Supplementary Material

Supplementary figure 1. PCR genotyping results on clonal ES cells and chimeric Ch7CDS mouse tails confirming target site insertion.

Supplementary figure 2. PCR genotyping results on clonal ES cells and chimeric Ai9CDS mouse tails confirming target gene insertion.

Supplementary figure 3. PCR genotyping results on clonal ES cells and chimeric TOPCDS mouse tails confirming target gene insertion.

Supplementary figure 4. UMAP and hierarchy tree of relevant clusters from OTTER classification of genetically-engineered mouse CIC::DUX4 sarcomas.

Supplementary table 1. List of all sgRNA sequences and primers used for genotyping.

Supplementary table 2. All scores from OTTER classification (> 0.1) of normal limb muscle, KP tumors, and CIC::DUX4 sarcomas from mice.

Supplementary table 3. List of 65 consensus CIC::DUX4 target genes in human CDS compiled from the literature.

Supplementary table 4. Coordinates for all 2,410 shared HA-CIC::DUX4 peaks with genomic annotation, paired gene expression data, and motif analysis.

Supplementary Methods

Derivation of transgenic animals

All animal experiments were approved by Duke University Animal Care and Use Committee, protocol number A014-22-01. Ch7CDS mice were generated using homology-independent CRISPR/Cas9-mediated targeted integration. A donor vector was designed to insert loxP sites and sequence from the human DUX4 TAD followed by an exogenous polyadenylation signal (bgh-PolyA) at the c-terminus of the endogenous Cic locus on chromosome 7 (ENSMUG00000005442). Suitable sgRNAs were identified (Supplementary table 1), validated in vitro, and then cloned into a CRISPR/Cas9 activator plasmid (Addgene #64073). G4 embryonic stem cells were co-transfected with the activator and donor plasmids and then selected in G418-containing media. Embryonic stem (ES) cell clones containing the conditional knock-in were identified by PCR and validated with Sanger sequencing before injecting into donor ICR mouse morulae. Morulae were transplanted into female nurse ICR mice for gestation and delivery of chimeric pups. Rosa26 Lox-STOP-Lox (LSL) Ai9-HA-CIC::DUX4 (Ai9CDS) mice were generated by subcloning an N-terminal 3x HA-tagged CIC::DUX4 fusion gene from Yoshimoto et al.[6] into a Rosa26 targeting construct (Addgene #21714). The sequence verified construct was then transfected into ES cells and selected in G418 media-containing. Clones containing the knock-in were identified by PCR, sequence validated, and then injected into donor ICR mouse morulae which were transplanted into female nurse ICR mice for gestation and delivery of chimeric pups. Rosa26 LSL TOPO-HA-CIC::DUX4 (TOPCDS) mice were generated by subcloning the LSL cassette from a Lox-STOP-Lox TOPO vector (Addgene #11584) into the Rosa26-Ai9-HA-CIC::DUX4 targeting construct. Transfection into ES cells, clonal selection, validation, and transplantation were carried out as above.

Genotyping

Genomic DNA (gDNA) was purified from ES cells and tail clips using the Quick-DNA Miniprep Kit (Zymo Research). Initial validation of target insertion was performed with primers designed to amplify across the 5' and 3' integration sites (Supplementary figure 1-3, Supplementary table 1). Several positive ES cell clones were selected (indicated by red font) and verified by Sanger sequencing prior to expansion and morulae aggregation. To look for recombination, gDNA was also purified from tumors and tumor-derived cell lines and primers were designed to amplify across the entire region between the loxP sites (Supplementary table 1). PCR was performed using TaKaRa LA Taq and optimized for amplicon size. PCR products from TOPCDS were gel purified, sanger sequenced, and analyzed using Geneious Prime.

Derivation of tumor cell lines

Mouse soft tissue sarcoma cell lines were generated as described previously[38]. Tumors were resected using aseptic technique from humanely euthanized animals. Tumor tissue was enzymatically and mechanically digested by serial pipetting, washed, and resuspended in sterile PBS. The cell suspension was filtered through a 70 μm cell strainer, pelleted, and resuspended in DMEM containing 10%FBS and 1X GA-1000 antibiotic (standard growth media). Cells were plated at high density in tissue culture treated flasks and assessed for cell death the following morning. Viable adherent cells were maintained in standard growth media and passaged with 0.25% trypsin-EDTA. Mycoplasma testing was preformed prior to sequencing experiments.

Western blot

Cells lines were maintained in standard growth media until 80% confluency. Using Trypsin-EDTA, the cells were lifted, collected in Hank's Balanced Salt Solution, and then pelleted by centrifugation (300xg for 3 minutes). Lysates were made using RIPA buffer (supplemented with 1% SDS, HALT protease inhibitor (Thermo Fisher Scientific), and Benzonase) and quantified

with Pierce BCA protein assay kits. Heat-denatured proteins were loaded onto a 10% Bis-Tris gel, run at 150v in 1x MES buffer, and then wet-transferred at 350mA for 1 hour at 4°C. All above steps were completed in a single day due to the unstable nature of the fusion protein. CIC::DUX4 (~260kD) was probed using an anti-HA antibody (Cell Signaling, 3724) at 1:1000 dilution and anti-DUX4 antibody (Abcam, ab124699) at 1:1000. Cre was detected using an anti-Cre antibody (Cell Signaling, 15036) at 1:1000 dilution with B-actin at 1:2000 (Sigma, A1978) as a loading control. Images were acquired on a LI-COR Odyssey CLx and processed using the Image Studio Software.

Immunohistochemistry

Tissue samples were fixed in 10% formalin/70% ethanol and embedded in standard paraffin blocks. Five micrometer sections were mounted and stained with hematoxylin and eosin (H&E) or antibodies. Chromogen amplification and detection of antibodies was performed using the Vectastain Elite ABC-HRP Kit and DAB Peroxidase Substrate Kit (Vector Labs). Antigen unmasking was performed with a citrate buffer pH 6.0 by a modified microwave retrieval method or boiling. Images were captured on a Leica inverted light microscope with DFC450 camera and processed using the Leica Application Suite. The following antibodies were used for immunohistochemistry: DUX4 (Thermo-Fisher, MA5-16147), HA (Cell Signaling, 3724), ETV4 (Proteintech 10684-1-AP), CD99 (Thermo-Fisher, bs-2523R), Pan-cytokeratin (abcam, ab9377), WT1 (Thermo-Fisher, PA5-116131), Sox10 (abcam, ab227680), Desmin (abcam, ab15200), and CD34 (Thermo-Fisher, 14-0341-82) at manufacturer recommended dilutions. All slides were reviewed by an expert sarcoma pathologist (DC) at Duke University. RNA-sequencing

RNA was extracted and purified from flash frozen tumors/normal tissues and cells using a Qiagen RNeasy kit following the manufacturer's instructions for fibrous tissue (Qiagen, Hilden,

Germany). High quality RNA (RIN >7) was divided into triplicate from which 150bp paired-end, rRNA-depleted, libraries were made using the Illumina TruSeq RNA library Prep Kit (Illumina, CA, USA). Libraries were quantified using the KAPA Library Quantification kit (KAPA Biosystems, MA, USA), multiplexed, clustered onto flowcells, and then sequenced using an Illumina HiSeq 4000 sequencer (or equivalent platform) by GENEWIZ (Azenta, NJ, USA). Raw sequencing reads were trimmed using Trimmomatic v0.39 (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3 MINLEN:36) and then aligned to the mm10 reference genome using default parameters in STAR v2.7.10a. FeatureCounts (Subread v2.0.3) was used to compile a count table from sorted and indexed BAM files which was loaded into DESeq2 to calculate differential expression. Deconvolution analysis was performed with SCDC v0.0.0.9000 (probability threshold= 0.7) using single cell RNA-seq reference datasets from Tabula Muris (https://doi.org/10.6084/m9.figshare.5821263.v3) manually curated into 12 cell ontology assignments as displayed in Figure 3d. To run OTTER, reads were aligned to GRCmm39 and translated to their human orthologs. To deal with gene drop out during ortholog conversion, missing genes were imputed with zeros. The resulting human genes matrices were inputted into OTTER. To increase model generalization for non-human samples, we strengthened OTTER's stochastic data augmentation procedure and introduced a hierarchical training procedure[39]. OTTER scores > 0.1 were reported (Supplementary table 2). Raw data is available on GEO under the accession number GSE241369.

ChIP-sequencing

ChIP-sequencing was performed in tumor-derived cell lines generated from Ai9CDS and TOPCDS mice using the Active Motif ChIP-IT High Sensitivity kit (ActiveMotif, CA, USA). Cells were seeded into 150mm dishes and cultured in standard growth medium. At 80% confluency, the cells were crosslinked in 37% formaldehyde (with methanol) for 15 minutes on

the dish and quenched with glycine. Washed cell pellets were manually lysed using a dounce homogenizer and the chromatin was fragmented using a Q125 sonicator (Qsonica, CT, USA) with the following settings: 25% amplitude, 30 seconds 'ON'/30 seconds 'OFF' for 20 minutes total. Separate immunoprecipitation reactions using an Anti-HA tag antibody (Abcam; ab9110) or anti-Histone H3K27ac antibody (ActiveMotif, 39134) were setup for overnight immunoprecipitation reactions at 4°C. ChIP DNA was bound to Protein G agarose beads, column purified, and eluted. 150bp, paired-end, DNA libraries for sequencing were prepared using the TruSeq DNA library Prep Kit (Illumina, CA, USA) and quantified using the KAPA Library Quantification kit (KAPA Biosystems, MA, USA). Libraries were sequenced on an Illumina HiSeq 4000 sequencer (or equivalent platform) by GENEWIZ (Azenta, NJ, USA). Raw sequencing reads were trimmed using Trimmomatic v0.39 and then aligned to the mm10 reference genome using Bowtie 2 (-q -t --no-mixed --no-discordant). Duplicate reads were marked and removed using Picard tools v2.18.2 and peaks were called with MACS3 (-B -f BAMPE -g 1.87e9 -q 0.01). Peak files were filtered against the ENCODE blacklisted regions (https://github.com/Boyle-Lab/Blacklist/tree/master/lists) and then annotated using ChIPseeker v3.17. De novo motif enrichment analysis on HA-CIC::DUX4 peaks was performed with HOMER v4.11. Super Enhancers were called from H3K27ac peaks using ROSE and used as input for CRCmapper to map the core regulatory circuitry[40]. Raw data is available on GEO under the accession number GSE241370.

Supplementary References

38. Oristian KM, Crose LES, Kuprasertkul N, Bentley RC, Lin Y-T, Williams N *et al.* Loss of MST/Hippo Signaling in a Genetically Engineered Mouse Model of Fusion-Positive Rhabdomyosarcoma Accelerates Tumorigenesis. *Cancer Res* 2018; **78**: 5513–20.

39. Mayouf M-S, Dupin De Saint-Cyr F. GH-CNN: A New CNN for Coherent Hierarchical Classification. In: Pimenidis E, Angelov P, Jayne C, Papaleonidas A, Aydin M (eds).

Artificial Neural Networks and Machine Learning – ICANN 2022. Springer Nature Switzerland: Cham, pp 669–681. (2022).

40. Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI *et al.* Models of human core transcriptional regulatory circuitries. *Genome Res* 2016; **26**: 385–96.

Supplementary Discussion

This hypothesis is supported by the high coincidence of a consensus ETS motif (GGAA) at CIC::DUX4 binding sites. Interestingly, although the most highly upregulated downstream targets of CIC::DUX4 are associated with peaks containing both CATT and GGAA motifs, genes associated with GGAA-only motifs are, on average, less highly upregulated in CDS perhaps signifying context-specific interactions. The role of ETV4 in CDS has been investigated before but the results are inconsistent. For example, in a transgenic zebrafish model of CDS, genetic loss of ETV4 impairs tumor formation[28]. In contrast, ETV4 knockdown in transformed CIC::DUX4-expressing NIH 3T3 cells had no effect on cell viability and tumor growth but did impair metastatic potential[25]. One explanation for the discrepancy may be a divergence in PEA3 subfamily (ETS transcription factors ETV1, ETV4, and ETV5) redundancy during evolution from zebrafish to human[33,34]. Another possibility relates to the species-specific distribution of ETS binding sites which has forestalled numerous attempts to model Ewing Sarcoma in mice[35]. To dissect the independent and overlapping roles of ETS1 and PEA3 subfamily genes in CDS, future work will use CRISPR/Cas9 to systematically and combinatorially delete these genes in cell lines and tumors. Of further interest is whether forced overexpression or stabilization of these same factors could be toxic to CDS cells in keeping with the 'Goldilocks phenomenon' of ETS family transcription factors described in Ewing Sarcoma[36,37].