Elucidating DUX4-mediated molecular mechanisms underlying FSHD pathophysiology using multiomics approaches

Zheng, D.

Chapter 5

DUX4 activates common and context-specific intergenic transcripts and isoforms

Dongxu Zheng[1], Anita van den Heuvel[1], Judit Balog[1], Iris M. Willemsen[1], Susan Kloet[1], Stephen J. Tapscott[2,3], Ahmed Mahfouz[1,4], Silvère M. van der Maarel[1*]

[1]Department of Human Genetics, Leiden University Medical Center, Leiden, Netherlands

[2]Division of Human Biology, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA

[3]Department of Neurology, University of Washington, Seattle, WA 98195, USA

[4]Delft Bioinformatics Lab, Delft University of Technology, Delft, Netherlands

Science Advances. 2025 May 7, 11(19):eadt5356.

https://doi.org/10.1126/sciadv.adt5356

Abstract

DUX4 regulates the expression of genic and nongenic elements and modulates chromatin accessibility during zygotic genome activation in cleavage stage embryos. Its misexpression in skeletal muscle causes facioscapulohumeral dystrophy (FSHD). By leveraging full-length RNA isoform sequencing with short-read RNA sequencing of DUX4-inducible myoblasts, we elucidate an isoform-resolved transcriptome featuring numerous unannotated isoforms from known loci and novel intergenic loci. While DUX4 activates similar programs in early embryos and FSHD muscle, the isoform usage of known DUX4 targets is notably distinct between the two contexts. DUX4 also activates hundreds of previously unannotated intergenic loci dominated by repetitive elements. The transcriptional and epigenetic profiles of these loci in myogenic and embryonic contexts indicate that the usage of DUX4-binding sites at these intergenic loci is influenced by the cellular environment. These findings demonstrate that DUX4 induces context-specific transcriptomic programs, enriching our understanding of DUX4-induced muscle pathology.

Introduction

DUX4, a pioneer transcription factor involved in zygotic genome activation (ZGA), is active during the cleavage stages in preimplantation embryos (1-3). Thereafter, it becomes repressed in most somatic tissues, including skeletal muscle (4, 5). Incomplete repression of DUX4 in skeletal muscle is the primary cause of facioscapulohumeral dystrophy (FSHD, MIM 158900) (5) where it initiates a cascade of pathological events, including aberrant RNA splicing processes (6), inflammation (7-9), and oxidative stress (10-13), eventually leading to muscle cell death. Despite the consensus on the causative role of DUX4 in FSHD pathology, its sporadic expression pattern in myonuclei confounded by a majority of nonexpressing nuclei makes it challenging to understand DUX4-associated transcriptomic changes comprehensively (14, 15).

Several studies have used short-read (SR) RNA sequencing (RNA-seq) to describe the transcriptional response to DUX4 in various cell models (16-18). Collectively, they have uncovered the transcriptomic changes activated by DUX4, deepening our understanding of FSHD pathogenesis. A study combining DUX4 chromatin immunoprecipitation sequencing (ChIP-seq) with SR RNA-seq in a myogenic cell model with DUX4 expression revealed that DUX4 binds to thousands of loci in the human genome, including known genes and transposable elements (TEs) (18). By activating long terminal repeats (LTRs), DUX4 creates novel transcription start sites (TSSs), potentially activating nearby genes (18, 19). Other studies have also reported the transcriptional activation of TEs and disruption of RNA splicing processes as molecular hallmarks of FSHD (4, 20, 21). Moreover, a study using a human embryonic stem cell (hESC) model overexpressing DUX4 identified numerous new promoters in intergenic regions with open chromatin state (22). These studies suggest that in FSHD, DUX4 induces substantial transcriptomic changes by creating new TSSs to activate downstream genes that are normally repressed in skeletal muscle and by affecting RNA processing resulting in the production of alternative transcripts.

Still, it remains challenging to comprehensively and precisely investigate transcriptomic changes by SR RNA-seq, especially in complex splicing situations, because this technology does not span full-length transcripts (23). Given the extensive alterations to the transcriptome inflicted by DUX4, including clear signs of affected RNA splicing, a full-length isoform sequencing (Iso-Seq) approach (24), complemented by SR RNA-seq providing additional sequencing depth, may better capture the “DUX4-induced transcriptome signature” and yield new insight into FSHD pathophysiology.

By combining PacBio Iso-Seq with SR RNA-seq on RNA harvested from immortalized myoblast cell lines engineered to express codon-altered DUX4 in an inducible manner (referred to as DUX4i), we obtained a comprehensive isoform-resolved transcriptome specific to the presence of DUX4 in a myogenic context. We identified a much more complex transcriptome than appreciated from prior SR RNA-seq studies, characterized by a great diversity of novel isoforms and intergenic transcripts. We validated our findings in publicly available RNA-seq datasets of human preimplantation embryos and FSHD muscle cells and found noticeable differences in isoform usage between DUX4-induced transcripts in preimplantation embryos

and skeletal muscle. Our study thus demonstrates that the DUX4-induced transcriptional landscape also depends on the cellular context.

Results

DUX4 establishes a unique and complex transcriptome in myoblasts

To obtain a comprehensive landscape of the full-length tome of DUX4-expressing myoblasts, we used doxycycline (DOX)-inducible DUX4 (DUX4i) human myoblast cell lines (iMB-clone-5 and iMB-clone-7). These cell lines were generated by introducing a DOX-inducible codon-altered DUX4 expression construct in an immortalized control myoblast cell line (see Materials and Methods for details). Both clones showed clear activation of DUX4 expression after 8 hours of DOX treatment (fig. S1A) and showed signs of DUX4-induced cell death within 24 hours of DOX treatment (fig. S1B). To balance DUX4 induction with cell viability, we treated DUX4i myoblasts with DOX for 16 hours, enabling us to capture the cascade of DUX4-triggered effects without substantial confounding effects from apoptosis. We performed PacBio Iso-Seq [referred to as long-read (LR) RNA-seq] on iMB-clone-5 before and after DOX treatment (referred to as DUX4− and DUX4+) (Fig. 1A and table S1). To offer sufficient support of splice junctions, we also performed SR RNA-seq of clone-5 and clone-7 at 0 and 16 hours of DUX4 induction with considerable sequencing coverage (average coverage: ~45 million reads per sample) (Fig. 1A and table S2). LR RNA-seq identified 57,071 and 85,165 nonredundant full-length isoforms in the DUX4− and DUX4+ samples, respectively (table S1). The transcriptomes from both conditions were aggregated to create a representative profile for the DUX4i myoblast model. After filtering the artifacts (see Materials and Methods), isoforms mapping to autosomes, sex chromosomes, and the mitochondrial genome (fig. S2A) were classified into different structural categories according to their splice junction match to the human reference transcriptome (GENCODE v39) (table S3) (25). Eventually, 84,710 nonredundant full-length isoforms associated with 13,541 genes across the two conditions were annotated (Fig. 1B). Among them, 35.6% were classified as full-splice matches (FSMs: the reference transcript and query isoform have the same number of exons and each internal junction matching), 26.6% were novel isoforms belonging to the novel not-in-catalog category (referred to as NNC: isoforms containing at least one novel splice site), and 25% to the novel incatalog (referred to as NIC: isoforms containing a new combination of known splice sites) (Fig. 1B).

To systematically compare the transcriptomes of the DUX4− and DUX4+ samples, we analyzed the isoforms separately according to their DUX4 condition (DUX4−, n = 45,304 isoforms; DUX4+, n = 58,820 isoforms; Fig. 1C, see Materials and Methods). Each condition showed comparable isoform length distributions, with the mean length in the DUX4− transcriptome being slightly longer than that in the DUX4+ transcriptome [mean length, 2,200 base pair (bp) in DUX4− and 2114 bp in DUX4+; Student's t test, P value <0.001] (fig. S2B). Under the DUX4− and DUX4+ conditions, 68.7 and 67% of isoforms, respectively, had more than two read counts, indicating considerable sequencing coverage (fig. S2C). However, the DUX4+ transcriptome displayed increased transcript diversity, as evidenced by a greater number of isoforms with lower read counts and isoforms unique to the DUX4+ condition (fig. S2, C and D). In addition, the DUX4+ myoblasts displayed a more complex transcriptome, characterized

by a higher proportion of NNC isoforms (28.9 versus 15.5%), incomplete splice match isoforms (referred to as ISM: isoforms that match a subsection of a known transcript, 8.8 versus 6.5%), intergenic isoforms (referred to as Intergenic: isoforms localized in intergenic regions, 2.2 versus 0.1%) and a lower proportion of FSM isoforms (34.4 versus 51.1%) (Fig. 1C). The SR RNA-seq data with deeper sequencing depth provided robust support for the isoforms identified by the LR RNA-seq data (Fig. 1D). Different from FSM isoforms, fewer intergenic (94.6 versus 98.2%), antisense (94.7 versus 99.1%), fusion (89.3 versus 97%), and genic (91 versus 95.3%) isoforms were supported by the SR RNA-seq data of the DUX4+ sample (Fig. 1D). This is possibly due to mapping issues in the SR RNA-seq. Moreover, fewer isoforms from the DUX4+ sample were supported by publicly available data of TSS [5′ Cap Analysis of Gene Expression (CAGE)] (26) and polyadenylation (polyA) motifs, suggesting the presence of novel TSS and polyA sites in the DUX4+ condition (Fig. 1D).

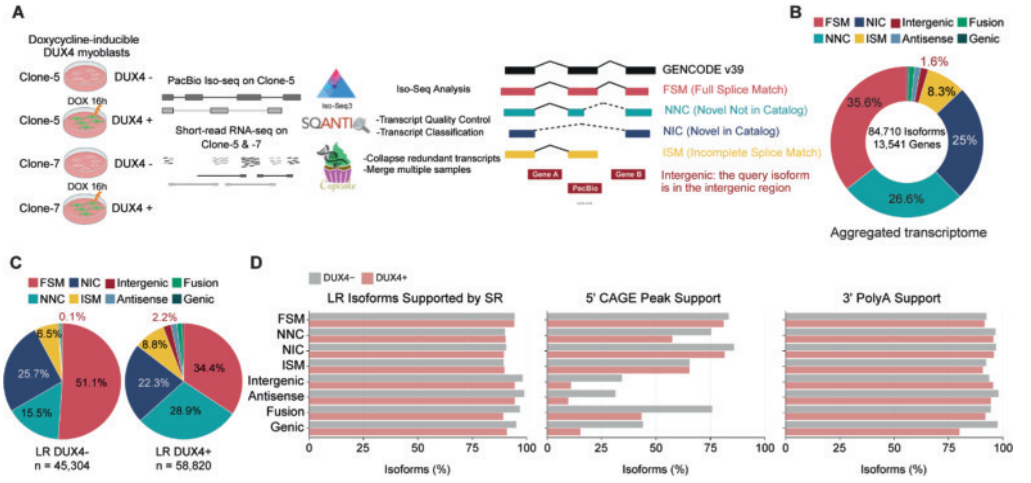

Fig. 1. The landscape of full-length transcriptome in DUX4i myoblasts.

(A) Schematic of experimental design and full-length isoform profiling by integrating LR and SR RNA-seq data. Full-length isoforms in LR RNA-seq are classified into different structural categories using SQANTI3 compared with the GENCODE (v39) isoform structure. Color codes represent each structural category. Gray dashed lines depict the novel splice junctions and black solid lines depict the known splice junctions. The figure is created with Biorender.com. (h, hours). (B) Donut plot showing the percentage of the aggregated full-length transcriptome classified into each structural category. (C) Pie charts, colored as in (B), showing the percentage of full-length isoforms classified into each structural category in the DUX4− and DUX4+ transcriptome separately. (D) Bar plots depicting the percentage of isoforms detected by SR RNA-seq data (left), isoform TSS supported by CAGE (middle), and isoform TT S supported by the presence of a poly(A) motif (right) in each structural category. Color codes represent the DUX4 conditions.

To assess the reliability of the DUX4i cell system in recapitulating FSHD pathology and the capability of Iso-Seq to capture typical DUX4-induced signatures, we validated three well-characterized hallmarks of DUX4 misexpression in muscle cells in our system. First, of 67

known core DUX4-target genes (27), 49 and 50 were detectable in LR and SR RNA-seq data of DUX4+ samples, respectively (fig. S3A). A few DUX4-target genes showed low expression in DUX4− samples, possibly due to the leakiness of the DUX4i expression construct. Second, DUX4 is known to disrupt the nonsense-mediated decay (NMD) pathway (17, 21, 28). The LR RNA-seq of the DUX4+ sample showed an increased number of NMD-targeted isoforms compared to the DUX4− sample (10.6 versus 5.9%; 6209 of 58,820 versus 2695 of 45,304) (fig. S3B). Compared to the DUX4− condition, NMD-targeted isoforms showed higher expression levels in both LR and SR RNA-seq data (fig. S3, C and D). Last, DUX4 activates transcripts from repetitive elements (REs) (2, 4, 18). Overall, more full-length isoforms emanating from REs were identified after DUX4 induction. Specifically, a higher percentage of transcripts were transcribed from LTRs (2.3 versus 1%) and long interspersed nuclear elements (LINEs) (2.4 versus 1.6%) with significantly higher expression levels, consistent with previous studies (fig. S3, E and F) (4, 18). This analysis confirms the typical DUX4-dependent transcriptome features observed in FSHD, validating the robustness of this cell model.

To assess the reproducibility and robustness of isoforms identified by LR RNA-seq, we performed a cross-dataset validation by analyzing isoform expression levels and presence in our SR RNA-seq dataset and the RNA-seq dataset of the DUX4i iMB135 myoblast cell line (17). A total of 87.6% (74,221 of 84,710) and 84.7% (71,728 of 84,710) of isoforms identified by LR RNA-seq were detectable in our SR RNA-seq dataset and the dataset from Jagannathan et al. (17), respectively. Correlation analysis of rlog-normalized isoform expression values revealed distinct clustering of DUX4+ samples and controls (fig. S4A), indicating consistent transcriptional changes induced by DUX4 despite differences in cell lines and sequencing protocols. Differential expression analysis at the isoform level identified 11,337 up-regulated and 5,091 down-regulated isoforms in our SR RNA-seq, and 7,785 up-regulated and 1,586 down-regulated isoforms in Jagannathan et al. (17) (adjusted P value <0.05 and |log2(fold change)| > 2) (fig. S4B and table S4). A substantial overlap of up-regulated isoforms was observed between the datasets, highlighting shared transcriptional responses (fig. S4C). Analysis of the structural categories of differentially expressed isoforms revealed broadly similar proportions of FSM, ISM, NIC, and NNC categories across datasets (fig. S4D). However, our SR RNA-seq captured a greater number of uniquely differentially expressed isoforms (fig. S4E), suggesting that while novel isoforms identified by LR RNA-seq were present in the independent dataset, many were not differentially expressed likely due to differences in sequencing strategies (paired-end versus single-end), read length (150 versus 100 bp), and sequencing depth (average coverage: ~45 million versus ~20 million reads). Enrichment analysis of genes associated with the differentially expressed isoforms in both datasets showed a high degree of overlap in significantly enriched Gene Ontology (GO) terms in Biological Processes (BP) (table S5), reflecting the consistency of BPs associated with DUX4 induction (fig. S4F). Collectively, these results validate the reproducibility of full-length isoforms in independent SR RNA-seq datasets derived from different cell lines. Our LR RNA-seq analysis demonstrates that the expression of DUX4 in myoblasts establishes a unique cellular environment characterized by a more complex transcriptome than previously anticipated.

Characteristics of novel isoforms from known loci

Full-length transcripts were categorized as known (FSM and ISM) or novel (NIC and NNC) isoforms based on their splice junction match to the reference transcriptome. ISM isoforms were excluded for further analysis since they are often sequencing artifacts or the products of degraded transcripts (29). A large proportion of FSM isoforms (43.6%; 13,172 of 30,193) were shared between the DUX4− and DUX4+ transcriptome, while 33% (9,973 of 30,193) and 23.3% (7,048 of 30,193) of FSM isoforms were unique to the DUX4− and DUX4+ samples, respectively (Fig. 2A). The reverse was true for the NIC and NNC categories: Both conditions expressed unique NIC and NNC transcripts (Fig. 2A), with more NIC (45 versus 38%; 9,508 of 21,150 versus 8,034 of 21,150) and particularly more NNC isoforms (68.9 versus 24.6%; 15,498/22,499 versus 5,524 of 22,499) being identified in the DUX4+ transcriptome. This suggests a shift from known to novel isoforms in the presence of DUX4.

To specify the differences between the two conditions, all isoforms were classified into three categories: DUX4− unique, DUX4+ unique, or common to both conditions. FSM isoforms uniquely expressed in DUX4− or DUX4+ samples had fewer exons than FSM isoforms common to both conditions (fig. S5A). NNC isoforms unique to DUX4+ had fewer exons than NNC isoforms in the other two categories. They also showed a reduced transcript, CDS, and open reading frame (ORF) length (fig. S5A). Novel isoforms were prone to acquiring new TSSs and transcription termination sites (TTSs), especially for the DUX4+ unique NNC isoforms (Fig. 2B). Last, quantifying the expression in the SR RNA-seq data revealed that common FSM isoforms were highly expressed in the DUX4− samples (fig. S5B). In contrast, among all structural categories, condition-specific isoforms were correspondingly highly expressed in SR RNA-seq data. This indicates consistency between the LR and SR data (fig. S5B).

The usage of new TSS and TTS was associated with a greater probability of a premature termination codon associated with NMD. Compared to the commonly expressed isoforms, the uniquely expressed isoforms in both conditions were more likely to be predicted as NMD targets, particularly under the DUX4+ condition (fig. S6A). We further analyzed the NMD-targeted isoforms classified as NIC and NNC, with a large proportion of them coming from the DUX4+ sample (fig. S6B). The genes corresponding to the overlapping isoforms between conditions were significantly enriched in GO terms associated with RNA processing (fig. S6C). Genes corresponding to condition-specific NMD-targeted isoforms showed similar enrichment patterns, primarily in RNA metabolism and translation processes (fig. S6D and table S6), suggesting that genes involved in these BPs tend to produce more NMD-targeted isoforms. To further characterize these NMD-targeted isoforms at the protein level, we extracted their predicted ORFs and queried them in the UniProt database (30). While most ORFs matched existing entries, isoforms unique to the DUX4+ condition showed a higher proportion of novel ORFs (DUX4+ unique: 11.6%, DUX4− unique: 7.5%; table S7) (31). To validate their translational potential, we analyzed a publicly available dataset including Ribosome profiling sequencing (Ribo-seq) and RNA-seq from DUX4i myoblasts (iMB135) treated with DOX for varying durations (0, 4, 8, and 14 hours) (21). After filtering for expression threshold [transcripts per million (TPM) > 0.5], we observed that both mRNA levels and ribosome footprint abundance of the remaining target isoforms increased significantly with DOX

treatment duration, accompanied by enhanced translation efficiency (fig. S6E) (see Materials and Methods). These findings suggest that genes involved in RNA processing and translation pathways could produce numerous NMD-targeted isoforms. Furthermore, these isoforms may have substantial translational potential and biological functions at the protein level.

DUX4 increases isoform complexity by alternative splicing

The LR RNA-seq allows for accurately characterizing alternative splicing events (ASEs). In the DUX4− and DUX4+ LR-seq transcriptomes, 32,247 and 43,494 ASEs (Fig. 2, C and D) were identified, respectively. While the DUX4− transcriptome showed a relatively balanced picture of known and novel ASEs among all classes, perhaps somewhat biased because of the leakiness of the expression construct, in the DUX4+ transcriptome, the proportion of novel ASEs was systematically high across all classes (Fig. 2D). The mutually exclusive exon (MX: 82.4%; 1,694 of 2,055), retained intron (RI: 76%; 2,236 of 2,941), and alternative last exon (AL: 76.5%; 4,200 of 5,489) classes comprised the highest proportions of novel ASEs in DUX4+ myoblasts. In addition, we found that a higher proportion of uniquely expressed isoforms obtained ASEs. Novel ASEs were significantly enriched in isoforms uniquely expressed in each condition, particularly in the DUX4+ sample (73.1% in DUX4− and 83.5% in DUX4+; prop.test, *P* value <0.001) (Fig. 2E). The genes corresponding to the isoforms with MX, RI, and AL were examined separately as they were the most enriched ASEs. There was negligible overlap among the isoforms within these three ASE classes, with only 93 loci present in all categories (fig. S7A). In the LR and SR RNA-seq data, these isoforms displayed higher expression levels in the DUX4+ samples (fig. S7B). Notably, enrichment analysis revealed an overrepresentation of GO BP terms related to RNA splicing and RNA metabolism within these gene sets (Fig. 2F and table S8). This suggests that the alternative splicing of genes associated with RNA processing itself may play a role in the increased ASEs observed in DUX4-expressing cells.

Next, we assessed the potential impact of novel isoforms on biological functions at the protein level. We predicted the ORFs of coding FSM, NIC, and NNC isoforms and queried them in the UniProt database (30) to identify the percentage of them with at least one match in the UniProt database. Higher proportions of NNC and FSM isoforms uniquely expressed in the DUX4− and DUX4+ conditions had no match in UniProt compared to the common isoforms (FSM: DUX4+ unique: 14.3%; DUX4− unique: 13.9% versus Common: 7.5%; NNC: DUX4+ unique: 12.6%; DUX4− unique: 12.8% versus Common: 9.9%), while more novel ORFs (identity score < 99%) (31) were observed in the uniquely expressed NIC (DUX4+ unique: 11.8%; DUX4− unique: 8.6%) and NNC (DUX4+ unique: 13.8%; DUX4− unique: 11.4%) isoforms specific to each condition (Fig. 2, G and H). Our analysis thus suggests that DUX4 expression in myoblasts initiates the transcription of novel coding isoforms that differ from the reference transcripts. These mRNA-level alterations may potentially result in functional dysregulation at the protein level.

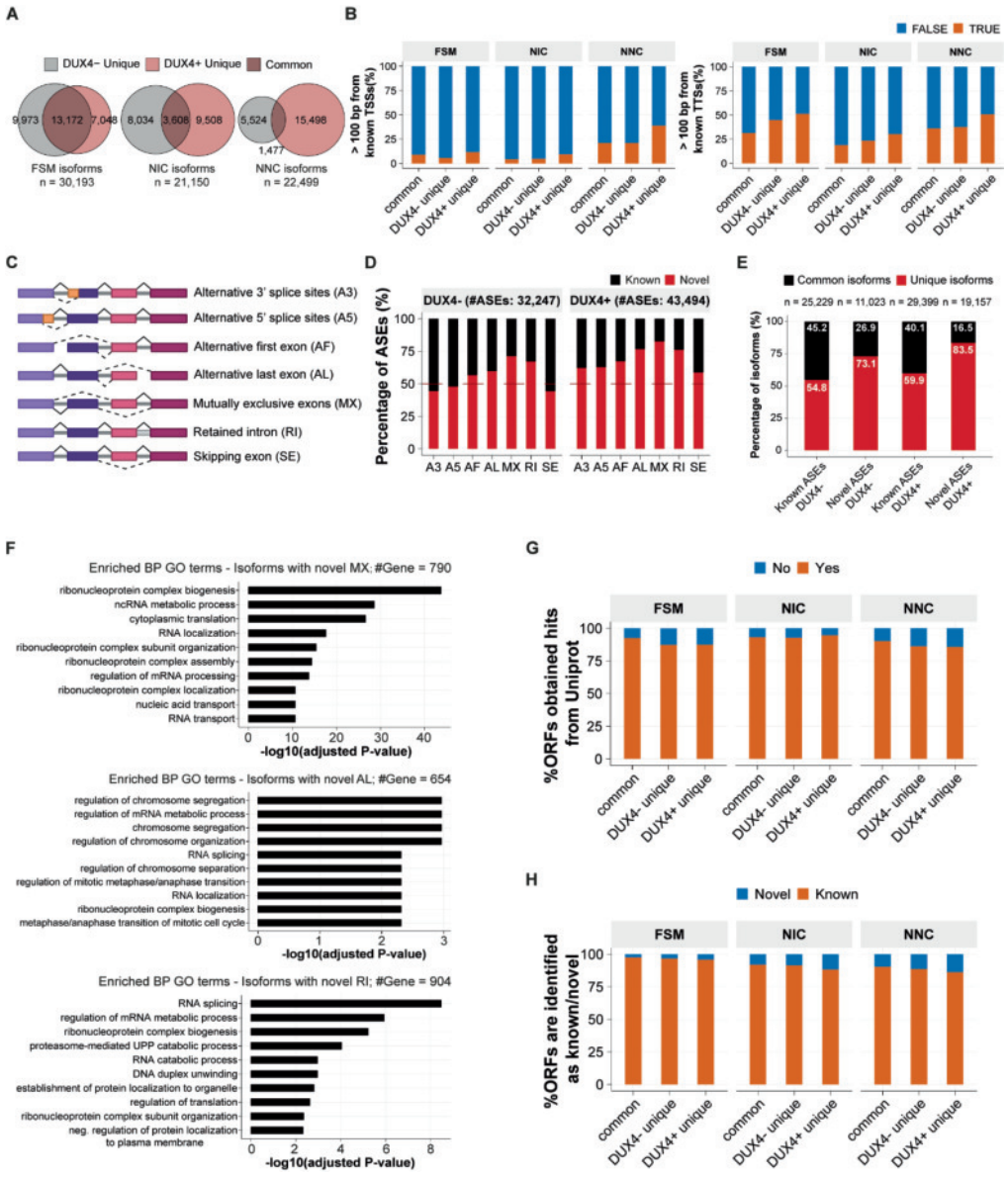

Fig. 2. Characteristics of novel isoforms from known loci.

(A) Venn diagrams showing the overlap of isoforms classified into FSM, NIC, and NNC between DUX4− and DUX4+ transcriptomes. Color codes are plotted for the subtypes of isoforms: uniquely expressed in DUX4−, uniquely expressed in DUX4+, and commonly expressed in both conditions. (B) Stacked bar plots displaying the percentage of isoforms in each subtype per structural category with or without: (left) over 100-bp shift in known TSSs compared with transcript in GENCODE; (right) over 100-bp shift in known TT Ss compared with GENCODE transcript. (C) Schematic of classification of alternative splicing events using SUPPA2. The figure is created with Biorender.com. (D) Stacked bar

plots depicting the percentage of known and novel ASEs in each category from the DUX4− and DUX4+ transcriptome. The dashed line in dark red represents the 50% to identify the ratio of known and novel events per type of ASE. (E) Stacked bar plots showing the percentage of commonly expressed and uniquely expressed isoforms with known or novel ASEs under each condition. Statistical significance was assessed using proportion test (*P* value <0.001). (F) Bar plots displaying GO terms [Biological Processes (BPs)] significantly enriched (adjusted *P* value <0.05) for genes associated with isoforms with novel RI, AL, and MX ASEs. (G) Stacked bar plot displaying the percentages of isoform-derived ORFs that obtained no match against the UniProt database, plotted by subtypes of isoforms in FSM, NIC, and NNC. (H) Stacked bar plot depicting the percentages of novel amino acid sequence identity for isoform-derived ORFs compared to their closest human protein isoform in the UniProt database, plotted by subtype of isoforms in FSM, NIC, and NNC. Novel ORFs show <99% identity with existing entries from UniProt.

Cellular context-specific binding patterns of DUX4 associate with isoform expression

Transcription factors play crucial roles in gene expression regulation, with their binding patterns and regulatory functions varying across cellular contexts. To investigate whether DUX4 regulates isoforms belonging to different structural categories in distinct cellular states, we performed an integrative analysis of DUX4 ChIP-seq data from three cell lines: DUX4i myoblasts (18), DUX4i hESCs (1), and DUX4i induced pluripotent stem cells (iPSCs) (2). Using 2-kb windows flanking the TSS, we annotated FSM, NIC, and NNC isoforms identified through LR RNA-seq. Pairwise comparisons revealed significant overlaps among FSM, NIC, and NNC isoforms containing DUX4 peaks (Fig. 3A). While a subset of isoforms across all three categories exhibited DUX4 peaks common to all cell lines, we observed notably higher overlap between DUX4i hESCs and DUX4i iPSCs across all isoform categories. This enhanced overlap might reflect their similar cellular states and higher number of identified peaks (myoblasts: 33,076, hESCs: 98,103, iPSCs: 166,514), or alternatively, their early embryonic state with more accessible genome-wide chromatin, whereas myoblasts, having established lineage commitment, may retain only core cell type-specific binding sites.

Peak annotation analysis revealed that DUX4 predominantly binds to distal intergenic regions and introns (fig. S8A). The distribution patterns of DUX4 peaks showed distinct cellular context specificity across isoform categories. Peaks from DUX4i myoblasts exhibited a more concentrated distribution pattern closer to TSS, while peaks from DUX4i hESCs and iPSCs were generally more distant from the TSSs of FSM and NIC isoforms. Notably, for NNC isoforms, peaks from all three cell lines clustered near the TTS (Fig. 3, B and C), suggesting that DUX4-binding patterns exhibit both cellular context and isoform category specificity, with perhaps the most robust binding sites preferentially occupied in a myogenic context. We further categorized isoforms with DUX4 peaks into eight groups based on their peak overlap patterns (fig. S8B). Analysis of RNA-seq data from all three cell lines revealed that isoforms generally showed relatively higher expression levels in cell lines where their DUX4 peaks were identified (fig. S8C), indicating cellular context–specific transcriptional regulation by DUX4.

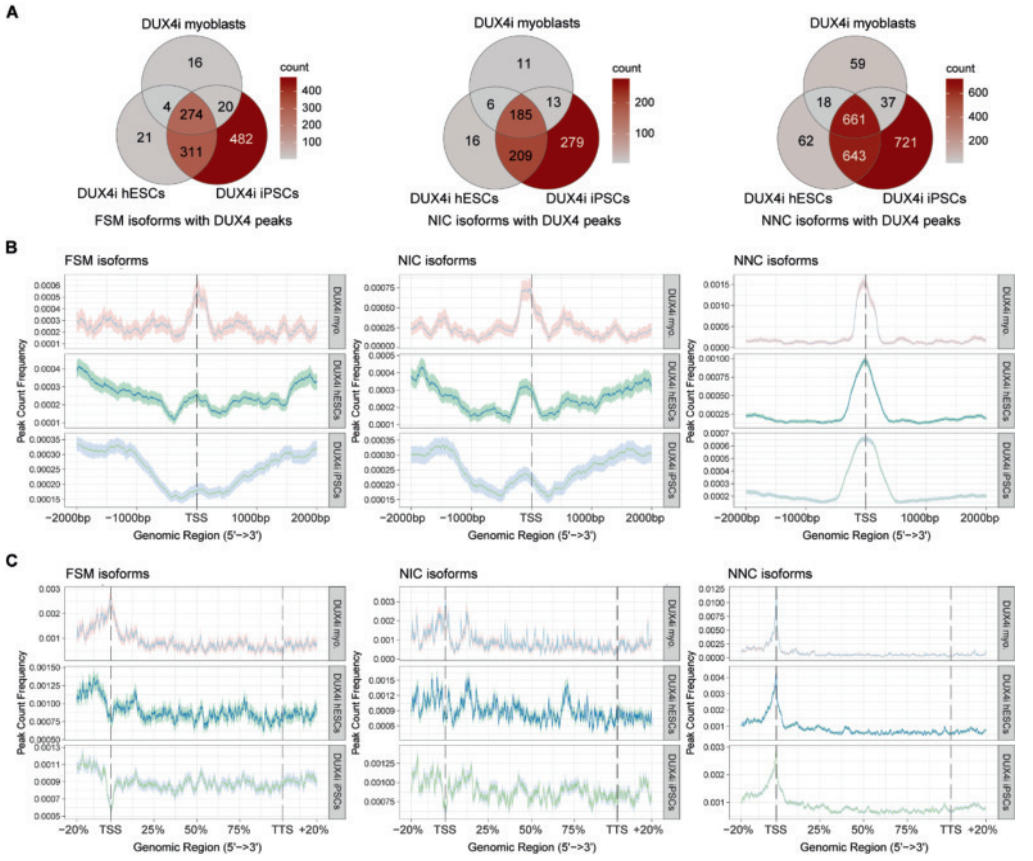

Fig. 3. Annotation of FSM, NIC, and NNC isoforms with DUX4 peaks.

(A) Venn diagrams showing the overlap of DUX4-bound isoforms across three cell lines. DUX4 peaks from three cell lines were annotated to isoforms classified into FSM, NIC, and NNC. Numbers indicate the count of isoforms with DUX4-binding sites in each cell line and their intersections. Pairwise hypergeometric tests revealed significant overlaps between any two cell lines (all *P* values <0.0001). (B) Peak density plots illustrating the genomic distribution of DUX4-binding sites around TSS (±2 kb) for FSM, NIC, and NNC isoforms. Each category contains three panels representing peak distributions from different cell lines. Shaded areas indicate confidence intervals estimated by bootstrap method. (C) Peak density plots showing DUX4-binding patterns across scaled transcript bodies (from TSS to TT S) plus upstream and downstream regions (20% of transcript length) for FSM, NIC, and NNC isoforms. Each panel represents peak distributions from different cell lines as indicated. Confidence intervals (shaded areas) were estimated using bootstrap method.

Isoform characterization of known loci demonstrates condition-specific isoform usage

Next, we investigated whether DUX4 activation influences transcript isoform usage of known genes by separating them into two main categories based on whether the gene was exclusively expressed in the DUX4+ condition. The analysis included 2,016 genes uniquely expressed in

DUX4+ and 7,344 genes commonly expressed in both conditions. These categories were further divided into three subtypes, detailing the utilization of known (FSM) or novel isoforms (NIC and NNC) (Fig. 4A): Category 1 (genes uniquely expressed in the DUX4+ condition): (a) only known isoforms expressed in DUX4+ myoblasts; (b) only novel isoforms expressed in DUX4+ myoblasts; (c) known and novel isoforms expressed in DUX4+ myoblasts. Category 2 (genes expressed in both DUX4− and DUX4+ conditions): (a) only known isoforms expressed in DUX4− myoblasts but known and novel isoforms expressed in DUX4+ myoblasts; (b) only known isoforms expressed in DUX4− myoblasts and only novel isoforms expressed in DUX4+ myoblasts; (c) known and novel isoforms expressed in DUX4− myoblasts but only novel isoforms expressed in DUX4+ myoblasts. Intriguingly, 29.5% (2,761 of 9,360) of genes showed altered isoform usage in the DUX4+ myoblasts, including 59.7% (40 of 67) of the core DUX4 target genes (tables S9 and S10). Particularly, we observed the highest proportions of genes falling in category 1-b (56.5%, 1,139 of 2,016), 1-c (12.9%, 261 of 2,016), and category 2-a (14%, 1,031 of 7,344), implying that these genes exhibit distinct splicing patterns when DUX4 is present. Correspondingly, the ASEs are enriched for the isoforms in these three subcategories (category 1-b: n = 3.595; category 1-c: n = 2,010; category 2-a: n = 10,426), predominated by SE and A3 (fig. S9). We then quantified the expression levels of the isoforms in each category in the LR and SR RNA-seq datasets. As expected, the isoforms in category 1 exhibited high expression levels in DUX4+ myoblasts in the SR RNA-seq data (Fig. 4B). The expression levels of isoforms in each category 2 subtype also showed consistency between the LR and SR RNA-seq (Fig. 4B).

Furthermore, we validated the translational potential of isoforms from each subcategory by analyzing the Ribo-seq and corresponding RNA-seq data from Campbell et al. (21). Category 1 isoforms showed higher expression levels, particularly at 8 and 14 hours posttreatment (fig. S10). Ribo-seq analysis revealed that coding isoforms exhibited substantial ribosome occupancy compared to noncoding isoforms, with this association further enhanced under DUX4+ conditions (fig. S10). By calculating the translation efficiency score through the normalization of ribosome footprint abundance to mRNA levels, we found that coding isoforms consistently displayed higher translation efficiency values than noncoding isoforms (fig. S10). These findings collectively demonstrate that these coding isoforms display substantial translational potential, as evidenced by their robust engagement with the translational machinery.

All genes linked to isoforms within categories 1 or 2 were used in enrichment analysis, yielding distinct outcomes for each main category (table S11). Category 1 genes, exclusively expressed in the DUX4+ condition, were enriched for GO terms associated with developmental processes (Fig. 4C), implying their potential relevance to the original role of DUX4 during early embryogenesis. Conversely, genes expressed in both conditions but exhibiting greater usage of novel isoforms in the DUX4+ condition were enriched for GO terms related to RNA metabolic processes, DNA damage, and apoptosis (Fig. 4C), suggesting potential dysfunctionality of these pathways due to ASEs.

Identification of cellular context–specific isoforms of DUX4 target genes

Earlier studies revealed that DUX4 is expressed sporadically in myonuclei *in vitro* and *in vivo* (6, 7, 14, 32-34). The expression of a core set of 67 DUX4 target genes is considered a reliable marker for DUX4 pathology (27). Compared to the DUX4− sample, 38 of the 67 core DUX4 target genes showed novel isoform expression in the DUX4+ condition (table S9). Given their natural expression during ZGA (4), we next asked whether DUX4 target gene isoform usage differs between cell types. We compared isoform usage in DUX4+ myoblasts to that observed during early embryogenesis, using a publicly available LR full-length transcriptome derived from human four-cell stage cells (35). In total, 49 and 50 of the 67 DUX4-target genes were detected in each transcriptome, respectively. Further, when considering a broader list of 146 DUX4-target genes (the DUX4 target genes 213 excluding DUX4 target genes 67) (27), 49 and 56 DUX4 target genes were detected, respectively (fig. S11). Some DUX4-target genes showed a similar isoform composition between both cell types, often expressing several isoforms that are not annotated in the reference transcriptome (GENCODE v39) (Fig. 4, D and E). However, the isoforms that contributed most to the total expression were noticeably different between both cell types (Fig. 4D). For instance, 62 isoforms of *TRIM43B* (12 ISM, 18 NIC, and 32 NNC) were detected in the DUX4+ transcriptome of myoblasts. In comparison, only 15 *TRIM43B* isoforms (3 FSM, 5 ISM, 4 NIC, and 3 NNC) were identified in the transcriptome of four-cell stage cells (Fig. 4D and fig. S12). In the DUX4+ myoblasts, the 32 NNC isoforms contributed most to the overall *TRIM43B* expression level. In contrast, the three FSM isoforms almost exclusively contributed to the expression from this locus in the four-cell stage cells. This analysis thus indicates that while DUX4 activates transcription of known loci, the splicing context, which is modulated by DUX4, results in the generation of distinct isoforms in a cellular context– specific manner.

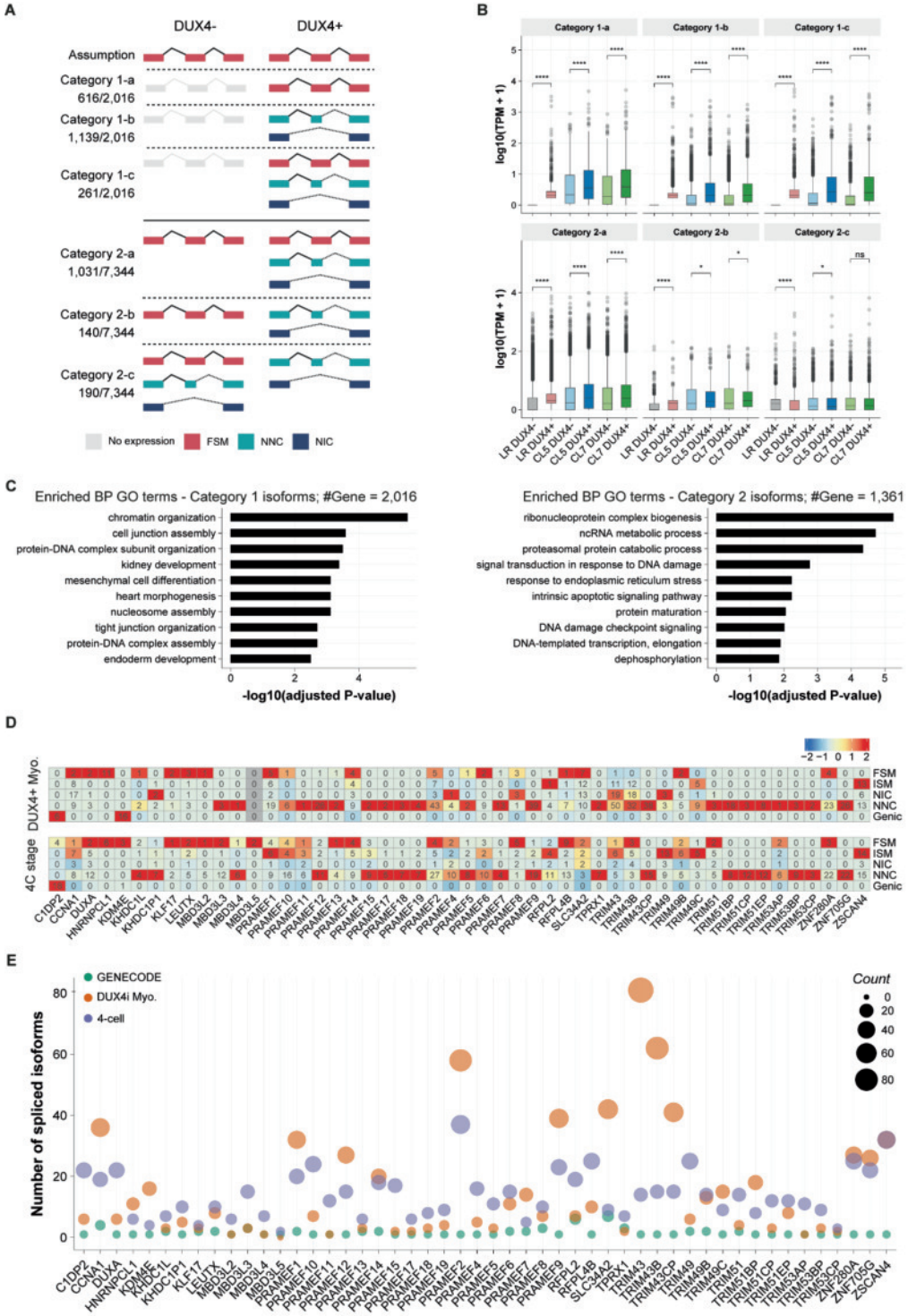
A
DUX4-
DUX4+
Assumption
Category 1-a
616/2,016
Category 1-b
1,139/2,016
Category 1-c
261/2,016
Category 2-a
1,031/7,344
Category 2-b
140/7,344
Category 2-c
190/7,344
No expression
FSM
NNC
NIC
B
Category 1-a
Category 1-b
Category 1-c
Category 2-a
Category 2-b
Category 2-c
log10(TPM + 1)
LR DUX4-
LR DUX4+
CL5 DUX4-
CL5 DUX4+
CL7 DUX4-
CL7 DUX4+
C
Enriched BP GO terms - Category 1 isoforms; #Gene = 2,016
chromatin organization
cell junction assembly
protein-DNA complex subunit organization
kidney development
mesenchymal cell differentiation
heart morphogenesis
nucleosome assembly
tight junction organization
protein-DNA complex assembly
endoderm development
-log10(adjusted P-value)
Enriched BP GO terms - Category 2 isoforms; #Gene = 1,361
ribonucleoprotein complex biogenesis
ncRNA metabolic process
proteasomal protein catabolic process
signal transduction in response to DNA damage
response to endoplasmic reticulum stress
intrinsic apoptotic signaling pathway
protein maturation
DNA damage checkpoint signaling
DNA-templated transcription, elongation
dephosphorylation
-log10(adjusted P-value)
D
DUX4+ Myo.
4C stage
FSM
ISM
NIC
NNC
Genic
E
GENECODE
DUX4i Myo.
4-cell
Count
Number of spliced isoforms

Fig. 4. Identification of isoform usage shift in DUX4-expressing cells.

(A) Schematic of all types of isoform compositions for each known gene under each condition. Color codes represent the structural categories: FSM, NIC, NNC, and isoforms without expression. The figure is created with Biorender.com. (B) Box plots showing the expression levels of isoforms associated with each isoform composition usage using LR RNA-seq data and SR RNA-seq data. *P* values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$, *$P < 0.05$, and $P > 0.05$ represented as "ns" (not significant). (C) Bar plots depicting the GO terms (BP) significantly (adjusted *P* value <0.05) enriched for genes with isoform in category 1 (left) and isoform in category 2 (right). (D) Heatmaps showing the expression levels of DUX4 core target genes from LR RNA-seq data obtained from DUX4i myoblasts (top) and four-cell stage cells (bottom). Expression at the gene level is calculated by summing up the normalized expression values of all isoforms for each gene. Color scales depict the expression level of each gene (Z score); red represents a high expression level; blue represents a low expression level. The number labeled in each cell represents the count of isoforms classified into each structural category for each DUX4 target gene. (E) Dot plot depicting the number of spliced isoforms in DUX4+ myoblasts and four-cell stage cells compared to the GENCODE transcriptome for the DUX4 target genes. Color codes represent each transcriptome.

We showed that the isoform usage of DUX4 target genes can be influenced by cellular context. We therefore further explored unique isoform usage in each condition and compared the isoforms of DUX4 target genes identified from the transcriptomes of DUX4 expressing DUX4i myoblasts and from four-cell stage cells. We found 71 and 62 isoforms (corresponding to 27 of 67 and 27 of 67 DUX4 target genes), respectively, containing at least one entirely novel exon in either the myogenic or the four-cell context (table S12). To verify their presence in cells with endogenous DUX4 expression, we analyzed publicly available RNA-seq data from (i) primary myotubes derived from patients with FSHD and healthy donors' biopsies (27) and (ii) human four-cell stage cells (36). The RNA-seq data of primary myotubes had considerably deep coverage (~143 million reads per sample). The analysis revealed that DUX4 target genes can produce isoforms that are specific to, or highly enriched in, a specific cellular context (Fig. 5A and fig. S13A). For example, isoform PB.76.171 of *PRAMEF9* with a novel exon was not detected in the four-cell LR-RNA generated transcriptome and reference transcriptome (GENCODE v39) (Fig. 5B). It was, however, highly expressed in FSHD primary myotubes but hardly detected in control primary myotubes and preimplantation embryos in the SR RNA-seq data (Fig. 5A). We also found that *SLC34A2* reported as a potential biomarker for FSHD (37, 38), uses an isoform (PB.12232.51) with a novel exon (Fig. 5B). The expression of this isoform also indicated its specificity in the myogenic context (Fig. 5A). Similarly, we observed the same phenomenon in the embryonic cleavage stage cells where *TRIM49C* uses an isoform (PB.9728.136) that includes a novel exon not detected in the transcriptome of DUX4i myoblasts and the reference transcriptome (fig. S13B).

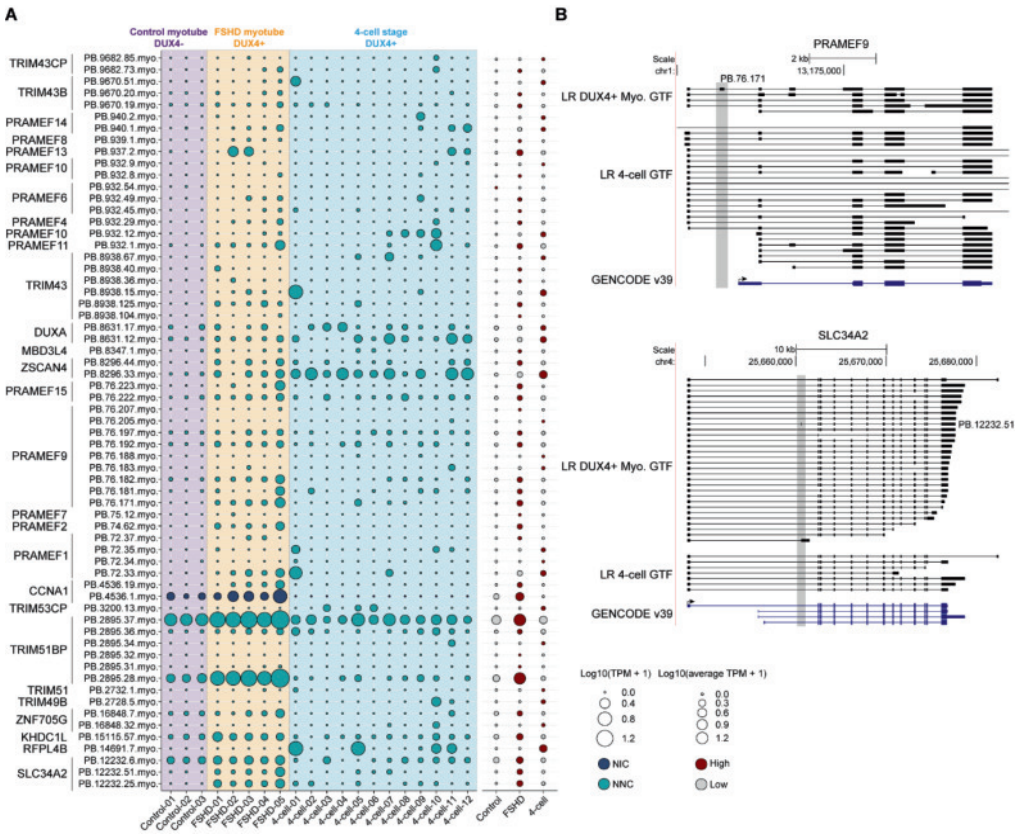

Fig. 5. Identification of cellular context-specific isoforms of DUX4 target genes.

(A) Dot plot showing expression levels of DUX4 target isoforms, identified in DUX4+ DUX4i myoblasts, across RNA-seq datasets from primary myotube cultures and four-cell stage cells. Color codes represent the structural categories and the size of dots represents the expression level of each isoform in each sample. Dark red represents the condition where the isoform shows the highest average expression. (B) Plots showing the isoforms of DUX4 target genes *PRAMEF9* and *SLC34A2*, with an entirely novel exon (PB.76.171 and PB.12232.51) compared to the transcriptome of four-cell stage cells. The shadow highlights the novel exons.

To assess the potential clinical relevance of our findings, we examined the expression levels of the myogenic context-enriched isoforms using an RNA-seq dataset from muscle biopsies of patients with FSHD (n = 37) and healthy donors (n = 26) (39). Several of these isoforms exhibited elevated expression levels in FSHD samples (fig. S14). Notably, isoform PB.2895.37, a novel isoform of *TRIM5BP*, showed significantly higher expression levels in FSHD samples compared to controls (fig. S14). Although this isoform displayed low expression levels in four-cell stage cells, these findings validate the translational value of our inducible cellular model and its potential application in understanding disease mechanisms.

In summary, while DUX4 activates germline genes involved in ZGA and, at the gene level, the enriched GO terms or pathways related to early embryonic development are frequently observed in FSHD (34, 40, 41), our analysis suggests that the transcriptomic changes associated with alternative splicing induced by DUX4 can be different between cellular contexts. These findings enhance our understanding of the embryonic signature in FSHD skeletal muscle and may contribute to the identification of novel isoform biomarkers specific to FSHD.

Identification and classification of unannotated intergenic isoforms

We observed a larger number of intergenic isoforms in DUX4+ myoblasts compared to the DUX4− condition (DUX4+: n = 1,275; DUX4−: n = 113) (Fig. 6A and table S13), which overall showed a higher expression level in the DUX4+ condition (fig. S15A). While most intergenic isoforms were confirmed by SR RNA-seq data, a higher proportion of DUX4+ uniquely expressed intergenic isoforms lacked support from SR RNA-seq data (101 of 1,260 versus 2 of 98), possibly because of mapping issues (fig. S15B). Only the intergenic isoforms detected in the DUX4+ condition (n = 1,275; corresponding to 652 loci) were retained for further analysis. To investigate whether DUX4 potentially drives the transcription of these intergenic isoforms, we mapped the DUX4-binding sites in the vicinity of the TSS using publicly available DUX4 ChIP-seq data of DUX4i myoblasts (18). Seventeen percent (217 of 1,275) of intergenic isoforms identified in DUX4+ myoblasts have one or more DUX4 ChIP-seq peaks within a 2-kb window centered around the TSS (Fig. 6, B and C). Further analysis of published DUX4 ChIP-seq datasets from DUX4i hESCs (1) and DUX4i iPSCs (2) revealed that 30.8% (393 of 1,275) and 42.2% (538 of 1,275) of intergenic isoforms have DUX4 peaks near their TSSs, respectively (Fig. 6, B and C). Across the three ChIP-seq datasets, 85.3% (185 of 217) of intergenic loci with DUX4-binding sites in a myogenic condition were consistently detected in the other two cell systems, and, overall, 45.3% (578 of 1,275) of all intergenic loci showed evidence for DUX4 binding in at least one of the studied cell systems (Fig. 6D). Moreover, we observed a higher degree of overlap between DUX4i hESCs and DUX4i iPSCs (Fig. 6D).

To further investigate whether there is evidence for cellular context-specific regulation of some of these intergenic loci by DUX4, we inferred the expression levels of intergenic isoforms from SR RNA-seq data of DUX4i myoblasts, hESCs, and iPSCs. Similar to the analysis performed for FSM, NIC, and NNC isoforms above, all DUX4-bound intergenic isoforms were classified into seven categories based on the evidence of DUX4 binding in each of the three cell lines, with some isoforms showing evidence of DUX4 binding in more than one dataset (fig. S15C). Category 8 includes all isoforms without DUX4 peaks. Overall, the intergenic isoforms identified in our myogenic cell model were also activated in hESCs and iPSCs after DUX4 induction, and the intergenic isoforms without evidence for DUX4 binding (category 8) showed relatively low expression levels compared to those with evidence for DUX4 binding, consistent with the results in the myogenic context. In our SR RNA-seq data of DUX4i myoblasts, the isoforms with DUX4 peaks according to the DUX4 ChIP-seq data from DUX4i myoblasts (categories 1, 4, 6, and 7) exhibited relatively higher expression levels than those with DUX4 peaks in the other two cell lines (fig. S15D). Consistently, the isoforms in categories 3, 5, and 7 enriched for DUX4 binding in iPSCs were highly expressed in DUX4+ iPSCs, and the isoforms in categories 2, 5, and 7 enriched for DUX4 binding in hESCs were highly expressed

in DUX4+ hESCs (fig. S15D). Although these DUX4i cell lines use different strategies to activate DUX4 and have variable induction times, our analysis uncovered evidence of cis regulation between DUX4 and its target intergenic loci, with a suggestion for context-specific DUX4 binding for a subset of loci.

It is known that DUX4 plays a role in chromatin regulation (22). Analysis of publicly available Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) data of DUX4i hESCs (22) revealed that 24.5% (312 of 1,275) of intergenic isoforms show increased chromatin accessibility in their TSS regions (Fig. 6C and fig. S16, A and B). Integrating these results with the DUX4 ChIP-seq data of DUX4i hESCs identified 55.5% (218 of 393) of intergenic isoforms with shared ChIP-seq and ATAC-seq peaks (fig. S16A). Motif enrichment analysis revealed a significant enrichment of DUX4-binding motifs within these ATAC-seq peaks (fig. S16C), suggesting that DUX4 may activate transcription by directly binding to these intergenic loci and by modulating their chromatin state (fig. S16D). Our analysis thus indicates a direct regulatory relationship between DUX4 and the transcriptional activation of these intergenic loci.

Last, we used DUX4 ChIP-seq data from DUX4i myoblasts to annotate and categorize all intergenic isoforms. As described above, in a myogenic context, we identified 17.1% (217 of 1,275) of intergenic isoforms, which condense into 205 loci upon DUX4 induction. For example, in the DUX4+ transcriptome, nine intergenic isoforms were identified from the locus PB.728 (chr1: 217,205,526217,231,944) bound by DUX4, with expression exclusively in the LR and SR RNA-seq data of DUX4+ samples (Fig. 6E). These intergenic isoforms were further classified into four categories based on their coding potential and direct regulation by DUX4 (fig. S17 and table S14; see Materials and Methods). Notably, intergenic isoforms regulated by DUX4 exhibited higher expression levels than those without DUX4 peaks (Fig. 6F), further affirming the direct regulation by DUX4.

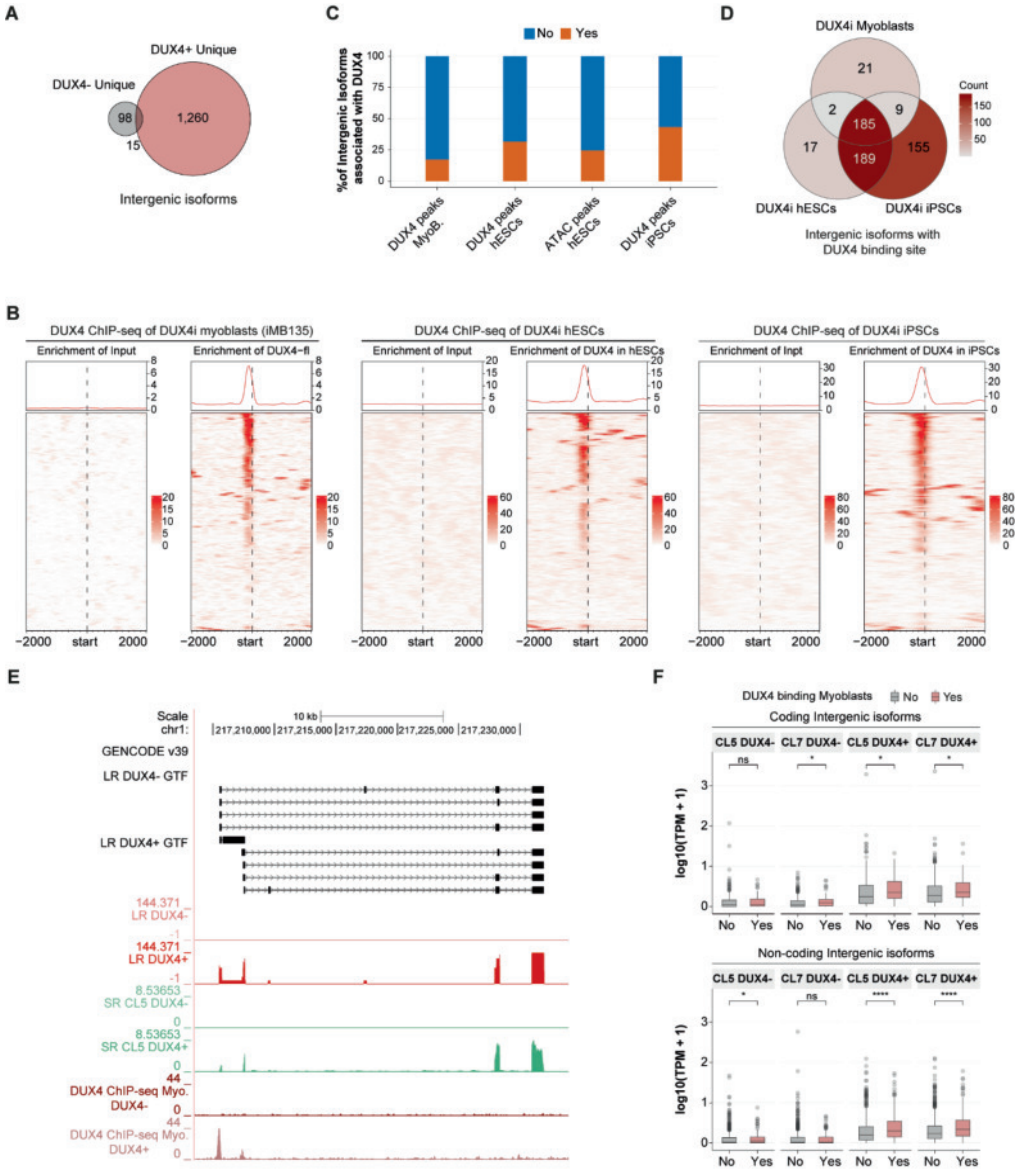

Fig. 6. Identification and classification of intergenic isoforms.

(A) Venn diagram showing the overlap between intergenic isoforms from DUX4− and DUX4+ transcriptomes. (B) Heatmaps depicting the signal intensities of DUX4 ChIP-seq peaks near TSS regions (±2 kb) of intergenic isoforms expressed in DUX4+ myoblasts using three publicly available DUX4 ChIP-seq datasets obtained from DUX4i myoblasts, DUX4i hESCs, and DUX4i iPSCs. Each row represents an individual isoform. The isoforms in each heatmap are independently clustered based on their binding patterns in the respective cell line. (C) Stacked bar plot displaying the percentage of intergenic isoforms with or without at least one DUX4-binding site or ATAC-seq peak. (D) Venn

diagram showing the overlapping intergenic isoforms with DUX4 ChIP-seq peak. Color scale represents the count of overlapping intergenic isoforms. Pairwise hypergeometric tests revealed significant overlaps between any two cell lines (all *P* values <0.0001). (E) UCSC genome browser visualization of an example of the intergenic locus with a DUX4-binding site. Color codes indicate the data in each track: black for transcriptomes of DUX4−, DUX4+, and GENCODE reference; bright red for LR RNA-seq data; blue for SR RNA-seq data; dark red for DUX4 ChIP-seq data from DUX4i myoblasts. (F) Box plots displaying the expression levels of intergenic isoforms with or without DUX4 ChIP-seq peak in each sample. *P* values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$, *$P < 0.05$, and $P > 0.05$ represented as ns (not significant).

Validation of intergenic isoforms in FSHD-related datasets

To confirm the existence of intergenic isoforms in independent samples, we reanalyzed two independent RNA-seq datasets from immortalized myoblast cell lines using distinct DUX4-induction strategies. One dataset [Young et al. (18)] included two immortalized myoblast cell lines (iMB135 and iMB54-1) expressing DUX4 through lentiviral delivery (18), while the other [Campbell et al. (21)] used iMB135 myoblasts with DUX4i expression through DOX treatments for 4, 8, and 14 hours (21). In both datasets, most intergenic isoforms (1,210 of 1,275 in Young et al. and 1,131 of 1,275 in Campbell et al.) were detectable (expression in TPM > 0) (Fig. 7A). Comparing the expression levels of intergenic isoforms with or without DUX4-binding sites in each sample, we observed significantly increased expression levels of intergenic isoforms associated with a DUX4binding site in DUX4+ samples from Young et al. (18) (Fig. 7B), consistent with the results from our SR RNA-seq data. Analyzing the RNA-seq dataset from Campbell et al. (21) revealed a significant increase in the expression of intergenic isoforms with prolonged treatment, suggesting a positive correlation between their expression levels and the presence of DUX4 (Fig. 7C).

We further validated these findings by analyzing two publicly available RNA-seq datasets from primary myotube cultures of eight patients with FSHD and six unaffected donors (27, 42). The average expression levels of the intergenic isoforms revealed a significant up-regulation in FSHD conditions compared to control conditions (Fig. 7D). However, substantial heterogeneity was observed among the samples of the Yao et al. (27) dataset, with FSHD-05 exhibiting higher expression of intergenic isoforms than other FSHD samples (fig. S18A, Heatmap). Considering the correlation between intergenic isoform expression and the DUX4 signature, we indeed found higher expression levels of DUX4 and its target genes in FSHD-05 (fig. S18A). Two FSHD samples exhibited elevated expression levels of intergenic isoforms in the RNA-seq data from Banerji et al. (42) (fig. S18B). Notably, only one FSHD sample (FSHD-01) showed detectable levels of DUX4 (fig. S18B). However, FSHD-02 and 03 showed higher expression levels of DUX4 target genes, correlating with elevated expression of the intergenic isoforms (fig. S18B). Although the intergenic isoforms showed relatively low expression levels with large variability likely due to the cellular heterogeneity in FSHD and the sporadic nature of DUX4, they were validated in independent DUX4i myogenic cell lines and confirmed in FSHD primary myotubes, confirming their existence in FSHD.

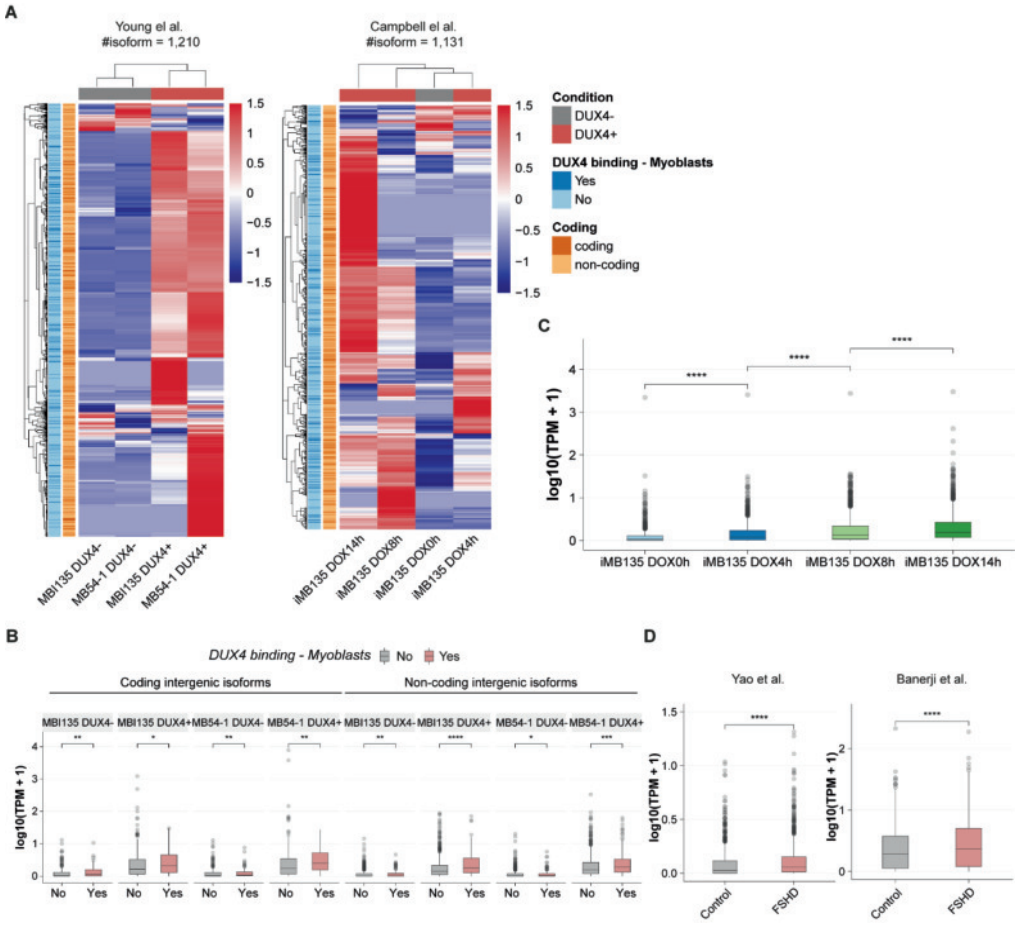

Fig. 7. Validation of intergenic isoforms in biologically relevant datasets.

(A) Heatmaps showing the expression levels of intergenic isoforms in two published RNA-seq datasets obtained from DUX4i immortalized myoblast cell lines through different DUX4 expression induction methods. Color scale depicts the expression level normalized in Z score. (h, hours). (B) Box plots displaying the expression levels of intergenic isoforms with or without the DUX4 ChIP-seq peak in each sample of RNA-seq data obtained from DUX4i iMB135 myoblasts. *P* values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, and $P > 0.05$ represented as ns (not significant). (C) Box plot showing the average expression levels of intergenic isoforms in each condition (0-, 4-, 8-, and 14-hour DOX treatment). *P* values are calculated the same as in (B). (D) Box plots showing the average expression levels of intergenic isoforms in FSHD and control samples [left, Yao et al. (27); right, Banerji et al. (42)]. *P* value is calculated the same as in (B).

Most intergenic isoforms originate from REs

Given that more than two-thirds of the human genome consists of REs (43) and that DUX4-binding sites are considerably enriched in REs genome-wide (18), we quantified the number of

REs within the intergenic isoforms. Intriguingly, 99.2% (1,265 of 1,275) of intergenic isoforms identified from the DUX4+ transcriptome contained at least one RE. Specifically, 55.5% (707 of 1,275), 92% (1,173 of 1,275), and 49.9% (636 of 1,275) of intergenic isoforms contained at least one LINE, LTR, or short interspersed nuclear element (SINE), respectively (table S15). LTR-containing intergenic isoforms showed substantial overlap with those containing LINEs or SINEs, which may suggest that LTRs were used as promoters by DUX4 to activate neighboring LINEs or SINEs to create spliced full-length mRNAs (Fig. 8, A and B). Further analysis showed that within the LINE family, L1 and L2 are the predominant subfamilies. Most retroelements in the LTR family belonged to ERVL-MaLR, ERVL, and ERV1, consistent with previous studies (18). In addition, in the SINE family, Alu and MIR are the subfamilies with the highest representation (Fig. 8C).

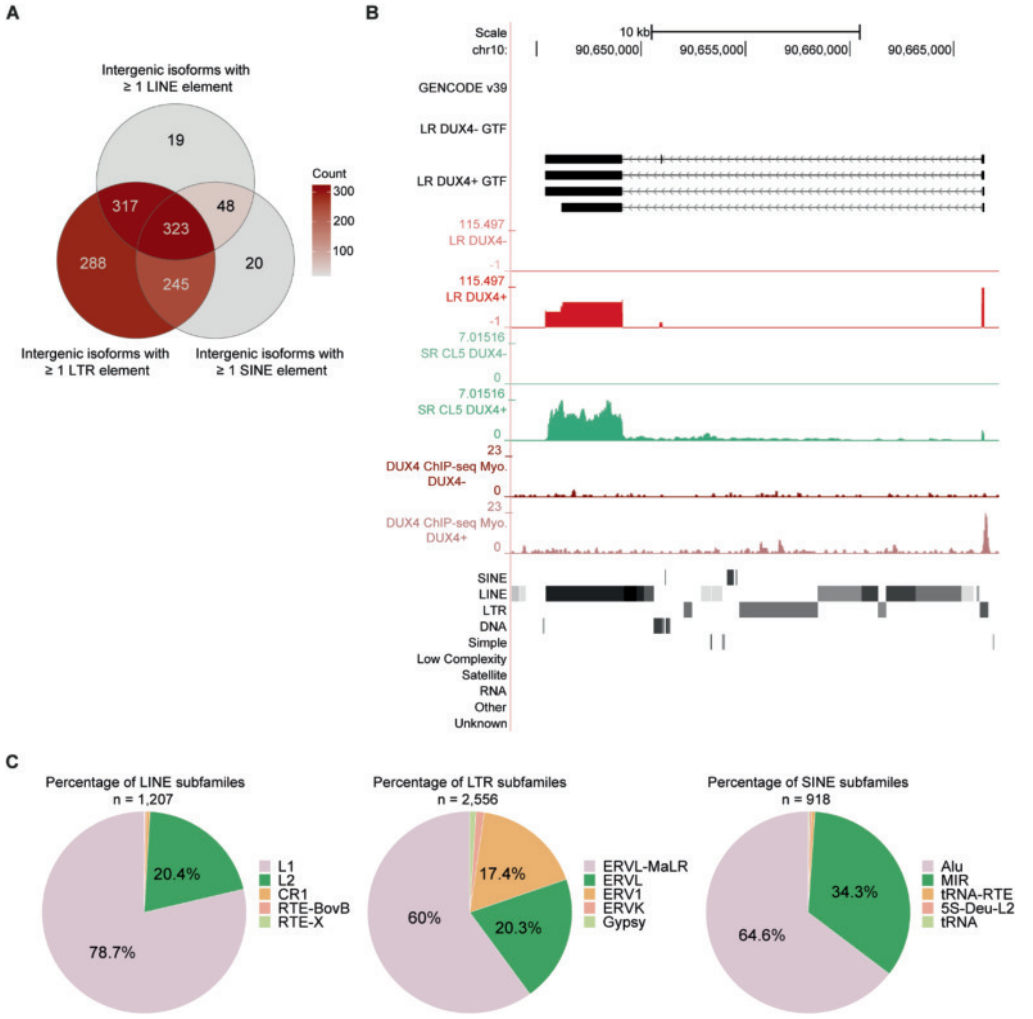

Fig. 8. Intergenic isoforms often originate from REs.

(A) Venn diagram showing the overlapping intergenic isoforms with REs. Color scale represents the count of overlapping intergenic isoforms. (B) UCSC genome browser visualization of an example intergenic locus with RepeatMasker track. Color codes indicate the data in each track: black for transcriptomes of DUX4−, DUX4+, and GENCODE reference; bright red for LR RNA-seq data; blue for SR RNA-seq data. (C) Pie charts showing the percentage of subfamilies for each RE. Color codes represent different subfamilies within each family.

To evaluate whether DUX4 peaks were nonrandomly enriched in REs, we performed permutation testing (n = 1,000). Our results showed that 76.6% (25,331 of 33,076) of DUX4 peaks overlapped with REs, accounting for 64.8% of the total peak regions. This overlap was significantly higher than expected by chance (P value <0.001 for both peak-count and base-pair overlap). Compared to randomly shuffled genomic regions of similar size and chromosomal

distribution, peaks showed a 1.2-fold enrichment (25,331 of 21,487) in overlap number and a 1.3-fold enrichment (3,922,147/3,036,240 bp) in base-pair coverage, indicating a modest but statistically significant preference of DUX4-binding sites for REs (fig. S19A).

Given that SINE elements, particularly Alu sequences, are preferentially enriched in gene-rich regions of the human genome (44), we further investigated the relationship between DUX4-binding sites and REs within genic regions. We annotated NIC and NNC isoforms identified under the DUX4+ condition with DUX4 peaks defined in DUX4i myoblasts. Among 320,177 analyzed exons, 6,858 (2.1%) showed substantial overlap with REs (overlap fraction >0.5). Of these RE-containing exons, 1,719 (25%) had at least one DUX4 peak within a 2-kb window upstream or downstream (table S16). Notably, 90% (1,547 of 1,719) of these repeats belonged to the LTR family (fig. S19B). These 1,719 exons corresponded to 1,475 NIC and NNC isoforms from 448 genes. For example, DUX4 binding to an LTR element upstream of *CCDC30* led to the generation of novel isoforms incorporating this LTR as the first exon. (fig. S19C). These findings demonstrate that DUX4 regulates both genic and intergenic targets through RE activation, particularly LTR elements, thereby contributing to isoform diversity.

Characterization of intergenic isoforms in preimplantation embryos

To characterize the intergenic isoforms under natural DUX4-positive conditions, we analyzed datasets from human preimplantation embryos and established the transcriptional and epigenetic landscape for these intergenic isoforms during early embryogenesis. Analyzing SR RNA-seq data from embryonic cells at different developmental stages (35) revealed that 71.5% (911 of 1,275) of intergenic isoforms were detectable at the four-cell stage (TPM > 0), and some showed stage-specific expression patterns in subsequent developmental stages (Fig. 9A and table S15). Moreover, predicted coding and noncoding intergenic isoforms with DUX4-binding sites showed substantially higher expression levels across all stages compared to those without DUX4-binding sites (fig. S20A). Analysis of public ATAC-seq datasets spanning early development from the four-cell to blastocyst stages (45) demonstrated that 10.9% (139 of 1,275) of intergenic isoforms have ATAC-seq peaks at the four-cell stage, with a decreasing number of ATAC-seq peaks in subsequent stages, suggesting the gain and loss of ATAC-seq peaks in different developmental stages (Fig. 9B; fig. S20, B and C; and table S15). The best matches to the *de novo* motif search based on the ATAC peaks near the TSS of intergenic loci at the four-cell stage were related to the double homeobox family of transcription factors further supporting the regulatory relationship between these intergenic loci and DUX4 (Fig. 9B). The matches of *de novo* motifs underlying the ATAC peaks at other developmental stages were variable, suggesting that additional transcription factors may regulate intergenic loci during early embryogenesis (Fig. 9B and fig. S20B).

Analysis of published reduced representation bisulfite sequencing (RRBS) DNA methylation data of early embryonic cells (46) showed a generally low methylation level within a 1-kb window surrounding the TSS of intergenic isoforms (Fig. 9C). In contrast, methylation levels notably increased in postimplantation embryos in these regions. Analysis of RRBS data of primary myoblasts derived from patients with FSHD2 and healthy donors (47) did not reveal significant differences in methylation levels near the TSS of intergenic isoforms (Fig. 9C) suggesting that these loci are developmentally regulated.

We integrated publicly available ChIP-seq and CUT&RUN data of relevant histone modifications to construct the epigenetic landscape for intergenic loci during early embryogenesis (48-50), especially focusing on the H3K4me3 and H3K27ac primarily associated with transcriptional activation (table S15). In the four-cell stage, 282 intergenic isoforms (250 intergenic loci) were enriched for H3K4me3 near their TSS (±2 kb). In contrast, at the eight-cell stage, 192 intergenic isoforms (180 loci) were enriched for H3K27ac. Notably, isoforms marked by H3K4me3 in the four-cell stage and by H3K27ac in the eight-cell stage displayed substantial overlap (Fig. 9D and fig. S21A). Although 264 intergenic isoforms (253 loci) had H3K9me3 marks in the four-cell stage, the intersection with isoforms marked by H3K4me3 was limited (Fig. 9D). Isoforms with H3K4me3 and/or H3K27ac marks displayed elevated expression levels compared to those with H3K9me3 marks during the four-cell/eight-cell stage (Fig. 9E), indicating an active transcriptional state.

Further investigation into the dynamic changes of these histone marks during early embryogenesis revealed that the intergenic isoforms marked by H3K4me3 during the four-cell stage lost this modification in the eight-cell stage. Conversely, the number of isoforms with H3K9me3 marks slightly increased with development (Fig. 9F). This switch in histone mark occupation correlated with a decrease in intergenic isoform expression during embryonic development (fig. S21B), suggesting an association between the intergenic loci and epigenetic marks of active transcription.

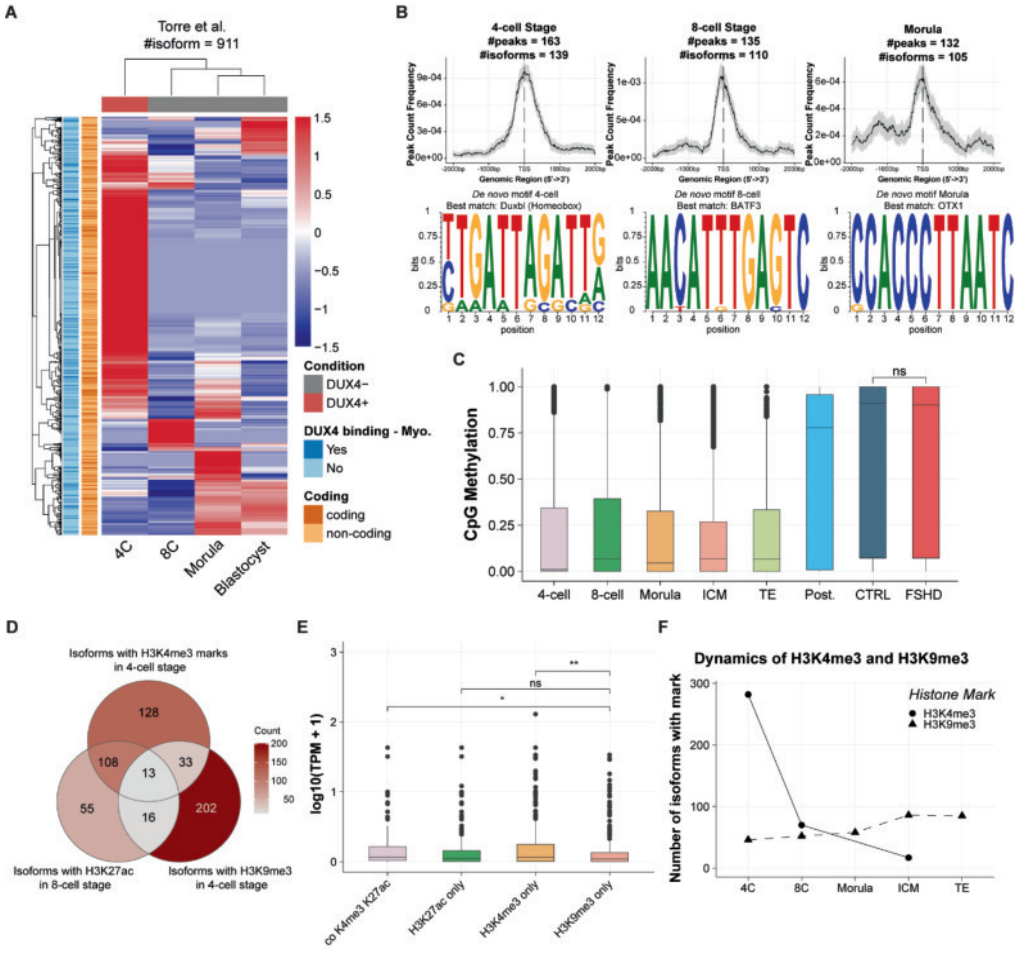

Fig. 9. Characteristics of intergenic loci in human preimplantation embryos.

(A) Heatmap showing the expression levels of intergenic isoforms in published RNA-seq data obtained from human preimplantation embryos. Color scale depicts the expression level normalized in Z score. (B) Density plots showing the signal intensity of ATAC-seq peaks near the TSS regions of intergenic isoforms using the published ATAC-seq data obtained from human four-cell stage cells, eight-cell stage cells, and morula. Sequence motif plots depicting the significantly (*P* value <0.05) enriched motifs for ATAC-seq peaks from each developmental stage. (C) Box plots depicting the methylation level of the 1-kb genomic region surrounding the TSS of intergenic isoforms, using published RRBS data obtained from cleavage stage cells, postimplantation embryos, and primary myoblasts derived from patients with FSHD and healthy donors. *P* value is calculated using unpaired Wilcoxon test to assess the difference in methylation level between FSHD and control myoblasts. Statistical significance is denoted as follows: $^{**}P < 0.01$, $^{*}P < 0.05$, and $P > 0.05$ represented as ns (not significant). (D) Venn diagram visualizing the overlapping isoforms with H3K4me3 marks in 4C, H3K27ac marks in 8C (eight-cell), and H3K9me4 marks in 4C (four-cell). Color scale shows the number of isoforms. (E) Boxplot showing the expression levels of intergenic isoforms with different histone modification marks. The P values are

calculated in the same way as in (C). (F) Line plot showing the dynamics of isoforms with different histone marks during early embryogenesis. The shape of the dot represents the type of histone marks.

Discussion

We present a comprehensive full-length isoform-resolved reference transcriptome for muscle cells expressing DUX4 and compare it to transcriptomes from different cell types under normal and pathological conditions in which DUX4 is expressed. This analysis revealed that while in all cell systems, DUX4 at first sight activates a largely comparable transcriptional program that also includes the transcriptional activation of yet unannotated intergenic loci enriched for REs, in detail, this program diversifies by cellular context-specific isoform usage.

By combining LR and SR RNA-seq, we uncovered 21,150 NIC and 22,499 NNC isoforms, representing transcripts previously not annotated in the reference transcriptome. This substantially broadens our knowledge of the skeletal muscle transcriptome in the presence of DUX4. Leveraging LR RNA-seq, we identified loci displaying preferential or unique isoform usage across different conditions. Genes involved in the RNA splicing process in DUX4+ myoblasts tend to use NIC or NNC isoforms, providing perspective for understanding disrupted RNA splicing as a prominent pathogenic hallmark of FSHD. Moreover, we observed differences in isoform usage of DUX4 target genes between DUX4-expressing myoblasts and DUX4-expressing embryonic cleavage stage cells. Considering the distinct cellular and epigenetic environments of somatic muscle cells and early embryonic cells is crucial, as complex tissue-specific regulatory mechanisms may contribute to differences in isoform usage. Current widely used transcriptome annotation files, such as GENCODE, often lack the level of isoform complexity highlighting the added value of LR sequencing to generate more accurate isoform-resolved transcriptomes. Our analysis of novel isoforms across different conditions suggests that skeletal muscle-specific isoforms have the potential to serve as biomarkers for FSHD. However, further validation through larger *in vivo* datasets, longitudinal studies, and correlation analyses with disease severity and progression is required.

Compared to reference transcripts, more novel isoforms tend to use new TSSs and TTSs. This may lead to the transcript being degraded as an NMD substrate. In particular, in muscle cells, DUX4 induces proteolytic degradation of UPF1, a key component of the NMD pathway, further leading to NMD inhibition and widespread accumulation of NMD-targeted mRNAs (28). This mechanism allows for the introduction of changes in the ORFs, potentially producing dysfunctional proteins. It has been reported that an NMD isoform of *SRSF3* significantly increases in expression following DUX4 induction in muscle cells and its truncated protein can trigger cytotoxicity, leading to apoptosis (21). Our findings regarding novel isoforms in DUX4+ myoblasts suggest that such events could be widespread, potentially offering insights into the DUX4-mediated cytotoxicity in skeletal muscle.

It has been reported that BPs involved in embryonic development were observed in FSHD (34, 41). However, most of the observations were based on the enrichment analysis at the gene level. Here, we first compared the transcriptomes derived from LR RNA-seq of DUX4i myoblasts and cleavage embryos, especially for the DUX4 target genes, which mostly overlap with ZGA genes. While, in general, there is consistency in the DUX4 transcriptional network of the

different cell types, an appreciable number of the DUX4 target gene isoforms show cellular context–dependent splicing junction usage. This suggests that, at the isoform level, the embryonic signature in FSHD myogenic cells differs from that of cleavage-stage embryos. We indeed identified isoforms of DUX4 target genes that are myogenic specific and verified them in primary FSHD myotube cultures, highlighting the advantages of full-length Iso-seq in capturing rare isoforms. Integrating LR and SR RNA-seq with considerable sequencing depth advances the identification of isoforms specific to tissues or diseases.

DUX4 binds to REs, such as LTRs, in DUX4-expressing muscle cells, forming new promoters to activate adjacent genes (18). A study using nanopore direct RNA-seq in DUX4-expressing human rhabdomyosarcoma cells further confirmed the production of LTR-fusion transcripts (19). In intergenic genomic regions, we provide evidence that DUX4 activates LTRs to form promoters, which in turn activates nearby RE families to form spliced isoforms. Our bioinformatic analysis identified 716 transcribed intergenic loci (DUX4−: 85; DUX4+: 652) directly or indirectly regulated by DUX4 that had never been annotated before. The DUX4-binding patterns of these intergenic loci partly overlap between the cell types studied but also show distinct differences across cell types, indicating cellular context–specific functional outcomes that are likely influenced by cofactor availability, the epigenetic landscape, and cellular context-specific promoters and enhancers. RE activation is pivotal in early embryogenesis, potentially initiating and driving ZGA (51-53), but a direct link with DUX4 remains to be explored.

We extensively used publicly available datasets to construct the transcriptional and epigenetic landscape of these unannotated intergenic loci during early embryogenesis. The stage-specific gene expression levels observed at these sites suggest that some intergenic loci may be regulated by DUX4 directly or by other factors induced by DUX4. This supports our previous findings, where analysis of snRNA-seq data from FSHD myotubes identified signals resembling those of the eight-cell stage and blastocysts (34). These intergenic loci are highly enriched for REs. During ZGA, the genome undergoes epigenetic reprogramming rendering REs active (54). The methylation differences between the cleavage stage and primary FSHD myoblasts may not play a crucial role in the transcription of these intergenic loci. The changes in histone marks such as H3K4me3 and H3K9me3 further support the active transcription of some intergenic loci in the four-cell stage. Histone modifications are dynamically regulated during myogenesis and further research on histone modifications in DUX4expressing muscle cells will aid in understanding DUX4-mediated cytotoxicity, its impact on normal myogenesis, and the active expression of intergenic loci in muscle cells.

In summary, the isoform-resolved transcriptome specific to DUX4+ myoblasts considerably improves the annotation of isoforms from both known and unknown loci, contributing to a deeper understanding of FSHD biology and unveiling the downstream events associated with the presence of DUX4 in different cellular contexts. In particular, combining LR and SR RNA-seq datasets shows great advantages in deciphering the transcriptomic changes induced by DUX4, offering new insights into DUX4mediated cytotoxicity.

Materials and Methods

Generation and validation of immortalized DOX-inducible DUX4 myoblast cell line

The DUX4i-immortalized myoblast cell lines used in this study were generated similarly to the previously generated DUX4i cell line described by Jagannathan et al. (17). In this current study, cells from a primary control myoblast cell line (32U 4qA/4qB, 161S haplotype) were first immortalized using retroviral transduction-based CDK4 and hTERT immortalization (plasmids provided by the Tapscott Lab; the CDK4 plasmid is based on a pLXSH vector backbone with standard CDK4 cDNA inserted by BamHI-NotI cloning, and the hTERT plasmid is based on a pLXSN vector backbone with hTERT cDNA inserted). Transduced cells were selected by hygromycin (20 μg/ml) and neomycin (40 μg/ml) selection. The polyclonal immortalized control myoblast cell line was then transduced with a tetracycline-inducible codon-altered DUX4 construct using lentiviral transduction [plasmid described by Jagannathan et al. (17) and now available via Addgene as pCW57.1-DUX4-CA, Addgene #99281]. Individual positive clones were selected by puromycin selection (1 μg/ml) and validated for DOX-induced DUX4 activation (8 hours, 1 μg/ml DOX) based on immunofluorescence microscopy imaging (fig. S1A) and DUX4-induced cell death within 24 hours (4 μg/ml DOX) (fig. S1B).

For immunofluorescence imaging, the cells were fixed by 2% formaldehyde cross-linking for 15 min, permeabilized with 1% Triton X-100 for 10 min, and stained for DUX4 (1:2000, ab124699, Abcam) and Hoechst33528 (1:1000, H3569, Thermo Fisher Scientific, Nuclei staining) using immunofluorescent staining. The immunofluorescence signal was imaged using a Nikon Eclipse Ti microscope.

Cell culture, DOX-induced DUX4 activation and RNA harvest

Myoblasts were cultured in Ham's F-10 Nutrient Mix with GlutaMAX Supplement (# 41550-021, Life Technologies, Waltham, Massachusetts, USA), supplemented with 20% heat-inactivated fetal bovine serum (#10270, Gibco/Life Technologies, Waltham, Massachusetts, USA), 1% penicillin-streptomycin (#15140122, Gibco/Life Technologies, Waltham, Massachusetts, USA), rhFGF (10 ng/ml; #C-60240, BioConnect, Huissen, Gelderland, the Netherlands), and 1 μM dexamethasone (#D2915, Sigma-Aldrich, St. Louis, Missouri, USA). On day 0 of DOX treatment, three million cells were plated in a 10-cm petri dish. On day 1, the cells were treated with DOX (2 μg/ml) for 16 hours. RNA was harvested by removing the cell culture medium and lysing the cells in 2 ml of Qiazol lysis reagent (Qiagen, catalog no. 79306). The lysed samples were split into two 1-ml aliquots and stored at −80°C.

SMRT-seq library construction and sequencing

One 1-ml aliquot of each Qiazol lysate (see above) was used for RNA isolation using the miRNeasy mini RNA isolation kit (Qiagen, catalog no. 217004), following the manufacturer's protocol, including a 30-min on-column deoxyribonuclease I (DNAse I) treatment (Qiagen, catalog no. 79256). RNA quality was assessed using an Agilent BioAnalyzer RNA Nano 6000 chip (catalog no. 5067-1511), with all samples having RNA integrity numbers (RINs) ranging from 9.2 to 9.5. The total RNA was used to generate the full-length cDNA using the NEBNext

Single Cell/Low Input cDNA Synthesis & Amplification Module and Iso-Seq Express Oligo Kit, following the manufacturer's instructions (protocol version, PN 101-763-800 version 02). The cDNA was used as input for SMRTbell library construction using the SMRTbell Express Template Preparation Kit v2.0 following the manufacturer (Pacific Biosciences). Sequencing was performed on Sequel-II using sequencing primer V4 and Binding kit 2.1 with a 24-hour movie time.

SR RNA-seq library construction and sequencing

One 1-ml aliquot of each Qiazol lysate (see above) was used for RNA isolation using the miRNeasy mini RNA isolation kit (Qiagen, catalog no. 217004), following the manufacturer's protocol, including a 30-min on-column DNAse I treatment (Qiagen, catalog no. 79256). RNA quality was assessed using an Agilent BioAnalyzer RNA Nano 6000 chip (catalog no. 5067-1511), with all samples having RIN/ RNA quality numbers ranging from 9.2 to 10. The RNA-seq libraries of all samples were generated with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB #E7760S/L) according to the manufacturer's protocol. Briefly, mRNA was isolated from total RNA using the oligo-dT magnetic beads. After the fragmentation of the mRNA, cDNA synthesis was performed followed by ligation with the sequencing adapters and PCR amplification of the resulting product. The quality and yield after sample preparation were measured using the Fragment Analyzer. Clustering and cDNA sequencing using the NovaSeq6000 were performed according to the manufacturer's protocols. Image analysis, base calling, and quality check were performed with the Illumina data analysis pipeline RTA3.4.4 and Bcl2fastq v2.20. Quality-filtered sequence tags are exported as the final fastq files for the downstream analysis.

LR RNA-seq data processing

The PacBio Iso-Seq3 pipeline (version 3.4.0) was used for processing the raw sequencing data obtained from each sample, leading to the generation of full-length non-concatemer (FLNC) reads. Initially, subreads underwent intramolecular error correction and polishing using the Circular Consensus Sequencing (CCS) algorithm (version 6.0.0). This step resulted in highly accurate CCS reads, with a minimum predicted accuracy exceeding 0.99, each requiring a minimum of three complete polymerase passes. Subsequently, the polished CCS reads were processed using the lima tool (version 2.2.0), which facilitated the removal of SMART-Seq primers and template-switching oligo sequences. In addition, the lima tool ensured the proper orientation of isoforms in the 5′ to 3′ direction. The refine and cluster modules from Iso-Seq3 were then applied to eliminate polyA tails and concatemers of each read and hierarchically cluster sequences, yielding FLNC reads ready for downstream analysis. These FLNC reads were subsequently mapped to the GRCh38 genome assembly to facilitate cross-validation with previous DUX4related RNA-seq datasets (GRCh38.p13-v39-GENCODE) using the splice-aware aligner minimap2 (version 2.17-r941) (55) with specific parameters (minimap2 -ax splice -uf --secondary = no -C5 -06,24 B4). Following mapping, the Cupcake tool (version 28.0.0) was used to collapse isoforms supported by at least two FLNC reads into nonredundant forms and filter out 5′ degraded isoforms. This process resulted in the creation of two distinct transcriptomes corresponding to the DUX4− and DUX4+ conditions. After calling abundance

from the processed isoforms using Cupcake, the DUX4− and DUX4+ transcriptomes were merged into a unified, comprehensive transcriptome specific to DUX4i myoblasts.

SR RNA-seq data processing

The raw fastq files of paired-end RNA-seq data for each sample underwent adapter trimming using Trim Galore (v0.6.7). Subsequently, a quality control (QC) check was performed using FastQC (v0.11.9) to assess the overall data quality. The trimmed RNA-seq data were then analyzed with SQANTI3 for isoform annotation and QC (see below).

Isoform annotation and QC

QC assessment and filtering were conducted using the SQANTI3 toolkit (v5.1.1) (56) with default settings for the full-length transcriptome and a publicly available full-length transcriptome of human four-cell stage cells (35). The references for QC involved the human genome and comprehensive gene annotation data downloaded from GENCODE v39. To complete this process, we used reference files for CAGE Peak data and polyA motif lists provided by SQANTI3. In addition, we supplied SQANTI3 with SR RNA-seq data to apply stringent filtering criteria, ensuring the credibility of full-length isoforms. Each isoform underwent evaluation against the GENCODE v39 reference, considering splicing patterns and splice junction support derived from SR RNA-seq data. This evaluation led to the categorization of isoforms into structural types, including FSM, ISM, NIC, NNC, genic, intron, antisense, fusion, and intergenic. Transcript-level and gene-level expression, measured in TPM, were calculated using Kallisto (v0.48.0) (57) for SR RNA-seq data, based on the transcriptome generated from long-read RNA-seq data.

Quantification of gene and isoform expression levels from SR RNA-seq data

Expression of isoforms and genes in the aggregated transcriptome derived from LR RNA-seq data was quantified using Kallisto (v0.48.0) from our SR RNA-seq data of DUX4i myoblast samples and publicly available data from the following studies: SR RNA-seq data of DUX4i myoblasts using lentivirus construction [Young et al. (18), Gene Expression Omnibus (GEO) database-ID GSE45883], SR RNAseq data of DUX4i myoblasts with time-series DOX-treatment [Campbell et al. (21), GEO-ID GSE178761], SR RNA-seq data of primary myotubes derived from patients with FSHD and healthy donors [Yao et al. (27), GEO-ID GSE56787], SR RNA-seq data of primary myotubes from patients with FSHD-1 and healthy donors [Banerji et al. (42), GEO-ID GSE153523], SR RNA-seq data of human preimplantation embryos [Torre et al. (35), GEO-ID GSE190544], SR RNA-seq data of DUX4i hESCs (Yoshihara et al. (58), ArrayExpress database ID E-MTAB-10569), SR RNA-seq data of DUX4i iPSCs [Hendrickson et al. 2, GEO-ID GSE95516], and single-cell RNA-seq data of human preimplantation embryos [Yan et al. (36), GEO-ID GSE36552]. In summary, the raw Sequence Read Archive (SRA) data of each sample was retrieved from the GEO database and converted into fastq files using fasterq-dump (v2.11.0). Subsequently, raw fastq files underwent adapter trimming using Trim Galore (v0.6.7) with specific parameters based on the data being single-end or paired-end sequencing. Trimmed reads for each sample were input into Kallisto for the quantification of expression levels normalized in TPM for each full-length isoform. Gene-level expression within the transcriptome was assessed using the R package tximport (59), based on the results obtained

from Kallisto. The trimmed reads were aligned to the GRCh38 reference genome (GENCODE v39) using STAR (60) in two-pass mapping mode. The gene-level expression of SR RNA-seq data was quantified using featureCounts (61), supplied with the GENCODE gene transfer format (GTF) annotation file of the GRCh38 reference genome. BigWig files displaying LR and SR RNA-Seq pileup were generated from the bam files using bamCoverage from deepTools (62), normalized in counts per million (CPM).

Differential expression analysis and correlation analysis of SR RNA-seq data

To quantify isoform expression, raw count values were obtained using Kallisto based on our full-length transcriptome derived from LR RNA-seq. We imported the raw count matrix into DESeq2, retaining isoforms with nonzero expression in at least one sample for subsequent analysis. After TMM normalization, differential expression analysis was performed at the isoform level. Significantly differentially expressed isoforms were identified using thresholds of |log2 fold change| > 2 and adjusted *P* value <0.05. To assess sample similarity, rlog-transformed data were used for Spearman correlation analysis.

Analysis of Ribo-seq data

Publicly available Ribo-seq data [Campbell et al. (21), GEO-ID GSE178760] was downloaded from the SRA database. The adapter in the fastq files was removed by cutadapt (v1.18). The trimmed reads were aligned against rRNA, snoRNA, and miRNA using Bowtie2 (v2.2.5). The reads aligned to these genomes were deleted, and the remaining reads were mapped to the GRCh38 reference genome (GENCODE v39) using STAR (v2.7.9a) in two-pass mode. After removing the duplicated reads using Samtools (v1.9), only the uniquely mapped reads were retained for the downstream analysis. The selected reads for each sample were imported into Kallisto (v0.48.0) for the quantification of expression levels normalized in TPM for each isoform within the full-length transcriptome. Translation efficiency was calculated as the ratio of ribosome-protected fragment abundance to corresponding mRNA levels. We first filtered out isoforms with low mRNA expression levels (TPM < 0.5) to avoid technical artifacts. Translation efficiency values were then computed by dividing normalized ribosome footprint counts by normalized mRNA expression levels for each isoform, providing a quantitative measure of translational potential.

Repeat annotation of the full-length transcriptome of DUX4i myoblasts

The nucleotide sequence of each isoform from the merged transcriptome was scanned for the presence of REs using RepeatMasker (v4.1.4) (63) with default parameters. The merged transcriptome was split into two transcriptomes according to the presence of isoforms under each condition. The percentage of REs was calculated in each transcriptome to assess the repeat content. The abundance of each TE in LR and SR RNA-seq data was assessed using TEtranscripts (v2.2.3) (64).

ASE analysis

The full-length transcriptome underwent comprehensive alternative splicing analysis using SUPPA2 (v2.3) (65) to identify ASEs including A3/A5 (alternative 3′ and 5′ splice sites), AF/AL (alternative first and last exons), SE (skipping exon), RI (retained intron), and MX

(mutually exclusive exon). Specifically, the generateEvent command from SUPPA2 was used to detect ASEs from GTF files containing FSM, NIC, and NNC isoforms expressed in DUX4− and DUX4+ samples, respectively. This command produced ioe output files, delineating local ASEs found in the GTF files. Novel splicing events were identified if all transcripts containing the events were novel isoforms (NIC or NNC), while events found in at least one known isoform (FSM) were categorized as known. GO terms (BP) enrichment analysis was performed on the genes associated with isoforms with novel ASEs for the DUX4+ sample using the R package Clusterprofiler (v4.2.2) (66).

Protein-level functional analysis of full-length isoform

The ORF nucleotide sequence for each coding FSM, NIC, NNC, and intergenic isoform was translated into its corresponding amino acid sequence using SQANTI3, with internal utilization of the software GeneMarkS-T. Subsequently, local alignment was performed through the blastp module of Diamond (v0.9.21) (67) against the human reference proteome to identify homologs in the UniProt database (30). ORFs with an identity score exceeding 99% were considered known, while those scoring below 99% were categorized as novel (31). In addition, for the ORFs of coding intergenic isoforms, PFAM domains were predicted (Protein families database). This prediction was executed using the hmmscan tool from hmmer (v.3.2.1) with default parameters. The selection of a single best ORF for each coding intergenic transcript was based on significant sequence homology and domain conservation to reference proteins.

Identification of cellular context–specific exons of DUX4 target genes

All the genomic coordinates of each exon of DUX4 target genes' isoforms (only FSM, NIC, and NNC) were extracted from the transcriptomes of DUX4i myoblasts and four-cell stage cells to generate bed files. We used bedtools (v2.30.0) to identify the exon with no overlap with others using the subtract function with parameter -A.

Analysis of publicly available ChIP-seq, ATAC-seq, and CUT&RUN data

Publicly available DUX4 ChIP-seq data from DUX4i myoblasts [Geng et al. (4), GEO-ID GSE33838], DUX4i hESCs [De laco et al. (1), GEO-ID GSE94322], and DUX4i iPSCs [Hendrickson et al. (2), GEO-ID GSE95515], along with ATAC-seq data from DUX4i hESCs [Vuoristo et al. (22), GEO-ID GSE171803] and preimplantation embryos [Liu et al. (45), SRP163205], were obtained from the SRA database. Following converting SRA data to fastq files, Trim Galore was applied to preprocess raw reads by trimming low-quality reads and removing adapters, using specific parameters based on single-end or paired-end sequencing data.

For DUX4 ChIP-seq data, processed reads were aligned to the GRCh38 reference genome (GENCODE v39) using Bowtie2 (v2.2.5) (68), and duplicated reads were subsequently removed using Samtools (v1.9) (69). Peak calling was performed using MACS2 (v2.2.6)

(70) with parameters -f BAMPE -g hs --bdg -q 0.05 for paired-end sequencing data and -f BAM -g hs --bdg -q 0.05 for single-end sequencing data. We noted that the DUX4 ChIP-seq data of three cell lines showed varying coverage and used different strategies for library construction. We first calculated the read coverage genome-wide with bin size = 10 kb for each sample using

the function multiBamSummary of Deeptools (v2.5.7). Then, we performed Spearman correlation analysis using the function plotCorrelation of Deeptools to show the similarity between samples. The result revealed a good correlation between hESCs and iPSCs with DUX4 induction, while all myogenic cells were clustered together (fig. S22A). Then, we down-sampled the samples of DUX4i hESCs and iPSCs by randomly extracting 50% of reads in each sample so that they had a comparable number of reads to the myoblast data. The Spearman correlation analysis showed highly conserved results compared to the results using original data (fig. S22B). Last, we only used the forward reads of ChIP-seq data of DUX4i hESCs and iPSCs and kept the first 49 bp of each read since the average length of reads in ChIP-seq data of DUX4i myoblasts is 49 bp. The retained fastq files were considered single-end sequencing data. We used the same pipeline to trim the data and mapped them to the human genome. The overall mapping rates of these single-end samples were comparable to those of the paired-end samples. After removing the duplicated reads, we counted the read coverage and performed the Spearman correlation analysis. The results were still comparable to the original data (fig. S22C).

For the ATAC-seq data, trimmed reads were mapped to the GRCh38 reference genome (GENCODE v39) using STAR in two-pass mode. ATAC-seq peaks were called using MACS2 with parameters -f BAM -g hs -B -q 0.01 --nomodel --nolambda --extsize 250, pooling replicates. Bigwig files were generated from filtered bam files using the bamCoverage command of Deeptools (normalized in CPM) to visualize peak signal intensity.

The liftover from UCSC utilities and CrossMap (0.7.0) (71) were used to convert bed and bigwig file coordinates from hg19 to hg38 for publicly available CUT&RUN datasets of human preimplantation embryos [Xia et al. (48), GEO-ID GSE124718]. The original results of H3K9me3 ChIP-seq data [Yu et al. (49), GEO-ID GSE176016] from human early embryonic cells, aligned to the hg38 reference genome, were directly used.

The identified peaks from ChIP-seq, ATAC-seq, and CUT&RUN datasets were used to annotate intergenic isoforms using the R package ChIPseeker (v1.30.3) (72). The ATAC-seq peaks within ±2 kb of the TSSs of intergenic isoforms were used as target sequences. The enrichment of known and *de novo* motifs was determined by findMotifsGenome.pl from HOMER (v4.11.1). The heatmaps showing the intensity of peaks were generated using the EnrichedHeatmap R package (v.1.16.0) (73).

Analysis of publicly available RRBS data

Publicly available RRBS data of human preimplantation embryos [Guo et al. (46), GEO-ID GSE49828] was downloaded from the SRA database. The RRBS data of primary myoblasts derived from patients with FSHD2 and healthy donors [Mason et al. (47)] was shared by the Tapscott Lab. Trim Galore (v0.6.7) was used to trim the adapter in the raw reads with the following parameters: -q 20 --phred33 --stringency 3 -j 4 --length 35 -e 0.1 --rrbs. Bismark (v0.24.1) (74) was used to align the trimmed reads to the GRCh38 reference genome (GENCODE v39). The methylation extractor module from Bismark was applied to generate a genome-wide report of cytosine methylation in the CpG context with the default setting. DNA methylation within the regions ±500 bp of TSSs of intergenic isoforms was calculated by the R package methylKit (v1.20.0) (75). For each region, the percentage of DNA methylation was

defined by dividing the number of identified Cs (methylated reads) by the total number of identified Cs and Ts (unmethylated reads) within the region.

Statistical analysis

The statistical analyses were performed using R software (version 4.1.3). Statistical significance was determined using the Student's *t* test (two-tailed) or unpaired Wilcoxon test. Statistical significance was defined as ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, and $P > 0.05$ was represented as ns (not significant). All Hypergeometric tests were performed to assess the significance of overlaps between datasets using R software (version 4.1.3), with the corresponding background set sizes.

Supplementary Materials

The PDF file includes:

Figs. S1 to S22

tables S1 to S3, S7 and S9

legends for tables S4 to S6, S8 and S10 to S16

Other Supplementary Material for this manuscript includes the following:

tables S4 to S6, S8 and S10 to S16

References and Notes

1. A. De laco, E. Planet, A. Coluccio, S. Verp, J. Duc, D. Trono, DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat. Genet. 49, 941-945 (2017).
2. P. G. Hendrickson, J. A. Doráis, E. J. Grow, J. L. Whiddon, J.-W. Lim, C. L. Wike, B. D. Weaver, C. Pflueger, B. R. Emery, A. L. Wilcox, D. A. Nix, C. M. Peterson, S. J. Tapscott, D. T. Carrell, B. R. Cairns, Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat. Genet. 49, 925-934 (2017).
3. J. L. Whiddon, A. T. Langford, C.-J. Wong, J. W. Zhong, S. J. Tapscott, Conservation and innovation in the DUX4-family gene network. Nat. Genet. 49, 935-940 (2017).
4. L. N. Geng, Z. Yao, L. Snider, A. P. Fong, J. N. Cech, J. M. Young, S. M. van der Maarel, W. L. Ruzzo, R. C. Gentleman, R. Tawil, S. J. Tapscott, DUX4 activates germline genes, retroelements, and immune mediators: Implications for facioscapulohumeral dystrophy. Dev. Cell 22, 38-51 (2012).
5. R. J. Lemmers, P. J. van der Vliet, R. Klooster, S. Sacconi, P. Camano, J. G. Dauwerse, L. Snider, K. R. Straasheijm, G. J. van Ommen, G. W. Padberg, D. G. Miller, S. J. Tapscott, R. Tawil, R. R. Frants, S. M. van der Maarel, A unifying genetic model for facioscapulohumeral muscular dystrophy. Science 329, 1650-1653 (2010).
6. A. M. Rickard, L. M. Petek, D. G. Miller, Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum. Mol. Genet. 24, 5901-5914 (2015).

7. L. H. Wang, S. D. Friedman, D. Shaw, L. Snider, C.-J. Wong, C. B. Budech, S. L. Poliachik, N. E. Gove, L. M. Lewis, A. E. Campbell, R. Lemmers, S. M. Maarel, S. J. Tapscott, R. N. Tawil, MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. Hum. Mol. Genet. 28, 476-486 (2019).
8. J. R. Dahlqvist, G. Andersen, T. Khawajazada, C. Vissing, C. Thomsen, J. Vissing, Relationship between muscle inflammation and fat replacement assessed by MRI in facioscapulohumeral muscular dystrophy. J. Neurol. 266, 1127-1135 (2019).
9. C.-J. Wong, L. H. Wang, S. D. Friedman, D. Shaw, A. E. Campbell, C. B. Budech, L. M. Lewis, R. Lemmers, J. M. Statland, S. M. van der Maarel, R. N. Tawil, S. J. Tapscott, Longitudinal measures of RNA expression and disease activity in FSHD muscle biopsies. Hum. Mol. Genet. 29, 1030-1043 (2020).
10. P. Dmitriev, Y. Bou Saada, C. Dib, E. Ansseau, A. Barat, A. Hamade, P. Dessen, T. Robert, V. Lazar, R. A. N. Louzada, C. Dupuy, V. Zakharova, G. Carnac, M. Lipinski, Y. S. Vassetzky, DUX4-induced constitutive DNA damage and oxidative stress contribute to aberrant differentiation of myoblasts from FSHD patients. Free Radic. Biol. Med. 99, 244-258 (2016).
11. P. Heher, M. Ganassi, A. Weidinger, E. N. Engquist, J. Pruller, T. H. Nguyen, A. Tassin, A.-E. Declèves, K. Mamchaoui, C. R. S. Banerji, J. Grillari, A. V. Kozlov, P. S. Zammit, Interplay between mitochondrial reactive oxygen species, oxidative stress and hypoxic adaptation in facioscapulohumeral muscular dystrophy: Metabolic stress as potential therapeutic target. Redox Biol. 51, 102251 (2022).
12. D. Bosnakovski, Z. Xu, E. J. Gang, C. L. Galindo, M. Liu, T. Simsek, H. R. Garner, S. Agha-Mohammadi, A. Tassin, F. Coppée, A. Belayew, R. R. Perlingeiro, M. Kyba, An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. EMBO J. 27, 2766-2779 (2008).
13. A. Turki, M. Hayot, G. Carnac, F. Pillard, E. Passerieux, S. Bommart, E. Raynaud de Mauverger, G. Hugon, J. Pincemail, S. Pietri, K. Lambert, A. Belayew, Y. Vassetzky, R. Juntas Morales, J. Mercier, D. Laoudj-Chenivesse, Functional muscle impairment in facioscapulohumeral muscular dystrophy is correlated with oxidative stress and mitochondrial dysfunction. Free Radic. Biol. Med. 53, 1068-1079 (2012).
14. A. Tassin, D. Laoudj-Chenivesse, C. Vanderplanck, M. Barro, S. Charron, E. Ansseau, Y.-W. Chen, J. Mercier, F. Coppée, A. Belayew, DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J. Cell. Mol. Med. 17, 76–89 (2013).
15. L. Snider, L. N. Geng, R. J. Lemmers, M. Kyba, C. B. Ware, A. M. Nelson, R. Tawil, G. N. Filippova, S. M. van der Maarel, S. J. Tapscott, D. G. Miller, Facioscapulohumeral dystrophy: Incomplete suppression of a retrotransposed gene. PLOS Genet. 6, e1001181 (2010).
16. D. Bosnakovski, M. D. Gearhart, E. A. Toso, E. T. Ener, S. H. Choi, M. Kyba, Low level DUX4 expression disrupts myogenesis through deregulation of myogenic gene expression. Sci. Rep. 8, 16957 (2018).
17. S. Jagannathan, S. C. Shadle, R. Resnick, L. Snider, R. N. Tawil, S. M. van der Maarel, R. K. Bradley, S. J. Tapscott, Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum. Mol. Genet. 25, 4419-4431 (2016).

18. J. M. Young, J. L. Whiddon, Z. Yao, B. Kasinathan, L. Snider, L. N. Geng, J. Balog, R. Tawil, S. M. van der Maarel, S. J. Tapscott, DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLOS Genet. 9, e1003947 (2013).
19. S. Mitsuhashi, S. Nakagawa, M. Sasaki-Honda, H. Sakurai, M. C. Frith, H. Mitsuhashi, Nanopore direct RNA sequencing detects DUX4-activated repeats and isoforms in human muscle cells. Hum. Mol. Genet. 30, 552-563 (2021).
20. S. Jagannathan, Y. Ogata, P. R. Gafken, S. J. Tapscott, R. K. Bradley, Quantitative proteomics reveals key roles for post-transcriptional gene regulation in the molecular pathology of facioscapulohumeral muscular dystrophy. eLife 8, e41740 (2019).
21. A. E. Campbell, M. C. Dyle, R. Albanese, T. Matheny, K. Sudheendran, M. A. Cortázar, T. Forman, R. Fu, A. E. Gillen, M. H. Caruthers, S. N. Floor, L. Calviello, S. Jagannathan, Compromised nonsense-mediated RNA decay results in truncated RNA-binding protein production upon DUX4 expression. Cell Rep. 42, 112642 (2023).
22. S. Vuoristo, S. Bhagat, C. Hyden-Granskog, M. Yoshihara, L. Gawriyski, E. M. Jouhilahti, V. Ranga, M. Tamirat, M. Huhtala, I. Kirjanov, S. Nykänen, K. Krjutskov, A. Damdimopoulos, J. Weltner, K. Hashimoto, G. Recher, S. Ezer, P. Paluoja, P. Paloviita, Y. Takegami, A. Kanemaru, K. Lundin, T. T. Airenne, T. Otonkoski, J. S. Tapanainen, H. Kawaji, Y. Murakawa, T. R. Burglin, M. Varjosalo, M. S. Johnson, T. Tuuri, S. Katayama, J. Kere, DUX4 is a multifunctional factor priming human embryonic genome activation. iScience 25, 104137 (2022).
23. S. A. Hardwick, A. Joglekar, P. Flicek, A. Frankish, H. U. Tilgner, Getting the entire message: Progress in isoform sequencing. Front. Genet. 10, 709 (2019)
24. A. Rhoads, K. F. Au, PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13, 278-289 (2015).
25. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. C. Sala, J. Chrast, F. Cunningham, T. D. Domenico, S. Donaldson, I. T. Fiddes, C. G. Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE: Reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res. 51, D766-D773 (2023).g
26. I. Abugessaisa, S. Noguchi, A. Hasegawa, A. Kondo, H. Kawaji, P. Carninci, T. Kasukawa, refTSS: A reference data set for human and mouse transcription start sites. J. Mol. Biol. 431, 2407-2422 (2019).
27. Z. Yao, L. Snider, J. Balog, R. J. Lemmers, S. M. Van Der Maarel, R. Tawil, S. J. Tapscott, DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum. Mol. Genet. 23, 5342-5352 (2014).
28. Q. Feng, L. Snider, S. Jagannathan, R. Tawil, S. M. van der Maarel, S. J. Tapscott, R. K. Bradley, A feedback loop between nonsense-mediated decay and the retrogene DUX4 in facioscapulohumeral muscular dystrophy. eLife 4, e04996 (2015).
29. M. Tardaguila, L. de la Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. Del Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M.

Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno-Manzano, A. Conesa, SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28, 396 411 (2018).
30. UniProt Consortium, UniProt: The Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523-D531 (2023).
31. D. F. T. Veiga, A. Nesta, Y. Zhao, A. Deslattes Mays, R. Huynh, R. Rossi, T.-C. Wu, K. Palucka, O. Anczukow, C. R. Beck, J. Banchereau, A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. Sci. Adv. 8, eabg6711 (2022).
32. A. van den Heuvel, A. Mahfouz, S. L. Kloet, J. Balog, B. G. van Engelen, R. Tawil, S. J. Tapscott, S. M. van der Maarel, Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum.Mol. Genet. 28, 1064-1075 (2019).
33. S. Jiang, K. Williams, X. Kong, W. Zeng, N. V. Nguyen, X. Ma, R. Tawil, K. Yokomori, A. Mortazavi, Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLOS Genet. 16, e1008754 (2020).
34. D. Zheng, A. Wondergem, S. Kloet, I. Willemsen, J. Balog, S. J. Tapscott, A. Mahfouz, A. van den Heuvel, S. M. van der Maarel, snRNA-seq analysis in multinucleated myogenic FSHD cells identifies heterogeneous FSHD transcriptome signatures associated with embryonic-like program activation and oxidative stress-induced apoptosis. Hum. Mol. Genet. 33, 284-298 (2024).
35. D. Torre, N. J. Francoeur, Y. Kalma, I. Gross Carmel, B. S. Melo, G. Deikus, K. Allette, R. Flohr, M. Fridrikh, K. Vlachos, K. Madrid, H. Shah, Y. C. Wang, S. H. Sridhar, M. L. Smith, E. Eliyahu, F. Azem, H. Amir, Y. Mayshar, I. Marazzi, E. Guccione, E. Schadt, D. Ben-Yosef, R. Sebra, Isoform-resolved transcriptome of the human preimplantation embryo. Nat. Commun. 14, 6902 (2023).
36. L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, F. Tang, Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131-1139 (2013).
37. S. Jagannathan, J. C. de Greef, L. J. Hayward, K. Yokomori, D. Gabellini, K. Mul, S. Sacconi, J. Arjomand, J. Kinoshita, S. Q. Harper, Meeting report: The 2021 FSHD International Research Congress. Skelet Muscle 12, 1 (2022).
38. A. L. Mueller, A. O'Neill, T. I. Jones, A. Llach, L. A. Rojas, P. Sakellariou, G. Stadler, W. E. Wright, D. Eyerman, P. L. Jones, R. J. Bloch, Muscle xenografts reproduce key molecular features of facioscapulohumeral muscular dystrophy. Exp. Neurol. 320, 113011 (2019).
39. A. van den Heuvel, S. Lassche, K. Mul, A. Greco, D. S. L. Granado, A. Heerschap, B. Küsters, S. J. Tapscott, N. C. Voermans, B. G. van Engelen, Facioscapulohumeral dystrophy transcriptome signatures correlate with different stages of disease and are marked by different MRI biomarkers. Sci. Rep. 12, 1426 (2022).
40. X. Kong, N. V. Nguyen, Y. Li, J. S. Sakr, K. Williams, S. Sharifi, J. Chau, A. Bayrakci, S. Mizuno, S. Takahashi, T. Kiyono, R. Tawil, A. Mortazavi, K. Yokomori, Engineered FSHD mutations results in D4Z4 heterochromatin disruption and feedforward DUX4 network activation. iScience 27, 109357 (2024).

41. A. E. Campbell, A. E. Belleville, R. Resnick, S. C. Shadle, S. J. Tapscott, Facioscapulohumeral dystrophy: Activating an early embryonic transcriptional program in human skeletal muscle. Hum. Mol. Genet. 27, R153-R162 (2018).
42. C. R. S. Banerji, M. Panamarova, P. S. Zammit, DUX4 expressing immortalized FSHD lymphoblastoid cells express genes elevated in FSHD muscle biopsies, correlating with the early stages of inflammation. Hum. Mol. Genet. 29, 2285-2299 (2020).
43. A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, D. D. Pollock, Repetitive elements may comprise over two-thirds of the human genome. PLOS Genet. 7, e1002384 (2011).
44. X.-O. Zhang, H. Pratt, Z. Weng, Investigating the potential roles of SINEs in the human genome. Annu. Rev. Genomics Hum. Genet. 22, 199-218 (2021).
45. L. Liu, L. Leng, C. Liu, C. Lu, Y. Yuan, L. Wu, F. Gong, S. Zhang, X. Wei, M. Wang, L. Zhao, L. Hu, J. Wang, H. Yang, S. Zhu, F. Chen, G. Lu, Z. Shang, G. Lin, An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat. Commun. 10, 364 (2019).
46. H. Guo, P. Zhu, L. Yan, R. Li, B. Hu, Y. Lian, J. Yan, X. Ren, S. Lin, J. Li, X. Jin, X. Shi, P. Liu, X. Wang, W. Wang, Y. Wei, X. Li, F. Guo, X. Wu, X. Fan, J. Yong, L. Wen, S. X. Xie, F. Tang, J. Qiao, The DNA methylation landscape of human early embryos. Nature 511, 606-610 (2014).
47. A. G. Mason, R. C. Slieker, J. Balog, R. J. L. F. Lemmers, C.-J. Wong, Z. Yao, J.-W. Lim, G. N. Filippova, E. Ne, R. Tawil, B. T. Heijmans, S. J. Tapscott, S. M. van der Maarel, SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes. Skelet. Muscle 7, 12 (2017).
48. W. Xia, J. Xu, G. Yu, G. Yao, K. Xu, X. Ma, N. Zhang, B. Liu, T. Li, Z. Lin, X. Chen, L. Li, Q. Wang, D. Shi, S. Shi, Y. Zhang, W. Song, H. Jin, L. Hu, Z. Bu, Y. Wang, J. Na, W. Xie, Y.-P. Sun, Resetting histone modifications during human parental-to-zygotic transition. Science 365, 353-360 (2019).
49. H. Yu, M. Chen, Y. Hu, S. Ou, X. Yu, S. Liang, N. Li, M. Yang, X. Kong, C. Sun, S. Jia, Q. Zhang, L. Liu, L. D. Hurst, R. Li, W. Wang, J. Wang, Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development. Cell Stem Cell 29, 1031-1050.e12 (2022).
50. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
51. T. S. Macfarlan, W. D. Gifford, S. Driscoll, K. Lettieri, H. M. Rowe, D. Bonanomi, A. Firth, O. Singer, D. Trono, S. L. Pfaff, Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature 487, 57-63 (2012).
52. J. W. Jachowicz, X. Bing, J. Pontabry, A. Bošković, O. J. Rando, M.-E. Torres-Padilla, LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat. Genet. 49, 1502-1510 (2017).
53. K. N. Schulz, M. M. Harrison, Mechanisms regulating zygotic genome activation. Nat. Rev. Genet. 20, 221-234 (2019).
54. W. Xia, W. Xie, Rebooting the epigenomes during mammalian early embryogenesis. Stem Cell Rep. 15, 1158-1175 (2020).
55. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100 (2018).

56. F. J. Pardo-Palacios, A. Arzalluz-Luque, L. Kondratova, P. Salguero, J. Mestre-Tomás, R. Amorín, E. Estevan-Morió, T. Liu, A. Nanni, L. McIntyre, E. Tseng, A. Conesa, SQANTI3: Curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat. Methods 21, 793-797 (2024).
57. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525-527 (2016).
58. M. Yoshihara, I. Kirjanov, S. Nykänen, J. Sokka, J. Weltner, K. Lundin, L. Gawriyski, E.-M. Jouhilahti, M. Varjosalo, M. H. Tervaniemi, T. Otonkoski, R. Trokovic, S. Katayama, S. Vuoristo, J. Kere, Transient DUX4 expression in human embryonic stem cells induces blastomere-like expression program that is marked by SLC34A2. Stem Cell Rep. 17, 1743-1756 (2022).
59. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Res 4, 1521 (2015).
60. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).
61. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-930 (2014).
62. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar, T. Manke, deepTools2: A next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160-W165 (2016).
63. A. F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657-663 (1999).
64. Y. Jin, O. H. Tam, E. Paniagua, M. Hammell, TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics 31, 3593-3599 (2015).
65. J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, E. Eyras, SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 19, 40 (2018).
66. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, G. Yu, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2, 100141 (2021).
67. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59-60 (2015).
68. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359 (2012).
69. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).
70. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).
71. H. Zhao, Z. Sun, J. Wang, H. Huang, J. P. Kocher, L. Wang, CrossMap: A versatile tool for coordinate conversion between genome assemblies. Bioinformatics 30, 1006-1007 (2014).

72. G. Yu, L.-G. Wang, Q.-Y. He, ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382-2383 (2015).
73. Z. Gu, R. Eils, M. Schlesner, N. Ishaque, EnrichedHeatmap: An R/Bioconductor package for comprehensive visualization of genomic signal associations. BMC Genomics 19, 234 (2018).
74. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571-1572 (2011).
75. A. Akalin, M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, C. E. Mason, methylKit: A comprehensive R package for the analysis of genome wide DNA methylation profiles. Genome Biol. 13, R87 (2012).

Acknowledgments: We thank L. Clemens-Daxinger for reviewing the manuscript and providing valuable and constructive feedback.

Funding: This study was supported by the Prinses Beatrix Spierfonds (W.OR19-06) to S.M.v.d.M. and US national institute of Arthritis and Musculoskeletal and Skin diseases (NIAMS) R01AR066248 to S.J.T. and S.M.v.d.M.

Author contributions:

Conceptualization: D.Z., A.v.d.H., S.M.v.d.M., and S.J.T.

Methodology: A.v.d.H., D.Z., S.M.v.d.M., A.M., and J.B.

Software: D.Z.

Validation: D.Z.

Formal analysis: D.Z.

Investigation: D.Z., J.B., and I.M.W.

Resources: A.v.d.H., S.M.v.d.M., and S.K.

Data curation: D.Z.

Writing—original draft: D.Z. and A.M.

Writing—review and editing: D.Z., S.M.v.d.M., A.v.d.H., A.M., S.J.T., and S.K. Visualization: D.Z.

Supervision: S.M.v.d.M., A.M., and A.v.d.H.

Project administration: S.M.v.d.M. and J.B.

Funding acquisition: S.M.v.d.M. and S.J.T.

Competing interests: S.M.v.d.M. and S.J.T. are board members of Renogenyx and have acted as consultants and/or are members of the advisory board for several pharmaceutical companies developing therapeutics for FSHD. S.M.v.d.M. is coinventor of several FSHD-related patent applications. All other authors declare that they have no competing interests.

Data and materials availability: The long-and short-read RNA sequencing data of DUX4i myoblasts that support the findings of this study are deposited in the European Genome-

phenome Archive (EGA) with the accession code EGAS50000000503 (LR RNA-seq data: EGAD50000000718; SR RNA-seq data: EGAD50000000719), under restricted access (due to privacy sensitivity of the raw sequencing data). Interested researchers should use the following links to request access: https://ega-archive.org/access/request-data/how-to-request-data/. All datasets under accession number EGAS50000000503 are available at the following link: https://ega-archive.org/studies/EGAS50000000503. The interested researchers should provide a description of the intended data use, including research purpose and data protection measures that will be implemented. Access requests will be reviewed by the corresponding author on the basis of legitimate scientific purpose and commitment to appropriate data protection. The review process typically takes 1 to 2 weeks. Upon approval, the requestors will receive credentials to download the dataset. No additional access forms are required beyond the information submitted through the EGA portal. The publicly available RNA sequencing data of DUX4i myoblasts are available in the GEO under accession number GSE45883 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45883). The publicly available RNA sequencing data of DUX4i myoblasts with time-series treatment are available in GEO under accession number GSE178759 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178759). The publicly available Ribosome profiling sequencing of DUX4i myoblasts with time-series treatment is available in GEO under accession number GSE178760 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178760). The two publicly available primary myotube RNA sequencing datasets are available in GEO under accession numbers GSE5678 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56787) and GSE153523 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153523). The publicly available long- and short-RNA sequencing data of human preimplantation embryos are available in GEO under accession numbers GSE190547 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190547) and GSE190544 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190544). The publicly available RNA sequencing data of DUX4i iPSCs are available in GEO under accession number GSE95516 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95516). The publicly available single-cell RNA sequencing data of human early embryos are available in GEO under accession number GSE36552 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552). The publicly available RNA sequencing data of DUX4i hESCs are available in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-10569 (https://ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10569?query=E-MTAB-10569).
The publicly available DUX4 ChIP-seq datasets of DUX4i myoblasts, hESCs, and iPSCs are available in GEO under accession numbers GSE33838 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33838), GSE94322 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94322), and GSE95515 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95515), respectively. The publicly available H3K9me3 ChIP-seq data of early human embryos are available in GEO under accession number GSE176016 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176016). The publicly available ATAC-seq data of DUX4i hESCs are available in GEO under accession number GSE171803 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171803). The publicly available ATAC-seq data of human preimplantation embryos are available in GEO under accession number PRJNA494280 (https://ncbi.nlm.nih.gov/bioproject/PRJNA494280).

The publicly available H3K27ac and H3K4me3 CUT&RUN data of human preimplantation embryos are available in GEO under accession number GSE124718 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124718). The publicly available RRBS data of human preimplantation embryos are available in GEO under accession number GSE49828 (https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49828). Tapscott Lab shared the RRBS data of patients with FSHD and healthy donors to support this study. All supporting scripts and data derived from this study have been deposited in Zenodo (https://doi.org/10.5281/zenodo.15003341) with restricted access to protect patient privacy and prevent potential re-identification of study participants. They are available from the corresponding author on request.

Supplementary figures

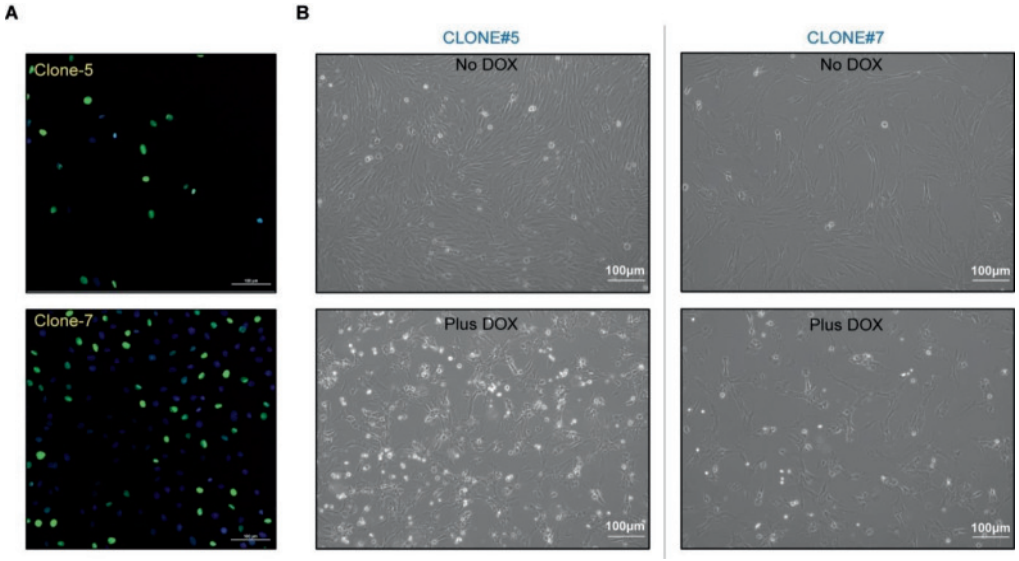

Fig. S1. Validation of doxycycline-inducible DUX4 myoblast cell lines.

(A) Immunofluorescence confocal microscopy image of the monoclonal DUX4i clones 5 and 7 after 8 hours of doxycycline treatment (1 ug/ml). Nuclei are stained in blue (Hoechst 33528) and DUX4 was stained in green. (B) Bright-field microscopy images of DUX4i clones 5 and 7 with and without 24 hours of doxycycline treatment (4 ug/ml). Scale bar = 100 µm

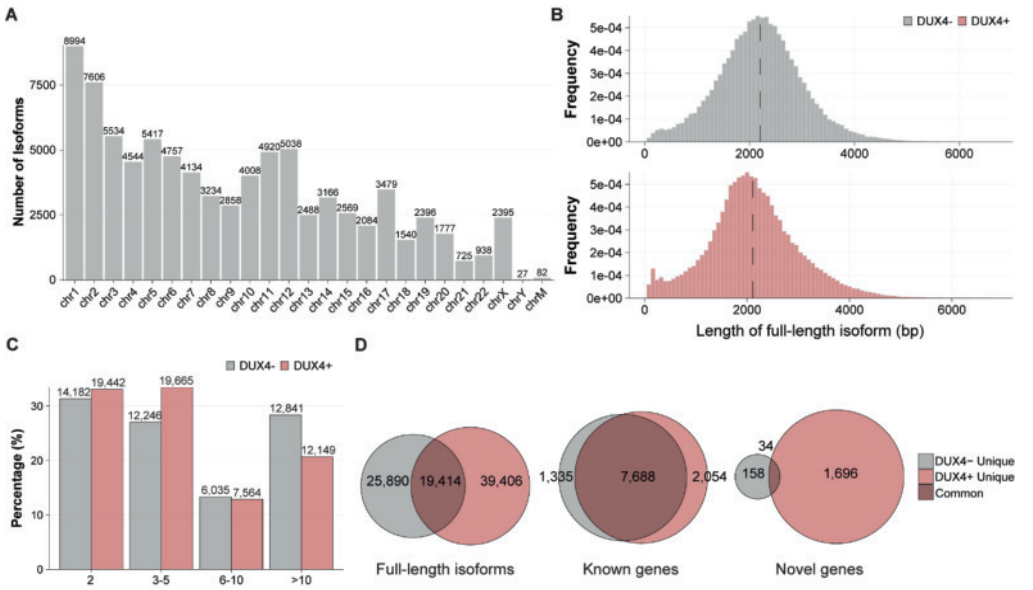

Fig. S2. Landscape of full-length transcriptome in DUX4i myoblasts.

(A) Bar plot showing the number of isoforms localized on autosomes, sex chromosomes, and mitochondrial DNA in the aggregated full-length transcriptome. (B) Density plot depicting the distribution of transcript lengths in the DUX4- and DUX4+ transcriptome. Dashed lines represent the mean of the transcript lengths. (C) Bar plot displaying the read count distribution per isoform in the DUX4- and DUX4+ transcriptome. (D) Venn diagrams showing the overlap between the DUX4- and DUX4+ transcriptome in all isoforms, all associated known genes, and all novel genes. Color codes represent the different categories.

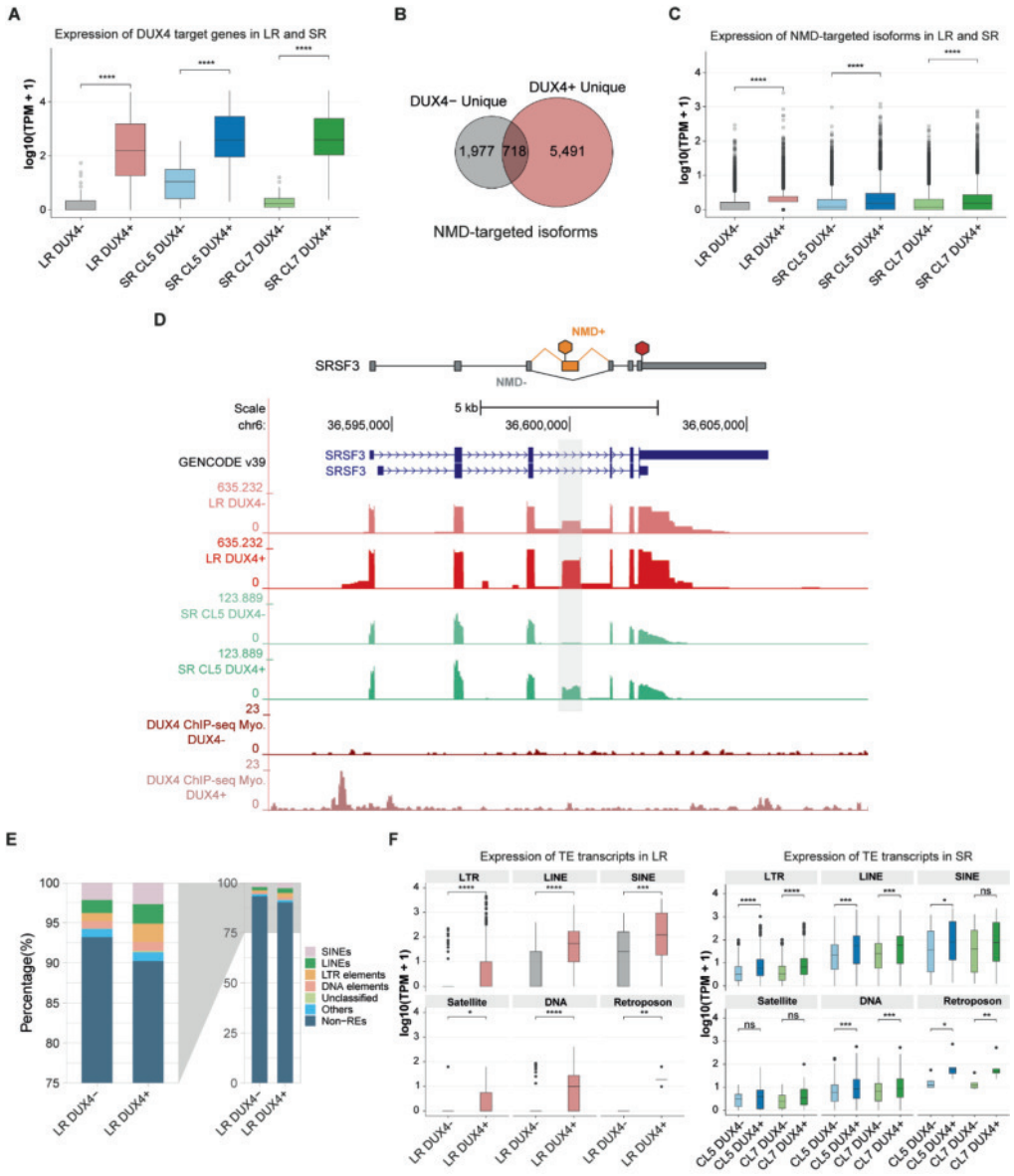

Fig. S3. Iso-Seq captures the hallmarks of DUX4 in DUX4i myoblasts.

(A) Box plot showing the expression levels of 67 core DUX4 target genes in the LR and SR RNA-seq data. *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$. (B) Venn diagram exhibiting the overlapping isoforms predicted as NMD targets between the DUX4- and DUX4+ LR transcriptome. (C) Box plot exhibiting the expression levels of isoforms with NMD label in each sample of LR and SR RNA-seq data. P-values are calculated the same as in (A), SR RNA-seq data are in blue and green. (D) Plot showing the structure of SRSF3 isoforms with NMD-targeted exon highlighted. The plot of SRSF3 transcript with NMD-targeted exon is created

with Biorender.com. Bright red tracks represent the coverage level in LR RNA-seq data. Blue tracks represent the coverage level in SR RNA-seq data. Dark red tracks represent the coverage level in DUX4 ChIP-seq data of DUX4i myoblasts. (E) Stacked bar plot showing the percentage of isoforms annotated with repetitive elements in the full-length transcriptome under DUX4- and DUX4+ conditions. (F) Box plots displaying the expression levels of TE transcripts belonging to different TE families per sample of LR and SR RNA-seq data. *P*-values are calculated the same as in (A). Statistical significance is denoted as follows: ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, and $P > 0.05$ represented as 'ns' (not significant).

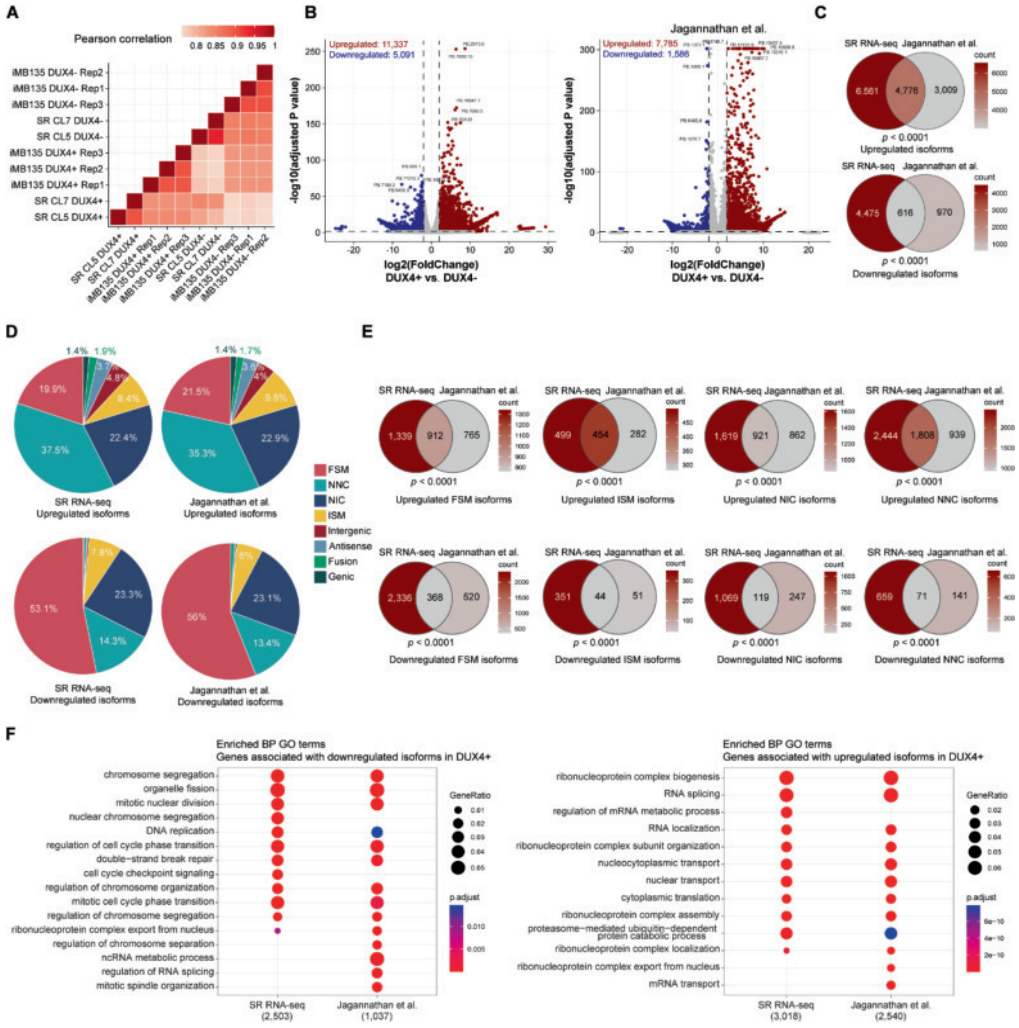

Fig. S4. Consistent transcriptomic changes observed across independent RNA-seq datasets.

(A) Heatmap showing Pearson correlation coefficients between samples in our data and the SR RNA-seq from Jagannathan et al. (17), with deeper red colors indicating higher correlation. (B) Volcano plots showing differentially expressed isoforms between DUX4+ and DUX4- conditions. Red and blue dots represent significantly upregulated and downregulated isoforms in the DUX4+ condition, respectively (|log2 fold change| > 2, adjusted *P*-value < 0.05). Grey dots indicate isoforms that did not meet the significance criteria. (C) Venn diagrams showing the intersection of upregulated and downregulated isoforms identified in our RNA-seq and Jagannathan et al. (17). Hypergeometric tests revealed significant overlaps for both upregulated and downregulated isoforms (both *P*-values < 0.0001). (D) Pie charts depicting the percentage distribution of different isoform categories among upregulated and downregulated isoforms identified in our RNA-seq data and Jagannathan et al. (17). (E) Venn diagrams showing the overlap of upregulate and downregulated isoforms for FSM, ISM, NIC, and NNC categories between our RNA-seq and Jagannathan et al. (17). Hypergeometric tests demonstrated significant

overlaps for all categories (all *P*-values < 0.0001). (F) Dot plots showing significantly enriched BP GO terms (adjusted *P*-value < 0.05) for genes associated with differentially expressed isoforms in our RNA-seq and Jagannathan et al. (17). Dot size represents gene count and color intensity indicates adjusted P-value.

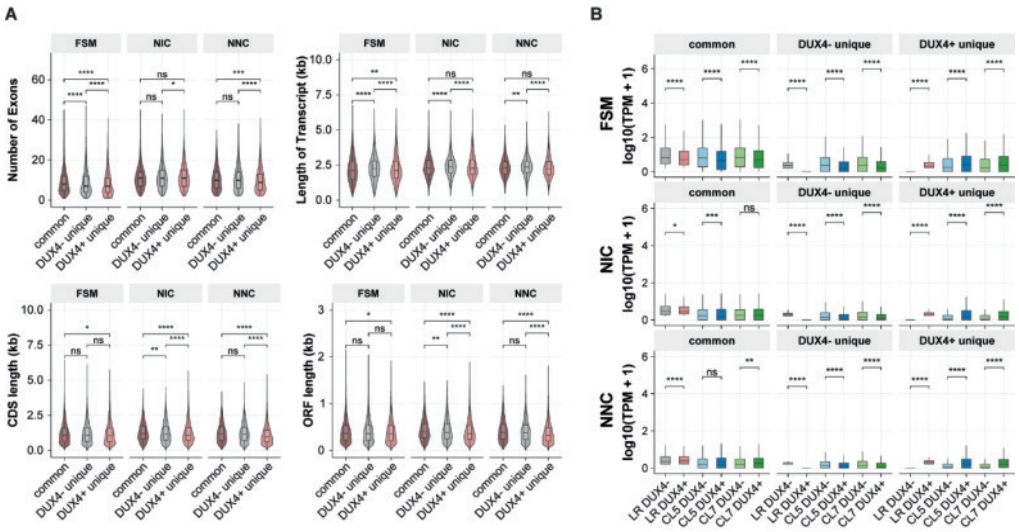

Fig. S5. Characteristics of novel isoforms specific to DUX4- and DUX4+ myoblasts.

(A) Violin plots showing exon number and transcript length of isoforms in each sub-type (uniquely expressed in DUX4-; uniquely expressed in DUX4+; commonly expressed) per structural category (FSM, NIC, NNC) (top row). Violin plots showing the CDS length, and ORF length of coding isoforms in each sub-type (uniquely expressed in DUX4-; uniquely expressed in DUX4+; commonly expressed) per structural category (FSM, NIC, NNC) (bottom row). *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, and $P > 0.05$ represented as 'ns' (not significant). (B) Box plots depicting the expression levels of FSM, NIC, and NNC isoforms in each sub-type per sample of LR and SR RNA-seq data. *P*-values are calculated the same as in (A).

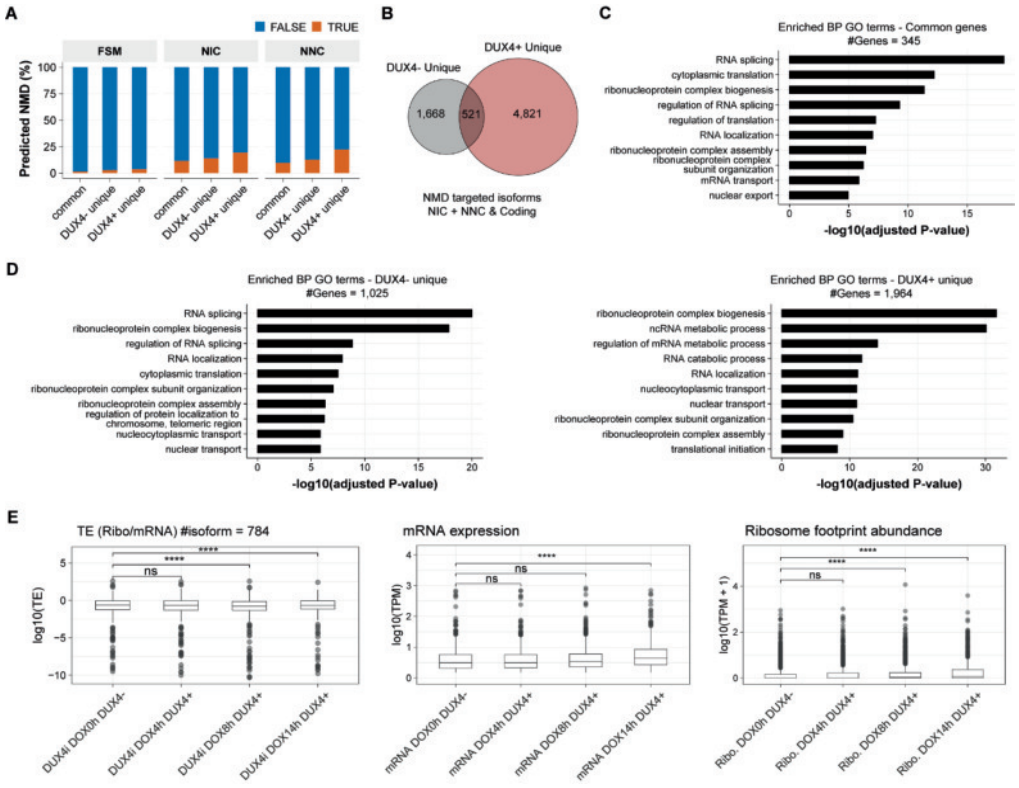

Fig. S6. Characteristics of NMD-targeted isoforms.

(A) Stacked bar plot displaying the percentages of NMD-targeted isoforms. (B) Venn diagram showing the intersected NMD-targeted isoforms (NIC and NNC) between the DUX4- and DUX4+ conditions. (C) Bar plots displaying GO terms (BP) significantly enriched (adjusted *P*-value < 0.05) for genes associated with commonly expressed NMD-targeted isoforms. (D) Bar plots displaying GO terms (BP) significantly enriched (adjusted *P*-value < 0.05) for genes associated with uniquely expressed NMD-targeted isoforms. (E) Box plots showing the distribution of (left) translation efficiency, (middle) mRNA expression levels, and (right) ribosome footprint abundance of the NMD-targeted isoforms classified into NIC and NNC. *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$ and $P > 0.05$ represented as 'ns' (not significant).

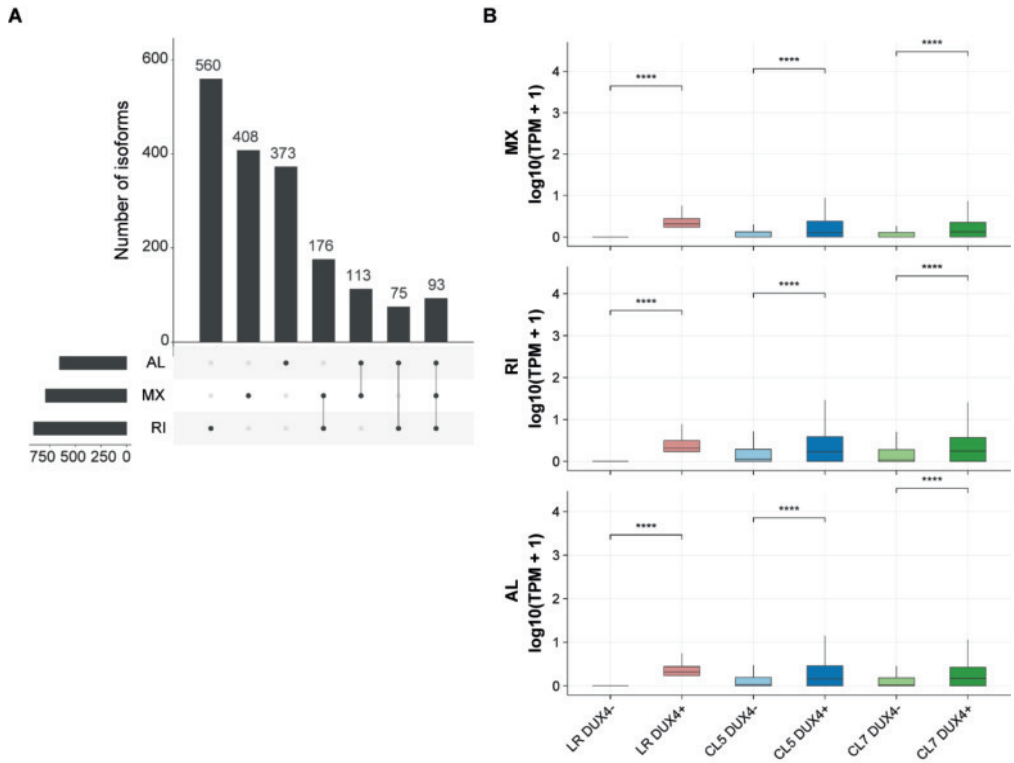

Fig. S7. Characteristics of isoforms with novel ASEs.

(A) UpSet plot showing the isoforms with novel AL, MX, and RI ASEs in DUX4+ transcriptome. (B) Box plots displaying the expression levels of isoforms with novel MX, RI, and AL ASEs per samples of LR and SR RNA-seq data. *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$.

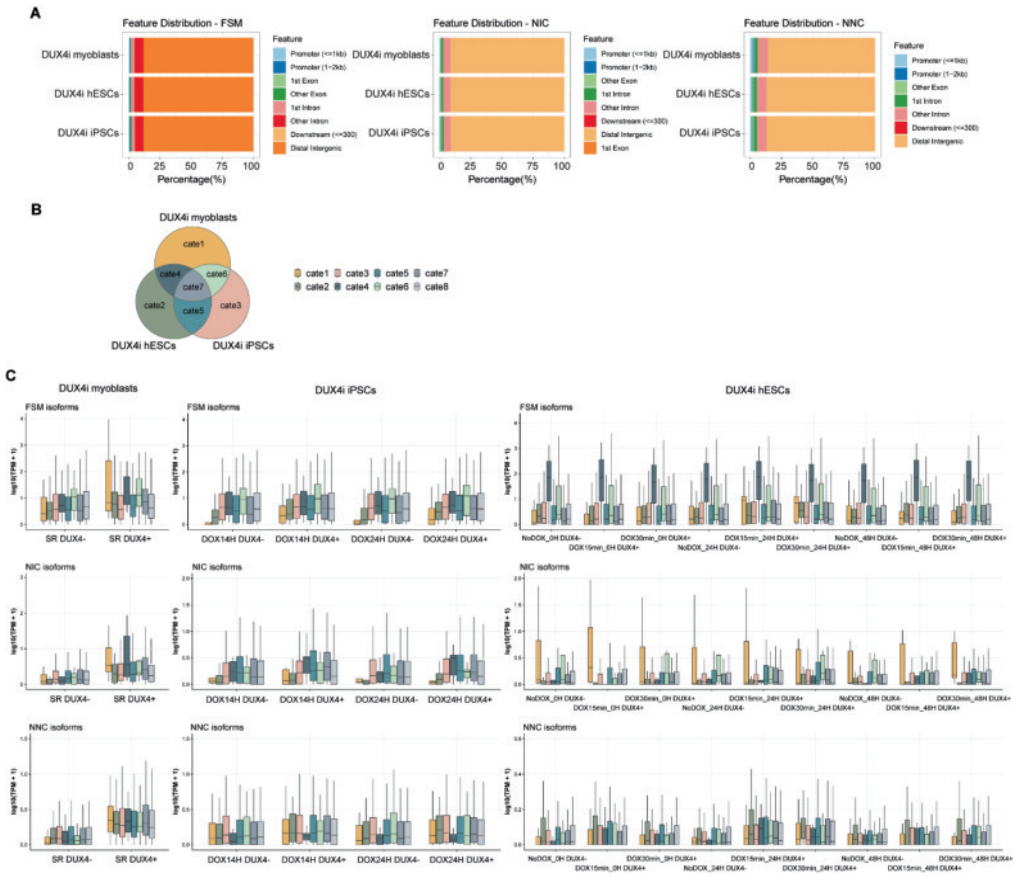

Fig. S8. Characteristics of FSM, NIC, and NNC isoforms with DUX4 peaks.

(A) Stacked bar plots showing the distribution of DUX4 peaks from three cell lines annotated to FSM, NIC, and NNC isoforms. Different colors represent distinct genomic regions. The x-axis indicates the percentage of peaks in each genomic region. (B) Venn diagram showing the overlap of isoforms bound by DUX4 across three cell lines. Color codes represent different categories. (C) Box plots showing the expression levels of isoforms in RNA-seq data from three cell lines. Isoforms are categorized based on the presence of DUX4 binding sites from different cell lines as shown in (B). Category 8 includes the isoforms without DUX4 peaks.

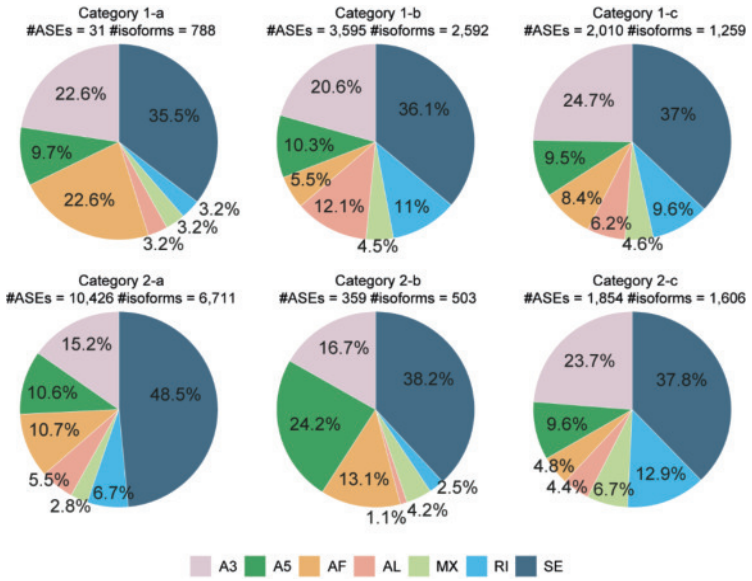

Fig. S9. Distribution of alternative splicing events across isoform subcategories.

Pie charts showing the proportion of different alternative splicing types in each sub-category. Color codes represent distinct alternative splicing events.

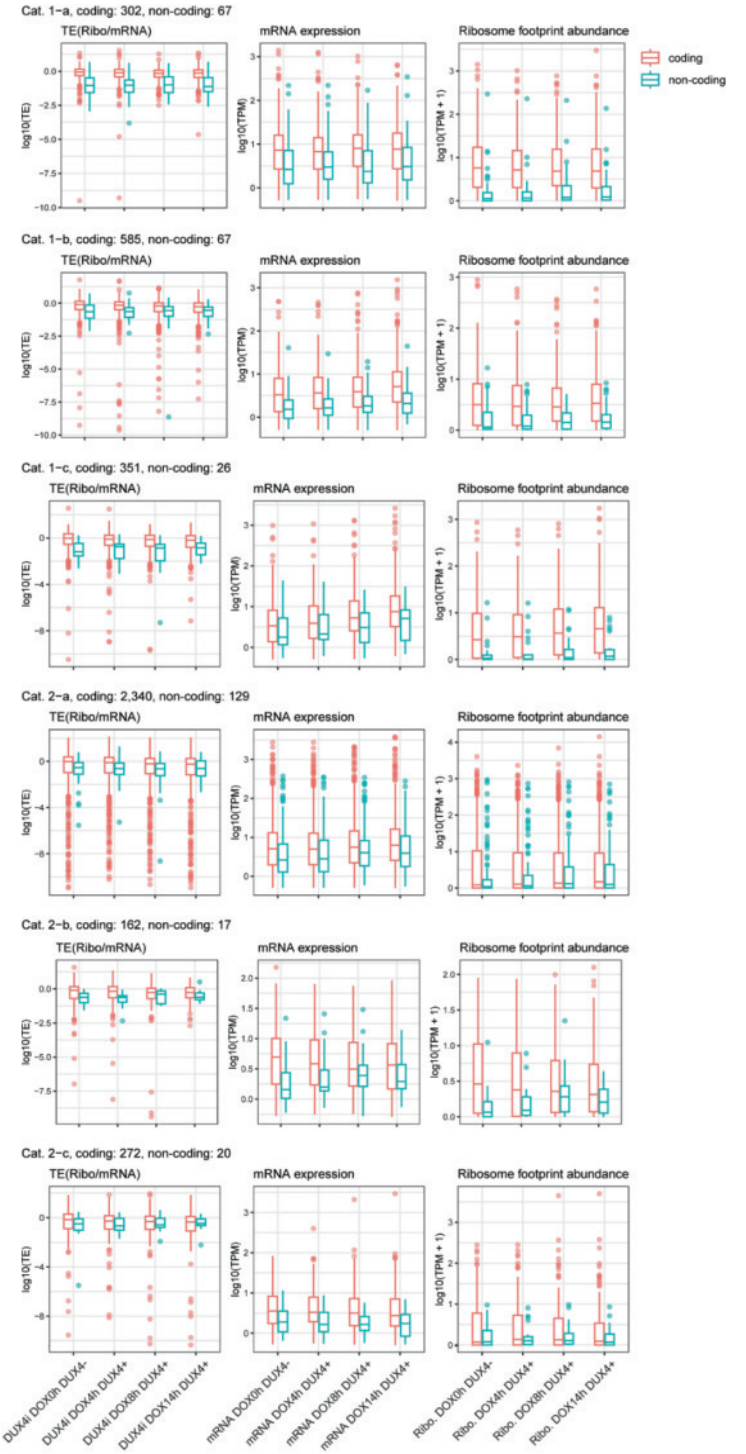
Cat. 1−a, coding: 302, non-coding: 67
Cat. 1−b, coding: 585, non-coding: 67
Cat. 1−c, coding: 351, non-coding: 26
Cat. 2−a, coding: 2,340, non-coding: 129
Cat. 2−b, coding: 162, non-coding: 17
Cat. 2−c, coding: 272, non-coding: 20
TE(Ribo/mRNA)
mRNA expression
Ribosome footprint abundance
log10(TE)
log10(TPM)
log10(TPM + 1)
coding
non-coding
DUX4- DOX0h
DUX4+ DOX4h
DUX4+ DOX8h
DUX4+ DOX14h

Fig. S10. Translational characteristics of isoform classified into different sub-categories.

Box plots showing translation efficiency, mRNA expression levels, and ribosome footprint abundance of coding and non-coding isoforms (mRNA TPM > 0.5) in each sub-category.

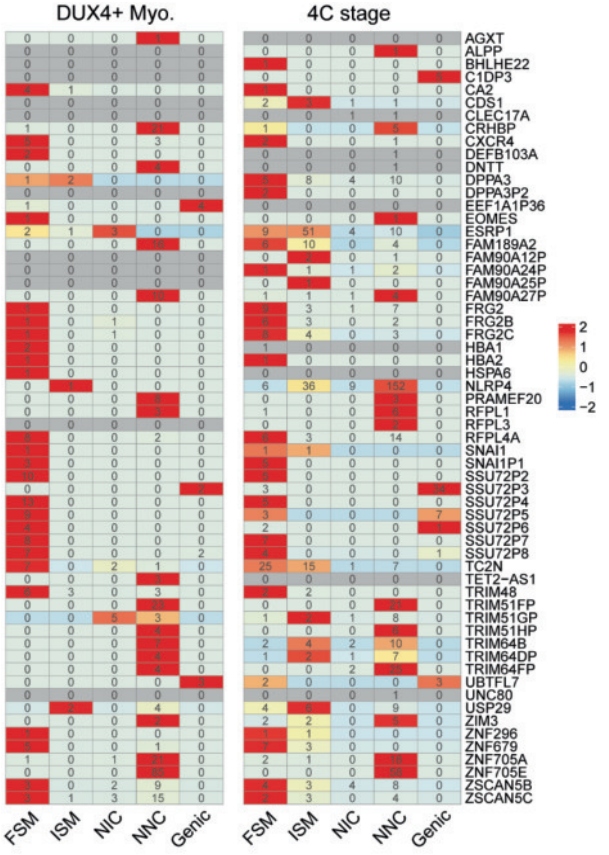

Fig. S11. Isoform usage and expression of genes from the set of 146 DUX4 target genes in DUX4i myoblasts and 4-cell stage.

Heatmaps showing the expression levels of DUX4 target genes from LR RNA-seq data obtained from DUX4i myoblasts (left) and 4-cell stage cells (right) (only keeping the detectable genes). Expression at the gene level is calculated by summing up the expression of isoforms from each gene. Color scales depict the expression level of each gene (Z-score): red represents a high expression level; blue represents a low expression level. The number labeled in each cell represents the count of isoforms classified into each structural category for each DUX4 target gene.

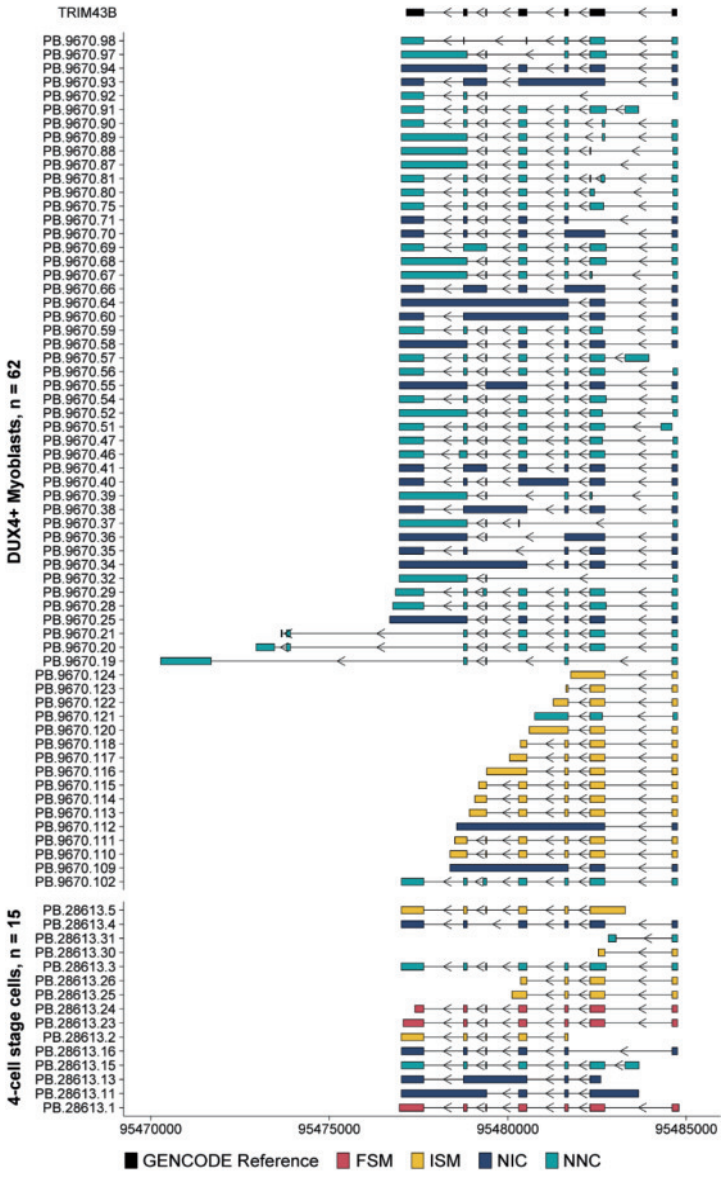

Fig. S12. Isoform usage of *TRIM43B* in DUX4i myoblasts and 4-cell stage cells.

Schematic overview showing that *TRIM43B* has a different isoform usage in DUX4+ myoblasts compared to 4-cell stage cells. Top; the *TRIM43B* transcript from GENCODE reference transcriptome v39; Middle; all transcripts of *TRIM43B* identified from the full-length transcriptome of DUX4+ myoblasts; Bottom; all transcripts of *TRIM43B* identified from the full-length transcriptome of human 4-cell stage cells. The plots are colored for structural categories. NNC and FSM isoforms contribute most to the gene expression of *TRIM43B* in DUX4+ myoblasts and 4-cell stage cells, respectively.

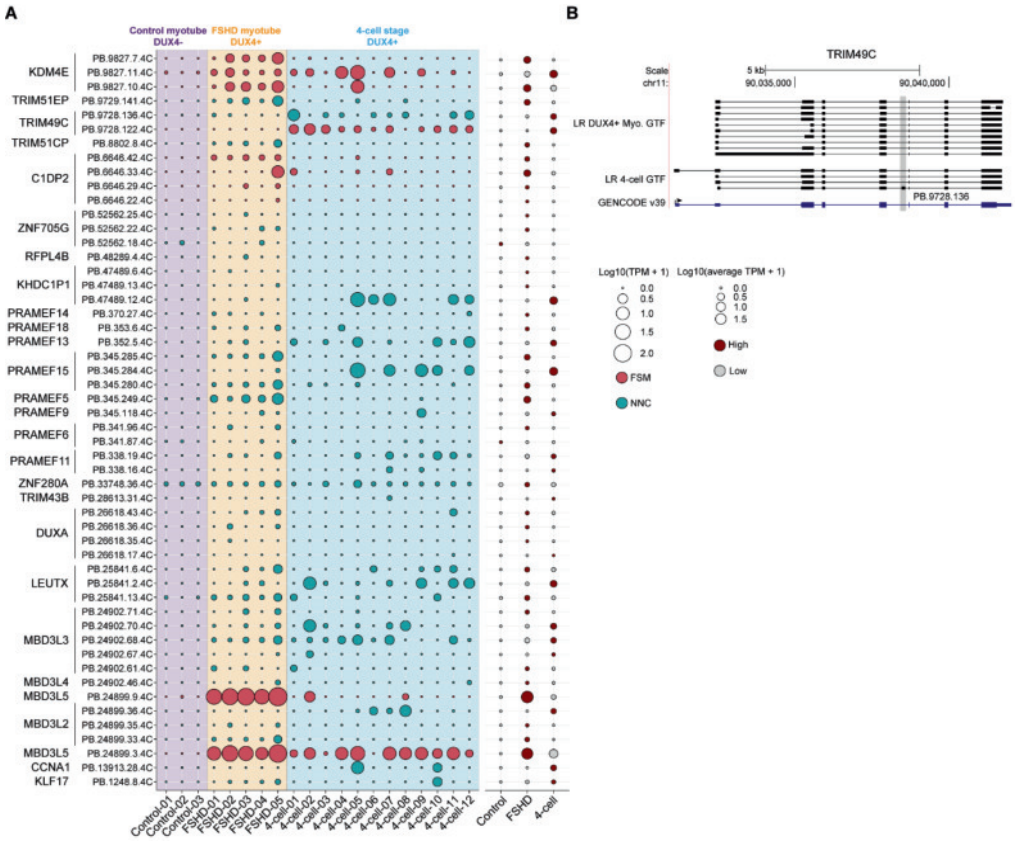

Fig. S13. Identification of cell type-specific isoforms of DUX4 target genes.

(A) Dot plot depicting the expression levels of isoforms of DUX4 target genes identified from 4-cell stage cells in RNA-seq of primary myotube culture and 4-cell stage cells. Color codes represent the structural category and the size of dots represents the expression level of each isoform in each sample. Dark red represents for which condition the isoform shows the highest average expression. (B) Plots showing the isoforms of DUX4 target genes: *TRIM49C* with an entirely novel exon (PB.9718.136) compared to the transcriptome of DUX4+ DUX4i myoblasts. The shadow highlights the target exon.

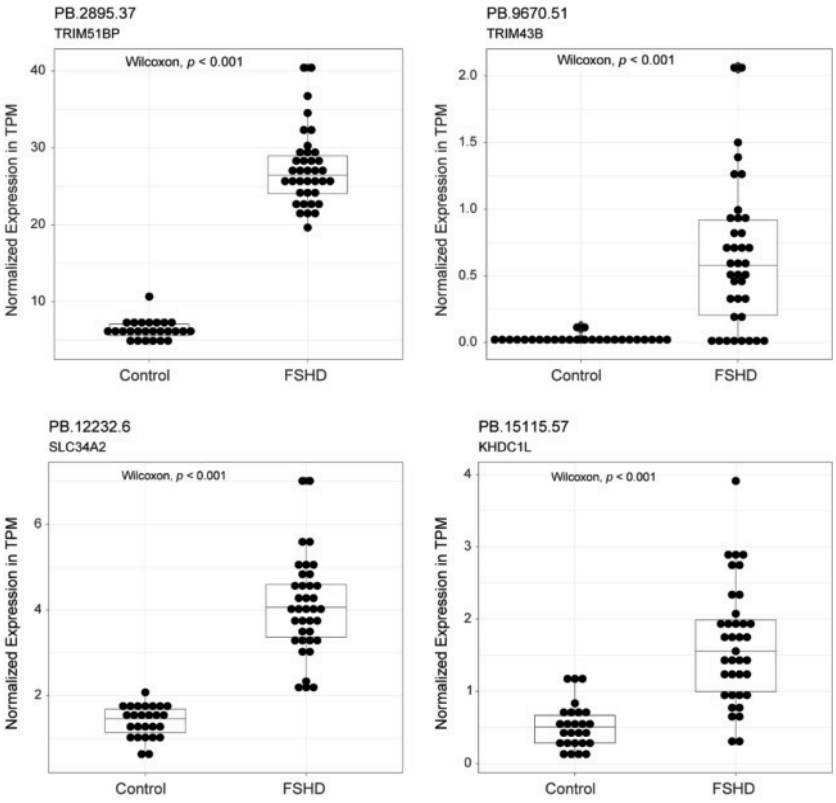

Fig. S14. Expression analysis of isoforms containing myogenic-specific exons in patient samples.

Box plots showing the expression levels of identified isoforms of DUX4 target genes in RNA-seq data from FSHD patients and healthy controls. *P*-values are calculated using unpaired Wilcoxon test.

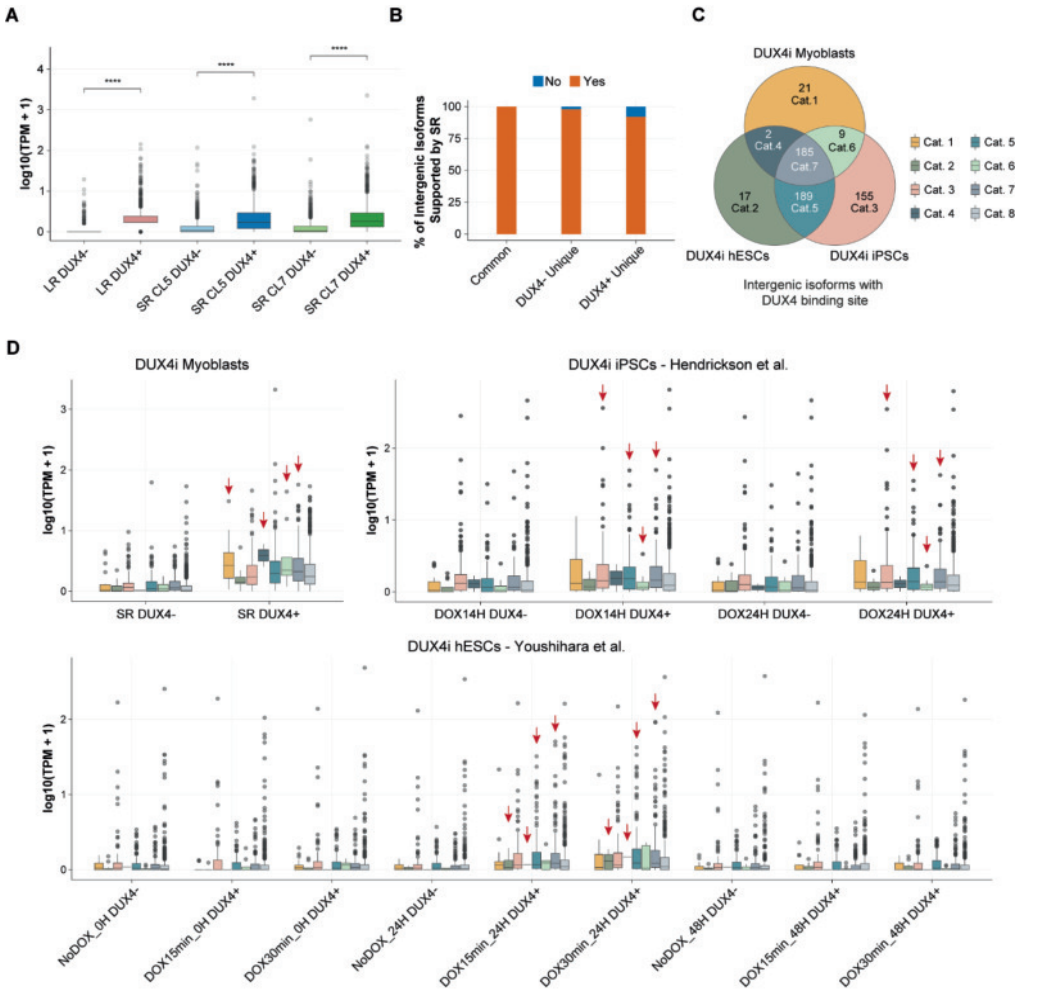

Fig. S15. Expression levels of intergenic isoforms with DUX4 binding site.

(A) Box plot showing the expression levels of intergenic isoforms in LR and SR RNA-seq datasets. *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$. (B) Stacked bar plot depicting the percentage of intergenic isoforms in each sub-type with support from SR RNA-seq data. (C) The same Venn diagram as Figure 5D with different color codes showing the categories of intergenic isoforms based on the DUX4 ChIP-seq data from different DUX4i cell lines. Category 1 to 7 include the intergenic isoforms with DUX4 binding sites and Category 8 includes the intergenic isoforms without DUX4 binding sites. (D) Box plots showing the expression levels of intergenic isoforms in the SR RNA-seq data of DUX4i myoblasts, iPSCs, and hESCs. Color codes represent each category of intergenic isoforms. The isoforms in Cat. 1, 4, 6, and 7 have DUX4 peaks in a myogenic context; the isoforms in Cat. 2, 4, 5, and 7 have DUX4 peaks in a hESC context; the isoforms in Cat. 3, 5, 6, and 7 have the DUX4 peaks in iPSCs. The red arrows highlight the categories with DUX4 peaks in the specific cell type.

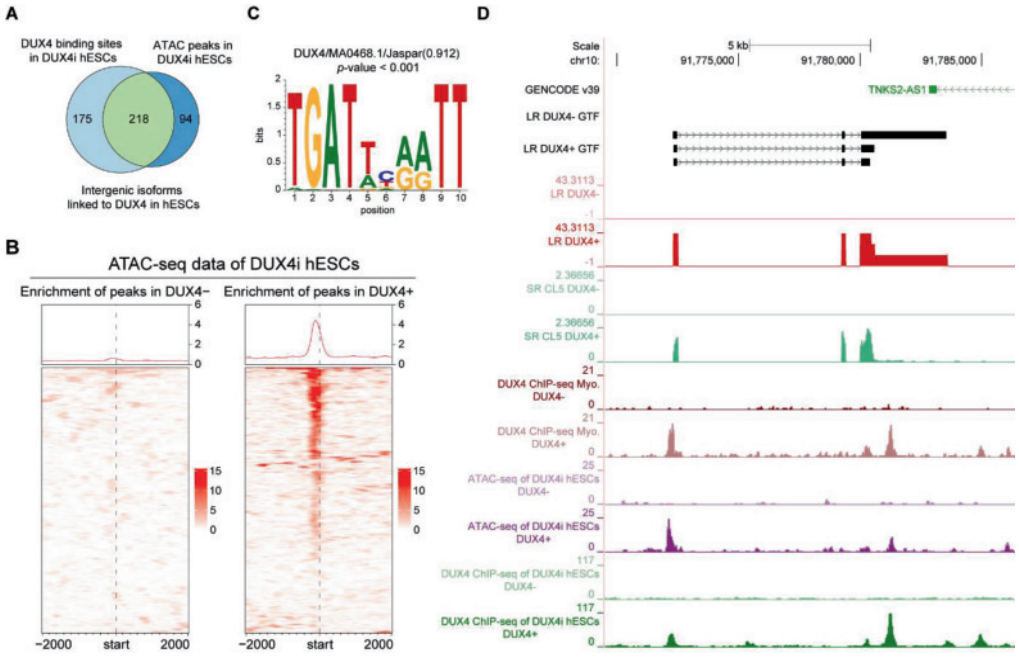

Fig. S16. Identification and classification of intergenic isoforms.

(A) Venn diagram showing the overlap of intergenic isoforms with DUX4 ChIP-seq peak or ATAC-seq peak in DUX4i hESCs. (B) Heatmaps exhibiting the signal density of ATAC-seq peaks near the TSS regions (±2 kb) of intergenic isoforms. (C) Motif plot showing the significant enrichment for the DUX4 binding motif underneath the ATAC peaks near the TSS regions of intergenic isoforms. (D) UCSC genome browser visualization of an example of the intergenic locus with DUX4 binding site and ATAC-seq peak in DUX4i hESCs. Color codes indicate the data in each track: black for transcriptomes of DUX4-, DUX4+, and GENCODE reference; bright red for LR RNA-seq data; blue for SR RNA-seq data; dark red for DUX4 ChIP-seq data from DUX4i myoblasts; purple for ATAC-seq data from DUX4i hESCs; dark green for DUX4 ChIP-seq data of DUX4i hESCs.

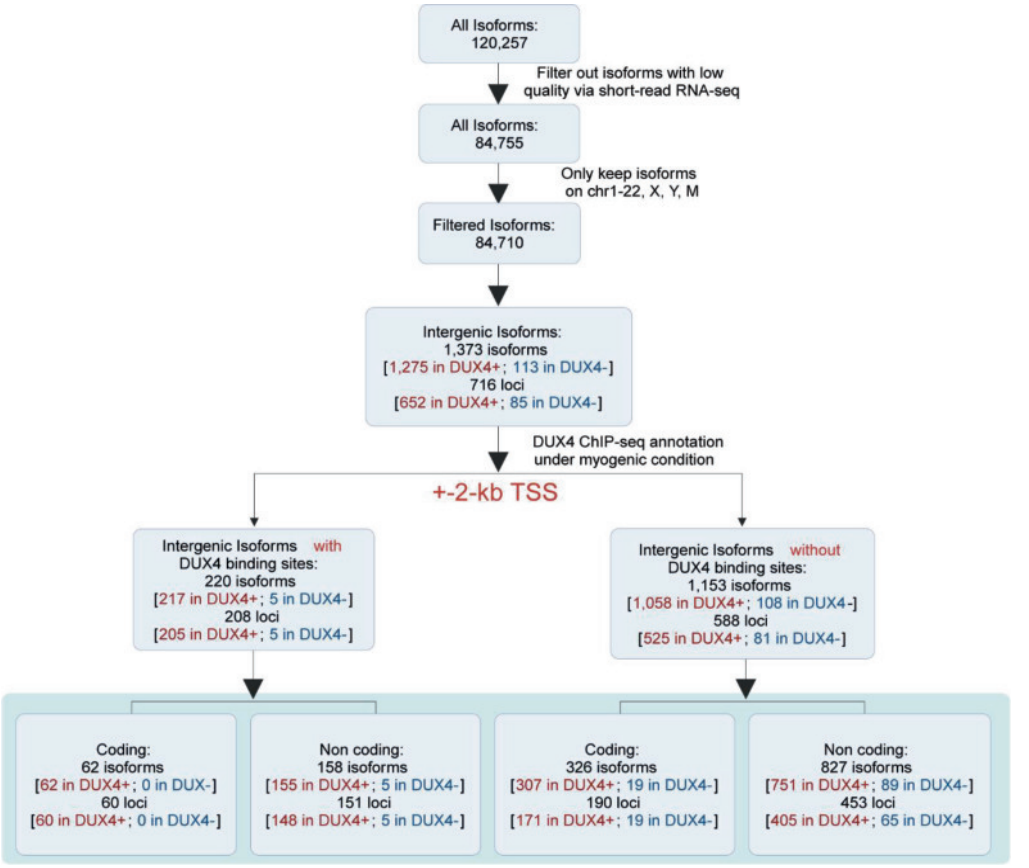

Fig. S17. Identification and classification of intergenic isoforms.

Schematic of workflow for classification of intergenic isoforms. All intergenic isoforms were classified into four sub-categories according to their coding potential and the presence of one or more DUX4 binding sites based on the DUX4 ChIP-seq of DUX4i myoblasts.

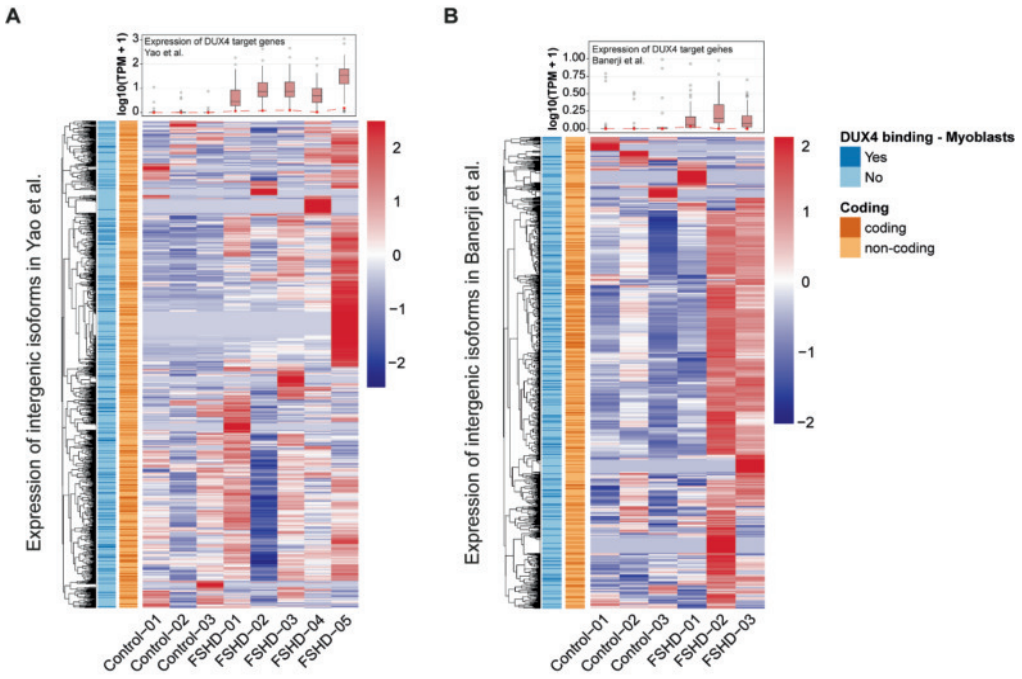

Fig. S18. Validation of intergenic isoforms in primary myotubes.

(A) Heatmap showing the expression levels of intergenic isoforms in published RNA-seq data sets obtained from primary myotube cultures derived from FSHD patients and healthy donors [Yao et al. (27)]. Color scale represents the expression levels that were row-normalized to the Z-score. Box plot on the top depicting the expression levels of DUX4 target genes in each primary myotube RNA-seq data set [Yao et al. (27)]. Red dot represents the expression levels of DUX4 in each sample. (B) Heatmap showing the expression levels of intergenic isoforms in published RNA-seq data obtained from primary myotube cultures derived from FSHD patients and healthy donors [Banerji et al. (42)]. Color scale represents the expression levels that were row-normalized to the Z-score. Box plot on the top depicting the expression levels of DUX4 target genes in each primary myotube RNA-seq data set [Banerji et al. (42)]. Red dot represents the expression level of DUX4 in each sample.

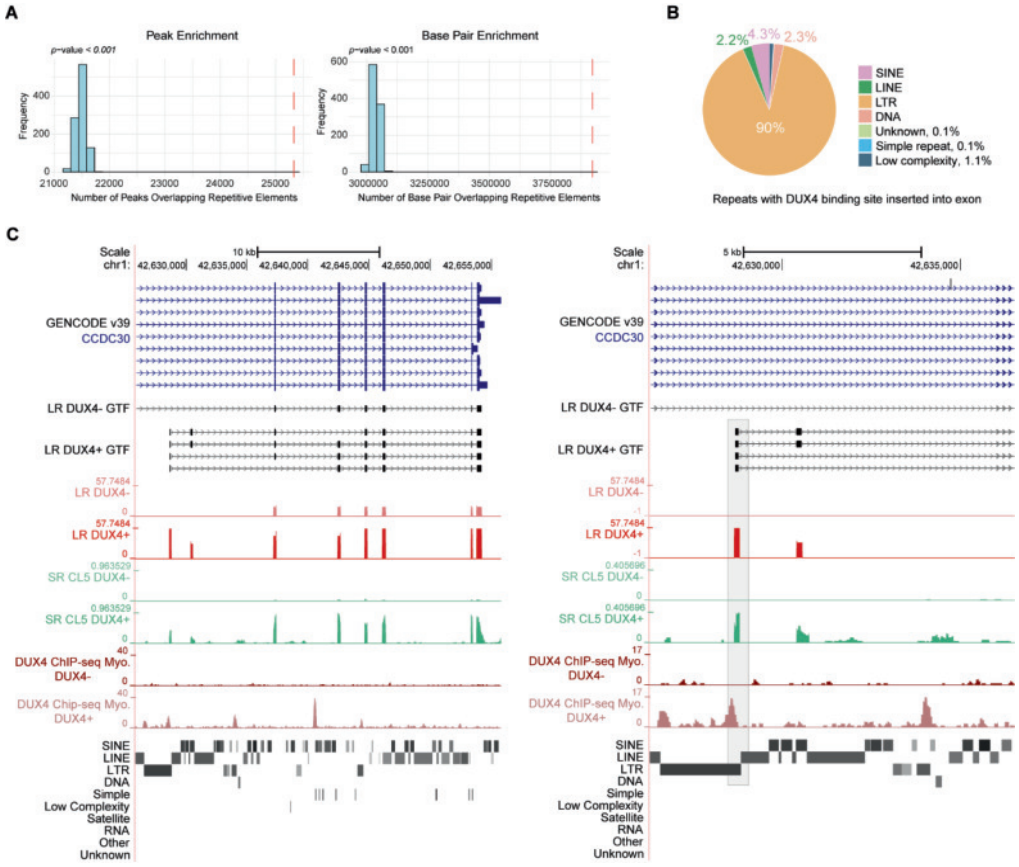

Fig. S19. Characterization of DUX4 binding patterns at repetitive elements and genic regions.

(A) Density plots showing the distribution of expected overlap between DUX4 peaks and repetitive elements from 1,000 permutation tests for (left) peak count and (right) base pair coverage. Red dashed lines indicate observed values, which are significantly higher than expected by chance (*P*-values < 0.001 for both metrics). (B) Pie chart showing the proportion of different repeat families that overlap with exons and contain DUX4 binding sites. Color codes represent different repeat families. (C) UCSC genome browser tracks showing CCDC30 containing isoforms with DUX4 peaks and incorporated LTR elements.

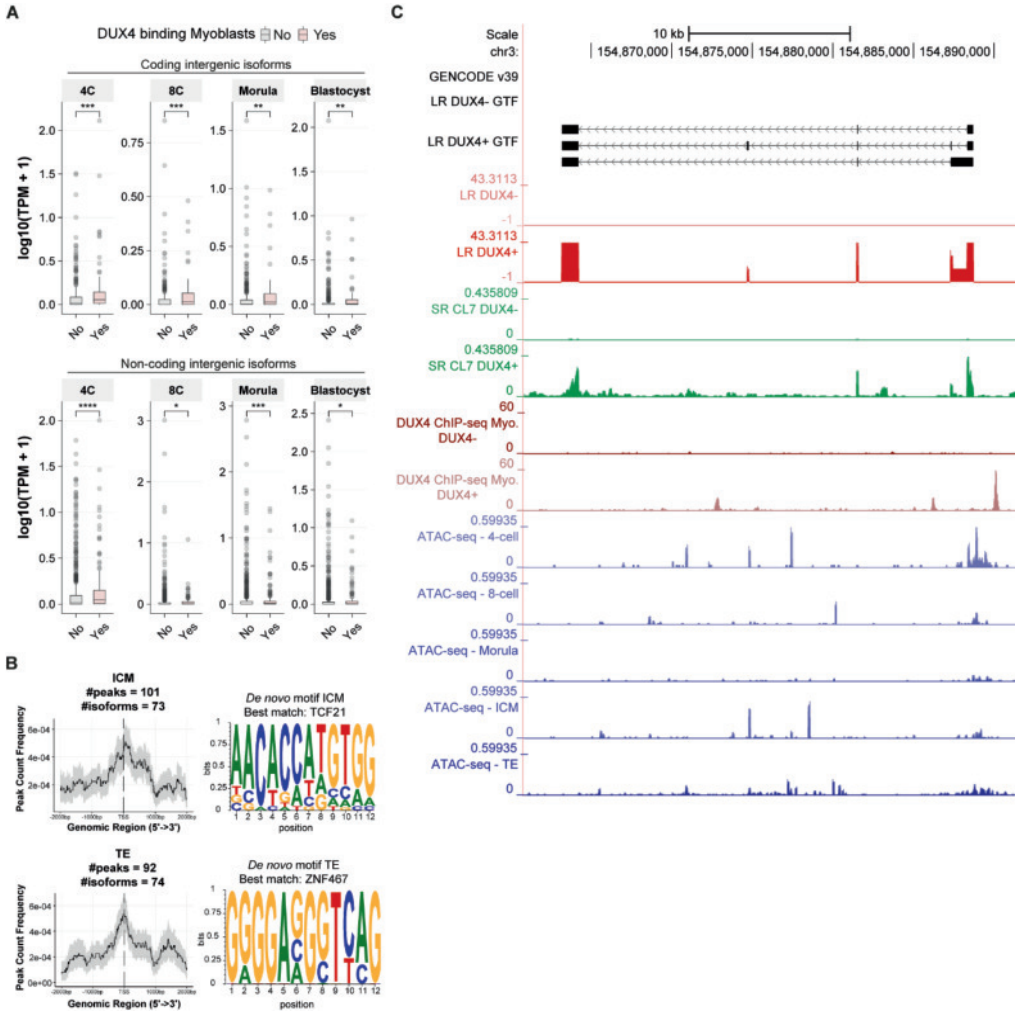

Fig. S20. Characterization of intergenic isoforms during early embryogenesis.

(A) Box plots showing the expression levels of intergenic isoforms, plotted by classification of intergenic isoforms, from the 4-cell stage to blastocysts. *P*-values are calculated using unpaired Wilcoxon test to assess the difference between the intergenic isoforms with versus without DUX4 peaks. Statistical significance is denoted as follows: ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$ and *$P < 0.05$. (B) Density plots showing the signal intensity of ATAC-seq peaks near the TSS regions of intergenic isoforms using the published ATAC-seq data obtained from ICM, and TE. Motif plots depict the significantly (*P*-value < 0.05) enriched motifs for ATAC-seq peaks from each developmental stage. (C) UCSC genome browser tracks showing the ATAC-seq peaks of one intergenic locus during early embryogenesis. The yellow and brown tracks represent the ATAC-seq data from 4-cell stage cells to TE.

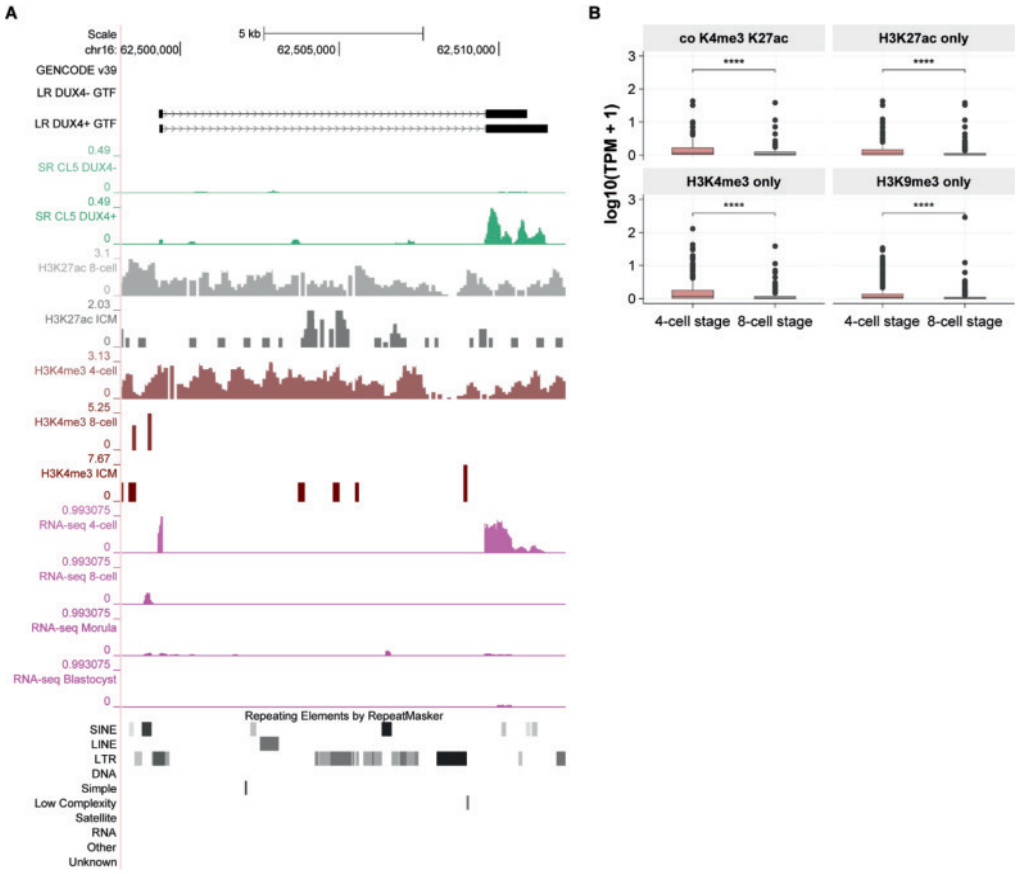

Fig. S21. Epigenetic landscape of intergenic isoforms during early embryogenesis.

(A) UCSC genome browser tracks showing the H3K27ac and H3K4me3 marks of one intergenic locus containing LTR and SINE elements. Grey color represents the H3K27ac CUT&RUN data of 8-cell stage cells and ICM. Dark red represents the H3K4me3 CUT&RUN data of 4-cell stage cells, 8-cell stage cells, and ICM. Purple color represents the RNA-seq of early embryonic cells. (B) Boxplots showing the expression levels of isoforms with different histone marks in 4C and 8C. *P*-values are calculated using unpaired Wilcoxon test. Statistical significance is denoted as follows: ****$P < 0.0001$.

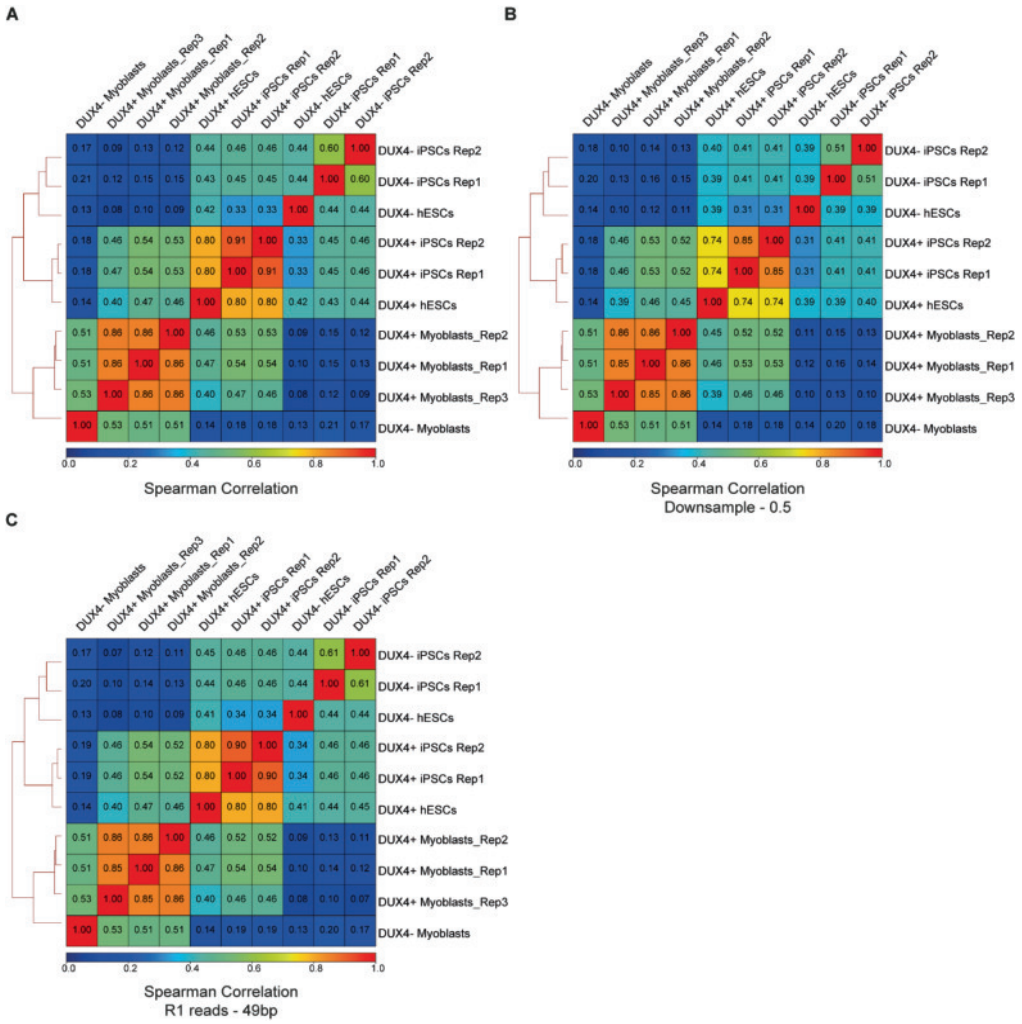

Fig. S22. Correlation analysis of DUX4 ChIP-seq data.

(A) Heatmap showing the Spearman correlations of read coverages genome-wide with bin size of 10 Kb between DUX4 ChIP-seq data of DUX4i myoblasts, hESCs, and iPSCs. Color codes represent the Spearman correlation: red is high and blue is low. (B) Heatmap showing the Spearman correlations of read coverages genome-wide with a bin size of 10 Kb between DUX4 ChIP-seq data of DUX4i myoblasts, hESCs, and iPSCs. The samples of DUX4i hESCs and iPSCs were scaled down by randomly extracting 50% of the reads. Color codes represent the Spearman correlation: red is high and blue is low. (C) Heatmap showing the Spearman correlations of read coverages genome-wide with a bin size of 10 kb between DUX4 ChIP-seq data of DUX4i myoblasts, hESCs, and iPSCs. The samples of DUX4i hESCs and iPSCs only used the R1 reads with the first 49 bp after trimming. Color codes represent the Spearman correlation: red is high and blue is low.

Supplementary tables

Table S1. Statistical overview for PacBio Iso-Seq

PacBio Iso-Seq		
	DUX4-	DUX4+
Nr of Circular Consensus Sequence (CCS) reads *Full-Length Non-artificial Concatemers*	3,517,410	3,832,832
#Total HiFi reads	85,953	124,177
#Uniquely mapped HiFi reads	85,847	124,032
Unique mapping rate (%)	99.9	99.9
#Non-redundant isoforms	57,071	85,165
Average length (bp)	2,200	2,114
Max. length (bp)	6,661	6,773

Table S2. Statistical overview for short-read RNA-seq data

short-read RNA-seq				
	CL5 DUX4-	CL5 DUX4+	CL7 DUX4-	CL7 DUX4+
#Total reads	58,191,313	26,077,073	50,049,837	46,720,311
#Uniquely mapped reads	52,854,467	21,344,087	45,430,605	40,529,166
Mapping rate (%)	90.8	81.9	90.8	86.8
Average mapped length	298.89	286.3	298.63	298.8

Table S3. SQANTI3 report – before and after rules filter

SQANTI3 report				Final transcriptome		
		Before rules filter	After rules filter	Aggregated transcriptome	DUX4- (FL.0h > 0)	DUX4+ (FL.16h > 0)
Summary	Unique Genes	25,157	13,586	13,541	9,438	11,921
	Unique Isoforms	120,257	84,755	84,710	45,304	58,820
Gene Classification	Annotated Genes	107,280	82,263	82,263	45,083	56,569
	Novel Genes	12,977	2,492	2,447	221	2,251
Isoform Classification	FSM	33,669	30,193	30,193	23,145	20,220
	ISM	11,234	7,043	7,043	2,957	5,149
	NIC	26,466	21,150	21,150	11,642	13,116
	NNC	33,009	22,499	22,499	7,001	16,975
	Genic Genomic	1,492	343	343	43	312
	Antisense	2,511	1,074	1,074	108	976
	Fusion	1,410	1,035	1,035	295	797
	Intergenic	10,465	1,418	1,373	113	1,275
	Genic Intron	1	0	0	0	0

* Annotated genes include FSM, ISM, NIC, and NNC isoforms.

* Novel genes include Genic Genomic, Antisense, Fusion, Intergenic, and Genic Intron isoforms.

Table S7. Statistical overview for NMD-targeted isoforms classified in NIC and NNC

	DUX4- unique	DUX4+ unique	common
#isoforms	1,688	4,821	521
#corresponding genes	1,025	1,964	435
#with hit in UniProt	1,609	4,635	500
#novel ORFs (identity score < 99%)	121	536	22
%novel ORF	7.5%	11.6%	4.4%

Table S9. Isoform usage in DUX4- and DUX4+

	subtypes	DUX4-	DUX4+	num. genes	DUX4 target genes 67	DUX4 target genes 146
Category 1: genes only expressed in DUX4+ sample	Cat. 1-a	no expression	normal	616	1	14
	Cat. 1-b	no expression	novel	1,139	20	12
	Cat. 1-c	no expression	normal + novel	261	13	8
Category 2: genes expressed in both samples	Cat. 2-a	normal	normal + novel	1,031	5	3
	Cat. 2-b	normal	novel	140	0	0
	Cat. 2-c	normal + novel	novel	190	2	0

*DUX4 target genes 146 was the DUX4 target genes 213 excluding DUX4 target genes 67.

Additional Table Captions

Table S4: Results of DE analysis

Table S5: Enriched BP GO terms of genes associated with upregulated and downregulated isoforms

Table S6: Enriched BP GO terms of genes associated with NMD-targeted isoforms

Table S8: Enriched GO BP terms of genes associated with isoforms with novel ASEs

Table S10: Genes with isoform usage shift in each category with expression level

Table S11: Enriched GO BP terms of genes in Category 1 and 2

Table S12: Characteristics of myogenic specific exons

Table S13: Genomic coordinates of intergenic loci

Table S14: Predicted coding intergenic isoforms with DUX4 peaks

Table S15: Full annotation of intergenic isoforms

Table S16: Annotation of DUX4 peaks and repetitive elements in genic region

Note: Tables S4-S6, S8 and S10-S16 are provided in the Auxiliary Supplementary Materials section.