ACMG guidelines for clinical diagnosis, and thence to improve patient care. We stress that the RNA-seq findings regarding variants in this paper do not meet the threshold for categorization as pathogenic or likely pathogenic. All individuals, even those who are considered healthy, contain many rare variants in disease-related genes. However, in the quest to elucidate the genetic causes of rare disease and increase diagnostic yield, the field must look beyond a simple monogenic mode of inheritance. As we learn more about how variants that are not exceedingly rare (but less common than most polymorphisms) contribute to rare disease

severity and heterogeneity, it will likely become necessary for additional or adapted guidelines to be developed in order to standardize how we interpret these variants in the context of a patient's personal variant profile.

CHAPTER 6. CONCLUSION

In the diagnosis of rare genetic disease, targeted RNA-seq has proven to be a powerful tool independent of or complementary to traditional genomic testing. In the past decade, NGS tools have become commonplace in the clinical diagnostic domain, a testament to their value and decreasing costs. WGS and WES have become regarded as the future of genetic testing due to their ability to be reanalyzed as research identifies new disease genes or reclassifies variants. Unfortunately, WGS and WES have shown only modest increases to diagnostic yield for most rare congenital diseases. As personalized medicine becomes a reality through therapies targeted to specific genes and variants, it becomes increasingly important to resolve VUS and identify aberrant splicing through functional assays, which DNA-sequencing methods are unable to do. This has always been a challenging part of rare disease research, as evaluating the pathogenicity of individual variants has typically been time-consuming and expensive.

RNA-seq is not only able to evaluate the effects of many VUS, but also can identify exonic variants in addition to structural events like aberrant splicing. I have shown here that aberrant splicing is an important contributor to the pathogenesis of neuromuscular and immune disorders. Targeted RNA-seq significantly improves the diagnostic yield of rare disease cohorts and can even phase variants without knowledge of parental alleles. In addition to validating RNA-seq as a valuable tool for clinical diagnostics, I would like to draw attention to a few other conclusions made along the way.

First, using a targeted gene panel is often the best option when you are able to connect a patient's phenotype to a category of disorders. The high read depth is unrivaled when it comes to confidence in variant calling and identification of splice events. Whole

transcriptome mRNA is too much information to sift through – bioinformaticians have to narrow it down to candidate genes as we simply do not know enough about splicing in each individual gene to evaluate aberrant events. Many genes are known to be more tolerant to LoF events, and this information translates to splicing analysis as well. One downside of targeted RNA-seq is the lack of available reference samples for comparison. However, sequencing a large enough disease cohort provides adequate information about panel-specific effects on sequencing to determine whether an unannotated event is abnormal.

Second, much work needs to be done to integrate the knowledge that splicing is a significant contributor to rare disease into traditional genomic analyses. It is unrealistic to believe that RNA sequencing will become favored over WGS and WES. In-silico splice predictor algorithms have improved and are being used in the interpretation of sequence variants in DNA, but they will never be as good as directly observing variant effect in RNA. Additional research needs to be done to expand the knowledge base regarding splicing in disease genes. One possibility to assist in this directive would be a database similar to ClinVar for reporting whether or not a variant was seen to affect splicing in mRNA studies.

As a final note, I'd like to discuss the state of rare disease RNA analysis methodology. When I first began my PhD research, bioinformatic analysis of human mRNA resembled the wild west. Unlike the well-developed and refined world of DNA-sequencing, few gold standards or widely accepted guidelines for analysis existed. In fact, many tools for RNA-seq data were simply the tools for DNA with minor alterations that failed to account for the nuances of RNA-seq analysis. The RNA analysis tools that did exist were built for application to large sets of replicate samples and did not perform well

for rare disease analysis, which typically has just one or two patients harboring the same mutation.

While I was developing my methods for analyzing RNA in rare disease patients, many others were doing the same. Time and time again, we have found RNA-seq improves diagnostics for Mendelian disease, doing so regardless of methodology used. At least one other study has found that using more than one splice analysis tool improves diagnostic yield even more. While I have found some of the newly developed tools to be helpful, in the current state of RNA data interpretation there is simply too much noise for anything to truly improve over manual analysis by a skilled interpreter of sequence data. One thing in particular I would like to see for this field is an improved method for variant calling in RNA that can accurately weed out false positive variant calls at exon-intron boundaries and could possibly take variable read depths at different genes and parts of genes (like exons vs introns) into account.

There are widely accepted standards for the analysis of DNA sequencing data, but we must keep in mind that they took time to develop. The analysis of RNA for the diagnostics of rare disease is still in its infancy, and the recent development of so many new bioinformatic tools for this purpose indicates the excitement behind this field. Continuing robust and comprehensive analysis of rare disease cohorts will not only provide additional diagnoses for the patients involved, but will advance our ability to best analyze this data.

APPENDIX A. SUPPLEMENTARY TABLE FOR CHAPTER 4

Table A.1 Variants and monocyte assay results for the dysferlinopathy cohort

Patient ID	RNA ID	Monocyte Assay Interpretation	Allele 1	Alelle 2	Alelle 3	Alelle 4	Diagnosis of other forms of MD	Comments
1		disease range	c.4886+1249G>T	c.1834C>T *	-			
2		disease range	c.1168_1180+1dup14	c.1168_1180+1dup14	-			
3		carrier range	c.2408G>A	-	-			
4		disease range	c.5668-7G>A *	c.5768-1415_5946+710del	-			
5		disease range	c.1620delA	c.1861G>C (p.G621R)	c.2690C>T			
6		normal range	-	-	-			
7		normal range	c.509C>A	-	-	-	*ANO5*: IVS1-1G>A c.191dupA	Other Gene: *ANO5*. 2 Pathogenic variants
8		normal range	-	-	-			
9		normal range	c.681C>T	-	-			
10		disease range	c.3444_3445delinsAA	c.4886+1249G>GT	-			
11		carrier range	-	-	-			
12		disease range	c.1392dupA *	c.5698_5699delAG *	-			
13		disease range	c.907-3C>A *	c.2641A>C	-			
14		disease range	-	-	-			
15	C1	disease range	c.3534C>T	-	-		* RNASeq identified *DNAJB6* gene UTR VUS c.-85G>T	Possible *DNAJB6* case with the 5'UTR VUS could have regulatory effect in gene expression
16		disease range	c.1168G>A	c.4886+1249G>T	c.4731G>A			
17		disease range	c.3444_3445delinsAA	c.4886+1249G>GT	-			
18		disease range	c.797T>G	c.1663C>T *	-			
19		disease range	c.2997G>T	c.5668-824C>T *	-			
20		carrier range	c.3892A>G	-	-			

Table A.1 continued

21		carrier range	c.3065G>A *	-	-			
22		disease range	c.4894G>T	c.4894G>T	c.1398-2A>G			
23		disease range	c.1392dupA *	c.5713C>T	-			
24		disease range	c.4200dupC	c.5509G>A *	-			
25		disease range	c.1481-1G>A *	c.5712C>T	-			
26		carrier range	-	-	-			
27		carrier range	-	-	-			
28		disease range	-	-	-			
29		carrier range	ND	ND	-			
30		carrier range	c.2290G>T	c.2638G>A	-			
31		disease range	c.5509G>A *	del Ex2_Ex3	-			
32		disease range	-	-	-		Homozygous for pathogenic variant in *FKRP* so likely has LGMD2I	*FKRP* Likely Pathogenic homozygous variant: c.826C>A
33		disease range	c.5979dupA	c.5979dupA	-			
34		carrier range	-	-	-			
35		disease range	c.857T>A	c.4886+1249G>T	-			
36		disease range	c.2643+1G>A *	c.3327_3328delGT	c.4577A>C *			
37		disease range	c.742C>T	c.5296G>A *				
38		disease range	c.353delT	c.3137G>A *				
39		disease range	c.1638+2T>	c.1642delG				
40		carrier range	-	-			Compound Heteroyzygote for 2 variants in *ANO5* so likely has 2L	Other Gene: *ANO5*: c.191dupA (pathogenic), c.749A>G (VUS)
41		carrier range	-	-				
42		normal range	c.4794G>T *	-				
43		normal range	-	-				
45		disease range	c.937+1G>A	c.5441G>A				
46		disease range	c.4090C>T	c.5296G>A *				
47		normal range	c.757C>T *	c.3892A>G				

Table A.1 continued

48		normal range	-	-	-	-	pathogenic variant found in *DMD* so likely has Duchenne/ Becker muscular dystrophy	*DMD* Pathogenic variant: c.1704+1G>A. Confirmed Becker's muscular dystrophy (DMD).
49		disease range	c.2071C>T	c.3113G>A *				
50		normal range	ND	ND				
51		normal range	ND	ND				
52		disease range	c.1392dupA *	c.755C>T				
53		disease range	del Ex2_3	del Ex2_3				
54		normal range	c.865T>C	c.4794G>T *				
55		disease range	c.1368C>A	c.1368C>A				
56		normal range	c.3992G>T *	c.3065G>A *				
57		disease range	c.937+1G>A	c.5441G>A				
58		disease range	c.2643+1G>A *	c.2643+1G>A *	c.4577A >C *			
59		disease range	c.663+1G>C	c.1284+2T>C	c.4374C >T			
60		normal range	-	-				
61		normal range	-	-				
62		disease range	c.4253G>A *	c.4253G>A *				
63		disease range	c.4253G>A *	c.4253G>A *				
64		disease range	c.2643+5G>A *	c.3113G>A *				
65		carrier range	ND	ND				
66		normal range	-	-				
67	C2	disease range	c.5341G>A *	c.3137G>A *				
68		disease range	c.3517dupT	c.5713C>T				
70		disease range	c.1639-6T>A *	c.5503A>G *				
71		normal range	-	-				
72		disease range	c.610C>T	c.5979dupA				
73		normal range	-	-				
74		carrier range	ND	ND				
75		disease range	c.937+1G>A	c.5266C>T				

Table A.1 continued

76		disease range	c.4200dupC	c.5509G>A *				
77		disease range	c.1053+1G>A	c.792G>C				
79		normal range	c.3534C>T	-				
80		disease range	c.2643+5G>A *	c.3113G>A *				
81		carrier range	ND	ND				
82		disease range	c.6124C>T *	c.2352_2355+1delGGAGG				
83		disease range	c.5979dupA	c.5979dupA				
84		normal range	-	-			Homozygous for pathogenic variant c.191dupA in *ANO5* so likely has 2L	*ANO5*; LGMD2L
85		disease range	c.2997G>T	c.5668-824C>T *				
87		normal range	-	-				
88		normal range	c.6197C>T	c.4794G>A				
89		disease range	c.610C>T	c.1053+1G>A *	c.1120G>C			
90		disease range	c.610C>T	c.1053+1G>A				
91		normal range	c.2902A>T	-				
92		normal range	ND	ND				
93		disease range	c.2779delG	c.2779delG				
94		carrier range	-	-				
95		disease range	c.5836_5839delCAGC *	c.5644C>T				
96		disease range	c.610C>T	c.2643+1G>A *	c.4577A>C *			
97		disease range	c.3702T>C	c.2790G>C	c.2643+1G>A *	c.4577A>C *		
98		disease range	c.5713C>T	c.5911T>C	c.1120G>C			
99		normal range	ND	ND				
101		normal range	-	-				Other Gene: *COL1A1* likely pathogenic variant: c.1724G>A (p.G575D). COL1A1 variants are inherited autosomal dominant

Table A.1 continued

102		ND	-	-				
103		disease range	del Ex2_3	del Ex2_3				
104		disease range	c.4685dupT	c.2162G>C				
115		normal range	-	-				
116		ND	-	-				
117		carrier range	c.3851T>C	c.681C>T				
118	C3	disease range	c.4742G>A	-				
119		carrier range	c.1120G>C					
120		disease range	c.265C>T	c.1956G>A				
121		disease range	c.1931-2delA	c.3349-2A>G				
122		carrier range	-	-			homozygous for a pathogenic variants in *FKRP* so likely has LGMD2I	*FKRP* pathogenic variant: c.826C>A
123		carrier range	-	-			homozygous for a pathogenic variant in *FKRP* so likely has LGMD2I	Other Genes: *FKRP* pathogenic variant: c.826C>A;
124		normal range	c.1120G>C	-				
125		disease range	c.855+2T>G	c.855+2T>G				
126	C4	normal range	c.2332C>T	-			* Identified by RNA-Seq;- *CAPN3* exon 17 pathogenic c.1981delA variant; *VCP* exon10 VUS c.1106T>C	Possibly either *CAPN3* (with 2nd variant not yet identified) or VCP (autosomal dominant).
127		disease range	c.755C>T	c.5444G>T				
128		normal range	c.626C>T	-				
129		disease range	c.5982_5989dup8	c.5982_5989dup8				
130	C5	disease range	c.3967C>G	-				
131		carrier range	c.4794G>T *	-			homozygous for pathogenic variant in *CAPN3* so likely has LGMD2A	*CAPN3* pathogenic variant: c.1993-G>A

Table A.1 continued

139		disease range	c.1343T>C	c.790G>T				
140		disease range	c.1180+5G>C	c.1180+5G>C				
141		disease range	c.3349-2A>G	c.5979dupA				
142		disease range	c.1368C>A	c.5342G>A				
143		normal range	c.2099G>A	-				
144		disease range	c.701G>A	c.1555G>A				
145		carrier range	c.3355G>A	-				
147		disease range	c.4794+1G>T	c.1663C>T *				
148		carrier range	c.1351A>G	-				
149		carrier range	c.6124C>T *	c.4794G>T *				
150		disease range	c.4228C>T	c.5609G>A				
151		disease range	c.5979dupA	c.5979dupA				
152		disease range	del Ex4 *	c.4434G>A				
153		carrier range	c.4024C>T	-				
154		disease range	c.1368C>A	c.1368C>A				
155		normal range	C.707A>C;	C.3760C>T				
156		normal range	c.758G>A	-				
157		carrier range	c.6022G>A	-				
158		disease range	c.2779delG	c.2779delG				
159		disease range	c.2779delG	c.2779delG				
160		disease range	c.2779delG	c.2779delG				
161		disease range	c.6056G>T	c.6056G>T				
162		disease range	c.5836_5839delCAGC *	c.1852G>C				
163		carrier range	c.4886+1249G>T	-				
164		normal range	c.5026G>T	-				
165		normal range	c.3277C>T	-				
166		disease range	c.2875C>T	c.5698_5699delAG *				
167		carrier range	c.221C>T	-	-	-	homozygous for a pathogenic variant in *ANO5* so likely has LGMD2L	Other Gene: *ANO5* homozygous pathogenic variant: c.191dupA
168		normal range	c.4134C>T	c.4267C>T				

Table A.1 continued

169		carrier range	c.5999G>A	-				
170		carrier range	c.1351A>G	-				
171		disease range	c.2643+1G>A *	c.4200dupC				
172		carrier range	c.842C>T	-				
173		disease range	c.3466T>C	c.3466T>C				
174	B1	disease range	c.1834C>T *	del Ex52 *	-	-		
175		carrier range	c.4787A>G	-				
176		normal range	c.2756G>A	-				
177		normal range	c.3624C>G	-				
178		disease range	c.5429G>A	c.757C>T *				
179		normal range	c.3487G>A	-	-	-	Other Gene: *COL6A1* pathogenic variant c.362A>G	Y (*COL6A1*; Bethlem myopathy/Ullrich muscular dystrophy)
180		carrier range	c.2902A>T	-				
181		carrier range	c.2516C>T	-				
182		carrier range	c.3624C>G	-				
183		disease range	c.6124C>T *	c.5302C>T				
184		disease range	c.2997G>T	c.2995T>C	c.4742G>A			
185		disease range	c.5902T>C	c.5902T>C				
187		carrier range	c.757C>T *	-				
188		carrier range	c.2614G>A	-				
189		normal range	c.1402C>T	c.4052A>G			other gene: *CAPN3* pathogenic variant c.1303G>A	Possibly *CAPN3* with the second variant not yet identified
190		carrier range	c.984C>T	-				
191	A12	carrier range	c.1517C>G *	c.4408C>T *				
192		carrier range	c.1351A>G	-				
193		disease range	c.5444G>T	del Ex52 *				
194		disease range	c.3967C>G	-				
195		disease range	c.2875C>T	c.5698_5699delAG *				
196		disease range	c.5698_5699delAG *	c.2875C>T				

Table A.1 continued

197		disease range	c.3137G>A *	c.701G>A				
198	B3	disease range	c.6216delC *	c.4509+1586dupG *				
199		disease range	c.1276G>A	c.5713C>T				
200		normal range	c.2872A>G	-				
201		disease range	c.2643+1G>A *	c.4872_4876delGCCCGinsCCCC				
202		disease range	c.1368C>A	c.3130C>T				
203	B4	disease range	c.5698_5699delAG *	del Ex4 *				
204		disease range	c.1053+1G>A *	c.5033G>A				
206		carrier range	c.1353G>A	-				
207		carrier range	c.1385G>A	-			2 pathogenic variants in *CAPN3* so likely has LGMD2A	other gene: 2 *CAPN3* pathogenic variants c.759_761delGAA, c.1468C>T
208		carrier range	c.617C>T	-				
209		carrier range	c.6063C>T	-				
210		disease range	c.2811-2A>C	c.5529G>A				
211		carrier range	c.5189T>C	-				
212		carrier range	c.4865C>T	-				
213		disease range	c.2105C>T	-				
214		carrier range	c.3243C>T	c.3624C>G				
215		disease range	c.2643+1G>A *	c.2643+1G>A *	c.4577A>C *	c.4577A>C *		
217		carrier range	-	-				
218		carrier range	c.1351A>G	c.2423G>A			other gene: *CAPN3* pathogenic variant c.245C>T	Possibly *CAPN3* with the second variant not yet identified
220		disease range	c.879_883dupGACAG	del Ex52 *				
223		disease range	c.4299C>G	c.5713C>T				
224		normal range	c.4253G>A *	-				
225		disease range	c.4497delT *	c.3444T>A *	c.4253G>A *			
227		disease range	c.2779delG	c.2348C>T	c.5963C>T			
229		disease range	c.3118C>T	c.3770G>A				
231		disease range	c.1834C>T *	c.3112C>T				

Table A.1 continued

232		normal range	c.1351A>G	-			other gene: *COL6A1* likely pathogenic variant c.1013G>A	Y (*COL6A1*; Bethlem myopathy/Ul rich muscular dystrophy)
233		disease range	c.1861G>A *	c.2643+1G>A *				
234		carrier range	c.4198C>G	-				
235		carrier range	c.3967C>G	-			pathogenic variant in *MYOT* so likely has LGMD1A	1 pathogenic variant in *MYOT*: c.179C>G. MYOT variants are autosomal dominant causing myofibrillar myopathy (OMIM# 604103). MYOT confirmed molecular diagnosis.
236		disease range	c.1129C>T	-				
237		normal range	c.2726G>T	-				
238		carrier range	c.1250A>G	-				
239		disease range	c.4880T>C	c.5509G>A *				
240		carrier range	c.3983C>T	-				
241		carrier range	c.5245C>T	-				
242	B5	disease range	c.1663C>T *	del Ex52 *				
243		disease range	c.3803G>A	c.1053+1G>A *				
244	C12	disease range	c.2929G>A	c.2929G>A	c.3022G>A	c.3022G>A		
245		disease range	c.1053+1G>A *	c.3803G>A				
246	B6	disease range	c.2875C>T *	del Ex52 *				
247	A1	disease range	c.2496_2499delGACA *	c.5668-7G>A *				
248	C6	carrier range	c.17T>A	c.4794G>T*				
249	A2	disease range	c.331C>T *	c.6124C>T *				
250	A3	disease range	c.2643+1G>A *	c.6124C>T *				
251	B7	disease range	c.3112C>T *	-				
252		disease range	c.6008G>A	c.3112C>T				
253		disease range	c.3112C>T	c.3112C>T				

Table A.1 continued

254		disease range	c.3112C>T	c.3112C>T				
255	C13	disease range	c.5526-7T>G *	c.5526-7T>G *	c.2079C>T	c.2079C>T		
256		disease range	c.3112C>T	c.6008G>A				
257		carrier range	c.3112C>T	-				
258	A14	disease range	c.4577A>C*	c.2643+1G>A *	c.3112C>T *			Pathogenic *DYSF* variant combination was not known
259		normal range	-	-				
261		disease range	c.3383dupT	c.3703-2A>G				
262		disease range	c.2875C>T	c.3349-2A>G				
263		normal range	ND	ND				
264	A4	disease range	c.1071delC *	c.5698_5699delAG *				
265		disease range	c.438T>C	c.937+4A>T	c.2779delG			
266		disease range	c.755C>T	c.5979dupA				
267		disease range	c.3041A>G	c.5526-1G>A				
268		normal range	c.3388G>A	-				
269		disease range	c.5077C>T	c.5698_5699delAG *				
270		carrier range	c.5216C>A	-				
271	C15	disease range	c.4060_4062delTCC	c.4439A>C				
272		disease range	c.2643+1G>A *	c.5077C>T				
273		disease range	c.5713C>T	c.1343T>C				
274	C7	normal range	c.469G>A *	-				
275		disease range	c.1096delA	c.1096delA				
276		disease range	c.487C>T	c.5979dupA				
277	B9	disease range	c.1053+1G>A *	del Ex25_29: c.2512_3174del *				
278		disease range	c.5077C>T	c.5698_5699delAG				
279		carrier range	ND	ND				
280		disease range	c.353delT	c.5444G>T				
281	A5	disease range	c.757C>T *	c.2894G>A *				
282		disease range	c.4168-1G>T	c.4168-1G>T	c.4090C>T *	c.4090C>T *		
283		normal range	c.4794G>T *	-				
284		normal range	c.1064A>G	c.2408G>A	c.2902A>T			
285		disease range	c.2643+1G>A *	c.2643+1G>A *				

Table A.1 continued

286		disease range	c.757C>T *	c.5444G>T				
287		disease range	c.2779delG	c.2779delG				
288		carrier range	ND	ND				
289		disease range	c.3041A>G	c.3041A>G				
290		disease range	c.5429G>A	c.5429G>A				
291		disease range	c.5302C>T	c.5302C>T	c.2452C >T			
292		disease range	c.5979dupA	c.5979dupA				
293		disease range	c.5884C>T	c.4199C>G	c.5026G >T			
294		disease range	c.1368C>A	c.5713C>T				
295		disease range	c.1094delA	c.6124C>T *				
296		disease range	C.764A>C	c.393_394delCC				
297		disease range	c.1368C>A	c.1368C>A				
298		disease range	c.1354-3_1354-2delCA	c.1354-3_1354-2delCA				
299		disease range	c.1368C>A	c.1368C>A				
300		disease range	c.1368C>A	c.1368C>A				
301		disease range	c.5979dupA	c.5979dupA				
302		disease range	c.5979dupA	c.5979dupA				
303		disease range	c.5979dupA	c.5979dupA				
304		disease range	ND	ND				
305		disease range	ND	ND				
306	C8	normal range	c.774C>G *	c.2902A>T *				
308	B10	carrier range	c.6124C>T *	c.4794G>T*				
309		normal range	-	-				
310		disease range	c.4434G>A	del Ex4 *				
311		disease range	c.2643+1G>A *	c.2643+1G>A *				
312		normal range	-	-				
313	B11	disease range	c.3517dupT *	c.5836_5839delCAGC *				
314	A6	disease range	c.673C>T *	c.673C>T *				
315		normal range	-	-				
316		disease range	c.4638+1G>A	c.757C>T *				
317	B12	disease range	c.3805dupG *	c.907-3C>A*				
325		disease range	c.1053+1G>A *	c.3517dupT				
326		normal range	ND	ND				

Table A.1 continued

327		disease range	c.5668-824C>T *	c.6124C>T *				
328		disease range	del Ex6	c.1639-6T>A *				
329		disease range	del Ex6	c.1639-6T>A				
330		disease range	c.3538delT	c.5366T>C				
331		disease range	c.4253G>A *	c.4253G>A *	c.4943A>G			
332		disease range	c.4253G>A *	c.4253G>A *				
333		disease range	c.1368C>A	c.1368C>A				
334	C9	disease range	c.2790G>C *	c.4024C>T *	c.4526T>G *			
335		disease range	c.3137G>A *	c.6038C>G				
336		carrier range	ND	ND				
337		disease range	c.1368C>A	c.1368C>A				
338		disease range	c.2643+1G>A *	c.5836_5839delCAGC *				
339		disease range	c.1053+1G>A *	c.3512_3513insT				
340	C16	disease range	-	-				
342	A11	carrier range	c.1392dupA *	c.1481-1G>A *				
343		disease range	c.5836_5839delCAGC *	c.5836_5839delCAGC *				
344		disease range	c.1834C>T *	c.465delA				
345	A13	disease range	c.2643+1G>A *	c.4577A>C*	c.5668-7G>A *			Pathogenic *DYSF* variant combination out of the 3 variants was not known
346		disease range	c.1392dupA *	c.3516_3517delTT				
347		disease range	c.533delG *	c.1861G>A *				
348		disease range	c.5668-7G>A *	c.5668-7G>A *				
349	A15	disease range	c.1171_1180+4dup14 *	c.1171_1180+4dup14 *				
350		disease range	c.3759_3768del10	c.2901_2904delCATG				
351	A9	disease range	c.533delG *	c.1861G>A *				
352		normal range	-	-			*FKRP* Homozygous c.826C>A pathogenic variant	Confirmed LGMD2I (*FKRP*)
353		normal range	c.4376A>G	-				
354		disease range	c.1053+1G>A *	c.610C>T				

Table A.1 continued

355		disease range	c.2643+1G>A *	c.2095C>T				
356	B8	disease range	c.5059T>C	c.4577A>C *	c.2643+1G>A *			
357		disease range	c.5429+2T>A	c.5429+2T>A				
358	B25	disease range	c.5181delA	c.1668_1669insGTT				
359		normal range	c.2423G>A	-				
360		disease range	c.2779delG	c.5594delG				
361	B27	disease range	c.5979dupA	c.5057+5G>A				
362	B29	disease range	c.5979dupA	-				
363		disease range	del Ex38	c.799_800delTT				
364		carrier range	c.857T>A	c.4794G>T *				
365		disease range	c.2643+1G>A *	c.1948delC	c.4577A>C *			
366		carrier range	-	-				
367		disease range	ND	ND				
368		ND	c.353delT	c.5668-824C>T *				
369	B13	ND	c.6124C>T *	c.5509G>A *				
370	A8	disease range	c.5509G>A *	c.5903G>A *				
371	B14	disease range	c.2163-2A>G *	del Ex23_24 *				
372	B15	disease range	c.4360G>T *	c.4756C>T *				
373	B16	disease range	c.5159delG *	c.125dupT *				
374		disease range	c.5159delG *	c.125dupT *				
375	A7	disease range	c.5668-824C>T *	c.5698_5699delAG *				
376	B17	disease range	c.863dupA *	c.3031+2T>C *				
377		ND	c.5871_5872delGT	c.5668-824C>T *				
378	B18	ND	c.2810+1G>A *	c.2811-20T>G *				
379		disease range	c.5698_5699delAG *	c.5668-824C>T *				
380	B21	disease range	c.4756C>T *	c.3113G>C *	c.3065G>A*			
381	C10	normal range	c.3065G>A*	c.3992G>T*			* *COL6A2* exon 25: c.1861G>A ; *COL6A2* exon 28: c.2893C>T	RNA-Seq identified variants in *COL6A2* found so likely has Bethlem myopathy
382		normal range	c.6124C>T *	c.5768-1G>C *				
383	B22	carrier range	c.6124C>T *	c.5768-1G>C *				
384	B2	ND	c.5022delT *	c.401C>T *	c.6196G>A *			
385	B20	ND	c.1663C>T *	c.1004G>C*	c.509C>A *			

Table A.1 continued

386	B19	disease range	c.4090C>T *	c.5296G>A*				
387	B23	disease range	c.1639-6T>A *	c.5503A>G*				
388		disease range	c.509C>A	c.5836_5839delCAGC *				
389	C14	ND	c.3904_4410del *	c.3904_4410del *				
390	B24	disease range	c.3112C>T *	c.3191_3196dupCGGAGG *	c.1180+5G>A *			
391	B28	ND	c.2077delC	c.4334-3C>A				
392	C11	disease range	c.2643+5G>A *	c.3113G>A *				
393	B30	ND	c.5429+1G>T	c.5057+5G>T				
394		ND	c.3516_3517del	c.4411-5C>G				
395	A16	ND	c.3112C>T	c.4577A>C	c.2643+1G>A			
396	A17	ND	c.3517dupT	c.3113G>A				
397	A10	ND	c.3041A>G	c.3041A>G	c.4820T>C	c.4820T>C		
398	B26	disease range	c.855+1delG	c.3031G>C				
399		ND	c.863dupA	c.3031+2T>C				
400	C17	ND	c.896G>A	c.1877T>C				
401	A18	ND	c.5429G>A	c.5429G>A				
402	A19	ND	c.3444T>A	c.4756C>T	c.3445G>A			Pathogenic *DYSF* combination of variants out of the three variants was not known due to lack of knowledge in phasing
403		ND	c.2077delC	c.3121C>T	c.6056G>A			Pathogenic *DYSF* combination of variants out of the three variants was not known due to lack of knowledge in phasing
404	A20	ND	del Ex25_29: c.2512_3174del *	c.2077delC				
405	C18	ND	c.3065G>A	c.4003G>A				
406	A21	disease range	c.2071C>T	c.3113G>A				

Table A.1 continued

100 (Control) Ethnicity : India		normal range	Normal Positive Control					
132 (control)		normal range	Normal Positive Control					
133 (Control) Ethnicity : Caucasian		normal range	Normal Positive Control					
134 (Control) Ethnicity : Caucasian		normal range	Normal Positive Control					
136 (Control) Ethnicity : Caucasian		normal range	Normal Positive Control					
137 (Control, Ethnicity : Caucasian)		normal range	Normal Positive Control					
216 (Ethnicity: Indian (south east Asian))		normal range	Normal Positive Control					
221 (Control) Ethnicity : South east Asian (Indian subcontinent)		normal range	Normal Positive Control					
222 (Control) Ethnicity : South east Asian (Indian subcontinent)		normal range	Normal Positive Control					
260 (Control) Ethnicity : South-East Asian (Indian subcontinent)		normal range	Normal Positive Control					

Table A.1 continued

307 (Control) Ethnicity : Caucasian		normal range	Normal Positive Control					
320 (Control) Ethnicity : Hispanic		normal range	Normal Positive Control					
321 (Control) Ethnicity : Caucasian (Australia)		normal range	Normal Positive Control					
322 (Control) Ethnicity : Caucasian (Spain)		normal range	Normal Positive Control					
324 (Control) Ethnicity : Caucasian		normal range	Normal Positive Control					

	P/LP variants
	VUS determined to be P/LP by RNA-seq
	Variants determined to be benign
	VUS
"-"	**no reportable *DYSF* variant were found in genetic testing. ND = not determined**
ND	**Not Determined. For genotype, information was not available. For %DYSF by monocyte assay, ND means informed consent or blood sample was not provided.**
Asterisk (*)	**Variant Visualized in RNA-seq**
	Confirmed diagnosis or close to diagnosis of Other Gene (not DYSF)

REFERENCES

1. Rankin J, Auer-Grumbach M, Bagg W, Colclough K, Duong NT, Fenton-May J, et al. Extreme phenotypic diversity and nonpenetrance in families with the LMNA gene mutation R644C. American journal of medical genetics Part A. 2008;146(12):1530-42.

2. Whicher D, Philbin S, Aronson N. An overview of the impact of rare disease characteristics on research methodology. Orphanet journal of rare diseases. 2018;13(1):1-12.

3. Navarrete-Opazo AA, Singh M, Tisdale A, Cutillo CM, Garrison SR. Can you hear us now? The impact of health-care utilization by rare disease patients in the United States. Genetics in Medicine. 2021;23(11):2194-201.

4. Grosse SD, Thompson JD, Ding Y, Glass M. The use of economic evaluation to inform newborn screening policy decisions: The Washington state experience. The Milbank Quarterly. 2016;94(2):366-91.

5. López-Bastida J, Oliva-Moreno J. Cost of illness and economic evaluation in rare diseases. Rare diseases epidemiology. 2010:273-82.

6. Angelis A, Tordrup D, Kanavos P. Socio-economic burden of rare diseases: a systematic review of cost of illness evidence. Health Policy. 2015;119(7):964-79.

7. Mazzucato M, Visonà Dalla Pozza L, Manea S, Minichiello C, Facchin P. A population-based registry as a source of health indicators for rare diseases: the ten-year experience of the Veneto Region's rare diseases registry. Orphanet journal of rare diseases. 2014;9(1):1-12.

8. Ferreira CR. The burden of rare diseases. American Journal of Medical Genetics Part A. 2019;179(6):885-92.

9. Bell SA, Tudur Smith C. A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of ClinicalTrials. gov. Orphanet journal of rare diseases. 2014;9(1):1-11.

10. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nature Reviews Genetics. 2018;19(5):253-68.

11. Liu Z, Zhu L, Roberts R, Tong W. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we? Trends in genetics. 2019;35(11):852-67.

12. Directors ABo. Clinical utility of genetic and genomic services: a position statement of the American College of Medical Genetics and Genomics. Genetics in medicine: official journal of the American College of Medical Genetics. 2015;17(6):505-7.

13. Fanen P, Wohlhuter-Haddad A, Hinzpeter A. Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. The international journal of biochemistry & cell biology. 2014;52:94-102.

14. Esquivel-Sada D, Nguyen MT. Diagnosis of rare diseases under focus: impacts for Canadian patients. Journal of community genetics. 2018;9(1):37-50.

15. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89.

16. Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. New england journal of medicine. 2018;379(22):2131-9.

17. Wu AC, McMahon P, Lu C. Ending the diagnostic odyssey—is whole-genome sequencing the answer? JAMA pediatrics. 2020;174(9):821-2.

18. Von der Lippe C, Diesen PS, Feragen KB. Living with a rare disorder: a systematic review of the qualitative literature. Molecular Genetics & Genomic Medicine. 2017;5(6):758-73.

19. Blöß S, Klemann C, Rother A-K, Mehmecke S, Schumacher U, Mücke U, et al. Diagnostic needs for rare diseases and shared prediagnostic phenomena: results of a German-wide expert Delphi survey. PLoS One. 2017;12(2):e0172532.

20. Zurynski Y, Deverell M, Dalkeith T, Johnson S, Christodoulou J, Leonard H, et al. Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. Orphanet Journal of Rare Diseases. 2017;12(1):1-9.

21. Bryson B, Bogart K, Atwood M, Fraser K, Locke T, Pugh K, et al. Navigating the unknown: A content analysis of the unique challenges faced by adults with rare diseases. Journal of Health Psychology. 2021;26(5):623-35.

22. Lasker JN, Sogolow ED, Sharim RR. The role of an online community for people with a rare disease: content analysis of messages posted on a primary biliary cirrhosis mailinglist. Journal of medical Internet research. 2005;7(1):e137.

23. Bogart KR, Irvin VL. Health-related quality of life among adults with diverse rare disorders. Orphanet journal of rare diseases. 2017;12(1):1-9.

24. Dimmock D, Caylor S, Waldman B, Benson W, Ashburner C, Carmichael JL, et al. Project Baby Bear: Rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care. The American Journal of Human Genetics. 2021.

25. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, Van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet. 2015;385(9975):1305-14.

26. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic acids research. 2017;45(D1):D840-D5.

27. Karczewski K, Francioli L. The genome aggregation database (gnomAD). MacArthur Lab. 2017.

28. Bahcall OG. ExAC boosts clinical variant interpretation in rare diseases. Nature reviews Genetics. 2016;17(10):584-.

29. MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, Abecasis G, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.

30. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. Nature Reviews Genetics. 2013;14(7):460-70.

31. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Science translational medicine. 2011;3(65):65ra4-ra4.

32. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine. 2015;17(5):405-23.

33. Bean LJ, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome medicine. 2017;9(1):1-3.

34. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42(D1):D980-D5.

35. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. American journal of human genetics. 2007;80(4):588.

36. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2016;44(D1):D862-D8.

37. Ormondroyd E, Mackley MP, Blair E, Craft J, Knight JC, Taylor J, et al. Insights from early experience of a Rare Disease Genomic Medicine Multidisciplinary Team: a qualitative study. European Journal of Human Genetics. 2017;25(6):680-6.

38. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. science. 2001;291(5507):1304-51.

39. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nature Reviews Genetics. 2010;11(8):559-71.

40. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Experimental & molecular medicine. 2018;50(8):1-8.

41. Simmonds N. Is it cystic fibrosis? The challenges of diagnosing cystic fibrosis. Paediatric respiratory reviews. 2019;31:6-8.

42. Ankala A, da Silva C, Gualandi F, Ferlini A, Bean LJ, Collins C, et al. A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. Annals of neurology. 2015;77(2):206-14.

43. Curry PD, Broda KL, Carroll CJ. The Role of RNA-Sequencing as a New Genetic Diagnosis Tool. Current Genetic Medicine Reports. 2021:1-9.

44. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics. 2017;18(10):599-612.

45. Phillips KA, Deverka PA, Trosman JR, Douglas MP, Chambers JD, Weldon CB, et al. Payer coverage policies for multigene tests. Nature biotechnology. 2017;35(7):614-7.

46. Grosse SD, Rasmussen SA. Exome sequencing: value is in the eye of the beholder. Genetics in Medicine. 2020;22(2):280-2.

47. Douglas MP, Parker SL, Trosman JR, Slavotinek AM, Phillips KA. Private payer coverage policies for exome sequencing (ES) in pediatric patients: trends over time and analysis of evidence cited. Genetics in Medicine. 2019;21(1):152-60.

48. Grant P, Langlois S, Lynd LD, Study G, Austin JC, Elliott AM, et al. Out-of-pocket and private pay in clinical genetic testing: a scoping review. Clinical Genetics. 2021.

49. Marshall CR, Bick D, Belmont JW, Taylor SL, Ashley E, Dimmock D, et al. The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. Genome Medicine. 2020;12(1):1-4.

50. Rockowitz S, LeCompte N, Carmack M, Quitadamo A, Wang L, Park M, et al. Children's rare disease cohorts: an integrative research and clinical genomics initiative. NPJ genomic medicine. 2020;5(1):1-12.

51. Lord J, Baralle D. Splicing in the diagnosis of rare disease: advances and challenges. Frontiers in Genetics. 2021;12:1146.

52. Wang Y, Liu J, Huang B, Xu YM, Li J, Huang LF, et al. Mechanism of alternative splicing and its regulation. Biomedical reports. 2015;3(2):152-8.

53. Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48. e24.

54. Caminsky N, Mucaki EJ, Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research. 2014;3.

55. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Human genetics. 1992;90(1):41-54.

56. Truty R, Ouyang K, Rojahn S, Garcia S, Colavin A, Hamlington B, et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. The American Journal of Human Genetics. 2021;108(4):696-708.

57. Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, et al. Pathogenicity and selective constraint on variation near splice sites. Genome research. 2019;29(2):159-70.

58. Rowlands C, Thomas HB, Lord J, Wai HA, Arno G, Beaman G, et al. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. Scientific reports. 2021;11(1):1-11.

59. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, et al. Annotating pathogenic non-coding variants in genic regions. Nature communications. 2017;8(1):1-11.

60. Tesoriero A, Wong E, Jenkins M, Hopper J, Brown M, Chenevix-Trench G, et al. Molecular characterization and cancer risk associated with BRCA1 and BRCA2 splice site variants identified in multiple-case breast cancer families. Human mutation. 2005;26(5):495-.

61. De La Hoya M, Soukarieh O, López-Perolio I, Vega A, Walker LC, van Ierland Y, et al. Combined genetic and splicing analysis of BRCA1 c.[594-2A> C; 641A> G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Human molecular genetics. 2016;25(11):2256-68.

62. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science translational medicine. 2017;9(386).

63. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nature medicine. 2019;25(6):911-9.

64. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome biology. 2020;21(1):1-16.

65. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome biology. 2016;17(1):1-19.

66. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. Frontiers in genetics. 2019;10:426.

67. Mertes C, Scheller IF, Yépez VA, Çelik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. Nature communications. 2021;12(1):1-13.

68. Schlieben LD, Prokisch H, Yépez VA. How machine learning and statistical models advance molecular diagnostics of rare disorders via analysis of RNA sequencing data. Frontiers in Molecular Biosciences. 2021;8.

69. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods. 2010;7(12):1009-15.

70. Li Y, Rao X, Mattox WW, Amos CI, Liu B. RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. PloS one. 2015;10(9):e0136653.

71. Ferraro NM, Strober BJ, Einson J, Abell NS, Aguet F, Barbeira AN, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. Science. 2020;369(6509).

72. Jenkinson G, Li YI, Basu S, Cousin MA, Oliver GR, Klee EW. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. Bioinformatics. 2020;36(17):4609-15.

73. Mehmood A, Laiho A, Venäläinen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. Briefings in bioinformatics. 2020;21(6):2052-65.

74. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. The American Journal of Human Genetics. 2019;104(3):466-83.

75. Gautam A, Donohue D, Hoke A, Miller SA, Srinivasan S, Sowe B, et al. Investigating gene expression profiles of whole blood and peripheral blood mononuclear cells using multiple collection and processing methods. PLoS One. 2019;14(12):e0225137.

76. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al. Measure transcript integrity using RNA-seq data. BMC bioinformatics. 2016;17(1):1-16.

77. Lissner MM, Thomas BJ, Wee K, Tong A-J, Kollmann TR, Smale ST. Age-related gene expression differences in monocytes from human neonates, young adults, and older adults. PloS one. 2015;10(7):e0132061.

78. Laing NG. Genetics of neuromuscular disorders. Critical reviews in clinical laboratory sciences. 2012;49(2):33-48.

79. Nguyen K, Bassez G, Bernard R, Krahn M, Labelle V, Figarella-Branger D, et al. Dysferlin mutations in LGMD2B, Miyoshi myopathy, and atypical dysferlinopathies. Human mutation. 2005;26(2):165-.

80. Echevarría L, Aupy P, Goyenvalle A. Exon-skipping advances for Duchenne muscular dystrophy. Human molecular genetics. 2018;27(R2):R163-R72.

81. Bousfiha A, Jeddane L, Picard C, Ailal F, Gaspar HB, Al-Herz W, et al. The 2017 IUIS phenotypic classification for primary immunodeficiencies. Journal of clinical immunology. 2018;38(1):129-43.

82. Picard C, Gaspar HB, Al-Herz W, Bousfiha A, Casanova J-L, Chatila T, et al. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. Journal of clinical immunology. 2018;38(1):96-128.

83. Condino-Neto A, Espinosa-Rosales FJ. Changing the lives of people with primary immunodeficiencies (PI) with early testing and diagnosis. Frontiers in immunology. 2018;9:1439.

84. Modell V, Orange JS, Quinn J, Modell F. Global report on primary immunodeficiencies: 2018 update from the Jeffrey Modell Centers Network on disease classification, regional trends, treatment modalities, and physician reported outcomes. Immunologic research. 2018;66(3):367-80.

85. Rubin Z, Pappalardo A, Schwartz A, Antoon JW. Prevalence and outcomes of primary immunodeficiency in hospitalized children in the United States. The Journal of Allergy and Clinical Immunology: In Practice. 2018;6(5):1705-10. e1.

86. De Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature genetics. 2017;49(2):256-61.

87. Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. Inflammatory bowel diseases. 2015;21(6):1428-34.

88. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics. 2011;43(11):1066-73.

89. Muise AM, Snapper SB, Kugathasan S. The age of gene discovery in very early onset inflammatory bowel disease. Gastroenterology. 2012;143(2):285-8.

90. Uhlig HH, Schwerd T, Koletzko S, Shah N, Kammermeier J, Elkadri A, et al. The diagnostic approach to monogenic very early onset inflammatory bowel disease. Gastroenterology. 2014;147(5):990-1007. e3.

91. Moran CJ, Walters TD, Guo C-H, Kugathasan S, Klein C, Turner D, et al. IL-10R polymorphisms are associated with very-early-onset ulcerative colitis. Inflammatory bowel diseases. 2013;19(1):115-23.

92. Ouahed J, Spencer E, Kotlarz D, Shouval DS, Kowalik M, Peng K, et al. Very early onset inflammatory bowel disease: a clinical approach with a focus on the role of genetics and underlying immune deficiencies. Inflammatory bowel diseases. 2020;26(6):820-42.

93. Sens J, Hoffmann D, Lange L, Morgan M, Falk C, Schambach A. Knock-out iPSCs for disease and therapy modeling of IL-10 associated primary immunodeficiencies. Human Gene Therapy. 2020.

94. Chakravorty S, Berger K, Arafat D, Nallamilli BRR, Subramanian HP, Joseph S, et al. Clinical utility of RNA sequencing to resolve unusual GNE myopathy with a novel promoter deletion. Muscle & nerve. 2019;60(1):98-103.

95. Wang X, Zhang B. Integrating genomic, transcriptomic, and interactome data to improve Peptide and protein identification in shotgun proteomics. J Proteome Res. 2014;13(6):2715-23.

96. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506-11.

97. Chakravorty S, Hegde M. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. Annu Rev Genomics Hum Genet. 2017.

98. Askanas V, Engel WK. New advances in the understanding of sporadic inclusion-body myositis and hereditary inclusion-body myopathies. Curr Opin Rheumatol. 1995;7(6):486-96.

99. Eisenberg I, Avidan N, Potikha T, Hochner H, Chen M, Olender T, et al. The UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase gene is mutated in recessive hereditary inclusion body myopathy. Nat Genet. 2001;29(1):83-7.

100. Huizing M, Krasnewich DM. Hereditary inclusion body myopathy: a decade of progress. Biochim Biophys Acta. 2009;1792(9):881-7.

101. Huizing M, Malicdan MCV, Krasnewich DM, Manoli I, Carrillo-Carrasco N. GNE Myopathy. In: Beaudet AL, Vogelstein B, Kinzler KW, Antonarakis SE, Ballabio A, Gibson KM, et al., editors. The Online Metabolic and Molecular Bases of Inherited Disease. New York, NY: The McGraw-Hill Companies, Inc.; 2014.

102. Argov Z, Yarom R. "Rimmed vacuole myopathy" sparing the quadriceps. A unique disorder in Iranian Jews. J Neurol Sci. 1984;64(1):33-43.

103. Sivakumar K, Dalakas MC. The spectrum of familial inclusion body myopathies in 13 families and a description of a quadriceps-sparing phenotype in non-Iranian Jews. Neurology. 1996;47(4):977-84.

104. Mori-Yoshimura M, Oya Y, Hayashi YK, Noguchi S, Nishino I, Murata M. Respiratory dysfunction in patients severely affected by GNE myopathy (distal myopathy with rimmed vacuoles). Neuromuscul Disord. 2013;23(1):84-8.

105. Chai Y, Bertorini TE, McGrew FA. Hereditary inclusion-body myopathy associated with cardiomyopathy: report of two siblings. Muscle Nerve. 2011;43(1):133-6.

106. Kimpara T, Imamura T, Tsuda T, Sato K, Tsuburaya K. [Distal myopathy with rimmed vacuoles and sudden death--report of two siblings]. Rinsho Shinkeigaku. 1993;33(8):886-90.

107. Nishino I, Malicdan MC, Murayama K, Nonaka I, Hayashi YK, Noguchi S. Molecular pathomechanism of distal myopathy with rimmed vacuoles. Acta Myol. 2005;24(2):80-3.

108. Hinderlich S, Stasche R, Zeitler R, Reutter W. A bifunctional enzyme catalyzes the first two steps in N-acetylneuraminic acid biosynthesis of rat liver. Purification and characterization of UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase. J Biol Chem. 1997;272(39):24313-8.

109. Varki A. Selectins and other mammalian sialic acid-binding lectins. Curr Opin Cell Biol. 1992;4(2):257-66.

110. Varki A. Diversity in the sialic acids. Glycobiology. 1992;2(1):25-40.

111. Varki A. Sialic acids as ligands in recognition phenomena. FASEB J. 1997;11(4):248-55.

112. Noguchi S, Keira Y, Murayama K, Ogawa M, Fujita M, Kawahara G, et al. Reduction of UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase activity and sialylation in distal myopathy with rimmed vacuoles. J Biol Chem. 2004;279(12):11402-7.

113. Sparks SE, Ciccone C, Lalor M, Orvisky E, Klootwijk R, Savelkoul PJ, et al. Use of a cell-free system to determine UDP-N-acetylglucosamine 2-epimerase and N-acetylmannosamine kinase activities in human hereditary inclusion body myopathy. Glycobiology. 2005;15(11):1102-10.

114. Malicdan MC, Noguchi S, Nonaka I, Hayashi YK, Nishino I. A Gne knockout mouse expressing human GNE D176V mutation develops features similar to distal myopathy with rimmed vacuoles or hereditary inclusion body myopathy. Hum Mol Genet. 2007;16(22):2669-82.

115. Malicdan MC, Noguchi S, Nonaka I, Hayashi YK, Nishino I. A Gne knockout mouse expressing human V572L mutation develops features similar to distal myopathy with rimmed vacuoles or hereditary inclusion body myopathy. Hum Mol Genet. 2007;16(2):115-28.

116. Bean LJ, Tinker SW, da Silva C, Hegde MR. Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. Hum Mutat. 2013;34(9):1183-8.

117. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4.

118. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

119. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

120. Chin EL, da Silva C, Hegde M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genet. 2013;14:6.

121. Jones MA, Bhide S, Chin E, Ng BG, Rhodenizer D, Zhang VW, et al. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med. 2011;13(11):921-32.

122. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.

123. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.

124. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 2016;17(1):103.

125. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

126. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotech. 2011;29(1):24-6.

127. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

128. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.

129. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

130. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Research. 2015;4:1070.

131. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368-75.

132. Hegde MR, Chin EL, Mulle JG, Okou DT, Warren ST, Zwick ME. Microarray-based mutation detection in the dystrophin gene. Hum Mutat. 2008;29(9):1091-9.

133. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.

134. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267-73.

135. Berg AT, Chakravorty S, Koh S, Grinspan ZM, Shellhaas RA, Saneto RP, et al. Why West? Comparisons of clinical, genetic and molecular features of infants with and without spasms. PLoS One. 2018;13(3):e0193599.

136. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

137. Voermans NC, Guillard M, Doedee R, Lammens M, Huizing M, Padberg GW, et al. Clinical features, lectin staining, and a novel GNE frameshift mutation in hereditary inclusion body myopathy. Clin Neuropathol. 2010;29(2):71-7.

138. Nalini A, Gayathri N, Nishino I, Hayashi YK. GNE myopathy in India. Neurol India. 2013;61(4):371-4.

139. Liewluck T, Pho-Iam T, Limwongse C, Thongnoppakhun W, Boonyapisit K, Raksadawan N, et al. Mutation analysis of the GNE gene in distal myopathy with rimmed vacuoles (DMRV) patients in Thailand. Muscle Nerve. 2006;34(6):775-8.

140. Tanboon J, Rongsa K, Pithukpakorn M, Boonyapisit K, Limwongse C, Sangruchi T. A Novel Mutation of the GNE Gene in Distal Myopathy with Rimmed Vacuoles: A Case with Inflammation. Case Rep Neurol. 2014;6(1):55-9.

141. Chaouch A, Brennan KM, Hudson J, Longman C, McConville J, Morrison PJ, et al. Two recurrent mutations are associated with GNE myopathy in the North of Britain. J Neurol Neurosurg Psychiatry. 2014;85(12):1359-65.

142. Nishino I, Carrillo-Carrasco N, Argov Z. GNE myopathy: current update and future therapy. J Neurol Neurosurg Psychiatry. 2015;86(4):385-92.

143. Celeste FV, Vilboux T, Ciccone C, de Dios JK, Malicdan MC, Leoyklang P, et al. Mutation update for GNE gene variants associated with GNE myopathy. Hum Mutat. 2014;35(8):915-26.

144. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

145. Hinderlich S, Weidemann W, Yardeni T, Horstkorte R, Huizing M. UDP-GlcNAc 2-Epimerase/ManNAc Kinase (GNE): A Master Regulator of Sialic Acid Synthesis. Top Curr Chem. 2015;366:97-137.

146. Garland J, Stephen J, Class B, Gruber A, Ciccone C, Poliak A, et al. Identification of an Alu element-mediated deletion in the promoter region of GNE in siblings with GNE myopathy. Molecular Genetics & Genomic Medicine. 2017;5(4):410-7.

147. Pogoryelova O, Cammish P, Mansbach H, Argov Z, Nishino I, Skrinar A, et al. Phenotypic stratification and genotype–phenotype correlation in a heterogeneous, international cohort of GNE myopathy patients: first report from the GNE myopathy disease monitoring program, registry portion. Neuromuscular Disorders. 2018;28(2):158-68.

148. Zhu W, Mitsuhashi S, Yonekawa T, Noguchi S, Huei JCY, Nalini A, et al. Missing genetic variations in GNE myopathy: rearrangement hotspots encompassing 5′ UTR and founder allele. Journal of human genetics. 2017;62(2):159-66.

149. Chabot B, Shkreta L. Defective control of pre–messenger RNA splicing in human disease. Journal of Cell Biology. 2016;212(1):13-27.

150. Bell SC, De Boeck K, Amaral MD. New pharmacological approaches for cystic fibrosis: promises, progress, pitfalls. Pharmacology & therapeutics. 2015;145:19-34.

151. Chen J, Weiss W. Alternative splicing in cancer: implications for biology and therapy. Oncogene. 2015;34(1):1-14.

152. Pistoni M, Ghigna C, Gabellini D. Alternative splicing and muscular dystrophy. RNA biology. 2010;7(4):441-52.

153. Vidak S, Foisner R. Molecular insights into the premature aging disease progeria. Histochemistry and cell biology. 2016;145(4):401-17.

154. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218).

155. Biamonti G, Bonomi S, Gallo S, Ghigna C. Making alternative splicing decisions during epithelial-to-mesenchymal transition (EMT). Cellular and Molecular Life Sciences. 2012;69(15):2515-26.

156. Love JE, Hayden EJ, Rohn TT. Alternative splicing in Alzheimer's disease. Journal of Parkinson's disease and Alzheimer's disease. 2015;2(2).

157. Pellagatti A, Armstrong RN, Steeples V, Sharma E, Repapi E, Singh S, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood. 2018;132(12):1225-40.

158. Rabin SJ, Kim JMH, Baughn M, Libby RT, Kim YJ, Fan Y, et al. Sporadic ALS has compartment-specific aberrant exon splicing and altered cell–matrix adhesion biology. Human molecular genetics. 2010;19(2):313-28.

159. Udd B, Krahe R. The myotonic dystrophies: molecular, clinical, and therapeutic challenges. The Lancet Neurology. 2012;11(10):891-905.

160. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet. 2017;390(10114):2769-78.

161. Furey TS, Sethupathy P, Sheikh SZ. Redefining the IBDs using genome-scale molecular phenotyping. Nature Reviews Gastroenterology & Hepatology. 2019;16(5):296-311.

162. Weiser M, Simon JM, Kochar B, Tovar A, Israel JW, Robinson A, et al. Molecular classification of Crohn's disease reveals two clinically relevant subtypes. Gut. 2018;67(1):36-42.

163. Digby-Bell JL, Atreya R, Monteleone G, Powell N. Interrogating host immunity to predict treatment response in inflammatory bowel disease. Nature Reviews Gastroenterology & Hepatology. 2020;17(1):9-20.

164. Lee JC, Lyons PA, McKinney EF, Sowerby JM, Carr EJ, Bredin F, et al. Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. The Journal of clinical investigation. 2011;121(10):4170-9.

165. Hyams JS, Thomas SD, Gotman N, Haberman Y, Karns R, Schirmer M, et al. Clinical and biological predictors of response to standardised paediatric colitis therapy (PROTECT): a multicentre inception cohort study. The Lancet. 2019;393(10182):1708-20.

166. Kugathasan S, Denson LA, Walters TD, Kim M-O, Marigorta UM, Schirmer M, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. The Lancet. 2017;389(10080):1710-8.

167. Marigorta UM, Denson LA, Hyams JS, Mondal K, Prince J, Walters TD, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. Nature genetics. 2017;49(10):1517.

168. West NR, Hegazy AN, Owens BM, Bullers SJ, Linggi B, Buonocore S, et al. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. Nature medicine. 2017;23(5):579-89.

169. Mo A, Krishnakumar C, Arafat D, Dhere T, Iskandar H, Dodd A, et al. African Ancestry Proportion Influences Ileal Gene Expression in Inflammatory Bowel Disease. Cellular and molecular gastroenterology and hepatology. 2020;10(1):203-5.

170. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative splicing signatures in RNA-seq data: Percent spliced in (PSI). Current protocols in human genetics. 2015;87(1):11.6. 1-.6. 4.

171. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

172. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

173. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research. 2019;47(D1):D766-D73.

174. Beauchemin N, Draber P, Dveksler G, Gold P, Gray-Owen S, Grunert F, et al. Redefined nomenclature for members of the carcinoembryonic antigen family. Experimental cell research. 1999;252(2):243-9.

175. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC bioinformatics. 2015;16(1):1-7.

176. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559-75.

177. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506(7487):185-90.

178. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43(1):11.0. 1-.0. 33.

179. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, et al. Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. Cell. 2019;178(3):714-30. e22.

180. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015;67(1):48.

181. Venkateswaran S, Marigorta UM, Denson LA, Hyams JS, Gibson G, Kugathasan S. Bowel location rather than disease subtype dominates transcriptomic heterogeneity in pediatric IBD. Cellular and molecular gastroenterology and hepatology. 2018;6(4):474.

182. Mowat AM, Agace WW. Regional specialization within the intestinal immune system. Nature Reviews Immunology. 2014;14(10):667-85.

183. Kelleher M, Singh R, O'Driscoll CM, Melgar S. Carcinoembryonic antigen (CEACAM) family members and inflammatory bowel disease. Cytokine & growth factor reviews. 2019;47:21-31.

184. Nagaishi T, Chen Z, Chen L, Iijima H, Nakajima A, Blumberg R. CEACAM1 and the regulation of mucosal inflammation. Mucosal immunology. 2008;1(1):S39-S42.

185. Chen L, Chen Z, Baker K, Halvorsen EM, da Cunha AP, Flak MB, et al. The short isoform of the CEACAM1 receptor in intestinal T cells regulates mucosal immunity and homeostasis via Tfh cell induction. Immunity. 2012;37(5):930-46.

186. Dery KJ, Gaur S, Gencheva M, Yen Y, Shively JE, Gaur RK. Mechanistic control of carcinoembryonic antigen-related cell adhesion molecule-1 (CEACAM1) splice isoforms by the heterogeneous nuclear ribonuclear proteins hnRNP L, hnRNP A1, and hnRNP M. Journal of Biological Chemistry. 2011;286(18):16039-51.

187. Glas J, Seiderer J, Fries C, Tillack C, Pfennig S, Weidinger M, et al. CEACAM6 gene variants in inflammatory bowel disease. PLoS One. 2011;6(4):e19319.

188. Barrett JS, Irving P, Shepherd SJ, Muir JG, Gibson PR. Comparison of the prevalence of fructose and lactose malabsorption across chronic intestinal disorders. Alimentary pharmacology & therapeutics. 2009;30(2):165-74.

189. Neugebauer KM. Nascent RNA and the Coordination of Splicing with Transcription. Cold Spring Harbor perspectives in biology. 2019;11(8):a032227.

190. Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Human genomics. 2014;8(1):1-6.

191. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology. 2013;31(1):46-53.

192. Illumina Technical Support. Considerations for RNA-Seq read length and coverage [Internet]. 2017 [updated 2020 Nov 11. Available from: https://support.illumina.com/bulletins/2017/04/considerations-for-rna-seq-read-length-and-coverage-.html.

193. Chakravorty S, Berger K, Rufibach L, Gloster L, Emmons S, Shenoy S, et al. Combinatorial clinically driven blood biomarker functional genomics significantly enhances genotype-phenotype resolution and diagnostics in neuromuscular disease. medRxiv. 2021.

194. Nallamilli BRR, Chakravorty S, Kesari A, Tanner A, Ankala A, Schneider T, et al. Genetic landscape and novel disease mechanisms from a large LGMD cohort of 4656 patients. Annals of clinical and translational neurology. 2018;5(12):1574-87.

195. Bushby K, Straub V. One gene, one or many diseases?: Simplifying dysferlinopathy. Neurology. 2010;75(4):298-9.

196. Krahn M, Béroud C, Labelle V, Nguyen K, Bernard R, Bassez G, et al. Analysis of the DYSF mutational spectrum in a large cohort of patients. Human mutation. 2009;30(2):E345-E75.

197. Harris E, Bladen CL, Mayhew A, James M, Bettinson K, Moore U, et al. The Clinical Outcome Study for dysferlinopathy An international multicenter study. Neurology Genetics. 2016;2(4):e89.

198. Moore U, Jacobs M, James MK, Mayhew AG, Fernandez-Torron R, Feng J, et al. Assessment of disease progression in dysferlinopathy: A 1-year cohort study. Neurology. 2019;92(5):e461-e74.

199. Fanin M, Angelini C. Progress and challenges in diagnosis of dysferlinopathy. Muscle & nerve. 2016;54(5):821-35.

200. Guglieri M, Magri F, D'angelo MG, Prelle A, Morandi L, Rodolico C, et al. Clinical, molecular, and protein correlations in a large sample of genetically diagnosed Italian limb girdle muscular dystrophy patients. Human mutation. 2008;29(2):258-66.

201. Ghaoui R, Cooper ST, Lek M, Jones K, Corbett A, Reddel SW, et al. Use of whole-exome sequencing for diagnosis of limb-girdle muscular dystrophy: outcomes and lessons learned. JAMA neurology. 2015;72(12):1424-32.

202. Chakravorty S, Hegde M. Gene and variant annotation for Mendelian disorders in the era of advanced sequencing technologies. Annual review of genomics and human genetics. 2017;18:229-56.

203. Chakravorty S, Hegde M. Inferring the effect of genomic variation in the new era of genomics. Human mutation. 2018;39(6):756-73.

204. Wan A, Place E, Pierce EA, Comander J. Characterizing variants of unknown significance in rhodopsin: a functional genomics approach. Human mutation. 2019;40(8):1127-44.

205. Gallardo E, de Luna N, Diaz-Manera J, Rojas-García R, Gonzalez-Quereda L, Flix B, et al. Comparison of dysferlin expression in human skeletal muscle with that in monocytes for the diagnosis of dysferlin myopathy. PLoS One. 2011;6(12):e29061.

206. Gallardo E, Ankala A, Núñez-Álvarez Y, Hegde M, Diaz-Manera J, Luna ND, et al. Genetic and epigenetic determinants of low dysferlin expression in monocytes. Human mutation. 2014;35(8):990-7.

207. Ankala A, Nallamilli BR, Rufibach LE, Hwang E, Hegde MR. Diagnostic overview of blood-based dysferlin protein assay for dysferlinopathies. Muscle & nerve. 2014;50(3):333-9.

208. Dastur RS, Gaitonde PS, Kachwala M, Nallamilli BR, Ankala A, Khadilkar SV, et al. Detection of dysferlin gene pathogenic variants in the indian population in patients predicted to have a dysferlinopathy using a blood-based monocyte assay and clinical algorithm: a model for accurate and cost-effective diagnosis. Annals of Indian Academy of Neurology. 2017;20(3):302.

209. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature communications. 2017;8(1):1-11.

210. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nature genetics. 2013;45(6):580-5.

211. Pramono ZAD, San Lai P, Tan CL, Takeda Si, Yee WC. Identification and characterization of a novel human dysferlin transcript: dysferlin_v1. Human genetics. 2006;120(3):410-9.

212. Pramono ZAD, Tan CL, Seah IAL, See JSL, Kam SY, San Lai P, et al. Identification and characterisation of human dysferlin transcript variants: implications for dysferlin mutational screening and isoforms. Human genetics. 2009;125(4):413-20.

213. Lee H, Huang AY, Wang L-k, Yoon AJ, Renteria G, Eskin A, et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genetics in Medicine. 2020;22(3):490-9.

214. Falkenberg KD, Braverman NE, Moser AB, Steinberg SJ, Klouwer FC, Schlüter A, et al. Allelic expression imbalance promoting a mutant PEX6 allele causes Zellweger spectrum disorder. The American Journal of Human Genetics. 2017;101(6):965-76.

215. Chakravorty S, Hegde M. Clinical utility of transcriptome sequencing: toward a better diagnosis for Mendelian disorders. Clinical Chemistry. 2018;64(6):882-4.

216. Romero IG, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC biology. 2014;12(1):1-13.

217. Shen Y, Li R, Tian F, Chen Z, Lu N, Bai Y, et al. Impact of RNA integrity and blood sample storage conditions on the gene expression analysis. OncoTargets and therapy. 2018;11:3573.

218. Brouard J-S, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. Journal of animal science and biotechnology. 2019;10(1):1-6.

219. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome biology. 2014;15(6):1-15.

220. Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nature biotechnology. 2014;32(9):888-95.

221. Love M, Anders S, Huber W. Differential analysis of count data–the DESeq2 package. Genome Biol. 2014;15(550):10-1186.

222. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Research. 2015;4.

223. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC bioinformatics. 2016;17(1):1-13.

224. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature methods. 2017;14(2):135-9.

225. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods. 2013;10(12):1185-91.

226. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

227. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research. 2018;46(D1):D794-D801.

228. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43(1):11.0.1-.0.33.

229. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics. 2013;76(1):7.20. 1-7.. 41.

230. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164-e.

231. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14(2):178-92.

232. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative splicing signatures in RNA-seq data: Percent spliced in (PSI). Current protocols in human genetics. 2015;87(1):11.6.1-.6.4.

233. Klepikova AV, Kasianov AS, Chesnokov MS, Lazarevich NL, Penin AA, Logacheva M. Effect of method of deduplication on estimation of differential gene expression using RNA-seq. PeerJ. 2017;5:e3091.

234. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Scientific reports. 2016;6(1):1-11.

235. Zhou W, Chen T, Zhao H, Eterovic AK, Meric-Bernstam F, Mills GB, et al. Bias from removing read duplication in ultra-deep sequencing experiments. Bioinformatics. 2014;30(8):1073-80.

236. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Human genetics. 2013;132(10):1077-130.

237. Krahn M, Borges A, Navarro C, Schuit R, Stojkovic T, Torrente Y, et al. Identification of different genomic deletions and one duplication in the dysferlin gene using multiplex ligation-dependent probe amplification and genomic quantitative PCR. Genetic testing and molecular biomarkers. 2009;13(4):439-42.

238. Cacciottolo M, Numitone G, Aurino S, Caserta IR, Fanin M, Politano L, et al. Muscular dystrophy with marked Dysferlin deficiency is consistently caused by primary dysferlin gene mutations. European journal of human genetics. 2011;19(9):974.

239. Aoki M, Liu J, Richard I, Bashir R, Britton S, Keers S, et al. Genomic organization of the dysferlin gene and novel mutations in Miyoshi myopathy. Neurology. 2001;57(2):271-8.

240. Dyle MC, Kolakada D, Cortazar MA, Jagannathan S. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated RNA decay. Wiley Interdisciplinary Reviews: RNA. 2020;11(1):e1560.

241. Swaika A, Boczek NJ, Sood N, Guthrie K, Klee EW, Agrawal A, et al. Whole Exome Sequencing Leading to the Diagnosis of Dysferlinopathy with a Novel Missense Mutation (c. 959G> C). Case reports in genetics. 2016;2016.

242. Rosales XQ, Gastier-Foster JM, Lewis S, Vinod M, Thrush DL, Astbury C, et al. Novel diagnostic features of dysferlinopathies. Muscle & nerve. 2010;42(1):14-21.

243. Liewluck T, Milone M. Characterization of isolated amyloid myopathy. European Journal of Neurology. 2017;24(12):1437-45.

244. Suwinski P, Ong C, Ling MH, Poh YM, Khan AM, Ong HS. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Frontiers in genetics. 2019;10:49.

245. Burgess DJ. The TOPMed genomic resource for human health. Nature Reviews Genetics. 2021;22(4):200-.

246. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. Jama. 2014;312(18):1880-7.

247. Shin HY, Jang H, Han JH, Park HJ, Lee JH, Kim SW, et al. Targeted next-generation sequencing for the genetic diagnosis of dysferlinopathy. Neuromuscular Disorders. 2015;25(6):502-10.

248. Marcon AR, Bieber M, Caulfield T. Representing a “revolution”: how the popular press has portrayed personalized medicine. Genetics in Medicine. 2018;20(9):950-6.

249. Williams JR, Lorenzo D, Salerno J, Yeh VM, Mitrani VB, Kripalani S. Current applications of precision medicine: a bibliometric analysis. Personalized medicine. 2019;16(4):351-9.

250. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. Journal of human genetics. 2021;66(1):11-23.

251. Tangye SG, Al-Herz W, Bousfiha A, Chatila T, Cunningham-Rundles C, Etzioni A, et al. Human inborn errors of immunity: 2019 update on the classification from the International Union of Immunological Societies Expert Committee. Journal of clinical immunology. 2020;40(1):24-64.

252. Barmada A, Ramaswamy A, Lucas CL. Maximizing insights from monogenic immune disorders. Current opinion in immunology. 2021;73:50-7.

253. Gruber C, Bogunovic D. Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. Human genetics. 2020;139(6):745-57.

254. Flinn AM, Gennery AR. Primary immune regulatory disorders: Undiagnosed needles in the haystack? Orphanet journal of rare diseases. 2022;17(1):1-9.

255. Kerur B, Benchimol EI, Fiedler K, Stahl M, Hyams J, Stephens M, et al. Natural history of very early onset inflammatory bowel disease in north america: a retrospective cohort study. Inflammatory bowel diseases. 2021;27(3):295-302.

256. Zheng HB, De La Morena MT, Suskind DL. The Growing Need to Understand Very Early Onset Inflammatory Bowel Disease (VEO-IBD). Frontiers in immunology. 2021;12:1858.

257. Kelsen JR, Dawany N, Moran CJ, Petersen B-S, Sarmady M, Sasson A, et al. Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease. Gastroenterology. 2015;149(6):1415-24.

258. Bhuvanagiri M, Schlitter AM, Hentze MW, Kulozik AE. NMD: RNA biology meets human genetic medicine. Biochemical Journal. 2010;430(3):365-77.

259. Alfares A, Aloraini T, Alissa A, Al Qudsi A, Alahmad A, Al Mutairi F, et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genetics in Medicine. 2018;20(11):1328-33.

260. Costain G, Jobling R, Walker S, Reuter MS, Snell M, Bowdin S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. European Journal of Human Genetics. 2018;26(5):740-4.

261. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of clinical exome sequencing data. New England Journal of Medicine. 2019;380(25):2478-80.

262. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genetics in Medicine. 2017;19(2):209-14.

263. Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genetics in Medicine. 2018;20(10):1216-23.

264. Al-Murshedi F, Meftah D, Scott P. Underdiagnoses resulting from variant misinterpretation: Time for systematic reanalysis of whole exome data? European Journal of Medical Genetics. 2019;62(1):39-43.

265. Basel-Salmon L, Orenstein N, Markus-Bustani K, Ruhrman-Shahar N, Kilim Y, Magal N, et al. Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. Genetics in Medicine. 2019;21(6):1443-51.

266. Boycott KM, Ardigo D. Addressing challenges in the diagnosis and treatment of rare genetic diseases. Nature Reviews Drug Discovery. 2018;17(3):151-2.

267. Zhang P, Itan Y. Biological network approaches and applications in rare disease studies. Genes. 2019;10(10):797.

268. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nature communications. 2016;7(1):1-6.

269. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome research. 2012;22(10):2008-17.

270. Rowlands CF, Taylor A, Rice G, Whiffin N, Hall HN, Newman WG, et al. MRSD: A quantitative approach for assessing suitability of RNA-seq in the investigation of mis-splicing in Mendelian disease. The American Journal of Human Genetics. 2022.

271. Waugh J, Perry CM. Anakinra. BioDrugs. 2005;19(3):189-202.

272. Moreland L, Bate G, Kirkpatrick P. Abatacept. Nature Reviews Drug Discovery. 2006;5(3).

273. Li J, Kim SG, Blenis J. Rapamycin: one drug, many effects. Cell metabolism. 2014;19(3):373-9.

274. Stenton SB, Partovi N, Ensom MH. Sirolimus. Clinical pharmacokinetics. 2005;44(8):769-86.

275. Dorjbal B, Stinson JR, Ma CA, Weinreich MA, Miraghazadeh B, Hartberger JM, et al. Hypomorphic caspase activation and recruitment domain 11 (CARD11) mutations associated with diverse immunologic phenotypes with or without atopic disease. Journal of Allergy and Clinical Immunology. 2019;143(4):1482-95.

276. Salah S, Hegazy R, Ammar R, Sheba H, AbdelRahman L. MEFV gene mutations and cardiac phenotype in children with familial Mediterranean fever: a cohort study. Pediatric Rheumatology. 2014;12(1):1-7.

277. Federici S, Calcagno G, Finetti M, Gallizzi R, Meini A, Vitale A, et al. Clinical impact of MEFV mutations in children with periodic fever in a prevalent western European Caucasian population. Annals of the rheumatic diseases. 2012;71(12):1961-5.

278. Moradian MM, Sarkisian T, Amaryan G, Hayrapetyan H, Yeghiazaryan A, Davidian N, et al. Patient management and the association of less common familial Mediterranean fever symptoms with other disorders. Genetics in Medicine. 2014;16(3):258-63.

279. Zhang X, Lei F, Wang XM, Deng KQ, Ji YX, Zhang Y, et al. NULP1 alleviates cardiac hypertrophy by suppressing NFAT3 transcriptional activity. Journal of the American Heart Association. 2020;9(16):e016419.

280. Schnappauf O, Heale L, Dissanayake D, Tsai WL, Gadina M, Leto TL, et al. Homozygous variant p. Arg90His in NCF1 is associated with early-onset Interferonopathy: a case report. Pediatric Rheumatology. 2021;19(1):1-8.

281. Wrona D, Siler U, Reichenbach J. Novel Diagnostic Tool for p47phox-Deficient Chronic Granulomatous Disease Patient and Carrier Detection. Molecular Therapy-Methods & Clinical Development. 2019;13:274-8.

282. Lacout C, Haddad E, Sabri S, Svinarchouk F, Garçon L, Capron C, et al. A defect in hematopoietic stem cell migration explains the nonrandom X-chromosome inactivation in carriers of Wiskott-Aldrich syndrome. Blood. 2003;102(4):1282-9.

283. Lutskiy MI, Sasahara Y, Kenney DM, Rosen FS, Remold-O'Donnell E. Wiskott-Aldrich syndrome in a female. Blood, The Journal of the American Society of Hematology. 2002;100(8):2763-8.

284. Denson LA, Jurickova I, Karns R, Shaw KA, Cutler DJ, Okou DT, et al. Clinical and genomic correlates of neutrophil reactive oxygen species production in pediatric patients with Crohn's disease. Gastroenterology. 2018;154(8):2097-110.

285. Ruan J, Schlüter D, Naumann M, Waisman A, Wang X. Ubiquitin-modifying enzymes as regulators of colitis. Trends in Molecular Medicine. 2022.

286. Hu B, Zhang D, Zhao K, Wang Y, Pei L, Fu Q, et al. Spotlight on USP4: Structure, function, and regulation. Frontiers in Cell and Developmental Biology. 2021:148.

287. de Valles-Ibáñez G, Esteve-Sole A, Piquer M, González-Navarro EA, Hernandez-Rodriguez J, Laayouni H, et al. Evaluating the genetics of common variable

immunodeficiency: monogenetic model and beyond. Frontiers in Immunology. 2018;9:636.

288. Ostrowski J, Paziewska A, Lazowska I, Ambrozkiewicz F, Goryca K, Kulecka M, et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. Scientific reports. 2016;6(1):1-10.

289. Cuthbert AP, Fisher SA, Mirza MM, King K, Hampe J, Croucher PJ, et al. The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology. 2002;122(4):867-74.

290. Lesage S, Zouali H, Cézard J-P, Colombel J-F, Belaiche J, Almer S, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. The American Journal of Human Genetics. 2002;70(4):845-57.

291. Cho JH, Abraham C. Inflammatory bowel disease genetics: Nod2. Annu Rev Med. 2007;58:401-16.

292. Horowitz JE, Warner N, Staples J, Crowley E, Gosalia N, Murchie R, et al. Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of Early Onset Crohn's Disease. Scientific reports. 2021;11(1):1-10.

293. Abolhassani H, Vitali M, Lougaris V, Giliani S, Parvaneh N, Parvaneh L, et al. Cohort of Iranian patients with congenital agammaglobulinemia: mutation analysis and novel gene defects. Expert review of clinical immunology. 2016;12(4):479-86.

294. Yazdani R, Abolhassani H, Kiaee F, Habibi S, Azizi G, Tavakol M, et al. Comparison of common monogenic defects in a large predominantly antibody deficiency cohort. The Journal of Allergy and Clinical Immunology: In Practice. 2019;7(3):864-78. e9.

295. Mao C, Zhou M, Uckun FM. Crystal structure of Bruton's tyrosine kinase domain suggests a novel pathway for activation and provides insights into the molecular basis of X-linked agammaglobulinemia. Journal of Biological Chemistry. 2001;276(44):41435-43.

296. Väliaho J, Faisal I, Ortutay C, Smith CE, Vihinen M. Characterization of all possible single-nucleotide change caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. Human Mutation. 2015;36(6):638-47.

297. Uhlig HH. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. Gut. 2013;62(12):1795-805.

298. Yin Q, Lin S-C, Lo Y-C, Damo SM, Wu H. Tumor Necrosis Factor Receptor-Associated Factors in Immune Receptor Signal Transduction. Handbook of Cell Signaling: Elsevier; 2010. p. 339-45.

299. de Diego RP, Sancho-Shimizu V, Lorenzo L, Puel A, Plancoulaine S, Picard C, et al. Human TRAF3 adaptor molecule deficiency leads to impaired Toll-like receptor 3 response and susceptibility to herpes simplex encephalitis. Immunity. 2010;33(3):400-11.

300. Annunziata CM, Davis RE, Demchenko Y, Bellamy W, Gabrea A, Zhan F, et al. Frequent engagement of the classical and alternative NF-κB pathways by diverse genetic abnormalities in multiple myeloma. Cancer cell. 2007;12(2):115-30.

301. Cheng G, Cleary AM, Ye Z-s, Hong DI, Lederman S, Baltimore D. Involvement of CRAF1, a relative of TRAF, in CD40 signaling. Science. 1995;267(5203):1494-8.

302. Saha SK, Pietras EM, He JQ, Kang JR, Liu SY, Oganesyan G, et al. Regulation of antiviral responses by a direct and specific interaction between TRAF3 and Cardif. The EMBO journal. 2006;25(14):3257-63.

303. Brodeur SR, Cheng G, Baltimore D, Thorley-Lawson DA. Localization of the major NF-κB-activating site and the sole TRAF3 binding site of LMP-1 defines two distinct signaling motifs. Journal of Biological Chemistry. 1997;272(32):19777-84.

304. Zhu S, Pan W, Shi P, Gao H, Zhao F, Song X, et al. Modulation of experimental autoimmune encephalomyelitis through TRAF3-mediated suppression of interleukin 17 receptor signaling. Journal of Experimental Medicine. 2010;207(12):2647-62.

305. Lee H, Choi J-K, Li M, Kaye K, Kieff E, Jung JU. Role of cellular tumor necrosis factor receptor-associated factors in NF-κB activation and lymphocyte transformation by herpesvirus saimiri STP. Journal of virology. 1999;73(5):3913-9.

306. Haddow JB, Musbahi O, MacDonald TT, Knowles CH. Comparison of cytokine and phosphoprotein profiles in idiopathic and Crohn's disease-related perianal fistula. World journal of gastrointestinal pathophysiology. 2019;10(4):42.

307. Hucthagowder V, Meyer R, Mullins C, Nagarajan R, DiPersio JF, Vij R, et al. Resequencing analysis of the candidate tyrosine kinase and RAS pathway gene families in multiple myeloma. Cancer genetics. 2012;205(9):474-8.

308. Rothlin CV, Leighton JA, Ghosh S. Tyro3, Axl, and Mertk receptor signaling in inflammatory bowel disease and colitis-associated cancer. Inflammatory bowel diseases. 2014;20(8):1472-80.

309. Aggarwal V, Banday AZ, Jindal AK, Das J, Rawat A. Recent advances in elucidating the genetics of common variable immunodeficiency. Genes & diseases. 2020;7(1):26-37.

310. Lucas CL, Chandra A, Nejentsev S, Condliffe AM, Okkenhaug K. PI3Kδ and primary immunodeficiencies. Nature Reviews Immunology. 2016;16(11):702-14.

311. Swan DJ, Aschenbrenner D, Lamb CA, Chakraborty K, Clark J, Pandey S, et al. Immunodeficiency, autoimmune thrombocytopenia and enterocolitis caused by autosomal recessive deficiency of PIK3CD-encoded phosphoinositide 3-kinase δ. Haematologica. 2019;104(10):e483.

312. Li Z, Rotival M, Patin E, Michel F, Pellegrini S. Two common disease-associated TYK2 variants impact exon splicing and TYK2 dosage. PLoS One. 2020;15(1):e0225289.

313. Góth L, Rass P, Madarasi I. A novel catalase mutation detected by polymerase chain reaction-single strand conformation polymorphism, nucleotide sequencing, and Western blot analyses is responsible for the type C of Hungarian acatalasemia. Electrophoresis. 2001;22(1):49-51.

314. Gifford CA, Ranade SS, Samarakoon R, Salunga HT, De Soysa TY, Huang Y, et al. Oligogenic inheritance of a human heart disease involving a genetic modifier. Science. 2019;364(6443):865-70.

315. Rahme E, Joseph L. Estimating the prevalence of a rare disease: adjusted maximum likelihood. Journal of the Royal Statistical Society: Series D (The Statistician). 1998;47(1):149-58.

316. Auvin S, Irwin J, Abi-Aad P, Battersby A. The problem of rarity: estimation of prevalence in rare disease. Value in Health. 2018;21(5):501-7.

317. Goth L, Nagy T. Inherited catalase deficiency: is it benign or a factor in various age related disorders? Mutation Research/Reviews in Mutation Research. 2013;753(2):147-54.

318. Hamilton HB, Neel JV, Kobara TY, Ozaki K. The frequency in Japan of carriers of the rare “recessive” gene causing acatalasemia. The Journal of Clinical Investigation. 1961;40(12):2199-208.