Integrating Transcriptome Sequencing From Mendelian Disease Patients and Healthy Controls to Improve Genetic Variant Interpretation

Integrating transcriptome sequencing from Mendelian disease patients and healthy controls to improve genetic variant interpretation

A dissertation presented

by

Besse Beryl Cummings

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

April 2019

Dissertation Advisor: Daniel G. Macarthur Besse Beryl Cummings

Integrating transcriptome sequencing from Mendelian disease patients and healthy controls to improve genetic variant interpretation

Abstract

Whole exome and whole genome sequencing have become increasingly routine approaches in understanding the genetic basis of Mendelian diseases. Despite their success, the current diagnostic rate for genomic analyses across a variety of rare diseases is approximately 25-50%, and a genetic diagnosis does not yield a full understanding of disease pathology. A key challenge of genome-based diagnostics is that the capacity DNA sequencing technologies to discover genetic variants substantially exceeds our ability to interpret their functional and clinical impact.

One approach to improve the interpretation of genetic variation is to integrate functional genomic information such as transcriptome sequencing, which provides direct insight into transcriptional perturbations caused by genetic changes. Such approaches have already proven useful for elucidating mechanisms of cancer and common disease but have yet to be systematically applied to rare disease. Here, we present complementary approaches to integrate transcriptome sequencing into our understanding of the genetic etiology of Mendelian disease. We first present our work establishing the utility of transcriptome sequencing as a complementary diagnostic tool in Mendelian disease diagnosis. We then focus on developing and validating a transcript expression aware annotation metric which allows for the integration of publicly available population transcriptome datasets into clinical variant interpretation.

Acknowledgements

While I'm enormously proud of the work laid out in this thesis, I feel the acknowledgements chapter is really my time to shine. Those that have been along the ride with me will know I don't shy away from overly emotional statements. So, I recommend the weaker hearted skip ahead to avoid overly sentimental prose.

Graduate school has been the best academic experience of my life. I've learned so much about human genetics, computer science, data science, statistics, giving talks, writing manuscripts, being a productive lab member and colleague, and so much more. The scientific perspective of the Analytical and Translational Genetics Unit, with an emphasis on rigorous quality control, borderline brutal critical feedback that can be intimidating but is done out of love and caring for the scientist and the science, a focus on using statistics the right way, and a passion for integrating human genetics into clinical decision making, has now become my scientific perspective, and I hope to carry it with me and disseminate it to the best of my ability.

I sincerely believe that the ATGU's contribution to understanding the genetics of Mendelian and complex disease, and more broadly to human genetics, is massive. I feel so lucky to have been a part of this community, and to work alongside brilliant trainees and staff scientists that have always shared their expertise and time. It's been such a pleasure to learn from you all.

Of course, the environment would not be what it is without the three leaders of the ATGU: Mark Daly, Ben Neale and Daniel MacArthur. Mark, thank

you for your support during difficult times, your excitement about new data and science that is truly contagious, and your general presence and leadership. I've never realized someone could be so exceptionally smart, and also be so exceptionally kind.

I began to realize that I was learning as I started to understand Ben Neale's points during lab meeting. He is someone whose comments and presence has helped me learn a lot of science that has been outside the scope of my PhD.

Thank you so much to my supervisor and mentor Daniel MacArthur. I am privileged to be your first PhD student, and can't begin to list everything I've learned from you. If I were to try, I would discuss a pragmatic attitude towards science and a scientific career, an exquisite focus on detail, an emphasis on critical feedback, but above all a kind of brilliance that can't be taught but which I've tried so hard to learn. I will never forget the amazing, and slightly traumatic, experience of preparing for my first-ever scientific talk at Biology of Genomes in 2016. I believe a lab member told me that she could have given the talk by heart, as she saw me practice it so many times. That two-week period taught me everything I know about scientific presentations, and I appreciate it so much. You are a great mentor, and if I could go back I would consider myself lucky to be able to join your lab and go through this experience all over again. I wouldn't change much. Thank you for your effort, your guidance, and your caring.

I've been lucky enough to never have suffered from imposter syndrome, and for that I can only thank Hunt Willard. His belief in me from day one of

freshman year of college has given me the confidence and fuel to face the difficulties of graduate school. He will forever be a mentor, and I appreciate him so much.

There are so many people that I genuinely consider to be like family. Konrad Karczewski taught me most of the technical skills I learned in graduate school. His patience to questions I could have answered myself, his excitement towards innovative ideas, his at times misunderstood attitude that puts the science above all has made me (what I consider and hope to be) the thick-skinned, no-nonsense, trust-the-science female scientist I am today. He's one of the few I don't have the words to express my gratitude, but I hope will one day come to me in this friendship I suspect will last a lifetime.

Kaitlin Samocha has been a fierce role model, someone to look up to and admire, that never pretends things are easy, that always gives honest advice, and never shies away from discussing the difficulties. You are a beacon for women in science and I'm lucky to consider you a friend and mentor. Jack Kosmicki who wore the difficulties of graduate school on his sleeve, and was always willing to commiserate. I consider you a brother and thank you for your help along the way. Sherif Gerges who is bound to do amazing things, and with whom every conversation I have left me feeling a little more positive and hopeful towards the world than before. Jessica Alfoldi, always there with a funny tweet or meme, always telling me it was going to be ok, always the voice of reason. Thank you. Nilesh K. Raval, who supported me when it mattered the most. I sincerely don't think I would have had the success without your compassion,

companionship and encouragement. Emma Fridel, who has been through the thick and thin with me since our days eating french fries at the Loop. I hope you know how much you mean to me. Kumar Veerapan, who, possibly unknowingly, provided me with the support and distractions when I needed it the most. You light me up with your sense of positivity and love, thank you!

Thank you to my family who provided never-ending support and love. Dünya bir yana, siz bir yana. Sizi çok seviyourm.

Finally, again, thank you to the Broad Institute and the Analytical and Translational Unit at Massachusetts Hospital. Everyone who has helped along the way, Juha Karjailanen, Elise Valkanas, Anne O'Donnell-Luria, Fengmei Zhao, Monkol Lek, Taru Tukiainen, Laurent Francioli, and of course Jill Doucette, Carla Hammond and Beth Raynard and many unnamed others.

Thank you all so much!

Table of Contents

Chapter 1

Introduction

Genetic variation and disease

A major goal of human genetics research is the ability to interpret variation in the genome in a way that is meaningful to human health and predictive of disease. In other words, a central aim is the accurate interpretation of the functional and clinical impact of variation we observe in a given genome.

It has been stated that a majority of human diseases, with the exception of cases of trauma, have a genetic component [1]. Twin and pedigree studies have clearly shown that common diseases such as type II diabetes, psychiatric disorders and cancer have a strong heritable component [1,2]. However, understanding the existence of a genetic component of a disorder does not provide information about the specific genes that affect disease pathogenesis.

A clear understanding of the link between specific genetic variants and disease would theoretically not only allow prediction of individual risk for a given disorder, but also offer insight into disease etiology, which can be harnessed to develop targeted therapeutics [1,3].

Early efforts to identify genes associated with human disease adopted the use of linkage analysis, developed in fruit flies [4]. This involves identifying families in which a phenotype segregates, and tracking polymorphic markers through generations to identify correlation between a phenotype and a variant site [4,5]. The polymorphisms used have changed over the years, from restriction fragment length polymorphisms [6], to simple sequence repeats [7,8] and single nucleotide polymorphisms (SNPs) [9], each offering more resolution than the previous iteration [5]. Through linkage analysis, a candidate genomic region is identified, and further scrutinized to pinpoint causal

mutation(s) and genes underlying the phenotype. The method was mostly confined to identifying highly penetrant large-effect genetic variants in rare disorders with notable successes of this approach, entitled “positional cloning”, including identification of genes underlying hemochromatosis [9], lactose intolerance , Duchenne muscular dystrophy [10], cystic fibrosis [11], hereditary disposition to cancer [12], and other Mendelian disorders [5].

Linkage analyses in common diseases however, were unable to identify causative genes, which pointed to a possible polygenic model [4]. Instead, as extensive maps of human polymorphisms became available, identifying genetic loci in common diseases has relied on using common variation and the linkage disequilibrium structure in the genome, whereby groups of variants are inherited together on haplotypes, to associate tagging SNPs with disease status in cohorts of unrelated individuals [13]. Such genome-wide association studies (GWAS) are now performed routinely, using cohort sizes of hundreds of thousands and have tied thousands of genetic loci to numerous human phenotypes, primarily identifying loci with individually small effect sizes [14–19].

Genetics of Mendelian disease

Monogenic disorders that follow Mendelian inheritance patterns are broadly characterized as Mendelian disease [5,20]. While these disorders are individually rare, at incidences ranging from a handful of cases worldwide to one in a few thousands, they are collectively common, accounting for approximately 10% of pediatric hospital admissions and up to 20% of infant deaths [21–23]. In the United States, more than 25 million people are affected by Mendelian disorders and in addition to the burden of

suffering faced by rare disease families, each patient has been estimated to cost the health care system approximately five million dollars during their lifetime [24–26].

Diagnosing a child with a rare disease by conventional diagnostic testing and phenotypic features can be challenging [24]. As a result, rare disease families often endure long periods of uncertainty and emotional turmoil involving multiple hospital visits and changing diagnoses [27,28] . This distressing period defined as the time between initial health concerns for a child, and when a diagnosis is reached, has been termed the “diagnostic odyssey” and can take many years [29]: One European survey of eight rare diseases found that a quarter of families waited between 5 and 30 years [24,30] to receive a genetic diagnosis.

Achieving a genetic diagnosis for rare disease families is critical to not only tailor the care needs of the patient, but also to alleviate emotional distress caused by an unnamed disease. Such care needs can range from better understanding of disease management and prognosis for a child’s condition, to procuring physical or occupational therapy in the school system, and forming communities with other rare disease families [27,31,32]. It has been reported that a genetic diagnosis can also alleviate a family’s sense of guilt in the community. Carmichael et al. report a case of a school gym teacher pushing their child with a rare disorder to ‘get over it’ and ‘rise to the occasion’ when struggling to keep up in gym class. After a genetic diagnosis was achieved, accommodations were made for activities that the patient was unable to perform [27]. Another mother reports finally absolving herself from perceived faults she may have committed during pregnancy and early childcare that she believed resulted in her child’s condition before receiving the genetic diagnosis [27].

In addition to the many, unquantifiable psychosocial benefits of a genetic diagnosis, knowing the mutation and inheritance pattern of a disorder in a family can help inform reproductive counseling and family planning [23,33,34]. From a biological perspective, linking genes and their allelic series to phenotypes aids our understanding of biological pathways underlying disease and health [20,35]. This can inform therapeutic development in rare disease, which in recent years has yielded success stories such as those for cystic fibrosis [36–39] and spinal muscular atrophy [40,41]. As therapeutic development continues, a genetic diagnosis can also be important for entry into clinical trials [42].

Scientists have been mapping Mendelian disease genes and identifying highly penetrant mutations for the past 40 years [4,5,24]. The early successes of linkage mapping and positional cloning led to the discovery of *CYBB* underlying chronis granulomatous disease in 1986 [24] and *CFTR* underlying cystic fibrosis in 1989 [5].The following decade saw the discovery of 42 additional Mendelian disease genes using these methods [24]. The arrival of the human genome sequence in the early 2000s and development in DNA sequencing technology greatly expanded the known Mendelian disease gene catalog [43].

Despite the importance of genetic diagnosis for rare disease families and the tremendous headway made in mapping Mendelian diseases and mutations, at present over half of patients with a suspected genetic disorder do not receive a genetic diagnosis [23,24] and the majority of genes underlying Mendelian diseases remain unknown. For example, while more than 80% of genes display detectable phenotypes upon homozygous inactivation in mice [44], suggesting analogously that many human

genes are likely to be implicated in disease phenotypes, less than 20% of human genes have been linked to any phenotype [43].

The role of next-generation sequencing in Mendelian disease

Although traditional mapping approaches have led to great insight into many genes underlying Mendelian disorders, they can be laborious and miss important classes of genetic variation [1,3,35]. In contrast, the introduction of next-generation sequencing technologies has rapidly accelerated the pace of gene discovery in Mendelian disease [24].

The advent of exome sequencing has greatly enhanced our capacity to identify variants that explain many Mendelian diseases in both known and novel disease genes. Chong et al. have reported that the pace of gene discovery has increased from ~166 to ~236 per year from the period of 2005-2009 to 2010-2014 [24]. This is driven by a shift toward increasing use of whole exome and whole genome sequencing (WES and WGS, respectively). In fact, since 2014, WES and WGS have resulted in almost three times as many discoveries as conventional methods [24].

Whole exome sequencing (WES) utilizes capture technologies to enrich for protein-coding regions that make up 1-2% of the human genome and is a cost-effective alternative to sequencing the complete genome [45,46]. Whole genome sequencing (WGS) is an alternative option as it uncovers virtually all the genetic variation in a person's genome. Studies have shown that WGS outperforms WES at covering protein-coding regions, since WES has historically been limited by its capture efficiency [47,48]. WGS also offers improved identification of structural variants missed by WES [49,50].

Despite its benefits, WGS is substantially (~3x) more expensive than WES, and our ability to interpret the pathogenicity of variants is currently largely confined to protein coding sections of the genome. Results from studies aiming to use WGS in Mendelian disease diagnosis have so far concluded that variants uncovered by WGS are largely identifiable and interpretable by genotyping arrays (for CNVs) and WES[24,51]. However, these studies have focused on heterogeneous Mendelian diseases with the goal of identifying novel genes.

There remains considerable debate on the relative value of high-quality WES versus WGS for genetic diagnosis and gene discovery in the field [52–55]. The view I've reached in preparation for this thesis is that for genetic diagnosis of an individual patient, the comprehensive nature of WGS will allow for fewer false-negatives and should be preferred, whereas for gene discovery for cohorts of rare disease patients, WES offers a more cost-effective approach that allows access to the interpretable regions of the genome.

The genetic diagnosis gap: What are we missing?

While exome sequencing is a current mainstay in Mendelian disease diagnosis, the success rate of detecting the causal variant with WES is far from complete, ranging from 15-50% [56–59] . The molecular diagnostic rate varies widely based on several factors including the age of onset of the disease, inheritance mode and genetic heterogeneity [24,60]. For example an approximate rate of 30-40% has been reported for neuromuscular disorders [56,61] and familial dilated cardiomyopathy [62] whereas the diagnostic rate for pediatric diseases with more complex presentations presented to the NIH Undiagnosed

Disease Program was ~11% [24,63]. Studies have also shown that of cases that achieved a molecular diagnosis, 53% were autosomal dominant, of which 87% were diagnosed with a *de novo* mutation, followed by 34% for autosomal recessive cases, 12% for X-linked inheritance and 0.2% for mitochondrial inheritance [58]**.** In addition, the average molecular diagnosis rates across a variety of adult rare disease patients was shown to be lower at 17.5%, than that for a primarily pediatric population [60].

The tremendous progress made in rare disease genetics and the disruptive impact of DNA sequencing begs the question of what current diagnostic technologies are missing in genetic diagnoses [64]. Akin to the missing heritability question in common diseases, explained below, several possible explanations have emerged. Firstly, as discussed, it is likely that many Mendelian disease genes remain to be discovered [24]. Individual patients arriving at various centers across the world may present as unique cases, for whom establishing a novel disease gene mutation as causative is extremely difficult, given that every individual harbors many private benign variants. This is known as the “n of 1” problem. For some cases, identifying a single additional unrelated case with a putatively pathogenic variant in the same gene and overlapping disease presentation can provide sufficient evidence to implicate the gene and provide a diagnosis for the patient [65]. Therefore, discovery of all Mendelian phenotype - gene relationships requires infrastructures of genetic variant interpretation and data sharing between centers. Many national efforts have been established towards the goal of identifying novel gene - phenotype relationships in Mendelian disease including Finding of Rare Disease Genes (FORGE) Canada [66], the Wellcome Trust Deciphering Developmental Delay Study [67], and the Centers for Mendelian Genomics[68]. It has been

shown that re-analysis of previously undiagnosed cases using novel disease gene information, has a positive impact on diagnostic rates, underlining the importance of novel gene discovery to improve genetic diagnosis [69–71].

In addition to the impact of novel disease gene discovery on genetic diagnosis rate, cases involving non-Mendelian inheritance patterns are also likely missed by current analytical approaches. For example, facioscapulohumeral muscular dystrophy (FSHD) is characterized by the activation of *DUX4* which is found in the D4Z4 microsatellite repeats occurring on the subtelomeric arm of chromosome 4q35 [72–74]. In individuals with FSHD type 1, contraction of the array below 10 repeats results in inefficient epigenetic silencing of the region, resulting in *DUX4* expression. In contrast, individuals with FSHD type 2 harbor inactivating mutations in *SMCHD1*, an epigenetic silencer, which also results in *DUX4* expression. Interestingly, neither the repeat contraction nor the *SMCHD1* loss-of-function is sufficient to cause the disease. Both subtypes also require the mutations to arise on the permissive 4qA haplotype which contains a polyadenylation signal, to produce functional *DUX4* mRNA [72]. Thus, both genetic subtypes of FSHD have a digenic inheritance model. Other cases Mendelian phenotypes such as retinitis pigmentosa, and Bardet-Biedl syndrome have also shown to result from digenic inheritance [74]. In some cases, oligogenic inheritance, in which several rare variants across more than two genes result in the disease phenotype, when one of the variants itself is not causative have been suggested [75,76]. In addition, the role of polygenic risk conferred by common variants has been explored in a cohort of approximately 7,000 Mendelian disease patients with developmental delay and intellectual disabilities [77]. Similarly large cohort sizes of rare disease patients will be

required for an unbiased understanding of the role of multigenic inheritance patterns in Mendelian diseases.

Incomplete penetrance is one possible mechanism which can result in false negatives for rare disease diagnosis. Penetrance is defined as the proportion of people with a causative genotype who display the clinical characteristic associated with the genotype and is likely due to a combination of genetic and environmental factors [78]. For rare disease diagnosis, a dominant mutation may be overlooked if it has been transmitted from a seemingly unaffected parent, resulting in a decrease in the diagnosis rate [79]. Given that our understanding of pathogenic Mendelian mutations is based on evaluating the genome of rare disease patients, it has been suggested that the penetrance of these mutations has likely been overestimated [80]. Evaluation of putatively disease-causing mutations in individuals that do not carry rare disease will be informative to improve estimates of genetic variation [61].

While much work remains to be done in mapping the genetic architecture of rare diseases such as continued discovery of novel disease genes, identifying genes exhibiting non-Mendelian inheritance patterns and elucidating the role of incomplete penetrance, a seemingly simple but critical component to improving genetic diagnosis is improving rare variant interpretation. Currently, while both WES and WGS can miss important classes of variation, a larger issue is that our ability to identify genetic variation far outstrips our capacity to interpret what we identify [81,82].

Many classes of genetic variation remain difficult to interpret. For example, although most missense variants are expected to be benign [4], countless gain- and loss-of-function missense variants have been implicated in rare disease [83]. In another

example, while it is well understood that mutations that disrupt the canonical two base pair GT/AG splice motifs have disruptive effects on transcription and are regarded as deleterious, variants found outside the four base pair canonical splice junction have often been ignored or characterized as variants of unknown significance (VUS)[84] similar to noncoding variants identified via WGS[61]. A central aim of this thesis is to improve rare variant interpretation in Mendelian disease diagnosis, which is extensively discussed in chapters 2 and 3.

The role of next generation sequencing in common disease

The approach to mapping genes to common diseases has historically relied on the common disease - common variant hypothesis, which proposes variants with population allele frequencies over 1% will contribute to disease susceptibility [4]. This hypothesis, which led to the so-called "GWAS era", was grounded in population genetic models of the interplay between recent human expansion, which results in most variation in the human genome being common, with the late-onset nature of common diseases allowing for mildly deleterious alleles to rise in population frequency [4].

The early days of the GWAS approach, applied to many complex disorders, yielded tens of loci associated with common human phenotypes. However, a perplexing observation was the relatively marginal proportion of phenotypic variation explained via the discovered loci [85]. In other words, most common variants conferred small amounts of relative risk (at about 1.1-1.5-fold) and explained a tiny fraction of the estimated heritability [85]. This observation was termed the "missing heritability" problem and it was suggested that increasing sample sizes, evaluating unappreciated variant classes (such

as structural variants), modeling genetic epistasis and other methodological improvements would be required to address the full spectrum of genetic effects on common disease.

As genotyping approaches have improved with more resolution, and DNA sequencing in large numbers has become attainable, one avenue of research for common disorders has been the exploration of the role of rare variants in common disease [85–88] with the underlying intuition being that due to selection, rare variation will have higher impact on disease risk, towards the spectrum of Mendelian variants, and that this may help account for the unexplained heritability in common disease [89].

Two distinctions exist between common and rare variant association tests. Firstly, due to the numerous nature of rare variation, sequencing is preferred to catalog rare variants. Secondly, because individual variants are rare, direct association tests between a rare variant and a phenotype are implausible and variants must be grouped into categories for burden testing [45,87,88]. Studies have aggregated variants based on genes, gene sets and pathways [88]. Recent years have seen the amassing of large exome sequence cohorts of common diseases, which will continue to allow linking genes or groups of genes to common disease [90–93]. In this thesis, we hypothesize that interpretation of individual rare variants that are discovered will become important to inform analyses and understand disease etiology, and this is discussed in chapter 3.

Tools available for rare variant interpretation

Establishing a causal relationship between a gene or variant can have different meanings for rare and common diseases. For the purposes of this thesis, we focus on

rare variants with high effect sizes, that can be implicated in both rare as well as common diseases, the latter of which is most likely to be identified through rare variant association tests including gene burden testing. While the methods discussed in the thesis can certainly be applied to interpretation of high frequency variants, we will not discuss the interpretation of low-effect size variants commonly identified in GWAS.

Guidelines to establish causality for rare variants

Guidelines for implicating genes and variants as causative for disease have been developed and revisited by the human genetics community over the years [82,94]. In 2013, the American College of Medical genetics, along with the Association for Molecular Pathology and the College of American Pathologists joined forces at a workshop to revise and reinstate standards and guidelines for the interpretation of sequence variants[94]. The resulting framework aims to place a variant on the spectrum of pathogenic, likely pathogenic, uncertain significance, likely benign and benign. Lines of evidence for interpretation are considered from supporting, moderately supporting, strongly supporting and very strongly supporting. For example, identification of a null variant, defined as nonsense, frameshift, canonical splice site-disrupting, single or multi-exon deletions, in a gene with a known loss-of-function mechanism of disease, is considered very strong evidence of pathogenicity. In contrast, in-frame insertions or deletions in repetitive regions are considered supporting evidence of benign impact. In addition to assessment of variant classes, lack of segregation evidence in cases where paternity and maternity are confirmed is considered strong evidence against pathogenicity [94]. Such guidelines allow for a systematic assessment of variants that is

consistent across laboratories and analysts, and it is recommended that all evidence, not only those that support the final verdict are presented[82,94,95].

When attempting to follow proposed guidelines, implicating a pathogenic variant in a Mendelian disease patient theoretically still requires the assessment of every rare variant in the patient, of which there are many thousands[96]. This problem therefore requires tools for variant prioritization. Below, we discuss two important tools for prioritizing many variants: *in silico* prediction algorithms and population allele frequencies, the latter of which has been made increasingly accessible over the years through the publication of large databases of human genetic variation.

In silico prediction tools

In silico prediction tools aim to predict the functional impact of sequence variation using a variety of information such as conservation, location of a variant in the gene, or the biochemical properties of possible amino acid changes [94,96]. A variety of tools exists, including comprehensive tools that can be applied to all genetic variants[97,98], or those that are aimed specifically at the prediction of the effect of coding variation such as missense[99,100] or splice-affecting variants[101,102] in addition to tools aimed at noncoding variants such as those in untranslated regions[103] (UTR) or predicted transcription factor-binding sites [104].

In silico prediction tools are useful guiding tools for variant interpretation. Such tools have consistently shown to globally differentiate between benign variants and those identified in databases of pathogenic variation, such as ClinVar [81,98] and are highlighted in the 2013 ACMG guidelines. For example, the prediction of no impact on

gene product via *in silico* algorithms is considered supporting evidence of benign impact [94]. However, for a given variant, the prediction of damaging does not equate to clinically pathogenic [96]. For example, using missense variants of unknown significance in *BRCA1* it has been shown that all tools suffer from poor specify and sensitivity making them unsuitable for pathogenicity prediction [105]. In another study, it was shown that prediction tools can have as high as a 30% false positive rate for calling benign variation pathogenic [106]. The damaging prediction of noncoding variants are known to be even less accurate than their coding counterparts due to insufficient understanding of the regulatory machinery encoded in DNA [96,107]. Therefore, it is generally recommended against using a single prediction tool as the sole source evidence to make clinical assertions [94,96] and predictions from different *in silico* tools are often combined and used as a single piece of evidence in variant interpretation.

Large-scale databases of genetic variation

One of the most useful pieces of information analysts have about a variant is its allele frequency in the general population. Indeed, a variant allele frequency of over 5% is considered stand-alone support for the interpretation of variants as benign, and an allele frequency greater than expected for the disorder is considered strong support [94].

Until 2014, approximately 3,000 genome sequences from diverse populations in 1000 Genomes project and approximately 6,500 exomes from NHLBI Exome sequence project were used to available the population variant frequency [108]. In 2014, the Exome Aggregation Consortium (ExAC) made summary variant-level data from approximately 60,000 exome sequences publicly available [109]. The dataset was released, in part, to

aid variant interpretation for rare diseases and accordingly passed through stringent quality control to remove related individuals and samples known to have a rare disease [109]. This dataset is referred to as a set of “ostensibly healthy” individuals without rare disease [110] and can be thought of as a population cohort.

The use of ExAC as a reference dataset for clinical variant interpretation to filter variants too common to plausibly cause a highly penetrant severe disease, was shown to be sevenfold more powerful than ESP [109]. Using a 1% allele frequency cutoff in both the entire dataset and in South Asian or Latino individuals, two populations that were underrepresented in reference databases, the authors reassigned 126 previously pathogenic variants to benign or likely benign. In one case, a variant associated with a severe recessive Mendelian liver disease was found to be present in homozygous state in 4 ExAC individuals. The variant was identified in a North American Indian pediatric cohort through linkage mapping, subsequent Sanger sequencing and functional analyses[111]. Phenotypic follow-up of the 4 ExAC individuals showed no signs of a severe Mendelian liver disorder, and the variant was reclassified as benign, highlighting the relative importance of population frequency evidence in comparison to linkage mapping and functional analyses [109].

Five years after the publication of the ExAC manuscript, its successor, the Genome Aggregation Consortium released summary data from over 140,000 individuals, this time a combination of approximately 125,000 WES and 15,000 WGS samples[112]. Both the ExAC and gnomAD datasets are presented in an intuitive interface to enable clinical geneticists and biologist to explore variants and genes of interest [113]. Other large databases of human genetic variation have also been released including

WES data from 50,726 individuals in the DiscovEHR study [114] and 53,831 genomes from the NHLBI TopMed program. The UK Biobank, a national prospective cohort with approximately 500,000 participants who have contributed physical and health data, aims to provide exome sequencing on all of its participants, an effort that is currently 10% complete [115,116].

It is difficult to overstate the impact of large-scale databases on Mendelian variant interpretation. Countless genetic diagnosis and gene discovery papers have been published using the ExAC and gnomAD resources [96] and the ExAC resource has been cited 3,965 times to date since its publication in 2016. In addition to the initial reclassification in the ExAC manuscript of high frequency variants previously annotated as pathogenic, disease-area specific studies have used the resource to reclassify previously reported variants as benign [117–119]. Multiple studies have reported that proposed pathogenic variation with strong prior support, such as segregation data, are more likely to be rare or absent in reference databases, whereas those with a weaker evidence basis are more likely to have many carriers [80,117].

Despite the tremendous positive impact of large genetic reference databases, they can also pose new variant interpretation challenges. In some cases, well-established pathogenic variation can be observed in healthy individuals in ExAC, and requires careful follow-up analyses. In one published case, a *de novo* loss-of-function variant in *ASXL1* was observed in a 6 year old child with presumed Bohring-Opitz syndrome. Germline inactivating mutations in *ASXL1* have previously been established to cause Bohring-Opitz syndrome, and this specific variant had previously been reported as *de novo* in another patient with a closely resembling clinical presentation[120].

However, a perplexing observation was the presence of this exact same variant in ExAC in 7 samples. Furthermore, ExAC individuals were also shown to carry other *ASXL1* loss-of-function variants previously reported to cause severe disease, with the most common two variants present in 132 and 118 individuals. Manual evaluation of the *de novo* mutation observed in the initial in ExAC samples revealed considerable allele imbalance, in which the two haplotypes differ from the expected 50%, and which is in line with the variants being somatic mosaic in ExAC individuals [120]. In other words, presence of this disease-causing variant in ExAC individuals was shown to not be germline and therefore not presumed to be disease-causing. Further evaluation of the two other disease-causing variants, each seen in over 100 ExAC samples were shown to be in line with variant calling errors, namely two frameshift variants occurring at the end of a homopolymer run, likely representing PCR artifacts [120]. This case study of a single gene cautions against blind filtering of variants based only on the ExAC frequency and highlight the importance of careful curation of variants in population databases that we do not expect to occur.

Large scale databases have begun to allow unbiased estimates of penetrance [80]. Specifically, the gnomAD database, a cohort depleted for rare disease, can help answer the penetrance question for a variety of diseases. However, as exemplified in the *ASXL1* case, variants observed in population databases should pass through careful quality control and ideally manual curation for such estimates to be as precise as possible. Chapter 3 of this thesis discusses one underappreciated error mode, and presents a tool to improve variant interpretation in such cases, which we expect will improve estimates of penetrance.

Transcriptome sequencing: a functional genomics tool

The terms exome and genome refer to the full set of all exonic regions, and the complete genome, respectively. Similarly, the transcriptome is defined as the type and quantity of all transcripts in a cell [121]. Transcriptome sequencing, also known as RNA sequencing, or RNA-seq, refers to the experimental procedure of generating complementary DNA from RNA molecules and performing next-generation sequencing [122,123]. RNA-seq is a high throughput method with high technical reproducibility[122] and offers base-level resolution of the RNA molecules present in a cell or tissue, and information about gene expression patterns, splicing, allele imbalance, and the variants present on RNA molecules, which can be germline, somatic or due to RNA editing [122]. While several RNA-seq approaches are available, including those that aim to analyze long noncoding RNA (lnRNA)[124], micoRNA (miRNA)[125], short interfering RNA (siRNAs)[126], small nucleolar RNA (snoRNA) [127], circular RNA (circRNA)[128] or others, this thesis primarily focuses on the sequencing of human polyadenylated mRNAs.

Quantifying gene expression with RNA-seq

Profiling the gene expression levels in a cell type, tissue or organism is central to understanding new biological processes. Sequencing RNA molecules in a given sample, allows the use of the number of sequencing reads identified per gene to be utilized as a proxy for the level of expression of the gene [129].

The number of sequencing per gene is subject to normalization to account for gene length, and sequencing depth in a sample. The reads per kilobase per million mapped reads (RPKM), the fragments per kilobase per million mapped reads (FPKM)

and the transcript per million mapped reads (TPM) are the three most commonly used gene expression units [130]. RPKM involves dividing the number of reads per gene with the gene length and total sequencing depth in the sample, using a million as a scalar [130]. FPKM uses the same transformation, but the numerator is the number of paired end reads instead of single reads. The difference between TPM and RPKM/FPKM is the order of the division, in which TPM calculates sample scaling factors before dividing read counts and gene lengths whereas RPKM/FPKM first divides read counts by gene lengths [130]. In other words, the TPM metric normalizes for gene length first, and sequencing depth second.

The TPM/RPKM/FPKM units allow for the comparison of gene expression across different samples. This can provide insight on the differing transcriptional landscape across tissue or cell types. One main approach in which gene expression values are used is to compare different biological conditions to identify “differentially expressed” genes [130]. However, given the normalization methods described result in proxies of gene expression, as RNA-seq represents a sampling of the true mRNA molecules present in a sample, statistical tests are required to identify the extent to which a gene is differentially expressed [130,131].

Traditionally RNA-seq has involved short-read sequencing, with read lengths between 25-100 bps. While this is useful for assigning a read to a gene of origin, one drawback of the approach is that it does not assess full-length isoforms, which represent the true biological unit of the transcriptome. However, methods have been developed for the probabilistic assignment of each sequencing read to its transcript of origin. Such methods, which rely on the available gene and transcript annotation,

account for confounding factors such as isoform length and GC content of the isoforms, which can affect mapping and 3' bias which is the preferential coverage of bases close to the 3' end of a transcript due to RNA degradation [132–136]. Recently developed long-read sequencing technologies aim to sequence full isoforms instead of short fragments, and are discussed in the final chapter of this thesis.

Evaluating alternative splicing with RNA-seq

The excision of introns from pre-mRNA in mRNA processing is called splicing and is an essential biological process in eukaryotic organisms [137,138]. Alternative splicing, which is the process wherein certain exons are skipped, creates protein diversity from the approximately 20,000 genes in the human genome [139].

In RNA-seq data, short read sequences can be confined within an exon or untranslated region, or they may map to exon-exon junctions, indicating the two exons are spliced together during mRNA maturation. Splice-aware mapping algorithms can successfully identify such junction reads, allowing for single base pair resolution insight into patterns of alternative splicing and exon inclusion rates [140].

Similar to differential expression analysis, algorithms exist for differential splicing analysis between conditions [130]. In addition, genetic variants that disrupt splicing have been linked to a wide variety of common disorders such as cancer and neurodegenerative disease [141,142]. Splice-affecting variants have been linked to several Mendelian disorders [143–145]. A canonical rare disease example is Hutchinson Gilford progeria syndrome in which a heterozygous *de novo* synonymous variant in *LMNA* results in gain of splicing form within exon 11, resulting in truncation of the protein

product [145]. It is estimated that over 70% of patients with this disorder have this recurrent mutation [145]. Other cases of splice-affecting mutations that do not occur in the essential 2 base pair junction have been reported, however many such cases are likely to remain undiagnosed.

Variant calling with RNA-seq

Given that germline variants will be present on mRNA products of a gene, RNA-seq can be thought of as a form of exome sequencing in genes that are expressed in the sample [146]. For some disorders, such as cancer, mutations in expressed regions may be of greater interest, as they are more likely to affect cellular function, and therefore identifying genetic variation via RNA-seq can be useful [147,148]. In addition to germline genetic variation, identifying variant information from RNA-seq can offer additional information to DNA sequencing, as it can identify somatic variation occurring only in a given tissue that may not be detectable in the tissue used for WES/WGS, and it can provide insight into post-transcriptional processes, such as RNA editing [146,149,150].

Varying sensitivity and specificity metrics have been reported for genetic variant identification for RNA-seq data. One study reports that approximately 70% of coding variants were identified by RNA-seq [149], while other reports 92% of all expected variants in expressed exons could be detected at > 10 x coverage [146]. In addition, 98% of variants identified with RNA-seq were reported to also be captured by WES/WGS [149]. An additional study reported 95% and 80% sensitivity for single nucleotide mutations and indels, respectively [151] highlighting the potential utility of RNA-seq data as a complementary tool for variant detection to DNA sequencing [147].

Identifying allele imbalance with RNA-seq

In addition to the ability of RNA-seq to identify genetic variation, it can also distinguish between expression of haplotypes [152]. In other words, at heterozygous sites in the human genome, RNA-seq can help detect unequal expression of the two chromosomal gene copies [153]. This approach is called allele-specific expression (ASE) analysis and can reveal insight into genetic imprinting, X-inactivation, truncating mutations causing nonsense-mediated decay, and allele specific transcription induced by genetic changes [153–155]. Several tools and frameworks have been developed to account for biases that can affect ASE estimates, such as mapping bias, quality control of read counts, and statistical tests to assign significance to ASE estimates [152,154]. In the case of variant interpretation, ASE can help prioritize genes where unequal expression between haplotypes exists, without identifying the causative mechanism of ASE, and prioritize the locus for follow-up interpretation.

Overview of large-scale population transcriptome datasets

As discussed in this chapter, interpreting the functional impact of variation in the genome is a central goal of human genetics, but is difficult to do based on DNA sequence alone. Therefore, integrating functional genomics information, such as RNA-seq, to link genetic variation to molecular phenotypes such as gene expression and splicing, as an intermediate link between genetics and disease has been a focus of the human genetics' community [156]. In this effort, several large-scale transcriptome datasets have been generated to decipher the effect of genetic variation on cellular phenotypes.

Here we summarize some of the larger datasets and touch on efforts to harmonize the project specific transcriptome sequencing efforts.

A canonical dataset of human genetic variation is the 1,000 Genomes data, which provided genome sequencing on diverse populations [157]. The Geuvadis Consortium produced RNA-seq from lymphoblastoid cells lines from 465 individuals in the 1,000 Genomes Project, with approximately 89-95 samples per five represented populations: Utah Residents with Northern and Western European Ancestry, Finnish from Finland, British in England and Scotland, Toscani in Italy, and Yoruba in Nigeria. The project showed that integrating RNA and DNA sequencing data can help uncover the landscape of regulatory variation, including variants that have effects on gene expression (commonly referred to as expression quantitative trait loci, or eQTLs), which can help further resolve GWAS loci that have been associated with disease.

The Cancer Genome Atlas (TCGA) has produced RNA-seq data from over 8,000 tissue samples over 30 cancer types and normal tissues to characterize gene expression alterations across and within cancer types [158,159] to complement their DNA sequencing efforts from tumors and normal samples. This effort has helped explore biomarkers of cancer subtypes and stages within a cancer subtype [160,161], the effect of splice disruptions in cancer pathogenesis [162–164], and the identification of somatic variation in tumors [165,166].

Geuvadis offered transcriptome sequencing from a diverse range of populations in a single tissue, and TCGA offered a disease-specific transcriptome dataset across a variety of tissues. The Genotype Tissue Expression Consortium (GTEx) in contrast aimed to provide transcriptome sequencing across several human tissue types in a

broad range of individuals. The consortium, whose initial goals were to create a resource to enable systematic analyses of genetic variation and disease, to fine-map GWAS associations, and to provide a biobank of tissues for other assays, has to date released transcriptome sequencing samples from 714 donors in 53 tissue types, corresponding to a total of 11,688 samples in the version 7 of the dataset [167,168]. The version 8 dataset, to be released in 2019, will include data from over 900 donors corresponding to a transcriptome sequencing dataset of almost 20,000 samples [Kristin Ardlie, personal communication]. The GTEx dataset has allowed for the characterization of gene expression patterns across tissues [169], development of statistical methods and analyses of expression and splicing QTLs and the characterization of their link to common and rare disease [137,141,168,170–174]

In addition to the consortium-based large-scale transcriptome sequencing efforts, several project specific RNA-seq datasets have been produced. Currently, it is estimated that there over 70,000 such project specific RNA-seq datasets [175]. Several efforts to joint-process and harmonize these datasets for combined analyses have been made including the TOIL pipeline, which presents a uniform pipeline for data alignment and gene and transcript quantification [176] as well as recount2, which allows researchers to search for keywords in project abstracts and download gene, isoform and junction quantifications [175]. These databases include RNA-seq data that is not available in the Geuvadis, TCGA or GTEx datasets such as fetal brain expression data and experimental data from stem cell experiments, and can be a useful resource to complement the larger datasets.

Bibliography

1. Collins, F. S. & McKusick, V. A. Implications of the Human Genome Project for medical science. *JAMA* **285**, 540–544 (2001).

2. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).

3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

4. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).

5. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228–237 (2003).

6. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).

7. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).

8. Litt, M. & Luty, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**, 397–401 (1989).

9. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).

10. Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509–517 (1987).

11. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).

12. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).

13. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).

14. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new

loci associated with body mass index. *Nat. Genet.* **42**, 937 (2010).

15. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

16. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

17. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).

18. Consortium, S. W. G. of T. P. G. & Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

19. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

20. Chakravarti, A. Genomic contributions to Mendelian disease. *Genome Res.* **21**, 643–644 (2011).

21. Berry, R. J., Buehler, J. W., Strauss, L. T., Hogue, C. J. & Smith, J. C. Birth weight-specific infant mortality due to congenital anomalies, 1960 and 1980. *Public Health Rep.* **102**, 171–181 (1987).

22. Scriver, C. R., Neal, J. L., Saginur, R. & Clow, A. The frequency of genetic disease and congenital malformation among patients in a pediatric hospital. *Can. Med. Assoc. J.* **108**, 1111–1115 (1973).

23. Boycott, K. M. *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).

24. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).

25. Christianson, A., Howson, C. P., Modell, B. & Others. March of Dimes: global report on birth defects, the hidden toll of dying and disabled children. *March of Dimes: global report on birth defects, the hidden toll of dying and disabled children.* (2005).

26. Angelis, A., Tordrup, D. & Kanavos, P. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. *Health Policy* **119**, 964–979 (2015).

27. Carmichael, N., Tsipis, J., Windmueller, G., Mandel, L. & Estrella, E. 'Is it Going to Hurt?': The Impact of the Diagnostic Odyssey on Children and Their Families. *J. Genet. Couns.* **24**, 325–335 (2015).

28. Madeo, A. C., O'Brien, K. E., Bernhardt, B. A. & Biesecker, B. B. Factors associated with perceived uncertainty among parents of children with undiagnosed medical conditions. *Am. J. Med. Genet. A* **158A**, 1877–1884 (2012).

29. Engel, P. A., Bagal, S., Broback, M. & Boice, N. Physician and patient perceptions regarding physician training in rare diseases: the need for stronger educational initiatives for physicians. *J Rare Dis* **1**, 1–14 (2013).

30. Faurisson, F. Survey of the delay in diagnosis for 8 rare diseases in Europe: EurordisCare2. *European Organisation for Rare Diseases Web site* (2004).

31. Makela, N. L., Birch, P. H., Friedman, J. M. & Marra, C. A. Parental perceived value of a diagnosis for intellectual disability (ID): A qualitative comparison of families with and without a diagnosis for their child's ID. *Am. J. Med. Genet.* **149A**, 2393–2402 (2009).

32. Lasker, J. N., Sogolow, E. D. & Sharim, R. R. The role of an online community for people with a rare disease: content analysis of messages posted on a primary biliary cirrhosis mailinglist. *J. Med. Internet Res.* **7**, e10 (2005).

33. Selkirk, C. G., Veach, P. M., Lian, F., Schimmenti, L. & LeRoy, B. S. Parents' Perceptions of Autism Spectrum Disorder Etiology and Recurrence Risk and Effects of their Perceptions on Family Planning: Recommendations for Genetic Counselors. *Journal of Genetic Counseling* **18**, 507–519 (2009).

34. Evers-Kiebooms, G., Denayer, L. & Van den Berghe, H. A child with cystic fibrosis: II. Subsequent family planning decisions, reproduction and use of prenatal diagnosis. *Clin. Genet.* **37**, 207–215 (1990).

35. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12**, 228 (2011).

36. Gräber, S. *et al.* Effects of Lumacaftor-Ivacaftor Therapy on CFTR Function in Phe508del Homozygous Patients with Cystic Fibrosis. *Cystic fibrosis* (2018). doi:10.1183/13993003.congress-2018.pa3415

37. Donaldson, S. H. *et al.* Tezacaftor/Ivacaftor in Subjects with Cystic Fibrosis and F508del/F508del-CFTR or F508del/G551D-CFTR. *American Journal of Respiratory and Critical Care Medicine* **197**, 214–224 (2018).

38. Van Goor, F. *et al.* Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18843–18848 (2011).

39. Ramsey, B. W. *et al.* A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N. Engl. J. Med.* **365**, 1663–1672 (2011).

40. Prakash, V. Spinraza—a rare disease success story. *Gene Ther.* **24**, 497 (2017).

41. Singh, R. N., Singh, N. N., Singh, N. K. & Androphy, E. J. Spinal muscular atrophy (SMA) treatment via targeting of SMN2 splice site inhibitory sequences. *US Patent* (2010).

42. Aartsma-Rus, A., Ginjaar, I. B. & Bushby, K. The importance of genetic diagnosis for Duchenne muscular dystrophy. *J. Med. Genet.* **53**, 145–151 (2016).

43. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–7 (2005).

44. Ayadi, A. *et al.* Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm. Genome* **23**, 600–610 (2012).

45. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).

46. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19096–19101 (2009).

47. Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S. & Girirajan, S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci. Rep.* **7**, 885 (2017).

48. Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31–46 (2010).

49. Tan, R. *et al.* An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Human Mutation* **35**, 899–907 (2014).

50. Kadalayil, L. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief. Bioinform.* **16**, 380–392 (2015).

51. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47**, 717–726 (2015).

52. Biesecker, L. G. & Green, R. C. Diagnostic Clinical Genome and Exome Sequencing. *New England Journal of Medicine* **370**, 2418–2425 (2014).

53. Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum. Mutat.* **36**, 815–822 (2015).

54. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–51 (2010).

55. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome

sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5473–5478 (2015).

56. Ankala, A. *et al.* A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. *Ann. Neurol.* **77**, 206–214 (2015).

57. Calvo, S. E. *et al.* Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci. Transl. Med.* **4**, 118ra10 (2012).

58. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014).

59. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880–1887 (2014).

60. Posey, J. E. *et al.* Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet. Med.* **18**, 678–685 (2016).

61. Lek, M. & MacArthur, D. The Challenge of Next Generation Sequencing in the Context of Neuromuscular Diseases. *J Neuromuscul Dis* **1**, 135–149 (2014).

62. Sweet, M., Taylor, M. R. G. & Mestroni, L. Diagnosis, prevalence, and screening of familial dilated cardiomyopathy. *Expert Opin Orphan Drugs* **3**, 869–876 (2015).

63. Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: Insights into rare diseases. *Genetics in Medicine* 1 (2011). doi:10.1097/gim.0b013e318232a005

64. Frésard, L. & Montgomery, S. B. Diagnosing rare diseases after the exome. *Cold Spring Harb Mol Case Stud* **4**, (2018).

65. Philippakis, A. A. *et al.* The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).

66. Beaulieu, C. L. *et al.* FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am. J. Hum. Genet.* **94**, 809–817 (2014).

67. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).

68. Bamshad, M. J. *et al.* The Centers for Mendelian Genomics: A new large-scale initiative to identify the genes underlying rare Mendelian conditions. *American Journal of Medical Genetics Part A* **158A**, 1523–1525 (2012).

69. Ewans, L. J. *et al.* Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genet. Med.* **20**, 1564–1574 (2018).

70. Wenger, A. M., Guturu, H., Bernstein, J. A. & Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet. Med.* **19**, 209–214 (2017).

71. Al-Nabhani, M. *et al.* Reanalysis of exome sequencing data of intellectual disability samples: Yields and benefits. *Clin. Genet.* **94**, 495–501 (2018).

72. Tawil, R., van der Maarel, S. M. & Tapscott, S. J. Facioscapulohumeral dystrophy: the path to consensus on pathophysiology. *Skelet. Muscle* **4**, 12 (2014).

73. Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–1653 (2010).

74. Lupski, J. R. Digenic inheritance and Mendelian disease. *Nat. Genet.* **44**, 1291–1292 (2012).

75. Gonzaga-Jauregui, C. *et al.* Exome Sequence Analysis Suggests that Genetic Burden Contributes to Phenotypic Variability and Complex Neuropathy. *Cell Rep.* **12**, 1169–1183 (2015).

76. Brooks, A. S., Oostra, B. A. & Hofstra, R. M. W. Studying the genetics of Hirschsprung's disease: unraveling an oligogenic disorder. *Clin. Genet.* **67**, 6–14 (2005).

77. Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).

78. Coll, M. *et al.* Incomplete Penetrance and Variable Expressivity: Hallmarks in Channelopathies Associated with Sudden Cardiac Death. *Biology* **7**, (2017).

79. Cooper, D. N., 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. *Hum. Genet.* **132**, 1077–1130 (2013).

80. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).

81. Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).

82. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

83. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

84. Spurdle, A. B. *et al.* Prediction and assessment of splicing alterations: implications

for clinical testing. *Hum. Mutat.* **29**, 1304–1313 (2008).

85. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

86. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genetics* **7**, e1001322 (2011).

87. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455–64 (2014).

88. Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A. & Neale, B. M. Discovery of rare variants for complex phenotypes. *Hum. Genet.* **135**, 625–634 (2016).

89. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

90. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).

91. Singh, T. *et al.* The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **49**, 1167–1173 (2017).

92. Satterstrom, F. K. *et al.* ASD and ADHD have a similar burden of rare protein-truncating variants. doi:10.1101/277707

93. Cardinale, C. J., Kelsen, J. R., Baldassano, R. N. & Hakonarson, H. Impact of exome sequencing in inflammatory bowel disease. *World J. Gastroenterol.* **19**, 6721–6729 (2013).

94. Richards, S. *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.* **17**, 405–424 (2015).

95. Strande, N. T., Brnich, S. E., Roman, T. S. & Berg, J. S. Navigating the nuances of clinical sequence variant interpretation in Mendelian disease. *Genet. Med.* **20**, 918–926 (2018).

96. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).

97. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

98. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of

human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

99. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

100. Adzhubei, I. A., Schmidt, S. & Peshkin, L. ramensky Ve, Gerasimova A., Bork P., Kondrashov AS, Sunyaev Sr. *Nat. Methods* **7**, 248–249 (2010).

101. Eng, L. *et al.* Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. *Hum. Mutat.* **23**, 67–76 (2004).

102. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).

103. Huang, Z. & Teeling, E. C. ExUTR: a novel pipeline for large-scale prediction of 3'-UTR sequences from NGS data. *BMC Genomics* **18**, 847 (2017).

104. Jayaram, N., Usvyat, D. & R Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* (2016). doi:10.1186/s12859-016-1298-9

105. Ernst, C. *et al.* Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Med. Genomics* **11**, 35 (2018).

106. Vihinen, M. & Niroula, A. How good are pathogenicity predictors in detecting benign variants? doi:10.1101/408153

107. Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am. J. Hum. Genet.* **99**, 595–606 (2016).

108. Auer, P. L. *et al.* Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *Am. J. Hum. Genet.* **99**, 791–801 (2016).

109. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

110. Landstrom, A. P. *et al.* Interpreting Incidentally Identified Variants in Genes Associated With Catecholaminergic Polymorphic Ventricular Tachycardia in a Large Cohort of Clinical Whole-Exome Genetic Test Referrals. *Circulation: Arrhythmia and Electrophysiology* **10**, (2017).

111. Chagnon, P. *et al.* A missense mutation (R565W) in cirhin (FLJ14728) in North American Indian childhood cirrhosis. *Am. J. Hum. Genet.* **71**, 1443–1449 (2002).

112. Karczewski, K. J., Francioli, L. C., Tiao, G. & Cummings, B. B. Variation across

141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* (2019).

113. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).

114. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).

115. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

116. Van Hout, C. V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* 572347 (2019). doi:10.1101/572347

117. Bennett, C. A., Petrovski, S., Oliver, K. L. & Berkovic, S. F. ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. *Neurol Genet* **3**, e163 (2017).

118. Kobayashi, Y. *et al.* Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med.* **9**, 13 (2017).

119. Paludan-Müller, C. *et al.* Analysis of 60 706 Exomes Questions the Role of De Novo Variants Previously Implicated in Cardiac Disease. *Circ. Cardiovasc. Genet.* **10**, (2017).

120. Carlston, C. M. *et al.* Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Hum. Mutat.* **38**, 517–523 (2017).

121. Nagalakshmi, U., Waern, K. & Snyder, M. RNA-Seq: a method for comprehensive transcriptome analysis. *Curr. Protoc. Mol. Biol.* **89**, 4–11 (2010).

122. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).

123. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).

124. Yamada, A. *et al.* A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer. *Sci. Rep.* **8**, 575 (2018).

125. Hu, Y., Lan, W. & Miller, D. Next-Generation Sequencing for MicroRNA Expression Profile. *Methods Mol. Biol.* **1617**, 169–177 (2017).

126. Ow, M. C., Lau, N. C. & Hall, S. E. Small RNA library cloning procedure for deep sequencing of specific endogenous siRNA classes in Caenorhabditis elegans. *Methods Mol. Biol.* **1173**, 59–70 (2014).

127. Krishnan, P. *et al.* Profiling of Small Nucleolar RNAs by Next Generation Sequencing: Potential New Players for Breast Cancer Prognosis. *PLoS One* **11**, e0162622 (2016).

128. Jakobi, T. & Dieterich, C. Deep Computational Circular RNA Analytics from RNA-seq Data. *Methods Mol. Biol.* **1724**, 9–25 (2018).

129. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).

130. Li, W. V. & Li, J. J. Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology* **6**, 195–209 (2018).

131. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

132. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

133. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

134. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).

135. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification with kallisto. *Nat. Biotechnol.* **34**, 525–527 (2016).

136. Zheng, W., Chung, L. M. & Zhao, H. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* **12**, 290 (2011).

137. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

138. Brett, D., Pospisil, H., Valcárcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nat. Genet.* **30**, 29–30 (2002).

139. Wang, Y. *et al.* Mechanism of alternative splicing and its regulation. *Biomed Rep* **3**, 152–158 (2015).

140. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

141. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

142. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).

143. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta* **1792**, 14–26 (2009).

144. Colapietro, P. *et al.* NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum. Genet.* **113**, 551–554 (2003).

145. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome. *Nature* **423**, 293–298 (2003).

146. Quinn, E. M. *et al.* Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One* **8**, e58815 (2013).

147. Neums, L. *et al.* VaDiR: an integrated approach to Variant Detection in RNA. *Gigascience* **7**, (2018).

148. Chepelev, I., Wei, G., Tang, Q. & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.* **37**, e106 (2009).

149. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).

150. Oczkowicz, M., Szmatoła, T., Piórkowska, K. & Ropka-Molik, K. Variant calling from RNA-seq data of the brain transcriptome of pigs and its application for allele-specific expression and imprinting analysis. *Gene* **641**, 367–375 (2018).

151. Prodduturi, N., Bhagwate, A., Kocher, J.-P. A. & Sun, Z. Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine. *BMC Med. Genomics* **11**, 67 (2018).

152. Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31**, 2497–2504 (2015).

153. Zhuo, Z., Lamont, S. J. & Abasht, B. RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. *Sci. Rep.* **7**, 11944 (2017).

154. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome*

Biol. **16**, 195 (2015).

155. Korir, P. K. & Seoighe, C. Inference of allele-specific expression from RNA-seq data. *Methods Mol. Biol.* **1112**, 49–69 (2014).

156. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

157. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

158. Peng, L. *et al.* Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Scientific Reports* **5**, (2015).

159. Wang, Q. *et al.* Unifying cancer and normal RNA sequencing data from different sources. *Sci Data* **5**, 180061 (2018).

160. Huo, T., Canepa, R., Sura, A., Modave, F. & Gong, Y. Colorectal cancer stages transcriptome analysis. *PLoS One* **12**, e0188697 (2017).

161. Rau, A., Flister, M., Rui, H. & Auer, P. L. Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics* **35**, 62–68 (2019).

162. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* **7**, 45 (2015).

163. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211–224.e6 (2018).

164. Kim, D. *et al.* Population-dependent Intron Retention and DNA Methylation in Breast Cancer. *Mol. Cancer Res.* **16**, 461–469 (2018).

165. Spurr, L. *et al.* Systematic pan-cancer analysis of somatic allele frequency. *Sci. Rep.* **8**, 7735 (2018).

166. Coudray, A., Battenhouse, A. M., Bucher, P. & Iyer, V. R. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* **6**, e5362 (2018).

167. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

168. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. doi:10.1101/074450

169. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

170. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics* **50**, 956–967 (2018).

171. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747–1751 (2017).

172. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).

173. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).

174. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).

175. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).

176. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).

Chapter 2

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

Abstract

Exome and whole-genome sequencing are becoming increasingly routine approaches in Mendelian disease diagnosis. Despite their success, the current diagnostic rate for genomic analyses across a variety of rare diseases is approximately 25-50%[1–4]. Here, we explore the utility of transcriptome sequencing (RNA-seq) as a complementary diagnostic tool in a cohort of 50 patients with genetically undiagnosed rare muscle disorders. We describe an integrated approach to analyze patient muscle RNA-seq, leveraging an analysis framework focused on the detection of transcript-level changes that are unique to the patient compared to over 180 control skeletal muscle samples. We demonstrate the power of RNA-seq to validate candidate splice-disrupting mutations and to identify splice-altering variants in both exonic and deep intronic regions, yielding an overall diagnosis rate of 35%. We also report the discovery of a highly recurrent *de novo* intronic mutation in *COL6A1* that results in a dominantly acting splice-gain event, disrupting the critical glycine repeat motif of the triple helical domain. We identify this pathogenic variant in a total of 27 genetically unsolved patients in an external collagen VI-like dystrophy cohort, thus explaining approximately 25% of patients clinically suggestive of collagen VI dystrophy in whom prior genetic analysis is negative. Overall, this study represents a large systematic application of transcriptome sequencing to rare disease diagnosis and highlights its utility for the detection and interpretation of variants missed by current standard diagnostic approaches.

Introduction

RNA sequencing as a diagnostic tool for Mendelian disease

RNA is the direct functional output of genetic variation on gene expression, making it a useful tool to assess the pathogenicity of variants as well as to identify genetic lesions that may elude DNA sequencing technologies. Given that DNA sequence changes first manifest in the way genes are expressed, we hypothesized that studying transcriptional changes in the affected tissue of patients with Mendelian disease would provide valuable insights into the cause of disease.

RNA-seq is the current state of the art technology for transcriptomics research, allowing the analysis of transcripts at single base pair resolution [5]. We hypothesized that the use of RNA-seq will empower Mendelian disease diagnosis by validating candidate pathogenic variants uncovered by DNA sequencing and by identifying new causal variants where DNA sequencing alone does not provide a definitive molecular diagnosis.

RNA-seq can provide several insights that are currently missed by DNA sequencing. Transcriptional aberrations, such as skipping of an exon, have been shown in many cases to explain Mendelian disorders[6–10]. Such aberrations are often due to canonical splice site variants that obliterate efficient splicing at an exon-intron junction, but can also be caused by mutations in the extended splice site region or exonic splice enhancer motif [7,8,11]. Currently, it is well understood that mutations that disrupt the canonical GT/AG splice motifs have detrimental effects on transcription. However, other classes of splice mutations have often been ignored or characterized as variants of

unknown significance (VUS)[12]. This is a prime motivating example for the use of RNA sequencing in diagnosis given that RNA-seq can help identify splicing patterns in patients that are missing in controls and sequencing including splice defects, somatic variation and allele-specific expression.

While exome sequencing capture technologies are constantly being improved, there are regions of the genome that WES does not efficiently sequence[13,14]. Hybridization based-capture technologies are limited by specific targets and capture efficiency of the probes, leaving many disease-relevant exons uncovered[14]. The input for RNA sequencing is the expressed transcripts, therefore it is not limited by probe-based capture systems. Therefore RNA-sequencing can complement WES through analysis of variants in regions poorly captured with WES.

In next-generation sequencing studies, DNA is usually derived from blood or saliva and not the affected tissue. In these cases, WES may also miss pathogenic somatic variants that are only found in the affected tissue. Somatic variants are mutations that occur post-zygotically and result in variants being present in a subset of tissues[15,16]. Somatic variation has already been linked to several Mendelian diseases including neuromuscular disorders[15–17]. We hypothesized that RNA-seq would be capable of identifying somatic variants present in an affected tissue that may be absent in blood or saliva-based WES.

Lastly, regulatory variants. or variants in the non-coding portion of a gene, remain difficult to interpret with DNA sequencing. It has been suggested that a large proportion of Mendelian regulatory variants may occur in the promoter region, resulting in reduced or eliminated expression of the gene from the affected copy[18].· In this case, only mRNA

transcribed from the unmodified promoter will be expressed and thus will be the only transcript picked up by RNA-seq. In these cases, we expect to see RNA-seq to show striking allelic imbalance in the gene, with heterozygous sites in the gene, identified by DNA sequencing, appearing homozygous in the RNA data [19].

Neuromuscular disorders as a model Mendelian disease for RNA-sequencing

Neuromuscular disorders (NMDs) are broadly characterized by progressive skeletal muscle weakness, fatigue, and loss of neuromotor capabilities. There is considerable clinical heterogeneity in neuromuscular disease with a wide spectrum of onset, rate of progression and clinical severity, making the broad class of disease difficult to diagnose based on phenotype alone[20,21]. WES has had a dramatic impact on both the understanding and clinical diagnosis of neuromuscular disorders. Over 150 genes have been associated to muscle disease and the current rate of genetic diagnosis in muscle disease patients is approximately 40-50% [20,22].

DNA sequence remains largely constant in tissue types, with the exception of somatic mutations. In contrast, recent large-scale studies have shown that gene expression and mRNA isoforms vary widely across tissue types and that up to 80% alternative splicing of pre-mRNA may be tissue-specific[23]. Therefore, sequencing the affected tissue is critical to correctly interpreting the effect of genetic variation on the transcriptomic landscape in muscle disease patients. Furthermore, our analysis of RNA-seq from 200 muscle biopsies from samples in the Genotype Tissue Expression Consortium (GTEx) demonstrates that genes associated with neuromuscular disease

are poorly expressed in blood, making blood-derived RNA-seq underpowered to detect relevant transcriptional aberrations that can cause muscle disease (see below).

Muscle biopsies are routine clinical practice for undiagnosed muscle disease patients that are used for histopathology screen as well as protein-based assays [24,25]. This current ease of access to affected skeletal muscle tissue coupled with the heterogeneity of neuromuscular disease makes it an attractive class of disorders for studies of transcriptome sequencing guided diagnosis.

Study design

To investigate the value of RNA-seq for diagnosis, we obtained primary muscle RNA from 63 patients with putatively monogenic muscle disorders. Thirteen of these cases had been previously diagnosed with variants expected to have an effect on transcription, such as loss-of-function or essential splice site variants, allowing us to validate the capability of RNA-seq to identify transcriptional aberrations (Appendix Table 2.1). The remaining cohort of 50 genetically undiagnosed patients included cases for whom DNA sequencing had prioritized variants predicted to alter RNA splicing or strong candidate genes, as well as cases with no strong candidates from genetic analysis.

We sought to explore the utility of transcriptome sequencing as a complementary diagnostic tool to exome and whole genome analysis. We reasoned that RNA-seq would allow us to interpret variants previously identified through genetic analysis and may pinpoint genetic lesions that may have eluded DNA sequencing. To interpret transcriptional aberrations seen in patients, we obtained a reference panel of RNA-seq data from skeletal muscle samples generated by the GTEx project[26]. Our framework

was based on identifying transcriptional aberrations present in patients that are missing in GTEx controls. We first validated the capacity of RNA-seq to resolve transcriptional aberrations in thirteen patients with prior genetic diagnosis and then analyzed the remaining fifty genetically undiagnosed patients to detect aberrant splice events and allele-specific expression and performed variant calling from RNA-seq data to identify pathogenic events or to prioritize genes for closer analysis.

Importance of sequencing the disease-relevant tissue

Recent large-scale studies have shown that gene expression and mRNA isoforms vary widely across tissues, indicating that for many diseases, sequencing the disease-relevant tissue will be valuable for the correct interpretation of genetic variation[23,27]. This is illustrated by the relative expression of known muscle disease genes in skeletal muscle, whole blood, and fibroblast samples from the Genotype Tissue Expression Consortium project (GTEx)[26,28] (Figure 2.1). The majority of the most commonly disrupted genes in muscle disease are poorly expressed in blood and fibroblasts, suggesting RNA-seq from these easily accessible tissues may be underpowered to detect relevant transcriptional aberrations in certain genes. For these reasons, we chose to pursue RNA-seq from primary muscle tissue biopsies, which are routinely performed as part of the diagnostic evaluation of undiagnosed muscle disease patients[25,29].

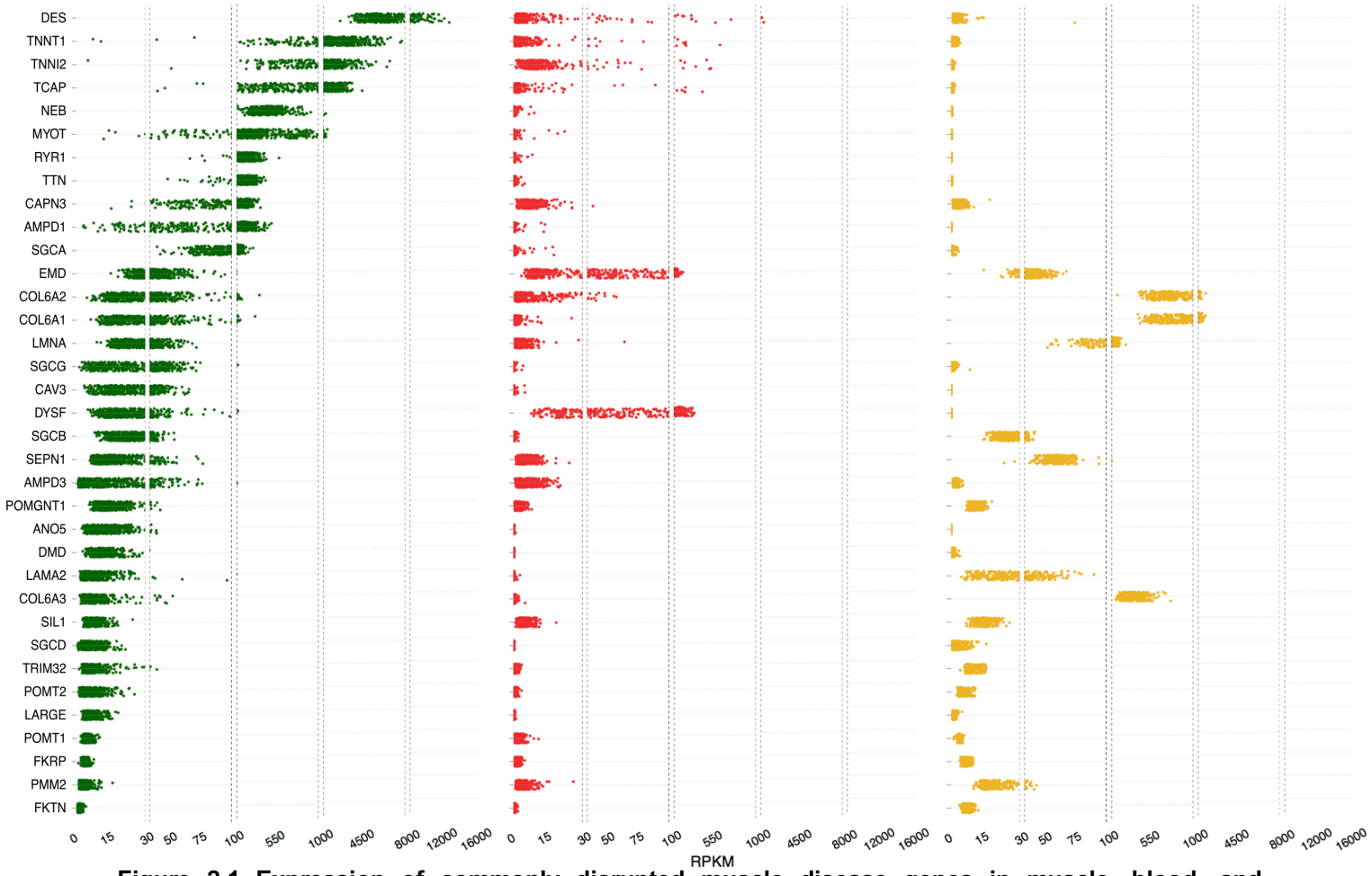


Figure 2.1 Expression of commonly disrupted muscle disease genes in muscle, blood, and fibroblasts. Expression of commonly disrupted neuromusucular disease genes in 430 muscle (green), 393 whole blood (red), and 283 fibroblast (yellow) GTEx samples shows these genes are relatively poorly expressed in more easily accessible blood and fibroblast tissues.

Materials and Methods

Clinical sample selection

Patient cases with available muscle biopsies were referred by clinicians from March 2013 through June 2016. Samples fell into four broad categories:

1. patients for whom previous genetic analysis had resulted in a diagnosis with at least one loss-of-function or essential splice site variant, serving as

positive controls to assess the capability of RNA-seq to identify the transcriptional effect of the variants (n = 13, patient IDs starting with ‘D’).

2. patients with candidate extended splice site variants that had been categorized as variants of unknown significance for which assignment of pathogenicity would result in a complete diagnosis for the patient (n=4, patient IDs starting with ‘E’).

3. patients for whom a strong candidate gene was implicated due to either a well-defined monogenic disease phenotype, such as patients with clear Duchenne muscular dystrophy evidenced by clinical diagnosis and loss of dystrophin expression (n=6), or to the presence of one pathogenic heterozygous variant identified in a gene matching the patient’s phenotype, without a second pathogenic variant in that gene (n= 6, patient IDs starting with ‘C’).

4. patients with no strong candidates based on previous genetic analysis such as exome or whole genome sequencing (n=34, patient IDs starting with ‘N’)

Patients that fit categories 2-4 are referred to as undiagnosed prior to RNA-seq and constitute the denominator for the 35% diagnosis rate. All patients had prior analysis of exome and/or whole genome sequencing data, except two cases (patients

E4 and D11) for whom targeted sequencing had identified a candidate extended and essential splice site variant, respectively. We favored cases with previous trio exome or whole genome sequencing: 29/63 patients had complete trios, with 3 additional patients having one parent sequenced. Although age of onset was not considered as an exclusion criterion, a majority of the patients in the cohort had a congenital or early-childhood onset primary muscle disorder.

Muscle biopsies or RNA were shipped frozen from clinical centers via a liquid nitrogen dry shipper and stored in liquid nitrogen cryogenic storage. Before submission to the sequencing platform, all muscle samples were visually inspected, photographed, cut into 50 µm sections on Leica CM 1950 model cryostat, and transferred to pre-chilled cryotubes in preparation for RNA extraction. When muscle arrived embedded in OCT, 8 µm transverse cryosections were mounted on positively charged Superfrost plus slides (VWR, 48311-703) and stained with hematoxylin and eosin (H&E) to assess the relative proportion of muscle versus fibrosis and adipose infiltration as well as the presence of overt freeze-thaw artifact**.** All samples analyzed with H&E showed muscle quality sufficient to proceed to RNA-seq.

Selection of GTEx controls

GTEx data were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap) under accession phs000424.v6.p1. From 430 available GTEx skeletal muscle RNA-seq samples, we selected 184 samples based on RNA Integrity (RIN) score (between 6 and 9), number of non-duplicate uniquely mapped read pairs (between 35M and 75M), and ischemic time (<12 hours) to remove any samples that were outliers for these quality

metrics. GTEx samples were further filtered to remove samples with known clinical conditions such as Klinefelter's syndrome or those for whom death followed after long or intermediate term illness or medical intervention (Hardy Scale 0, 3, or 4). Overall, approximately 80% of GTEx samples with available muscle RNA-seq are above the age of 40 (median age 54) and have BMI over 25 (median BMI 27). Thus we selected samples to enrich for younger GTEx donors to more closely match our patient cohort. All samples below the age of 50 were selected, resulting in 76 samples with high quality RNA-seq data. We then added older samples back on the criterion that their BMI was below 30. This resulted in a total of 184 GTEx control samples for our reference panel, with comparable male and female sample count (105 male and 79 female). This filtering method also enriched RNA-seq data from organ donors and surgical donors as opposed to postmortem samples (72% of selected GTEx controls are derived from surgical or organ donors vs 45% in the unfiltered dataset).

RNA sequencing, processing and quality control

RNA was extracted from muscle biopsies via the miRNeasy Mini Kit from Qiagen per kit instructions. All RNA samples were measured for quantity and quality. Samples had to meet the minimum cutoff of 250 ng of RNA and RNA Quality Score (RQS) of 6 to proceed with RNA-seq library prep. A fraction of samples falling below an RQS of 6 were also submitted for sequencing. All samples submitted had a range of RQSs between 3.5-8.

Sequencing was performed at the Broad Institute Genomics Platform using the same non-strand-specific protocol with poly-A selection of mRNA (Illumina TruSeq)

used in the GTEx sequencing project [28], to ensure consistency of our samples with GTEx control data. Paired end 76 bp sequencing was performed on Illumina HiSeq 2000 instruments, with sequence coverage of 50M or 100M. One sample (patient N33) was sequenced to higher depth at 500M reads allowing us to perform downsampling analysis of the effects of increasing RNA-seq depth.

GTEx BAM files downloaded from dbGaP were realigned after conversion to FASTQ files with Picard SamToFastq. Both patient and GTEx reads were aligned using Star 2-Pass[30] version v.2.4.2a using hg19 as the genome reference and Gencode V19 annotations[31]. Briefly, first-pass alignment was performed for novel junction discovery, and the identified junctions were filtered to exclude unannotated junctions with less than 5 uniquely mapped read support, as well as junctions found on the mitochondrial genome. These junctions were then used to create a new annotation file, and second-pass alignment was performed as recommended by the STAR manual to enable sensitive junction discovery. Duplicate reads were marked with Picard MarkDuplicates (v.1.1099).

Quality metrics for patient and GTEx RNA-seq data were obtained by running RNA-seQC[32] (v1.1.8) on STAR aligned BAMs. PCA on gene expression was performed based on RPKM values calculated by RNA-seQC. Two samples (D6 and N3) were removed due to outlier status in PCA, consistent with a high proportion of non-muscle tissue in the samples. For GTEx samples, the expression and exon-level read count data were downloaded from dbGAP under accession phs000424.v6. For PCA of exon inclusion metrics, we obtained PSI values for GTEx samples as described in [33] .

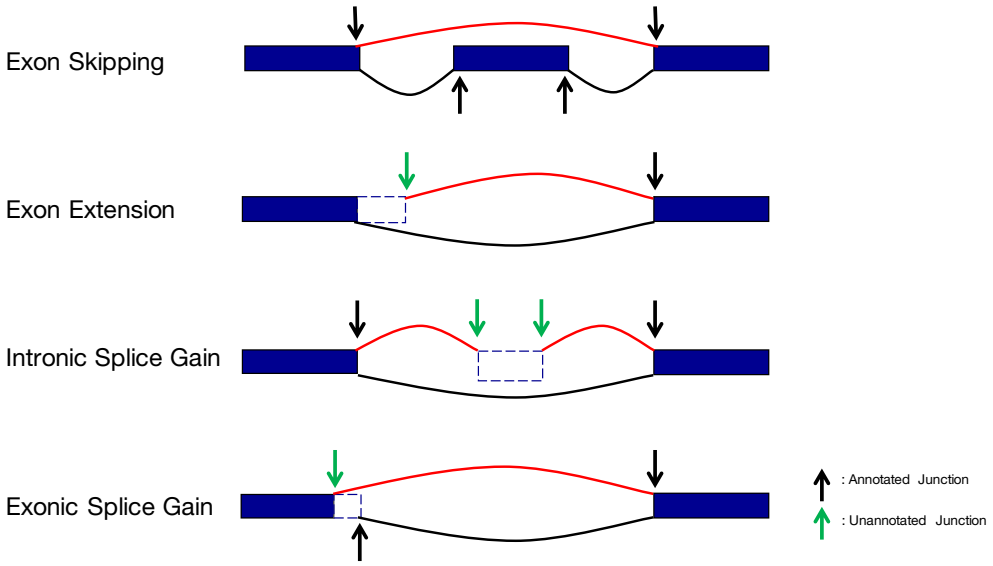
To ensure that patient DNA and RNA data were identity-matched, we compared variants identified in WES, WGS, and RNA-seq data. WES, WGS, and RNA-seq data were joint-genotyped for a set of ~5,800 common SNPs collated by Purcell et al.[34] using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 3.4[35]. We then calculated pairwise inheritance by descent (IBD) estimates between DNA and RNA-seq data using PLINK[36] (v1.08p). Relatedness coefficients for WES, WGS, and RNA-seq data from the same individual ranged from 0.67-1.00 across our samples (mean = 0.9), compared to a range of 0-0.18 (mean= 0.001) for non-matching individuals, confirming that the sources for DNA and RNA-seq were the same for each patient in our dataset.

Identification of pathogenic splice events

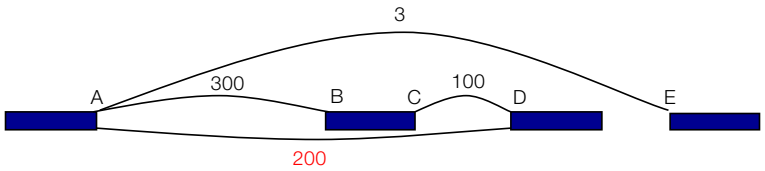
Splice junctions were identified from split-mapped reads, considering only uniquely aligned, non-duplicate reads that passed platform/vendor quality controls. For each splice junction we noted:

1. the genomic coordinates
2. the gene in which the junction was observed based on Gencode v.19
3. the number of samples in which the splice junction was observed
4. the number of total reads supporting the junction in 245 samples (184 GTEx and 61 patient)
5. the per-sample read support for the junction

A.



B.



Splice junction	Number of reads supporting splice junction	Type of Event
A-B	300	Canonical splicing
C-D	100	Canonical splicing
A-D	200	Exon skipping event (Unannotated exon-exon junction)
A-E	3	Likely mapping noise

Exon-Intron Junction	Number of reads at the exon-intron junction overlapping a canonical splicing junction
A	300
B	300
C	100
D	100

Splice Junction	Normalized Value
A-B	300/300 = 1
A-D	200/300 = .66
A-E	3/300 = 0.01
C-D	100/100 = 1

Figure 2.2 Overview of splice junction filtering approach. A. Types of possible aberrant splice events targeted for detection with splice junction discovery method. Among possible pathogenic splice events, such as exon skipping or intronic splice gain, shown in red, at least one exon-intron junction will be a part of a known annotated transcript, thus allowing for filtering of junctions that have no sharing with an annotated junction. **B.** An example of the normalization scheme for the case of heterozygous exon skipping. Exon-exon junctions A-B and C-D are wildtype splice events, whereas A-D represents an exon skipping event and A-E is likely mapping noise with low level read support. Every exon-exon junction is normalized by the maximum read support of a shared exon-intron junction that is annotated in Gencode v19. Here, the exon skipping event has 200 read support and the shared annotated exon-intron junctions have 100 and 300 read support. Therefore 200/300 is the normalized value that supports this event. In contrast, 3 reads support a splicing event between junction A-E, which is normalized by 300, therefore the normalized value is 0.01 and the junction is filtered.

We then performed local normalization of per-sample read support based on the support for the highest shared annotated junction (Figure 2.2A). For example, an exon-skipping event harbors two annotated exon-intron junctions, and we normalize this by the maximum of read count support for canonical splicing at these two wildtype junctions. This local normalization allows for filtering low-level mapping noise and accounts for stochastic gene expression and library size differences between samples (Figure 2.2B).

To identify pathogenic splice events, splice junctions in protein coding genes were filtered in terms of the number of samples a splice junction is present in and the number of reads and the normalized value supporting that junction. Specifically, we defined a sensitive cutoff at which an aberrant splice event is seen with at least 5% of the read support compared to the shared annotated junction, with at least 2 reads supporting the event. We also required a splice junction to contain at least one annotated exon-exon junction, indicating that the event was spliced into an existing transcript (Figure 2.2A). We performed analysis on a per-sample basis, each time requiring the normalized value of a given splice junction to be maximum in that sample and twice that of the next highest sample, allowing us to search for unique events in the patient.

All candidate pathogenic splice events were manually evaluated using the Integrative Genome Viewer (IGV)[37]. This resulted in the identification of aberrant splicing at 8/9 pathogenic essential splice site variants and resulted in the diagnosis of 10/17 patients in the study. A splice aberration was not observed around an essential splice site variant found in *TTN* in patient D5 due to insufficient number of reads

mapping to the local region. We extended filtering parameters to identify splice junctions present in fewer than 10 samples, but with high read support in each sample, allowing us to identify the intronic splice-gain event present in 4 patients in *COL6A1* (see below). We note that this approach would also identify putatively pathogenic splice aberrations for which there are GTEx carriers. The remaining 3 Duchenne muscular dystrophy patients were diagnosed through manual analysis of splicing patterns in *DMD* and resulted in the identification of splice disruption. Overlapping structural variants at these regions were confirmed by subsequent WGS.

Allele specific expression analysis

Allele counts for heterozygous variants were calculated from RNA-seq data using GATK ASEReadCounter package version 3.6[38]. Heterozygous variants that passed VQSR filtering were first extracted for each sample from exome sequencing VCFs (GTEx v6 VCF exome downloaded via accession phs000424.v6.p1) using GATK SelectVariants. The analysis was restricted to biallelic SNPs due to known issues in mapping bias in RNA-seq against indels[38]. Sample-specific VCFs and RNA-seq BAMs were inputted to GATK ASEReadCounter, requiring coverage in the RNA-seq data of each variant to be at least 20 reads, with a minimum base quality of 10 and counting only uniquely mapped reads.

To detect allele-specific expression unique to patients, we first calculated a distribution of allele balance in each gene based on GTEx reference and alternative allele counts and identified patients who fell outside of the 95% confidence interval for mean allele balance in the gene. This resulted in the identification of a median of 3

genes with allele imbalance in 189 neuromuscular disease genes. The method re-identified allele imbalance in all 4 cases where the patient was known to have a loss of function variant in trans with an additional variant. Allele imbalance in causative genes was observed in 5 diagnoses made in the study (patients E2, C1, C9, N22, and N25; fig. S16C, D). Due to hemizygosity on the X chromosome in 6 Duchenne muscular dystrophy males and the protein-level effect of pathogenicity in the 4 patients harboring the *COL6A1* intron inclusion, we would not have expected ASE in 10 diagnoses made in the study.

Variant calling from RNA-seq data

Variant calling was performed using GATK HaplotypeCaller following best practices guidelines (https://www.broadinstitute.org/gatk/guide/article?id=3891). BAMs aligned with STAR were processed for genotyping using GATK SplitNCigar reads, and variant calling was performed using HaplotypeCaller for each sample. The resulting VCFs were merged with GATK MergeVCFs and annotated with VEP v81[39]. The variant call set was uploaded onto *seqr* analysis platform (seqr.broadinstitute.org), and analysis was performed using the various inheritance patterns, functional annotation, and variant frequency in reference databases including ExAC[40] and 1000 Genomes[41].

Expression outlier analysis

We identified samples that are outliers for gene expression by calculating z-scores derived from $\log_2$(RPKM+1) for gene expression values obtained from RNA-seQC. We identified gene expression outliers with a z-score cutoff of 3, defining

samples that were both under and over expression outliers. This resulted in the identification of a median of 207, 37, and 2 genes per sample in all genes, OMIM genes[42], and neuromuscular disease genes, respectively. This method resulted in the identification of 1 of 12 causative genes as an expression outlier for samples previously diagnosed by DNA sequencing. This method also identified only 3/6 Duchenne or Becker's muscular dystrophy patients as expression outliers for *DMD,* suggesting expression outlier status analysis was underpowered in our study.

Identification of pathogenic variants in triplicate repeat regions

The triplicated regions of *NEB* and *TTN* (chr2:179517931-179528342 and chr2:179,517,939-179,528,317, respectively) contain repeats with high sequence similarity, resulting in low mapping quality scores and low-quality variant[43,44]. In order to improve variant detection in these regions, we first constructed pseudo-mini-references by masking the hg19 genome except for a given triplicate region using BEDtools[45] (v2.16.1) maskfasta (Figure 2.3A). We extracted reads mapped to the region from exome, split RNA-seq, and whole genome BAMs where available, using SAMtools[46] v1.3 and BEDTools bamtofastq. We then re-mapped the reads aligning to the full triplicate region to the masked reference containing only one triplicate component using BWA-mem[47] v.0.7.12 (Figure 2.3B) Variant calling was performed on the resulting BAMs using GATK HaplotypeCaller with ploidy 6 to account for reads originally aligned to 3 genomic regions being realigned to a single region. VCFs were then annotated with VEP v81[39].

We searched the resulting annotated VCFs for putative loss-of-function variants including nonsense, frameshift, essential splice, and extended splice site variants. For the two variants identified using this method, we performed manual evaluation of read data to ensure allele balance was in line with ploidy 6. For patient N25, 3,133/41,618 and 13/89 reads supported the nonsense variant in RNA-seq and WGS data, respectively (Figure 2.3C). For patient N22, 74/470 reads supported the presence of the frameshift variant (Figure 2.3D). Three references were constructed for each gene, masking two triplicate regions at a time to ensure putatively pathogenic variants were detected in all three cases. We performed remapping of all data types available for patients to confirm any putatively pathogenic variants were detected in all datasets. Both variants identified via this method were confirmed via genotyping. Code for remapping the triplicate region and a masked reference for the first triplicate region of *NEB* can be found at https://github.com/berylc/MendelianRNA-seq.

Splice site prediction

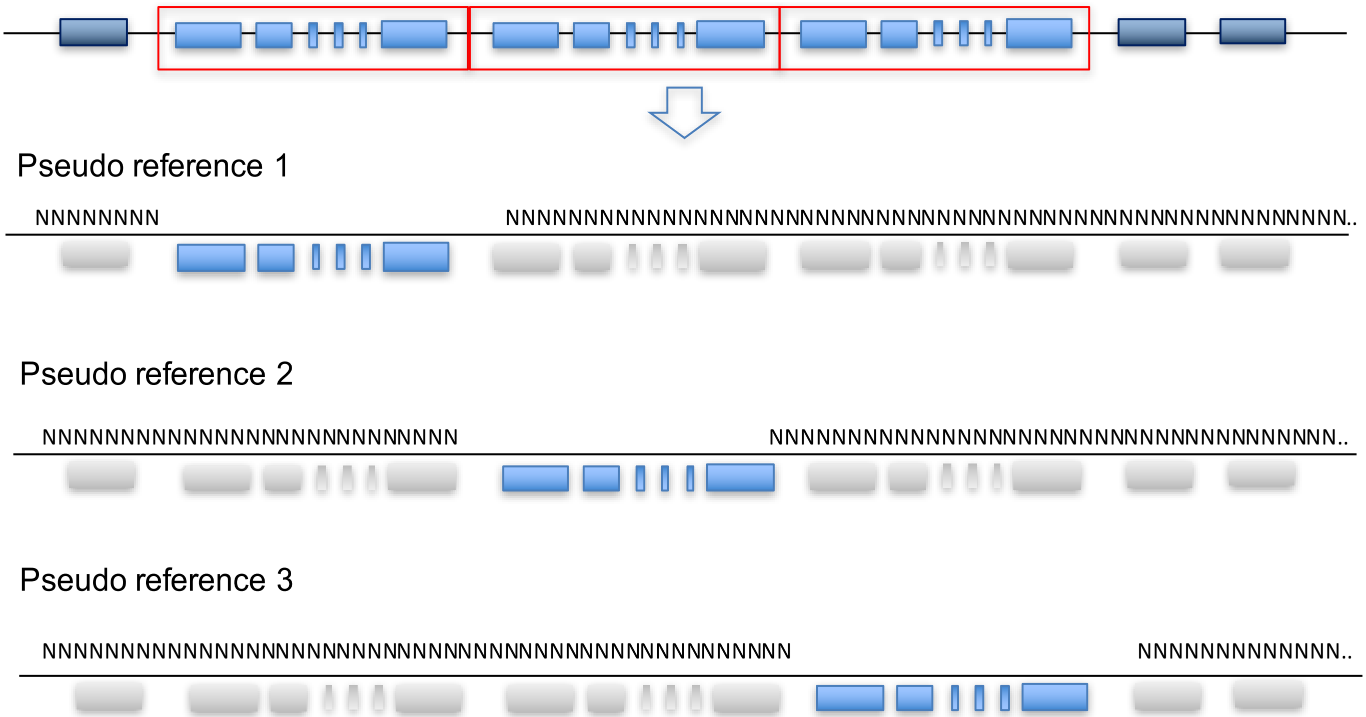
In silico splice site predictions were obtained using the Alamut Batch Software v.1.4.4 available from Interactive Biosoftware (http://www.interactive-biosoftware.com). MaxEntScan (MES)[48], GeneSplicer (GS)[49], Splice Site Prediction by Neural Network (NNSplice)[50], and Human Splicing Finder (HSF)[51] are integrated into this commercial software and were all evaluated for prediction of discovered splice affecting variants. While these tools predict the impact of splice-disrupting variants, Alamut also reports variants with predicted splice-gain effects in 6 categories: i. cryptic donor weakly activated; ii. cryptic donor strongly activated; iii. cryptic acceptor weakly activated; iv. cryptic acceptor strongly activated; v. new donor site; and vi. new acceptor site.

We ran Alamut predictions on 6 patients for whom extended splice site variants were evaluated or novel splice-creating variants were discovered with RNA-seq and where WES data were available (patients E1-3, C9, C11, and N22). The analysis was restricted to WES data as Alamut software was not amenable to parallel computation on the Linux RedHat version 4.4.7-17 cluster utilized at the Broad Institute with a Univa Grid Engine for Research (UGER) v.8.4.0 queuing system: Attempts to run the software on a WGS VCF

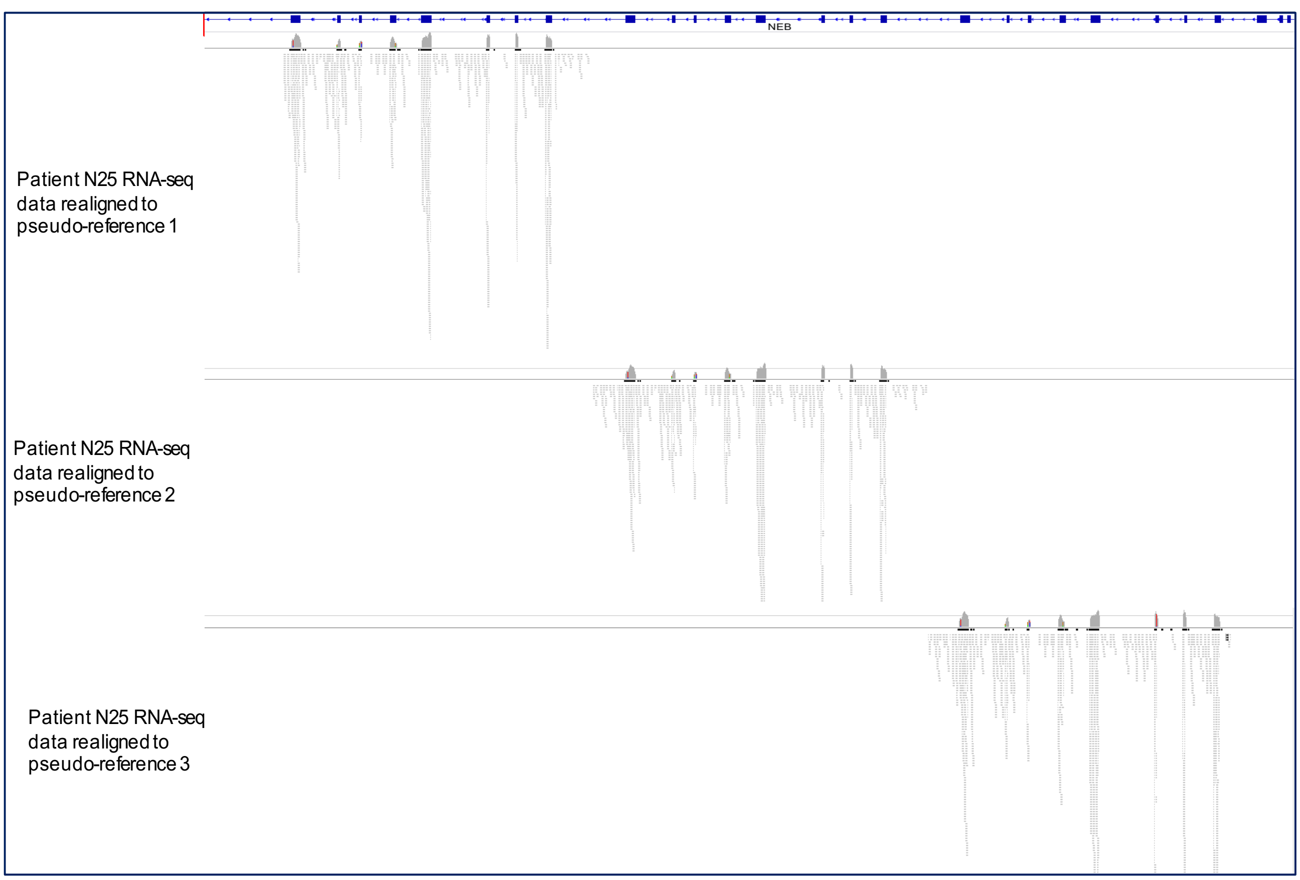
Figure 2.3: Overview of triplicate region remapping. A. Schematic of references built for remapping the triplicate region. Reads that have aligned with low quality are extracted and remapped to a pseudo-reference, and variant calling is performed with ploidy 6. This is performed for all three regions to ensure a given variant is called in all three remapping instances. **C.** IGV screenshot showing reads with mapping quality > 20 demonstrates that reads re-mapped to each of the three pseudo-references have high quality alignment. **D.** Nonsense variant identified in the NEB triplicate region of patient N25. While raw whole genome-sequencing reads aligning to the region do not show evidence of the variant, both re-mapped WGS and RNA-seq data show support for the variant. E) Frameshift variant in the TTN triplicate region in patient N22. Although RNA-seq data showed low coverage of the region, the variant was validated by genotyping.

(Figure 2.3 continued)

A.

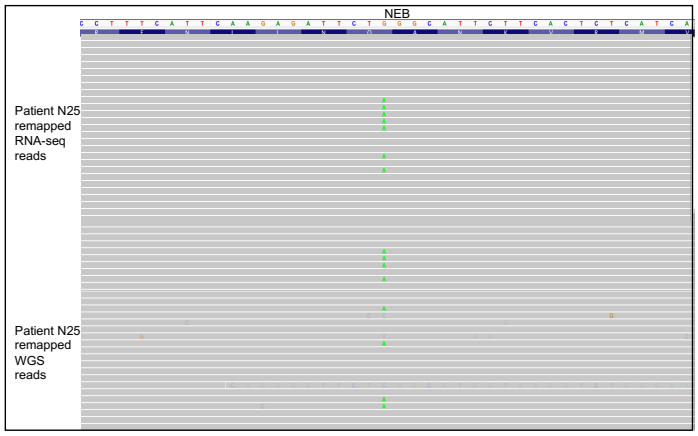


B.

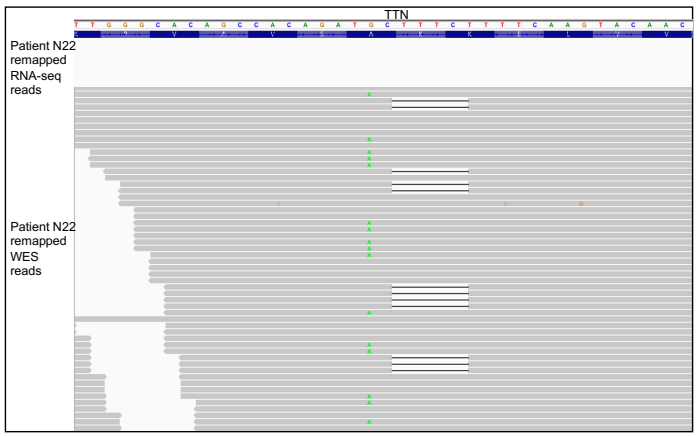


(Figure 2.3 continued)

C.



D.



were abandoned after over 2 months of computation had completed analysis on a single patient VCF only to chromosome 7.

We evaluated the strength of the canonical splice site with and without variants, defining the impact of a variant as (WT_{score} - $Variant_{score}$). We then identified the number of rare, high quality variants (VQSR filter PASS, GQ >10, global allele frequency < 1% in ExAC) that would be predicted to be as or more damaging than the evaluated extended splice site variant. In addition, we evaluated the total number of rare, high quality variants reported to result in gain of splicing on an exome-wide scale. We did not evaluate *in silico* predictions of variants that affect splice factor binding motifs (such as those discovered in patients C1 and N25 in the study) as it is estimated that ~75% of typical mRNA are spanned by at least one splice motif, indicating that tools querying the effect of variants on these motifs often lack specificity[52,53].

For two disruptive extended splice sites identified in the study (in patients E2 and C9), several *in silico* predictions showed a low number of exome-wide variants with a

score at or above the score of the variant, indicating that *in silico* predictions successfully prioritized these disruptive variants with high specificity. However, two extended splice site variants shown to have no effect on local RNA splicing (patients E1 and E3) also showed the same low number of variants predicted to be damaging for each tool tested, suggesting that use of these predictions alone could result in assignment of false pathogenicity. Splice prediction tools also showed poor specificity in identifying splice site-creating variants, predicting an average of ~140 rare splice-creating variants per exome. The low specificity for splice site-creating variants and the false positives observed for splice-disrupting variants show that while *in silico* splice predictions tools can be useful to prioritize variants for follow up analysis, they are currently insufficient to designate variants as causal for genetic diagnosis based on DNA information alone.

RT-PCR validation and Sanger sequencing of cDNA

The SuperScript® III First Strand Synthesis Kit (ThermoFisher Scientific, 18080051) was used to make cDNA from 50ng of RNA according to kit instructions. The Herculase II Fusion DNA polymerase (Agilent, 600679) was used for PCRs. Control cDNA was obtained from RNA extracted from muscle biopsies of other patients with splicing effects in unrelated genes. All PCR products were analyzed on a 2% agarose gel unless otherwise indicated.

Sequence determination of the cDNA boundaries of the intron inclusion for C7 was performed by PCR purifying the bands from the PCR of exon 55- intron inclusion and exon 56- intron inclusion. A second identical PCR reaction was performed followed by PCR

purification on the PCR product to amplify it enough for Sanger sequencing with the Exon 55R and Exon 56F primers. Sequence verification of the boundaries exonic splicing for C11 was performed by gel extraction and purification of the two bands from a 4% agarose gel (Qiagen, QIAEX II Gel Extraction Kit, 20021), a second PCR to further amplify the PCR product, PCR purification (Qiagen), and sequenced with the forward and reverse PCR primers.

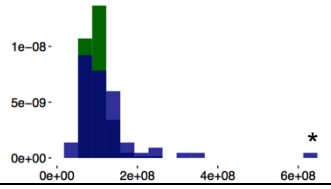
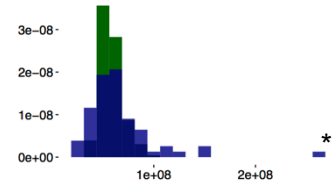
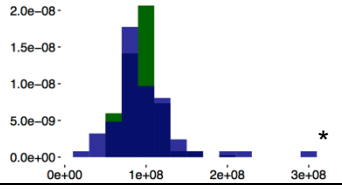
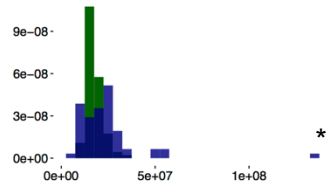
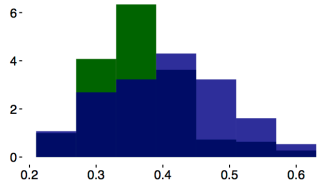
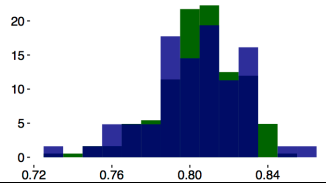
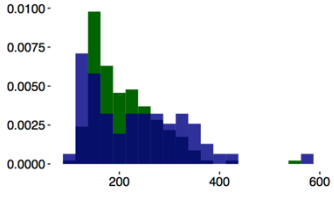
Patient C1 had patterns of clear nonsense-mediated decay with only the haplotype harboring the essential splice site variant being detectable in the RNA-seq data. However, it was detectable in patient fibroblasts transdifferentiated into skeletal myotubes via MyoD overexpression by RNA-seq and RT-PCR. The RT-PCR was designed to amplify between exons 6-8 of *POMGNT1*. Sequence confirmation of the new exon junction was performed by gel purification of the shorter band in C1 fibroblasts trans-differentiated into skeletal myotubes via MyoD overexpression from a 4% agarose gel (Qiagen, QIAEX II Gel Extraction Kit, 20021), a second PCR to further amplify the PCR product, PCR purification (Qiagen), and sequenced with the forward and reverse primers.

Results

Comparison of patient RNA-seq to a muscle RNA-seq reference panel

Patient muscle samples were sequenced using the same protocol as in the GTEx project[28] and analyzed using identical pipelines to minimize technical differences, with patients sequenced at or above the same coverage as GTEx controls. From 430 skeletal

Table 2.1 Comparison of quality metrics between patient and GTEx RNA-sequencing samples shows correspondence between patients and controls. The outlier (*) observed in the first four metrics corresponds to patient N33, who was sequenced at higher depth of 500M reads and thus has a higher number of mapped reads. P-values are based on a t-test.

Quality metric	Distribution	GTEx mean (n=184)	Patient mean (n=64)	p-value
Total purity filtered reads sequenced		9.81 x 10^7	12.2 x 10^7	0.041
Number of uniquely mapped reads		5.8 x 10^7	6.6 x 10^7	0.085
Estimated library size		9.68 x 10^7	9.55 x 10^7	0.597
Number of split reads		1.77 x 10^7	2.39 x 10^7	0.008
Proportion of duplicate reads		0.37	0.41	0.002
Exonic rate		0.81	0.80	0.671
Mean fragment length		209.5	232.4	0.0897

muscle RNA-seq samples available through GTEx, we selected a subset of 184 samples based on RNA-seq quality metrics including RNA integrity (RIN) score and ischemic time, as well as phenotypic features such as age, BMI, and cause of death to more closely match our patient samples.

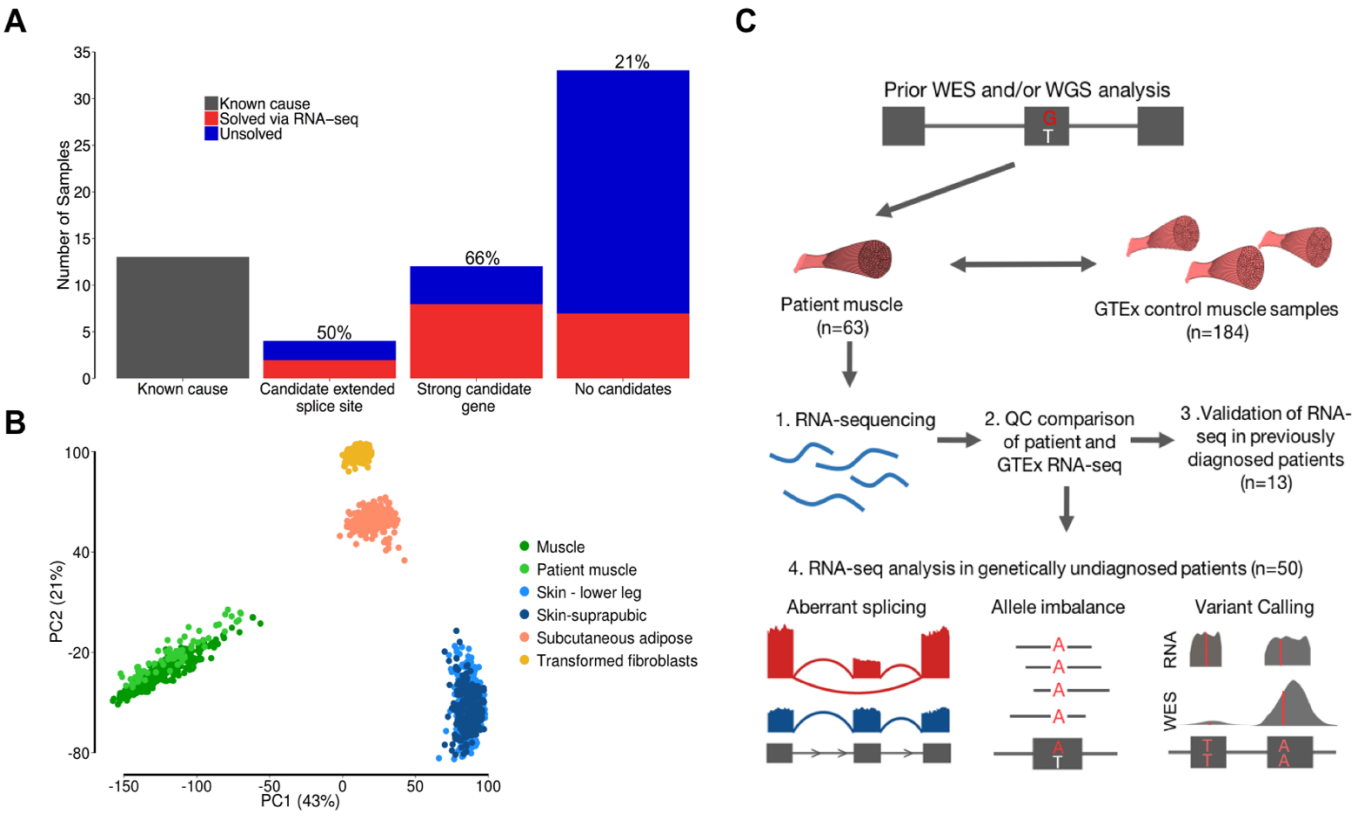
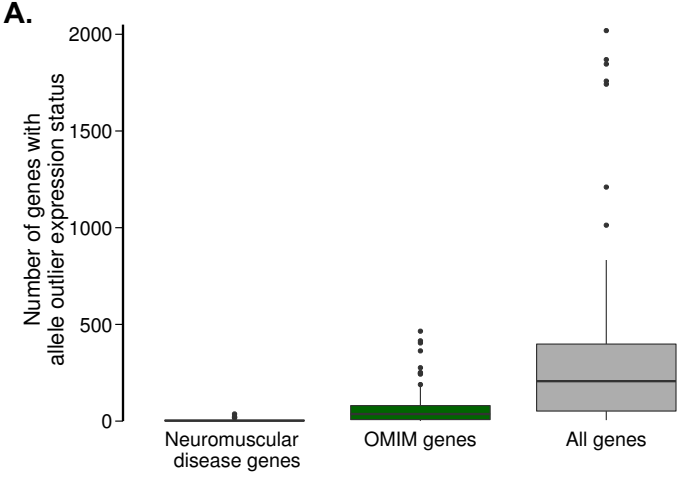


Figure 2.4 Experimental design and quality control. A. Overview of the number of samples that underwent RNA-seq. We performed RNA-seq on 13 previously genetically diagnosed patients, 4 patients in whom previous genetic analysis had identified an extended splice site variant of unknown significance (VUS), 12 patients in whom genetic analysis had identified a strong candidate gene, and 34 patients with no strong candidates from previous analysis. RNA-seq enabled the diagnosis of 35% of patients overall, with the rate, shown above the barplots, varying depending on previous evidence from genetic analysis. **B.** PCA based on gene expression profiles of patient muscle samples passing QC (n=61) and GTEx samples of tissues that potentially contaminate muscle biopsies shows that patient samples cluster closely with GTEx skeletal muscle. **C.** Overview of experimental set up and RNA-seq analyses performed. Our framework is based on identifying transcriptional aberrations that are present in patients and missing in GTEx controls. Upon ensuring that GTEx and patient RNA-seq data were comparable, we validated the capacity of RNA-seq to resolve transcriptional aberrations in previously diagnosed patients and performed analyses of aberrant splicing, allele imbalance, and variant calling in our remaining cohort of genetically undiagnosed muscle disease patients.

Comparison between our GTEx reference panel and patient muscle RNA-seq samples showed analogous quality metrics (Table 2.1). Principal component analysis (PCA) of expression and splicing profiles demonstrated that patient muscle RNA-seq closely resembled control muscle when compared to tissues that potentially contaminate muscle biopsies, such as skin or fat, despite variation in the site of muscle biopsy across patients (Figure 2.4B). Based on this clustering, we removed two samples from analysis because their expression patterns clustered more closely with GTEx adipose tissue than muscle, consistent with tissue contamination or late-stage degenerative muscle pathology. We also performed fingerprinting based on patient WES, WGS, and RNA-seq data to ensure the source of DNA sequencing and muscle RNA-seq data was the same individual.

A.



B.

Patient	Gene	Inheritance	Variant Class	Causative gene is an expression outlier
D1	*LARGE*	Recessive	missense deletion	No
D2	*PYROXD1*	Recessive	essential splice missense	No
D5	*TTN*	Recessive	essential splice frameshift	No
D7	*NEB*	Recessive	essential splice frameshift	No
D8	*NEB*	Recessive	essential splice essential splice	No
D9	*RYR1*	Recessive	nonsense essential splice	Yes
D10	*TTN*	Recessive	essential splice essential splice	No
D11	*COL6A3*	De novo dominant	essential splice	No
D3	*DYSF*	Recessive	missense frameshift	No
D4	*TOR1AIP1*	Recessive	frameshift missense	No
D12	*PLOD1*	Recessive	nonsense missense	No
D13	*TTN*	Recessive	nonsense nonsense	No

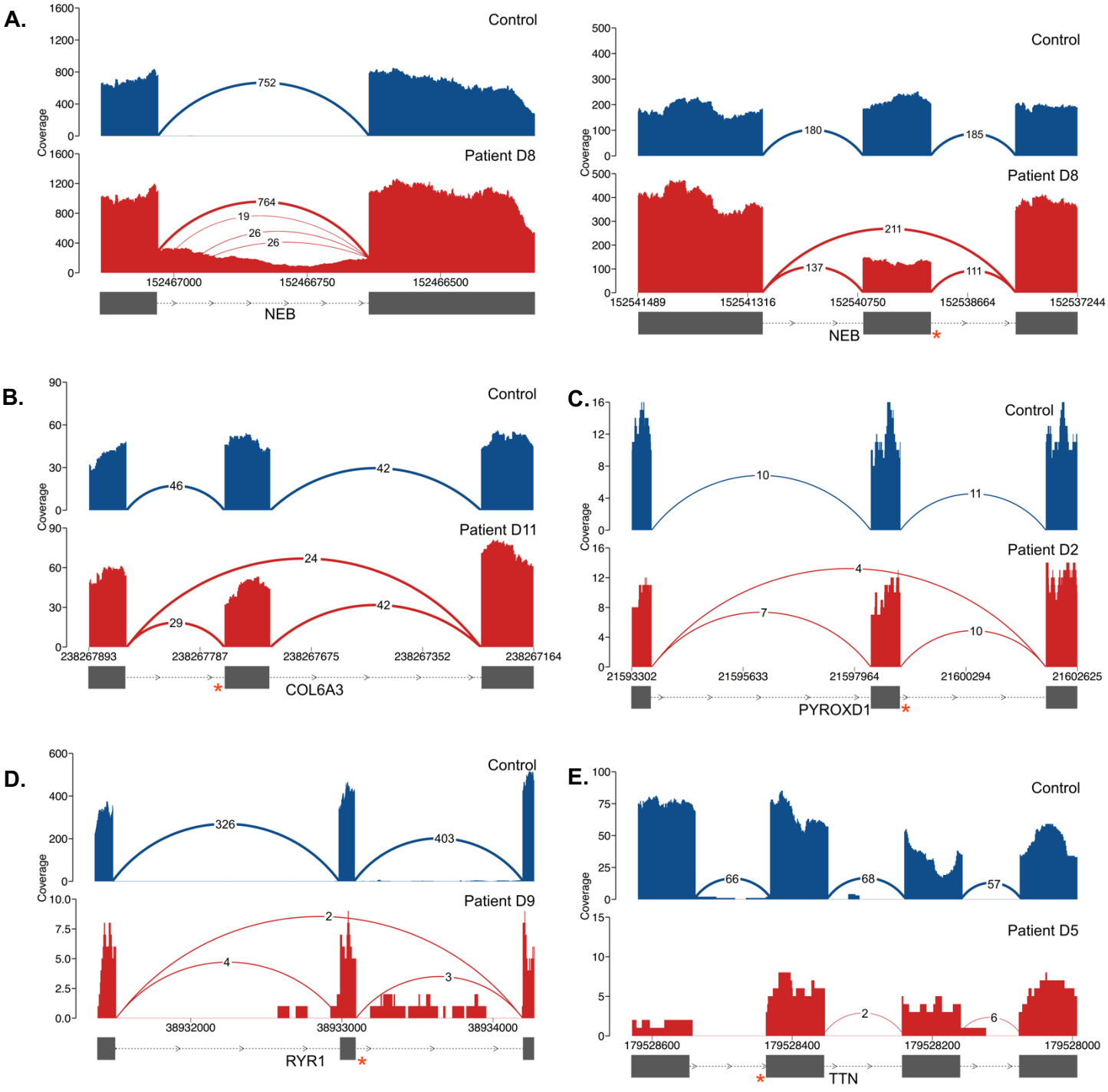
Figure 2.5 Overview of results from expression outlier analysis. **A.** The method used to identify samples that are outliers for gene expression resulted in a median of 207, 37, and 2 genes per sample in all genes, OMIM genes, and neuromuscular disease genes, respectively. **B.** For patients who were previously diagnosed via DNA sequencing, only one causative gene had expression outlier status, suggesting that our method to detect expression outlier genes was underpowered with 184 GTEx controls.

We explored the utility of analyzing patient RNA-seq data to detect aberrant splice events and allele-specific expression and performed variant calling from RNA-seq data to identify pathogenic events or to prioritize genes for closer analysis (Figure 2.4C). We also identified outlier gene expression status in patients; however, this analysis was underpowered to prioritize candidate genes in our study (Figure 2.5). The resulting diagnoses were made primarily through detection of aberrant splice events in patients, with information on gene-level allele imbalance playing a complementary role

In previously diagnosed cases, manual evaluation of pathogenic essential splice site variants revealed a splice aberration such as exon skipping or extension, demonstrating that RNA-seq can help resolve the effect of variants on transcription (Figure 2.6). To detect aberrant transcriptional events genome-wide, we developed an approach based on identifying high quality exon-exon splice junctions present in patients or groups of patients and missing in GTEx controls (code available at https://github.com/berylc/MendelianRNA-seq). We performed splice junction discovery from split-mapped reads, considering only those that were uniquely aligned and non-duplicate. To account for library size and stochastic gene expression differences between samples, we performed local normalization of read counts based on read support for overlapping annotated junctions (Figure 2.2). We then performed filtering of splice junctions based on the number of samples in which a splice junction is observed and the number of reads and normalized value supporting that junction in each sample. Our approach successfully re-identified all known pathogenic events in patients in whom manual evaluation had revealed aberrant splicing around splice variants previously identified through genomic testing. We defined filtering parameters that selectively

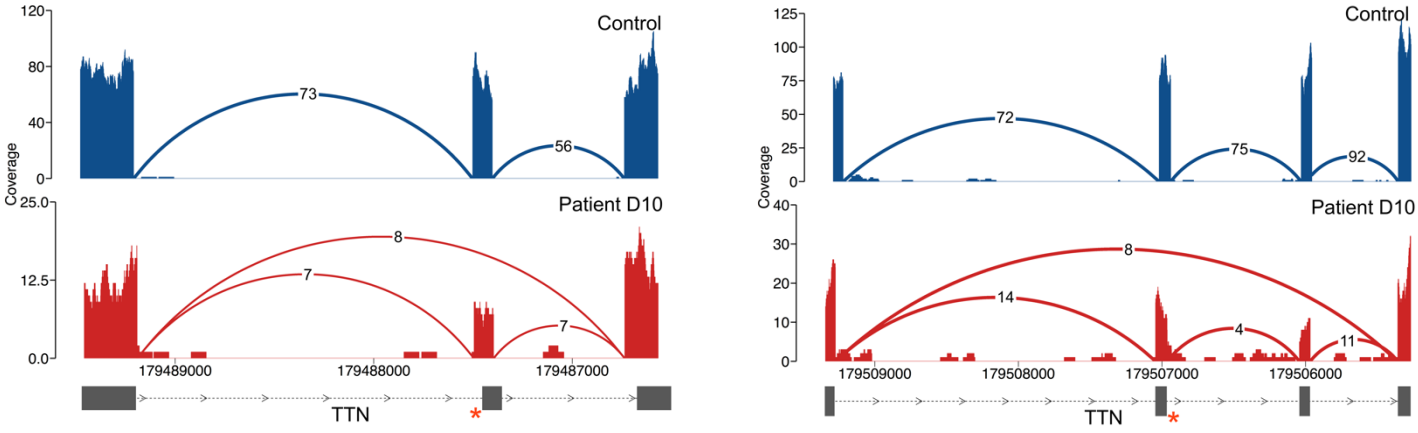
Figure 2.6 Evaluation of RNA-seq around pathogenic essential splice site variants previously identified via genetic analysis. All essential splice site variants are marked with red asterisks. **A.** intron inclusion and splicing from adjacent intact GT splice sites around one essential splice site variant (left) and exon skipping in a second essential splice found in *NEB* (right) in patient D8. **B.** Exon skipping caused by a *de novo* essential splice site variant found in *COL6A3* in patient D11. **C)** Exon skipping caused by an essential splice site variant in *PYROXD1* in patient D2. **D.** Exon skipping caused by an essential splice site variant in *RYR1* in patient D9. **E.** No effect seen around essential splice site variant in patient D5 due to decreased expression of *TTN* in the patient and insufficient coverage of the region **F.** Two exon skipping events caused by separate essential splice site variants in *TTN* in patient D10.

(Figure 2.6 continued)



(figure 2.6 continued)

F.



identified these previously known aberrant splice events and applied them to our remaining cohort of undiagnosed patients. This method resulted in the identification of a median of 5, 26, and 190 potentially pathogenic splice events per sample in ~190 neuromuscular disease associated genes, OMIM genes, and all genes respectively (Figure 2.7), which required manual curation to interpret pathogenicity and led to the diagnoses made in this study.

Overview of diagnoses made via RNA-seq

RNA-seq allowed the diagnosis of 17 previously unsolved families, yielding an overall diagnosis rate of 35% in this challenging subset of rare disease patients for whom extensive prior analysis of DNA sequencing data had failed to return a genetic diagnosis. We also identified splice disruption in other known and putatively novel disease genes in

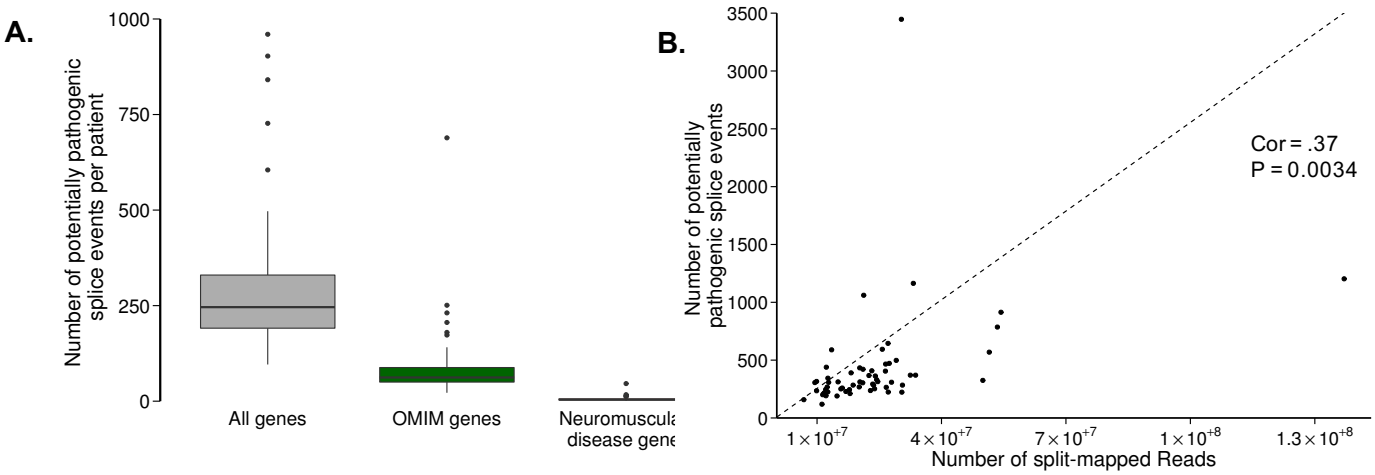


Figure 2.7 Number of potentially pathogenic splice events identified per patient. A) Median number of events identified via splice detection method was 105, 26, and 5 for all genes, OMIM genes and neuromuscular disease genes, respectively, allowing for manual inspection of RNA-seq data for a relatively low number of events to identify pathogenic splice aberrations. **B)** The number of identified potentially pathogenic splice junctions shows correlation with library size. This is likely explained by more splice junctions being identified in samples with higher coverage (>100 million reads) that are more deeply sequenced than GTEx controls. This includes low abundance annotated transcripts that are picked up by deeper sequencing and not identified in samples with lower sequencing depth.

several patients; however, due to unavailability of additional information, such as parental DNA, we could not pursue these cases further. Detection of aberrant splicing led to the identification of a broad class of both coding and non-coding pathogenic variants resulting in a range of splice defects such as exon skipping, exon extension, exonic and intronic splice gain, which were validated by RT-PCR analysis (Figure 2.8, Table 2.2). RNA-seq patterns also helped pinpoint three structural variants in *DMD* that were subsequently confirmed by WGS.

Cases diagnosed in this study highlight several key advantages of RNA-seq in rare disease diagnosis to confirm the pathogenicity of variants and to detect previously unidentified variation. In four patients with previously detected extended splice site variants of unknown significance (VUS), RNA-seq confirmed splice disruption in two patients (Figure 2.4A). The variants had no observable effect on local splicing patterns

in the remaining two patients, emphasizing the value of RNA-seq in ruling out non-pathogenic VUS.

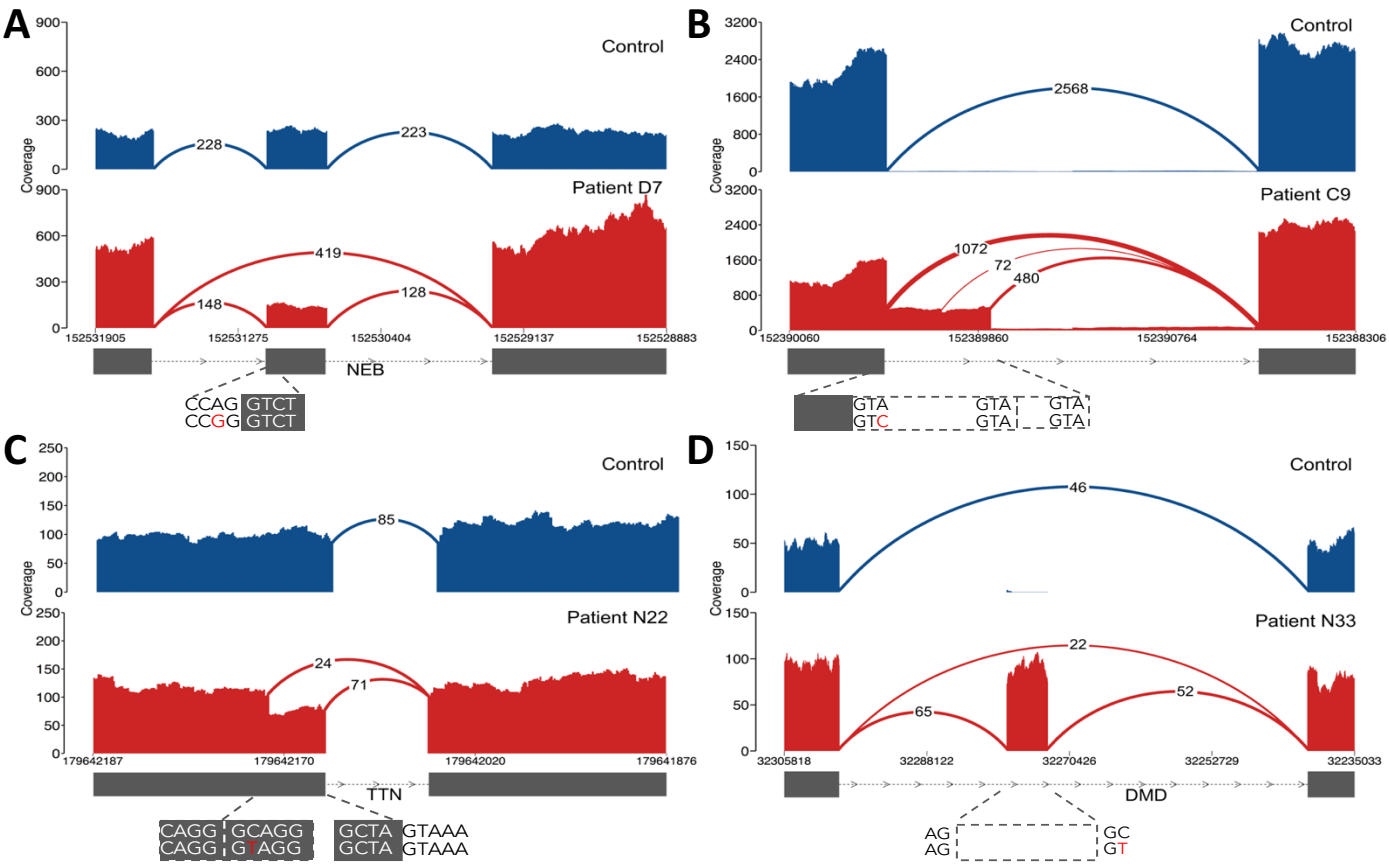


Figure 2.8 Types of pathogenic splice aberrations discovered in patients. RNA-seq identified a range of aberrations caused by both coding and non-coding variants such as **A)** exon skipping caused by an essential splice site variant in patient D7, **B)** exon extension caused by a donor +3 A>C extended splice site variant in nemaline myopathy patient C9, where disruption of splicing at the canonical splice site results in splicing from intact GTA motifs from the intron, **C)** exonic splice-gain caused by a C>T donor splice site-creating variant in patient N22 with a donor + 5-G sequence context, resulting in a stronger splice motif than the existing canonical splice site, and **D)** intronic splice gain in patient N33 caused by a C>T donor splice site-creating deep intronic variant**.** Evidence for wild type splicing in addition to the inclusion of the pseudo exon in the patient is in line with the milder Becker's muscular dystrophy phenotype. Splice aberrations shown in B, C, and D result in the introduction of a premature stop codon to the transcript.

RNA-seq also led to the identification of an additional disruptive extended splice site variant missed by exome sequencing. In a nemaline myopathy patient with one previously detected recessive frameshift variant in the *NEB* gene, RNA-seq identified an exon

Table 2.2: Diagnoses made in the study via patient muscle RNA-seq

Patient	Phenotype	Gene	Variants	Variant Class	Effect
E2	Nemaline myopathy	*NEB*	chr2:152,544,805 C>T chr2:152,520,057 C>T	essential splice, extended splice	exon skipping + exon extension, exon extension
C9	Nemaline myopathy	*NEB*	chr2:152,581,432 TG>T chr2:152,389,953 A>C	frameshift, extended splice	exon extension
E4	Fetal akinesia	*TTN*	chr2:179,586,600 CAT>C chr2:179,446,219 ATACT>A	frameshift, extended splice	exon skipping
C6	Duchenne muscular dystrophy	*DMD*	chrX:32,366,860 A>C	intronic variant	intronic splice-gain
N33	Myalgia, myoglobunuria	*DMD*	chrX:32,274,692 G>A	intronic variant	intronic splice-gain
C7	Becker muscular dystrophy	*DMD*	chrX:31,613,687 G>T	intronic variant	Intronic splice-gain
N29	Collagen VI-related dystrophy	*COL6A1*	chr21:47,409,881 C>T	intronic variant	intronic splice-gain
N30	Collagen VI-related dystrophy	*COL6A1*	chr21:47,409,881 C>T	intronic variant	intronic splice-gain
N31	Collagen VI-related dystrophy	*COL6A1*	chr21:47,409,881 C>T	intronic variant	intronic splice-gain
N32	Collagen VI-related dystrophy	*COL6A1*	chr21:47,409,881 C>T	intronic variant	intronic splice-gain
N25	Nemaline myopathy	*NEB*	chr2:152,355,017 G>T chr2:152,449,646G>A	intronic variant, nonsense	intronic splice-gain
C11	Congenital fiber-type disproportion	*RYR1*	chr19:38,958,362 C>T chr19:38,958,372 G>A	synonymous, missense	exonic splice gain
N22	Multi/minicore congenital myopathy	*TTN*	chr2:179,642,185 G>A chr2:179,523,240 CTTCT>C	missense, frameshift	exonic splice-gain
C1	Alpha dystroglycanopathy	*POMGNT1*	chr1:46,655,129 C>A chr1:46,660,532 G>A	essential splice, synonymous	exonic splice-gain, exon skipping
C3	Duchenne muscular dystrophy	*DMD*	chrX:31,790,694-31,798,498	inversion-deletion	exon skipping
C2	Duchenne muscular dystrophy	*DMD*	chrX:31,378,946-151,194,962	inversion	splice disruption
C4	Duchenne muscular dystrophy	*DMD*	chrX:32,521,820-35,180,380	inversion	splice disruption

extension event caused by an underlying variant at the +3 position of the donor site which led to the introduction of a premature stop codon to the transcript as the second recessive allele (Figure 2.8B). The exon harboring this variant was not captured in the exome kit used to screen the patient, underlining the utility of RNA-seq at complementing WES to identify previously undetected variants.

Synonymous and missense variants in large, variation-rich genes such as *TTN* are exceptionally challenging to interpret and are often filtered out in DNA sequencing pipelines[54,55]. With RNA-seq, we were able to assign pathogenicity to a missense variant in *TTN* and two synonymous variants in *RYR1* and *POMGNT1.* In patient N22, the identified missense variant created a GT donor splice site for which the consensus motif included a G nucleotide in the +5 position, known to contribute to the strength of the splice site[56–59]. The well-conserved donor +5-G motif was missing in the competing canonical splice site, thus resulting in a stronger novel splice site and gain of splicing from the exon body (Figure 2.8C). A similar mechanism was observed in *RYR1*, caused by a synonymous variant in a patient carrying a second pathogenic allele in the gene. In an additional patient carrying an essential splice site variant in *POMGNT1*, we identified a synonymous variant disrupting an exonic splice motif and resulting in exon skipping.

In eight cases, RNA-seq aided in the identification of non-coding pathogenic variants. We identified splice site-creating hemizygous deep intronic variants in *DMD* that resulted in the creation of a pseudo-exon and led to a premature stop codon in the coding sequence in three patients. Although RNA-seq from a patient with severe Duchenne muscular dystrophy showed only splicing to the pseudo-exon, wildtype splicing between annotated exons was observed in two patients with a milder Becker

muscular dystrophy phenotype, indicating the presence of residual functional *DMD* transcripts that explain the milder disease course (see below, under Identification of pathogenic noncoding variants with RNA-seq section). Such intronic variants are unobservable with WES and too abundant to be interpretable with WGS alone, emphasizing the utility of RNA-seq at resolving pathogenicity of these non-coding variants

In two patients with no strong candidates from WES and WGS (N22 and N25) we identified heterozygous splice disruption in two commonly disrupted recessive muscle disease genes, *NEB* and *TTN* [43,44]. These genes harbor regions with highly similar sequences, the so-called triplicate repeat regions. Due to high sequence similarity, the region has poor mapping quality, resulting in low quality variant calls that are filtered by most current diagnostic pipelines. To identify possible pathogenic variants in the triplicated regions of *NEB* and *TTN* in these two patients, we developed a method based on remapping the triplicate regions to a de-triplicated pseudo-reference and performing hexaploid variant calling (Figure 2.3). This method was applied to available WES/WGS and RNA-seq data for all patients and identified one novel nonsense and one novel frameshift variant in *NEB* and *TTN* in these two patients, which finalized their diagnoses (N25 and N22).

Resolving the effect of extended splice site variants with RNA-seq

In four patients, prior genetic analysis identified extended splice site variants of unknown significance (patients E1-E4). In two patients, RNA-seq supported pathogenicity for the variants whereas for the remaining two, RNA-seq showed no

aberrant splicing caused by the variants, ruling them out as pathogenic. We also identified an additional disruptive extended splice site variant in patient C9 (Figure 2.8B) that was previously missed by WES.

In patient E2, exome analysis identified an essential and extended splice site variant *in trans* in *NEB*, confirmed by segregation analysis. There were 2 individuals in ExAC carrying the extended splice site variant and none carrying the essential splice site variant (ExAC Allele Count = 2 and 0, respectively). RNA-seq showed exon skipping and extension caused by the essential splice site variant and intron inclusion around the donor +5 G>A extended splice site variant, leading to a premature stop codon (Figure 2.9A)

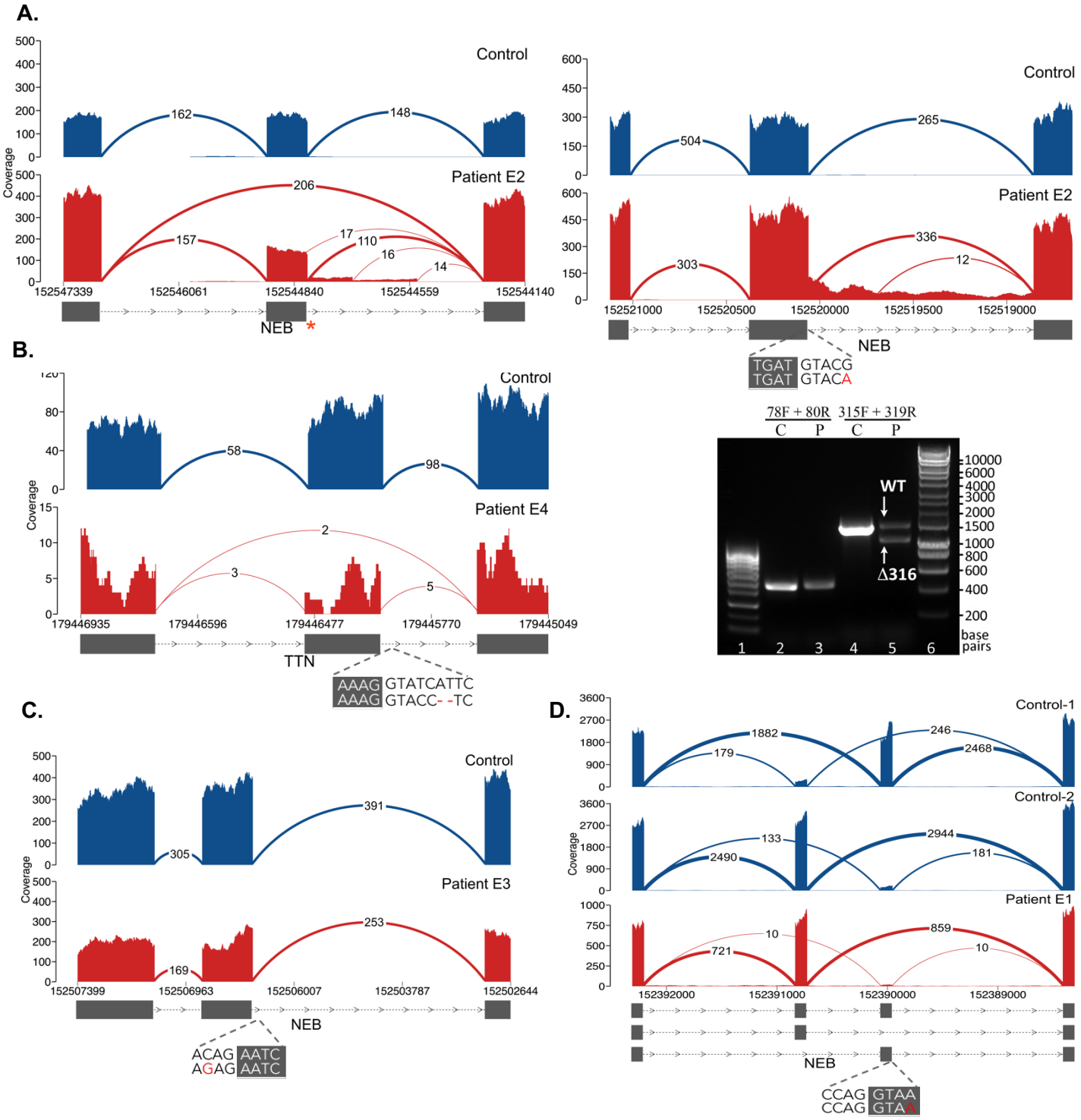


Figure 2.9 Resolving the effect of extended splice site variants with RNA-seq. A) Exon skipping and splicing from adjacent intact GT splice sites in patient E2 caused by an essential splice variant indicated by red asterisk (left). Intron inclusion and low levels of splicing from intact intronic splice site motifs around the extended splice site variant (right). **B)** Skipping of exon 316 caused by an extended splice variant in *TTN* in Patient E4 (left). Amplification of cDNA using primers between exons 315 and 319 (lanes 4 and 5) identified two amplicons: the upper band was wild-type (WT) sequence, with the lower band confirming skipping of exon 316 (right). **C)** No splicing defects were observed around the extended splice site variant identified in patient E3. **D)** No evidence of aberrant splicing around the extended splice site variant in patient E1. The splicing pattern in the patient suggests possible isoform switching, however, the same splicing pattern is observed in other patients who do not carry the extended splice site variant (Control-2 represents patient N22).

In patient E4, gene panel testing identified a frameshift and extended splice site variant *in trans* in *TTN*, confirmed by segregation analysis (ExAC AC = 0 for both variants). RNA-seq showed evidence for skipping of exon 316 harboring the extended splice site variant, which was subsequently confirmed by RT-PCR (Figure 2.9B).

In patient E3, carrying a nonsense and a donor +3 G>C extended splice site variant in *NEB* (ExAC AC = 0 for both variants), no splicing defects around the extended splice site were observed (Figure 2.9C). We considered the possibility of an aberrant splice event at the position causing complete nonsense-mediated decay, which may result in failure of RNA-seq to pick up any reads supporting the event. We observed a 28% mean allele balance in *NEB* in the patient. Due to unavailability of parental DNA, it was not possible to distinguish between the haplotypes resulting in nonsense-mediated decay. However, accounting for nonsense-mediated decay resulting from a splice aberration would still predict the presence of ~55 reads supporting the splice aberration. Therefore, based on a conservative interpretation taking into account allele balance in *NEB* as well as the local splice patterns of the extended splice site, the variant does not result in a local splice disruption.

In patient E1, trio WES identified a nonsense and extended splice site variant *in trans* in *NEB* (ExAC AC = 0 and 3, respectively). No clear evidence of aberrant splicing around the extended splice site variant was observed. Comparison to GTEx controls showed decreased splicing to the exon harboring the extended splice site variant, suggesting the possible presence of isoform switching. However, this splicing pattern is observed in other patients who do not carry the extended splice, indicating the variant is not causal for the splicing pattern observed (Figure 2.9D). We also considered the

possibility of an aberrant splice event causing complete nonsense-mediated decay and being missed by transcriptome sequencing. However, the 22% allele balance observed in *NEB* in the patient was against the paternal haplotype harboring the nonsense variant, excluding the possibility that a local splice aberration was missed by RNA-seq.

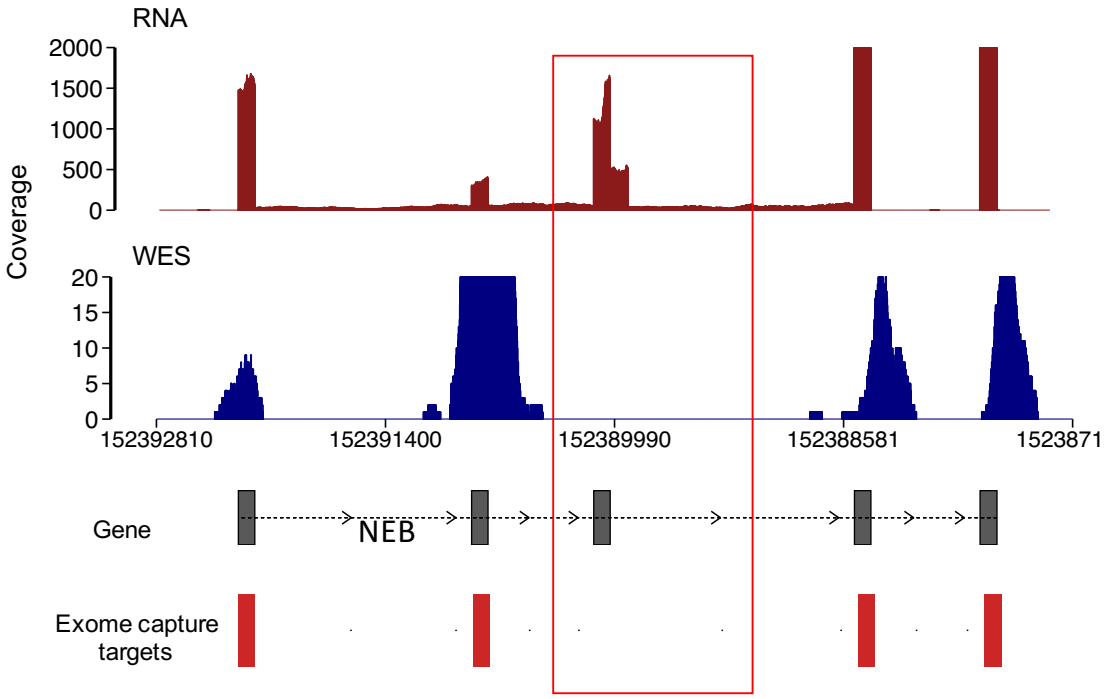


Figure 2.10 Coverage of exon harboring splice-disrupting variant identified in patient C9 in RNA-seq and WES. While previous exome analysis in the patient resulted in the identification of a frameshift variant in NEB, the extended splice site variant was not detected. This is due to the absence of a target for the exon harboring the pathogenic extended splice site variant in the exome capture kit used for WES.

In patient C9, for whom previous WES analysis had identified a recessive frameshift variant in *NEB,* RNA-seq identified a separate exon extension event in the gene (ExAC AC = 0 for both variants). The exon extension led to splicing from intact intronic splice motifs (Figure 2.8B), resulting in the introduction of a premature stop

codon to the transcript. The donor +3 A>C variant visible in the RNA-seq data was missed by WES analysis due to absence of a target probe for the exon in the capture kit utilized for the patient (Figure 2.10). We note that while a probe for the exon is included in more recent exome capture kits, identification of the variant in WES data would result in a VUS designation without further functional validation. Segregation analysis confirmed the variants were found *in trans* in both the proband and an affected sibling.

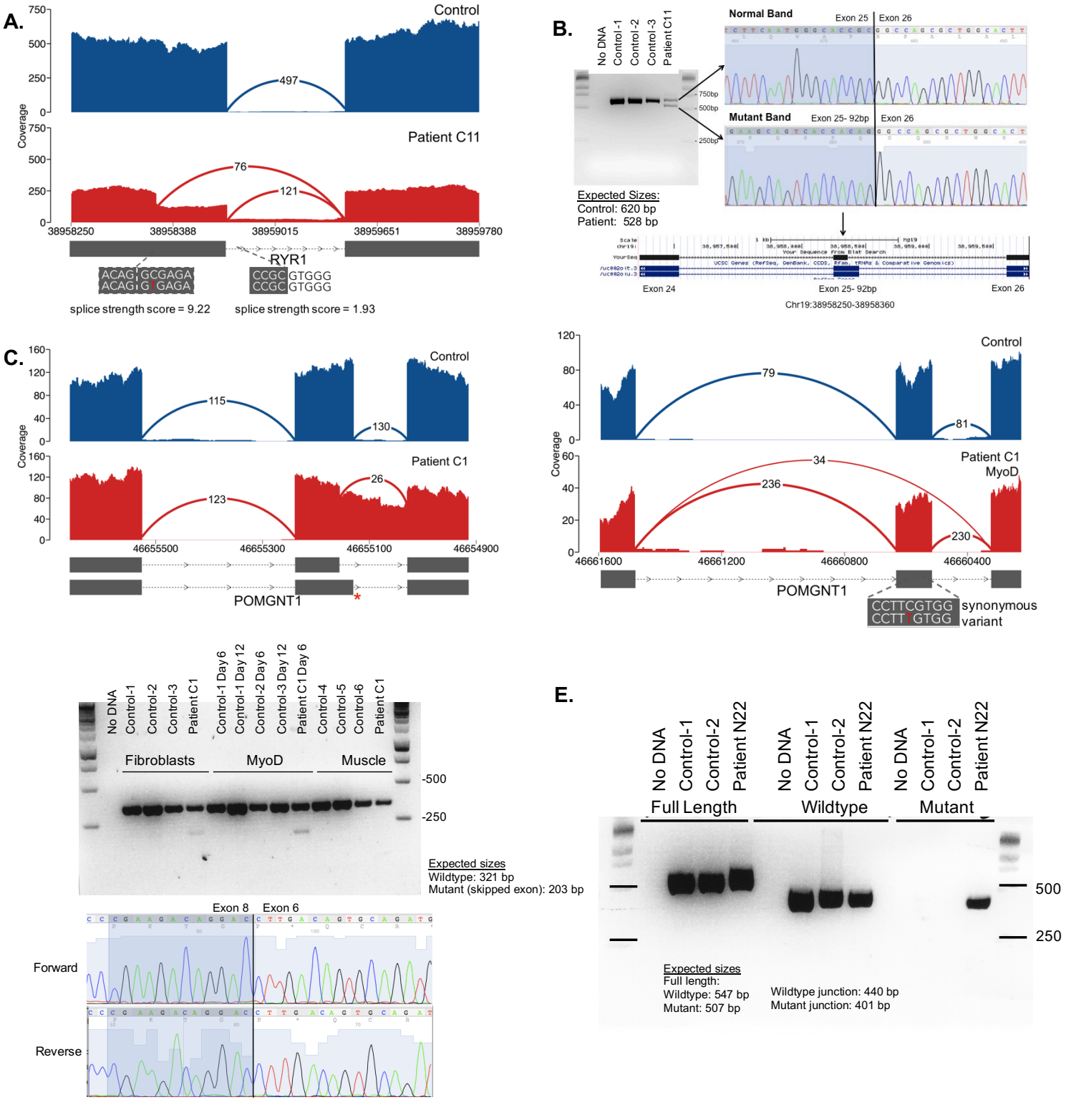
Assignment of pathogenicity to missense and synonymous variants with RNA-seq

In three patients, RNA-seq identified splicing defects caused by missense and synonymous variants (patients C11, C1 and N22).

In patient C11, a pathogenic missense variant in *RYR1* was identified through WES analysis (ExAC AC=2). RNA-seq identified an exonic splice gain event caused by a GC>GT splice site-creating synonymous variant (ExAC AC=0). We evaluated the strength of the competing splice sites with MaxEntScan and found that the newly created splice site had a 5x greater score, indicating a stronger splice site was created (Figure 2.11A). The splice event was confirmed via RT-PCR and sequencing of the shorter band corresponding to the exonic splice gain (Figure 2.11B). Segregation analysis confirmed the variants were found *in trans* in the proband.

Figure 2.11 Assignment of pathogenicity to missense and synonymous variants with RNA-seq. A) Exonic splice gain event in patient C11 caused by a C>T donor splice-site-creating variant. B) Confirmation of the exonic splice gain in patient C11. Controls for RT-PCR are muscle RNA from patients without pathogenic variants in RYR1 (patients N8, N22, and N33). C) Intron inclusion and splicing to a nearby intact splice site in patient C1 caused by an essential splice site variant indicated by red asterisk (left) and exon skipping caused by a synonymous variant (right). D) Confirmation of exon skipping in patient C1 by RT-PCR with a reduced level of nonsense-mediated decay observed in transdifferentiated myotubes (top). Sequencing of the lower band confirmed skipping of exon 7 (bottom). Controls for RT-PCR are muscle RNA from patients without pathogenic variants in POMGNT1 (Patients N8, N22, and N33) and fibroblasts from healthy samples. E) RT-PCR confirmation of exonic splice gain event in patient N22. Controls for RT-PCR are muscle RNA from patients without pathogenic variants in TTN (Patients N8 and N33).

(Figure 2.11 continued)



In patient C1, RNA-seq identified high levels of intron inclusion at an essential splice site variant in *POMGNT1* as well as splicing to a nearby intact splice site (ExAC AC = 13, Figure 2.11C). While the new aberrant splicing is seen on an annotated Gencode v.19 transcript, no other samples from our muscle RNA-seq dataset carry this splice event. We observed complete allele imbalance in the patient indicating the presence of strong nonsense mediated decay in muscle. We thus transdifferentiated fibroblasts available from the patient into skeletal myotubes via MyoD overexpression. RNA-seq from resulting myotubes showed evidence of exon skipping caused by a synonymous variant (ExAC AC = 2, Figure 2.11C)**.** The exon skipping event was confirmed by RT-PCR and sequencing of mutant band corresponding to the exon skipping (fig. S11D). Segregation analysis confirmed the variants were found *in trans* in the proband.

In patient N22, RNA-seq identified an exonic splice gain event in the patient caused by a GT donor splice site-creating missense variant (ExAC AC = 0, Fig. 2C). We also identified a frameshift variant in the *TTN* triplicate repeat region in the patient (Figure 2.3D). RT-PCR, designed to amplify the region with primers spanning the wild-type and mutant exon junctions, confirmed the exonic splice gain event (Figure 2.11D), and segregation analysis confirmed the variants were found *in trans* in the proband.

Identification of pathogenic noncoding variants with RNA-seq

In 8 patients, we identified deep intronic variants resulting in inclusion of pseudo-exons into the transcript (patients C6, C7, N25, N33, N29-32). All intronic variants

identified were missing from the 1000 Genomes dataset [41] as well as an internal dataset of 5,500 WGS samples.

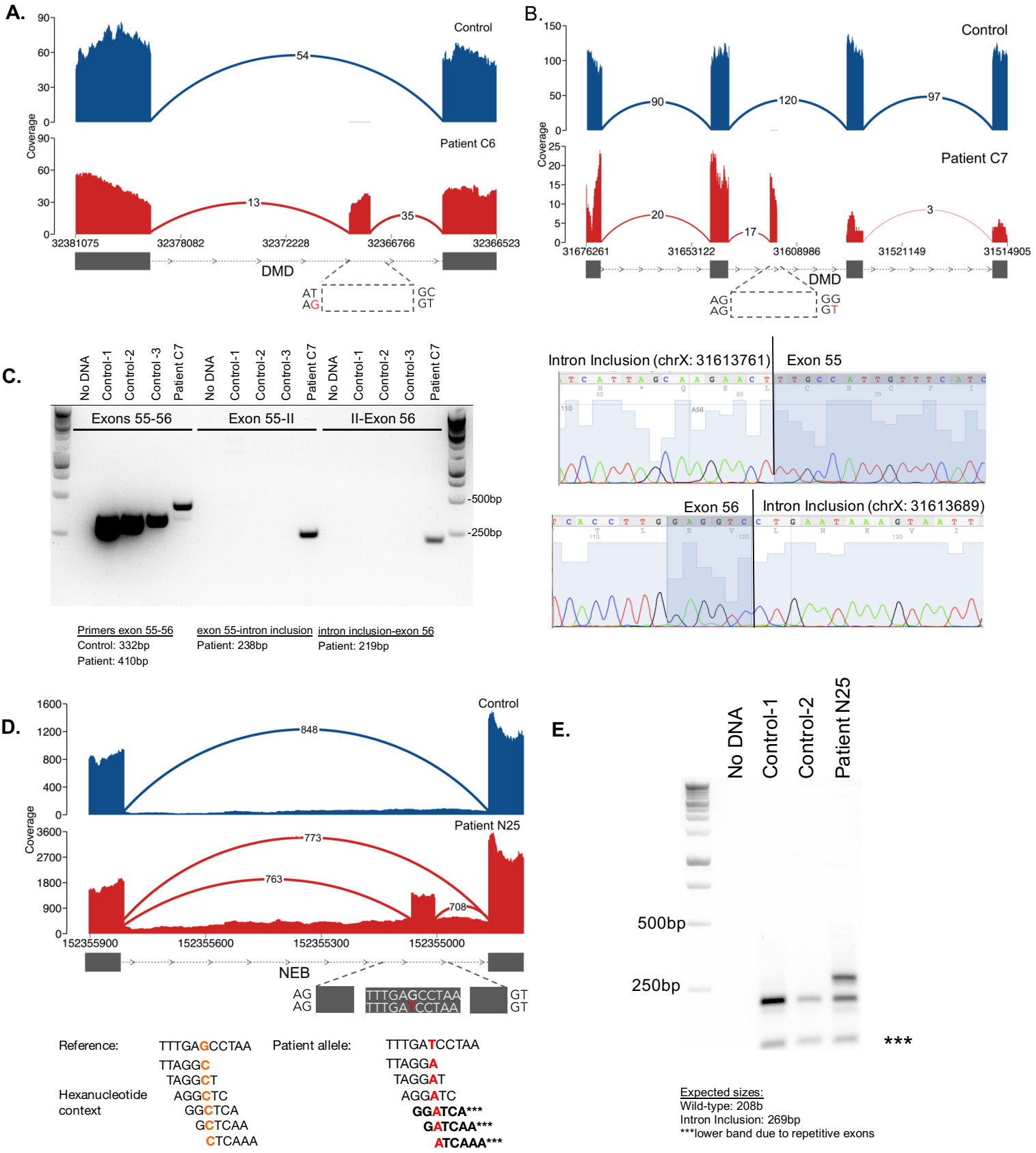
In patients C6, C7, and N33 we identified hemizygous splice site-creating variants in *DMD,* which resulted in the introduction of a premature stop codon to the dystrophin transcript. In Duchenne muscular dystrophy patient C6, the intron inclusion is caused by a hemizygous AT>AG acceptor splice site-creating intronic variant that pairs with an adjacent GT splice donor motif, defining an exonic boundary (Figure 2.12A). RNA-seq shows no evidence of wild type splicing between exons 37 and 38, in line with the patient's severe phenotype. This splicing pattern was confirmed by RT-PCR and Sanger sequencing of the mutant band[10].

In patient C7, RNA-seq identified splicing to an intronic region between exons 55 and 56 of *DMD,* which resulted in the introduction of premature stop codon to the dystrophin transcript (Figure 2.12B). To confirm the inclusion of the pseudo-exon and to identify its exact breakpoints, we performed RT-PCR and Sanger sequencing of the mutant band (Figure 2.12C). RT-PCR results confirmed the presence of the included pseudo-exon caused by a GG>GT donor splice site-creating variant as well as canonical splicing between exons 55 and 56, in line with the milder Becker muscular dystrophy diagnosis of the patient.

In patient N33, the inclusion of the pseudo-exon was caused by a GC>GT donor splice site-creating variant occurring between exons 43 and 44 (Figure 2.8D). RNA-seq

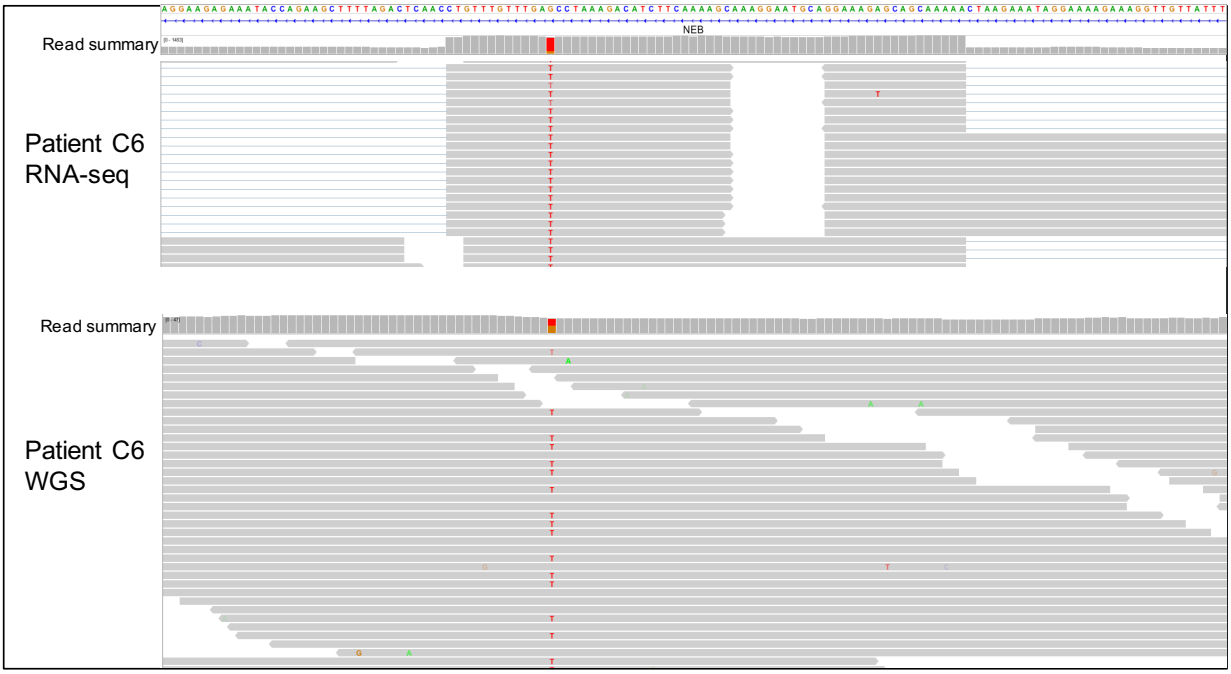
Figure 2.12 Identification of pathogenic noncoding variants with RNA-seq A. Inclusion of a pseudo-exon identified in DMD in patient C6. RNA-seq shows no evidence of wild type splicing between exons 37 and 38, in line with the patient's severe phenotype **B.** Inclusion of a psuedo-exon in DMD in Becker muscular dystrophy patient C7 **C.** Confirmation of pseudo-exon in patient C7 with RT-PCR (left) and Sanger sequencing of mutant band (right). RT-PCR shows the inclusion event is present only in the patient along with a fainter band supporting wild type splicing, indicating that full length dystrophin transcript without the pseudo-exon is being produced, in line with the milder phenotype of the patient. Controls for RT-PCR are patients without the DMD mutation (N8,N22,N33) **D.** inclusion of a pseudo exon in NEB in patient N25. The hexanucleotide context of the variant shows 3 splice enhancer motifs with the patient's allele and none in the reference, suggesting the variant is acting through creation of exonic splice enhancer motifs **E.** RT-PCR confirmation of pseudo-exon created in patient N25 with patients N22 and N8 as controls. **F.** allele balance in RNA-seq at the variant is heavily skewed toward the alternate allele (top), with ~50% allele balance in the WGS data from the patient (bottom) demonstrating the inclusion event is occurring on the variant haplotype **G.** RT-PCR confirmation of pseudo-exon created by a splice-site-creating variant in patient N33 H) Western blot showing reduced dystrophin levels in the patient, in line with a mild Becker muscular dystrophy diagnosis

(Figure 2.12 continued)

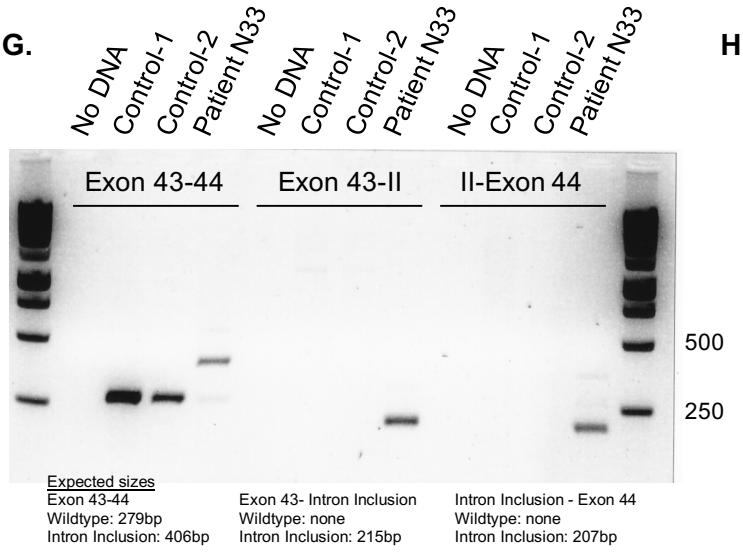


(Figure 2.12 continued)

F.



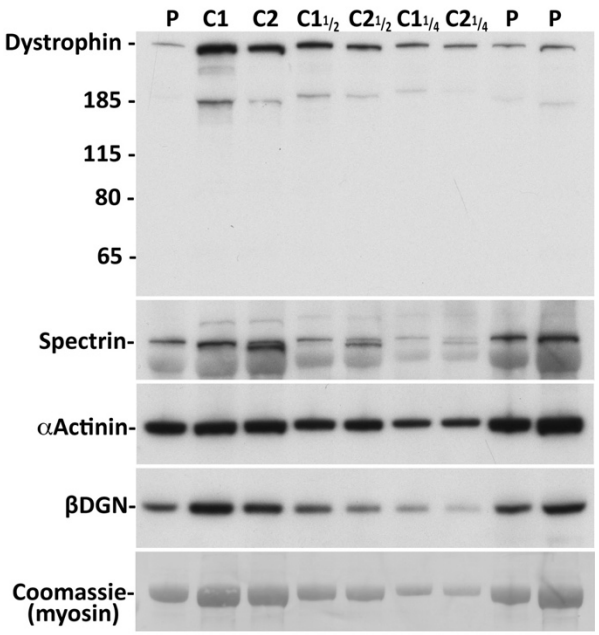
G.



Expected sizes

Exon 43-44	Exon 43- Intron Inclusion	Intron Inclusion - Exon 44
Wildtype: 279bp	Wildtype: none	Wildtype: none
Intron Inclusion: 406bp	Intron Inclusion: 215bp	Intron Inclusion: 207bp

H.



data supported inclusion of the pseudo-exon as well as canonical splicing between exons 43 and 44, indicating that full length dystrophin transcript without the pseudo-exon is being produced, in line with the milder phenotype of the patient. This splicing

pattern was confirmed by RT-PCR (Figure 2.12G), and segregation analysis confirmed the presence of the variant in the unaffected mother and affected brother. This resulted in a genetic diagnosis of previously unidentified mild Becker muscular dystrophy of the patient who had presented with myalgia and myoglobinuria. To confirm the Becker muscular dystrophy diagnosis, we performed immunoblot of skeletal muscle lysates from the patient (male, quadriceps, 21 yrs 11 m) and two age-matched controls (control 1: quadriceps, male 21 yrs 9 m; control 2: vastus lateralis, female, 23 years). Levels of dystrophin were reduced in the patient, relative to controls (Figure 2.12H). A standard curve of muscle lysate from controls estimates the levels of dystrophin in the affected individual to be around a quarter of levels observed in control muscle. Levels of beta-dystroglycan (DGN) were also reduced in the patient, consistent with secondary reduction of levels of other members of the dystrophin-dystroglycan complex.

In patient N25, RNA-seq identified inclusion of a pseudo-exon in *NEB*. The pseudo-exon is flanked by canonical GT/AG splice site motifs but a heterozygous variant absent in 1,000 Genomes and an internal dataset of 5,550 WGS samples is present in the pseudo-exon. We examined the hexanucleotide context of the position with the reference and alternate alleles, querying the RESCUE-ESE database (http://genes.mit.edu/burgelab/rescue-ese/) to identify possible splice enhancer or silencer motifs. We found 3 splice enhancer motifs with the patient's variant allele and none in the reference, suggesting the variant is acting through creation of exonic splice enhancer motifs for the inclusion of the pseudo-exon (Figure 2.12D). We also confirmed this inclusion of the pseudo-exon was missing in GTEx adipose (n=231), skin (n=241), and fibroblast (n=156) samples to ensure the event was not observed as a result of

contamination of the muscle tissue. Allele balance in RNA-seq at the position is heavily skewed toward the alternate allele, whereas allele balance for the variant in the WGS data from the patient is ~50%, demonstrating the inclusion event is occurring on the variant haplotype (Figure 2.12F). We also identified a nonsense variant in the *NEB* triplicate repeat region in the patient (Figure 2.3C). The inclusion of the pseudo-exon was confirmed with RT-PCR (Figure 2.12E), and segregation analysis confirmed the proband carried both variants, with the unaffected father carrying only the intronic variant.

Identification of aberrant splicing overlapping structural variants

In four patients (patients D1,C2-4), we identified aberrant transcriptional signatures around structural variants.

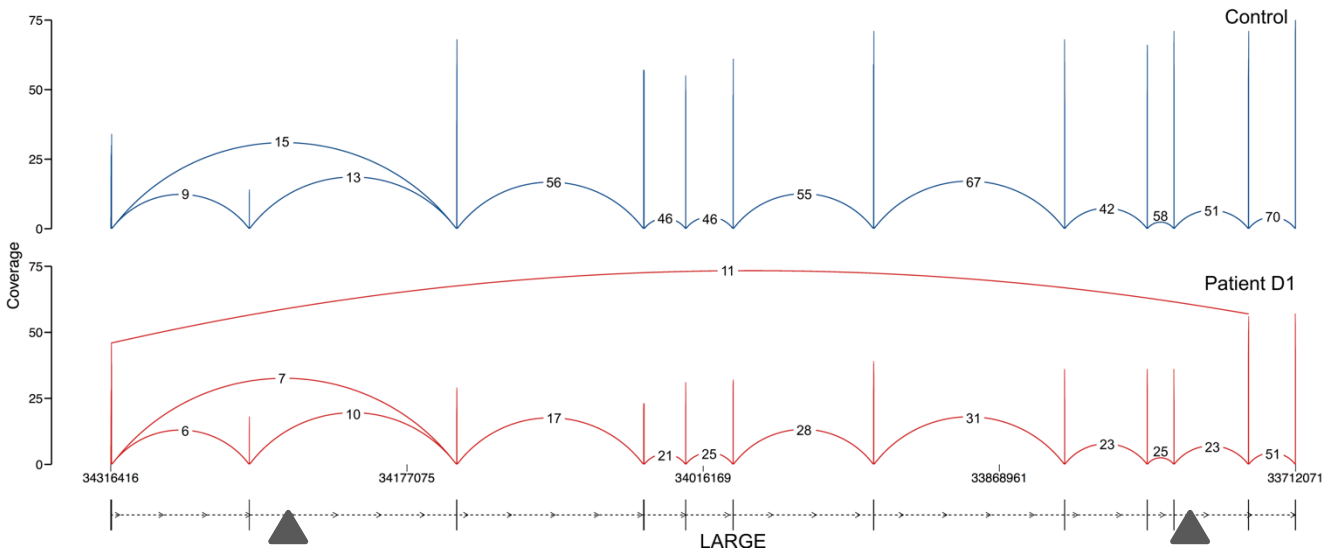
In patient D1, for whom previous WES and array CGH had identified a missense variant and a ~446 kb deletion in *LARGE,* we observed aberrant splicing between the exons flanking the deletion (Figure 2.6A).

In three previously undiagnosed Duchenne muscular dystrophy patients, we identified aberrant splicing in *DMD* that were confirmed to overlap structural variants. In patient C3, we identified skipping of exon 51 (Figure 2.13B) with no rare variants in the local genomic region

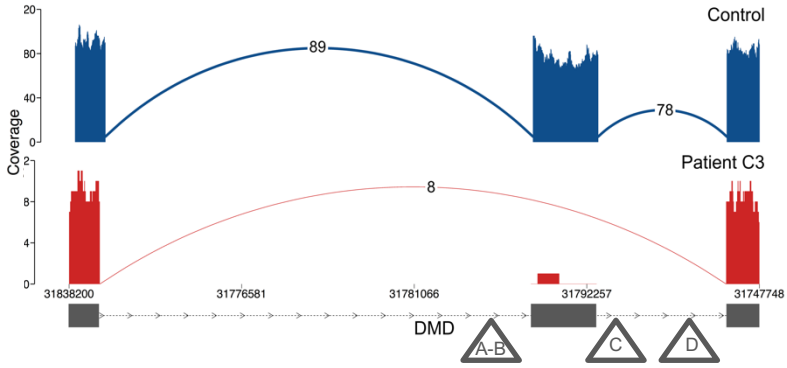
Figure 2.13 Identification of aberrant splicing overlapping structural variants with RNA-seq. A) Aberrant splicing in patient D1, around a ~450 kb deletion in *LARGE*. The breakpoints of the deletions are represented by gray triangles where deletion of the segment results in splicing between remaining exons in the gene. **B)** Skipping of exon 51 in patient C3 identified via RNA-seq is caused by a complex inversion-deletion event in *DMD* spanning ~7.8 kb, denoted by gray triangles. **C,D)** Schematic of structural rearrangement in patient C3 and design of primers to confirm the inversion-deletion event (C). Sanger sequencing of junction breakpoints confirms the event in the patient (D). **E)** Exon-level expression and **F)** splicing patterns in *DMD* in patients C2 and C4 show drop in coverage that are not seen in controls or other DMD patients such as patient C3, suggesting the presence of a structural variant. **G)** WGS identifies a 2.6 mb inversion in patient C4 and **H)** a 119 mb inversion in patient C2, with breakpoints overlapping the drop in coverage and disruption of splicing in both patients. **I,J)** Schematic of structural rearrangement, design of primers, and confirmation of the inversion through Sanger sequencing of junction breakpoints in patient C4 (I) and C2 (J).

(Figure 2.13 continued)

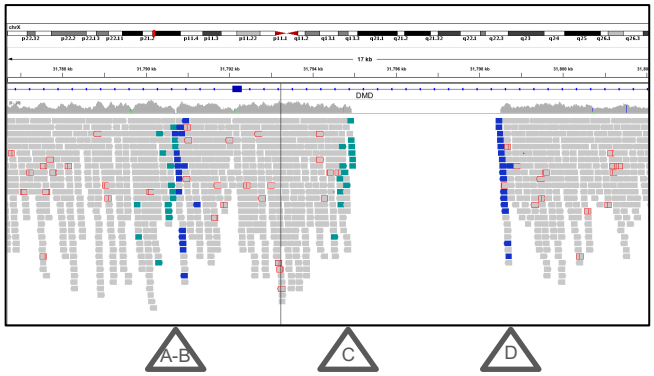
A.



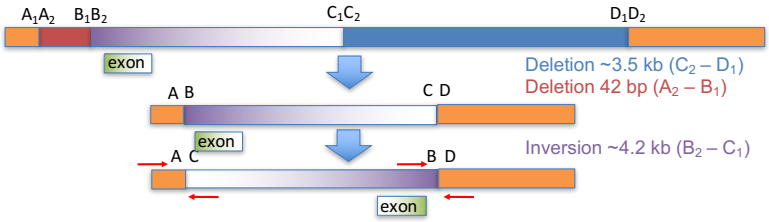
B.

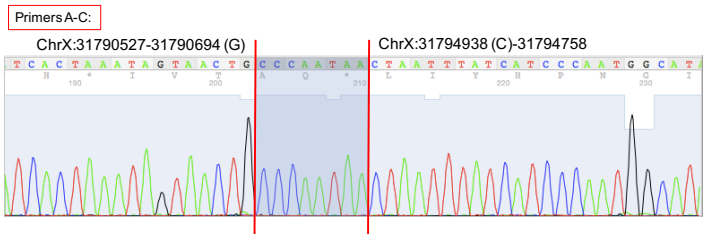


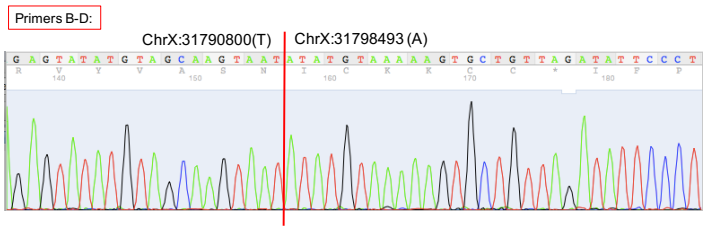
C.



D.

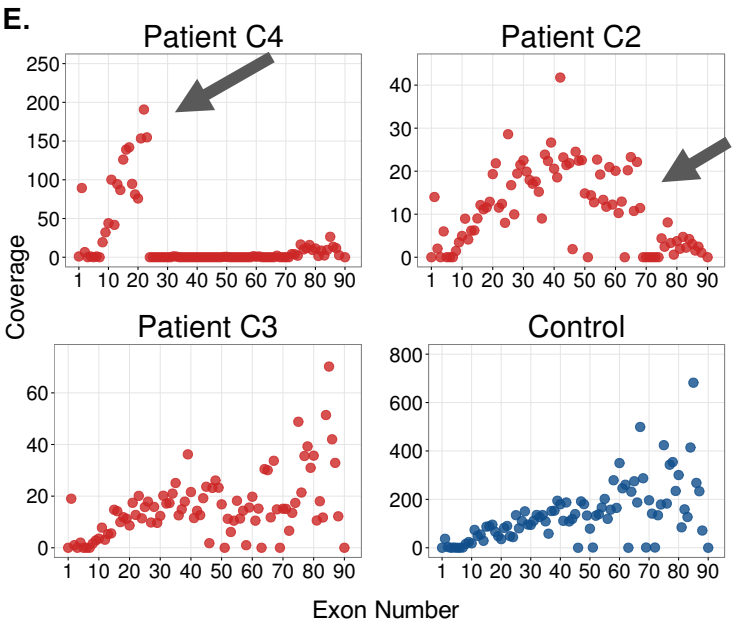


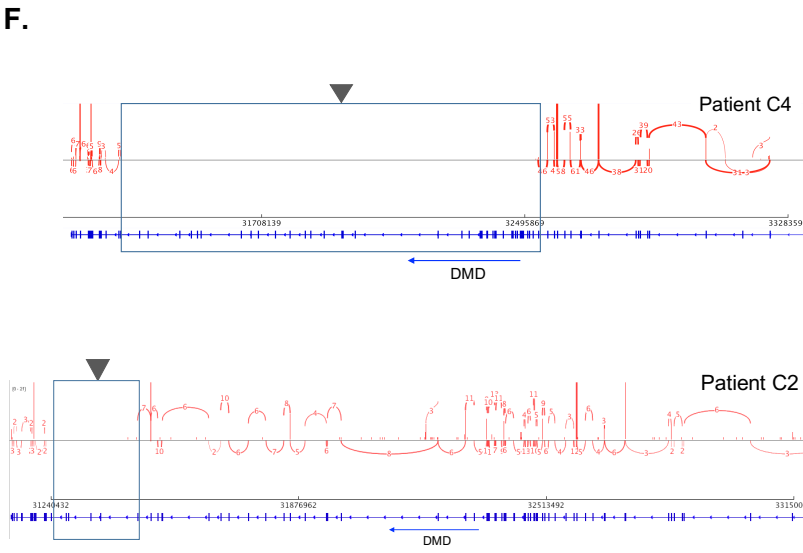


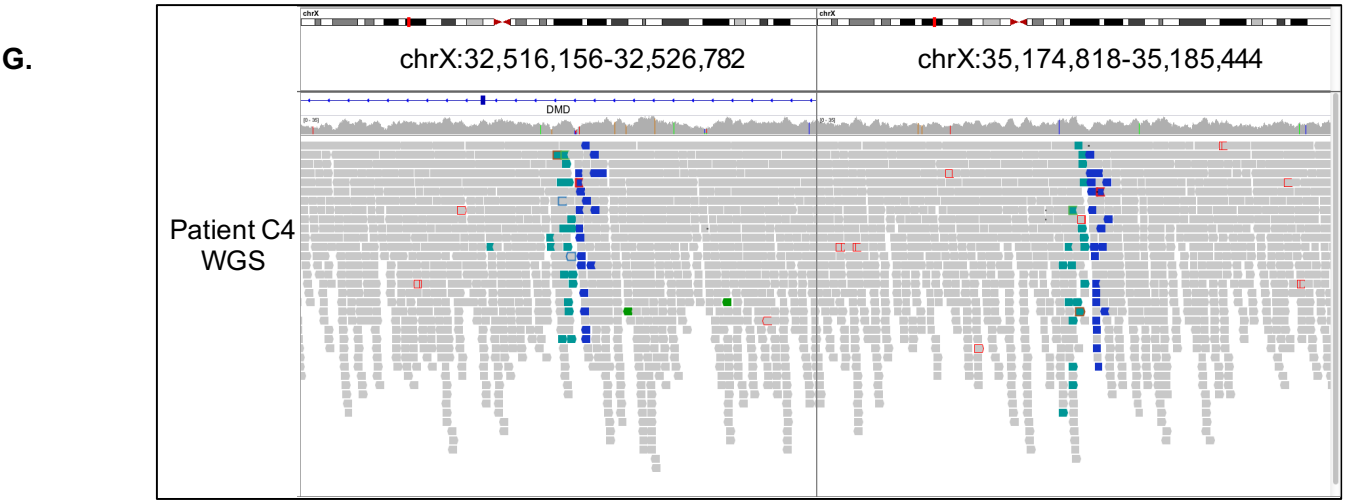


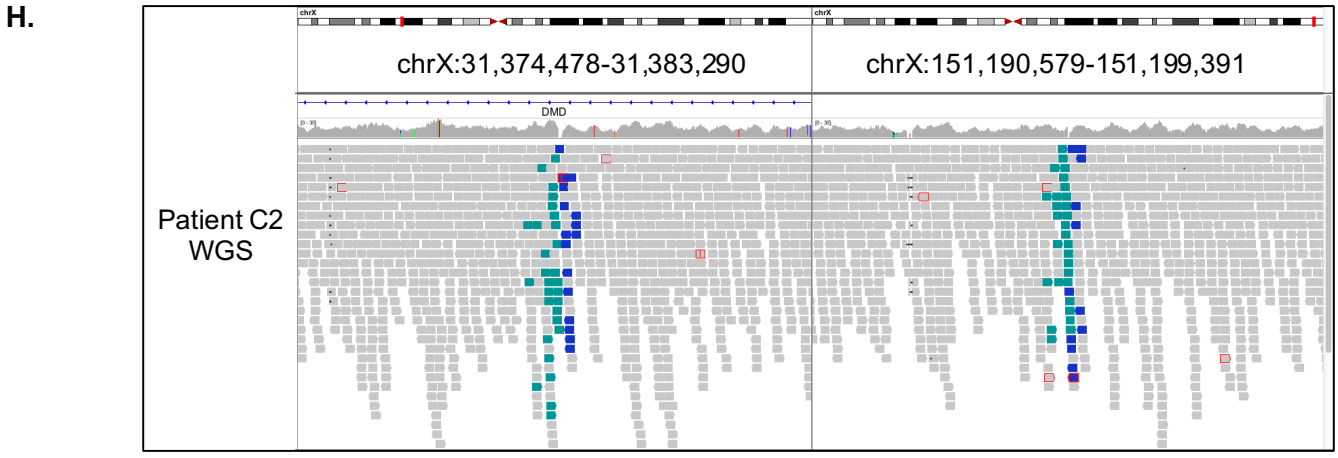
Primers	Chromosome	Strand	Start	End
AC	X	-	31794758	31794938
AC	X	+	31790527	31790694
BD	X	+	31798493	31798728
BD	X	-	31790737	31790800

(Figure 2.13 continued)



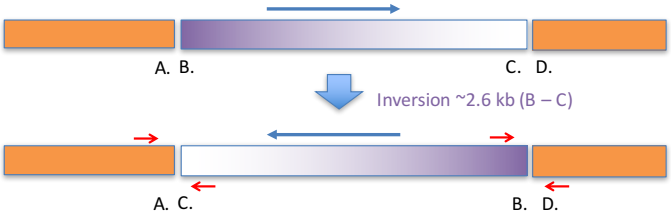


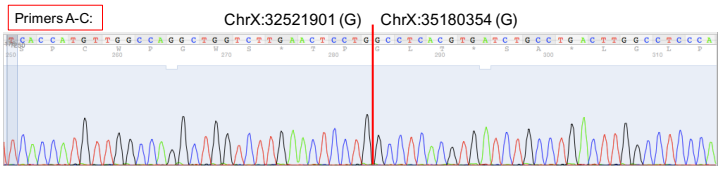




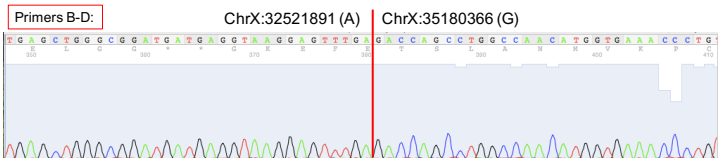
(Figure 2.13 continued)

I.

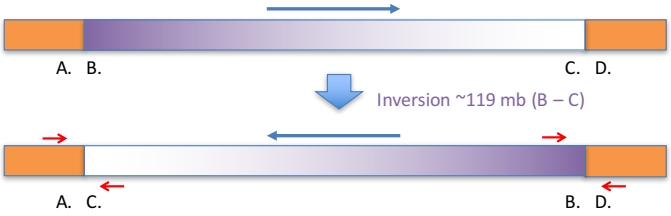


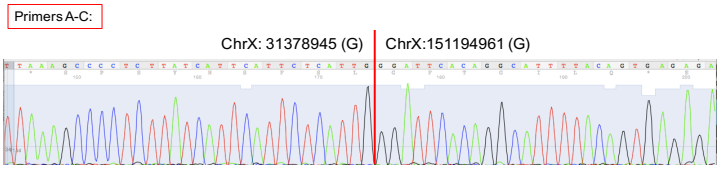


Primers	Chromosome	Strand	Start	End
AC	X	+	32521654	32521901
AC	X	-	35180151	35180354
BD	X	+	35180366	35180478
BD	X	-	32521891	32522022

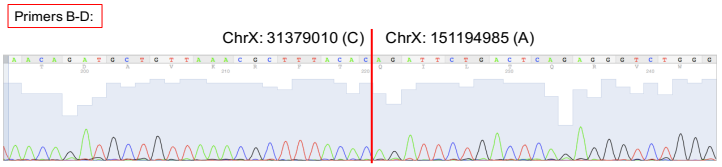


J.





Primers	Chromosome	Strand	Start	End
AC	X	-	151194733	151194961
AC	X	+	31378836	31378946
BD	X	+	151194985	151195213
BD	X	-	31379010	31379200



in WES data to explain the event, prompting us to pursue WGS in the patient. This identified an inversion of the region with an accompanying deletion (Figure 2.13C). This inversion encompasses exon 51 in *DMD,* disrupting GT/AG splice site signals and resulting in the out-of-frame skipping of the exon. The event was confirmed through Sanger sequencing of putative junction breakpoints primed from opposite strands (Figure 2.13D).

In patients C4 and C2, we observed a drop in exon level expression and splicing patterns around exons 19 and 62 of *DMD,* respectively. Similar drops in coverage were

missing in other DMD patients such as C3 and GTEx controls (Figure 2.13E,F). With no plausible candidate variants identified in the regions with WES, we pursed WGS and identified two inversion events. In patient C4, the drop exon expression and splicing overlapped the breakpoint of a 2.6 mb inversion around exon 19, whereas in patient C2 the drop overlapped a 119 mb inversion around exon 62 (Figure 2.13G,H). Sanger sequencing of junction breakpoints confirmed the inversion events in both patients (Figure 2.13I,J).

Identification of a recurrent splice site-creating variant in collagen VI-related dystrophy.

A notable example of the power of transcriptome sequencing is our discovery of a genetic subtype of severe collagen VI-related dystrophy, which is caused by mutations in one of three collagen 6 genes (*COL6A1, COL6A2*, and *COL6A3*)[25]. In four patients who had previously tested negative with deletion/duplication testing and fibroblast cDNA sequencing of the collagen VI genes as well as clinical WES and WGS, we identified an intron inclusion event in *COL6A1* using RNA-seq (Figure 2.14A). The splicing-in of this intronic segment, which is missing in GTEx controls and all other patients in our cohort, is caused by a donor splice site-creating GC>GT variant that pairs with a cryptic acceptor splice site 72 bp upstream, creating an in-frame pseudo-exon (Figure 2.14B). This variant is missing in the 1000 Genomes Project dataset [41] as well as an in-house dataset of 5,500 control WGS samples. The resulting inclusion of 24 amino acids occurs within the N-terminal triple-helical collagenous G-X-Y repeat region

of the *COL6A1* gene, the disruption of which has been well-established to cause dominant-negative pathogenicity in a variety of collagen disorders[60].

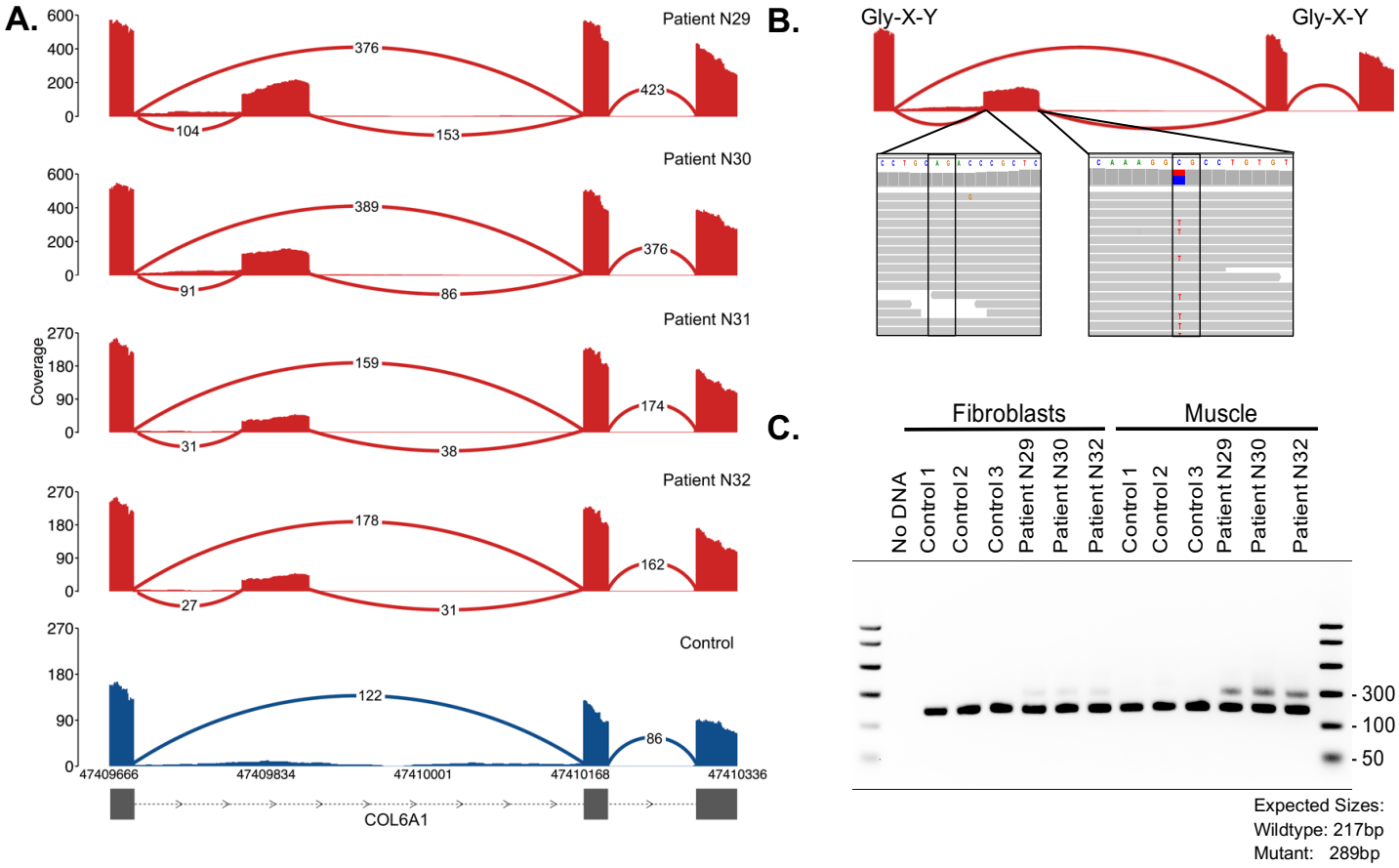


Figure 2.14 Identification of a recurrent splice site-creating variant in four collagen VI-related dystrophy patients. A. Splicing in of the pseudo-exon was observed in four patients in our cohort (red) and missing in all other patients and GTEx samples (blue). **B.** Inclusion of the 24 amino acid segment is caused by a C>T donor splice site-creating variant which pairs with a AG splice acceptor site 72 bp upstream. The variant is found in a CpG nucleotide context, which likely explains its recurrent *de novo* status, and disrupts the Gly-X-Y repeat motifs of *COL6A1*. **C.** The inclusion event is observable in RT-PCR amplicons from patient muscle but is found at comparatively lower levels in cultured dermal fibroblasts derived from the patients, explaining why the pathogenic event was missed in all four patients through previous fibroblast cDNA sequencing.

Of note, cDNA analysis shows that the aberrant transcript is observable in muscle but in much smaller amounts in cultured dermal fibroblasts, making the event identifiable by muscle transcriptome analysis despite being previously missed by fibroblast cDNA

sequencing (Figure 2.14C). Using this information, we genotyped the variant in a larger, genetically undiagnosed collagen VI-like dystrophy cohort and identified 27 additional patients carrying the intronic variant. We confirmed that the variant had occurred as an independent *de novo* mutation in all 16 families for whom trio DNA was available. Based on this screening, we estimate that up to a quarter of all cases clinically suggestive of collagen VI-related dystrophy but negative by exon-based sequencing are due to this recurrent *de novo* mutation.

Screening of additional collagen VI-related dystrophy patients for the *COL6A1* chr21:47,409,881 mutation

The *COL6A1* chr21:47,409,881 mutation identified by RNA-seq in four patients in our study was subsequently screened via genomic sequencing in 637 patients across seven diagnostic centers. From the results of this screening, we can begin to estimate the frequency of this mutation among patients without previously identified mutations in *COL6A1*, *COL6A2,* or *COL6A3.* These estimated frequencies are inclusive of those cases identified via RNA-seq and those cases identified via genomic sequencing (out of a total of 641 patients). The estimated frequencies (listed below) vary depending on the particular diagnostic setting, highlighting the importance of clinicians and scientists with disease-specific expertise (in this instance, collagen VI-specific expertise) working together to arrive at deeply phenotyped and histotyped patients.

1. For patients screened by centers with clinical and research expertise in collagen VI (n=3), where DNA samples were identified for screening by both

clinicians and scientists based on a clinical phenotype suggestive of collagen VI deficiency and muscle immunohistochemical findings of decreased or mislocalized collagen VI, the *COL6A1* mutation was identified in a total of 18/81 (22%) of patients (2/10; 15/68; 1/3 across three centers)

2. For patients screened by laboratories specializing in collagen VI research (n=2), where DNA samples were identified for screening without further input of clinicians with expertise in collagen VI, the *COL6A1* mutation was identified in 6/149 (4%) of patients (1/59; 5/90 across two centers)

3. For patients screened by diagnostic laboratories where all available DNA samples (referred for sequencing of the collagen 6 genes and without previously identified mutations in *COL6A1*, *COL6A2,* or *COL6A3*) were screened (n=2) without further input of clinicians or scientists with expertise in collagen VI, the *COL6A1* mutation was identified in 7/411 (1.7%) of patients (3/361; 4/50 across two centers)

Evaluation of splice prediction algorithms and RNA-seq in alternative tissues

Exons harboring the pathogenic variants identified in this study show low coverage in GTEx whole blood and fibroblast samples, indicating that a majority of these diagnoses likely could not have been made using RNA-seq from these tissues (Figure 2.15). Furthermore, many of the diagnoses made in this study could not have been made on genotype information alone, as splice prediction algorithms alone are

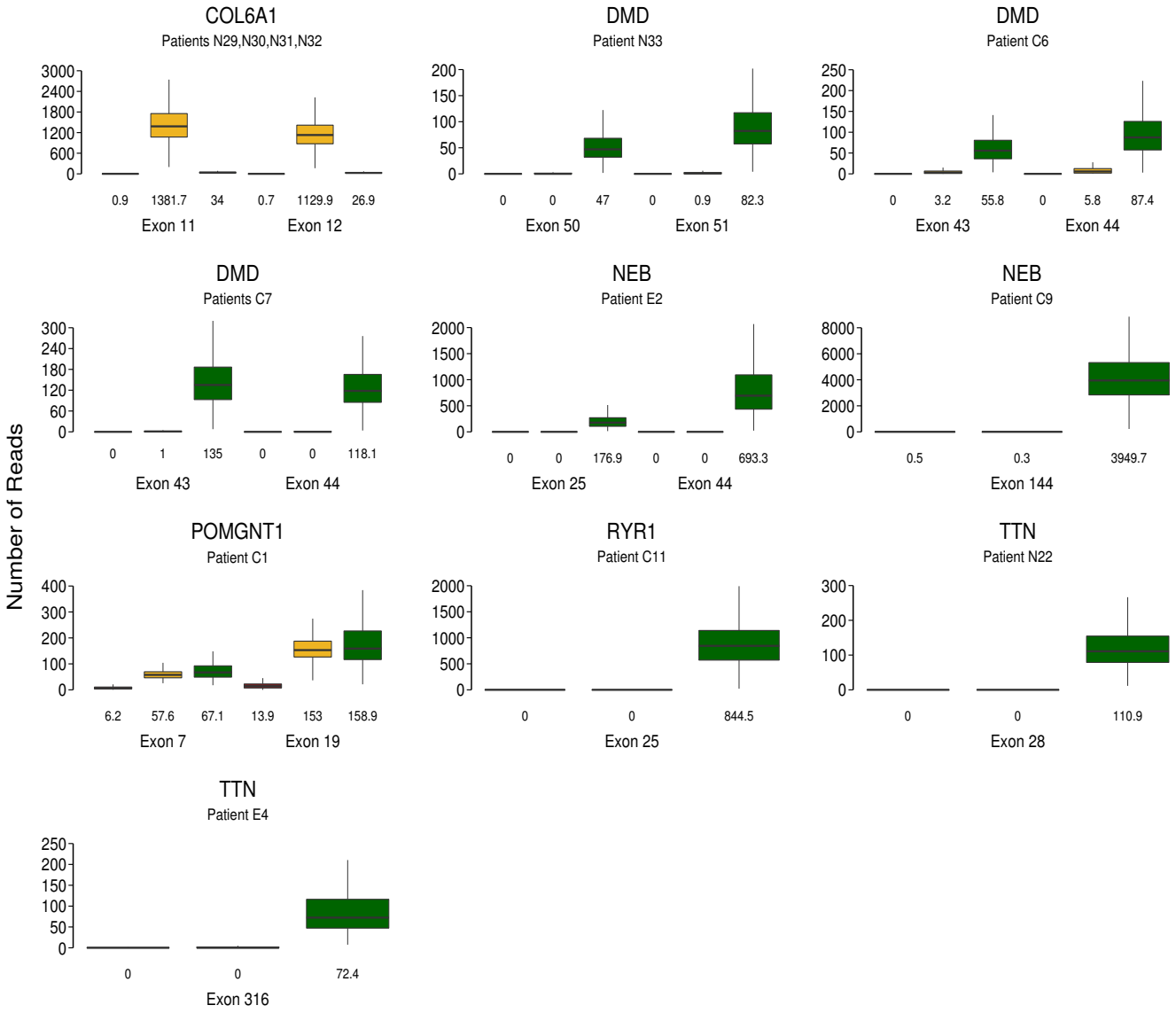


Figure 2.15 Comparison of the number of reads aligning to exons harboring pathogenic variants identified in the study in GTEx muscle, whole blood, and fibroblast tissues. We assayed the number of reads aligning to each exon in which a pathogenic variant was identified in a patient in GTEx muscle (n=430, green), whole blood (n=393, red), and fibroblast (n=284, yellow) samples in order to evaluate whether diagnoses made in the study could have been made via RNA-seq from these tissues. The data illustrate overall low coverage of the affected exons in fibroblasts and whole blood, suggesting that a large portion of variants identified in this study may not have been identified based on sequencing of these tissues. Numbers under boxplots represent the median number of reads aligning to the exon in each tissue

currently insufficient to classify variants as causal[12,61]. Although existing *in silico* algorithms correctly predicted disruption for the two extended splice site variants of unknown significance in our study, they also generated false positive predictions for the remaining two extended splice site variants with no effect on splicing. In addition, existing algorithms showed poor specificity in identifying splice site-creating coding variants, identifying on average over 100 putative splice site-creating rare variants (<1% population frequency in ExAC) exome-wide (Figure 2.16).

Discussion

Our results show that RNA-seq is valuable for the interpretation of coding as well as non-coding variants, and can provide a substantial increase in diagnosis rate in patients for whom exome or whole genome analysis has not yielded a molecular diagnosis. In our cohort, RNA-seq led to the diagnosis of 66% of patients where clinical phenotyping and DNA sequencing prioritized a strong candidate gene. In comparison, through identifying aberrant splice events found in patients and missing in GTEx controls, we were able to diagnose 21% of patients with no strong candidates from WGS or WES.

Our work illustrates the value of large multi-tissue transcriptome data sets such as GTEx to serve as a reference to facilitate the identification of extreme splicing or allele balance outlier events in patients. In the case of muscle disorders, our diagnoses were made primarily through direct identification of aberrations in splicing using the GTEx skeletal-muscle RNA-seq dataset as a reference panel. Our present work focused on identifying such

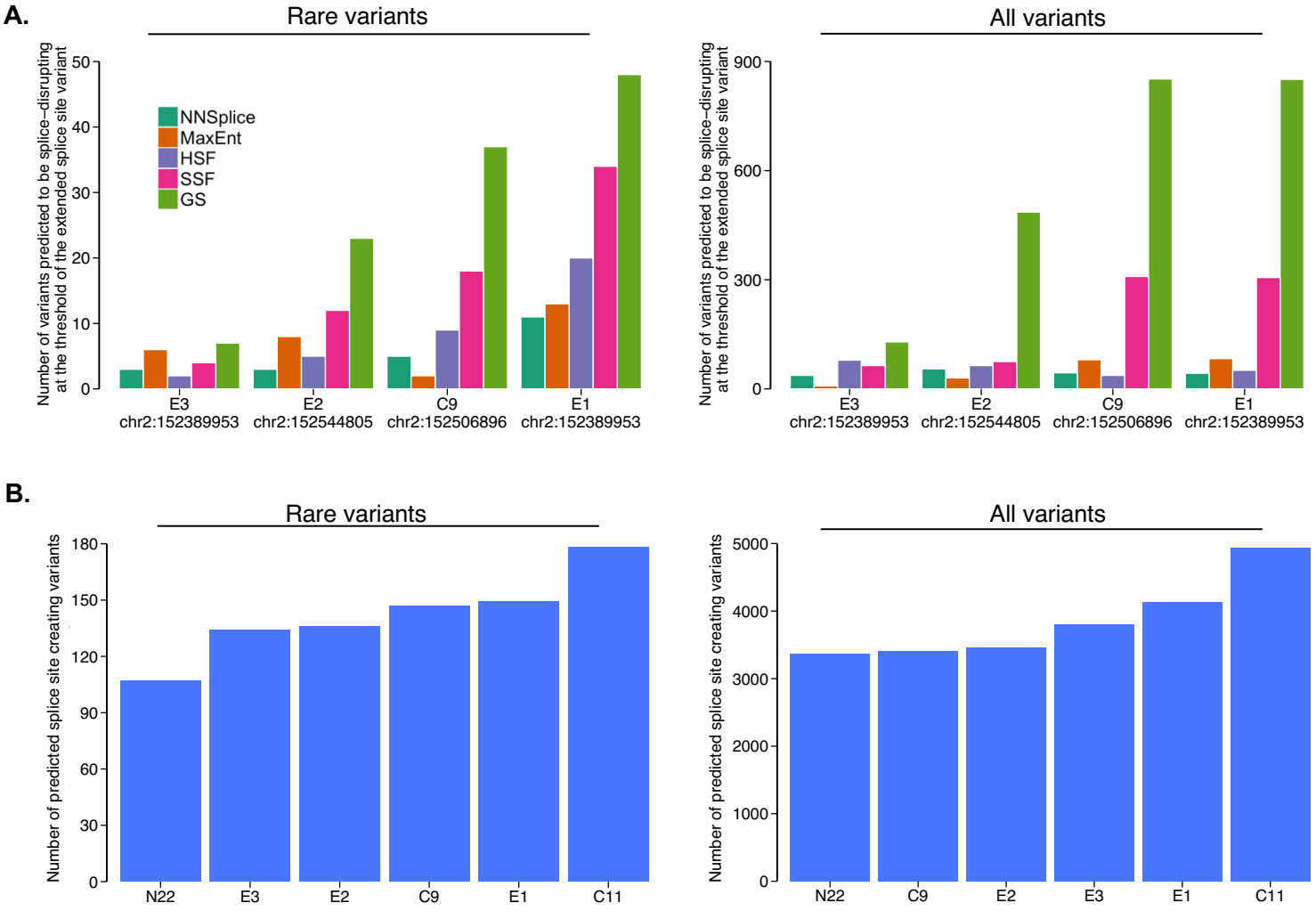


Figure 2.16 Evaluation of splice prediction algorithms. A. The number of rare (ExAC AF <1%, GQ>10) and all variants predicted to be disrupting at or above the threshold of the extended splice site variants evaluated with RNA-seq for patients E1, E2, E3, and C9. Few rare variants are predicted to be as damaging as the pathogenic extended splice site variants confirmed to disrupt splicing with RNA-seq in patients E2 and C9, indicating that these variants would likely be identified based on *in silico* predictions. However, extended splice site variants that do not disrupt local splicing patterns in patients E3 and E1 also show this pattern, illustrating that use of in silico predictions could lead to false pathogenciity assignments. **B.** An average of 140 rare (ExAC AF <1%, GQ>10) and ~2000 total variants per patient are predicted to result in gain of splicing, indicating that *in silico* predictions currently lack the specificity to make pathogenicity assignment on DNA sequence information alone.

aberrations in known muscle disease genes, and the considerably lower number of putatively pathogenic events identified in neuromuscular disease genes versus all genes underlines the advantage of a candidate gene list for this analysis Further improvements in filtering identified splice junctions to obtain a smaller list of candidate events will be useful to expand this work for new disease gene discovery. In addition, with increasing sample sizes and improvements in methods, RNA-seq can also be used to identify somatic variants and to detect regulatory variants upstream, through analysis of expression status and allelic imbalance.

Access to the disease-relevant tissue for many Mendelian disorders remains a major barrier for the use of transcriptome sequencing in genetic diagnosis. The RNA-seq framework developed in this study can be adapted for rare diseases where biopsies are available, such as Mendelian disorders affecting heart, kidney, liver, skin, and other tissues. For disorders where biopsy of the disease-relevant tissue is unattainable, analyses are possible through identification of proxy tissues using databases such as GTEx and careful consideration of the expression status of the relevant genes in the proxy tissue. Alternatively, the framework developed in this study can also enable diagnoses through reprogramming patient cells into induced pluripotent stem cells and differentiation into disease-relevant tissues of interest.

Evaluation of existing splice prediction algorithms for the splice-disrupting variants identified in the study highlights that information on DNA sequence alone does not currently match the ability of RNA-seq to identify the transcriptional consequences of variants on a genome-wide scale. The diagnoses made in our study with RNA-seq, particularly the discovery of the highly recurrent mutation in *COL6A1,* demonstrates that

other such cryptic splice-affecting variants may contribute substantially to undiagnosed diseases that have evaded prior detection with exome or whole genome analysis.

Overall, this work suggests that RNA-seq is a valuable component of the diagnostic toolkit for rare diseases and can aid in the identification of new pathogenic variants in known genes as well as new mechanisms for Mendelian disease.

Significance

Our ability to turn the dizzying rate of WES data into practical clinical knowledge has relied on careful pipelines of data processing, annotation, and interpretation. Several tools have been developed for processing DNA sequencing data for Mendelian diagnosis including tools for quality controlling data, aligning sequences to the reference and efficient identification of polymorphisms in a genome. Guidelines have also been developed to aid interpretation of variant-level pathogenicity [62]. At the time of the project the Exome Aggregation Consortium had release a dataset of over 60,000 ostensibly healthy indivdiuals as a population reference[40], a number which has now increased to over 120,000 with the Genome Aggregation Database Consortium (gnomAD)[63]. Previously, no reference panels or pipelines existed for the purpose of Mendelian disease diagnosis using RNA sequencing. This project has built such frameworks by using existing RNA-seq processing tools as well as building novel methods for detection of pathogenic transcription events.

The integration of patient transcriptome data to improve diagnosis is now being widely adopted by the rare disease research community, evidenced by several publications in the past year[64–68] as well as increase of poster abstracts that mention the

use of RNA-seq in genetic diagnosis in conferences such as Genomics of Rare Disease. In fact, the 2018 American Society of Human Genetics meeting featured a platform session entitled “Using RNA-seq to prioritize Mendelian variants” featuring talks from the Emory University Clinical Sequencing platform and the Undiagnosed Disease Network and the 2019 meeting will feature a workshop “RNA-seq for Mendelian disease diagnostics: A hands-on tutorial through bioinformatic tools and workflows”, emphasizing the growing interest in using transcriptome sequencing methods to improve variant interpretation. This workshop will be co-host by me and representatives from the labs of Stephen Montgomery and Julien Gagneur.

Our group is also scaling the use of transcriptome sequencing for genetic diagnosis where tissue is available-as part of the Broad Center for Mendelian Genomics and the Rare Genomes Project, with a combined goal of DNA sequencing for over 5,000 rare disease families in the next few years, combined with RNA-seq in hundreds cases where appropriate tissue can be obtained.

We have taken care to ensure the methods in our study can be adopted and improved upon by other groups. This included the publication of a blog post accompanying our manuscript available at macarthurlab.org/blog that discusses study design considerations for cohort-level RNA-seq as well as the commands to run to reproduce our analyses (blog post is attached in the Appendix). As of April 2019, this post had been viewed over 4,000 times.

Lastly, while our focus has been to improve genetic diagnosis, our discovery of the unexpectedly common *COL6A1* intronic variant is now under consideration for therapeutic development. Our collaborators at the NIH have developed antisense

oligonucleotides targeting the pseudo-exon and have shown successful repression in patient-derived fibroblasts without affecting *COL6A1* expression[69]. Further development has the potential to lead to an FDA fast-tracked therapy and benefit the collagen dystrophy patient community. Importantly, this emphasizes the utility of RNA-seq to identify treatable splice-defects.

Author contributions

Beryl B. Cummings: conceived and designed experiments, analyzed transcriptome sequencing data, wrote the analysis text

Daniel G. MacArthur: conceived and designed experiments, writing edits, general guidance

Jamie L. Marshall: designed and performed RT-PCR and Sanger sequence validation experiments, provided comments on the manuscript

Ying Hu, Adam Bournazos, Mark Davis: designed and performed RT-PCR and Sanger sequence validation experiments, provided comments on the manuscript

Taru Tukiainen: Analyzed transcriptome sequencing data, provided comments on the manuscript

Monkol Lek: aided the analysis of exome and whole-genome data, general guidance

Sandra Donkervoort, Ying Hu, A. Reghan Foley, Veronique Bolduc, Carsten G. Bonnemann: aided the analysis of exome and whole-genome data, performed follow-up analyses in collagen VI dystrophy cohort.

Daniel Birnbaum, Ben Weisburd, Fengmai Zhao, Konrad J Karczewski, Anne O'Donnell-Luria: . contributed reagents/materials/analysis tools

Leigh Waddell, Gina L. O'Grady, Elicia Estella, Hemakumar Reddy, Sandra A Sandaradura, Anna Sarkozy, Hernnann Gonorazky, Kristal Claeys, Himanshu Joshi, Emily C Oates, Rhoula Ghaoui, Nigel Laing, Ana Topf, Peter B Kang, Alan H Beggs, Kathryn North, Volker Straub, James J. Dowling, Francesco Muntoni, Nigel F Clark, Sandra Cooper, Carsten G. Bonnemann: provided patient samples and clinical information, when needed aided in variant interpretation for patient sequence variants.

Bibliography

1. Ankala, A. *et al.* A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. *Ann. Neurol.* **77**, 206–214 (2015).

2. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014).

3. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47**, 717–726 (2015).

4. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).

5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

6. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta* **1792**, 14–26 (2009).

7. Oliveira, J. *et al.* Novel synonymous substitution in POMGNT1 promotes exon skipping in a patient with congenital muscular dystrophy. *J. Hum. Genet.* **53**, 565–572 (2008).

8. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome. *Nature* **423**, 293–298 (2003).

9. Colapietro, P. *et al.* NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum. Genet.* **113**, 551–554 (2003).

10. Gonorazky, H. *et al.* RNAseq analysis for the diagnosis of muscular dystrophy. *Ann Clin Transl Neurol* **3**, 55–60 (2016).

11. Slaugenhaupt, S. A. *et al.* Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. *Am. J. Hum. Genet.* **68**, 598–605 (2001).

12. Spurdle, A. B. *et al.* Prediction and assessment of splicing alterations: implications for clinical testing. *Hum. Mutat.* **29**, 1304–1313 (2008).

13. Parla, J. S. *et al.* A comparative analysis of exome capture. *Genome Biol.* **12**, R97 (2011).

14. Chilamakuri, C. S. R. *et al.* Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* **15**, 449 (2014).

15. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic

variation, and neurological disease. *Science* **341**, 1237758 (2013).

16. Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* **31**, 382–392 (2015).

17. Donkervoort, S. *et al.* Mosaicism for dominant collagen 6 mutations as a cause for intrafamilial phenotypic variability. *Hum. Mutat.* **36**, 48–56 (2015).

18. Ma, M. *et al.* Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics* **16 Suppl 8**, S3 (2015).

19. Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31**, 2497–2504 (2015).

20. Ghaoui, R. *et al.* Use of Whole-Exome Sequencing for Diagnosis of Limb-Girdle Muscular Dystrophy: Outcomes and Lessons Learned. *JAMA Neurol.* **72**, 1424–1432 (2015).

21. Laing, N. G. Genetics of neuromuscular disorders. *Crit. Rev. Clin. Lab. Sci.* **49**, 33–48 (2012).

22. Lek, M. & MacArthur, D. The Challenge of Next Generation Sequencing in the Context of Neuromuscular Diseases. *J Neuromuscul Dis* **1**, 135–149 (2014).

23. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

24. Ghaoui, R., Clarke, N., Hollingworth, P. & Needham, M. Muscle disorders: the latest investigations. *Intern. Med. J.* **43**, 970–978 (2013).

25. Bönnemann, C. G. *et al.* Diagnostic approach to the congenital muscular dystrophies. *Neuromuscul. Disord.* **24**, 289–311 (2014).

26. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. doi:10.1101/074450

27. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

28. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

29. McDonald, C. M. Clinical approach to the diagnostic evaluation of hereditary and acquired neuromuscular diseases. *Phys. Med. Rehabil. Clin. N. Am.* **23**, 495–563 (2012).

30. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–

21 (2013).

31. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

32. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

33. Schafer, S. *et al.* Alternative splicing signatures in RNA-seq data: Percent spliced in (PSI). *Curr. Protoc. Hum. Genet.* **87**, 11–16 (2015).

34. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).

35. Auwera, G. A. V. der *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* 11.10.1–11.10.33 (2013). doi:10.1002/0471250953.bi1110s43

36. Schaffer, A. Faculty of 1000 evaluation for PLINK: a tool set for whole-genome association and population-based linkage analyses. *F1000 - Post-publication peer review of the biomedical literature* (2009). doi:10.3410/f.1162373.622875

37. Robinson, P. & Jtel, T. Z. Integrative genomics viewer (IGV): Visualizing alignments and variants. *Computational Exome and Genome Analysis* 233–245 (2017). doi:10.1201/9781315154770-17

38. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).

39. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

40. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

41. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

42. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–7 (2005).

43. Bang, M.-L. *et al.* The complete gene sequence of titin, expression of an unusual≈ 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.* **89**, 1065–1072 (2001).

44. Kiiski, K. *et al.* A recurrent copy number variation of the NEB triplicate region: only

revealed by the targeted nemaline myopathy CGH array. *Eur. J. Hum. Genet.* **24**, 574–580 (2016).

45. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

48. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).

49. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).

50. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. *J. Comput. Biol.* **4**, 311–323 (1997).

51. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).

52. Mersch, B., Gepperth, A., Suhai, S. & Hotz-Wagenblatt, A. Automatic detection of exonic splicing enhancers (ESEs) using SVMs. *BMC Bioinformatics* **9**, 369 (2008).

53. Chasin, L. A. Searching for Splicing Motifs. *Advances in Experimental Medicine and Biology* 85–106 (2007). doi:10.1007/978-0-387-77374-2_6

54. Begay, R. L. *et al.* Role of Titin Missense Variants in Dilated Cardiomyopathy. *Journal of the American Heart Association* **4**, (2015).

55. Richards, S. *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.* **17**, 405–424 (2015).

56. Roca, X., Krainer, A. R. & Eperon, I. C. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev.* **27**, 129–144 (2013).

57. Rivas, M. A. *et al.* Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).

58. Lord, J. *et al.* Pathogenicity and selective constraint on variation near splice sites. *Genome Res.* **29**, 159–170 (2019).

59. Zhang, S. *et al.* Base-Specific Mutational Intolerance Near Splice-Sites Clarifies

Role Of Non-Essential Splice Nucleotides. doi:10.1101/129312

60. Butterfield, R. J. *et al.* Position of glycine substitutions in the triple helix of COL6A1, COL6A2, and COL6A3 is correlated with severity and mode of inheritance in collagen VI myopathies. *Hum. Mutat.* **34**, 1558–1567 (2013).

61. Duzkale, H. *et al.* A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453–463 (2013).

62. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

63. Karczewski, K. J., Francioli, L. C., Tiao, G. & Cummings, B. B. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* (2019).

64. Kernohan, K. D. *et al.* Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. *Hum. Mutat.* **38**, 611–614 (2017).

65. Ulirsch, J. C. *et al.* The Genetic Landscape of Diamond-Blackfan Anemia. *Am. J. Hum. Genet.* **104**, 356 (2019).

66. Al-Hashim, A., Gonorazky, H. D., Amburgey, K., Das, S. & Dowling, J. J. A novel intronic mutation in MTM1 detected by RNA analysis in a case of X-linked myotubular myopathy. *Neurology Genetics* **3**, e182 (2017).

67. Hamanaka, K. *et al.* RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. *Genet. Med.* (2018). doi:10.1038/s41436-018-0360-6

68. Fresard, L., Smail, C., Smith, K. S., Ferraro, N. M. & Teran, N. A. Identification of rare-disease genes in diverse undiagnosed cases using whole blood transcriptome sequencing and large control cohorts. *BioRxiv* (2018).

69. Bolduc, V. *et al.* A recurrent COL6A1 pseudoexon insertion causes muscular dystrophy and is effectively targeted by splice-correction therapies. *JCI Insight* **4**, (2019).

Chapter 3

Development and validation of transcript-expression aware annotation to improve rare variant discovery and interpretation

Abstract

The acceleration of DNA sequencing in patients and population samples has resulted in unprecedented catalogues of human genetic variation, but the interpretation of rare genetic variants discovered using such technologies remains extremely challenging. A striking example of this challenge is the existence of disruptive variants in dosage-sensitive disease genes, even in apparently healthy individuals. Through manual curation of putative loss of function (pLoF) variants in haploinsufficient disease genes in the Genome Aggregation Database (gnomAD) [1], we show that one explanation for this paradox involves alternative mRNA splicing, which allows exons of a gene to be expressed at varying levels across cell types. Currently, no existing annotation tool systematically incorporates this exon expression information into variant interpretation. Here, we develop a transcript-level annotation metric, the proportion expressed across transcripts (pext), which summarizes isoform quantifications for variants. We calculate this metric using 11,706 tissue samples from the Genotype Tissue Expression project [2] (GTEx) and show that it clearly differentiates between weakly and highly evolutionarily conserved exons, a proxy for functional importance. We demonstrate that expression-based annotation selectively filters 22.4% of falsely annotated pLoF variants found in haploinsufficient disease genes in gnomAD, while removing less than 4% of high-confidence pathogenic variants in the same genes. Finally, we apply our expression filter to the analysis of *de novo* variants in patients with autism spectrum disorder (ASD) and developmental disorders and intellectual disability (DD/ID) to show that pLoF variants in weakly expressed regions have effect sizes similar to those of synonymous variants, while pLoF variants in highly expressed exons

are most strongly enriched among cases versus controls. Our annotation is fast, flexible, and generalizable, making it possible for any variant file to be annotated with any isoform expression dataset, and will be valuable for rare disease diagnosis, rare variant burden analyses in complex disorders, and curation and prioritization of variants in recall-by-genotype studies.

Introduction

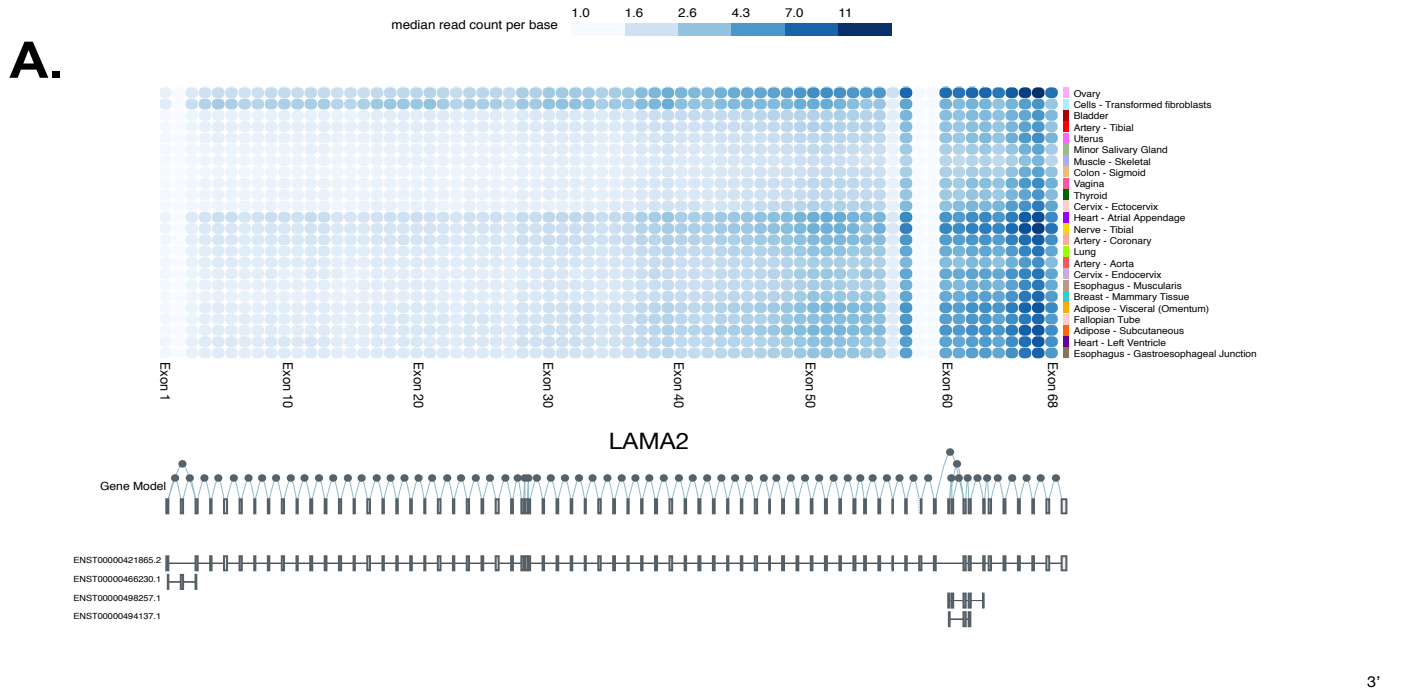
Alternative splicing as a source of variability for variant interpretation

A primary challenge in the use of genome and exome sequencing to predict human phenotypes is that our capacity to identify genetic variation exceeds our ability to interpret their functional impact [3,4]. One underappreciated source of variability for variant interpretation involves differences in alternative mRNA splicing, which enables exons to be expressed at different levels across tissues. These expression differences mean that variants in different regions of a gene can have different phenotypic outcomes depending on the isoforms they affect. For example, variants occurring in an exon differentially included in two isoforms of *CACNA1C* with diverse tissue expression patterns result in distinct types of Timothy syndrome [5]. Pathogenic variants in the isoform that exhibits multi-tissue expression result in a multi-system disorder [5–7], whereas those on the isoform predominantly expressed in heart result in more severe and specific cardiac defects [8]. In addition, Mendelian variants have been found on tissue-specific isoforms [9,10] and isoform expression levels in *TTN* have been used to show that pLoF variants found in healthy controls occur in exons that are absent from

dominantly expressed isoforms, whereas those in dilated cardiomyopathy patients occur on constitutive exons [11] , emphasizing the utility of exon expression information for variant interpretation.

3' bias prevents the use of read-pileup at exons as a proxy for expression

The advent of large-scale transcriptome sequencing datasets, such as GTEx [2], provides an opportunity to incorporate cross-tissue exon expression into variant interpretation. However, the current formats of these databases do not readily allow for unbiased estimation of exon expression. The GTEx web browser offers information on exon-level read pileup across tissues, but this approach is confounded by technical artifacts such as 3' bias [12] (preferential coverage of bases close to the 3' end of a transcript; Figure 3.1A). Such systematic biases mean that simple exon-level coverage in a transcriptome dataset cannot be used as a reliable proxy for exon expression, especially in longer genes (Figure 3.1B). In contrast, isoform quantification tools provide estimates of isoform expression levels that correct, albeit imperfectly [13,14], for confounding by 3' bias as well as other technical artifacts such as isoform length, isoform GC content, and transcript sequence complexity [14–16].



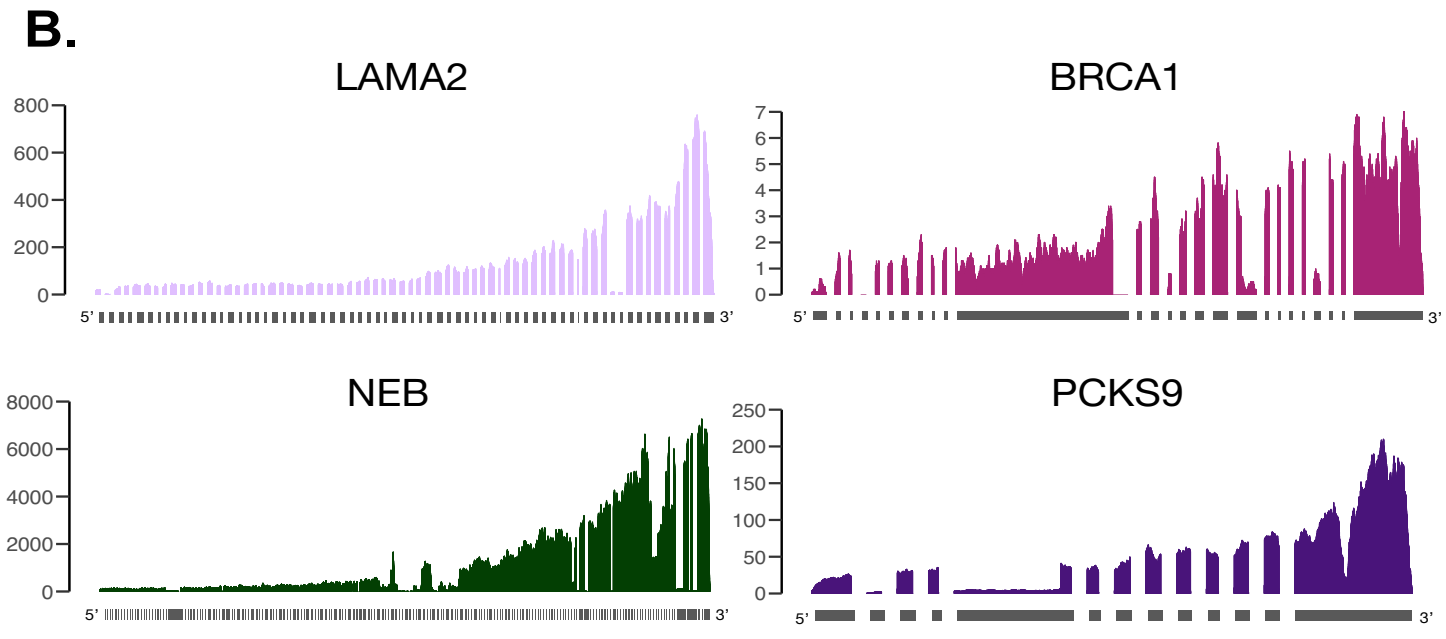


Figure 3.1 Technical artifacts in transcriptome sequencing experiments prevent the use of read pileup at exons as an unbiased proxy for expression A. Example of exon expression information on the GTEx web browser (gtexportal.org) for *LAMA2,* which has 3 annotated transcripts. Blue-gray gradient represents median read count per base. For example, while exons 5 and 55 are annotated on a single transcript, the mean read count for the exons are 0.8 and 3.25 in GTEx ovary, respectively, reflecting the confounding effect of 3' bias **B.** Examples of 3' bias in genes of varying lengths and expression levels shows 3' bias is pervasive. Base-level coverage of uniquely mapped reads were calculated in 10 random GTEx samples per tissue using samtools depth in tissues where the genes are highly expressed. Plots show (1) *LAMA2* in tibial nerve (2) *BRCA1* in mammary tissue (breast) (3) *NEB* in skeletal muscle and (4) *PCSK9* in liver, all of which display 3' bias.

Presence of pLoF variants in dosage sensitive disease genes in public datasets

Genome-based diagnostics offer unprecedented catalogues of human genetic variation, but the interpretation of rare genetic variants discovered using such technologies remains extremely challenging [3,4]. As the human genetics community continues to amass human sequence data, an emerging paradoxical finding is the presence disruptive variants in dosage-sensitive disease genes in ostensibly healthy individuals. In the gnomAD database, we identify 401 high-quality pLoF variants that pass both sequencing and annotation quality filters, in 61 haploinsufficient disease genes where heterozygous pLoF variants are established to cause severe developmental delay phenotypes with high penetrance. Given the severity of these phenotypes and their extremely low worldwide prevalence, ranging from 1 in 10,000 to less than 1 in a million, very few, if any true pLoF variants would be expected to be found in the gnomAD population. As such, most or all of these observed pLoF variants are likely to be errors [17]. However, to the extent that current quality filters are able to aid the interpretation of such variants, they appear to be high quality.

Materials and Methods

Datasets and code used in the study

We utilized the gnomAD v.2.1.1 sites Hail 0.2 (https://hail.is) table which is accessible publicly at gs://gnomad-public/release/2.1.1 and at https://gnomad.broadinstitute.org. The GTEx v7 gene and isoform expression data were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap) under accession

phs000424.v6.p1. The GTEx pipeline for isoform quantification is available publicly (https://github.com/broadinstitute/gtex-pipeline/) and briefly involves 2-pass alignment with STAR v2.4.2a [18], gene expression quantification with RNA-SeQC v1.1.8 [19], and isoform quantification with RSEM v1.2.22 [14]. The LOEUF constraint file was downloaded from gs://gnomad-resources/lof_paper/. Variants used in all gnomAD analyses in the manuscript passed random forest filtering, and all pLoF variants were annotated as high confidence (HC) by LOFTEE v.1.0, which is described in [1]. All files used in the analyses in the manuscript are available in gs://gnomad-public/papers/2019-tx-annotation/. Scripts to QC the gnomAD dataset are available at https://github.com/macarthur-lab/gnomad_qc and the scripts to generate files for the analyses are available at https://github.com/macarthur-lab/tx_annotation.

Curation of pLoF variants in haploinsufficient developmental disease genes

For identification of haploinsufficient developmental delay genes, we selected genes curated by the ClinGen Dosage Sensitivity Working Group [20]; 58 of the 61 genes had a score of 3 with sufficient evidence for pathogenicity, while two genes (*CHAMP1*, *CTCF*) had a score of 2 (some evidence) and one gene (*RERE*) was not yet scored. The penetrance of pathogenic variants in each gene was reviewed in the literature, and only genes with >75% reported penetrance were included. These conditions are those too severe to expect to see an individual in gnomAD (likely unable to consent for a study without guardianship). The 61 genes include 50 autosomal genes of high severity and high penetrance and 11 genes on chromosome X where the phenotype is expected

to be severe or lethal in males and moderate to severe in females. The resulting gene list is available at

gs://gnomad-public/papers/2019-tx-annotation/data/HI_genes_100417.tsv.

We extracted pLoF variants, defined as essential splice acceptor, essential splice donor, stop gained, and frameshift variants, identified in the 61 haploinsufficient disease genes from the gnomAD v2.1.1 [1] exome and genome sites tables, and considered only those pLoF variants that passed random forest filtering in the gnomAD dataset, and were annotated as high confidence (HC) by LOFTEE v1.0 [1]. We then performed manual curation of 401 pLoF variants using a web-based curation portal to identify any reason a pLoF may have been a variant calling or annotation error, and categorized the likelihood of each variant being a true LoF.

For manual curation, evidence to refute a true LoF variant was categorized into the following groups: mapping error, strand bias, reference error, genotyping error, homopolymer sequence, in-frame multi-nucleotide variant or frame-restoring indel, essential splice site rescue, minority of transcripts, weak exon conservation, last exon, and other annotation error. All possible reasons to reject a LoF consequence were flagged, even when a single criterion would categorize the variant as not LoF. Variants were then categorized as LoF, likely LoF, likely not LoF, and not LoF based on criteria outlined in Table 3.1.

Table 3.1 Summary of criteria for LoF verdicts of 401 pLoF in 61 haploinsufficient disease genes identified in gnomAD

LoF	Likely LoF	Likely not LoF	Not LoF
GQ > 99, absence of any evidence to refute a LOF consequence and one of two criteria met: AB>35 and read depth >15	Low complexity region	Re-initiation by downstream methionine in first coding exon	Minority of coding RefSeq transcripts (except when exon well conserved)
	Allele balance ≤ 35% but >25%	Mapping ambiguity (UCSC)	Variant falls in last coding exon
	QC ≤ 20	PhyloCSF weak	Weak conservation of exon
	GC rich region	In frame splice site rescue between 6 and 21 base pairs from the intron/exon boundary and validated by Alamut	Frame restoring indel
	Strand bias in regions where coverage is skewed towards a strand	Homopolymer repeat	In phase multi-nucleotide variant abolishing stop codon
	Read depth <15	Falls within terminal exon although will disrupt >25% but <50% of coding sequence	Complete splice site rescue within 6 base pairs and validated in Alamut
	Any single other transcript error	Minority of transcripts but exon well conserved	Complex mapping and assembly error
	Potential splice site rescue >21 BP away and weakly supported by Alamut	Combination of multiple flags (e.g. On 50% of RefSeq transcripts and partial loss of exon conservation; homopolymer repeat and mapping error etc.)	Reference error
	Splice rescue within 21 BP but very weak signal as per Alamut	Strand bias despite equal coverage of forward and reverse strand across region	Combination of multiple flags (e.g. in frame splice rescue within 21 bp and minority of transcripts with exon well conserved; mapping ambiguity and PhyloCSF weak etc.)
	Falls within terminal exon although will disrupt >50% of coding sequence		

A summary of the manual curation flags for the variants are available in Table 3.2. Technical errors comprised genotyping errors, strand biases, reference errors, and repetitive regions that could be detected by visual inspection of reads in the Integrative Genomics Viewer [21] (IGV) and from the UCSC genome browser [22]. Genotyping errors comprised skewed allele balances (conservative cutoff of ≤35%), low complexity sequences, GC rich regions, homopolymer tracts (≥ 6 base pairs or ≥ 6 trinucleotide repeats) and low quality metrics (genotype quality, or GQ, < 20). Strand bias was flagged when a variant was skewed preferentially on the forward or reverse strand, or when the majority (>90%) of a given strand covered a region; this was often observed

around intron/exon boundaries. Strand biases despite balanced coverage of the forward and reverse strands were weighted towards likely not LoF, whereas a strand bias due to skewed strand coverage was weighted alongside other genotyping errors. Reference errors were uncommon, but typically presented as by a small deletion in a coding exon, curated by GENCODE as a <5 base pair intron, which is biologically impossible used by GENCODE curators to restore and open reading frame. Most genotyping errors and strand biases in isolation were not deemed critical in deciding whether a variant was likely not LoF or not LoF, with the exception of allele balance ≤25%. Mapping errors were often identified by an enrichment of complex variation surrounding a variant of interest. Furthermore, the UCSC browser [22] was used to highlight mapping discrepancies, such as self-chain alignments, segmental duplications, simple tandem repeats, and microsatellite regions.

Table 3.2 Summary of manual curation flags for 401 pLoF in 61 haploinsufficient disease genes identified in gnomAD

Mapping error	Strand Bias	Reference error	Genotyping error	Homopolymer	MNV	Essential splice site rescue	Minority of transcripts	Weak exon conservation	Last exon	Other annotation error
Human self-chained repeats (UCSC)	Variant present on >90% of either forward or reverse strand	Reference contains small (<5bp) deletion within exon	Allele balance ≤ 35%	Repeat of ≥ 6 base pairs	In phase MNV that abolishes stop codon	In frame splice site rescue within 36 bp and validated by Alamut	Variant falls on <50% of coding NCBI RefSeq transcripts	Flagged by PhyloCSF as weak	Variant falls within the terminal coding exon	Variant falls on exactly 50% of NCBI coding RefSeq transcripts
Tandem repeats (UCSC)			GC rich region	Repeat of ≥ 6 trinucleotide repeats	Frame restoring indel			Entire exon is weakly conserved upon visual inspection in the UCSC browser	Variant falls within 50 base pairs of penultimate coding exon	Exon is partially conserved, specifically lacking conservation of variant/transcript
Segmental dups (UCSC)			GQ <20						Variants affects >25% coding sequence	Re-initiation by downstream methionine in first coding exon
Complex variation in region e.g. multiple indels, SNVs			Low complexity sequence							Stop codon occurs upstream of variant
			Low read depth <15							

In-frame multi-nucleotide variants (MNVs), essential splice site rescue, and frame-restoring insertion-deletions are rescue events that are predicted to restore gene function. MNVs were visualized in IGV and cross checked with codons from the UCSC browser; in-frame MNVs that rescued stop codons were scored as not LoF. Essential splice site rescue occurs when an in frame alternative donor or acceptor site is present nearby, predicted to result in only a small loss or gain of sequence from the transcript. Thirty-six base pairs upstream and downstream of the splice variant were assessed for splice site rescue. Cryptic splice sites within 6 base pairs of the splice variant were considered a complete rescue, rendering the variant not LoF. Rescue sites > 6 base pairs away but within +/- 20 base pairs were weighted with less confidence, scoring as likely not LoF. All potential splice site rescues were validated using Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/). Frame-restoring indels were identified by scanning approximately +/- 80 base pairs from the annotated indel and counting any insertions/deletions to assess if the frame would be restored.

Transcript errors encompass issues surrounding alternative transcripts, variants within a terminal coding exon, poorly conserved exons, and re-initiation events. Coding variants that occupied the minority (<50%) of NCBI coding RefSeq transcripts[23] for a given gene were considered not LoF. These variants often affected poorly conserved exons, as determined by PhyloP [24], PhyloCSF [25], and visualization in the UCSC browser [22]. The only exception to the minority of transcript criteria were cases where the exon was well conserved, which relegated the categorization to likely not LoF. Variants within the last coding exon, or within 50 base pairs of the penultimate coding exon were also considered not LoF, unless 25% < x <50% of the coding sequence was affected, in

which case the variant was deemed likely not LoF. If >50% of the coding sequence was disrupted by a variant in the last exon, this was deemed likely LoF. Other transcript errors included: re-initiation errors; upstream stop codons of a given LoF variant; variants that fell on exactly 50% of coding RefSeq[23] transcripts; and/or partial exon conservation. Re-initiation events were flagged when a methionine downstream of the variant in the first coding exon was predicted to restart transcription, and were predicted to be likely not LoF. Variants occurring after a stop codon in the last coding exon were considered not LoF, particularly across the region of the exon or transcript in question. Error categories were grouped for a summarized figure as follows: Minority of transcripts and weak exon conservation were grouped as transcript errors, genotyping errors and homopolymers as sequencing errors, essential splice rescue and MNV grouped as rescue and strand bias was included in other annotation errors.

The criteria above were strictly adhered throughout and manual curation was performed by two independent reviewers to ensure maximum consistency and minimize human error. Any discordance in curation was re-curated by both curators together and resolved. Full results of manual curation are available in gs://gnomad-public/papers/2019-tx-annotation/results.

Calculation of transcript-expression aware annotation

We first imported the GTEx v7 isoform quantifications into Hail (hail.is) and calculated the median expression of every transcript per tissue. This precomputed summary isoform expression matrix is available for GTEx v7 in gs://gnomad-public/papers/2019-tx-annotation/data/. We also import and annotate a variant file with

the Variant Effect Predictor (VEP) version 85 [26] against Gencode v19 [27], implemented in Hail with the LOFTEE v1.0 plugin.

We use the transcript consequences VEP field to calculate the sum of isoform expression for variant annotations, i.e. the annotation-level *expression across transcripts* (ext). For variants that have multiple consequences for one transcript (for example, a SNV that is both a missense and a splice region variant on one transcript) we use the worst consequence, ordered by VEP (in this example, missense takes precedence over splice region). We filter the consequences to those only occurring on protein coding transcripts. Full ordering of the VEP consequences is available at: https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html).

We then sum the expression of every transcript per variant, for every combination of consequence, LOFTEE filter, and LOFTEE flag for every tissue (Figure 3.2A). For example, if a SNV is synonymous on ENST1, a LOFTEE HC stop-gained on ENST3 and ENST4, and LOFTEE low-confidence (LC) stop gained variant on ENST 5 and ENST6, the ext values will be synonymous: ENST1, stop-gained HC: ENST 3 + ENST4, and stop-gained LC: ENST5 + ENST6 per tissue. This can be computed with the tx_annotate() function by setting the tx_annotation_type to “expression”. We foresee the non-normalized ext values to be useful when only considering one tissue of interest.

To allow for taking average expression values across tissues of interest, we normalize the expression value for a given value to the total expression of the gene on which the variant is found. This is carried out by dividing the ext value with the median gene expression value per tissue in transcripts-per-million (TPM) from RNASEQC v1.1.8 [19] (Figure 3.2B), which is publicly available via gtexportal.org. The resulting

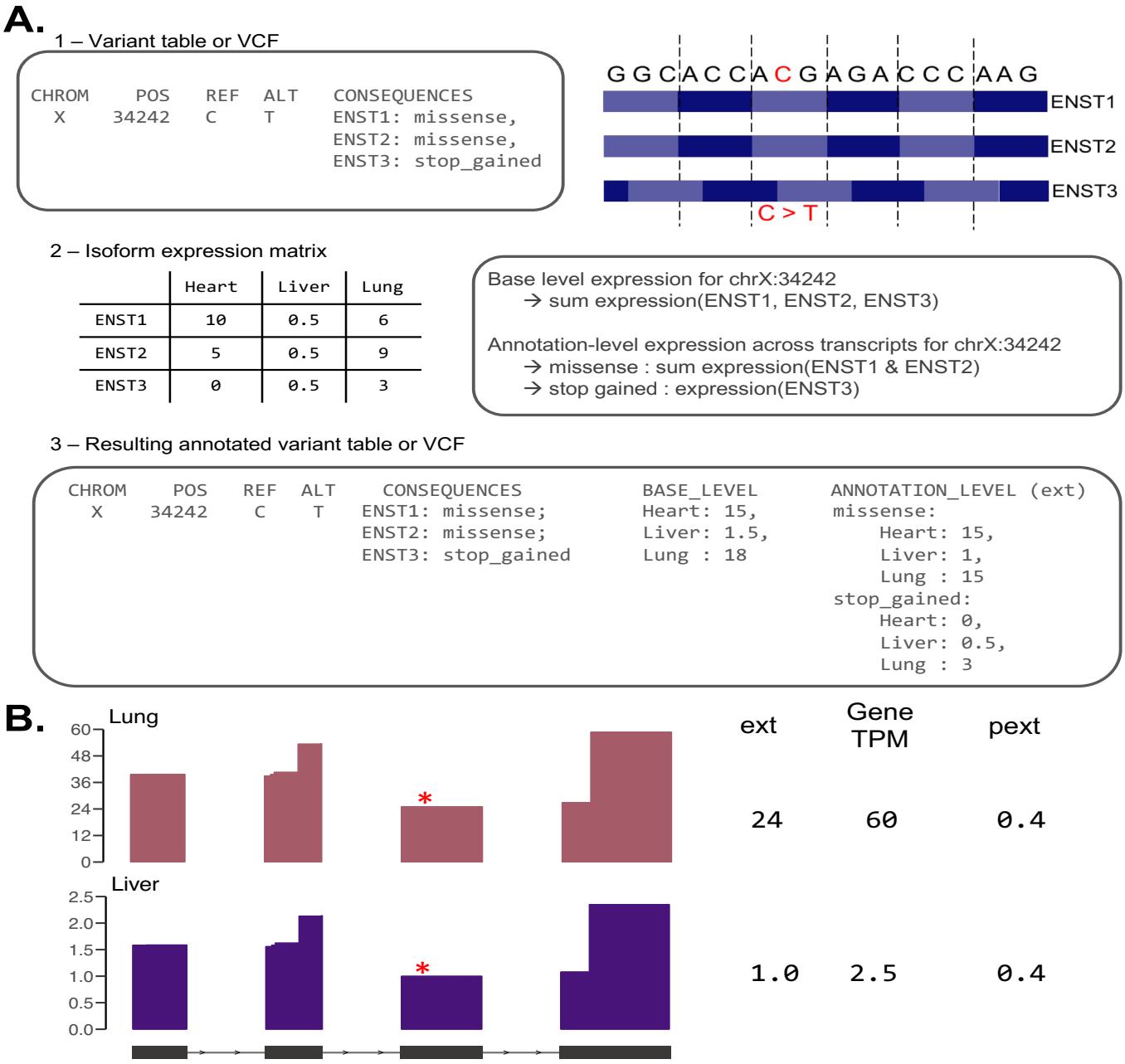


Figure 3.2 Details of calculating transcript-expression annotation. A. A SNV can have different consequences across annotate transcripts. For example, an SNV on a region with three annotated transcripts, can have a missense effect on two transcripts and a nonsense effect on one transcript. The base-level expression, mainly to be utilized for quick visualization of variant expression in genes, is calculated as the sum of the three transcripts. The annotation level expression across transcript (ext) metric defines the expression of a variant as the sum of the expression of transcripts on which an annotation exists. In this example, the expression value for the missense variant will be the sum of the expression on transcripts where the variant is a missense (ENST1 and ENST2) and the value for the nonsense will the the sum of the expression of transcripts where the variant is a nonsense (ENST3). **B.** To account for gene expression differences between tissues, we normalize the ext value by the gene TPM in the tissue to calculate the proportion expressed across transcript (pext) value used in the manuscript. This allows for combining pext values across tissues to for example, get the mean pext value across GTEx.

pext (proportion expression across transcript) value can be interpreted as the proportion of the total transcriptional output from a gene that would be affected by the variant annotation in question. If the gene expression value (and thus the denominator) in a given tissue is 0, the pext value will not be available for that tissue. When taking averages across tissues, such unavailable pext values are not considered (ie. when taking the mean across tissues, we remove NAs). This value can be computed with the tx_annotate() function by setting the tx_annotation_type to “proportion”. For the analyses in this manuscript, we remove reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes, vagina), cell lines (transformed fibroblasts, transformed lymphocytes) and any tissue with less than one hundred samples (bladder, brain Cervicalc-1 spinal cord, brain substantia nigra, kidney cortex, minor salivary gland) resulting in the use of 38 GTEx tissues.

The full transcript-expression aware annotation pipeline, implemented in Hail 0.2, is fully available at https://github.com/macarthur-lab/tx_annotation with commands laid out for analyses in the manuscript. Passing a Hail table through the tx_annotate() function returns the same table with a new field entitled “tx_annotation” which provides either the ext or pext value per variant-annotation pair, depending on parameter choice. We provide a helper function to extract the worst consequence and the associated expression values for these annotations. All analyses in the manuscript are based on the worst consequence of variant, ordered by VEP [26].

Functional validation of transcript-expression aware annotation

Conservation analysis was performed using phyloCSF [25] scores using the same file utilized for the LOFTEE plugin, available publically in gs://gnomad-public/papers/2019-tx-annotation/data/phylocsf_data.tsv.gz . We denoted exons with a phyloCSF max open reading frame (ORF) score > 1000 as highly conserved and those with phyloCSF max ORF score < -100 as lowly conserved and evaluated their average usage in GTEx.

Using the base-level pext values that are used in the gnomAD browser, we filtered to intervals with high or low conservation, and calculated the average pext value in the interval. To evaluate regions with low conservation but high expression, we identified genes harboring unconserved regions with the pext value > 0.9 for pathway enrichment analysis and used the web browser for FUMA GENE2FUNC feature [28], which incorporates Reactome [29], KEGG [30], Gene Ontology [31] (GO) as well as other ontologies.

Analysis of pext values for LOFTEE flags and the MAPS calculation were performed utilizing the gnomAD v2.1.1 exome dataset. Calculation of MAPS scores was previously described in Lek et al. 2016 [32] and is implemented as a Hail module, as described in Karczewski et al. 2019 [1]. MAPS is a relative metric, and so cannot be compared across datasets, but is a useful summary metric for the frequency spectrum, indicating deleteriousness as inferred from rarity of variation (high values of MAPS correspond to lower frequency, suggesting the action of negative selection at more deleterious sites). The MAPS scores were calculated on the gnomAD v.2.1.1 dataset

partitioning upon the LOEUF score and expression bin. The script for generating MAPS scores is available at https://github.com/macarthur-lab/tx_annotation/blob/master/analyses/maps_submit_per_class.py

Manual evaluation of unexpressed regions in haploinsufficient developmental delay genes using the GENCODE workflow

As an orthogonal evaluation of regions flagged as unexpressed with the pext metric, we identified any region in 61 haploinsufficient disease genes with a pext value < 0.1 in all GTEx tissues and in GTEx brain samples, due to the relevance of brain tissues for these disorders, regardless of mutational burden in gnomAD. The resulting list of 128 regions was evaluated by the HAVANA manual annotation group of the GENCODE project [27].

The manual evaluation first established whether the transcript model corresponding to the region in question was correct in terms of structure, comparing exon / intron combinations, and the accuracy of splice sites against the RNA evidence supporting the model. Second, the functional biotype of each model was reassessed; in particular, whether the decision to annotate the model as protein-coding in GENCODE v19 was appropriate. Note that GENCODE models that incorporate alternative exons or exon combinations in comparison to the 'canonical' isoform are likely to be annotated as coding if they contain a prospective CDS that is considered biologically plausible, based on a mechanistic view of translation.

We binned cases into three main categories, according to confidence in both the accuracy and potential functional relevance of the overlapping models: (1) 'error', where the model was seen to have an incorrect transcript structure and/or a CDS that conflicted with updated GENCODE annotation criteria (these annotations had been or will be changed in future GENCODE releases based on this evaluation); (2) 'putative', where the model structure and CDS satisfied our current annotation criteria, although we judged the potential of the transcript represented to encode a protein with a functional role in cellular physiology to be nonetheless speculative (these have been maintained as putative protein-coding transcripts in GENCODE); (3) 'validated', where we believe it is highly probable that the model represents a true protein-coding isoform. High confidence in the validity of the CDS was based on comparative annotation, i.e. the observation of CDS conservation and also the existence of equivalent transcript models in other species. GENCODE also annotates transcript models as 'nonsense-mediated decay (NMD) and 'non-stop decay' (NSD), where a translation is found that is predicted to direct the RNA molecule into cellular degradation programs. While it has been established that such 'non-productive' transcription events can play a role in gene regulation and thus disease, the interpretation of variants within NMD and NSD CDS remains challenging [33]. These models were therefore classed in a separate category.

Gene list comparisons

To evaluate the filtering power of the pext metric for Mendelian variants, we evaluated the number of variants that would be filtered with an average GTEx pext cutoff of 0.1 (low expression) in the ClinVar[34] and gnomAD datasets. We downloaded

the ClinVar VCF from the ClinVar FTP (version dated 10/28/2018), imported it into Hail, annotated it with VEP v85 against Gencode v19, and added pext annotations with the tx_annotate() function. All evaluated variants were annotated as HC by LOFTEE v1.0, and ClinVar variants were filtered to those marked as pathogenic, with no conflicts, and reviewed with at least one star status.

For variants in 61 haploinsufficient genes, we identified any variant identified in at least one individual with any zygosity in both datasets. For variants identified in autosomal recessive disease genes, we used a list of 1,183 OMIM disease genes deemed to follow a recessive inheritance pattern by Blekhman et al. [35] and Berg et al. [36] (available as https://github.com/macarthur-lab/gene_lists/blob/master/lists/all_ar.tsv). We compared the pext value for all pLoF variants identified in ClinVar versus any variant in a homozygous state in at least one individual in the gnomAD exome or genome datasets. Finally, we used a LOEUF cutoff of 0.35 to denote constrained genes, and compared any synonymous or pLoF variant in these genes in the gnomAD exome or genome datasets.

De novo and rare variant analysis

De novo variants were collated from previously published studies. We collected *de novo* mutations identified in 5,305 probands from trio studies of intellectual disability/developmental disorders (Hamdam et al [37]: n = 41, de Ligt et al [38]: N = 100, Rauch et al [39]: N = 51, Deciphering Development Delay Study [40] : n = 4,293, Lelieveld et al [41]: n = 820), 1,073 probands with congenital heart disease with co-morbid developmental delay (Sifrim et al [42]: n = 512, Chih Jin et al [43]: 561), 6,430 ASD

probands, and 2,179 unaffected controls from the Autism Sequencing Consortium[44]. We also utilized a previously published dataset of variants in 8,437 cases with ASD and/or attention-deficit/hyperactivity disorder and 5,214 controls from the Danish Neonatal Screening Biobank [45]. In this analysis, we analyzed pLoF variants identified in highly constrained genes (first LOEUF decile) with a combined total allele count of ≤10 in cases and controls.

We annotated both *de novo* and rare variants with VEP v85 against Gencode v19 and added pext annotations with the tx_annotate() function. We then calculated the average pext metric across 11 GTEx brain samples and binned them as low (pext < 0.1), medium (0.1 ≤ pext ≤ 0.9) or high (pext > 0.9) expression. We then calculated the number of pLoF, missense, and synonymous variants per pext expression bin. To obtain case-control rate ratios and the 95% confidence intervals for *de novo* variant analyses, we used a two-sided Poisson exact test on counts [46]. To obtain the odds ratio for the rare variant analysis in ASD/ADHD, we used the Fisher's exact test for count data.

Isoform quantifications via salmon

To evaluate whether use of a different isoform quantification tool would affect results, we compared the results of *TCF4* base-level expression, MAPS and comparison of the number of variants filtered in haploinsufficient developmental disease genes in ClinVar vs gnomAD using RSEM quantifications used in this study with quantifications using salmon v.0.12 [16]. Due to the intractability of re-quantifying the entire GTEx dataset, we downloaded and requantified 151 GTEx brain – cortex CRAM

files from the V7 dataset. We first converted CRAMs to fastq files using Picard 2.18.20 and ran salmon with the "salmon quant –i index -fastq1 – fastq2 –minAssignedFrag1 –validateMappings" command. The index was created with the "salmon index –t transcript.fa –type quasi –k 31" command using the GENCODE v19 protein-coding and lncRNA transcripts FASTA files. The existing GTEx RSEM isoform quantifications were filtered to the same GTEx brain – cortex samples. The WDL script for the quantification pipeline is available at : gs://gnomad-public/papers/2019-tx-annotation/results/salmon_rsem/salmon.wdl and the commands to obtain results for each individual analysis at https://github.com/macarthur-lab/tx_annotation/blob/master/analyses/rsem_vs_salmon.py.

Transcript expression aware annotation with a fetal isoform expression dataset

While our analyses were based on transcript expression aware annotation from the GTEx v7 dataset, we provide necessary files for pext annotation with the Human Brain Development Resource (HBDR) fetal brain dataset [47] in gs://gnomad-public/papers/2019-tx-annotation/data/. HBDR includes 558 samples from varying brain subregions across developmental time points. We downloaded HDBR sample fastq files from European Nucleotide Archive (study accession PRJEB14594) and obtained RSEM isoform quantification on HBDR fastqs using the GTEx v7 quantification pipeline, publicly available at https://github.com/broadinstitute/gtex-pipeline/) which briefly involves 2-pass alignment with STAR v2.4.2a [18] and and isoform quantification with RSEM v1.2.22 [14]. The dataset was used for the analysis of baselevel expression values in *SCN2A* in Figure 3.9D (see below). The commands to obtain the results is available

in https://github.com/macarthur-lab/tx_annotation/blob/master/analyses/get_scn2a_baselevel_fetal.py

Results

Contribution of alternative splicing to pLoF annotation

We find that isoform diversity is a contributor to the paradoxical finding of disruptive variants in dosage-sensitive disease genes in ostensibly healthy individuals. Manual curation of these 401 pLoF variants in severe haploinsufficient disease genes reveals common error modes that result in likely misannotation of pLoFs, with diversity of transcript structure, mediated by alternative mRNA splicing, emerging as a major consideration (Figure 3.3 and Figure 3.4). Specifically, we curate 240 out of 401 pLoF variants in haploinsufficient genes as not LoF, followed by 66 variants as likely not LoF. The remaining 95 variants were broken down as 41 being likely LoF and 43 as LoF. We further analyzed the flags in the 306 variants categorized as not LoF or likely not LoF. Allowing for multiple flags per variants, we find that variants found on weakly conserved exons, which are often enriched for false exon annotations, and those occurring on a minority of transcripts (see Methods) to be the most common error mode, with 190 of the variants carrying the flags. This was followed by 142 variants flagged as a genotyping error, and 104 as mapping error (Figure 3.4). This indicates that correct transcript annotation can be vital for correct interpretation of variant functional effect. However, no existing tools systematically incorporate information on transcript expression into variant interpretation.

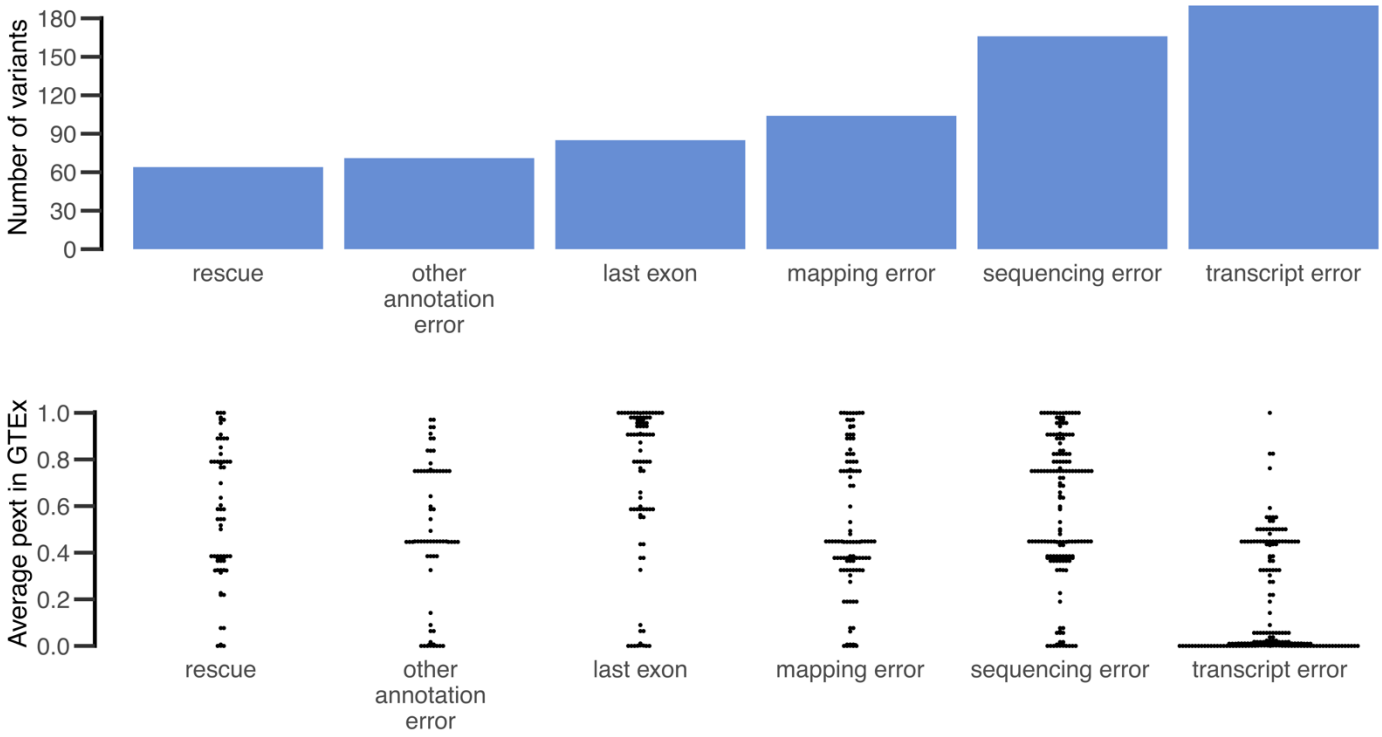


Figure 3.3 Curation of pLoF variants in haloinsufficient disease genes found in gnomAD reveals transcript errors as a major confounding error mode in variant annotation. We identified and manually curated 401 pLoF variants in the gnomAD dataset in 61 haploinsufficient severe developmental delay genes and flagged any reason the pLoF may not be a true LoF variant. Top plot shows the frequency of each error mode present in the 306 variants classified as unlikely to be a true LoF. Transcript errors emerge as a major putative error mode in the annotation of these pLoF variants. Beeswarm plot on bottom shows the average pext score across GTEx tissues presented in the manuscript for each variant in the error categories. This shows that pext values are discriminately lower for variants that are annotated as possible transcript errors.

Development of transcript-expression aware annotation

We utilized isoform-level quantifications from 11,706 tissue samples from the GTEx v7 dataset to derive an annotation-specific expression metric. For each tissue, we annotate each variant with the expression of every possible consequence across all transcripts, which can be used to summarize expression in any combination of tissues of interest. We first compute the median expression of a transcript across tissue samples, and define the expression of a given variant as the sum of the expression of all transcripts for which the

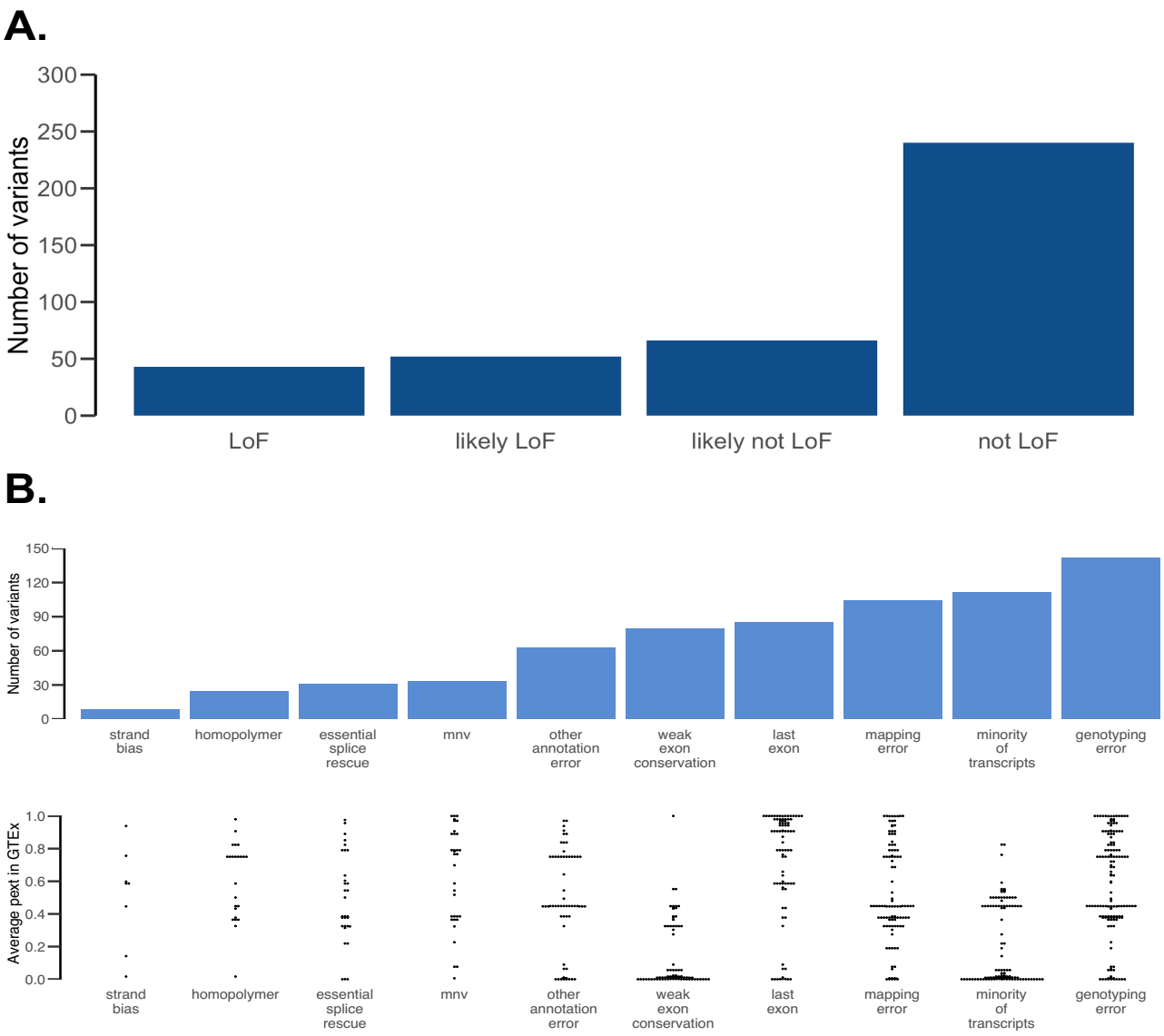


Figure 3.4 Details of manual curation of 401 pLoF variants in 61 haploinsufficient developmental disease genes A. Distribution of curation verdicts for the 401 pLoF variants. We categorized 240 variants (76%) as not being LoF, 66 as likely not LoF, 52 as likely LoF and 43 as LoF **B.** Full distribution of the flags refuting true LoF status for 306 not LoF and likely not LoF variants (top) and their corresponding pext score in GTEx (bottom). A variant with multiple flags is assigned to each flag as in Figure 1 (ie. double counted). Minority of transcripts and weak exon conservation were grouped as transcript errors, genotyping errors and homopolymers grouped as sequencing errors, essential splice rescue and MNVs grouped as rescue and strand bias was included in other annotation errors. While the pext values are randomly distributed for other error modes, they are enriched for lower values in transcript errors. Criteria for curation for each verdict and flag in Tables 3.1, 3.2, respectively.

variant has the same annotation (Figure 3.2A, Figure 3.5). By normalizing the expression of the annotation to the total gene expression, we define a metric (proportion expression across transcripts, or pext), which can be interpreted as a measure of the proportion of the total transcriptional output from a gene that would be affected by the variant annotation in question (Figure 3.2B).

Gene-based visualization of the pext score

The pext metric allows for quick visualization of the expression of exons across a gene. Figure 3.5B shows *TCF4,* a haploinsufficient gene in which heterozygous variants result in Pitt-Hopkins syndrome [48] a highly penetrant disorder associated with severe developmental delay. This gene harbors 20 unique high quality pLoF mutations across 56 individuals in the gnomAD database. All 20 variants lie on exons with no evidence of expression across the GTEx dataset (Figure 3.5B) indicating that functional TCF4 protein can be made in the presence of these variants. This visualization is now available for all genes in the gnomAD browser (gnomad.broadinstitute.org), and can aid in rapid identification of variants occurring on exons with little to no evidence of expression in GTEx.

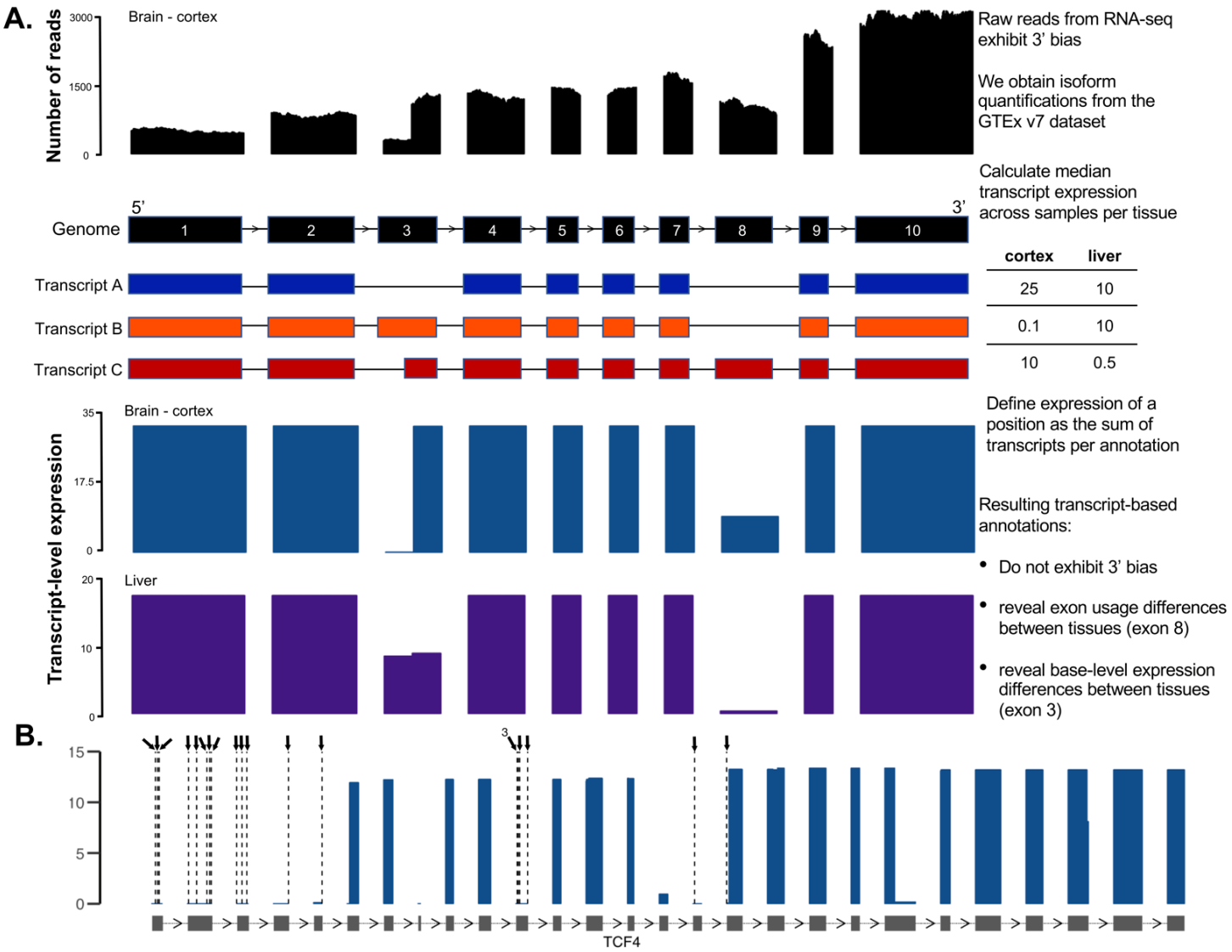


Figure 3.5 Summary of transcript-expression based annotation method A. Overview of transcript aware annotation. Most genes have many annotated isoforms, which can have varying expression patterns across tissues. Utilizing number of reads aligning to exonic regions in transcriptome datasets as a proxy for exon expression (top panel black) has confounding effects, due to 3' bias. In this example, while exon 3 and 8 have markedly different expression levels in Brain – Cortex, the number of reads aligning to the two exons are similar, masking exon usage differences. Transcript-aware annotation defines the expression of every variant as the sum of transcripts that have the same annotation. The resulting transcript-level expression plots do not exhibit 3' bias, and reveal exon usage differences, such as those in exons 3 and 8, across tissues. **B.** Example of utility of transcript-expression based annotation. There are 20 high quality pLoF variants in the haploinsufficient developmental delay gene *TCF4* in gnomAD, annotated as dashed lines and arrows. All 20 variants have no evidence of expression in the GTEx dataset, suggesting functional TCF4 protein can be made in the presence of these variants.

Functional validation of pext with conservation

To explore whether expression-based annotation marks functionally important regions, we compared the distribution of the pext metric in evolutionarily conserved and unconserved regions using phyloCSF [25]. Exons with patterns of multi-species conservation consistent with coding regions have higher phyloCSF scores, and should exhibit detectable expression patterns, whereas regions with lower scores will be enriched for incorrect exon annotations, which are expected to have little evidence of expression in a population transcriptome dataset. As expected, we observe significantly lower expression for unconserved regions, and near-constitutive expression in highly conserved regions (Figure 3.6A and Figure 3.7A). This difference remains statistically significant after correcting for exon length (logistic regression $p < 1.0 \times 10^{-100}$), which can influence both phyloCSF scores and isoform quantifications, indicating that transcript expression-aware annotation marks functionally relevant exonic regions.

While the metrics are associated, we find that pext provides orthogonal information to conservation for variant interpretation. For example, regions with low evidence of conservation but high expression (in Figure 3.6A) are enriched for genes in immune-related pathways (Figure 3.8), which are selected for diversity but represent true coding regions. In addition, the pext value is higher for pLoF variants annotated as high confidence (HC) by the Loss of Function Transcript Effect Estimator [1] (LOFTEE) with no additional flags than those flagged as having found on unlikely open reading frames or weakly conserved regions (Figure 3.6B, Figure 3.7B). However, LOFTEE-HC variants with no flags can also have low pext values, suggesting transcript-expression

aware annotation adds additional information to the currently available interpretation toolkit.

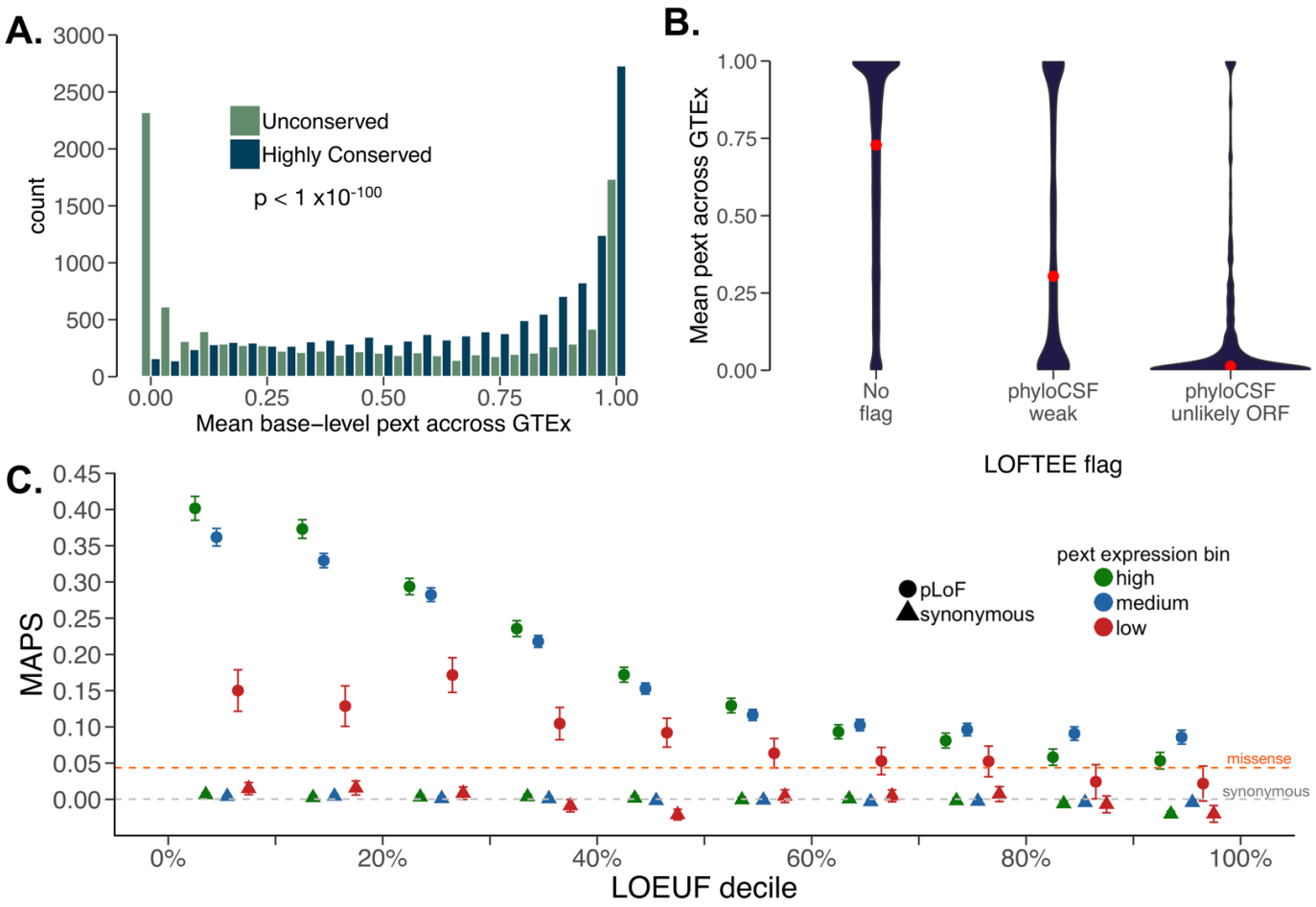


Figure 3.6 Functional validation of transcript-expression based annotation A. We define highly conserved and unconserved regions and compared the expression status of these regions across GTEx. Highly conserved regions are enriched for having near-constitutive expression whereas unconserved regions are enriched for having little to no usage across GTEx. This difference is significant after correcting for gene length (logistic regression p value < 1 x 10^{-100}). We note that unconserved regions with high levels of expression (pext > 0.9) are enriched for immune-related genes, which are selected for diversity and thus have low conservation, but represent true coding regions. **B.** Transcript-expression based annotation recapitulates, and adds information to, existing interpretation tools. LOFTEE-HC pLoF variants in gnomAD with no flags are enriched for higher pext values, whereas HC pLoF variants falling on low phyloCSF or unlikely ORF regions are enriched for low expression. However HC-pLoF variants can also be filtered based on a low pext score. Red dots represent median pext value across GTEx **C.** Nonsynonymous variants found on near-constitutive regions tend to be more deleterious. We compared the mutability adjusted proportion singleton (MAPS) score for variants with low (<0.1), medium (0.1 ≤ pext ≤ 0.9) and high (pext > 0.9) expression. Variants with near-constitutive expression have a higher MAPS score, indicating higher deleteriousness than those with little to no evidence of expression. Dashed grey and orange line represent MAPS values for all gnomAD missense and all synonymous variants, respectively.

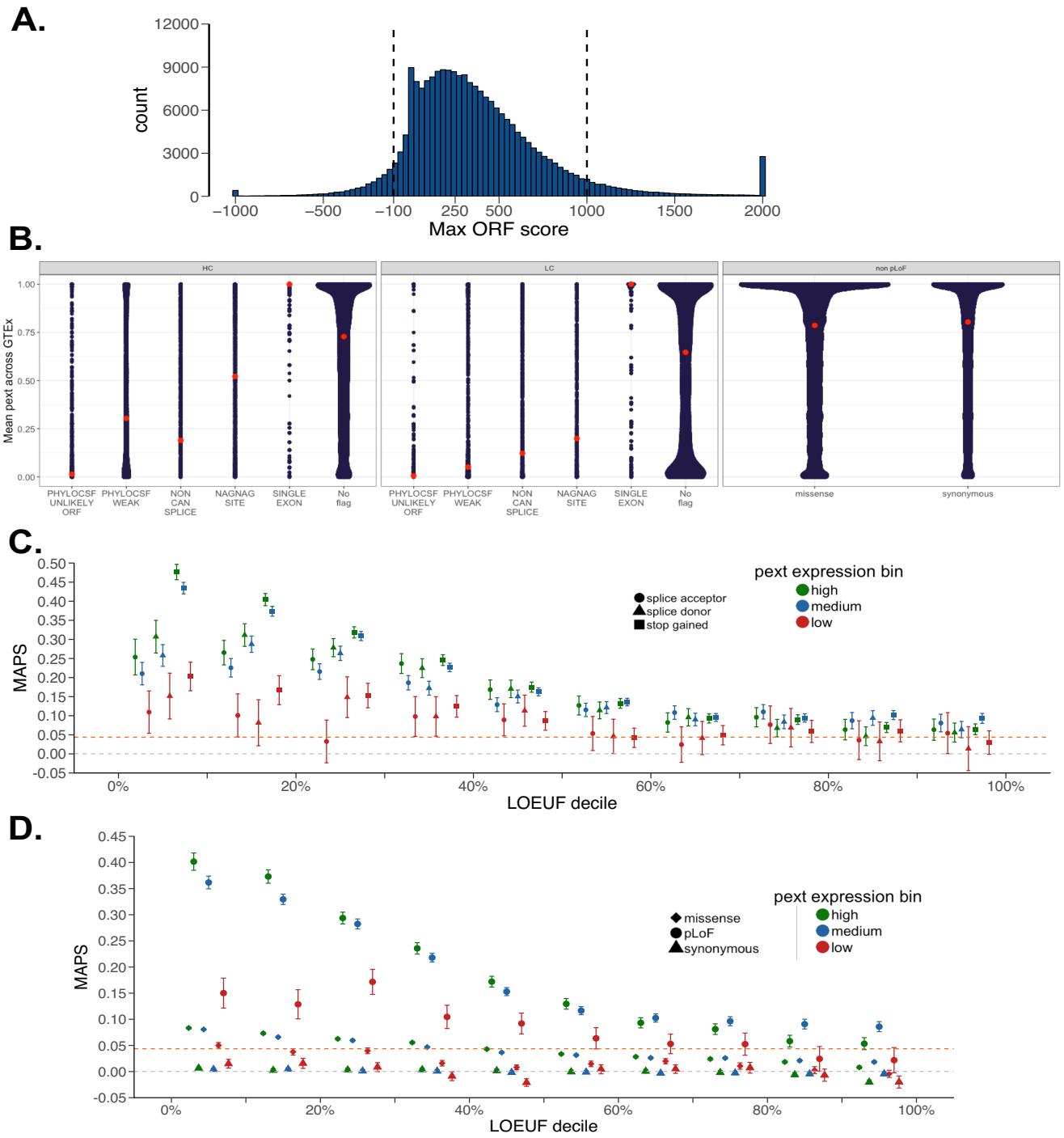


Figure 3.7 Functional validation of pext A. Distribution of max ORF scores from phyloCSF across the genome. We denoted exons with a maximum phyloCSF open reading frame (ORF) score > 1000 as highly conserved and those with maximum phyloCSF ORF score < -100 as unconserved. Max ORF scores were capped at -1000 and 2000 for plotting. **B.** Sina plots of pext distribution in all gnomAD exome variants, partitioned on LOFTEE flags and filters (filters denoted as gray bars above plots). Red dots denote median average pext value per category **C.** MAPS score for pLoF variants broken down by specific pLoF consequence shows consistent differences in MAPS for each pLoF category between high, medium and low pext expression bins. **D.** MAPS score including missense variants shows consistent skew between variants found on high (>0.9) , medium (0.1≤ x ≤ 0.9) and low (<0.1) average GTEx pext expression bins.

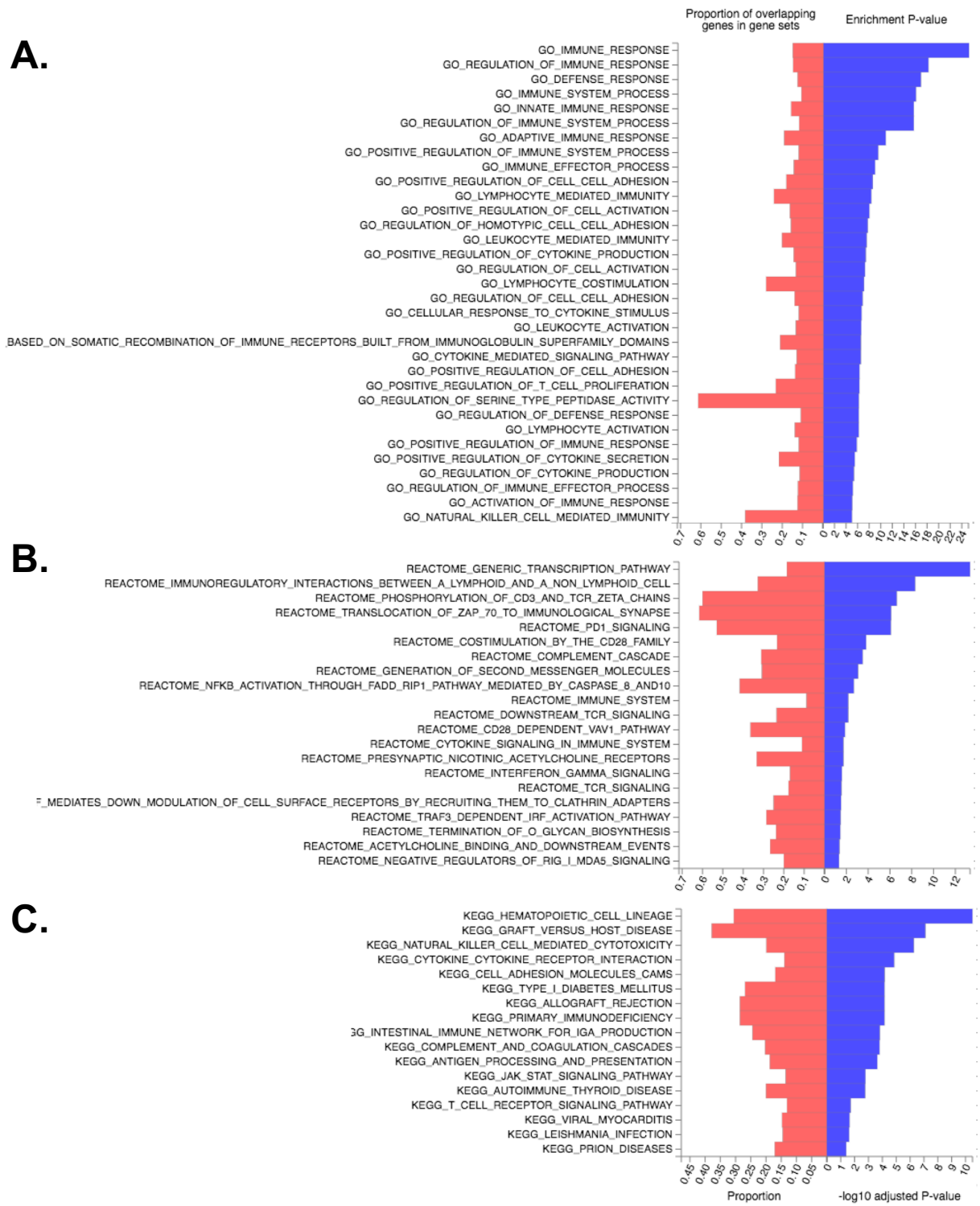


Figure 3.8 Results from FUMA GENE2FUNC analysis in unconserved regions with high expression values We ran pathway analysis on 1,310 genes harboring 2,414 regions with low conservation (phyloCSF < -100) but high expression (pext > 0.9) shown in Figure 3A using the FUMA GENE2FUNC web browser. Results from **A)** Gene Ontology Biological Processes **B)** Reactome pathways and **C)** KEGG pathways show that these regions are enriched for immune pathways, which are selected for diversity but represent true coding regions, emphasizing the orthogonal information provided by pext over conservation alone.

Manual evaluation of low pext regions in haploinsufficient genes using GENCODE standards

We undertook manual evaluation of 128 regions marked as unexpressed (pext < 0.1 in all tissues and in GTEx brain) in 61 haploinsufficient genes following the GENCODE manual annotation workflow [27] to evaluate the annotation quality in these coding sequence (CDS) regions. A third of flagged regions were associated with low quality models that have been removed or switched to non-coding biotypes in subsequent GENCODE releases (Figure 3.9A-C) while 70% of the remaining regions correspond to models that satisfy only minimum criteria for inclusion in the gene set, corresponding to 'putative' annotations that lack markers for CDS functionality. Nonetheless, we find support for some highly conserved CDS', several of which show evidence of transcription in fetal tissues, underlining the importance of incorporating multiple isoform expression datasets for interpretation (Figure 3.9D).

Stratifying the mutability adjusted proportion singleton (MAPS) score with pext

Nonsynonymous variants found on constitutively expressed regions would be expected to be more deleterious than those on regions with no evidence of expression. To test this, we defined expression bins based on the average pext value across GTEx tissues where an average pext value less than 0.1 was defined as low (or unexpressed), above 0.9 as high (or near-constitutive) and intermediate values as medium expression. We compared the mutability-adjusted proportion singleton (MAPS), a measure of negative selection on variant classes [32], partitioned on the LoF Observed Upper-bound Fraction (LOEUF) decile, a measure of constraint against pLoF variants in

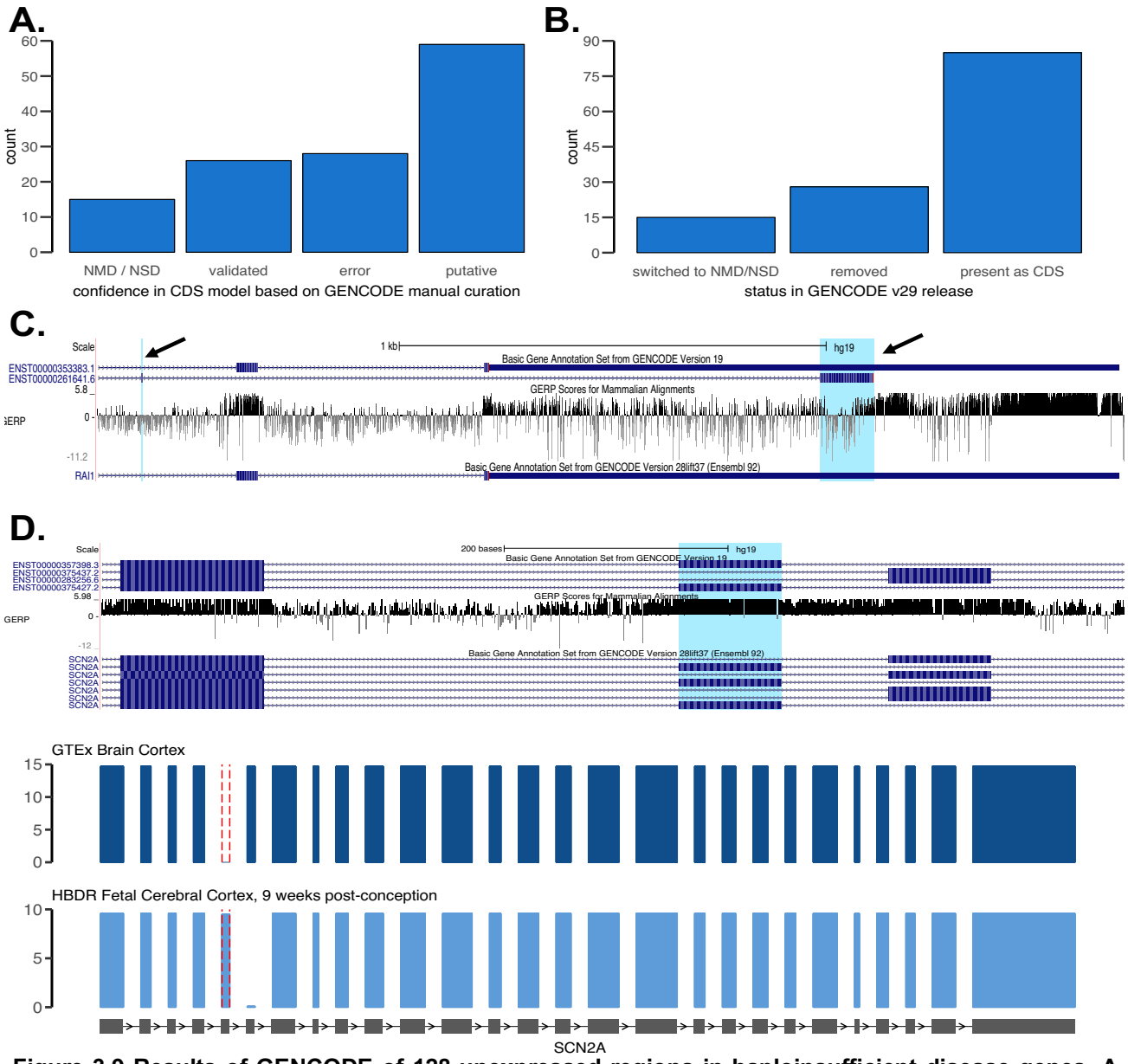


Figure 3.9 Results of GENCODE of 128 unexpressed regions in haploinsufficient disease genes. A. Summary of confidence in the CDS models tagged as unexpressed in GTEx based on expert manual evaluation. The major curation mode was putative annotation, where regions meet minimal annotation criteria but the coding potential of the region remains speculative. This was followed by regions that were marked as errors and non-coding regions, and have since been removed or are marked for removal based on this analysis. **B.** Summary of current annotation status of the regions in GENCODE v29. While some regions have been removed, or have been switched to a noncoding biotype, a majority remain in subsequent annotation sets **C.** An example of two erroneous gencode v19 CDS' in *RAI1* (chr17:17712481-17712483 and chr17:17714069-17714194, highlighted in blue) flagged by pext as unexpressed. The regions exhibit poor conservation and represents an incorrectly computationally predicted microexon and it's downsteam CDS, likely due to a poor quality cDNA alignment **D.** An example of a likely-coding CDS in *SCN2A* (chr2:166165675-166165766) which is well-conserved (gene model on top). While the region is unexpressed in GTEx, it exhibits considerable expression in fetal tissues from the Human Brain Developmental Resource (shown on bottom), highlighting the importance of incorporating multiple isoform datasets for accurate interpretation.

the gnomAD dataset [1] in each of these expression bins. MAPS scores differed substantially between pLoF variants found on low-expressed and high-expressed regions in genes intolerant to pLoF variation (Figure 3.6C and Figure 3.7C, D). This skew of nonsynonymous variation in high-expressed regions suggests that variation arising in such exons tends be more deleterious, whereas nonsynonymous variants on regions with low expression are similar to missense variants in their inferred deleteriousness.

Use of pext can aid Mendelian variant interpretation

To evaluate the utility of transcript expression-based annotation in Mendelian variant interpretation, we assessed the number of variants that would be filtered based on a pext cutoff of <0.1 (low expression) across GTEx tissues for three gene sets. Firstly, we evaluated high-quality pLoF variants in the 61 manually curated haploinsufficient genes in gnomAD and ClinVar [34]. The low pext expression bin resulted in filtering of 22.4% of pLoF variants in haploinsufficient developmental delay genes in gnomAD, but only 3.8% of high quality pathogenic variants in ClinVar (Figure 3.10A; $p = 5.2 \times 10^{-34}$; Methods). We next compared pLoF variants in autosomal recessive disease genes found in a homozygous state in at least one individual in gnomAD and any pLoF variant in these genes in ClinVar and observed similar results: expression-based annotation filters 35% of variants in gnomAD while only filtering 3.8% of variants in ClinVar (Figure 3.10B; $p = 2.3 \times 10^{-81}$).

Finally, we evaluated gnomAD pLoF variants in genes that are constrained against pLoF variation [1] (LOEUF score < 0.35). Given that these genes are depleted for

loss-of-function variation in the general population, we expect the observed pLoF variants in these genes to be enriched for annotation errors. We compared the proportion filtered to synonymous variants in the same genes, which we expect to be neutrally distributed. Our metric removes 17.2% of pLoF variants in high pLI genes, but only 5.6% of synonymous variants (Figure 3.10C; $p < 1.0 \times 10^{-100}$).

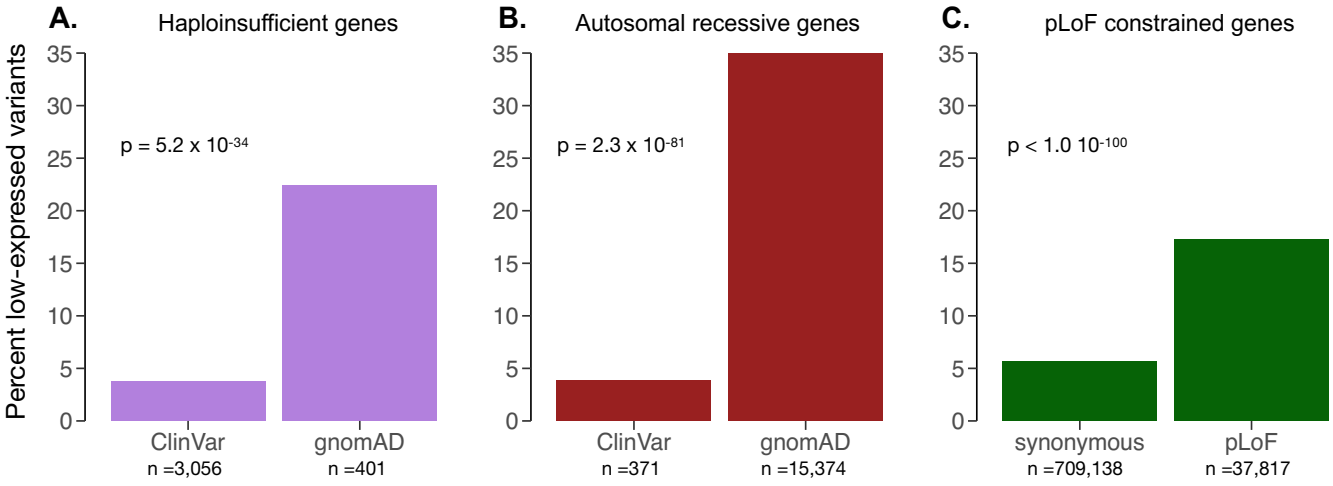


Figure 3.10 Transcript-expression based annotation aids Mendelian variant interpretation. A. Comparison of the proportion of high quality pLoF variants filtered in a curated list of 61 haploinsufficient developmental delays genes in gnomAD vs ClinVar with a cutoff of average pext across GTEx ≤ 0.1 (low expression). Expression-based filtering results in removal of 22.4% of gnomAD pLoFs and less than 3.8% of a curated high-confidence set of pLoFs in ClinVar. **B**. Expression-based annotation filters 35% of pLoF variants found in gnomAD in a homozygous state in at least one individual, and 3.8% of any pLoF variants found in the same genes in ClinVar. **C**. We extended this filtering approach to pLoF and synonymous variants in gnomAD pLoF-intolerant genes (defined by LOEF < 0.35). This filters 17.2% LoF and 5.6% of synonymous variants. Numbers below bar plots indicate the total number of high quality variants considered in each group. For pLoFs only LOFTEE-HC variants were considered, p-values calculated from fisher's exact test for counts.

Use of pext can improve power in gene burden testing analyses

To explore the benefits of this approach for rare variant analyses, we applied pext binning to burden testing of *de novo* variants in patients with developmental delay / intellectual disability or autism spectrum disorder using a set of 23,970 *de novo* variants collated from several studies including the Deciphering Developmental Disorders (DDD)

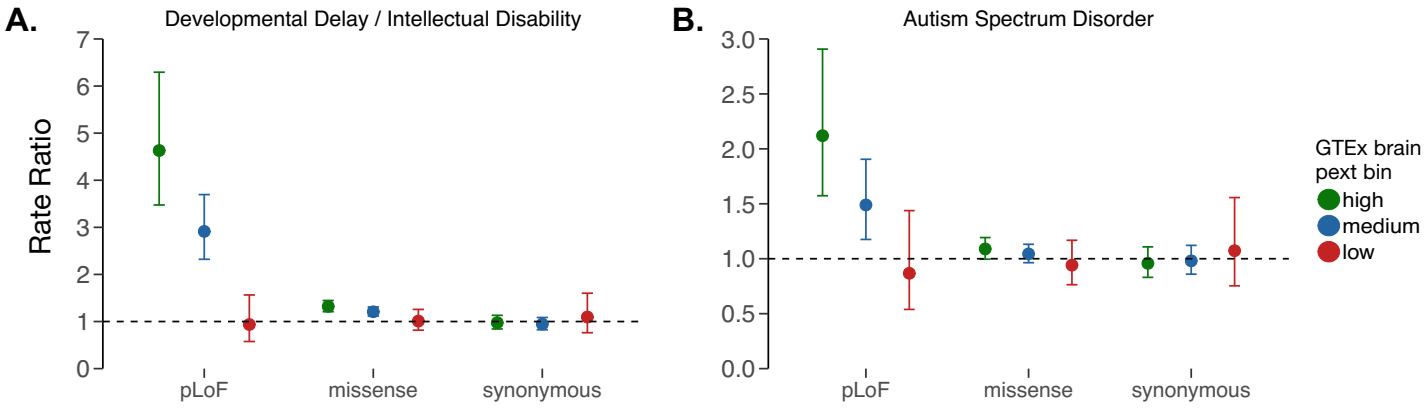


Figure 3.11 Application of transcript-expression based annotation to *de novo* variant analyses in **A.** developmental delay and/or intellectual disability (DD/ID) and **B.** autism spectrum disorder (ASD). We find that *de novo* pLoF variants found on constitutively expressed regions in GTEx brain tissues have larger effect sizes than *de novo* LoF variants in weakly expressed regions in both disorders. Strikingly, *de novo* pLoF variants found on regions with little evidence for expression are as equally distributed in cases vs controls as *de novo* synonymous variants, suggesting such variants can be removed from gene burden testing analyses to boost discovery power. The high pext expression bin contains 45.6% and 40.7%, and the low expression bin contains 4.5% and 8.2% of *de novo* pLoF variants found in DD/ID and ASD cohorts, respectively. Rate ratio represents estimate from the Poisson exact test.

project and the Autism Sequencing Consortium (ASC) [37–41,44]. We find that *de novo* pLoF variants in DD/ID patients in low-expressed regions have effect sizes similar to those of synonymous variants (rate ratio, denoted as RR, of low-expressed pLoFs = 0.94, p = 0.81) whereas pLoF variants in highly expressed regions have much larger effect sizes (RR = 4.63, p = 3.6 x 10^{-38}; Figure 3.11A). This observation is consistent for *de novo* variants in autism (RR for low-expressed pLoFs = 0.87, p = 0.54; RR for high-expressed pLoFs = 2.12, p = 8.2 x 10^{-8}; Figure 3.11B) and congenital heart disease with co-morbid neurodevelopmental delay [42,43] (Figure 3.12A) as well as rare variants (AC ≤ 10) identified in highly constrained genes in the large iPSCYH case/control study of Danish patients with autism spectrum disorder and attention-deficit/hyperactivity disorder [45] (Figure 3.12B). Overall, we consistently observe low-expressed pLoF

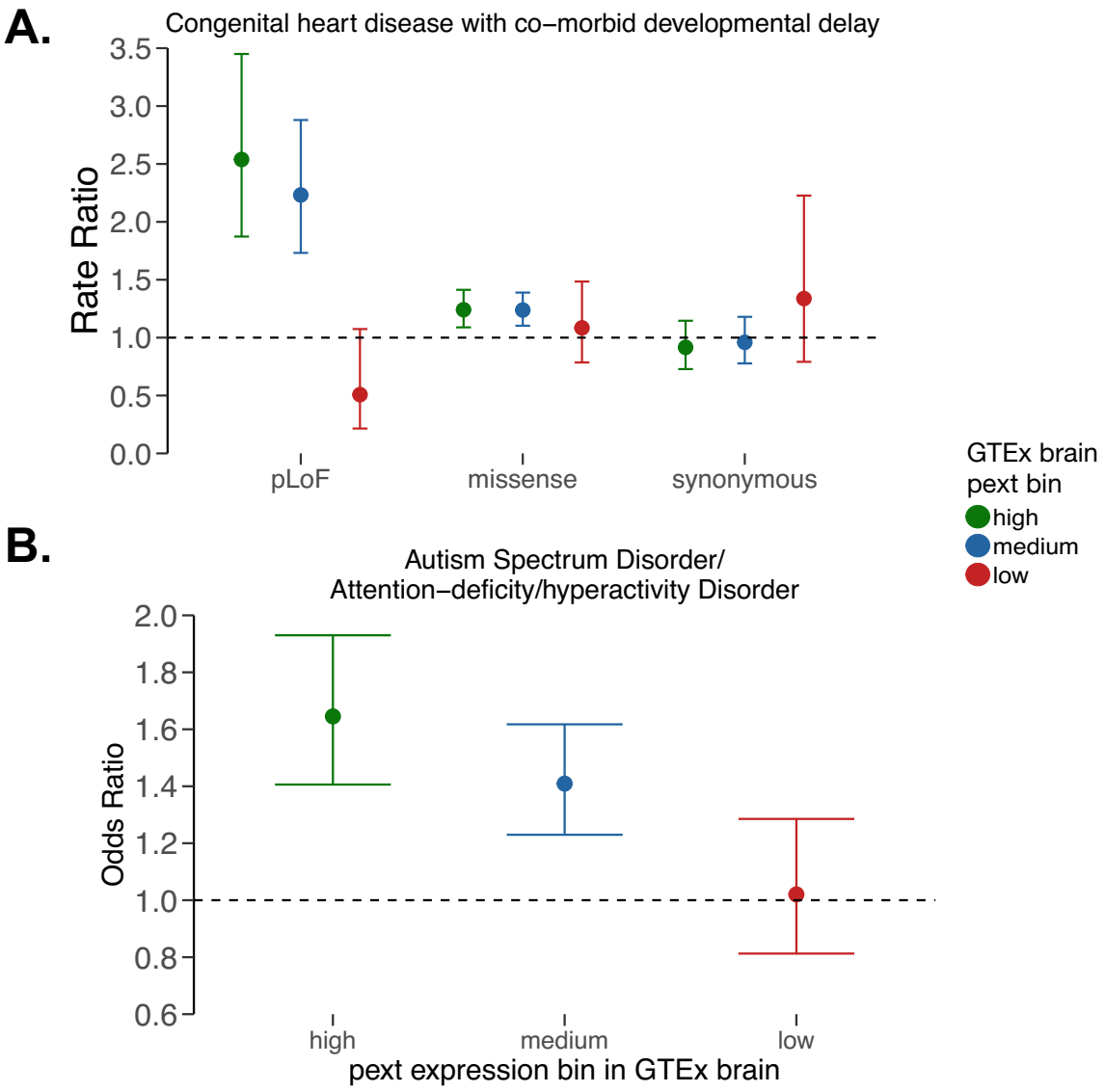


Figure 3.12 Application of transcript-expression based annotation to *de novo* and rare variant analysis in additional datasets A. Using *de novo* variants identified in probands with congenital heart disease and co- morbid developmental delay we find a consistent effect of *de novo* pLoF variants found on high expressed regions in GTEx brain having larger effect sizes than *de novo* LoF variants in weakly expressed regions. Once again, *de novo* pLoF variants found on regions with little evidence for expression are similarly distributed in cases and controls as *de novo* synonymous variants, suggesting such variants can be removed from gene burden testing analyses to boost discovery power. Rate ratio represents estimate from the poisson test. **B.** Rare pLoF variants (combined AC in cases and controls ≤ 10) identified in highly constrained genes (first decile in LOEUF) portioned upon pext expression bins show that those with high expression in GTEx brain have higher effect sizes than those identified in low-expressed regions, which are equally distributed in cases and controls. Odds ratio represents estimate from Fisher's exact text on counts.

variants exhibiting effect sizes similar to those of synonymous variants, and pLoF variants in constitutive regions having larger effect sizes, suggesting that incorporating transcript expression-aware annotation in rare variant studies can boost power for gene discovery.

Discussion

We have described the development and validation of a transcript expression-based annotation framework to integrate results from transcriptome sequencing experiments into clinical variant interpretation. While our initial analysis utilizes GTEx, our method can be used with any isoform expression dataset to annotate any variant file rapidly in the scalable software framework Hail (https://hail.is). For example, annotation of >120,000 gnomAD individuals with GTEx takes under an hour using 60 cores, at a cost of about $5 on public cloud compute, which can be further scaled to larger datasets. In addition, the annotations we provide are flexible: while we have described the use of average transcript-level expression across many tissues, alternative approaches such as using minimum expression across any tissue, may prove useful depending on variant interpretation goals.

We note that while this metric successfully discriminates between near-constitutive and low expression levels, which are useful for prioritizing and filtering variants respectively, regions with intermediate expression levels are more challenging to interpret. Regions tagged as low expression are often corroborated by expert opinion of CDS curation, but domain knowledge of a gene will outperform this summary metric.

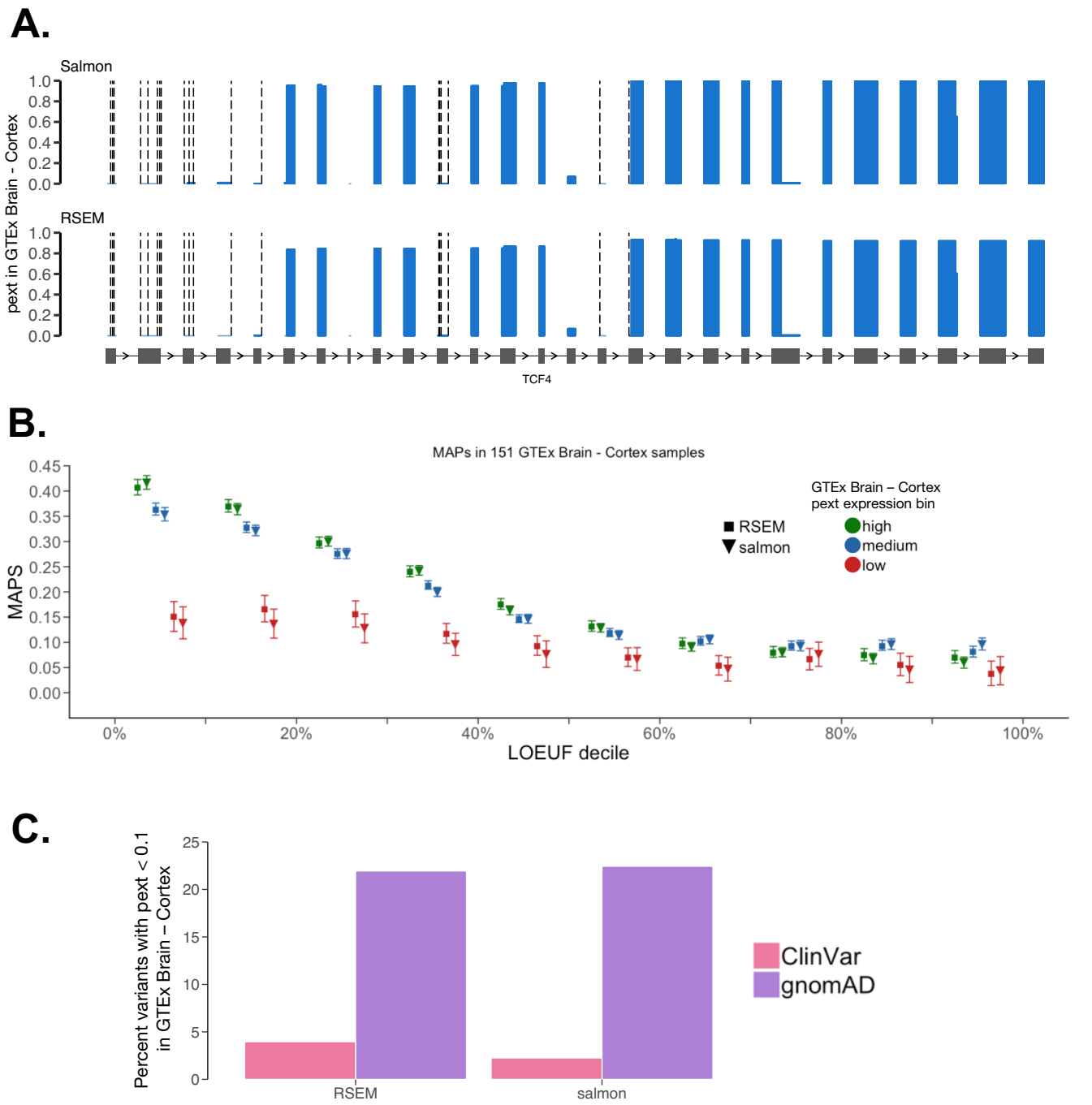


Figure 3.13 Comparison of key results using Salmon vs RSEM. Using isoform quantifications on 151 GTEx Brain – Cortex samples, we compared the results from analyses in the manuscript using quantifications from salmon and RSEM. Results from the analyses were consistent, underlining that the pext calculation is robust to isoform quantification tool used. **A.** The baselevel pext metric in *TCF4* using the two quantification tools. The 20 pLoF variants identified in gnomAD in *TCF4*, denoted by dashed lines, lie on unexpressed regions in Brain – Cortex samples using salmon or RSEM **B.** No significant difference in the MAPs score is seen for pLoF variants in pext expression bins with RSEM and salmon quantifications. **C.** The number of pLoF variants filtered with a Brain – Cortex pext cutoff of 0.1 in gnomAD vs ClinVar was similar, with results from quantification from salmon filtering 52 fewer ClinVar variants (out of 3,056 variants).

An important caveat in our approach is the imprecision of isoform quantification methods using short-read transcriptome data. However we note that repeating key analyses in the manuscript with a different isoform quantification tool showed consistent results (Methods, Figure 3.13), suggesting robustness to the precise pipeline used. The utility of this framework will increase as our ability to quantify isoform expression across tissues improves, including refinement of methods and gene models, as well as availability of long-read RNA-seq data from human tissues. In addition, improvement of single-cell RNA-seq technologies and generation of data across human tissues will provide insight into cell type-specific exon usage for incorporation into variant interpretation [49].

Significance

The development and validation of the pext score allows for quick and practical integration of population transcriptome datasets into interpretation of both variants in rare diseases, as well as statistical analyses in disorders with complex genetic architecture. Prior to our work, assessing the expression status of a region harboring a variant has been based on visual inspection of expression statuses of exons, and mental recalculation and correction of 3'bias. For gene burden testing in complex diseases such as schizophrenia, type 2 diabetes and others, choosing a variant annotation to include in a statistical test has often relied on selecting the consequence on the canonical transcript or to identify the worst consequence across all transcripts. Our method allows for a data-driven approach to selecting consequences for such statistical tests, which we show can improve power in analyses.

The generic and flexible method we developed allows for integration of any isoform expression dataset of interest with any variant file. The code used to generate pext scores is available as open source software and we have written detailed instructions on applying the method (Appendix 3.1). In addition, we provide a precomputed file of the transcript-expression value for every possible single nucleotide variant in the human genome. This allows for the integration of any combination of over 70,000 human RNA-seq samples that have been deposited into public repositories, a continuously growing number [50].

The pext metric has already proven useful in variant curation for drug target identification [51] and for filtering variants for identification of human knockouts [1]. Overall, we foresee this metric to be incorporated into variant interpretation in a Mendelian disease pipelines, rare variant burden analyses, and the prioritization of variants for recall-by-genotype studies.

Author contributions

Beryl B. Cummings: conceived and designed experiments, developed and validated the pext metric, wrote the analysis text

Konrad J. Karczewski: quality control and analysis of gnomAD dataset, development of LOEUF score used in the manuscript, help with coding in Hail, general guidance, critical review of manuscript

Daniel G. MacArthur: conceived and designed experiments, writing edits, general guidance

Mark J. Daly: general guidance, aid in writing manuscript, feedback on burden testing analyses

Jack A. Kosmicki: Collating de novo variant data for developmental delay and autism, help with de novo burden test analyses, analysis suggestions

Eleanor G. Seaby, Moriel Singer-Berk: Curation of pLoF variants in gnomAD, aid in writing method section

Jonathan M Mudge: GENCODE curation of region in haploinsufficient developmental delay genes with low pext, aid in writing methods section

Nicholas A. Watts, Matthew Solomonson: Integration of pext data into the gnomAD browser

Juha Karjailenen: Sharing method to perform salmon quantification

Timothy Poterba, Cotton Seed: Development of Hail, aid in developing pext module

F. Kyle Satterstrom: Sharing analysis and data for Danish Neonatal Screening Biobank for ADHD and ASD

Jessica Alfoldi: Aggregation of gnomAD dataset, general guidance, writing edits.

Bibliography

1. Karczewski, K. J., Francioli, L. C., Tiao, G. & Cummings, B. B. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* (2019).

2. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. doi:10.1101/074450

3. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

4. Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).

5. Dick, I. E., Joshi-Mukherjee, R., Yang, W. & Yue, D. T. Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca2+-dependent inactivation. *Nat. Commun.* **7**, 10370 (2016).

6. Splawski, I. *et al.* CaV1.2 Calcium Channel Dysfunction Causes a Multisystem Disorder Including Arrhythmia and Autism. *Cell* **119**, 19–31 (2004).

7. Liao, P. & Soong, T. W. CaV1.2 channelopathies: from arrhythmias to autism, bipolar disorder, and immunodeficiency. *Pflugers Arch.* **460**, 353–359 (2010).

8. Splawski, I. *et al.* Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8089–96; discussion 8086–8 (2005).

9. Nousbeck, J. *et al.* A mutation in a skin-specific isoform of SMARCAD1 causes autosomal-dominant adermatoglyphia. *Am. J. Hum. Genet.* **89**, 302–307 (2011).

10. Guven, A. & Tolun, A. TBC1D24 truncating mutation resulting in severe neurodegeneration. *J. Med. Genet.* **50**, 199–202 (2013).

11. Roberts, A. M. *et al.* Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Science Translational Medicine* **7**, 270ra6–270ra6 (2015).

12. Zheng, W., Chung, L. M. & Zhao, H. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* **12**, 290 (2011).

13. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).

14. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

15. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).

16. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

17. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

18. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

19. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

20. Riggs, E. R. *et al.* Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Hum. Mutat.* **39**, 1650–1659 (2018).

21. Robinson, P. & Jtel, T. Z. Integrative genomics viewer (IGV): Visualizing alignments and variants. *Computational Exome and Genome Analysis* 233–245 (2017). doi:10.1201/9781315154770-17

22. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

23. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

24. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

25. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–82 (2011).

26. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

27. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

28. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

29. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–7 (2011).

30. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

31. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

32. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

33. da Costa, P. J., Menezes, J. & Romão, L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *Int. J. Biochem. Cell Biol.* **91**, 168–175 (2017).

34. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

35. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).

36. Berg, J. S. *et al.* An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**, 36–44 (2013).

37. Hamdan, F. F. *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet.* **10**, e1004772 (2014).

38. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).

39. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).

40. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

41. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).

42. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).

43. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).

44. Satterstrom, F. K. *et al.* Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. bioRxiv, 484113. (2018).

45. Satterstrom, F. K. *et al.* ASD and ADHD have a similar burden of rare protein-truncating variants. doi:10.1101/277707

46. Przyborowski, J. & Wilenski, H. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* **31**, 313–323 (1940).

47. Lindsay, S. J. *et al.* HDBR expression: a unique resource for global and individual gene expression studies during early human brain development. *Front. Neuroanat.* **10**, 86 (2016).

48. Sweatt, J. D. Pitt–Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. *Exp. Mol. Med.* **45**, e21 (2013).

49. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).

50. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).

51. Minikel, E. V., Karczewski, K. J., Martin, H. C. & Cummings, B. B. Evaluating potential drug targets through human loss-of-function genetic variation. *BioRxiv* (2019).

Chapter 4

Discussion

Summary of results

As stated in the introduction, a main goal of human genetics research is accurately predicting the functional and clinical consequences of variation in the genome. While DNA sequencing technologies have had a remarkable impact on linking genetic variation to both common and rare disease, an emerging issue has been the fact that our ability to discover genetic variation outstrips our ability to interpret its functional impact [1,2]. In this thesis, we employed a functional genomic tool, transcriptome sequencing, in a practical and accessible way to aid in the interpretation of genetic variation in disease. To that aim, we first use RNA-seq directly on affected tissue samples from patients with Mendelian neuromuscular disease to aid in the interpretation of effect variants identified by prior WES/WGS on transcription. We developed tools and frameworks to identify splice aberrations, allelic imbalance and expression outlier status unique to a patient or groups of patients, and performed variant calling on patient RNA-seq data. Our results established RNA-seq as a useful complementary approach to DNA sequencing, with the diagnosis of 17 patients in our initial cohort of 50. This study was among the first of its kind to establish the utility of a functional genomics tool in rare disease diagnosis, and was published in April 2017 in *Science Translational Medicine* [3]. To date, it has received 122 citations.

Upon establishing the utility of RNA-seq for rare variant interpretation in rare disease, we considered the use of RNA-seq in cases where patient tissue was unavailable, as would be the case for Mendelian neurological disorders. We hypothesized that population transcriptome datasets would be greatly valuable to interpret genetic variants identified in rare disease patients, however the state of such

large-scale functional datasets were not in a format to allow analysts and clinicians to seamlessly integrate this information into variant interpretation. This was further motivated by the surprising observation of a large number of disrupting variants in the gnomAD database in dosage sensitive disease genes. Our hypothesis was based on the assumption that these variants were likely enriched for annotation errors but our current toolkit for systematic interpretation failed to filter them. We therefore hypothesized that these variants were likely enriched in regions of misannotated exons, or due to alternative mRNA splicing, were not acting in the relevant tissues. To test this hypothesis, we developed a summarized metric of isoform expression using data from the largest cross-tissue transcriptome sequencing inventory, GTEx. Starting from isoform quantifications in this dataset, we developed the proportion expressed across transcript (pext) metric, that represents the proportion of expression from a gene that is attributable to the given variant annotation. We tested pext using orthogonal functional metrics such as conservation, and established its utility for selectively filtering variants that are enriched for annotation errors. We integrated the pext score into the gnomAD browser, which receives an average of 70,000 page views per week. In addition, the transcript expression-aware annotation metric is available as an open source tool for scientists to integrate any isoform expression matrix of interest with any variant file. The resulting manuscript is currently available on biorXiV and has in review at a peer-reviewed journal as part of the gnomAD project manuscript bundle [4].

Overall, the two chapters of this thesis represent a practical way to integrate one type of functional genomics information into rare variant interpretation both in Mendelian and common disorders. Taken together, this allows scientists, clinicians and analysts

access to information that was previously unavailable and the development of the projects offer a few lessons that are discussed below.

Emerging concepts

The importance of splice variants in undiagnosed rare disease patients

In chapter 2, we set out to explore transcriptional aberrations in genetically undiagnosed Mendelian disease patients. We evaluated splicing, allele imbalance, expression outlier status and performed variant calling to identify putative pathogenic somatic variation. The diagnoses made in the study were all based on identification of variants affecting splicing. This was an unexpected result. While it may be possible that splice-affecting variants were prioritized due to the ease of interpreting their loss-of-function impact, versus for example, disruptive variants in the untranslated regions that may result in allele imbalance but are difficult to pinpoint in the DNA sequence even when allele imbalance is observed, this observation highlighted the importance of variants that affect splicing in undiagnosed cases.

In a parallel effort, Kremer et al. performed RNA-seq on fibroblast samples in patients with rare mitochondriopathies [5]. They focused on the identification of aberrant expression, aberrant splicing and ASE and were able to genetically diagnose 5 out 48 patients in their cohort with prior negative WES. Interestingly, the authors note that a majority of the newly diagnosed cases arose from identification of defective splicing. In one case, the authors identified a possible novel disease gene, *TIMMDC1*, with an intronic mutation causing inclusion of a pseudo-exon, similar to the *COL6A1* example in

this thesis. In total they identified the mutation in three unrelated families, highlighting that other previously underappreciated diagnosis may be accessible through considering splicing.

To assess the contribution of noncanonical splice aberrations to genetic diagnosis of developmental delay, which represent the largest single class of monogenic disorders, Lord et al. evaluated mutational constraint and pathogenicity of variants affecting the extended splice junction, defined as up to 30 bases around the exon-intron junction [6]. They identified important positions outside of the essential 2 base intronic region that were constrained against variation. Specifically, they estimate a variant at the G nucleotide of the donor + 5 position has an approximately 80% chance of being deleterious. Similarly, the donor -1 G nucleotide positions was shown to be constrained against variation, and these results were replicated in a parallel study by Zhang et al. [7]. The authors then evaluated *de novo* variants in their patient cohort of 7,833 probands, 5,907 of which had been previously undiagnosed following trio WES. They identified a diagnostic *de novo* mutation in the extended splice site region in 18 patients. Comparing this number to the number of patients diagnosed with pathogenic variants at the essential splice site, they observe that 27% of splice disrupting mutations in this cohort fall in the non-canonical positions. An analysis of pathogenic or likely pathogenic variants in ClinVar reveals that while 83.5% of variants fall in the essential splice site region, only 16.5% fall in the non-canonical splice region with high constraint. Taken together, the authors calculate that there is a 35-40% under ascertainment of disruptive non-canonical splice variants. It is important to note that this analysis focuses only on this extended splice region, and does not include exonic mutations that may

disrupt splicing, or those occurring at the branchpoint or splice-site creating deep intronic mutations, and the authors note that the 35-40% under-ascertainment estimate is likely conservative.

In a comprehensive analysis of the genetic architecture of Diamond Blackfan Anemia, Ulirsch et al. identified aberrant splicing caused by non-canonical splice mutations in 6 subjects in a cohort of 472 individuals. While the absolute number of diagnoses made via extended splice mutations is not high, the proportional diagnosis rate after standard diagnostic methods evaluating nonsense, essential splice and frameshift mutations is high, leading the authors to conclude that probands lacking typical gene mutations may harbor such cryptic mutations in known genes.

An observation made in this thesis, and which is echoed in Hurles et al. and Kremer et al. as well as the previously existing ACMG guidelines for variant interpretation [8], is that existing tools to predict splice-disrupting variants from DNA sequence lack specificity. However, based in part on our work, the work discussed above, and many other studies now evaluating the importance of splice-disrupting mutations in Mendelian disease, new tools for splice-prediction have been published. In one case, Jaganathan et al. used a deep learning approach to predict the splice-disrupting effect of any class of variant including exonic and deep intronic variation [9]. They report that 75% of predicted splice-disrupting variants do have an effect on splicing, with the score provided by the algorithm tracking with the rate of disruption. They observe a 71% sensitivity for near-exonic variation, and 41% for deep intronic mutation, highlighting that deep intronic variants with splice effects are more challenging to predict. The authors do identify a subset of intronic and synonymous mutations with

high scores, that are comparable in deleteriousness to frameshift, stop gain and essential splice site variants. By analyzing the gnomAD database, they predict that the average human carries ~5 rare functional cryptic splice mutations, compared to ~11 rare loss of function variants. Finally, the authors use a set of *de novo* variants identified in patients with developmental delay and autism to estimate that splice-affecting variants are estimated to account for 9% of developmental delay diagnoses and 11% of autism. In summary, the authors develop an improved splice affecting variant prediction algorithm with a high rate of validation. However, the rate of sensitivity, especially for certain classes of genetic variation has room for improvement and validation via manual cDNA analysis or RNA-seq is still recommended before concluding pathogenicity. As such tools continue to improve, estimates of the contribution of splice-affecting mutations in Mendelian disease will be available will gain precision.

The use of RNA-seq in genetic diagnosis and gene discovery

Several questions emerge from ours and Kremer et al.'s work on using RNA-seq for genetic diagnosis: What is the approximate diagnostic rate we can expect with RNA-seq in a given rare disease cohort? Do results indicate RNA-seq should be routinely used in diagnosis? In which cases will this approach be most useful? What are the study design considerations to maximize diagnostic yield? Will RNA-seq be useful for gene discovery?

The diagnostic rate in our cohort of 50 exome and/or genome undiagnosed individuals was 35%. Kremer et al reported a 10% diagnosis rate. In an additional study, Fresard et al employed RNA-seq in a cohort of 56 patients with a variety of severe

undiagnosed cases. Through performing RNA-seq on blood, they reported an 8.5% diagnosis rate [10]. Both Kremer et al. and this study acknowledged the higher yield in our study was likely due to the better genetic characterization of neuromuscular disorders versus mitochondrial disorders. However, it is important to note that all three studies employed various analysis strategies. While our work was primarily focused on evaluating splicing, Kremer et al. and Fresard et al. emphasized using RNA-seq for the identification of expression outliers. The three studies also differed significantly in their approaches within the same analysis strategy. The two studies that highlighted expression outlier status employed differing techniques, and while we developed our own analysis pipelines for splice aberrations, Kremer et al. used a previously published algorithm called Leafcutter for splice aberration detection [11]. Currently, these three studies are the main systematic uses of RNA-seq for diagnosis; however given the inherently different genetic architecture of the rare diseases studied, and the varying analysis strategies and tools employed, it is difficult to compare the diagnosis rates and conclude what are over and underestimates. However, we believe the success of these studies will prompt more systematic uses of RNA-seq in rare disease cohorts and such estimates will improve.

It is important to emphasize the use of RNA-seq as a complementary diagnostic tool for interpretation and not a replacement of WES or WGS. RNA-seq only allows for the assessment of expressed genes in a given tissue at a particular time point, and is subject to artifacts such as ischemic time [12]. This prevents the reliable identification of germline genetic variants with RNA-seq, and overall variant calling from RNA-seq carries approximately 80% sensitivity [13–15]. Other subtler effects can also result in false

negatives with RNA-seq: Consider a case in which a basally expressed gene carries two disease-causing loss-of-function variants *in trans* in a patient. If both variants result in nonsense mediated decay, the expression of the gene should expectedly be depleted. Therefore RNA-seq may fail to capture any molecules for this gene. In this case, RNA-seq analysis may identify the gene as an expression outlier, but without paired DNA sequencing, there will not be enough data to evaluate the gene in RNA-seq for diagnosis. Therefore, we again highlight RNA-seq as a useful complementary tool for cases for which standard diagnostic tools including WES and/or WGS has not yielded a molecular diagnosis.

A critical consideration to employ RNA-seq is which tissues are obtainable from patients. Again, in our study, due to the availability of muscle tissue from patients based on routine biopsies for diagnosis, the affected tissue was available. We do note however, that our study involved the collection of a variety of muscle subtypes such as biceps and deltoid. Fresard et al. employed RNA-seq on blood [10], Kremer et al on patient fibroblasts [5], Sankaran et al on lymphoblastoid cell lines [16]. In additional studies of single cases or smaller cohorts, Oliver et al. employed blood RNA-seq in a patient with multiple osteochondromas [17], and Hamanaka et al. employed muscle RNA-seq in a cohort of 10 patients with nemaline myopathy [18]. Both our manuscript and that of Hamanaka et al. evaluated alternative tissues as a proxy of the muscle transcriptome and found that many of the most important muscle diseases genes were not expressed in blood or fibroblasts [3,18]. In contrast, Kremer et al. showed that mitochondrial disease genes were expressed in fibroblasts, and Sankaran et al. reasoned that the ubiquitously expressed nature of the genes involved in Diamond Blackfan anemia would enable

analysis in cell lines. Because Fresard et al. analyzed a cohort of patients with severe syndromic diseases, they showed that blood can be useful as a proxy. All studies discussed share a simple strategy which is the careful evaluation of available tissues. In other words, the proxy tissue used will depend on the disease of interest, the expression status of commonly disrupted genes of the disorder across tissues, and (critically) the clinical availability of that tissue from patients.

It is important to highlight that only considering the expression status of a set of previously established disease genes in a candidate proxy tissue may not be sufficient. Our understanding of tissue-specific splicing patterns and their effect on diagnosis rates in proxy tissues is limited. While studies have suggested the splicing is secondary to gene expression for cellular identity [19] we hypothesize firstly that this is likely an underestimate brought on by short-read RNA-seq and secondly that tissue-specific effects of splice disrupting variants will need more consideration. One useful resource that would be much anticipated would be a proxy tissue database based on the GTEx project. In such a case, evaluation of co-expression and co-splicing networks across GTEx tissues could greatly informed proxy tissue decisions in Mendelian diagnosis.

For a variety of Mendelian disorders, the affected tissue will never be attainable, such as for neurodevelopmental diseases. Even in the case of neuromuscular disorders, as WES or WGS continue to rise as frontline diagnostic strategies, muscle biopsies will become more infrequent [20,21]. In addition to considering proxy tissues, *in vitro* strategies may be useful in such cases. Recently, Gonorazky et al. generated myotubes from transdifferentiated patient fibroblasts and employed RNA-seq in a cohort of 25 patients, yielding a diagnosis rate of 36% [22]. For retinal diseases, Buskin et al.

generated patient-specific retinal organoids and retinal pigment epithelium from induced pluripotent stem cells [23]. Such *in vitro* approaches can be low throughput, but alleviate the necessity of invasive access to patient tissue.

The use of transcript-expression aware annotation

Given how recently our transcript-expression aware annotation work was published, it is difficult to predict its adoption in the community. However, we believe our method allows access to tissue-specific exon expression patterns, and based on our work, is useful for filtering falsely annotated variants. It also aids in the choice of annotation for a given variant in rare variant analyses for complex disorders. So far, this method has been employed to filter homozygous loss-of-function variants in the gnomAD cohort, and has removed approximately 30% of such variants [24]. This means that analysts are able to curate approximately 1,000 fewer putative loss-of-function variants, as our metric tracks closely with conservation. In an additional example, Minikel et al. have used the pext score to deeply curate loss-of-function variants identified in potential drug targets in the gnomAD dataset. They have found in one example, using transcript-expression aware annotation, filters almost all loss-of-function variants in *MAPT* which has important implications of using inhibitory drugs against the gene product [25].

Future directions and improvements

Much work is needed to continue to assess the utility of functional genomics in Mendelian disease diagnosis. For example, what is the role of other functional genomics approaches for variant interpretation? In Kremer et al., metabolomics was used in conjunction with RNA-seq to identify metabolites that were completely depleted in patient samples [5]. A recent study by Aref-Eshghi et al. employs genome-wide DNA methylation analysis in peripheral blood in a cohort of neurodevelopmental disease and congenital anomalies to show it aids variant interpretation [26]. Such functional tools are going to be integral to evaluate the contribution of noncoding variation to the diagnostic gap in Mendelian disease.

Transcriptome sequencing will be useful to address the diagnostic gap by offering insight into mechanisms of variable penetrance. By employing RNA-seq and allele specific expression analysis, Castel et al. have shown the effect of *in cis* genetic variants on the expression of coding variants. Employed more broadly in Mendelian patients, this approach will offer insight into the molecular underpinnings of penetrance and variable expressivity [27].

One caveat in our transcript-expression aware annotation pipeline is the shortcoming of short-read RNA-seq data at isoform quantification. A fundamental drawback of this approach is that it cannot assess full length isoforms. Instead, isoform expression measures from short read data are probabilistic and not directly quantified. Long read RNA-seq is an emerging technology that allows capture of full-length isoforms, thus bypassing many of the flaws associated with short read RNA-seq. This technology has just recently reached the brink of scalable application to larger sample

sets and has not yet been applied at scale to human tissues or used for the analysis of functional genetic variants. We believe generating long read RNA-seq data will be invaluable towards characterizing isoform diversity across human tissues, which can inform proxy tissues for genetic diagnosis, and to improve our understanding of the tissue-specific effects of variants, and their link to human disease.

While our focus in the application of transcriptome sequencing has been on tissues, single-cell technologies have gained massive traction over the last few years [28]. While many of these technologies rely on sequencing 3' ends of transcripts [29], full length isoform capture in single cells [30] would allow the evaluation of the cell-type specific effects of genetic variation.

Bibliography

1. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
2. Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).
3. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, (2017).
4. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant discovery and interpretation. *bioRxiv* 554444 (2019). doi:10.1101/554444
5. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
6. Lord, J. *et al.* Pathogenicity and selective constraint on variation near splice sites. *Genome Res.* **29**, 159–170 (2019).
7. Zhang, S. *et al.* Base-Specific Mutational Intolerance Near Splice-Sites Clarifies Role Of Non-Essential Splice Nucleotides. doi:10.1101/129312
8. Richards, S. *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.* **17**, 405–424 (2015).
9. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
10. Fresard, L., Smail, C., Smith, K. S., Ferraro, N. M. & Teran, N. A. Identification of rare-disease genes in diverse undiagnosed cases using whole blood transcriptome

sequencing and large control cohorts. *BioRxiv* (2018).

11. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
12. Zhu, Y., Wang, L., Yin, Y. & Yang, E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Sci. Rep.* **7**, 5435 (2017).
13. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).
14. Prodduturi, N., Bhagwate, A., Kocher, J.-P. A. & Sun, Z. Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine. *BMC Med. Genomics* **11**, 67 (2018).
15. Quinn, E. M. *et al.* Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One* **8**, e58815 (2013).
16. Ulirsch, J. C. *et al.* The Genetic Landscape of Diamond-Blackfan Anemia. *Am. J. Hum. Genet.* **104**, 356 (2019).
17. Oliver, G. R. *et al.* RNA-Seq detects a SAMD12-EXT1 fusion transcript and leads to the discovery of an EXT1 deletion in a child with multiple osteochondromas. *Mol Genet Genomic Med* **7**, e00560 (2019).
18. Hamanaka, K. *et al.* RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. *Genet. Med.* (2018). doi:10.1038/s41436-018-0360-6
19. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science*

348, 660–665 (2015).

20. Aartsma-Rus, A., Ginjaar, I. B. & Bushby, K. The importance of genetic diagnosis for Duchenne muscular dystrophy. *J. Med. Genet.* **53**, 145–151 (2016).
21. McDonald, C. M. Clinical approach to the diagnostic evaluation of hereditary and acquired neuromuscular diseases. *Phys. Med. Rehabil. Clin. N. Am.* **23**, 495–563 (2012).
22. Gonorazky, H. D. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
23. Buskin, A. *et al.* Human iPSC-Derived RPE and Retinal Organoids Reveal Impaired Alternative Splicing of Genes Involved in Pre-mRNA Splicing in *PRPF31* Autosomal Dominant Retinitis Pigmentosa Type 11. *SSRN Electronic Journal* doi:10.2139/ssrn.3155753
24. Karczewski, K. J., Francioli, L. C., Tiao, G. & Cummings, B. B. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* (2019).
25. Minikel, E. V., Karczewski, K. J., Martin, H. C. & Cummings, B. B. Evaluating potential drug targets through human loss-of-function genetic variation. *BioRxiv* (2019).
26. Aref-Eshghi, E. *et al.* Diagnostic Utility of Genome-wide DNA Methylation Testing in Genetically Unsolved Individuals with Suspected Hereditary Conditions. *Am. J. Hum. Genet.* **104**, 685–700 (2019).
27. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory

variation contributes to disease risk. *Nature Genetics* **50**, 1327–1334 (2018).

28. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
29. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
30. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4259

Appendix

Explanation of the appendix

In the publication of our work on improving genetic diagnosis in Mendelian disease with transcriptome sequencing, we published an accompanying blog post for those who would like to apply our methods in their own work. This blog post is presented as Appendix 2.1. Similarly, in our publication of a transcript expression aware annotation tool, we made the method available as open source, wrote a detailed Github repository on using the tool, and detailed the analyses in the manuscript. This work is presented as Appendix 3.1

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing – a walk through

This post summarizes our recent manuscript on the application of transcriptome sequencing (RNA-seq) to the diagnosis of patients with Mendelian diseases, and provides a practical walk-through of our framework, methods and the Github code accompanying the paper).

Why RNA-seq for genetic diagnosis?

The current rate of genetic diagnoses across a variety of Mendelian disorders is approximately 25-50%. This means that more than half the families that come into the clinic searching for a genetic cause for their disease fail to receive a diagnosis.

There are a variety of reasons current diagnostic rates for Mendelian disorders are far from perfect. In some cases, the pathogenic variants are in genes that have not yet been established in the literature to cause the particular disorder, and with a single case, there isn't enough evidence to make the diagnosis. There may also be complex inheritance patterns, such as digenic causes, to disorders that we have so far been underpowered to uncover. In addition, there are some key classes of variants where improvements in methods are still needed, such as calling structural variants from exome data or somatic variant discovery.

However, perhaps the most common driver for missed diagnoses is our inability to successfully functionally interpret the variants we see in patient DNA. This is especially true for variants we identify in whole genome sequencing (WGS) since our understanding of the non-coding genome remains limited, and the sheer number of these variants is overwhelming: **in the gnomAD WGS database, every European**

carries an average of 7,067 variants that are not found in anyone else. This means that even frequency filtering with gnomAD leaves us with too many candidate variants for which the functional impact is unknown.

This is where RNA-seq comes in. RNA-seq offers a layer of functional information on top of what we know from the genetic analysis, and can help us begin to interpret some of the variants we identify with exome or whole genome sequencing, or identify new variants that may elude these technologies.

In our project we set about using RNA-seq to improve the diagnosis of a cohort of patients with a variety of severe, undiagnosed muscle diseases. We set out to look for splicing defects, allele imbalance, expression outlier status and to do variant calling directly on RNA-seq data. Our goal was to identify variants that may not have been captured with DNA-sequencing or to identify non-coding variants with functional impact that we may not have been properly interpreted. In the end, we were able to genetically diagnose 17 out of 50 undiagnosed patients, primarily through discovery of splice aberrations.

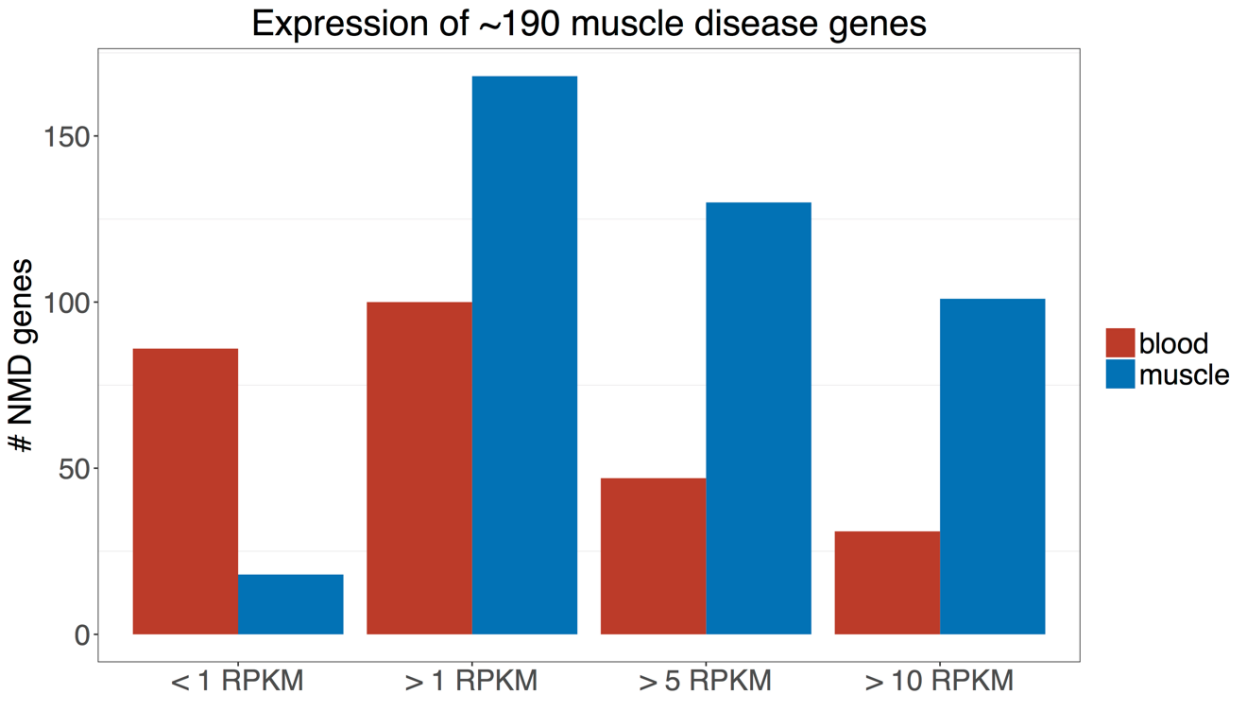
Some considerations for study design

Before applying RNA-seq to cohorts of undiagnosed Mendelian disease patients there are some critical questions to inform study design including i. what tissue to sequence ii. how many patients and controls to begin with and iii. what protocol and read depth to use for patient RNA-seq.

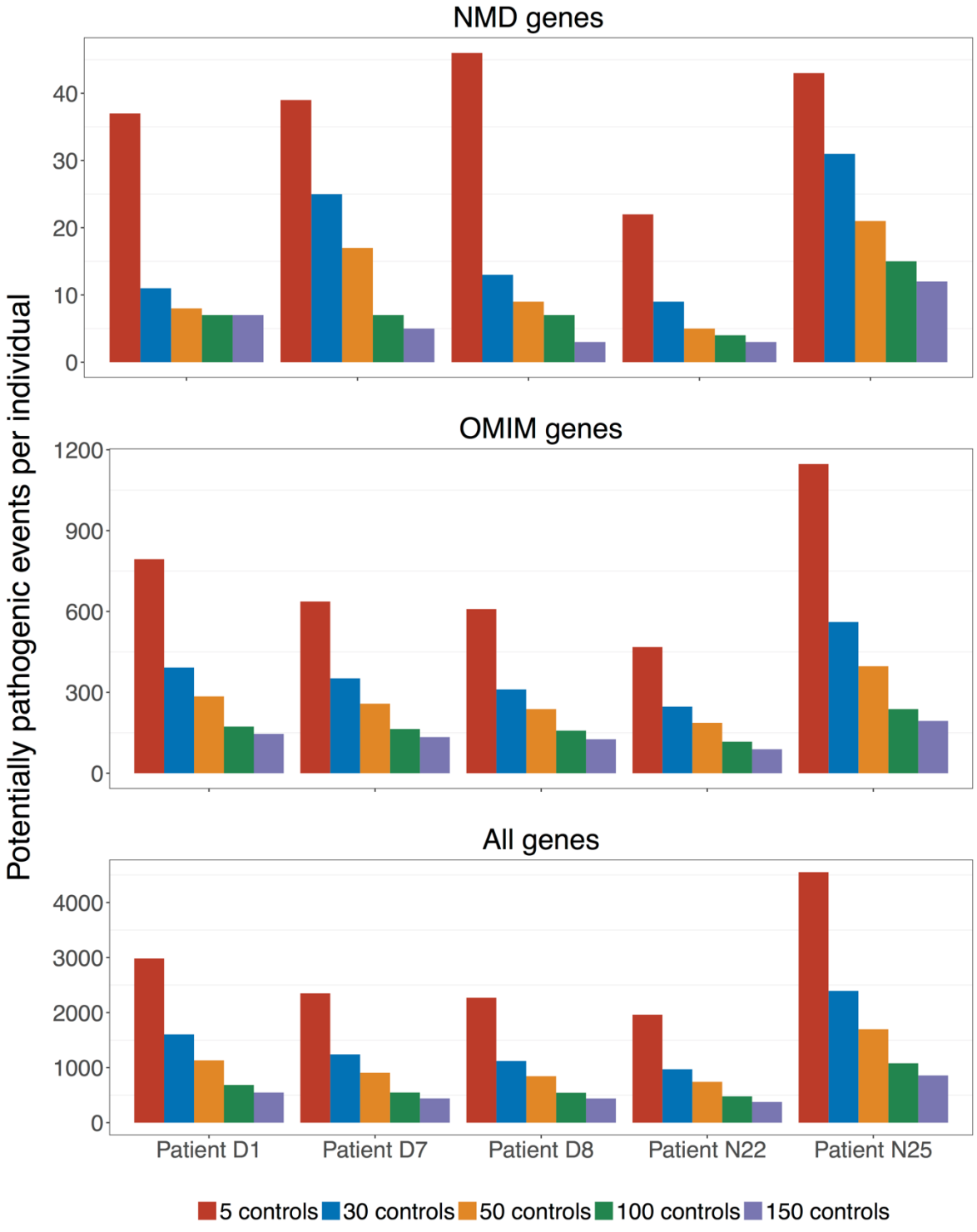
Based on multi-tissue transcriptome studies like GTEx, it is becoming increasingly clear that gene expression and splicing profiles can vary widely between

tissues. Therefore to identify aberrations in these profiles, it is ideal to go after the disease-relevant tissue.

Consider the comparison below, showing the expression of ~190 Mendelian neuromuscular disease genes in GTEx blood and muscle, which suggests that almost half of these genes are below 1 RPKM in blood, a cutoff below which it is difficult to see enough reads in 50 million-paired end RNA-seq dataset to identify splice aberrations. This suggests that blood RNA-seq is a poor proxy for muscle for the purposes of diagnosis. We highly recommend performing a similar analysis based on your genes of interest in order to choose a proxy tissue for RNA-seq, if the disease-relevant tissue is unavailable.



Secondly, increasing the number of controls will increase your power to filter out non-deleterious junctions. We applied the framework laid out in the manuscript to identify “potentially pathogenic junctions” in 5 patients while increasing the number of GTEx controls. This shows that having few controls results in identifying more events that are seemingly unique to the patient. It also underlines the greater filtering power of having a specific gene list for the disorder of interest. Here, we recommend leveraging GTEx by choosing a proxy tissue represented in the dataset. If this is not possible, we would definitely recommend sequencing your own set of controls, and starting with at least ~20-30 samples.


NMD genes
OMIM genes
All genes
Potentially pathogenic events per individual
0
10
20
30
40
0
300
600
900
1200
0
1000
2000
3000
4000
Patient D1
Patient D7
Patient D8
Patient N22
Patient N25
5 controls
30 controls
50 controls
100 controls
150 controls

Lastly, the protocol and read depth is dependent on whether you will be able to incorporate GTEx data into your analysis, in which case we recommend staying close to the GTEx protocol. Our patient samples were sequenced with 50 or 100 million paired-end reads and at a minimum, we recommend 50M paired-end reads. However, the precise impact of read depth will be dependent on the expression of your genes of interest, considering that expression of the disrupted gene in the patient may be decreased (for a gene in which recessive LOF mutations result in disease) and that larger genes will be more dramatically impacted by 3' bias, which will lower your ability to have enough reads at the 3' end of a gene to look for splice aberrations.

A reference panel of control tissue RNA-seq

Mendelian muscle disorders are a major disease focus in our lab. They are also a very practical place to test the value of RNA-seq for diagnosis: the collection and storage of frozen muscle biopsies is currently routine clinical practice for undiagnosed patients as they are used in protein studies, meaning that high-quality RNA from a disease-relevant tissue is available for a very large proportion of undiagnosed patients with these diseases.

One of the most powerful tools in Mendelian disease diagnosis are large-scale reference databases to look up the population frequency of a variant of interest (**cough* gnomAD). This allows for filtering out events that are too common in the population to plausibly result in a Mendelian phenotype. We needed similar reference databases to filter events we identified in our patient muscle RNA-seq, so we turned to the Genotype Expression Project (GTEx) dataset, which is a large multi-tissue transcriptome sequencing effort that has sequenced across ~50 tissue types in ~600

individuals. The GTEx inventory includes skeletal-muscle RNA-seq, so we integrated their data into our framework, which eliminated a need to obtain and sequence muscle tissue from healthy controls.

At the offset of the project, just over 400 skeletal-muscle RNA-seq samples were available from GTEx. We sub-selected 184 controls from GTEx that had high quality RNA-seq as well as phenotypic features that more closely matched our patient samples (see Methods section of the manuscript for more details).

Quality controlling patient RNA-seq data

We performed three main quality controls: i. technical QC ii. comparing gene expression profiles to GTEx samples to assess tissue quality iii. sample matching to ensure the source of RNA-seq was the proband for which we had prior information. This included fingerprinting comparison for WES/WGS /RNA-seq from the proband as well as checking to see if the sex entered for the patient validated in the RNA-seq, to ensure there were no sample mix-ups.

For technical QC, we obtained metrics by running RNA-seQC with gencode annotations obtained from the GTEx project.

For the tissue check, we used GTEx skeletal-muscle samples as well as tissues that potentially contaminate muscle biopsies such as skin or adipose to run principal component analysis (PCA) with our patient samples. Initially, we ran the PCA by using all genes, but now to QC each batch that comes in, we use tissue-preferentially expressed genes identified by GTEx, which produces similar results, but runs faster. A list of tissue-preferentially expressed genes are available in supplementary table S5 of this GTEx manuscript.

To validate the sex of the individual, we compared the average chrY and XIST expression and also clustered samples based on sex-preferentially expressed genes from the same GTEx paper as above.

There is example code and data in the Github page to check the clustering pattern of a randomly selected patient sample. You can see and run the code by cloning the repo https://github.com/berylc/MendelianRNA-seq.

in MendelianRNA-seq/QC

For the muscle check:

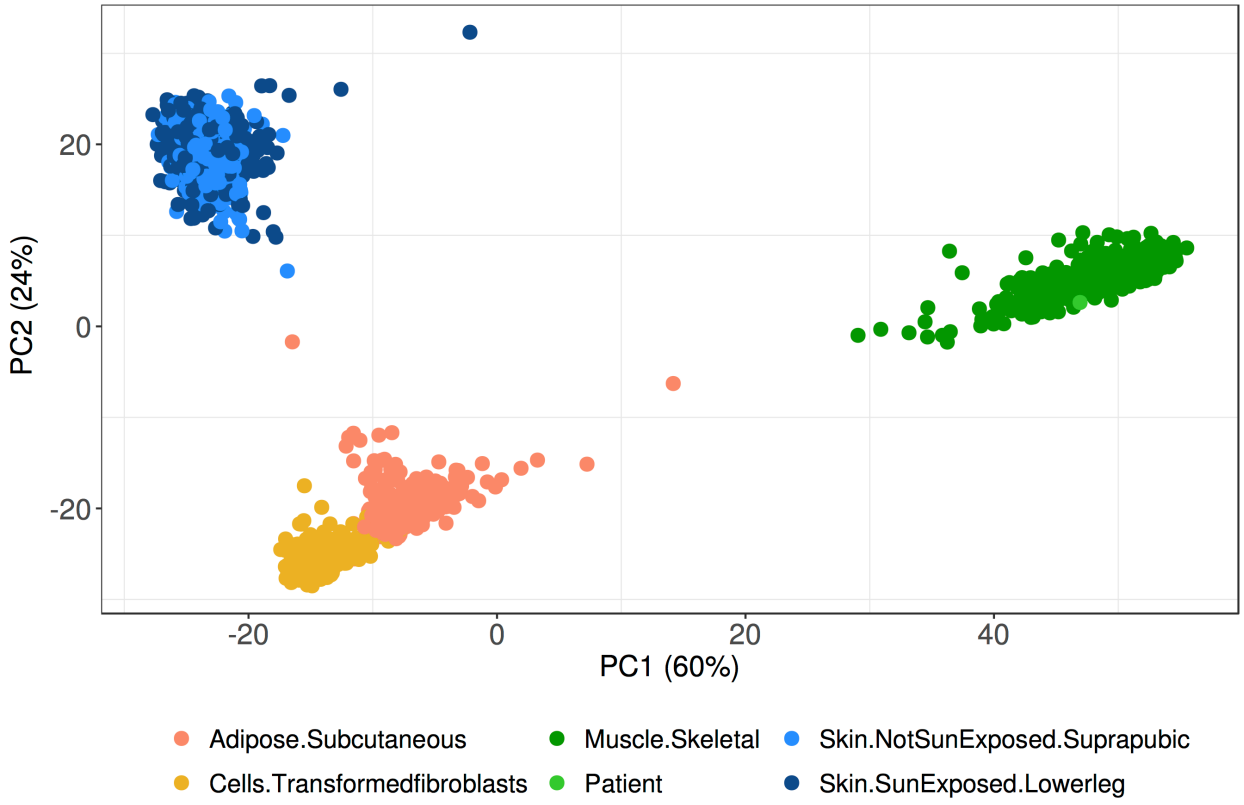
```
$ Rscript MuscleCheck.R -patient_rpkm ../example/genes.rpkm.gct
-out_file walkthrough.tissue
```

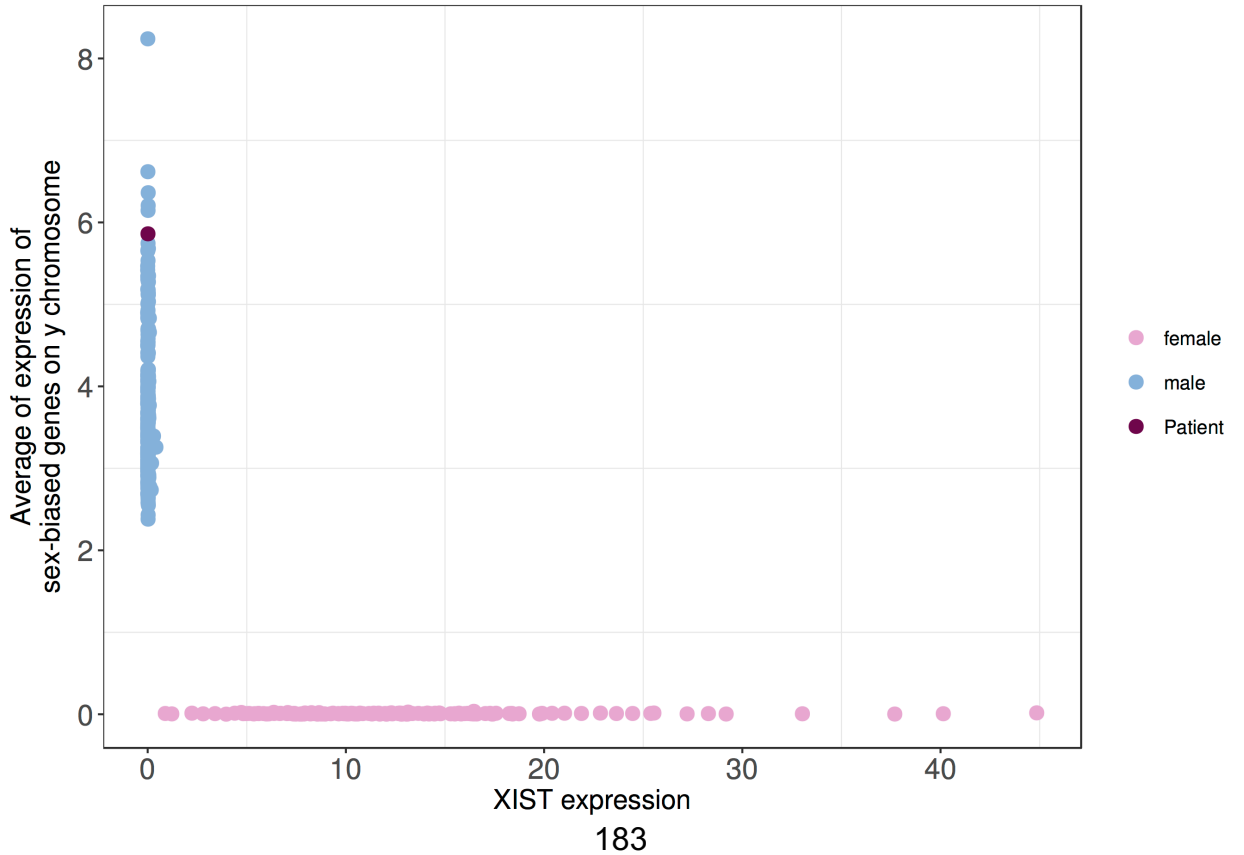
and for the Sex Check:

```
$ Rscript SexCheck.R -patient_rpkm ../example/genes.rpkm.gct
-out_file walkthrough.sex
```

Both commands create PDF files with plots for manual inspection and output text to indicate whether the samples cluster with muscle as well as their respective sex. In MuscleCheck.R, adding -writePCADat outputs the PC coordinates of all samples, in order to identify samples that may look like outliers by manual inspection.

Running these commands produces the plots below, which validate the RNA-seq sample is muscle and that the sample is male:





Default GTEx files to run this code for muscle RNA-seq are stored in Github. For the tissue check, you need an expression matrix of control tissues of interest and relevant tissue-preferentially expressed genes (ie. tissue_preferential_genes_fibs.msck.skn.adbp.txt and gtex_expression_tissue_preferential_fibs.msck.skn.adbp.txt in MendelianRNA-seq/data). For the sex check you need an expression matrix with sex-preferentially expressed genes (gtex_expression_sex_biased.txt).

If you'd like to build a similar QC framework for a different set of tissues of interest, you can follow the code to create the GTEx expression matrices for sex and tissue-preferential genes:

in MendelianRNA-seq/data

Get the required GTEx files:

```
$ wget -i gtex_file_urls
```

>Go through make_gtex_files.R and change the tissue names to create the files for your own tissues of interest.

Lastly, to ensure the sample for WES/WGS and RNA-seq are the same, we used PLINK to look at IBD estimates from ~5,800 common SNPs collated by Purcell et al. This code can be used to check relatedness within family WES/WGS/RNA-seq data as well

in MendelianRNA-seq/QC

-Note that to genotype RNA-seq files, you will need to split your BAMs first. You can do this by using GATK SplitNCigarReads.

-You will also need to have GATK and PLINK installed, and point the scripts to a human genome reference fasta file.

1)Generate GVCFs from all bams you're interested in checking

For a single BAM:

$ sh MakeGVCF.sh \

/path/to/gatk.jar \

/path/to/Homo_sapiens_assembly19.fasta \

/path/to/your.bam

2)Make a list of all output GVCFs and joint genotype for ~5,800k common SNPs

$ls -1 *vcf.gz > gvcfs.list

$ sh JointGenotype.sh \

/path/to/gatk.jar \

/path/to/Homo_sapiens_assembly19.fasta \

gvcfs.list

3)Make PLINK TFAM and TPED files for PLINK

$ zcat out.joint.vcf.gz | ./MakeTPED.pl > out.tped

$ zcat out.joint.vcf.gz | ./MakeTFAM.pl > out.tfam

4)Run PLINK

$ plink --noweb --tfile out --genome

>This will produce a plink.genome file, which has IBD values for your samples. We assessed the PI_HAT column to check relatedness. In our cohort the PI_HAT value for

WES, WGS, and RNA-seq data from the same individuals ranged from 0.67-1.00 (mean = 0.9), compared to a range of 0-0.18 (mean= 0.001) for non-matching individuals.

(Re)processing patient and GTEx data

We downloaded and decrypted Tophat aligned BAM files from the GTEx dbGAP and realigned them with STAR 2-pass. We were specifically interested in unannotated splice events (ie. splice aberrations like exon skipping or intron inclusion) so we decided to align with STAR which we reasoned would be more sensitive to detect such events (relevant paper comparing several alignment methods found here).

In order to be as sensitive as possible to detect splice events with low-level read support, we concatenated 1st pass junctions identified across all samples and fed these junctions into the Star 2nd Pass alignment. Please note that these steps are also laid out in the STAR manual for 2-pass alignment.

-You will need to have Picard and STAR installed for realignment

-You will also need to have created a STAR genome file for your RNA-seq protocol (see "Generating genome indices" in the STAR manual)

In MendelianRNA-seq/Reprocessing

1)If you're starting off with Tophat BAMs, turn them into fastqs

$ sh BamToFastq.sh /path/to/picard.jar /path/to/your.tophat.bam

2)Run STAR 1st pass to identify junctions

> 1stPassScript.sh is a wrapper around GeneralAlignment.sh which includes pre-specified alignment parameters for differing read lengths/sequencing types (stranded/unstranded, single-end/paired-end) etc. You can modify this script to specify the read length in your samples. It's currently set up to align 76 bp unstranded paired-end RNA-seq (ie. GTEx RNA-seq)

```
$ sh ./1stPassScript.sh \
/path/to/your.tophat.bam \
/path/to/directory/containing/sample/fastq/directory \
/path/to/STAR/executable \
/path/to/STAR/genome
```

3)Concatenate all your junctions from the first pass alignment and filter the junctions to remove splice junctions on the mitochondrial genomes and unannotated junctions with less than 5 reads

```
$ cat *tab > all.SJout.tab

$ awk '$1!="MT"' all.SJout.tab | awk '$6~1' > final.filtered.SJout.tab

$ awk '$1!="MT"' all.SJout.tab | awk '$6~0' | awk 'int($7)>5' >> final.filtered.SJout.tab
```

4)Create a new STAR genome by aligning one sample.

> This is the same as 1stPassScript.sh except now we add the junction file we created final.filtered.SJout.tab. This will align the one sample, and create a new genome file, which you will feed into the next step:

```
$ sh 2ndPassScript_CreateGenome.sh \
/path/to/your.tophat.bam \
/path/to/directory/containing/sample/fastq/ \
/path/to/STAR/executable \
/path/to/STAR/genome \
final.filtered.SJout.tab
```

5)Align all other samples using the new genome file created by the last step

```
$ sh 2ndPassScript_AllOthers.sh \
/path/to/your.tophat.bam \
/path/to/directory/containing/sample/fastq/ \
/path/to/STAR/executable
/path/to/STAR/newly/created/STAR/genome
```

6)Mark Duplicates with Picard

$ sh MarkDuplicates.sh /path/to/picard.jar /path/to/your.new.star.bam

>Note that if you have only a handful of samples, you can simply run (4) for all your samples, instead of doing a 2-step approach. This will create identical STAR genome for each samples, which you can delete.

Splice junction discovery and filtering

Our primary goal with splice junction analysis was to be able to identify splicing events that were found in one patient or groups of patients, and largely missing in GTEx controls. When we started the project, there were no easily adaptable tools that served this purpose. For example, DEXSEQ is a tool used for differential expression analysis, but it performs differential exon usage analysis to find global differences between experimental groups. Our goal was not to identify general exon-usage differences between diseased and healthy skeletal-muscle but to identify specific aberrations in specific individuals and look for the underlying genetic variants.

A year or so after we started the project, a software called Leafcutter was published, which can be used to identify sample-specific splice junctions. This paper used Leafcutter to identify sample-specific splice junctions, and was able to identify one sample with aberrant splicing out of 48 patients.

We developed our own pipeline that is composed of three steps i. Splice junction discovery from split reads ii. Normalization of read support for junctions based on local canonical splicing iii. Filtering splice junctions and spot-checking.

1- Splice junction discovery

We identified splice junctions supported by uniquely mapped split reads. This approach takes a gene annotation file and a list of BAM files as input, and identifies all the junctions in the samples.

>We'll run a mini-example of this using a subsetted bams from some patients in the manuscript (we use the subset so the data are unidentifiable) While the code is set up to run across many genes and bam files, we will run it on bams subsetted down to three exons in NEB, where one patient carries an essential splice site variant, to get a sense of how the scripts work.

In MendelianRNA-seq/Analysis

```
$ sh SpliceJunctionDiscovery.sh ../example/example.gene.NEB.list
../example/patient.bam.list
```

```
$ head All.example.gene.NEB.list.splicing.txt
```

```
Gene Type Chrom Start End NTimesSeen NSamplesSeen Samples:NSeen

NEB protein_coding 2 152518855 152520062 1 1 Patient.D1.small:1

NEB protein_coding 2 152355006 152782552 1 1 Patient.D1.small:1

NEB protein_coding 2 152544892 152544894*1D17 1 1 Patient.E2.small:1
```

>The first three column names are self-explanatory; Start and End refer to the boundaries of the splice junction (ie Start and End are both exon-intron junctions), NTimesSeen is the total number of reads in the dataset that support that junction, NSamplesSeen is the number of samples the junction is seen in and the final column lays out the number of reads supporting the junction in each sample (sorted by read support so the sample with the highest read support is at the end).

>Note that if you are doing this on bams you've aligned with a different method, you must assign unique mapping quality to be 60 (vs. the default 255). You can do this while you're aligning with STAR using --outSAMmapqUnique 60 or you can use the GATK tool ReassignOneMappingQuality

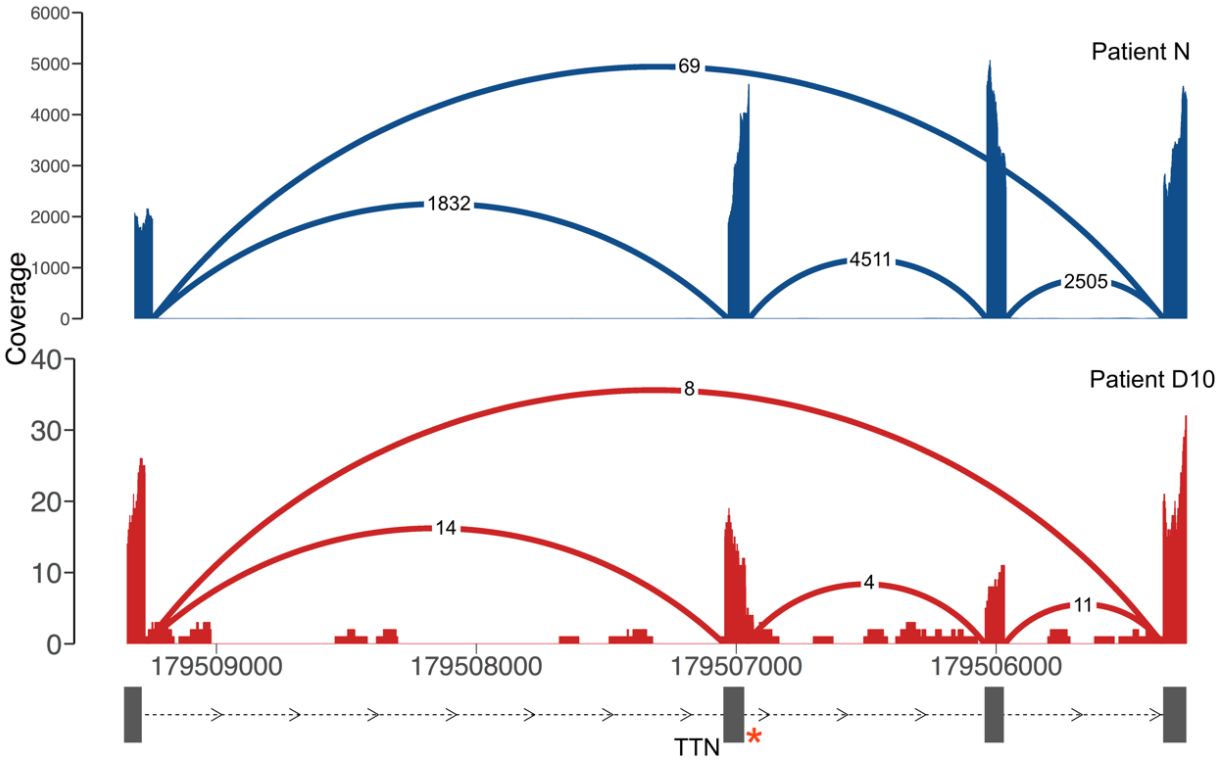
1- Splice junction normalization

To identify potentially pathogenic splice evente, we may want to identify junctions that are unique to one individual. However, a pathogenic splice junction that only occurs in one sample in the whole dataset may still have read support in other samples, due to mapping noise. If a gene is highly expressed and/or if some samples have more reads sequenced, this may result in junctions that are present at very low levels in a sample to still to have high read support.

Here is a real example of a splice junction that is present in two individuals:

Gene	Splice junction	Total Read Support	Number of Samples	Sample : Read Support
TTN	2:179505357-179509275	78	2	Patient D10: 8, Patient N: 69

There are 69 reads in Patient N and 8 reads in Patient D10 that support this splice junction. It's present in only two individuals and is in a known muscle disease gene, so it may be potentially interesting. From the read support data alone, it looks like it may be a real junction in Patient N and possibly mapping noise in Patient D10.



Here is a Sashimi plot of the local region in the two samples:

In fact, while there is higher read support for the junction in Patient N, the read support for the junction is only 1.5-4% of that of the canonical junctions. In contrast, the junction has 8 reads supporting it in Patient D10 but this constitutes ~50% of the read support for canonical reads, more in line with this being a heterozygous event. Patient D10 carries a heterozygous essential splice site variant (denoted by the red asterisk) which is causing exon skipping and this is more likely to be a low-level event or mapping noise in Patient N. The difference in read support between the two samples can be explained by the fact that Patient N was sequenced at 100M read depth

compared to 50M in Patient D10, and that Patient D10 carries two LOF variants in TTN, which is decreasing expression of the gene.

In order to be able to access information on the *relative* support for a junction in the context of library size and gene expression, we normalize the splice junctions by the overlapping canonical junctions. This is explained in detail in the manuscript, specifically in Supplementary Figure 5. Note that you don't have to run this normalization to be able to do filtering in the next step, but it does increase your filtering power.

in MendelianRNA-seq/Analysis

```
$ python NormalizeSpliceJunctionValues.py -splice_file
All.example.gene.NEB.list.splicing.txt --normalize > All.NEB.normalized.splicing.txt
```

This modifies the splice junction file in a few ways: it adds a column indicating the proportion of read support for the junction compared to the read support of the overlapping junction (ie. it would return 0.57 for Patient D10 and 0.04 for Patient N above). It also adds a column indicating whether both, one or neither exon-intron junctions are annotated. Both junctions being annotated can indicate a canonical splice event or exon skipping, one junction being annotated can indicate an exon extension, intronic splice-gain or exonic splice gain, and neither being annotated can indicate a structural variant

3- Splice junction filtering and visualization

The final step is to filter the file of junctions to identify potentially deleterious splice events. There are several different ways to do this filtering, such as looking for events that are only seen in one sample, with high read support. Alternatively you can look for splice junctions seen in many individuals, but only seen in one individual with read support higher than say, 20 reads (this would be one way to filter out mapping noise). You can also look for splice events that have over 100 reads supporting in the entire dataset, but seen in less than 5 individuals (to potentially identify groups of patients that have the splice event). You can also utilize the normalization scheme developed above and only look for splice events that are seen at a level of 30% of the overlapping canonical junctions.

In other words, there are many way to look at the data to identify putatively pathogenic events, and the junctions you'd like to pull out will depend on your experimental design. Below we give a few examples that recover the exon skipping event in Patient E2.

in MendelianRNA-seq/Analysis

$ python FilterSpliceJunctions.py -h

Should give you all the currently built in parameters to splice junction filtering along with their description

>Identify splice junctions that have at least 10 reads supporting the event:

$ python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -n_read_support 10

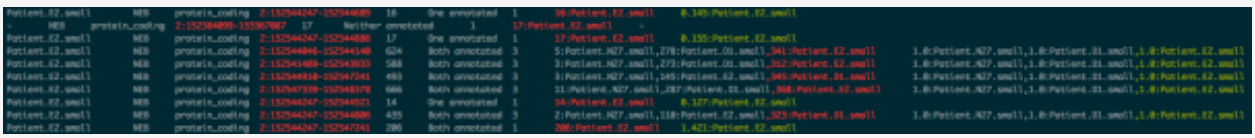
```
NEB     protein_coding  2:152544247-152544689   16      1       16:Patient.E2.small
NEB     protein_coding  2:152384099-153367087   17      1       17:Patient.E2.small
NEB     protein_coding  2:152544247-152544886   17      1       17:Patient.E2.small
NEB     protein_coding  2:152544046-152544140   624     3       5:Patient.N27.small,278:Patient.D1.small,341:Patient.E2.small
NEB     protein_coding  2:152541489-152543933   588     3       3:Patient.N27.small,273:Patient.D1.small,312:Patient.E2.small
NEB     protein_coding  2:152544910-152547241   493     3       3:Patient.N27.small,145:Patient.E2.small,345:Patient.D1.small
NEB     protein_coding  2:152547339-152548378   666     3       11:Patient.N27.small,287:Patient.D1.small,368:Patient.E2.small
NEB     protein_coding  2:152544247-152544521   14      1       14:Patient.E2.small
NEB     protein_coding  2:152544247-152544886   435     3       2:Patient.N27.small,110:Patient.E2.small,323:Patient.D1.small
NEB     protein_coding  2:152544247-152547241   286     1       206:Patient.E2.small
```

>The coloring scheme will highlight the samples with the highest read support. This requires colorama and if you have Anaconda installed, it should run. If not, you can comment out “from colorama import.. “ at the beginning of the script. In this case, you can add the -print_simple argument, which will print without coloring.

>Notice that the command above did not recover information about the normalized values of junction read support. This is set up so you can filter junctions without having to run NormalizeSpliceJunctionValues.py

>To include the normalized values run:

$ python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -include_normalized



>In some cases, you may want to filter out junctions where a samples have less than x number of reads. While in this case because we are only dealing with a 3-exon region,

we don’t have much noise, generally this helps reduce the number of non-deleterious events you identify.

$ python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -n_read_support 10 -include_normalized -filter_n_reads 5

> Notice that sample N27 only had 3 reads aligning to junction 2:152544910-152547241 (total read support for the junction : 493). When we filtered samples with less than 5 reads, this removed N27 from the line, and also decreased the number of samples the junction is seen in from 3 to 2, and subtracted the read support for the sample (total read support for the junction became 490). We usually only set filter_n_reads to 2 or 3, to filter out very low-level events.

>Identify splice junctions where the highest normalized read support is seen in a specific sample:

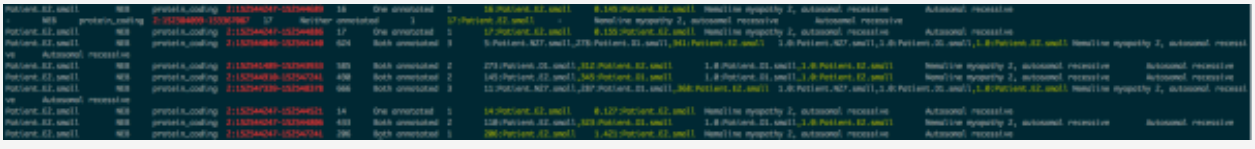
$ python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -n_read_support 10 -include_normalized -sample_with_highest_normalized_read_support Patient.E2.small

>Or simply identify junctions that are only seen in one sample

$python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -n_read_support 10 -include_normalized -n_samples 1

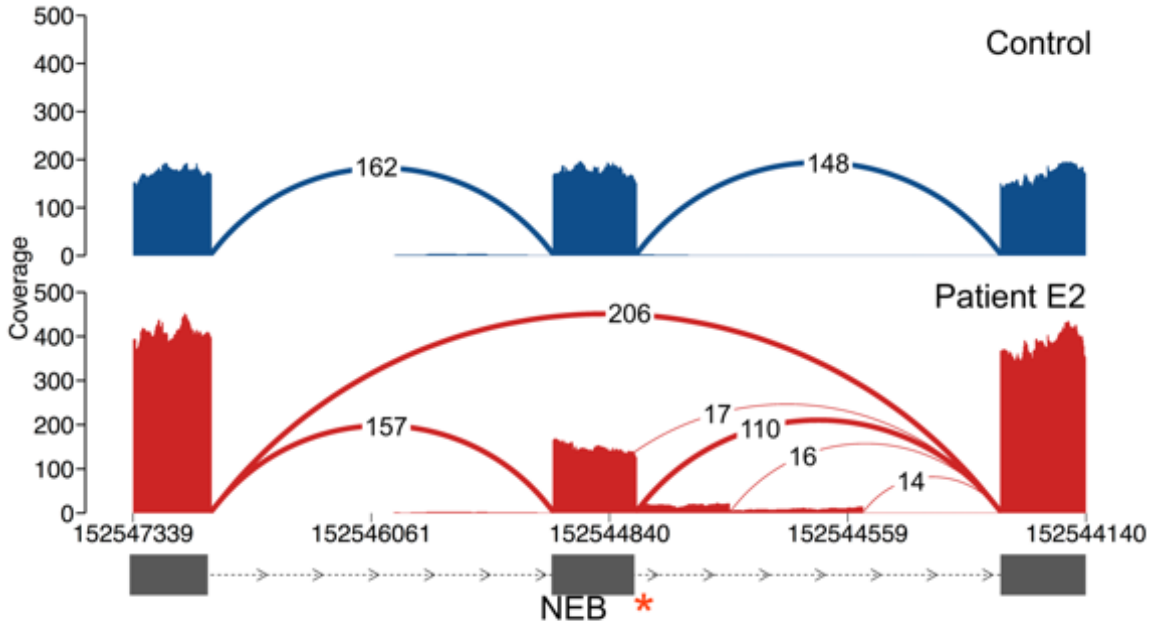
> You can also add OMIM information to the junction file:

$ python FilterSpliceJunctions.py -splice_file All.NEB.normalized.splicing.txt -n_read_support 10 -include_normalized -filter_n_reads 5 -add_OMIM



> To only show junctions occurring in OMIM genes you can add -only_OMIM. You can also only show junctions in specific genes of interest by adding say, -genes NEB,TTN,CAPN3.

In this example, searching for junctions that are present in one sample leads us to the essential splice site in this patient. In fact, all four splice junctions that we identify as only being present in Patient E2 are due to this single variant, which is abolishing splicing at the canonical junction, resulting in both exon skipping as well as splicing from intact motifs within the intron.



Future plans

We hope that the steps described above will give those interested a start to applying splice junction analysis to their own patient cohorts.

However, what's needed long-term is a fully functional (and largely automated) tool for RNA-seq-guided diagnosis that is usable by non-bioinformaticians. To that end, we are currently working on the development of a web-based tool to perform such analyses and visualize splice junctions, without requiring any command-line experience. We'll announce that tool here as soon as it's ready.

We are also continuing to apply RNA-seq as part of the Broad Center for Mendelian Disease Genetics, with a focus on novel disease gene discovery. In this effort, we have expanded RNA-seq out to other areas and tissue types such kidney biopsy RNA-seq from patients with Mendelian forms of kidney disease. We are also interested in exploring the use of proxy tissues and are performing fibroblast RNA-seq for a variety of Mendelian disorders. If you are interested in submitting genetically undiagnosed patients for RNA-seq as part of the CMG, please visit https://cmg.broadinstitute.org/Apply.

Transcript expression-aware annotation

Welcome to the repository for the "Transcript expression-aware annotation improves rare variant discovery and interpretation" manuscript. Here we'll outline how to get transcript expression values for your variant file and isoform expression expression matrix of interest, and outline the commands and code to recreate analyses in the pre-print.

Applying transcript expression aware annotation to your own dataset

You will need

1. A variant file that has columns for chrom, pos, ref and alt
2. An isoform expression matrix
3. The ability to use Hail locally or on a cloud platform

You can have additional columns in your variant file, which will be maintained, and only new columns of transcript-expression annotation will be added. Your isoform expression matrix must start with two columns : 1. transcript_id and 2. gene_id. The remaining columns can be any samples or tissues. If you have biological replicates, they should be numbered with a '.' delimiter (e.g. MuscleSkeletal.1, MuscleSkeletal.2, MuscleSkeletal.3).

Instructions to set up Hail can be found in the Hail docs

If you're unable to set up Hail in your local environment, we have released the pext values for every possible SNV in the genome: gs://gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvs.tx_annotated.021819.tsv.bgz

Please be aware that while we don't expect any issues, the files may be iterated upon until publication of the manuscript!

We will walk through an example of annotating *de novo* variants in autism and developmental delay / intellectual disability with the GTEx v7 dataset.

0) Start a cluster and a Hail environment

We recommend using cloud tools from Neale lab for Google Cloud.
You will need the gnomAD and tx-annotation init scripts, which are both publically available. To start a cluster:

```
cluster start tutorial --worker-machine-type n1-highmem-8 --spark 2.2.0 --version 0.2
--init gs://gnomad-public/tools/inits/master-init.sh,gs://gnomad-public/papers/2019-
tx-annotation/tx-annotation-init.sh --num-preemptible-workers 8
```

At the top of your script specify `from tx_annotation import *` which will start a Hail environment, and import necessary parts of the gnomAD repository.

1) Prepare the variant file

The variant file we'll be using for the tutorial is available at : gs://gnomad-public/papers/2019-tx-annotation/data/asd_ddid_de_novos.txt

This is what the first line of the file looks like :

DataSet	CHROM	POSITION	REF	ALT	GENE NAME	VEP_functional_class
ASC_v15_VCF	1	94049574	C	A	BCAR3	splice_donor_variant

Again, we will only use the chrom, pos, ref, alt columns, and will add additional columns. The VEP columns in the file are based on the canonical transcript, so we will re-VEP.

In order to add pext values, you must annotate with VEP. This is how to import the file into Hail, define the variant field, vep, and write the MT. Note that this VEP configuration will also annotate with LOFTEE v.1.0

1 - Import file as a table

```
rt = hl.import_table("gs://gnomad-public/papers/2019-tx-annotation/data/asd_ddid_de_novos.txt")
```

2 - Define the variant in terms of chrom:pos:ref:alt and have Hail parse it, which will create locus and alleles fields

```
rt = rt.annotate(variant=rt.CHROM + ':' + rt.POSITION + ":" + rt.REF + ":" + rt.ALT)
rt = rt.annotate(** hl.parse_variant(rt.variant))
rt = rt.key_by(rt.locus, rt.alleles)
```

3 - Make a MT from the Table, and repartition for speed (rule of thumb is ~2k variants per partition)

```
mt = hl.MatrixTable.from_rows_table(rt)
mt = mt.repartition(10)
```

4 - VEP and write out the MT

```
annotated_mt = hl.vep(mt, vep_config)
annotated_mt.write("gs://gnomad-public/papers/2019-tx-annotation/results/de_novo_variant/asd_ddid_de_novos.vepped.021819.mt")
```

2) Prepare the isoform expression file

We'll use the GTEx v7 isoform expression file. Here is what the header of the file looks like

transcript_id	gene_id	GTEX-1117F-0226-SM-5GZZ7	GTEX-1117F-0426-SM-5EGHI	GTEX-1117F-0526-SM-5EGHJ
ENST00000373020.4	ENSG00000000003.10	26.84	4.13	13.54

We've replaced the sample names with unique tissue names, so that samples with the same tissue are labelled as WholeBlood.1, WholeBlood.2, WholeBlood.3 etc:

transcript_id gene_id Adipose-Subcutaneous.1 Muscle-Skeletal.2 Artery-Tibial.3
ENST00000373020.8 ENSG00000000003.14 26.32 3.95 13.23

transcript_id	gene_id	Adipose-Subcutaneous 1	Muscle-Skeletal. 2	Artery-Tibial. 3
ENST00000373020.4	ENSG00000000003.10	26.84	4.13	13.54

We first need to get the median expression of all transcripts per tissue. This can be carried out using the `get_gtex_summary()` function (the function name is a misnomer, as it can work on non-GTEx files).

```
gtex_isoform_expression_file = /path/to/text/file/with/isoform/quantifications
gtex_median_isoform_expression_mt = /path/to/matrix_table/file/you/want/to/create
get_gtex_summary(gtex_isoform_expression_file,gtex_median_isoform_expression_mt )
```

If you'd like to get mean isoform expression accross tissues and not median, add get_medians = False to the command. If you want to also export the median isoform expression per tissue file as a tsv, add make_per_tissue_file = True

Unfortunately, we can't share the per-sample GTEx RSEM file as it requires dbGAP approval. However, running this on the GTEx v7 dataset creates: gs://gnomad-public/papers/2019-tx-annotation/data/GTEx.V7.tx_medians.110818.mt which is the file used for the analyses in the manuscript and the file you can use for your annotation GTEx v7 annotation.

At this point, you'll also need a separate file with gene expression values per tissue, with the tissue names matching the median isoform expression file. For the manuscript, we directly imported gene expression values provided by GTEx, which were created using RNASeQC, from the GTEx portal website. They are available here: gs://gnomad-public/papers/2019-tx-annotation/data/GTEx.v7.gene_expression_per_gene_per_tissue.120518.kt

3) Add pext values

All you have to do at this point is import your VEP'd variant matrix table, and run the tx_annotate() function!

1 - Import VEP'd variant MT, and median isoform expression MT:

```
mt, gtex = read_tx_annotation_tables(ddid_asd_de_novos, gtex_v7_tx_summary_mt_path, "mt")
```

2 - Run tx_annotation

```
ddid_asd = tx_annotate_mt(mt, gtex,
                          tx_annotation_type = "proportion",
                          filter_to_csqs=all_coding_csqs)
```

This command by default will remove certain GTEx tissues with <100 samples, reproductive tissues, or cell lines (specified in `tx_annotation_resources` and in the manuscript).

- If you don't want to remove these tissues (or if you are not working with GTEx) specify `tissues_to_filter = None`.
- If you'd like to get the non-normalized ext values instead of pext, specify `tx_annotation_type = "expression"`.
- Not specifying `filter_to_csqs=all_coding_csqs` will add pext values to non-coding variants (which may be desired behavior based on your goals). Note that splice variants are considered coding variants here. The description of coding csqs is available in `tx_annotation_resources` as `all_coding_csqs`.

- If you're only interested in getting pext for a certain group of genes, you can specify that with `filter_to_genes`. This will return the same file, but will only add the pext values to the genes of interest. An example of adding pext while specifying genes is below (under the ClinVar - gnomAD comparison section)

The function returns your variant MT with a new field called `tx_annotation` (ddid_asd_de_novos_with_pext above).

At this point, you can choose what annotation you want to use for a given variant (for example, you may be interested in any pLoF variant, or variants found on certain set of transcripts, or just variants found on the canonical transcript - the last of which sort of defeats the point of using this method). In the manuscript we used the worst consequence accross transcripts, which is the context for which we see this method being most powerful. If you'd also like to use the worst consequence, and pull out pext values for the worst consequence, we have helper functions available:

4) Optional post-processing to pull out pext values for the worst consequence annotation

At this point you will remove all variants that did not receive a pext value (e.g. if you specific `filter_to_csqs = all_coding_csqs` this will remove noncoding variants). At this point, we don't support the OS annotation in LOFTEE, which add pLoF annotations to missense and synonymous variants (for example, a synonymous variant can be called LOFTEE HC in the latest LOFTEE release if it's predicted to affect splicing). We therefore replace OS annotations with the original annotation (ie. we replace the HC for

a synonymous variant with ""). Finally, we extract the worst consequence, and create one column per tissue.

1 - Remove variants that did not receive a pext annotation (ie. noncoding variants)

```
ddid_asd = ddid_asd.filter_rows(~hl.is_missing(ddid_asd.tx_annotation))
```

2 - Overwrite LOFTEE OS variants with original variant annotation

```
ddid_asd =
ddid_asd.annotate_rows(tx_annotation=ddid_asd.tx_annotation.map(fix_loftee_beta_nonlofs))
```

3 - Pull out worst consequence

```
ddid_asd = pull_out_worst_from_tx_annotate(ddid_asd)
```

At this point you can write out the file with `ddid_asd.rows().export("out_file")`

This will create the transcript annotated *de novo* variant file used in Figure 4 of the manuscript. We've exported the result of this code snippet here:

gs://gnomad-public/papers/2019-tx-annotation/results/de_novo_variant/asd_ddid_de_novos.tx_annotated.proprotion.021819.tsv.bgz

Analyses in manuscript

Here we'll detail the commands for obtaining pext values for some of the analyses in manuscript. This will go over the analysis of:

- Getting baselevel expression values for a gene (Figure 2B)
- Comparison of highly conserved and unconserved regions (Figure 3A)
- Comparison of % variant filtered with pext < 0.1 in haploinsufficient disease genes (Figure 4A)

Note that scripts for these and other analyses are availabel in `/analyses/` folder in this repository. The paths to the files are available in `tx_annotation_resources.py` If you find something is missing, please e-mail me at berylc@broadinstitute.org

Getting baselevel expression values

The idea here is that you annotate the expression of a given *position* as opposed to a variant consequence pair. The baselevel pext value will always be higher than any of the variant annotation pext values, because the base value is just the sum of the expression of protein coding transcripts that overlap the coding base. This baselevel value is what we show in the gnomAD browser. Just because a position has a high baselevel value though, *does not* mean that say, a pLoF at that position has a high pLoF value.

We get these baselevel values by using the sites table of all possible variant in the genome. We sum of the expression of all transcripts overlapping that base, where there's a coding consequence.

- TCF4

```
from tx_annotation import *
mt, gtex = read_tx_annotation_tables(context_ht_path, gtex_v7_tx_summary_mt_path, "ht")
gene_baselevel= get_baselevel_expression_for_genes(mt, gtex, gene_list = {'TCF4'})
gene_baselevel.export("gs://gnomad-public/papers/2019-tx-annotation/results/TCF4.baselevel.ext.021319.tsv.bgz")
```

The resulting file is used in Fig2B and Supp Fig 4.

You can specify any number of genes you want in `gene_list`. If you don't specify any genes, it will annotate all positions in the exome.

- SCN2A using fetal isoform expression

```
hbdr_fetal_path = "gs://gnomad-public/papers/2019-tx-annotation/data/HBDR.RSEM.sample_specific.tx_medians.021719.mt"
mt, hbdr_fetal = read_tx_annotation_tables(context_ht_path, hbdr_fetal_path, "ht")
gene_baselevel= get_baselevel_expression_for_genes(mt, hbdr_fetal, gene_list = {'SCN2A'})
```

This file was used in Supp Fig 6D.

Comparison of pext in highly conserved and unconserved regions

1 - Read in baselevel expresison and phyloCSF files

```
phylocsf = hl.import_table(phylocsf_file_path, impute = True)
all_baselevel_ht = hl.read_table(all_baselevel_ht_path)
phylocsf = phylocsf.annotate(chrom = phylocsf.chromosome_name.replace("chr",""))
```

Note that all_baselevel_ht_path is the file used to create the tx_annotation tracks in the gnomAD browser

2 - Define regions of high and low conservation, and filter remaning regions

```
phylocsf = phylocsf.annotate(conservation_type = hl.case(missing_false=True)
                             .when(phylocsf.max_score > 1000, "high")
                             .when(phylocsf.max_score < -100, "low")
                             .default('filter'))
phylocsf = phylocsf.filter(phylocsf.conservation_type != "filter")
```

3 - Make intervals in the phyloCSF file phylocsf = phylocsf.annotate(chrom = phylocsf.chromosome_name.replace("chr",""))

```
phylocsf = phylocsf.annotate(
    interval = hl.interval(hl.locus(phylocsf.chrom, phylocsf.start_coordinate),
hl.locus(phylocsf.chrom, phylocsf.end_coordinate)),
    interval_name = phylocsf.chrom + ":" + hl.str(phylocsf.start_coordinate) + "-" +
hl.str(phylocsf.end_coordinate) )
phylocsf = phylocsf.key_by(phylocsf.interval)
```

4 - Filter the baselevel expression file to the intervals of high or low conservation in the phyloCSF file

```
all_baselevel_ht = all_baselevel_ht.annotate(**phylocsf[all_baselevel_ht.locus])
all_baselevel_ht=
all_baselevel_ht.filter(hl.is_defined(all_baselevel_ht.conservation_type), keep=True)
```

5- Get mean pext in these intervals

```
mean_proportion_in_interval = (all_baselevel_ht.group_by(
symbol = all_baselevel_ht.symbol,
 ensg = all_baselevel_ht.ensg,
 enst = all_baselevel_ht.transcript_id,
 = all_baselevel_ht.interval_name,
 conservation_type = all_baselevel_ht.conservation_type).
                                aggregate(mean_of_mean_pext =

hl.agg.filter(~hl.is_nan(all_baselevel_ht.mean_prop_conservation),
hl.agg.mean(all_baselevel_ht.mean_prop_conservation))))
```

6 - Export the file for plotting

```
mean_proportion_in_interval.export("gs://gnomad-public/papers/2019-tx-
annotation/results/conservation.phylocsf.vs.pext.021219.tsv.bgz")
```

Comparison of % variant filtered with pext < 0.1 in haploinsufficient disease genes (Figure 4A)

This also serves as an example of annotating only a subset of genes in a variant table.

Here we will annotate variants in HI genes in the gnomAD exomes sites HT, the gnomAD genomes sites HT, and the ClinVar HT with pext values.

```
out_dir = "gs://gnomad-public/papers/2019-tx-
annotation/results/gene_list_comparisons/"
```

1 - Import HI genes

```
hi_genes = import_gene_list(curated_haploinsufficient_genes, gene_column="ENSGID",
ensg=True)
```

There are two options for importing gene lists, either importing ENSG IDs, or importing gene symbols. `gene_column`refers to the column in the file that contains your gene

names, if the values are ENSGs, specify `ensg = True`. You can specify `peek = True` if you'd like to just like to import the gene list file and take a peek without doing anything.

2 - Annotate gnomAD exomes

```
mt, gtex = read_tx_annotation_tables(gnomad_release_mt_path,
gtex_v7_tx_summary_mt_path, "ht")
mt = mt.filter_rows(hl.len(mt.filters) == 0)
mt_gnomad_hi = tx_annotate_mt(mt, gtex,"proportion",
                              filter_to_csqs=lof_csqs,
                              filter_to_genes=hi_genes, gene_column_in_mt="gene_id")
mt_gnomad_hi = mt_gnomad_hi.filter_rows(~hl.is_missing(mt_gnomad_hi.tx_annotation))
mt_gnomad_hi = pull_out_worst_from_tx_annotate(mt_gnomad_hi)
mt_gnomad_hi.rows().export("%sHI_genes.gnomad.exomes.r2.1.tx_annotated.021519.tsv.bgz
" %out_dir)

- `gene_column_in_mt` is one of either `gene_id` (ENSG) or `gene_symbol` and tells
the function which VEP field to look to filter to genes.

- `mt = mt.filter_rows(hl.len(mt.filters) == 0)` filters variants to only those that
are RF PASS.
```

3 - Annotate gnomAD genomes

```
mt_genomes, gtex = read_tx_annotation_tables(gnomad_genomes_release_mt_path,
gtex_v7_tx_summary_mt_path, "ht")
mt_genomes = mt_genomes.filter_rows(hl.len(mt_genomes.filters) == 0)
mt_gnomad_genomes_hi = tx_annotate_mt(mt_genomes, gtex,"proportion",
                              filter_to_csqs=lof_csqs,
                              filter_to_genes=hi_genes, gene_column_in_mt="gene_id")
mt_gnomad_genomes_hi =
mt_gnomad_genomes_hi.filter_rows(~hl.is_missing(mt_gnomad_genomes_hi.tx_annotation))
mt_gnomad_genomes_hi = pull_out_worst_from_tx_annotate(mt_gnomad_genomes_hi)
mt_gnomad_genomes_hi.rows().export("%sHI_genes.gnomad.genomes.r2.1.tx_annotated.02161
9.tsv.bgz" %out_dir)
```

4 - Annotate ClinVar

```
clinvar_mt, gtex = read_tx_annotation_tables(clinvar_ht_path,

gtex_v7_tx_summary_mt_path, "ht")

mt_clinvar_hi = tx_annotate_mt(clinvar_mt, gtex,"proportion",

                               filter_to_csqs=lof_csqs, filter_to_genes=hi_genes,

                               gene_column_in_mt="gene_id")

mt_clinvar_hi =

mt_clinvar_hi.filter_rows(~hl.is_missing(mt_clinvar_hi.tx_annotation))

mt_clinvar_hi = pull_out_worst_from_tx_annotate(mt_clinvar_hi)

mt_clinvar_hi = mt_clinvar_hi.annotate_rows(**mt_clinvar_hi.info)

mt_clinvar_hi = mt_clinvar_hi.drop("vep", "tx_annotation","info")

mt_clinvar_hi.rows().export("%sHI_genes.clinvar.alleles.single.b37.tx_annotated.02151

9.tsv.bgz" %out_dir)
```

The fields are dropped to save space