Supplementary materials

Transcriptome-based variant calling and aberrant mRNA discovery enhance diagnostic efficiency of neuromuscular diseases

Supplementary Materials and Methods

RNA-seq variant calling and prioritization

RNA-seq variant calling was performed after GATK's best practices with modification. STAR aligner (v.2.7.1a) with two-pass mode was used to map raw reads into the hg19 reference genome. Picard tools marked duplicate reads in the resulting BAM files. GATK's (v.3.7.0) SplitNCigarReads was utilized to extract exon segments and hard clip intronic regions along with reassigning read mapping quality. Base quality recalibration was performed with GATK's BaseRecalibrator. GATK's Haplotypecaller in GVCF mode was used to call variants and filter variants with phred-scale confidence score below 20. GVCF files were combined into cohort-GVCF file and genotyping was performed with GATK's GenotypeGVCFs. Variants were filtered with following filters (FS > 30.0, QD < 2.0) and normalized by bcftools (v1.3). SnpEff and SnpSift was used to annotate variants with information from 1000Genomes, gnomAD, ExAC, KOVA, UK10K, ClinVar, OMIM, and in-house database.

Variant prioritization was performed with Exomiser (v.12.1.0), after its performance was compared with those of Divine, and DeepPVP (Table S3). First, using GATK Select Variants and bcftools, indels exceeding 70bp and variants with RNA-seq coverage < 5 were excluded. Phenotype information including clinical symptoms, laboratory test results, and pathology reports was converted into the HPO terms. Exomiser was used to further filter variants and prioritize genes. Resulting variants with allele frequency > 1% in 1000Genomes, TOPMED, UK10K, EXAC, or GNOMAD exomes were removed. Variants annotated in 5'/3'UTR, noncoding transcript exon, upstream/downstream gene, intergenic, or intron were removed. Inheritance filter was set as default (UNDEFINED). Both omimPrioritiser and hiPhivePrioritiser in Exomiser were used to score each gene. We evaluated whether pathogenic gene identified by WES was prioritized within the top 10 candidates ranked by EXOMISER_GENE_COMBINED_SCORE (Table S3).

Aberrant splicing analysis

The LeafcutterMD software was employed to detect aberrant splicing events. RNA-seq BAM files were converted into junction files by quantifying the usage of each intron. Introns in junction files were clustered into intron clusters with default parameters. Outlier intron excision events were detected by modeling intron usage into mathematical model (Dirichlet-multinomial model) and outlier detection process. Cluster *P*-values were multiple-corrected by Benjamini and Hochberg method and adjusted *P*-values < 0.05 were selected as a candidate of aberrant splicing events. Finally, calls residing in muscle disease-related genes were manually inspected and ggsashimi was used for visualization.

Allele specific expression analysis

In order to perform allele specific expression analysis, we adapted the python script presented by Castel *et al*. In short, heterozygous variants with a coverage depth >20 and a minor allele frequency between 0.3 and 0.7 were extracted from the normalized unfiltered WES VCF file. The positions of these variants were converted into a BED file, and Samtools mpileup was used on RNA-seq BAM files. Information from the WES VCF file and mpileup output file carrying the expression information for all heterozygous exome variants covered in RNA-seq were combined into a final file. Minor allele frequency of variants in RNA seq VCF (MAF < 0.3 or MAF > 0.7) were used to define allelic imbalance. Variants showing allele specific expression patterns were manually examined for clinical relevance.

Gene expression outlier analysis

Gene expression outliers were detected by OUTRIDER package in R. RNA-seq raw counts matrix were combined into one matrix and used as an input file. After removing low expressed genes

with filterExpression function, OUTRIDER function was used to calculate adjusted *P*-values for each gene. Outlier genes with adjusted p value < 0.05 were manually inspected for correlation with patient clinical symptoms.

ASO design, primary myoblast culture, and transfection

Twenty-nine antisense oligonucleotides (ASOs), 20-24 nucleotides (nts) in length, were designed (13 ASOs for CDC_NM54.1 and 16 ASOs for CDC_NM55.1) to target splice donors, splice acceptors, or exonic splicing enhancers of each cryptic exon using the published splicing prediction algorithms (Table S8). The ASOs had full 2'-*O*-methoxyethyl and phosphorothioate chemical modifications, and were synthesized at Microsynth (Balgach, Switzerland). An FDA-approved drug against spinal muscular atrophy Nusinersen and a non-targeting oligonucleotide with the same chemical modifications as the 29 ASOs were also synthesized as controls. Due to the limited supply of cells, 12 of them which were selected for subsequent experiment. For off-target prediction, each ASO sequence and its variations (allowing internal and terminal mismatches) were aligned on the UCSC hg19 reference genome and the refSeq transcriptome using BWA (version: 0.7.17). None of the 12 ASOs that were used in the experiment predicted off-targets with functional significance (i.e., antisense to the exonic sequences or introns within 1 kb from the nearby exons defined by the refSeq), even after allowing internal mismatches up to 2 nts or end-trimming by up to 4 nts.

Patient primary myoblasts were isolated from muscle biopsy tissue. Muscle tissue was chopped in skeletal muscle growth media (CC-3245, Lonza, Basel, Switzerland) and further digested in Trypsin-EDTA. After centrifugation, chopped muscle pieces were explanted to culture dish to allow cell outgrowth for 2-3 weeks. When cells reached 70-80% confluency, cells were seeded in 24-well plate coated with 0.5% Matrigel (354234, Corning, Corning, NY, USA) at a density of 3 x 10^4 cells per well. Reaching 90% confluency, medium was switched to DMEM (SH30243.01,

Hyclone, Logan, UT, USA) supplemented with 5% fetal bovine serum and penicillin-streptomycin to facilitate differentiation. Cells were transfected with Lipofectamine 3000 (3 μl/ml, L3000001, Invitrogen, Waltham, Massachusetts) along with 2 μl/mL P3000, and 100/200/400 nM ASO. 48 hours after transfection, total RNA was isolated and reverse transcribed with oligo dT and random hexamers for subsequent PCR and Sanger sequencing confirmation.

Unsupervised clustering and cell type deconvolution

Non-negative matrix factorization (NMF) was performed following a previous study utilizing R package NMF. The trimmed mean of M-values was used for normalization of the raw counts matrix. The top 25% most-variable genes were selected based on absolute deviation from the median. To optimize cluster number *k*, 50 random initialization and NMF runs were performed with *k* ranging from 2 to 20. Stability of clustering was evaluated following the cophenetic correlation coefficient and dispersion coefficient. The chosen cluster number was $k = 11$ with which 500 random initialization and NMF runs were performed and the best factorization was used for further analysis. Marker genes for each cluster were extracted using a gene scoring algorithm implemented in the NMF package. GO analysis was performed using Goseq. Muscle cell types were deconvoluted using Bisque with input marker genes from a published muscle single-cell RNA-seq study (Rubenstein AB, Smith GR, Raue U et al. Single-cell transcriptional profiles in human skeletal muscle. Sci Rep 2020; 10: 229). R^2 and *P*-value were calculated from linear regression model for fibrosis/adipose score by fibroadipogenic progenitor (FAP) cell abundance, and reciprocal of (age + 1) by satellite cell abundance. Custom Python and R scripts were used to parse and visualize results.

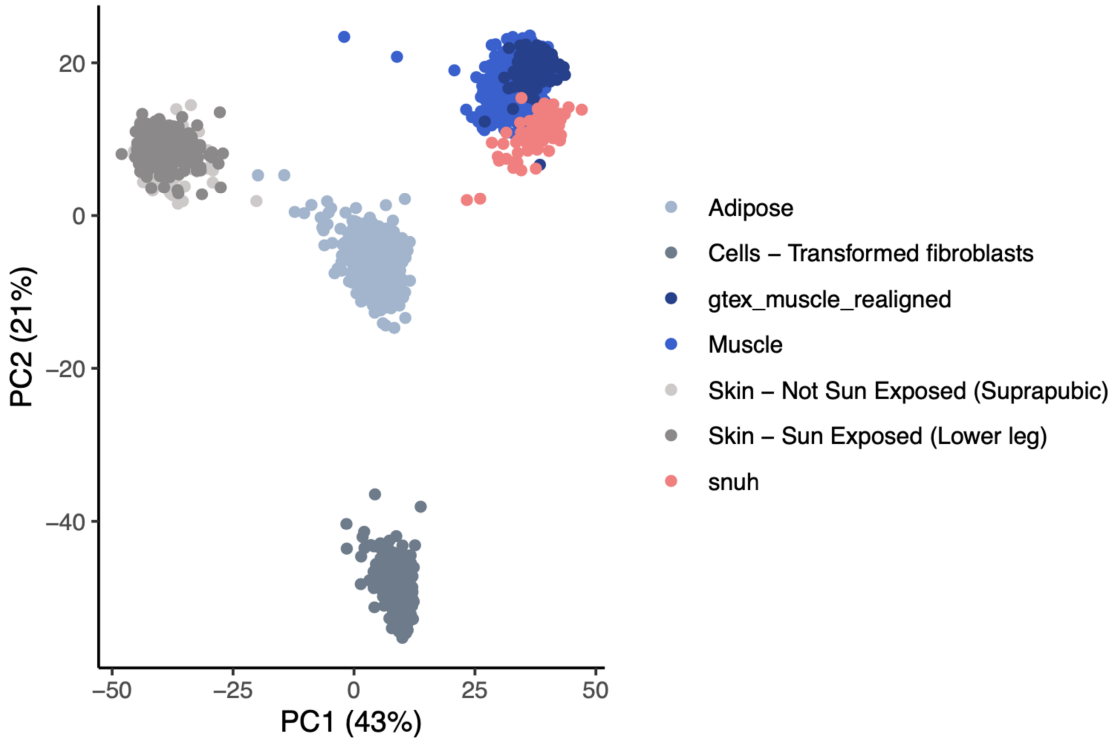

Figure S1. PCA on transcriptome data generated from GTEx and our cohort. PCA plot of transcriptome data of muscle, fibroblasts, and skin from GTEx and transcriptome data of muscle from our patient cohort (“snuh”). Muscle transcriptome from our study co-clustered with the GTEx muscle set.

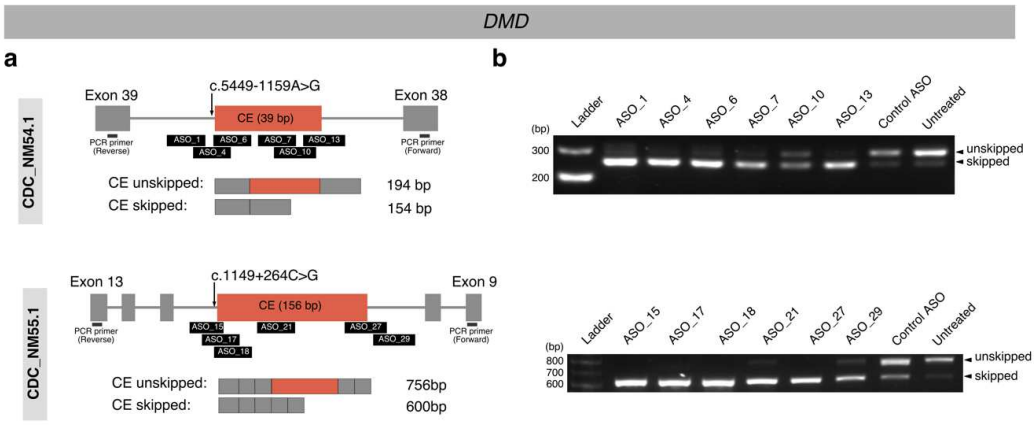

Figure S2. Different ASOs targeting the CEs consistently induced CE skipping. (A) ASOs were designed to target different sites within or near CEs of *DMD* in patients CDC_NM54.1 (upper) and CDC_NM55.1 (lower). Note that a different primer set was used for the evaluation of CDC_NM55.1 patient's *DMD* transcript than those used in Fig. 2g. (B) RT-PCR products from myoblasts treated with different ASOs (200 nM) were visualized by agarose gel electrophoresis.

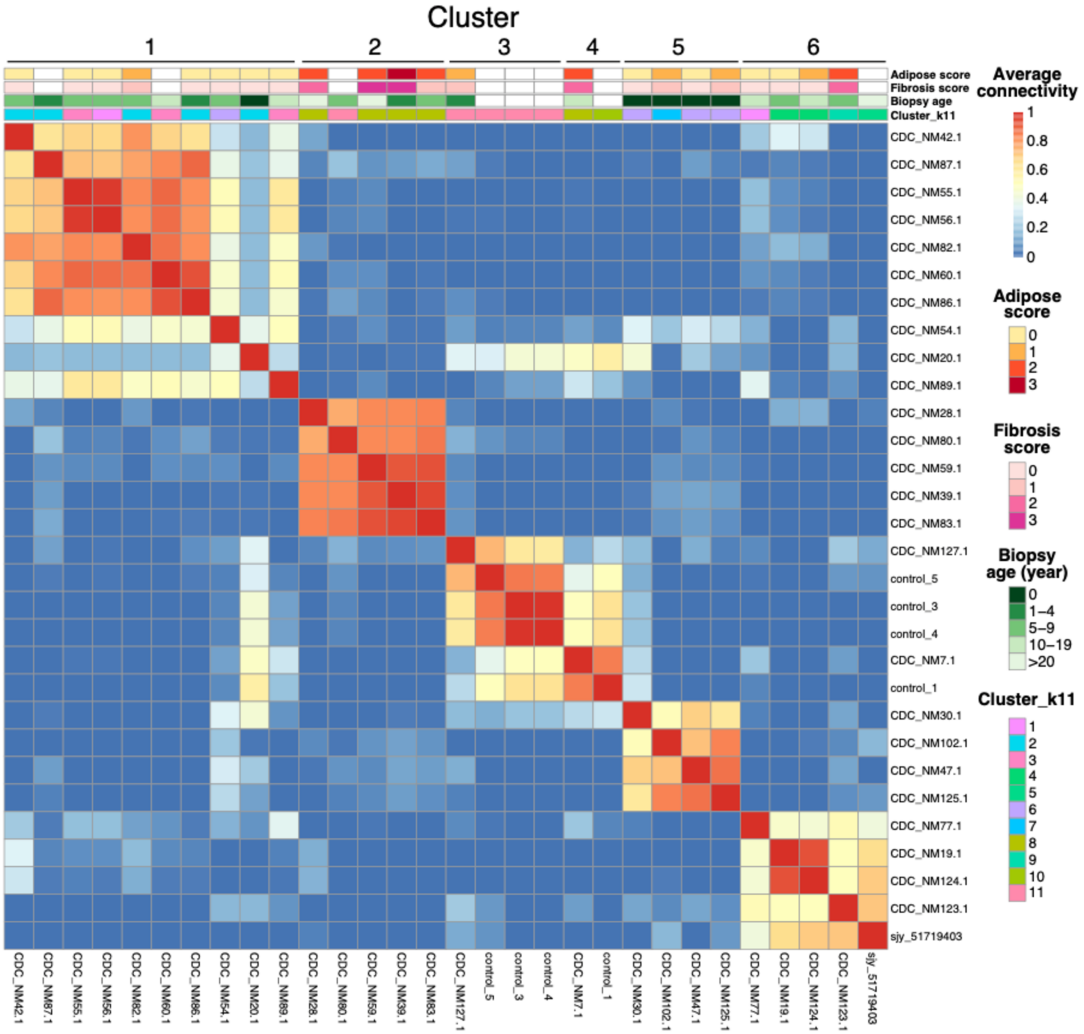

Figure S3. NMF clustering result using samples with definite answers. Twenty-six samples with identified genetic causal factors and four control samples are plotted for an independent NMF clustering.

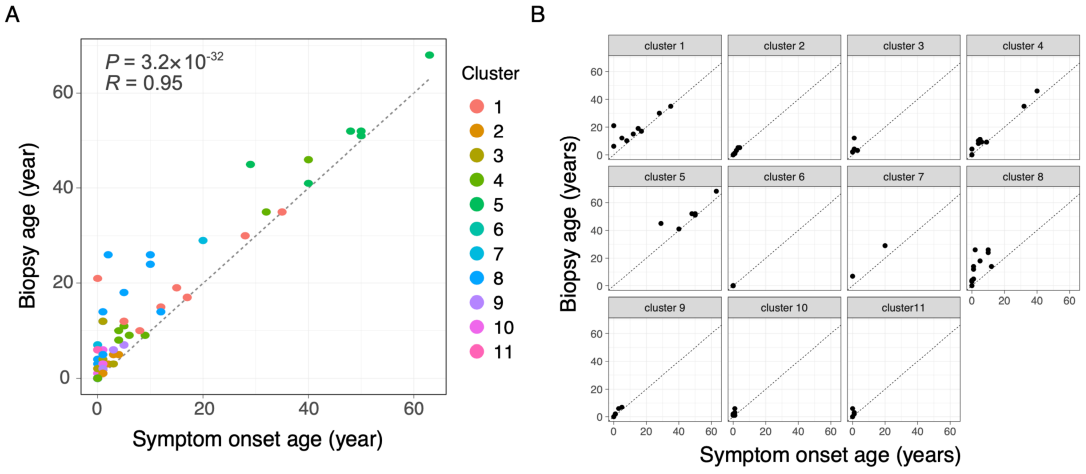

Figure S4. Differences in the ages of symptom onset and biopsy. (A) Scatter plots of symptom onset ages and biopsy ages across all samples with age data (also shown in Fig. 4e). (B) Samples divided by cluster.

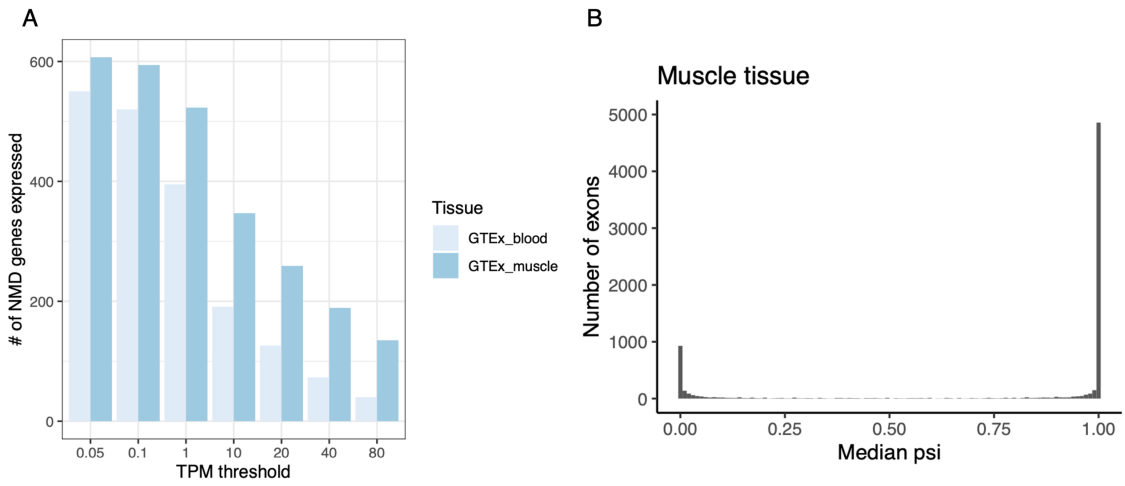

Figure S5. Expression of NMD genes in muscle and blood tissue. Curated list of NMD genes (*n* = 641) were downloaded from the 2021 version of gene table of neuromuscular disorders. (A) Expression of NMD genes were assessed in normal muscle and blood tissues, downloaded from GTEx. (B) Distribution of exon skipping percent-spliced-in (PSI) metrics of each exon in NMD genes in normal muscle tissue (GTEx).