Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

Beryl B Cummings[1,2,3], Jamie L Marshall[1,2], Taru Tukiainen[1,2], Monkol Lek[1,2,4,5], Sandra Donkervoort[6], A. Reghan Foley[6], Veronique Bolduc[6], Leigh Waddell[4,5], Sarah Sandaradura[4,5], Gina O'Grady[4,5], Elicia Estrella[7], Hemakumar M Reddy[7], Fengmei Zhao[1,2], Ben Weisburd[1,2], Konrad J Karczewski[1,2], Anne O'Donnell-Luria[1,2], Daniel Birnbaum[1,2], Anna Sarkozy[8], Ying Hu[6], Hernan Gonorazky[9], Kristl Claeys[10], Himanshu Joshi[4,5], Adam Bournazos[4,5], Emily Oates[4,5], Roula Ghaoui[4,5], Mark Davis[11], Nigel Laing[11,12], Ana Topf[13], GTEx Consortium, Peter Kang[7], Alan Beggs[14], Kathryn N North[15], Volker Straub[13], James Dowling[9], Francesco Muntoni[8], Nigel F Clarke[4,5], Sandra T Cooper[4,5], Carsten G Bonnemann[6], Daniel G MacArthur[1,2]

Correspondence: danmac@broadinstitute.org

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 neuromuscular 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 the largest systematic application of transcriptome sequencing to rare disease diagnosis to date and highlights its utility for the detection and interpretation of variants missed by current standard diagnostic approaches.

The primary challenge of genome-based diagnostics is that the capacity of WES and WGS to discover genetic variants substantially exceeds our ability to interpret their functional and clinical impact [5-7]. One approach to improve the interpretation of genetic variation is to integrate functional genomic information such as RNA-seq, which provides direct insight into transcriptional perturbations caused by genetic changes [8, 9]. Such approaches have already proven useful for elucidating mechanisms of cancer and common disease [10, 11] but have yet to be systematically applied to rare disease diagnosis.

Here we describe the application of this technology to the diagnosis of patients with a range of primary muscle disorders, including myopathies and muscular dystrophies, using RNA obtained from affected muscle tissue (Supplementary Table 1). 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 [12, 13]. 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) (Figure 1, Supplementary Figure 1) [14]. The majority of the most commonly disrupted genes in neuromuscular disease are poorly expressed in blood and fibroblasts, suggesting RNA-seq from these easily accessible tissues will often be underpowered to detect relevant transcriptional aberrations. Fortunately, primary muscle tissue is available for a substantial fraction of undiagnosed neuromuscular disease patients as biopsies are routinely performed as part of the diagnostic evaluation [15, 16].

To investigate the value of RNA-seq for diagnosis we obtained primary muscle RNA from 63 patients with putatively monogenic neuromuscular 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 (Supplementary Table 2). 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 (Figure 2a, see Methods).

Patient muscle samples were sequenced using the same protocol as the GTEx project [14] and analyzed using identical pipelines to minimize technical differences, with patients sequenced at or above the same coverage as GTEx controls. From 430 skeletal 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, ischemic time, as well as phenotypic features such as age, BMI and cause of death to more closely match our patient samples.

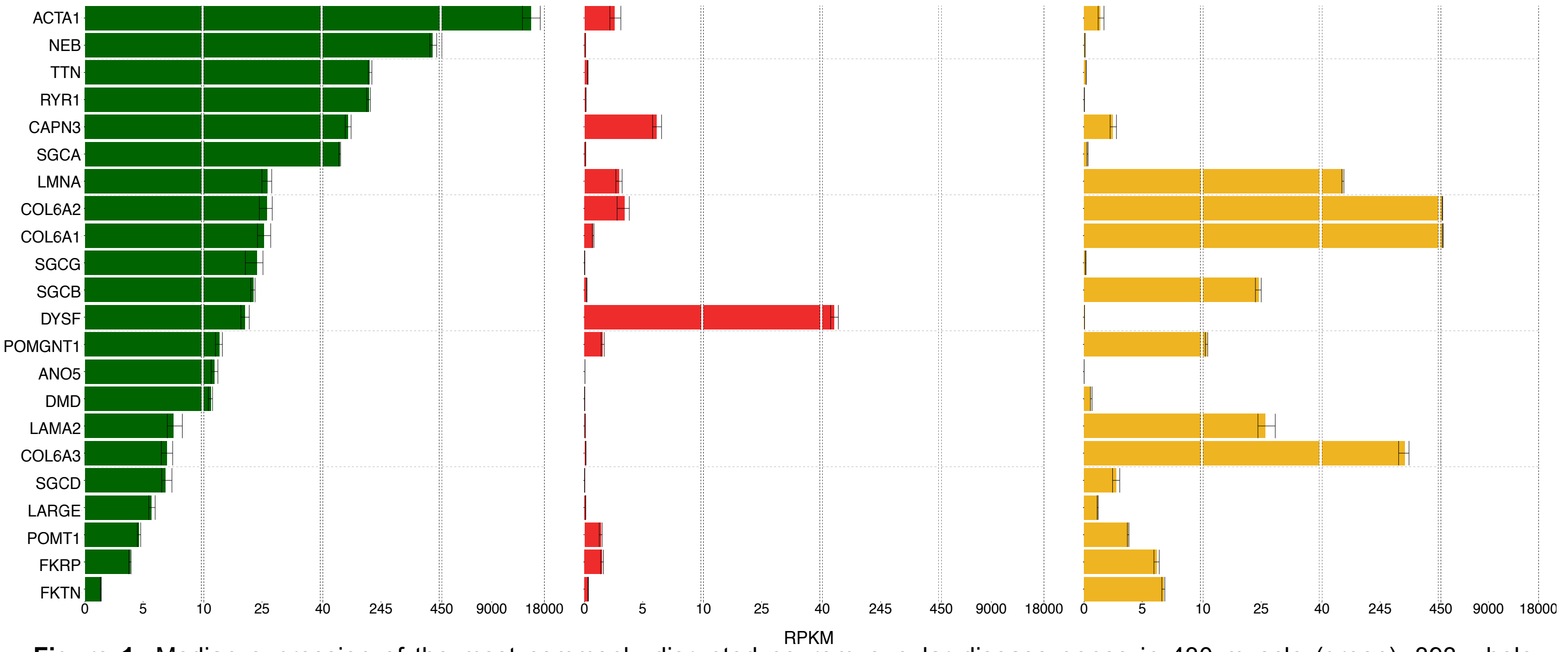

Figure 1: Median expression of the most commonly disrupted neuromusucular disease genes in 430 muscle (green), 393 whole blood (red), and 283 fibroblast (yellow) GTEx samples shows that many of these genes are poorly expressed in more accessible blood and fibroblast tissues, indicating that transcriptome analysis of these tissues may be underpowered to detect relevant transcriptional aberrations. Error bars represent 95% confidence intervals based on bootstrapping samples 10,000 times.

Comparison between our GTEx reference panel and patient muscle RNA-seq samples showed analogous quality metrics (Supplementary Table 3). Principal component analysis of expression and splicing profiles demonstrated patient muscle RNA-seq closely resembled control muscle when compared to tissues that potentially contaminate muscle biopsies, such as skin or fat (Figure 2b, Supplementary Figure 2a). Based on this clustering, we removed two samples from analysis for which expression patterns clustered more closely with GTEx adipose tissue than muscle, consistent with tissue contamination or late-stage degenerative muscle pathology (Supplementary Figure 2b). 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.

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 2c). 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.

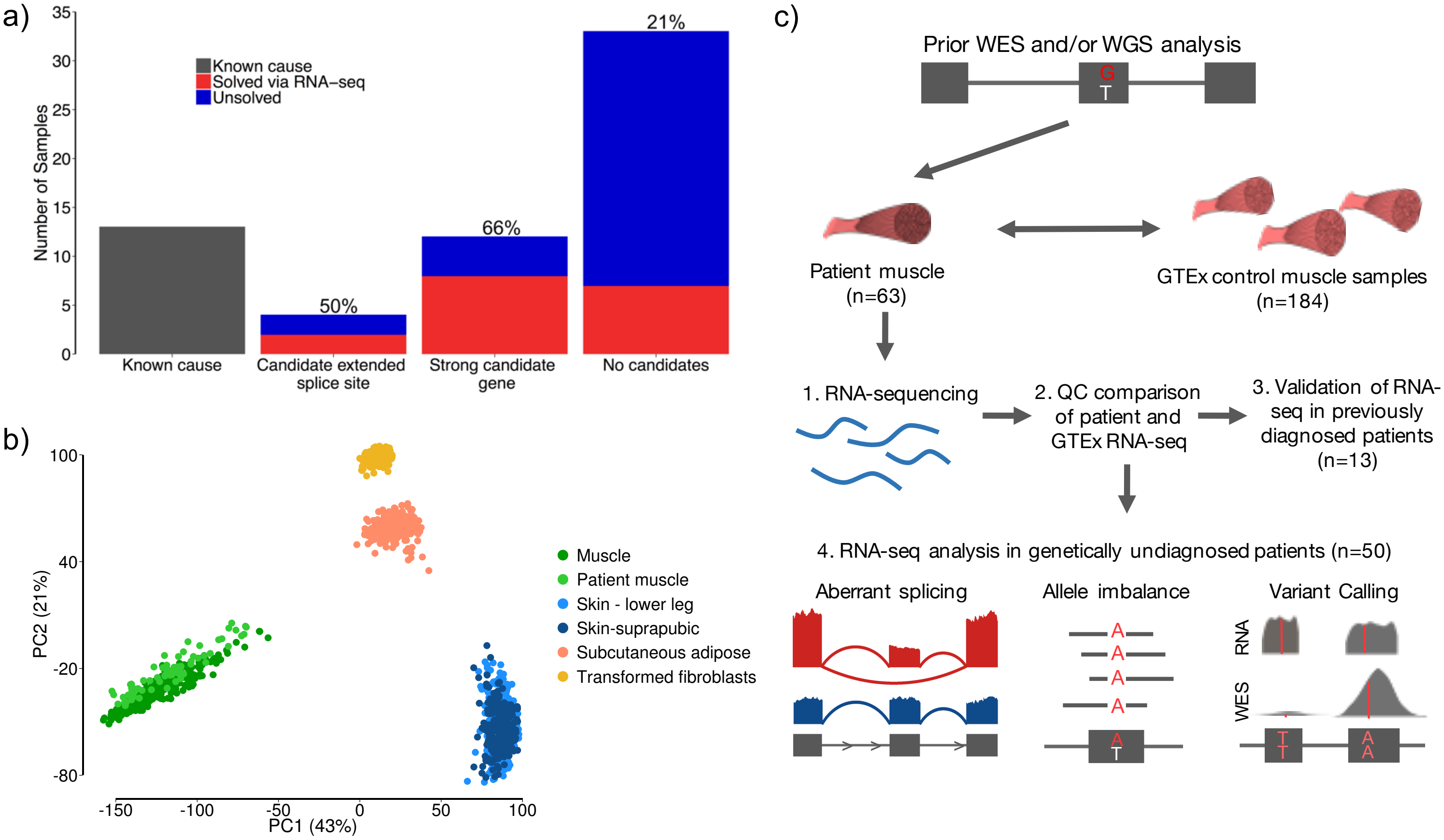

Figure 2: Experimental design and quality control **a)** Overview of the number of samples that underwent RNA-seq. We performed RNA-seq on 13 previously genetically diagnosed patients, 4 patients in whom previous genetic analysis had identified an extended splice site variant of unknown significance (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 varying depending on previous evidence from genetic analysis **b)** Principal component analysis based on gene expression profiles of patient muscle samples passing QC (n=61) and GTEx tissue samples that potentially contaminate muscle biopsies shows 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 present in patients that are missing in GTEx controls. Upon ensuring GTEx and patient RNA-seq data are 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.

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 (Supplementary Figure 3a-f). To detect such 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. 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 (Supplementary Figure 4a, b). 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 identified these previously known aberrant splice events and applied them to our remaining cohort of undiagnosed patients (Methods). This method resulted in the identification of a median of 5 potentially pathogenic splice events per sample in ~190 neuromuscular disease associated genes (Supplementary Figure 5), which require manual curation to interpret pathogenicity and led to the diagnoses made in this study.

Two commonly disrupted muscle disease genes, *NEB* and *TTN*, harbor regions with highly similar sequences, the so-called triplicate repeat regions [17, 18]. Due to high sequence similarity, reads aligning to the region have poor mapping quality, resulting in low quality variant calls that are filtered by most current diagnostic pipelines. In order to identify possible pathogenic variants in the triplicated regions of *NEB* and *TTN*, we developed a method based on remapping the triplicate regions to a de-triplicated pseudo-reference and performing hexaploid variant calling (Supplementary Figure 6a-c). This method was applied to available WES/WGS and RNA-seq data for each patient and identified one novel nonsense and one novel frameshift variant in *NEB* and *TTN* respectively, which contributed to the diagnosis of two patients (Supplementary Figure 6d, e).

RNA-seq led to 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. Detection of aberrant splicing events 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 (Figure 3, Table 1). RNA-seq patterns also helped pinpoint three structural variants in *DMD* that were subsequently confirmed by WGS (Supplementary Figure 7).

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 3a, Supplementary Figure 8a, b). 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 (Supplementary Figure 8c, d).

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 extension event caused by an underlying variant at the +3 position of the donor site as the second recessive allele (Figure 3b). The exon harboring this variant was not targeted for capture in the exome used to screen the patient (Supplementary Figure 9), underlining the utility of RNA-seq at complementing WES to identify previously undetected variants

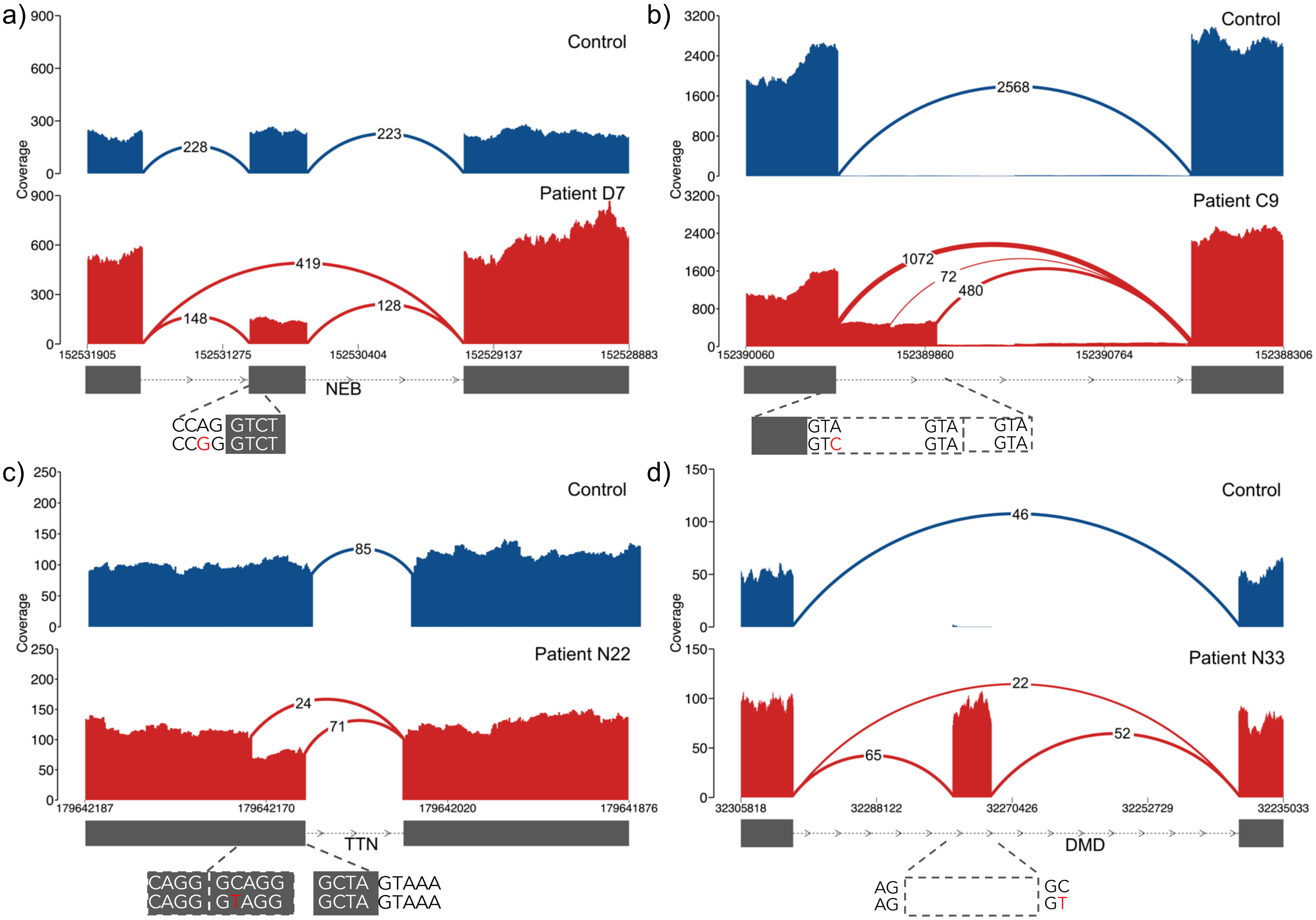

Figure 3: Types of pathogenic splice aberrations discovered in patients. RNA-seq identified a range of aberrations caused by both coding and non-coding variants such as **a)** exon skipping caused by an essential splice site variant in patient D8 **b)** exon extension caused by a donor +3 A>C extended 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 **d)** intronic splice gain in patient N33 caused by 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 muscular dystrophy phenotype. Events b, c and d result in the introduction of a premature stop codon to the transcript.

Synonymous and missense variants in large, variation rich genes such as *TTN* are exceptionally challenging to interpret and are often filtered in DNA sequencing pipelines [19, 20]. With RNA-seq we were able to assign pathogenicity to a missense variant in *TTN* and two synonymous variants in *RYR1* and *POMGNT1* in patients who were previously found to have an established pathogenic recessive variant but were missing a second pathogenic allele (Supplementary Figure 10). In a patient harboring a frameshift variant in *TTN*, 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 [21, 22]. 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 3c). A similar mechanism was observed in *RYR1*, caused by a synonymous variant (Supplementary Figure 10a,b). 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 (Supplementary Figure 10 c,d).

Table 1: 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, 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

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 (Figure 3d, Supplementary Figure 11). While RNA-seq from a patient with severe Duchenne muscular dystrophy showed only splicing to the pseudo-exon (Supplementary Figure 11a), 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. 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.

A notable example of the power of transcriptome sequencing is our discovery of a novel genetic subtype of severe collagen VI-related dystrophy, which is caused by mutations in one of three collagen 6 genes (*COL6A1, COL6A2* and *COL6A3*) [15]. In four patients negative for prior 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 4a). 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 4b). This variant is missing in the 1000 Genomes Project dataset [23] 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 [24]. Interestingly, cDNA analysis shows that the aberrant transcript is observable in muscle but at much lower levels in cultured dermal fibroblasts, making the event identifiable by muscle transcriptome analysis despite being previously missed by fibroblast cDNA sequencing (Figure 4c). 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 14 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 (Supplementary Text). In an accompanying manuscript (Bolduc et al. 2016) we describe the consistently severe clinical features of patients carrying this variant, thus defining a severe new genetic subtype of collagen VI-related dystrophy, which lends itself to an antisense oligonucleotide-based splice-modulating therapeutic approach to block the inclusion of the pseudo-exon.

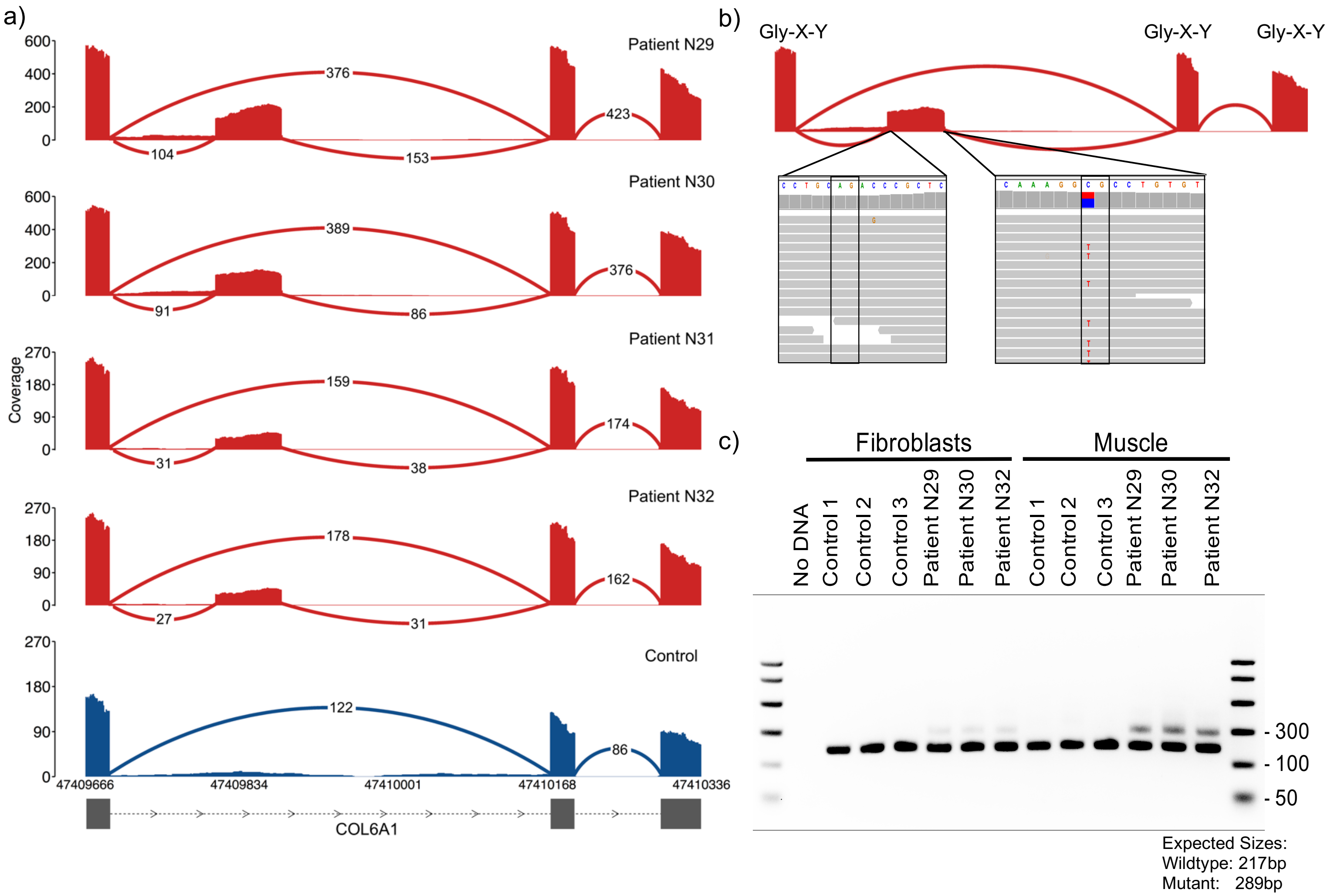

Figure 4: Identification of a recurrent splice site-creating variant in four collagen VI-related dystrophy patients **a)** Splicing in of the pseudo-exon was observed in four patients in our cohort and missing in all other patients and GTEx samples. **b)** Inclusion of the 24 amino acid segment is caused 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, likely explaining why the pathogenic event was missed in all four patients by previous fibroblast cDNA sequencing.

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 (Supplementary Figure 12). The low levels of the pathogenic intron inclusion event in *COL6A1* in fibroblasts where the gene is typically highly expressed further highlights the benefit of performing RNA-seq on the disease-relevant tissue, even in cases where the relevant genes are well expressed in a proxy tissue. Furthermore, many of the diagnoses made in this study could not have been made on genotype information alone, since splice prediction algorithms alone are currently insufficient to classify variants as causal [25,26]. While existing *in silico* algorithms predicted disruption for two extended splice site variants identified in our study, they also generated false positive predictions for two variants with no effect on splicing (Supplementary Text, Supplementary Figure 13a). In addition, existing algorithms showed poor specificity in identifying splice site creating coding and non-coding variants, indicating information on DNA sequence alone currently does not match the ability of RNA-seq to directly identify the transcriptional consequence of variants on a genome-wide scale (Supplementary Figure 13b).

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. The RNA-seq framework developed in this study can be adapted for other rare diseases where biopsies are available; for disorders where biopsy of the affected tissue is unattainable, analyses are possible through identification of proxy tissues using databases such as GTEx, or through reprogramming of patient cells into induced pluripotent stem cells and differentiation into disease-relevant tissues of interest.

In the case of neuromuscular disorders our diagnoses were made primarily through direct identification of aberrations in splicing; however, with increasing samples sizes and improvements in methods, RNA-seq can also be useful to identify somatic variants and to detect regulatory variants upstream, through analysis of expression status and allelic imbalance. Our work also 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. 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 novel pathogenic variants in known genes as well as new mechanisms for Mendelian disease.

Methods

Clinical sample selection

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

1. patients where 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 patients' 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. 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. Prior to submission all muscle samples were visually inspected, photographed, cut into 50μm sections on Leica CM 1950 model cryostat, and transferred to pre-chilled cryotubes in preparation for RNA extraction. When muscle arrived embedded in OCT, 8μm 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.

RNA sequencing

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 250ng 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. 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 [14], 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 to permit downsampling analysis of the effects of increasing RNA-seq depth.

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). A full list of GTEx sample IDs used as the reference panel can be found in Supplementary Table 4.

RNA sequencing alignment and quality-control

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 version v.2.4.2a using hg19 as the genome reference and Gencode V19 annotations. 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 (v1.1.8) on STAR aligned BAMs [27]. Principal component analysis (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 sample (Supplementary Figure 2b). 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 all samples as described in [28].

To ensure 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. [29] using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 3.4. We then calculated pairwise inheritance by descent (IBD) estimates between DNA and RNA-seq data using PLINK (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 sources for DNA and RNA-seq were the same for each patient in our dataset.

Exome and whole sequencing

Whole exome sequencing on DNA samples (>250 ng of DNA, at >2 ng/μl) was performed using Illumina or Agilent SureSelect v2 exome capture. The exome sequencing pipeline included sample plating, library preparation (2-plexing of samples per hybridization), hybrid capture, sequencing (76 bp paired reads) and sample identification QC check. Hybrid selection libraries covered >80% of targets at 20x with a mean target coverage of >80x. The exome sequencing data were de-multiplexed and each sample's sequence data were aggregated into a single Picard BAM file. Whole genome sequencing was performed on 500 ng to 1.5 ug of genomic DNA using a PCR-free protocol. These libraries were sequenced on the Illumina HiSeq X10 with 151 bp paired-end reads and a target mean coverage of >30x.

Exome and genome sequencing data were processed through a Picard-based pipeline, using base quality score recalibration (BQSR) and local realignment at known indels. The BWA aligner was used for mapping reads to the human genome build 37 (hg19). Single Nucleotide Polymorphism (SNPs) and insertions/deletions (indels) were jointly called across all samples using GATK HaplotypeCaller. Default filters were applied to SNP and indel calls using the GATK Variant Quality Score Recalibration (VQSR) and variants were annotated using Variant Effect Predictor (VEP v78); additional information on this pipeline is provided in Supplementary Section 1 of [30]. The variant call set was uploaded on to the *seqr* analysis platform (seqr.broadinstitute.org) to perform variant filtering using inheritance patterns, functional annotation and variant frequency in reference databases including ExAC [30] and 1000 Genomes [23].

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.

We then performed local normalization of per-sample read support based on the support for the highest shared annotated junction (Supplementary Figure 4a). 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 (Supplementary Figure 4b).

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. 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 the event was spliced into an existing transcript (Supplementary Figure 4a). 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). 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 (Supplementary Figure 3e). We extended filtering parameters to identify splice junctions present in less 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* (Figure 4a). 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 (Supplementary Figure 7). Code for splice junction discovery, normalization and filtering is available on https://github.com/berylc/MendelianRNA-seq. List of OMIM and neuromuscular disease genes used for splice detection and ASE analysis can be found at https://github.com/macarthur-lab/omim and https://github.com/berylc/MendelianRNA-seq)

References

1. Ankala, A., et al., *A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield.* Annals of Neurology, 2015. **77**(2): p. 206-214.
2. Yang, Y., et al., *Molecular findings among patients referred for clinical whole-exome sequencing.* JAMA, 2014. **312**(18): p. 1870-1879.
3. Taylor, J.C., et al., *Factors influencing success of clinical genome sequencing across a broad spectrum of disorders.* Nature Genetics, 2015. **47**(7): p. 717-726.
4. Chong, J.X., et al., *The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities.* The American Journal of Human Genetics, 2015. **97**(2): p. 199-215.
5. MacArthur, D., et al., *Guidelines for investigating causality of sequence variants in human disease.* Nature, 2014. **508**(7497): p. 469-476.
6. Goldstein, D.B., et al., *Sequencing studies in human genetics: design and interpretation.* Nature Reviews Genetics, 2013. **14**(7): p. 460-470.
7. Lek, M. and D. MacArthur, *The Challenge of Next Generation Sequencing in the Context of Neuromuscular Diseases.* Journal of Neuromuscular Diseases, 2014. **1**(2).
8. Wang, Z., M. Gerstein, and M. Snyder, *RNA-seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.
9. Byron, S.A., et al., *Translating RNA sequencing into clinical diagnostics: opportunities and challenges.* Nature Reviews Genetics, 2016.
10. Jung, H., et al., *Intron retention is a widespread mechanism of tumor-suppressor inactivation.* Nature genetics, 2015.
11. Li, Y.I., et al., *RNA splicing is a primary link between genetic variation and disease.* Science, 2016. **352**(6285): p. 600-604.
12. Melé, M., et al., *The human transcriptome across tissues and individuals.* Science, 2015. **348**(6235): p. 660-665.
13. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-476.
14. The GTEx Consortium *The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-660.
15. Bönnemann, C.G., et al., *Diagnostic approach to the congenital muscular dystrophies.* Neuromuscular Disorders, 2014. **24**(4): p. 289-311.
16. McDonald, C.M., *Clinical approach to the diagnostic evaluation of hereditary and acquired neuromuscular diseases.* Physical Medicine and Rehabilitation Clinics of North America, 2012. **23**(3): p. 495-563.

17. Kiiski, K., et al., *A recurrent copy number variation of the NEB triplicate region: only revealed by the targeted nemaline myopathy CGH array.* European Journal of Human Genetics, 2015.
18. 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.* Circulation research, 2001. **89**(11): p. 1065-1072.
19. 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.* Genetics in Medicine, 2015. **17**(5): p. 405-423.
20. Begay, R.L., et al., *Role of Titin Missense Variants in Dilated Cardiomyopathy.* Journal of the American Heart Association, 2015. **4**(11): p. e002645.
21. Roca, X., A.R. Krainer, and I.C. Eperon, *Pick one, but be quick: 5' splice sites and the problems of too many choices.* Genes & development, 2013. **27**(2): p. 129-144.
22. Rivas, M.A., et al., *Effect of predicted protein-truncating genetic variants on the human transcriptome.* Science, 2015. **348**(6235): p. 666-669.
23. The 1000 Genomes Project Consortium *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
24. 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.* Human Mutation, 2013. **34**(11): p. 1558-1567.
25. Spurdle, A.B., et al., *Prediction and assessment of splicing alterations: implications for clinical testing.* Human mutation, 2008. **29**(11): p. 1304-1313.
26. Duzkale, H., et al., *A systematic approach to assessing the clinical significance of genetic variants.* Clinical genetics, 2013. **84**(5): p. 453-463.C
27. DeLuca, D.S., et al., *RNA-SeQC: RNA-seq metrics for quality control and process optimization.* Bioinformatics, 2012. **28**(11): p. 1530-1532.
28. Schafer, S., et al., *Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI).* Current Protocols in Human Genetics: p. 11.16. 1-11.16. 14.
29. Purcell, S.M., et al., *A polygenic burden of rare disruptive mutations in schizophrenia.* Nature, 2014. **506**(7487): p. 185-190.
30. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-291.

Author affiliations

1. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA. USA.
2. Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
3. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA
4. School of Paediatrics and Child Health, University of Sydney, Sydney, Australia
5. Institute for Neuroscience and Muscle Research, Kids Research Institute, The Children's Hospital at Westmead, Sydney, Australia.
6. Neuromuscular and Neurogenetic Disorders of Childhood Section, Neurogenetics Branch, National Institute of Neurological Disorders and Stroke/National Institutes of Health, Bethesda, Maryland, USA
7. Division of Pediatric Neurology, Department of Pediatrics, University of Florida College of Medicine, Gainesville, Florida, USA
8. Dubowitz Neuromuscular Centre, UCL Institute of Child Health, London, UK
9. Division of Neurology, Hospital for Sick Children, Toronto, Ontario, Canada
10. Department of Neurology, University Hospitals Leuven and University of Leuven (KU Leuven), Leuven, Belgium
11. Department of Diagnostic Genomics, PathWest Laboratory Medicine, Perth, Australia

12. Harry Perkins Institute of Medical Research, University of Western Australia, Perth, Australia
13. The John Walton Muscular Dystrophy Research Centre, MRC Centre for Neuromuscular Diseases, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom.
14. Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, USA
15. Murdoch Children’s Research Institute, The Royal Children’s Hospital, Parkville, Australia

Author contributions

B.B.C, T.T., D.G.M conceived and designed the experiments. B.B.C and T.T analyzed RNA-seq data. J.L.M, Y.H, A.Bo, and M.D. performed validation experiments. B.B.C, M.L, S.D, A.R.F, L.W, S.S, G.O’G, H.M.R, E.O, R.G, S.T.C and C.G.B. analyzed exome and whole-genome data. S.D, A.R.F, V.B, L.W, S.S, G.O’G, E.E, H.M.R, A.S, H.G, K.G.C, E.O, R.G, N.G.L, A.T, A.Be, P.B.K, K.N.N, V.S, J.D, F.M, N.F.C, S.T.C. and C.G.B. provided patient samples and clinical information. F.Z, B.W, K.J.K, A.O’D-L, D.B. and H.J. contributed reagents/materials/analysis tools. J.L.M, T.T, M.L, S.D, A.R.F, V.B, L.W, S.S, K.J.K, A-O’D-L, E.O, N.G.L, A.T, J.D, C.G.B and S.T.C. critically evaluated the manuscript. B.B.C. and D.G.M. wrote the manuscript.

Acknowledgements

We thank Hayley Brooks, Danielle Sookiasian, Meganne E. Leach, Daniel Ezzo, Jahannaz Dastgir, Anne Rutkowski, Carla Grosmann, Chamindra Konermans, Sophia Ceulemans, Mary-Lynn Chu, Ellen Moran and Katherine Matthews for sample collection and quality control. We thank Carrie Miceli, Stan Nelson Viktor Rusu and David Altshuler for sharing of control cell lines and plasmids.

This project was supported by funding from the Broad Institute’s BroadIgnite and Broad*next*10 programs. B.B.C. is supported by the NIH GM096911 Training Grant. T.T. is supported by the Academy of Finland, Finnish Cultural Foundation, Orion-Farmos Research Foundation, and Emil Aaltonen Foundation. M.L. Australian National Health and Medical Research Council CJ Martin Fellowship, Australian American Association Sir Keith Murdoch Fellowship and the MDA/AANEM Development Grant. L.W., S.S., N.L., N.C., and K.N. are supported by National Health and Medical Research Council of Australia (1080587, 1075451, 1002147, 1113531, 1022707, 1031893). K.J.K. is supported by NIGMS Fellowship (F32GM115208). A.H.O.-L. is supported by National Institutes of Health under Ruth L. Kirschstein National Research Service Award 4T32GM007748. P.K., E.E. and H.K.M are supported by NIH R01NS080929.