

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

**A family of double-homeodomain transcription factors
regulates zygotic genome activation in placental mammals**

Alberto De Iaco, Evarist Planet, Andrea Coluccio, Sonia Verp, Julien Duc, and
Didier Trono*

School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL),
1015 Lausanne, Switzerland

*Correspondence to:

Didier Trono
Ecole Polytechnique Fédérale de Lausanne (EPFL)
School of Life Sciences
SV-LVG Station 19, CH-1015 Lausanne, Switzerland

Phone: +41 (0)21 693 1761

Email: didier.trono@epfl.ch

27 **In metazoan embryos, transcription is mostly silent for a few cell**
28 **divisions, until release of a first major wave of embryonic transcripts by**
29 **so-called zygotic genome activation (ZGA) ¹. Maternally provided ZGA-**
30 **triggering factors have been identified in *Drosophila melanogaster* and**
31 ***Danio rerio* ^{2,3}, but their mammalian homologues are still undefined.**
32 **Here, we reveal that the DUX family of transcription factors ^{4,5} is**
33 **essential to this process in human and mouse. First, human *DUX4* and**
34 **murine *Dux* are both expressed prior to ZGA in their respective species.**
35 **Second, both orthologues bind the promoters and activate the**
36 **transcription of ZGA genes. Third, *Dux* knockout in mouse embryonic**
37 **stem cells (mESCs) prevents their cycling through a 2-cell-like state.**
38 **Finally, zygotic depletion of *Dux* leads to impaired early embryonic**
39 **development and defective ZGA. We conclude that DUX proteins are key**
40 **inducers of zygotic genome activation in placental mammals.**

41 *Dux* genes encode for double-homeodomain proteins and are conserved
42 throughout placental mammals ^{4,5}. Human *DUX4*, the intronless product of an
43 ancestral *DUXC*, is nested within the D4Z4 macrosatellite repeat of
44 chromosome 4 as an array of 10 to 100 units ⁶. *DUX4*, *DUXC*, and *Dux* genes
45 from other placental mammals display the same repetitive structure, with
46 *DUX4* from primates and *Afrotheria* and *DUXC* from cow and other
47 *Laurasiatheria* localizing at telomeric or pericentromeric regions, and murine
48 *Dux* tandem repeats lying adjacent to a mouse-specific chromosomal fusion
49 point that resembles a subtelomeric structure ^{4,5}.

50 Overexpression-inducing mutations in *DUX4* are associated with facio-
51 scapulo-humeral dystrophy (FSHD), the third most common muscular

52 dystrophy ^{7,8}, and forced DUX4 production in human primary myoblasts leads
53 to upregulation of genes active during early embryonic development ⁹. Based
54 on this premise, we analyzed publicly available RNA-seq datasets
55 corresponding to this period, focusing on *DUX4* and the 100 genes most
56 upregulated in *DUX4*-overexpressing muscle cells (Figure 1A, Table S1) ^{10,11}.
57 *DUX4* RNA was detected from oocyte to 4-cell (4C) stage, while transcripts
58 from its putative targets emerged on average at 2-cell (2C) and peaked at 8-
59 cell (8C) stages, as previously defined for human ZGA ¹². Transcripts
60 upregulated in *DUX4*-overexpressing muscle cells ¹¹ were also enriched at 8C
61 stage (Supplementary Figure 1AB), and upon clustering genes according to
62 their patterns of early embryonic expression (Figure 1B) we could delineate i)
63 1517 genes, the transcripts of which were already detected in oocytes,
64 plateaued up to 4C and abruptly dropped afterwards (maternal gene cluster);
65 ii) 94 genes and 124 genes, the expression of which started at 2C, and
66 peaked at 4C and 8C, respectively, before decreasing briskly, consistent with
67 early ZGA genes (2-4C and 2-8C gene clusters); and iii) 1352 genes
68 expressed only from 4C, peaking at 8C, and then decreasing progressively,
69 as expected for late ZGA genes (4-8C gene cluster). Only the two early ZGA
70 clusters (2-4C and 2-8C) were highly enriched for genes upregulated in
71 *DUX4*-overexpressing myoblasts (Figure 1C, Supplementary Figure 1C).

72 Chromatin cannot be reliably analyzed from the very low number of cells that
73 make up an early embryo, but ChIP-seq data obtained in *DUX4*
74 overexpressing human embryonic stem cells (hESCs) (Figure 2AB,
75 Supplementary Figure 2) and myoblasts ⁹ (Supplementary Figure 3) revealed
76 a marked enrichment of the transcription factor around the annotated

77 transcriptional start site (TSS) region of early ZGA genes (2-4C and 2-8C
78 clusters), but not of zygotic (maternal) and late ZGA (4-8C) genes.
79 Interestingly, several genes were not bound on their annotated TSS, but on
80 neighboring sequences, and their transcription was found to start near this
81 DUX4 binding site (Supplementary figure 4). It was previously demonstrated
82 that DUX4 drives expression of many of its target genes from alternative
83 promoters ¹¹. Upon examining publicly available single-cell RNA sequencing
84 data quantifying the far 5'-ends of transcripts (TFEs) in early human
85 development ¹³, we correspondingly found that the TFE of 24 out of 31 early
86 ZGA genes overlapped with DUX4 binding sites (Figure 2CD, Supplementary
87 figure 3CD). DUX4 was also recruited to several groups of transposable
88 elements (TEs), notably endogenous retroviruses such as HERVL, MER11B
89 and C, the expression of which increased at ZGA (Supplementary figure 2BC).
90 Furthermore, DUX4 overexpression in hESCs led to early ZGA genes
91 induction, as previously observed in myoblasts (Figure 2E) ¹¹.

92 Dux and DUX4 have largely conserved amino acid sequences, in particular
93 within the two DNA-binding homeodomains and the C-terminal region,
94 previously described as responsible for recruiting p300/CBP (Supplementary
95 Figure 5B) ¹⁴. The murine *Dux* tandem repeat encodes two main transcripts,
96 full-length *Dux* (or *Duxf3*) and a variant named *Gm4981* lacking the first
97 homeodomain (Supplementary Fig. 5A). Both *Dux* and *Gm4981* are
98 expressed in mouse embryos prior to ZGA-defining genes and transposable
99 elements (e.g. murine ERVL or MERVL) at the middle 2C stage, indicating
100 that their products likely are functional homologues of DUX4 (Figure 3A) ¹⁵.
101 To consolidate these results, we turned to mESCs, a small percentage of

102 which displays at any given time a 2C-like transcriptome in culture, with
103 expression of ZGA genes notably from the MERVL promoter^{16,17}. Upon
104 analyzing single-cell RNA-seq data from 2C-like mESCs¹⁸, we confirmed that
105 *Dux* transcripts were markedly enriched, as were early ZGA RNAs such as
106 *Zscan4*, *Zfp352* and *Cml2* (Figure 3B and Supplementary Figure 6). We used
107 CRISPR/Cas9-mediated genome editing to delete the *Dux*-containing
108 macrosatellite repeat in mESCs expressing a GFP reporter under control of a
109 MERVL promoter. This resulted in a complete absence of GFP⁺ 2C-like cells,
110 and in the loss of a large fraction of 2C-like cell-specific transcripts (Figure
111 3CD, Supplementary Figure 7). Overexpression of *Dux* but not *DUX4* rescued
112 the 2C-like state in the mESC KO clones (Figure 3EF, Supplementary Figure
113 8 and 9), albeit not in all cells where *Dux* was produced (Figure 3G).
114 Interestingly, both murine *Dux* and human *DUX4* were able to induce the
115 transcription of ZGA genes in the human 293T cell line (Supplementary Figure
116 10).

117 Upon depletion of the transcriptional repressor TRIM28 (tripartite motif-
118 containing protein 28; KAP1) from mESCs, expression of 2C-specific genes
119 increased as previously observed¹⁷, as did levels of *Dux* transcripts (Figure
120 4A, Supplementary Figure 11BCD). Remarkably, this phenotype was
121 completely abrogated in *Dux*-depleted mESCs (Figure 4BC, Supplementary
122 Figure 9 and 11ABCD). Correspondingly, we found that TRIM28 associates
123 with the 5'-end of the *Dux* gene and that tri-methylation of histone 3 lysine 9
124 (H3K9me3), a canonical marker of TRIM28-mediated repression, was
125 enriched on the *Dux* locus and lost upon knockdown of the heterochromatin
126 inducer (Figure 4D, Supplementary Figure 11EF).

127 Finally, we addressed the role of Dux during murine early embryonic
128 development. For this, we injected zygotes with plasmids encoding for the
129 Cas9 nuclease and either the two guide RNAs (sgRNAs) used to generate
130 *Dux* KO mESCs or a non-targeting sgRNA control. We then determined the
131 RNA profile of 2C embryos around 7 hours after the first cell division or
132 monitored their *ex vivo* development into blastocysts over 4 days (Figure 5A).
133 We found that Dux-depleted embryos presented a major differentiation defect,
134 most failing to reach the morula/blastocyst stage, and did not exhibit
135 transcriptional changes typical of ZGA, such as induction of *MERVL*, *Zscan4*
136 and several other tested early ZGA genes, and drop in *Mpo* maternal
137 transcript (Figure 5BC, Supplementary Figure 12).

138 In sum, our data reveal *DUX* genes as key regulators of early embryonic
139 development. The demonstrated ability of DUX4 to recruit the p300/CBP
140 complex and to induce local chromatin relaxation ¹⁴ as well as the mechanism
141 of action of Zelda, a master inducer of ZGA in *Drosophila* ^{19,20}, suggest that
142 DUX proteins could act as pioneer factors for transcriptional activation, by
143 opening chromatin around the TSS of early ZGA genes to facilitate access for
144 other transcription factors. Still, the genomic recruitment of pioneer factors
145 such as OCT4, NANOG and KLF4 can be hampered if heterochromatin marks
146 are overly abundant at their target loci ²¹. Many murine ZGA genes are
147 expressed from the LTR of endogenous retroviruses, which in mESC cells are
148 typically enriched in repressive marks ¹⁷. It could be that, at any given time,
149 these marks are relieved in only a small percentage of mESC in culture. What
150 drives this fluctuation remains to be determined. As well, what controls
151 expression of *DUX* genes themselves is yet to be defined, although the

152 conserved genomic localization of all placental mammal *DUX* orthologs close
153 to telomeric and subcentromeric regions suggests that this genomic context,
154 characterized by high levels of repression, might be of primary relevance ^{4,5,22}.
155 *DUX* genes seem indeed to become expressed only during events associated
156 with major chromatin relaxation, for instance in early embryos and upon loss
157 of repression of the D4Z4 macrosatellite repeat in myoblasts of FSHD patients
158 ^{23,24}. Our data indicate that TRIM28 plays a major role in murine *Dux*
159 repression, but the only mild increase in cells entering the 2C state when it is
160 depleted (around 5% of mESCs) and the demonstrated ability of several other
161 transcriptional modulators (e.g. SETDB1, EHMT2, HP1, CHAF1A/B, RYBP,
162 KDM1A) to prevent cycling of mESCs through this state indicate that control
163 of the *Dux* macrosatellite repeat is most likely multifactorial ^{16,25-29}. Broad de-
164 repression of the human and murine *DUX*-containing repeats could similarly
165 occur right after fertilization in either species. Future investigations of the
166 chromatin state of these loci in early embryos will shed light on the epigenetic
167 changes responsible for this process and on the nature of their molecular
168 mediators.

169

170 **Materials and methods**

171

172 *Cell lines and tissue culture*

173 mESC WT and KO for *Trim28* ³⁰, and E14 mESCs containing the MERVL
174 regulatory sequence driving expression of a 3XturboGFP-PEST ¹⁶ were
175 cultured in feeder-free conditions on 0.1% gelatin-coated tissue culture plates
176 in 2i medium, a N2B27 base medium supplemented with the MEK inhibitor,

177 PD0325901 (1 μ M), the GSK3 β inhibitor CHIR99021 (3 μ M) and LIF. E14
178 mESCs express the main markers of pluripotency (RNA-seq). H1 ESCs
179 (WA01, WiCell) were maintained in mTesRI (StemCell Technologies) on hES-
180 qualified Matrigel (BD Biosciences). 293T cells were maintained in DMEM
181 supplemented with 10% FCS. All cells were regularly checked for the absence
182 of mycoplasma contamination.

183

184 *Plasmids and lentiviral vectors*

185 The MT2/gag sequence was amplified from the pGL3 plasmid ²⁹, and the
186 human PGK promoter from pRRLSIN.cPPT.R1R2.PGK-GFP.WPRE ³⁰, to be
187 cloned upstream of luciferase in pGL4.20. Table S2 shows the primers used
188 to obtain truncations of the MT2/gag sequence. Single guide RNAs (sgRNAs)
189 targeting sequences flanking the 5' and 3' of the *Dux*-containing
190 macrosatellite repeat were cloned into px459 (version 2) using a standard
191 protocol ³¹. Table S2 shows the primers used to clone the sgRNAs. The
192 pLKO.1-puromycin shRNA vector was used for the Trim28 knock-down ³⁰.
193 The pLKO.1 vector was further modified to express blasticidin-S-deaminase
194 drug resistance cassette in place of the puromycin N-acetyltransferase. The
195 resulting pLKO.1-blasticidin backbone was used to clone shRNAs against the
196 murine *Dux* transcript. The sequence of the primers used to clone the *Dux*
197 shRNA is shown in Table S2. The *Gm4981* cDNA was cloned from the
198 genome of E13 mESCs while codon-optimized h*DUX4* and m*Dux* were
199 synthesized (Invitrogen). *Gm4981*, *DUX4*, *Dux* and *LacZ* cDNAs were cloned
200 in the pAIB HIV-1-based transfer vector encoding also for blasticidin
201 resistance using the In-Fusion® HD Cloning Kit (Clontech) ³². pMD2-G

202 encodes the vesicular stomatitis virus G protein (VSV-G). The minimal HIV-1
203 packaging plasmid 8.9NdSB carrying a double mutation in the capsid protein
204 (P90A/A92E) was used to achieve higher transduction of the lowly permissive
205 mESCs³³.

206

207 *Production of lentiviral vectors, transduction and transfection of mammalian*
208 *cells*

209 Lentiviral vectors were produced by transfection of 293T cells using
210 Polyethylenimine (PEI) (Sigma, Inc)³³. To generate stable KDs, mESCs were
211 transduced with empty pLKO.1 vector or vectors containing the shRNA
212 targeting *Kap1* or *Dux* transcripts³⁰. Cells were selected with 1 µg/ml
213 puromycin or 3 µg/ml blasticidin starting one day after transduction. hESCs
214 expressing LacZ and DUX4 were generated by transfecting the corresponding
215 AIB plasmids with *TransIT*®-LT1 Transfection Reagent (Mirus Bio LLC), while
216 nucleofection (Amaxa™ P3 Primary Cell 4D-Nucleofector™ X Kit) was used
217 to engineer mESC expressing LacZ, DUX4, *Dux* and Gm4981.

218

219 *Creation of Dux KO mESC lines*

220 E14 mESCs containing the MERVL regulatory sequence driving expression of
221 a 3XturboGFP-PEST were co-transfected with px459 plasmids encoding for
222 Cas9, the appropriate sgRNAs and puromycin resistance cassette by
223 nucleofection (Amaxa™ P3 Primary Cell 4D-Nucleofector™ X Kit). 24 hours
224 later, the cells were selected for 48 hours with 1 µg/ml puromycin, single-cell
225 cloned by serial dilution, expanded and their DNA was extracted to detect the

226 presence of WT and/or KO alleles. Three WT and three homozygous *Dux* KO
227 clones were selected and used in this study.

228

229 *Luciferase assay*

230 293T or E14 mESCs were cotransfected with the various pGL4.20 derivatives,
231 the renilla plasmid and the pAIB transfer vector encoding either for LacZ, Dux,
232 Gm4981 or DUX4 using Lipofectamine 3000 (Invitrogen). Luciferase activity
233 was quantified 24h after transfection. Firefly luciferase activity was normalized
234 to the activity of *Renilla* luciferase. Light emission was measured on a
235 luminescence plate reader.

236

237 *Immunofluorescence assay*

238 mESC clones expressing an HA-tagged Dux protein were fixed for 20 min
239 with 4% paraformaldehyde, permeabilized for 5 min with 0.1% Triton-X 100,
240 and blocked for 30 min with 1% BSA in PBS. Cells were then incubated for 1
241 hour with anti-HA.11 (Covance) or anti-NANOG (Active Motif) or anti-SOX2
242 (Active Motif) antibodies diluted in PBS with 1% BSA. After 3 washes, the
243 cells were incubated with anti-mouse (HA) or anti-rabbit (NANOG, SOX2)
244 Alexa Fluor 647-conjugated secondary antibodies for 1 hour and washed
245 again three times. Every step until this point, was carried with cells in
246 suspension. Pelleted cells were then resuspended in VECTASHIELD®
247 Mounting Medium with DAPI (Vector Laboratories) and mounted on the
248 coverslip. The slides were viewed with a Zeiss LSM700 confocal microscope.

249

250 *Fluorescence-activated cell sorting (FACS)*

251 FACS analysis was performed with a BD FACScan system. Trim28 knock-
252 down mESCs containing the MT2/gag-GFP reporter were subjected to FACS
253 sorting with AriaII (BD Biosciences).

254

255 *Standard PCR, RT-PCR and RNA sequencing*

256 For the genotyping of *Dux* WT and KO alleles, genomic DNA was extracted
257 with DNeasy Blood & Tissue Kits (QIAGEN) and the specific PCR products
258 were amplified using PCR Master Mix 2X (Thermo Scientific) combined with
259 the appropriate primers (design in Supplementary Figure 6A; primer
260 sequences in Table S2).

261 Total RNA from cell lines was isolated using the High Pure RNA Isolation Kit
262 (Roche). cDNA was prepared with SuperScript II reverse transcriptase
263 (Invitrogen). Ambion Single Cell-to-CT kit (Thermo Fisher) was used for RNA
264 extraction, cDNA conversion and mRNA pre-amplification of 2C stage
265 embryos. Primers listed in Supplementary Table S2 were used for SYBR
266 green qPCR (Applied Biosystems). Library preparation and 150-base-pair
267 paired-end RNA-seq were performed using standard Illumina procedures for
268 the NextSeq 500 platform (GSE94325).

269

270 *ChIP and ChIP sequencing*

271 ChIP and library preparation were performed as described previously ³⁰.
272 DUX4-HA ChIP was done using the anti-HA.11 (Covance) antibody.
273 Sequencing of Trim28 and H3K9me3 ChIP was performed with Illumina
274 HiSeq 2500 in 100-bp reads run. Sequencing of DUX4 was performed with
275 Illumina NextSeq 500 in 75-bp paired-end reads run.

276

277 *RNA-seq datasets preprocessing*

278 Single-cell RNA-Seq of human and mouse early embryo development
279 (GSE36552 and GSE45719 respectively), single-cell RNA-Seq of 2C-like cells
280 (E-MTAB-5058), DUX4 overexpression in human myoblasts (GSE45883), and
281 KAP1 KO (GSE74278) datasets were downloaded from different repositories
282 (GEO, and ArrayExpress) ^{34,35}. Reads were mapped to the human genome
283 (hg19) or mouse genome (mm9) using TopHat (v2.0.11) ³⁶ in sensitive mode
284 (the exact parameters are: tophat -g 1 --no-novel-juncs --no-novel-indels -G
285 \$gtf --transcriptome-index \$ transcriptome --b2-sensitive -o \$localdir \$index
286 \$reads1 \$reads2). Gene counts were generated using HTSeq-count.
287 Normalization for sequencing depth and differential gene expression analysis
288 was performed using Voom ³⁷ as it has been implemented in the limma
289 package of Bioconductor ³⁸. TEs overlapping exons were removed from the
290 analysis. Counts per TE integrant (genomic loci) were generated using the
291 multiBamCov tool from the bedtools software ³⁹. Normalisation for sequencing
292 depth was performed using Voom, with total number of reads on genes as
293 size factor. To compute total number of reads per TE family, counts on all
294 integrants of each family were added up.

295

296 *Analysis of single cell expression data from human and mouse embryonic*
297 *stages*

298 For every embryonic stage we performed a statistical test to find the genes
299 that had a different expression level compared to the other stages ¹⁰, using a
300 moderated F-test (comparing the interest group against every other) as

301 implemented in the limma package of Bioconductor. Genes were selected as
302 expressed in a specific stage if having a significant p-value (<0.05 after
303 adjusting for multiple testing with the Benjamini and Hochberg method) and
304 an average fold change respective to the other embryonic stages bigger than
305 10. We additionally removed all genes exhibiting a 1.1-fold higher expression
306 in any of the embryonic stages compared to the stage analyzed (Suppl.
307 Figure 1A). Note that with this approach a gene can be marked as expressed
308 in more than one stage. Codes are available on demand.

309

310 *Correspondence between DUX4 overexpression and single cell expression*
311 *data from human embryonic stages*

312 For every stage, we classified the genes in 4 patterns of expression by
313 performing a hierarchical clustering (with Pearson correlation as distance and
314 complete agglomeration method). Figure 1B shows the 2 most relevant
315 patterns derived from the 4C and 8C stages.

316 Expression of the genes identified with this method was then compared
317 between DUX4- and GFP-overexpressing human myoblast cells. For a gene
318 to be considered differentially expressed, a p-value (after multiple testing
319 correction with the Benjamini and Hochberg method) lower than 0.05 and a
320 fold change bigger than 2 were imposed. A moderated t-test was used for the
321 statistical test, as implemented in the limma package of Bioconductor.

322

323 *ChIP-seq data processing*

324 ChIP-seq dataset of DUX4 overexpressed in human myoblasts (GSE94325)
325 was downloaded from GEO. Reads were mapped to the human genome

326 assembly hg19 using Bowtie2 using the sensitive-local mode ⁴⁰. SICER was
327 used to call histone mark peaks ⁴¹. For the ones that are not histone marks,
328 we used MACS (with default parameters) when the data was single-end and
329 MACS2 (the exact parameters are: macs2 callpeak -t \$chipbam -c \$tibam -f
330 BAM -g \$org -n \$name -B -q 0.01 --format BAMPE) when the data was
331 paired-end ⁴². Both, SICER peaks with an FDR above 0.05 and MACS peaks
332 with a score lower than 50 were discarded. RSAT was used for motif
333 discovery and to compute motif abundance ⁴³. To compute the percentage of
334 bound TE integrants in each family, we used bedtools suite.

335

336 *Coverage plots*

337 ChIP-seq signals on features of interest were extracted from the bigWigs
338 beforehand normalized for sequencing depth (reads per hundred millions).
339 Each signal was then smoothed using a running average of window 75bp for
340 DUX4, 250bp for Trim28, and 500bp for H3K9me3. Finally, the mean and
341 standard error of the mean of the signals were computed and plotted for each
342 set of features of interest. Scripts are available on demand.

343

344 *Pronuclear injection of mouse embryos*

345 Pronuclear injection was performed according to the standard protocol of the
346 Transgenic Core Facility of EPFL. In summary, B6D2F1 mice were used as
347 egg donors (5 weeks old). Mice were injected with PMSG (10 IU), and HCG
348 (10 IU) 48 hours after. After mating females with B6D2F1 males, zygotes
349 were collected and kept in KSOM medium pre-gassed in 5% CO₂ at 37 °C.
350 Embryos were then transferred to M2 medium and microinjected with 10

351 ng/ μ g of either a px459 plasmid containing a non-targeting sgRNA or a mix of
352 the two plasmids used to obtain the KO in mESCs, in injection buffer (10mM
353 Tris HCl pH7.5, 0.1mM EDTA pH8, 100mM NaCl). After microinjection,
354 embryos were cultured in KSOM medium at 37 °C in 5% CO₂ for 4 days. In
355 each of three independent experiments, 5 embryos per condition were
356 collected around 7 hours after first cell division (2C formation) for qPCR
357 analysis, and differentiation of the remaining embryos was followed. At day 4,
358 all the fertilized embryos (between 16 to 23 per condition) were classified for
359 their developmental state. Randomization and blind outcome assessment
360 were not applied. All animal experiments were approved by the local
361 veterinary office and carried out in accordance with the EU Directive (2010/63/
362 EU) for the care and use of laboratory animals.

363

364 *Sample sizes and statistical tests*

365 We used non-parametric statistical tests (2-sided Wilcoxon test), when we
366 had enough sample size (low-cell number qPCR). Otherwise we used a 2-
367 sided unpaired t-test (standard qPCR and FACS). Fisher's exact test was
368 used to test for differences in proportions in contingency tables.

369

370 **Data availability**

371 RNA-seq and ChIP-seq data generated in this study have been deposited in
372 the NCBI Gene Expression Omnibus (GEO) under accession number
373 GSE94325.

374

375 **Acknowledgments**

376 We thank T. Macfarlan and M.E. Torres-Padilla for sharing reagents and for
377 helpful discussions, the Transgenic and Gene Expression Core Facility
378 (EPFL) and S. Offner for technical assistance. This work was financed
379 through grants from the Swiss National Science Foundation, the Gebert-Rüf
380 Foundation, the INGENIUM grant (FP7 MC-ITN INGENIUM 290123), and the
381 European Research Council (ERC 268721 and ERC 694658) to D.T.

382

383 **Author contributions**

384 A.D.I and D.T. conceived the project, designed the experiments, analyzed the
385 data and wrote the manuscript; A.D.I., A.C. and S.V. carried out the
386 experiments; E.P. and J.D. performed the bioinformatics and statistical
387 analyses.

388

389 **Conflict of interest**

390 The authors declare that they have no conflict of interest.

391

392 **References**

- 393 1 Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation
394 during the maternal-to-zygotic transition. *Annual review of cell and*
395 *developmental biology* **30**, 581-613, doi:10.1146/annurev-cellbio-
396 100913-013027 (2014).
- 397 2 Liang, H. L. *et al.* The zinc-finger protein Zelda is a key activator of the
398 early zygotic genome in *Drosophila*. *Nature* **456**, 400-403,
399 doi:10.1038/nature07388 (2008).
- 400 3 Lee, M. T. *et al.* Nanog, Pou5f1 and SoxB1 activate zygotic gene expression
401 during the maternal-to-zygotic transition. *Nature* **503**, 360-364,
402 doi:10.1038/nature12632 (2013).
- 403 4 Leidenroth, A. *et al.* Evolution of DUX gene macrosatellites in placental
404 mammals. *Chromosoma* **121**, 489-497, doi:10.1007/s00412-012-0380-y
405 (2012).
- 406 5 Clapp, J. *et al.* Evolutionary conservation of a coding function for D4Z4, the
407 tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy.

- 408 *American journal of human genetics* **81**, 264-279, doi:10.1086/519311
409 (2007).
- 410 6 Hewitt, J. E. *et al.* Analysis of the tandem repeat locus D4Z4 associated
411 with facioscapulohumeral muscular dystrophy. *Human molecular genetics*
412 **3**, 1287-1295 (1994).
- 413 7 Wijmenga, C. *et al.* Chromosome 4q DNA rearrangements associated with
414 facioscapulohumeral muscular dystrophy. *Nature genetics* **2**, 26-30,
415 doi:10.1038/ng0992-26 (1992).
- 416 8 Gabriels, J. *et al.* Nucleotide sequence of the partially deleted D4Z4 locus
417 in a patient with FSHD identifies a putative gene within each 3.3 kb
418 element. *Gene* **236**, 25-32 (1999).
- 419 9 Geng, L. N. *et al.* DUX4 activates germline genes, retroelements, and
420 immune mediators: implications for facioscapulohumeral dystrophy.
421 *Developmental cell* **22**, 38-51, doi:10.1016/j.devcel.2011.11.013 (2012).
- 422 10 Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation
423 embryos and embryonic stem cells. *Nature structural & molecular biology*
424 **20**, 1131-1139, doi:10.1038/nsmb.2660 (2013).
- 425 11 Young, J. M. *et al.* DUX4 binding to retroelements creates promoters that
426 are active in FSHD muscle and testis. *PLoS genetics* **9**, e1003947,
427 doi:10.1371/journal.pgen.1003947 (2013).
- 428 12 Vassena, R. *et al.* Waves of early transcriptional activation and
429 pluripotency program initiation during human preimplantation
430 development. *Development* **138**, 3699-3709, doi:10.1242/dev.064741
431 (2011).
- 432 13 Tohonen, V. *et al.* Novel PRD-like homeodomain transcription factors and
433 retrotransposon elements in early human development. *Nature*
434 *communications* **6**, 8207, doi:10.1038/ncomms9207 (2015).
- 435 14 Choi, S. H. *et al.* DUX4 recruits p300/CBP through its C-terminus and
436 induces global H3K27 acetylation changes. *Nucleic acids research*,
437 doi:10.1093/nar/gkw141 (2016).
- 438 15 Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq
439 reveals dynamic, random monoallelic gene expression in mammalian cells.
440 *Science* **343**, 193-196, doi:10.1126/science.1245316 (2014).
- 441 16 Ishiuchi, T. *et al.* Early embryonic-like cells are induced by
442 downregulating replication-dependent chromatin assembly. *Nature*
443 *structural & molecular biology* **22**, 662-671, doi:10.1038/nsmb.3066
444 (2015).
- 445 17 Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with
446 endogenous retrovirus activity. *Nature* **487**, 57-63,
447 doi:10.1038/nature11244 (2012).
- 448 18 Eckersley-Maslin, M. A. *et al.* MERVL/Zscan4 Network Activation Results
449 in Transient Genome-wide DNA Demethylation of mESCs. *Cell reports* **17**,
450 179-192, doi:10.1016/j.celrep.2016.08.087 (2016).
- 451 19 Sun, Y. *et al.* Zelda overcomes the high intrinsic nucleosome barrier at
452 enhancers during *Drosophila* zygotic genome activation. *Genome research*
453 **25**, 1703-1714, doi:10.1101/gr.192542.115 (2015).
- 454 20 Schulz, K. N. *et al.* Zelda is differentially required for chromatin
455 accessibility, transcription factor binding, and gene expression in the

- 456 early *Drosophila* embryo. *Genome research* **25**, 1715-1726,
457 doi:10.1101/gr.192682.115 (2015).
- 458 21 Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the
459 pluripotency reprogramming factors' initial engagement with the genome.
460 *Cell* **151**, 994-1004, doi:10.1016/j.cell.2012.09.045 (2012).
- 461 22 Perrod, S. & Gasser, S. M. Long-range silencing and position effects at
462 telomeres and centromeres: parallels and differences. *Cellular and*
463 *molecular life sciences : CMLS* **60**, 2303-2318, doi:10.1007/s00018-003-
464 3246-x (2003).
- 465 23 van der Maarel, S. M., Tawil, R. & Tapscott, S. J. Facioscapulohumeral
466 muscular dystrophy and DUX4: breaking the silence. *Trends in molecular*
467 *medicine* **17**, 252-258, doi:10.1016/j.molmed.2011.01.001 (2011).
- 468 24 Wu, J. *et al.* The landscape of accessible chromatin in mammalian
469 preimplantation embryos. *Nature* **534**, 652-657,
470 doi:10.1038/nature18606 (2016).
- 471 25 Maksakova, I. A. *et al.* Distinct roles of KAP1, HP1 and G9a/GLP in
472 silencing of the two-cell-specific retrotransposon MERVL in mouse ES
473 cells. *Epigenetics & chromatin* **6**, 15, doi:10.1186/1756-8935-6-15 (2013).
- 474 26 Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in
475 enhancer activity and telomere elongation. *Genes & development* **28**,
476 2103-2119, doi:10.1101/gad.248005.114 (2014).
- 477 27 Schoorlemmer, J., Perez-Palacios, R., Climent, M., Guallar, D. & Muniesa, P.
478 Regulation of Mouse Retroelement MuERV-L/MERVL Expression by REX1
479 and Epigenetic Control of Stem Cell Potency. *Frontiers in oncology* **4**, 14,
480 doi:10.3389/fonc.2014.00014 (2014).
- 481 28 Walter, M., Teissandier, A., Perez-Palacios, R. & Bourc'his, D. An epigenetic
482 switch ensures transposon repression upon dynamic loss of DNA
483 methylation in embryonic stem cells. *eLife* **5**, doi:10.7554/eLife.11418
484 (2016).
- 485 29 Macfarlan, T. S. *et al.* Endogenous retroviruses and neighboring genes are
486 coordinately repressed by LSD1/KDM1A. *Genes & development* **25**, 594-
487 607, doi:10.1101/gad.2008511 (2011).
- 488 30 Ecco, G. *et al.* Transposable Elements and Their KRAB-ZFP Controllers
489 Regulate Gene Expression in Adult Tissues. *Developmental cell* **36**, 611-
490 623, doi:10.1016/j.devcel.2016.02.024 (2016).
- 491 31 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system.
492 *Nature protocols* **8**, 2281-2308, doi:10.1038/nprot.2013.143 (2013).
- 493 32 De Iaco, A. *et al.* TNPO3 protects HIV-1 replication from CPSF6-mediated
494 capsid stabilization in the host cell cytoplasm. *Retrovirology* **10**, 20,
495 doi:10.1186/1742-4690-10-20 (2013).
- 496 33 De Iaco, A. & Luban, J. Cyclophilin A promotes HIV-1 reverse transcription
497 but its effect on transduction correlates best with its effect on nuclear
498 entry of viral cDNA. *Retrovirology* **11**, 11, doi:10.1186/1742-4690-11-11
499 (2014).
- 500 34 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI
501 gene expression and hybridization array data repository. *Nucleic acids*
502 *research* **30**, 207-210 (2002).
- 503 35 Kolesnikov, N. *et al.* ArrayExpress update--simplifying data submissions.
504 *Nucleic acids research* **43**, D1113-1116, doi:10.1093/nar/gku1057 (2015).

- 505 36 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the
506 presence of insertions, deletions and gene fusions. *Genome biology* **14**,
507 R36, doi:10.1186/gb-2013-14-4-r36 (2013).
- 508 37 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock
509 linear model analysis tools for RNA-seq read counts. *Genome biology* **15**,
510 R29, doi:10.1186/gb-2014-15-2-r29 (2014).
- 511 38 Gentleman, R. C. *et al.* Bioconductor: open software development for
512 computational biology and bioinformatics. *Genome biology* **5**, R80,
513 doi:10.1186/gb-2004-5-10-r80 (2004).
- 514 39 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for
515 comparing genomic features. *Bioinformatics* **26**, 841-842,
516 doi:10.1093/bioinformatics/btq033 (2010).
- 517 40 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.
518 *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 519 41 Zang, C. *et al.* A clustering approach for identification of enriched domains
520 from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952-1958,
521 doi:10.1093/bioinformatics/btp340 (2009).
- 522 42 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology*
523 **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
- 524 43 Thomas-Chollier, M. *et al.* RSAT: regulatory sequence analysis tools.
525 *Nucleic acids research* **36**, W119-127, doi:10.1093/nar/gkn304 (2008).
526
527
528

529 **Figure Legends**

530 **Figure 1. DUX4 promotes transcription of genes expressed during early**
531 **ZGA**

532 **(A)** Comparative expression during early human embryonic development of
533 *DUX4* (red) and the top 100 genes upregulated upon *DUX4* overexpression in
534 human primary myoblasts (blue, full line average, dashed lines 95%
535 confidence interval around the mean). Oo, oocyte; Zy, zygote; 2C, 4C, 8C,
536 corresponding n-cell stage; Mo, morula; Bl, blastocyst. **(B)** Cluster of genes
537 differentially expressed during early embryonic development were selected
538 from the previously identified subsets of genes (Supplementary Figure 1A)
539 based on high expression at 4C (upper panels) and 8C (lower panels). Blue
540 and dotted line delineate mean and 95% confidence, respectively. **(C)**
541 Expression of genes from each cluster illustrated in (B) when *DUX4* is
542 ectopically expressed in human primary myoblasts. Lower parts of the panels
543 depict the fold change expression of genes within these clusters, all randomly
544 distributed along the y-axes, with kernel density plotted in the upper part.

545

546 **Figure 2. DUX4 binds TSSs of genes expressed during early ZGA and**
547 **activates their expression in hESCs.**

548 **(A)** Average coverage normalized for sequencing depth of ChIP-seq signal of
549 *DUX4* (blue) when overexpressed in hESCs in a window of 5 kb from the
550 annotated TSS of genes belonging to the 2-4C and 2-8C clusters from Figure
551 1B. Total input is represented in gray (line, average; shade, standard error of
552 the mean). **(B)** Fraction of genes belonging to each cluster from Figure 1B with
553 a *DUX4* peak within 5 kb of their annotated TSS. Fisher's exact test was

554 performed to compare maternal vs. 2-4C and 2-8C ($p= 3.54e^{-61}$ and $p= 2.23e^{-}$
 555 13 respectively) **(C)** Average coverage of ChIP-seq signal of DUX4 (blue)
 556 when overexpressed in hESCs within 5 kb of TFE of transcripts specifically
 557 upregulated at oocyte-to-4C and 4C-to-8C transitions. Total input is
 558 represented in gray (line, average; shade, standard error of the mean). **(D)**
 559 Fraction of TFE from oocyte-to-4C ($n=32$) and 4C-to-8C ($n=128$) transitions
 560 that have a DUX4 peak overlapping with their 5' end. Fisher's exact test was
 561 performed to compare 4C-to-8C vs. oocyte-to-4C TFEs ($p= 4.48e^{-17}$). **(E)**
 562 Comparative expression in hESCs of three genes activated at ZGA (*ZSCAN4*,
 563 *MBD3L2* and *DUXA*) and two control housekeeping genes (*ACTB* and *TBP*)
 564 24 hours after transfection with plasmids expressing LacZ (green squares) or
 565 DUX4 (blue circles). Expression was normalized to *ACTB*. Horizontal lines
 566 represent the mean. *** $p \leq 0.001$, unpaired t-test.

567

568 **Figure 3. Dux is necessary for formation of 2C-like mESCs.**

569 **(A)** Comparative expression of the two alternative transcripts of *Dux*, *Dux*
 570 (pink) and *Gm4981* (orange), with genes (blue) and transposable elements
 571 (MERVL; green) specifically expressed during murine ZGA. Full lines
 572 represent the average and dashed lines the 95% confidence interval around
 573 the mean **(B)** Single-cell RNA-sequencing comparison between mESCs
 574 sorted for expression of both tomato and GFP reporters driven by MERVL
 575 and *Zscan4* promoters, respectively (revelators of 2C-like cells), and the
 576 double negative population. Average gene expression was quantified and fold
 577 change between positive and negative cells was plotted. Dots are randomly

578 distributed along the y-axes. The upper plot represents the kernel density
579 estimate of middle-2C stage (blue line) and the rest of the genes (gray line).
580 The *Dux* macrosatellite repeat was deleted in mESCs carrying a MERVL-GFP
581 reporter by CRISPR/Cas9-mediated excision. **(C)** Fraction of GFP⁺ cells in
582 WT or *Dux*-deleted cells. **(D)** RNA sequencing analysis of WT and *Dux* KO
583 mESC clones. The dot plot displays the average gene expression of three
584 independent clones from each cell type. **(E)** GFP expression in *Dux* KO (blue
585 circles) and WT (green squares) mESC clones carrying an integrated
586 MERVL-GFP reporter, and transiently expressing *LacZ*, *DUX4*, *Dux* or
587 *Gm4981* transgenes. **(F)** RNA sequencing analysis of *Dux* KO mESC clones
588 transiently expressing *Dux* or control. The dot plot displays the average gene
589 expression of two independent clones from each cell type. **(G)** *Dux* KO
590 mESCs carrying an integrated MERVL-GFP reporter and transiently
591 expressing a HA-tagged form of *Dux* were stained for HA and
592 immunofluorescence was detected by confocal microscopy. DAPI, blue; GFP,
593 green; HA, red. Horizontal bars in **(C)** and **(E)** represent the mean. *** $p \leq$
594 0.001, unpaired t-test.

595

596 **Figure 4. TRIM28 regulates formation of 2C-like mESCs by repressing**
597 ***Dux* expression**

598 **(A)** RNA sequencing analysis of WT and *Trim28* KO mESCs. Average gene
599 expression was quantified and fold change between KO and WT cells plotted.
600 Dots are randomly distributed along the y-axes. The upper plot represents the
601 kernel density estimate of genes specifically expressed in 2C-like mESCs
602 (green line) and the rest of the genes (gray line). **(B)** WT (blue circles) and

603 *Dux* KO (green squares) mESC clones carrying an integrated MERVL-GFP
604 reporter were transduced with lentiviral vectors encoding for shRNAs targeting
605 *Trim28* or a control. 4 days later GFP expression was quantified. Horizontal
606 lines represent the mean. *** $p \leq 0.001$, unpaired t-test. **(C)** RNA sequencing
607 of *Trim28*-depleted or control *Dux* KO mESC clones. The dot plot represents
608 the average gene expression of three independent KO clones transduced with
609 lentiviral vectors encoding for a control or a *Trim28*-specific shRNA. **(D)**
610 Average coverage of ChIP-seq signal of *Trim28* (top plot; blue lines; two
611 replicates) and H3K9me3 (bottom plot; two replicates) in control (red lines)
612 and *Trim28* KD mESCs (green line) around the *Dux* gene. Total input is
613 represented in gray. ChIP-seq reads were mapped on the genome, before
614 focusing the analysis on a 500bp window around the main *Dux* gene.
615 H3K9me3 peaks over the *Dux* macrosatellite repeat were only called in the
616 control KD mESCs (Sicer; false discovery rate 0.05)

617

618 **Figure 5. *Dux* is necessary for mouse early embryonic development**

619 **(A)** Schematic of the *Dux* loss-of-function experiment in mouse pre-
620 implantation embryos. Zygotes were first injected in the pronucleus with
621 plasmids encoding for the Cas9 nuclease and sgRNAs targeting the flanking
622 region of the *Dux* macrosatellite repeat or a non-targeting sgRNA, then were
623 either **(B)** monitored for their ability to differentiate *ex vivo* or **(C)** collected at
624 2C-stage for mRNA quantification. **(B)** Average percent of embryos reaching
625 the morula/blastocyst stages (white) or failing to differentiate (delayed/dead
626 embryos, black; defective morula/blastocyst, grey) 4 days after pronuclear
627 injection. The plot represents an average from 3 independent experiments

628 with 16 to 23 embryos for each condition. Fisher's exact test was performed to
629 compare the embryonic stage of control against *Dux* KO ($p= 1.54e^{-10}$) **(C)**
630 Comparative expression of *Dux*, early ZGA genes (*Zscan4*, *Sp110*,
631 *B020004J07Rik*, *Dub1*, *Tdpoz4*, *Eif1a*, *Tcstv3*, *Cml2*), 2C-restricted TE
632 (MERVL, the LTR and int regions of which are detected with MT2_mm and
633 MERVL-int primers, respectively), a gene (*Mpo*), the expression of which
634 decreases at ZGA, 2 genes (*Actb*, *Zbed3*) stably expressed during pre-
635 implantation embryonic development and a control TE (IAPEz) in 15 2C stage
636 embryos (5 from each of 3 independent experiments) 15-24 hours after
637 pronuclear injection with plasmids expressing Cas9 and control or *Dux*-
638 specific sgRNAs. Boxes depict the 25 and 75 percentiles, line in the boxes
639 represents the median. Expression was normalized to *Actb*. * $p \leq 0.05$ ** $p \leq$
640 0.01, *** $p \leq 0.001$, Wilcoxon test.