Conservation and innovation in the DUX4-family gene network

Jennifer L Whiddon1,2, Ashlee T Langford1, Chao-Jen Wong1, Jun Wen Zhong1 & Stephen J Tapscott1

Facioscapulohumeral dystrophy (FSHD; MIM 158900, MIM 158901) is caused by misexpression of the DUX4 transcription factor in skeletal muscle1. Animal models of FSHD are hindered by incomplete knowledge regarding the conservation of the DUX4 transcriptional program in other species2–5. Despite the divergence of their binding motifs, both mouse DUX and human DUX4 in mouse and human muscle cells, respectively, activate genes associated with cleavage-stage embryos, including MERVL and ERVL–MaLR retrotansposons. We found that human DUX4 expressed in mouse cells maintained modest activation of cleavage-stage genes driven by conventional promoters but did not activate MERVL-promoted genes. Thus, the ancestral DUX4-regulated genes are characteristic of cleavage-stage embryos and are driven by conventional promoters, whereas divergence of the DUX4 and DUX homeodomains correlates with retrotansposon specificity. These results provide insight into how species balance conservation of a core transcriptional program with specificity. These results provide insight into how species balance conservation of a core transcriptional program with specificity. These results provide insight into how species balance conservation of a core transcriptional program with specificity.

Although the transcriptome of human cells expressing DUX4 is known6–9, the transcriptome of mouse cells expressing DUX4 has been largely unknown9. Both proteins are encoded by retrogenes derived by the retroposition of DUX mRNA9–11, and both proteins induce apoptosis when they are expressed in cultured human and mouse muscle cells12,13. Recent studies expressing DUX in human muscle cells12 or DUX4 in mouse cells4,13 have shown a partial overlap of regulated genes and a similar consensus binding site12; however, these two proteins have diverged significantly at the sequence level, including their homeodomains. Determination of the degree of similarity in their transcriptional programs might improve understanding of the rapid evolutionary divergence of Dux and DUX4, and inform the development of mouse models of FSHD, a disease that still lacks treatment options.

To compare the DUX transcriptome with the previously published DUX4 transcriptome in FSHD muscle cells, we generated RNA-seq and chromatin immunoprecipitation coupled to sequencing (ChIP-seq) data sets for DUX expressed in mouse skeletal muscle cells (Online Methods). We observed increased expression of 962 genes and decreased expression of 204 genes (Fig. 1a and Supplementary Tables 1 and 2). In these data, the most upregulated genes were normally expressed in mouse 2-cell (2C) embryos (for example, Zscan4a, Zscan4b, Zscan4c, Zscan4d, and Zscan4e; and Testv1 and Testv3)14–16; therefore, we used gene set enrichment analysis (GSEA) to compare our data to 2C-like embryonic stem cells17. The top of the DUX transcriptome was significantly enriched in the 2C-like gene signature (258/469 genes in the 2C-like gene signature contributed to the GSEA core enrichment; normalized enrichment score (NES) = 12.56, P < 0.001; Fig. 1b, Supplementary Table 3 and Supplementary Fig. 1). In addition, direct targets of DUX (i.e., genes whose RNA showed an increase in expression of fourfold or more and had a ChIP-seq peak within 1 kb of the annotated transcription start site (TSS)) were enriched in the 2C-like gene signature, on the basis of hypergeometric testing (60 direct targets in the 2C-like signature/189 total direct targets; 16-fold more direct targets in the 2C-like gene signature than the 3.7 genes expected by chance, P = 9.1 × 10−36), including Zscan4a, Zscan4b, Zscan4c, Zscan4d, Zscan4e, and Zscan4f; Testv1 and Testv3; Usp17lb and Usp17ld; Pramef25; and Zip352. We further confirmed that robust induction of both the Pramef25 and Zscan4c reporter constructs depended on intact DUX-binding sites (Supplementary Figs. 2a, 2b and 3a, b). ChIP-seq peaks at the TSSs of each of the five Zscan4-cluster genes supported the hypothesis that DUX directly binds and activates each Zscan4-cluster gene (Supplementary Fig. 3c–h).

Although there are two MERVL elements in the Zscan4 locus, we did not observe RNA-seq reads that spliced from these MERVLs to any Zscan4-encoding gene (Supplementary Fig. 3i–j). Importantly, the published 2C-like signature includes Dux itself, and Dux RNA is expressed in mouse embryonic stem cells (J.L.W., unpublished data). Gene ontology analysis also identified ‘embryo development’ among the significantly enriched terms (Supplementary Table 4). Together, these results demonstrated that DUX directly regulates many genes in the 2C-like transcriptome in myoblasts.

Despite the considerable sequence divergence in their two DNA-binding homeodomain regions (Fig. 1c), we found that DUX and DUX4 (ref. 18) activated orthologous genes in myoblasts of their respective species, including genes in the mouse 2C-like gene signature. For this analysis, we considered only genes with simple 1:1 mouse-to-human orthology, according to HomoloGene19. GSEA
Figure 1 DUX and DUX4 activate a gene signature specific to early cleavage-stage embryos in the muscle cells of their respective species. (a) DUX transcriptome in C2C12 mouse muscle cells. Red dots are genes with an absolute log2(fold change) ≥2 and adjusted P ≤0.05. Normalized counts were calculated by DESeq2 (normalized count = read count/size factor, with size factors estimated with the median-of-ratios method25). Control samples were uninduced cells of the same cell line. (b) GSEA of the 2C-like gene signature17. The x axis shows the log2(fold change)-ranked DUX transcriptome. The enrichment score increases when a gene in the DUX transcriptome is also in the 2C-like gene set, and a black vertical bar is drawn in the plot at bottom; the enrichment score decreases when a gene is not in the 2C-like gene set. The P value was empirically determined on the basis of 1,000 permutations of ranked gene lists. (c) Human DUX4, mouse DUX, and canine DUX homeodomain alignments (%), percentage amino acid identity; asterisk, four predicted DNA-contacting residues. (d) GSEA of the top 500 most upregulated genes in DUX4-expressing human cells. The x axis shows the log2(fold change)-ranked DUX4 transcriptome in mouse cells. This cross-species comparison required limiting both gene set and transcriptome to 1:1 mouse-to-human orthologs. The converse comparison is shown in Supplementary Figure 4a. (e) GSEA of the human orthologs of the mouse 2C-like gene signature. The x axis shows the log2(fold change)-ranked DUX4 transcriptome in human muscle cells. Both the gene set and transcriptome were limited to 1:1 mouse-to-human orthologs. The mouse 2C-like gene signature has 469 genes in total, 297 of which have simple 1:1 mouse-to-human orthology.

Figure 2 Despite binding-motif divergence and general transcriptome divergence, the DUX4 transcriptome in mouse muscle cells is enriched in the 2C-like gene signature. (a) DUX- and DUX4-binding motifs, as derived de novo from ChIP-seq peaks with the MEME algorithm. DUX4 ChIP–seq data were previously published6 but reanalyzed by using the methods of this study. The analysis showed divergence in the first half of the motif and conservation of the second half of the motif. E values listed reflect an estimate of the expected number of motifs, with the given motif’s log likelihood ratio (or higher) and with the same width and site count that would be expected in a similarly sized set of random sequences (in which each position in each sequence is independent, and letters are chosen according to the background letter frequencies). The histogram to the right shows that 578 peaks out of the 600 used to generate the DUX motif carry a match to the motif and that the motifs are centrally located within each ChIP–seq peak. A DUX4 histogram is also shown. (b) GSEA of the human 2C-like gene signature. The x axis shows the log2(fold change)-ranked DUX4 transcriptome in mouse cells. Because the mouse 2C-like gene signature and this DUX4 transcriptome were both identified in mouse cells, neither gene set nor transcriptome was limited to genes with 1:1 mouse-to-human orthology.
In this context, DUX4 showed the same binding motif as that in human cells (Supplementary Fig. 5a), increased expression of 582 genes and decreased expression of 428 genes (Supplementary Fig. 5b). Although DUX4 regulated many genes that were not orthologous to DUX-regulated genes and overall showed little similarity to the DUX transcriptome (Supplementary Fig. 5c), the genes that were upregulated in both the DUX and DUX4 transcriptomes were enriched in 2C-like genes, as assessed by hypergeometric testing ($P = 1.07 \times 10^{-11}$), and GSEA showed significant enrichment in the 2C-like gene signature activated by DUX4 in mouse cells (NES = 4.25, P ≤ 0.05; Fig. 2b and Supplementary Table 7). The activation of this signature, however, was not as robust as that of DUX-activated cell line; data previously published 18. (Fig. 3b) Example of a DUX ChIP–seq peak in MERVL (MT2 element in RepBase nomenclature). Track height is 200 repeatName, as defined by RepeatMasker. Red color indicates differential expression with absolute log$_2$(fold change) ≥ 1 and adjusted $P ≤ 0.05$. Numbers in parentheses are log$_2$(fold change). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. (b) As in a for DUX4-expressing human muscle cells compared with uninduced cells of the same cell line; data previously published 18. (d) Example of a DUX ChIP–seq peak in MERVL (MT2 element in RepBase nomenclature). Track height is 200 reads for all tracks. The mm10 genome location is chromosome (chr) 15: 52742953–52744319. (e) Luciferase assay comparing the activation of a 2C-active MERVL element reporter with DUX, DUX4 or an empty vector. The MERVL element contains a match to the DUX motif and was mutated, as shown in the schematic to the right. The full sequence is shown in Supplementary Figure 6d. Activation of the mutated MERVL reporter is also shown. Data shown are mean fold change over empty vector for three cell cultures prepared in parallel for each condition. luc, luciferase. Error bars, s.e.m. The nonmutated MERVL reporter activation experiment was repeated on three separate occasions and yielded consistent results. The mutated MERVL reporter experiment was performed on one occasion.

Figure 3 DUX, but not DUX4, activates transcription of repetitive elements characteristic of the early embryo in mouse muscle cells. (a) Expression levels of repeats during DUX expression in mouse cells compared with uninduced cells of the same cell line, organized by repeat class. Long terminal repeat (LTR) elements organized by family are shown in Supplementary Fig. 6a–c. Each dot is a repeatName, as defined by RepeatMasker. Red color indicates differential expression with absolute log$_2$(fold change) ≥ 1 and adjusted $P ≤ 0.05$. Numbers in parentheses are log$_2$(fold change). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. (b) As in a for DUX4-expressing human muscle cells compared with uninduced cells of the same cell line; data previously published 18. (c) As in a for DUX4-expressing human muscle cells compared with uninduced cells of the same cell line; data previously published 18. (d) Example of a DUX ChIP–seq peak in MERVL (MT2 element in RepBase nomenclature). Track height is 200 reads for all tracks. The mm10 genome location is chromosome (chr) 15: 52742953–52744319. (e) Luciferase assay comparing the activation of a 2C-active MERVL element reporter with DUX, DUX4 or an empty vector. The MERVL element contains a match to the DUX motif and was mutated, as shown in the schematic to the right. The full sequence is shown in Supplementary Figure 6d. Activation of the mutated MERVL reporter is also shown. Data shown are mean fold change over empty vector for three cell cultures prepared in parallel for each condition. luc, luciferase. Error bars, s.e.m. The nonmutated MERVL reporter activation experiment was repeated on three separate occasions and yielded consistent results. The mutated MERVL reporter experiment was performed on one occasion.
retroelement caused robust expression of the adjacent gene (Fig. 4d), in agreement with our previous finding that DUX4 expressed in human cells binds ERVL–MaLRs and uses them as alternative promoters. That DUX4 bound and activated the transcription of specific endogenous retrotransposon elements in the mouse genome that were not activated by DUX suggested that homeodomain divergence can selectively activate preexisting subsets of endogenous retrotransposons and induce the expression of adjacent genes.

The above results indicated that DUX and DUX4 have maintained the ability to regulate a set of 2C-like genes in mouse cells despite considerable divergence of their homeodomains; however, this conservation does not extend to the retrotransposons activated by each protein. We used chimeric proteins to identify the regions of DUX and DUX4 responsible for this partial conservation of function (Fig. 5a). The chimera with the DUX homeodomains and the DUX4 C terminus (MMH) matched the transcriptional activity of DUX (Fig. 5b–d), thus indicating that the transcriptional divergence between DUX and DUX4 mapped to the region containing the two homeodomains.

To determine the relative contribution of each homeodomain, we introduced each human homeodomain individually into DUX to create MHM and HMM chimeras (Fig. 5a). Neither MHM nor HMM activated transcription of MERVL-promoted genes (Fig. 5b). However, for 2C-like genes with conventional promoters, the individual DUX4 homeodomains showed different capacities to substitute for the corresponding DUX homeodomain; furthermore, MHM, as compared with HMM, consistently showed stronger activation of the target genes (Fig. 5c,d). We confirmed MHM and HMM expression and stability by using a reporter assay (Supplementary Fig. 9a). We also performed reciprocal experiments in human cells and again observed that the second homeodomains were more equivalent than the first homeodomains (Fig. 5e–f), thus indicating that the similarity of the second homeodomain was important to maintaining the functional conservation of the 2C-like gene signature at conventional promoters.

To further explore the evolutionary conservation of the DUX4 family’s ability to activate an early embryonic gene signature, we assessed the canine DUXC gene. Both Dux and DUX4 are retroposed copies of an ancestral DUXc mRNA, and neither mice nor humans have retained DUXC9–11 (Fig. 1c). When expressed in mouse muscle cells, canine DUCX did not activate MERVL-promoted genes (Fig. 5b) but did activate transcription of 2C-like genes with conventional promoters (Fig. 5c–d), thus again indicating that the ancestral DUX4-like gene activated genes characteristic of early cleavage-stage embryos independently of retrotransposon-promoted genes.

Figure 4 DUX and DUX4 use different types of LTR elements as alternative promoters for protein-coding genes. (a) Histogram in which black bars are counts of genes in the 2C-like signature that are MERVL promoted and activated by the indicated factor; white bars are genes that were detected by RNA-seq but that were not upregulated compared with control samples; and gray bars are genes with no reads in RNA-seq. MERVL-promoted genes for this plot were determined on the basis of the presence of an MT2-type element overlapping the annotated TSS of a gene in the published 2C-like gene signature17. (b) Histogram showing the number of genes in the 2C-like signature for which the indicated factor bound a MERVL (MT2-type) element, on the basis of ChIP-seq data, and there was at least one RNA-seq read that connected the ChIP-seq peak range to an annotated exon in mouse muscle cells, termed ‘peak-associated genes’ (PAGs). A schematic of PAGs that overlap MERVLs is shown at right. Two examples of PAGs that start in MERVL (MT2-type) elements are shown in Supplementary Figure 7a,b. (c) LTR-family distribution of PAGs that overlap any LTRs (CHIP-seq peak in an LTR with at least one RNA-seq read that connects the CHIP-seq peak range to an annotated exon). Although DUX and DUX4 both have PAGs that start in ERVL–MaLRs, the ERVL–MaLRs are predominantly different (only 1/31 DUX4 PAGs in ERVL–MaLRs was also identified as a DUX PAG). (d) Two examples of DUX4 binding an LTR and inducing novel transcription. Gray box, LTR element. Track height in reads is shown in parentheses below each browser screenshot.
Our current study shows that DUX and DUX4, when expressed in muscle cells, activate genes associated with an early 2C-like program, in agreement with results from a recent study suggesting that DUX and DUX4 may regulate the 2C-like program in early embryos\(^5\). Despite the divergence of their homeodomains and binding sequences, these factors have maintained their ability to activate the 2C-like gene signature within their own species. However, they have diverged in their ability to activate subsets of retrotransposons, thus suggesting that they are under evolutionary pressure to maintain activation of endogenous genes and a subset of beneficial retrotransposon-driven genes, but to diverge away from the activation of retrotransposons driving deleterious genes. Genes regulated by all DUX4-family factors probably represent the core ancestral network, whereas retrotransposon-promoted genes probably contribute species-specific additions. Such comparisons are particularly relevant to FSHD, a disease for which non-primate animal models remain imperfect. The observation that the expression of both DUX4 and DUX leads to apoptosis in mouse muscle cells supports the use of DUX4 in mice as a model of FSHD\(^8,24\). The cellular toxicity exhibited by cross-species expression might be due to the few classes of genes robustly activated (such as members of the PRAME family), the aggregate action of the larger number of genes moderately activated (such as the 2C/cleavage-stage signature), or the ability of each factor to activate different classes of retrotransposons and repetitive elements in different species. Nonetheless, because the pathophysiologic mechanisms of FSHD remain poorly understood, our study suggests that homeodomain divergence might homeodomains in mouse muscle cells of various classes, defined below. Data show means for 3 separate cell cultures for each condition; error bars, s.e.m. The experiments in b and d were also repeated on three separate days and showed consistent results. The experiments in c were completed on one occasion. (b) 2C-like genes with MERVL promoters. (c) 2C-like genes with conventional promoters that are induced by DUX4 and DUX. (d) 2C-like genes with conventional promoters that are induced only by DUX. (e) Schematic of reciprocal set of chimeric proteins; HHM is the two DUX homeodomains and the DUX4 C terminus; MMH is DUX with homeodomain (HD) 2 from DUX4; HMM is DUX with HD1 from DUX4. (f) RT–qPCR data for DUX4-target genes in human rhabdomyosarcoma cells. Data show means for 3 separate cell cultures for each condition; error bars, s.e.m. These experiments were completed on one occasion.

Figure 5 Transcriptional divergence between DUX4 and DUX maps to the two DNA-binding homeodomains. (a) Schematic of chimeric proteins; MMH is the two DUX homeodomains and the DUX4 C terminus; MHM is DUX with homeodomain (HD) 2 from DUX4; HMM is DUX with HD1 from DUX4. (b–d) RT–qPCR data for 2C-like genes in mouse muscle cells of various classes, defined below. Data show means for 3 separate cell cultures for each condition; error bars, s.e.m. The experiments in b and d were also repeated on three separate days and showed consistent results. The experiments in c were completed on one occasion. (b) 2C-like genes with MERVL promoters. (c) 2C-like genes with conventional promoters that are induced by DUX4 and DUX. (d) 2C-like genes with conventional promoters that are induced only by DUX. (e) Schematic of reciprocal set of chimeric proteins; HHM is the two DUX homeodomains and the DUX4 C terminus; MMH is DUX4 with HD2 from DUX4; HMM is DUX with HD1 from DUX4. (f) RT–qPCR data for DUX4-target genes in human rhabdomyosarcoma cells. Data show means for 3 separate cell cultures for each condition; error bars, s.e.m. These experiments were completed on one occasion.

| URLs | bcl2fastq Conversion Software version 1.8.4, http://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html; PANTHER user manual, http://pantherdb.org/help/PANTHER_user_manual.pdf; Gene Expression Omnibus (GEO), https://www.ncbi.nlm.nih.gov/geo/. |

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper. 

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Shadle (Fred Hutchinson Cancer Research Center) for making and sharing the pCW57.1-DUX4 virus-like particles used to generate the DUX4-inducible C2C12 clonal cell lines; P. Hendrickson (Huntsman Cancer Institute) for 2C::Tomato Reporter (Addgene plasmid no. 40281), from which we cloned the MERVL used to create the MERVL reporter in this study; and D. Root (Broad Institute) for the gift of Addgene plasmid no. 41393. We thank G. Kardon for critical guidance in the project and also thank members of the research groups of S. Tapsott and H. Malik for helpful discussions and critical reading of the manuscript. This work was supported by NIH NIA/MS R01AR045203 (S.J.T.), National Science Foundation Graduate Research Fellowship Program DGE-1256082 (J.L.W.), University of Washington Interdisciplinary Training in Genome Sciences T32 HG00035 from the National Human Genome Research Institute (J.L.W.), and Friends of FSH Research (J.L.W. and S.J.T.).

AUTHOR CONTRIBUTIONS

J.L.W. and S.J.T. conceived and designed the study and wrote the manuscript, with the assistance and final approval of all authors. J.L.W. performed and interpreted RNA-seq, ChIP-seq, RT–qPCR, and luciferase experiments. A.T.L. performed RT–qPCR and luciferase experiments. C.-J.W. performed bioinformatic data analyses. J.W.Z. contributed to the creation and characterization of the MMH-chimera cell line.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
1. Tawil, R., van der Maarel, S.M. & Tapscott, S.J. Facioscapulohumeral dystrophy: the path to consensus on pathophysiology. Skelet. Muscle 4, 12 (2014).
2. Lek, A., Rahimov, F., Jones, P.L. & Kunkel, L.M. Emerging preclinical animal models for FSHD. Trends Mol. Med. 21, 295–306 (2015).
3. Wallace, L.M. et al. DUX4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. Ann. Neurol. 69, 540–552 (2011).
4. Krom, Y.D. et al. Intrinsic epigenetic regulation of the D4Z4 macrosatellite repeat in a transgenic mouse model for FSHD. PLoS Genet. 9, e1003415 (2013).
5. Dandapat, A. et al. Dominant lethal pathologies in male mice engineered to contain an X-linked DUX4 transgene. Cell Rep. 8, 1484–1496 (2014).
6. Geng, L.N. et al. DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. Dev. Cell 22, 38–51 (2012).
7. Young, J.M. et al. DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLoS Genet. 9, e1003947 (2013).
8. Bosnakovski, D., Daughters, R.S., Xu, Z., Slack, J.M. & Kyba, M. Biphasic myopathic phenotype of mouse DUX, an ORF within conserved FSHD-related repeats. PLoS One 4, e7003 (2009).
9. Leidenroth, A. et al. Evolution of DUX gene macrosatellites in placental mammals. Chromosoma 121, 489–497 (2012).
10. Leidenroth, A. & Hewitt, J.E. A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol. Biol. 10, 364 (2010).
11. Clapp, J. et al. Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. Am. J. Hum. Genet. 81, 264–279 (2007).
12. Eidahl, J.O. et al. Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. Hum. Mol. Genet. 25, 4577–4589 (2016).
13. Knopp, P. et al. DUX4 induces a transcriptome more characteristic of a less-differentiated cell state and inhibits myogenesis. J. Cell Sci. 129, 3816–3831 (2016).
14. Falco, G. et al. Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Dev. Biol. 307, 539–550 (2007).
15. Zhang, W. et al. Zfp206 regulates ES cell gene expression and differentiation. Nucleic Acids Res. 34, 4780–4790 (2006).
16. Macfarlan, T.S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature 487, 57–63 (2012).
17. Akiyama, T. et al. Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. DNA Res. 22, 307–318 (2015).
18. Jagannathan, S. et al. Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum. Mol. Genet. 25, 4419–4431 (2016).
19. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44, D7–D19 (2016).
20. Bailey, T.L. et al. MEME SUITE; tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009).
21. Noyes, M.B. et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133, 1277–1289 (2008).
22. Peaston, A.E. et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev. Cell 7, 597–606 (2004).
23. Hendrickson, P.G. et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat. Genet. http://dx.doi.org/10.1038/ng.3844 (2017).
24. Bosnakovski, D. et al. An isogenic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. EMBO J. 27, 2766–2779 (2008).
25. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).
ONLINE METHODS

General statistical methods. Standard statistical tests were used as described for each individual application. Three separate cell cultures for each condition were used for RNA-seq and RT–qPCR as indicated. The ChiP–seq studies were multiple singleton experiments using several antibodies that immunoprecipitated the same binding domain, as described. No statistical methods were used to predetermine sample size.

RNA sequencing (RNA-seq). C2C12 mouse myoblasts (ATCC CRL-1772) were grown in DMEM (Gibco/Life Technologies) supplemented with 10% FBS (Thermo Scientific) and 1% penicillin–streptomycin (Life Technologies). The cells were obtained from the American Type Culture Collection (ATCC) and passaged without losing the ability to differentiate into myotubes but were not routinely checked for mycoplasma. We cloned the Dux transgene into the pCW57.1 lentiviral vector, a gift from D. Root (Addgene plasmid no. 41393), which has a doxycycline-inducible promoter. Dux and DUX4 transgenes were codon-altered to decrease the overall CpG content, because doing so has been shown to enhance transgene expression of the inducible DUX4 vector18. To create monoclonal cell lines, we first transduced pCW57.1-Dux into 293T cells (ATCC CRL-3216), along with the packaging and envelope plasmids pMD2.G and psPAX2, by using Lipofectamine 2000 reagent (Thermo Fisher). Viral-like particles containing pCW57.1-DUX4 were a gift from S. Shadel and were prepared in a similar manner. C2C12 cells were plated at low density and transduced with lentivirus particles at a low multiplicity of infection (MOI <1) in the presence of polybrene. Cells were selected and maintained in 2.6 μg/ml puromycin. Individual clones were isolated with cloning cylinders approximately 7 d after transfection and were chosen for analysis on the basis of robust transgene expression after 2 μg/ml doxycycline treatment for 36 h (DUX4 and DUX) or 18 h (MMH).

Three separate cell cultures were prepared for each condition, and total RNA was extracted from whole cells with a NucleoSpin RNA kit (Machery-Nagel), according to the manufacturer’s instructions. Total RNA integrity was verified with an Agilent 2200 TapeStation (Agilent Technologies) and quantified with a Trinean DropSense96 spectrophotometer (Caliper Life Sciences). RNA-seq libraries were prepared from total RNA with a TrueSeq RNA Sample Prep v2 kit (Illumina) and a Sciclone NGSx Workstation (PerkinElmer). Library size distributions were validated with an Agilent 2200 TapeStation (Agilent Technologies). Additional library quality control, blending of pooled indexed libraries, and cluster optimization were performed with a Qubit 2.0 Fluorometer (Life Technologies–Invitrogen). RNA-seq libraries were pooled (14-plex) and clustered onto two flow-cell lanes. Sequencing was performed with an Illumina HiSeq 2500 sequencing instrument (100-bp read-length (SR100) sequencing). Image analysis and base calling were performed with Illumina’s Real Time Analysis version 1.18 software, and this was followed by ‘demultiplexing’ of indexed reads and generation of FASTQ files with Illumina’s bcl2fastq Conversion Software version 1.8.4.

RNA-seq data analysis. Most data analysis was accomplished by using R (development version 3.4.0) and Bioconductor (3.3.0). First, reads of low quality were filtered out. Then, remaining reads were aligned to the reference genome (mm10 assembly) with TopHat26 (version 2.1.0), which is a splice-aware aligner that utilizes Bowtie2 (version 2.2.3) during the alignment process. Reads were allowed to map to up to 20 locations. Reads overlapping known genes in the UCSC Genome Browser were counted with the summaizeOverlaps function in the GenomicAlignments package, and differential gene expression was determined with DESeq2, which calculated P values with the Wald test and adjusted P values for multiple testing with the procedure of Benjamini and Hochberg. DESeq2 estimates variance for each gene with the average expression level across all samples27. GSEA was performed with the GSEA-re-ranked module of the Broad Institute’s GenePattern28 algorithm. Specifically, we used 1,000 gene-list permutations to determine the P value and the classic scoring scheme29. Because we compared data against only one gene set (from Akiyama et al.17), we did not correct for multiple tests. GSEA plot interpretation is described in the legend of Figure 1b. The negative control is shown in Supplementary Figure 1. Gene Ontology (GO) analysis was done with the Gene List Analysis tool of the PANTHER Classification System30 (version 10.0), which calculated P values with the binomial statistic, as described in the PANTHER user manual. Repeat-element analysis was accomplished with an in-house R package named rmskStats (version 0.99.0), which counts reads falling completely within RepeatMasker–annotated repeat elements. To account for reads aligning to multiple repetitive genome positions, rmskStats adjusts the count of a read to the fraction of the number of reported alignments (the NH column). For example, a read that maps to five locations counts as 0.2 read at each location. With these count results along with the statistical significance calculated by DESeq2, rmskStats then applies hypergeometric tests to infer the enrichment or depletion of families and classes of repeat elements. Reads that supported repeats being used as alternative promoters or alternative first exons were identified, and activation scores were calculated as described previously7, with one difference: we retained reads that linked ChiP–seq peaks to annotated exons regardless of whether they spliced across an intron.

Chromatin immunoprecipitation coupled to sequencing (ChiP–seq). All ChiP–seq experiments were performed with the monoclonal cell lines described in the RNA-seq section above, with a doxycycline-inducible system and codon-altered transgenes. To determine DUX4-binding sites in the mouse genome, we compared DUX4-expressing cells, 24 h after induction, immunoprecipitated with a 50:50 mixture of MO488 and MO489 antibodies to DUX4 (previously described in Geng et al.31), and DUX4-expressing cells immunoprecipitated with an antibody to an HA tag (originally, cat. no. PRB-101C, Covance; currently cat. no. 923501, BioLegend), which was not present in these cells. We performed ChiP–seq for DUX with two complementary approaches. First, we immunoprecipitated DUX4 from DUX-expressing cells 24 h after induction, by using two commercially available antibodies to DUX to analyze samples from a doxycycline-inducible C2C12 clonal cell line prepared as described for RNA-seq (A-19, cat. no. sc-385089, and S-20, cat. no. sc-385090, Santa Cruz Biotechnology) and compared the results with a mock pulldown with anti-IgG (cat. no. 315-005-003, Jackson Immunoresearch). Validation of commercial antibodies is available on the manufacturers’ websites. Second, we created a monoclonal population of cells with the doxycycline-inducible vector expressing a chimeric protein that fused the codon-altered DUX homeodomains with the codon-altered DUX4 C terminus (MMH). The MMH chimera maintained the DNA-binding domain of DUX and the -terminal epitopes of DUX4, thus allowing us to use the same DUX4 antiserum to immunoprecipitate the MMH chimera and DUX4 (Supplementary Fig. 1a). MMH immunoprecipitation was done 18 h after induction. We confirmed that the MMH chimera retained the DUX DNA binding specificity by comparing the ChiP–seq peaks of the MMH chimera with those of DUX. Although the antibodies to DUX had a lower signal-to-noise ratio, and thus identified fewer peaks, the majority of the peaks identified by the antibody to DUX were a subset of the MMH-chimera-identified peaks (Supplementary Fig. 1b). ChiP–seq with the A-19 antibody to DUX resulted in 2,400 peaks, 97.5% of which overlapped a peak in the MMH-chimera data set (51,356 peaks). Similarly, ChiP–seq with a second antibody to DUX, S-20, resulted in 628 peaks, 96.7% of which overlapped with a peak in the MMH-chimera data set. Furthermore, the MEME motif-prediction algorithm predicted nearly identical motifs from A-19 peaks and MMH-chimera peaks (Supplementary Fig. 1c), and the Pearson coefficient between the MMH chimera and DUX transcriptomes was 0.7847 (Supplementary Fig. 1d). We therefore used the ChiP–seq data set from the MMH chimera for all the analyses described in the main text, because of the superior signal-to-noise ratio compared with that of the commercially available antiserum to DUX.

Cross-linked ChiP was performed similarly to previously described methods for other transcription factors31,32. Briefly, ~106 cells were fixed in 1% formaldehyde for 11 min, quenched with glycone, lysed, and then sonicated to generate DNA fragments of 150–600 bp. The soluble chromatin was diluted 1:10 and precleared with protein A/G beads for 2 h. The remaining chromatin was incubated with primary antibody overnight, and then protein A/G beads were added for an additional 2 h. The beads were washed and then de-cross-linked overnight.

ChiP samples for DUX- and DUX4-expressing cells were validated by RT–qPCR and then prepared for sequencing, per the NuGEN Ovation Ultralow library system protocol with direct-read barcodes. ChiP–seq libraries were prepared from immunoprecipitated samples with an Ovation Ultralow Library System kit (NuGEN Technologies). Library size distributions were validated.
with an Agilent 2200 TapeStation (Agilent Technologies). Additional library quality control, blending of pooled indexed libraries, and cluster optimization were performed with a Qubit 2.0 Fluorometer (Life Technologies–Invitrogen). ChIP–seq libraries were pooled (12-plex) and clustered onto two flow-cell lanes. Sequencing was performed with an Illumina HiSeq 2500 in rapid-run mode with single-read, 100-bp read-length (SR100) sequencing. DUX4 ChIP–seq was performed at a separate time from DUX ChIP–seq.

ChIP samples for MHH-expressing cells were validated by RT–qPCR and then prepared for sequencing per the NEBNext DNA Library Prep kit (NEB, E7370L) instructions. Adaptor-ligated DNA was then size-selected and purified with AMPure XP beads (Beckman Coulter, A63881). Libraries were quantified, pooled, and sequenced on an Illumina HiSeq 2500 instrument in 125-bp, paired-end mode.

**ChIP–seq data analysis.** Image analysis and base calling were performed with Illumina’s Real Time Analysis version 1.18 software, followed by ‘demultiplexing’ of indexed reads and generation of FASTQ files with Illumina’s bcl2fastq Conversion Software version 1.8.4. With R (development version 3.4.0) and Bioconductor (version 3.3.0), low-quality reads that contained at least one N in the sequence were filtered out, and the tails were trimmed after 2–5 nt had quality encoding less than 4 (phred score 20). Further filtering included elimination of reads with <36 nt. The retained reads were then aligned to mm10 DNA from mock pulldown samples, as described above for negative controls, and only peaks with q <0.01 were considered. MACS2 calculated q values from P values with the Benjamini–Hochberg procedure. De novo motif prediction was done with MEME-Chip 4.11.2 (refs. 20, 35,36), on the basis of the top 600 peaks identified by MACS, ranked by q value, under the expectation of zero or one motif occurrence per sequence and requiring motifs to be between 5 and 15 nt in length. The Find Individual Motif Occurrences (FIMO) component of the MEME-Chip suite was used to identify good matches to the top predicted binding motif for DUX4 and DUX genome wide.

**Transient transfection and dual luciferase assay.** Transient DNA transfections of C2C12 cells were performed with SuperFect (Qiagen) according to the manufacturer’s specifications. Briefly, 16,000 cells were seeded per well of a 24-well plate on the day before transfection, which was done with 1,020 ng total DNA and 5 μl SuperFect per well. Cells were cotransfected with a pcS2 expression vector carrying the effector construct indicated (500 ng/well), a pcS2 expression vector carrying Renilla luciferase (20 ng/well) and a pGL3-basic reporter vector (500 ng/well) carrying a test-promoter fragment upstream of the firefly luciferase gene. Cells were lysed 24 h after transfection in Passive Lysis Buffer (Promega). Luciferase activities were quantified with reagents from the Dual-Luciferase Reporter Assay System (Promega), according to the manufacturer’s instructions. Light emission was measured with a BioTek Synergy2 luminometer. Luciferase data are given as mean fold change over empty vector ± s.e.m. for three separate cell cultures per condition.

**Code availability.** Code supporting the findings of this study is available at https://github.com/TapscottLab/Dux4FamilyGeneNetwork/ and https://github.com/TapscottLab/rmskStats/.

**Data availability.** The data generated in this publication have been deposited in the Gene Expression Omnibus database and are accessible through GEO series accession number GSE87282. The RNA-seq data for DUX4-expressing human myoblasts were previously published and are accessible through GEO series accession number GSE85461. The ChIP–seq data for DUX4 in human myoblasts were previously published and are accessible through GEO series accession number GSE33838.

26. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
27. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
28. Reich, M. et al. GenePattern 2.0. Nat. Genet. 38, 500–501 (2006).
29. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005).
30. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res. 44, D136–D142 (2016).
31. Conery, M.L., Yao, Z., Zheng, J.W., Groudine, M. & Tapscott, S.J. Distinct activities of MyoD and MyoD indicate separate roles in skeletal muscle lineage specification and differentiation. Dev. Cell 36, 375–385 (2016).
32. Cao, Y. et al. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev. Cell 18, 662–674 (2010).
33. Liao, Y., Snyder, G.K. & Shi, W. Subread aligner: fast, accurate and scalable read mapping by seed-and-extend. Nucleic Acids Res. 41, e108 (2013).
34. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).
35. Ma, W., Noble, W.S. & Bailey, T.L. Motif-based analysis of large nucleotide data sets using MEME-Chip. Nat. Protoc. 9, 1428–1450 (2014).
36. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36 (1994).
37. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210 (2002).