Published as: *Nat Genet*. 2017 June ; 49(6): 925–934.

Conserved roles for murine DUX and human DUX4 in activating cleavage stage genes and MERVL/HERVL retrotransposons

Peter G. Hendrickson[1], **Jessie A. Doráis**[1,2], **Edward J. Grow**[1], **Jennifer L. Whiddon**[3], **Jong-Won Lim**[3], **Candice L. Wike**[1], **Bradley D. Weaver**[1], **Christian Pflueger**[1], **Benjamin R. Emery**[2], **Aaron L. Wilcox**[2], **David A. Nix**[1], **C. Matthew Peterson**[2], **Stephen J. Tapscott**[3,#], **Douglas T. Carrell**[2,#], and **Bradley R. Cairns**[1,#]

[1]Department of Oncological Sciences, Huntsman Cancer Institute and Howard Hughes Medical Institute, Salt Lake City, UT, USA

[2]Departments of Obstetrics and Gynocology, and Surgery. University of Utah School of Medicine, Salt Lake City, UT, USA

[3]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

To better understand transcriptional regulation during human oogenesis and pre-implantation development, we defined stage-specific transcription, which revealed the cleavage stage as highly distinctive. Here, we present multiple lines of evidence that a eutherian-specific, multi-copy retrogene, *DUX4*, encodes a transcription factor which activates hundreds of endogenous genes (e.g. *ZSCAN4*, *ZFP352*, *KDM4E*) and retroviral elements (MERVL/HERVL-family) that defines the cleavage-specific transcriptional programs in mouse and human. Remarkably, mouse *Dux* expression is both necessary and sufficient to convert mouse embryonic stem cells into two-cell embryo-like ('2C-like') cells, measured here by the reactivation of '2C' genes and repeat elements, the loss of POU5F1 protein and chromocenters, and by the conversion of the chromatin landscape (assessed by ATAC-seq) to a state strongly resembling mouse two-cell embryos. Taken together, we propose mouse DUX and human DUX4 as major drivers of the cleavage/'2C' state.

[#]Corresponding authors: brad.cairns@hci.utah; douglas.carrell@hsc.utah.edu; stapscot@fredhutch.org.

Data Accession

All sequencing data has been deposited to GEO and can be found under the series accession number GSE85632.

Author Contributions

IRB processing, patient consent, patient management and sample selection/processing was overseen by J.A.D., D.T.C., and C.M.P., with processing by clinical staff (B.R.E., and A.L.W.) in clinical (non-federally funded) facilities. Following cDNA and library preparation, subsequent sequencing and transcriptome analyses, along with all molecular and functional approaches were overseen by B.R.C., with contributions from S.J.T. Experiments were performed, analyzed, and statistically evaluated by P.G.H., with contributions from E.J.G., J.L.W., J-W.L., C.L.W., B.D.W., C.P., B.R.E., and D.A.N. The manuscript was written by P.G.H. and B.R.C.

Competing Financial Interests

The authors have no competing financial interests.

Introduction

Mammalian pre-implantation development is a fascinating and complex developmental time that involves major changes in chromatin structure and transcriptional activity. Several events that occur specifically during cleavage stage (2-cell, 4-cell, and 8-cell embryos) are critical for embryonic success, including embryonic genome activation (EGA), epigenetic reprogramming (e.g. DNA demethylation and chromatin remodeling), and restoration of telomere length[1]. Despite their importance, our understanding of their mechanisms and upstream regulation remain limited. Here, KDM4-family H3K9 demethylase enzymes are involved in heterochromatin de-repression[2, 3], and the ZSCAN4 transcription factor family in the sister chromatid exchange (T-SCE) mechanism needed for telomere elongation[4, 5]. The mRNAs for *KDM4E* and *ZSCAN4* are not maternally inherited, and are expressed exclusively during cleavage stage; however, which transcription factor(s) enable cleavage-specific expression, and how they are linked mechanistically to EGA are major unanswered questions.

Remarkably, these gene families and many other cleavage-specific genes in mice have exapted retrotransposons – specifically cleavage-specific MERVL elements – for their coordinated expression[6, 7]. Curiously, MERVL and MERVL-linked genes are also spontaneously reactivated in a rare subpopulation of pluripotent mouse embryonic stem cell (mESC), termed the '2C-like' cell[8]. Coincident with MERVL reactivation, '2C-like' cells acquire the unique molecular and developmental features and functions of totipotent cleavage-stage cells[9–11], prompting interest in defining upstream regulatory factors.

Our initial efforts sought to define the changes in transcription/transcript abundance that accompany human egg and pre-implantation embryo development, and the datasets we present here provide a deep resource for future studies. Our analyses revealed the cleavage stage as highly unique, similar to observations made in mouse, and our *in silico* analyses suggested upstream regulatory involvement of a cleavage-specific homeodomain transcription factor called DUX4 (Fig. 1). The *DUX4* gene has been extensively characterized for its causal involvement in the disease facioscapulohumeral muscular dystrophy (FSHD) whereby its improper expression in myoblasts activates genes and retrotransposons normally expressed in human embryos, triggering apoptosis[12, 13]. Here, we provide multiple lines of evidence that DUX4 and its mouse ortholog, DUX, share central roles in driving cleavage-specific gene expression (including *ZSCAN4*, *KDM4E*, *PRAMEF,* etc.), ERVL-family retrotransposon transcription, and chromatin remodeling. Taken together, DUX4 appears to reside at the top of a transcriptional hierarchy initiated at EGA that helps drive important developmental events during mammalian embryogenesis.

Results

Transcriptomes of oocytes and pre-implantation development

Samples from seven stages of human oogenesis and early embryogenesis were donated from consented patients undergoing in vitro fertilization (IVF) in accordance with Institutional Review Board (IRB) guidelines and approval (Fig. 1a, left panel). Blastocyst embryos were manually separated into ICM and mural trophectoderm by laser dissection (Fig. 1a, right

panel). To minimize variation, all samples were processed together. For each, total RNA was divided (providing two technical replicates) and processed in parallel using a transposase-based library method to sequence total RNA without 3' bias[14]. To maximize dataset utility, we performed deep RNA sequencing (RNA-seq) using a paired-end 101bp sequencing format. Replicates were highly concordant (spearman correlation, r>0.92), and yielded on average ~76 million unique, stranded, mappable reads (Supplementary Table 1). Importantly, read coverage from transcription start site (TSS) to transcription termination site (TTS) was exceptionally well-balanced compared to prior work (Fig. 1b, Supplementary Fig. 1a), making these new datasets the most comprehensive transcriptomes of human oocyte and pre-implantation embryonic development to date.

PCA and clustering analyses reveal a unique cleavage-stage transcriptome

Collectively, 19,534 (33.3%) of the 58,721 genes annotated by Ensembl were expressed across our sample series (count>10) (Supplementary Table 2). Remarkably, 17,335 (88.7%) were differentially expressed (fold change>2; FDR<0.01) in at least one stage by adjacent stage pairwise analyses. To examine developmental order, we performed principal component analysis (PCA) using all genes of moderate-to-high expression (9,734; Fragments Per Kilobase Per Million [FPKM] >1). The top three principal components effectively separated the sampled stages, while replicates of the same stage remained closely associated (Fig. 1c). Here, separation distances within the PCA map represent the extent to which developmental transitions are accompanied by major changes in transcript abundance. Notably, the stages of oocyte development (along with the pronuclear stage) co-localize along a short temporal arc, consistent with progressive but moderate changes in transcript abundance. In contrast, the cleavage-stage replicates were clearly distinct, consistent with new transcription after embryonic genome activation (EGA). An additional major change involves transition to the morula stage, which appears strikingly similar to trophectoderm replicates, whereas the ICM replicates form a distinct separate group. K-means algorithims were used to cluster genes based on their temporal expression and enrichment (Fig. 1d, Supplementary Table 3). Stage-specific gene sets pertaining to the immature egg (Cluster 1), cleavage (Cluster 4), and ICM (Cluster 7) stages were identified and contained genes of both known (e.g. *FIGLA*, *ZSCAN4*, and *NANOG*) and unknown specificity and developmental function.

Examination of alternative splicing and novel transcription

Overall, our transcription profiles were consistent with prior single cell datasets[15, 16] (Supplementary Fig. 1b). However, improvements in read coverage balance and directionality enabled the discovery of new novel transcription (Supplementary Fig. 1c, Supplementary Table 4) and splice isoform expression during pre-implantation development (Supplementary Fig. 1d–f, Supplementary Table 5). Together, these datasets yield extensive new information providing a major resource for future studies.

A *DUX4* binding motif is enriched upstream of cleavage-specific genes

We then addressed a key question in pre-implantation embryo development – what transcription factors drive stage-specific gene expression? To identify candidates, we performed *de novo* motif calling on the promoters of genes in clusters 1, 4, and 7 (Fig. 1e,

Supplementary Fig. 1g). The most highly enriched motif was associated with cluster 4 genes and matched the predicted binding site of a transcription factor known as DUX4 (p= 1e–11) (Fig. 1f). *DUX4* is one of three coding DUX (double homeobox) genes in humans, which also includes *DUXA* and *DUXB*[17]. The DUX gene family is a member of the paired (PRD)-like class of homedomains, that includes ARGFX, *LEUTX*, *DPRX*, and *TPRX1,* all of which show signs of rapid evolution/divergence and an involvement in human EGA[18–23].

DUX4 potently activates cleavage-specific genes and repetitive elements

DUX4 mRNA and protein are restricted to the 4-cell stage (early EGA) (Fig. 2a, Supplementary Fig. 2a) preceding the transient expression/enrichment of other 'PRD-like' genes during the 8-cell and morula stages (Supplementary Fig. 2b,c). To identify DUX4 transcriptional targets we overexpressed it in human induced pluripotent stem cells (iPSC) and performed RNA sequencing (RNA-seq). Compared to luciferase controls, induction of *DUX4* for 14 or 24hrs via dox administration led to significant differential expression (FC>2; FDR<0.01) of 163 and 193 genes, respectively (Supplementary Fig. 2d, Supplementary Table 6) –all of which were upregulated except one (*ZNF208*). Remarkably, as a group this gene set (which included notable DUX/PRD-like factors listed above) showed robust and transient expression in the cleavage stage embryo (Fig. 2b, Supplementary Fig. 2e).

The most highly activated gene was *ZSCAN4*, a defining cleavage-stage gene in both human and mouse[24]. Based on previous ChIP-sequencing data from human myoblasts (MB), *ZSCAN4* is directly bound by DUX4 and contains four distinct DUX4 binding sites. To test for direct DUX4 activity in embryonic stem cells (hESCs) we developed a luciferase reporter using the 2kb promoter (LP) sequence for *ZSCAN4* (Fig. 2c). Transient co-transfection with *DUX4* induced luciferase expression >2,000-fold. However, in contrast to prior work[22], transient co-transfection with *DUXA* had no effect. Omitting three of the four DUX4 binding sites (LP-3xmut) greatly reduced activation, whereas eliminating the proximal Alu elements (SP), previously implicated in *ZSCAN4* activation via DUXA[21, 22], had no affect. Thus, *ZSCAN4* activation is specifically controlled by the direct binding of DUX4 to its predicted binding sites.

In addition to activating gene expression, introduction of DUX4 also led to an increase in transcripts derived from ACRO1 and HSATII satellite repeats, which are also enriched in cleavage-stage embryos (Supplementary Fig. 2f,g). Most striking, however, was the strong induction of HERVL retrotransposons (Fig. 2d) which are selectively transcribed in the cleavage stage, consistent with previous findings[25]. In keeping with endogenous targets like *ZSCAN4*, DUX4 ChIP-sequencing (ChIP-seq) peaks in myoblasts are highly enriched in activated LTR and satellites repeats suggesting that the observed effects are direct[12, 13]. To confirm and extend, we repeated the DUX4 ChIP-seq experiment in human iPSCs post 24hr *DUX4* (or luciferase) expression. At standard statistical thresholds (qval<0.01), we observed more than 200,000 peaks (vs. control) shared between two technical replicates. At high thresholds (qval<10^{-20}) we observed 65,728 shared peaks- 50,674 (77%, p<1e–300) of which overlap with the 63,795 peaks previously identified in myoblasts (Supplementary Fig. 2h, Supplementary Table 7). Using GREAT[26], we next determined direct DUX4 targets. Of

the 739 cleavage-stages genes we identified, at least 25% (191, pval=0.01) were directly occupied by DUX4 in iPSCs; including those encoding prominent cleavage-stage transcription factors (TF), chromatin modifiers (CM), and post-translational modification enzymes (PTE) many of which were also markedly upregulated following *DUX4* expression in iPSCs (Fig. 2e, Supplementary Fig. 2i). Unique reads revealed significant DUX4 enrichment at activated LTR elements (e.g. MLT2A1, MLT2A2) and HSATII satellites (Supplementary Fig. 2j), consistent with prior findings and the notion of direct repeat element activation. Taken together, our work supports roles for DUX4 in direct activation of a transcriptional program at EGA which helps de-repress germ cell heterochromatin and coordinate gene expression for ensuing lineage decisions (Fig. 2f).

Functional conservation of *DUX* proteins in defining the cleavage stage transcriptome in mammals

As genetic tools and genomic datasets involving cleavage stage transcription and chromatin dynamics are only available for mouse, we turned here to test whether *DUX4* displays conserved and central roles in mammalian embryogenesis. Our analysis of prior RNA-seq datasets[27] revealed cleavage-stage specific transcription of a weakly conserved *DUX4* homolog in mouse, called *Dux*[28–30] (Fig. 3a, Supplementary Fig. 3a). Notably, *Dux* is transiently and specifically expressed in early 2-cell stage mouse embryos (Fig. 3a), one cell cycle earlier than *DUX4* expression in human embryos but consistent with the onset of EGA.

To test whether *Dux* expression can function as an early embryonic transcriptional activator, we initially expressed it in myoblasts and performed qRT-PCR. Like *DUX4*, *Dux* robustly activated the expression of key cleavage-specific genes such as *Zscan4*, *Zfp352*, and *Tcstv1* (Supplementary Fig. 3b). To extend these findings transcriptome-wide in a developmentally relevant cell-type, we next transfected mESCs with a dox-inducible *Dux* expression construct (codon altered to ensure robust expression). RNA-seq on a non-clonal population revealed the upregulation of 123 genes (FC>2, FDR<0.01) (Fig. 3b), including notable retrotransposons (e.g. MERVL and its LTR, MT2_Mm) with no genes being significantly downregulated (Supplementary Table 8). This cohort of differentially expressed genes is transiently and specifically expressed in the mouse cleavage-stage embryo (Fig. 3c) and contains several orthologs (e.g. *Zscan4*, *Prame*f, *Ubtfl1*, *Kdm4e*) of genes enriched in human cleavage stage, and directly activated by DUX4 in iPSCs. Thus, *Dux* appears to operate as a functional ortholog of *DUX4* in mouse, regulating gene expression during EGA.

Conversion of mESCs to ‘2C-like’ cells by *Dux* expression

We next tested whether *Dux* could convert mESCs to a state that resembles the 2-cell mouse embryo (‘2C-like’). ‘2C-like’ cells are a rare metastable subpopulation of mESCs previously identified and isolated by their spontaneous reactivation of MERVL, a murine-specific retrotransposon otherwise only expressed in the 2-cell stage mouse embryo[31–33] (Fig. 3d, top panel). Remarkably, MERVL reactivation in mESCs, revealed by the expression of a MERVL-linked fluorescent protein (MERVL::tdTomato or MERVL::GFP) is linked to the acquisition of molecular and functional features that are specific to the totipotent cleavage

embryo, including the expression of early embryonic (2C) genes[8], the loss of POU5F1, and the disaggregation and reformation of constitutive heterochromatin into chromocenters[9].

Accordingly, we find *Dux* (Fig. 3d, bottom panel) and DUX-induced genes strongly upregulated in MERVL-expressing cells (Fig. 3e). To evaluate whether *Dux* could drive conversion of mESCs to the '2C-like' state, we then stably integrated our dox-inducible *Dux* construct (or luciferase control) into MERVL::GFP reporter mESCs and expanded clonal cell lines (Fig. 3f, left panel). Using flow cytometry to count the number of GFP-positive (GFP^{pos}) cells post dox-induction (24hrs), we observed conversion efficiencies in *Dux*-expressing clones ranging from 10–74% GFP^{pos}, with the most efficient clone exhibiting a >500-fold increase compared to controls (Fig. 3f, middle panel). Live imaging fluorescent microscopy confirmed this observation (Fig. 3f, right panel) and further revealed dose dependency (Supplementary Fig. 3c).

Dox-induced cells were then either sorted by FACS into GFP^{neg} and GFP^{pos} populations, or left unsorted (versus 'no dox' control), and subjected to RNA-seq (Supplementary Fig. 3d). These two approaches yielded a highly significant overlap ($p<1e–300$) of differentially expressed genes (DEGs) resulting in the unbiased clustering of sorted and unsorted *Dux*-expressing cells (Supplementary Fig. 3e, Supplementary Tables 9, 10). Notably, *Dux* transgene RNA levels correlated with dox induction and with conversion to a GFP^{pos} state. Although transgene expression in the induced cells exceeded the expression of endogenous *Dux* RNA in spontaneously fluctuating '2C-like' cells (Supplementary Fig. 3f), the transcriptional profiles were highly similar ($r=0.78$) (Fig. 3g). Together, these data indicate DUX as a potent transcriptional activator of '2C-like' genes and retrotransposons (Supplementary Fig. 3g). To further determine whether *Dux* expression imposed other attributes of the '2C-like' state, we examined the status of POU5F1 protein and chromocenters. Here, our IHC results demonstrated a complete loss of POU5F1 (despite no change in mRNA) in GFP^{pos} cells, coinciding with the loss of chromocenters (Fig. 3h). Thus, *Dux* expression appears to elicit in mESCs multiple molecular/biological features of '2C-like' cells, implicating DUX as the driver of '2C-like' conversion.

Dux is necessary for induction of '2C-like' cells

Depletion of *Chaf1a*, the p150 subunit of the chromatin assembly factor 1 complex (CAF-1) (Supplementary Fig. 4a) also induces the conversion of mESCs to a '2C-like' state[9], prompting an examination of the relationship between CAF-1 and *Dux* in this process. To begin, we examined prior RNA-seq datasets of mESCs following CAF-1 depletion; this revealed striking *Dux* upregulation (11–18 fold) in CAF-1 depleted mESCs (Fig. 4a, top panel). Moreover, the downstream targets of DUX (determined in our *Dux* overexpression studies) composed the most highly activated genes in the CAF-1 depleted datasets (Fig. 4a, bottom and right panel; Supplementary Fig. 4b).

We next determined whether *Dux* was necessary for *Chaf1a* knockdown-mediated entry into a '2C-like' state. To test, we transfected mESCs containing the MERVL::GFP reporter with siRNA pools targeting *Dux* mRNA (si308 and si309) and/or a previously validated siRNA against *Chaf1a*. First, depletion of *Dux* alone (si308) was sufficient to reduce the spontaneous conversion of mESCs to a '2C-like' state (Supplementary Fig. 4c, left panel),

and we confirm prior results showing that depletion of *Chaf1a* alone leads to a >20-fold increase (Supplementary Fig. 4c, right panel). Interestingly, co-transfection of mESCs with siRNA against *Dux* and *Chaf1a* nearly abolished the inductive effect of *Chaf1a* knockdown alone (Fig. 4b). To examine the extent to which entry into the '2C-like' state was inhibited, we repeated the knockdowns and isolated RNA for sequencing. First, knockdown of *Chaf1a* alone greatly altered gene expression, resulting in the upregulation of 2,229 genes (FC>2, FDR<0.01) including *Dux* and other prominent '2C-like' genes and repetitive elements (Fig. 4c, Supplementary Fig. 4d, Supplementary Table 11). Moreover, co-depletion of *Chaf1a* and *Dux* prevented the activation of 605–824 (27–36%, with si309 or si308, respectively) of the original 2,229 upregulated genes including 123 of 422 '2C' genes induced by *Chaf1a* knockdown (~29%; hypergeometric probability p=2.1e–65) and notable '2C-like' genes and repetitive elements: *Zscan4*, *Zfp352*, *Tcstv3*, MERVL, and GSAT (Supplementary Fig. 4e–g). Based on this data, we defined the 824-gene cohort as '*Dux*-dependent' and the remaining 1404-gene cohort as '*Dux*-independent'. Remarkably, while the '*Dux*-independent' cohort lacks developmental stage enrichment, the '*Dux*-dependent' cohort is predominantly expressed in the 2-cell stage embryo (Supplementary Fig. 4h). Thus, conversion of mESCs to a '2C-like' state - either spontaneous or through CAF-1 knockdown – is dependent on *Dux* (Supplementary Fig. 4i).

Dux expression coverts the chromatin landscape of mESCs to one strongly resembling early 2-cell mouse embryos

New genomics methodologies, namely ATAC-seq, enable the determination of open versus closed chromatin genome-wide[34]. Cleavage stage chromatin undergoes extensive reorganization to facilitate EGA and the conversion of gametes into totipotent embryos, supported by the distinctive ATAC/chromatin profiles recently revealed in early 2-cell stage embryos[35]. To further characterize *Dux* function, we next tested whether its expression could convert the chromatin in mESCs to a landscape resembling that of an early 2-cell stage embryo. Accordingly, we performed ATAC-seq on sorted MERVL:: GFP^{pos} and MERVL:: GFP^{neg} cells post 24hrs dox-induced *Dux* expression. After calling peaks in each condition, regions of significantly different ATAC-sensitivity (log10 likelihood ratio > 3) were identified. Here, we identified 6,071 regions (>500bp in length) that gained ATAC signal in GFP^{pos} cells compared to GFP^{neg} cells (ATAC-gained) and 4,231 regions that lost ATAC signal (ATAC-lost) (Fig. 5a, Supplementary Table 12). Remarkably, not only did the ATAC signal in these regions resemble that seen in early embryos, but unbiased correlation clustering based on genome-wide ATAC-signal clustered the '2C-like' cells with early 2-cell stage (Supplementary Fig. 5a). In contrast to the 9,131 common peaks found primarily at gene promoters, the ATAC-gained regions were mostly in intergenic space (Fig. 5b), with the majority (64.5%, P<0.001) directly overlapping a MERVL element. Using metagene analysis, we show that *Dux*-induced '2C-like' cells exhibit extensive and specific opening of chromatin at MERVL elements, mimicking that of an early 2-cell stage embryo (Fig. 5c). To determine the number and precise location of the MERVL instances that become open following *Dux* expression, we re-analyzed our ATAC-seq analysis using only unique reads. Here, although the number of called ATAC-gained regions was severely reduced, a still significant fraction (27%, p<0.001) overlapped a MERVL element (Supplementary Fig. 5b). Furthermore, while the ATAC-gained regions were located near genes highly and

significantly expressed in '2C-like' cells, the regions that lost ATAC sensitivity were generally located near genes displaying moderate downregulation (Supplementary Fig. 5c). Taken together, these data demonstrate that *Dux*-induced '2C-like' cells acquire chromatin accessibility at MERVL elements, which are used specifically in 2-cell stage embryos to regulate the gene expression program at EGA.

DUX occupancy is strongly correlated with '2C' gene expression and open chromatin

To determine if the observed changes in gene expression and chromatin architecture in '2C-like' cells is due to direct DUX binding, we localized DUX in mESCs by ChIP-seq. As no ChIP-grade antibody for DUX is available, we created a 3xHA-tagged *Dux* expression construct and isolated a new clonal MERVL::GFP mESC line. As with earlier clones, our HA-tagged clone displayed high conversion efficiency (60% GFP^{pos} 24hrs post dox-induction) and expression of HA-*Dux* coincided with the acquisition of key '2C-like' features (Supplementary Fig. 3h,i). The HA ChIP-seq yielded ~19,000 peaks shared between two biological replicates over input (qval<0.05), occupying 3,881 genes highly enriched in the MGI gene expression signature 'Two-cell stage embryo" (Fig. 6a, Supplementary Fig. 6a). Importantly, many of the 3,881 DUX-occupied genes (~20%) were also activated following *Dux* overexpression in mESCs and were identified by prior studies as markers of the '2C' and '2C-like' state (Fig. 6b,c). Conservative analyses using unique reads revealed at least 53% of all MERVL-LTRs (MT2-Mm) and at least 37% of the regions that gain ATAC-sensitivity in '2C-like' cells are directly bound by DUX in mESCs (Supplementary Fig. 6b,c)

Using the top 10,000 peak summits based on enrichment score, we further identified a consensus DUX binding motif (Supplementary Fig. 6d), with the top hit (WGATTYAATCW) scoring an E-value of 2.0e–7234. Notably, this motif was highly enriched (adj. pvalue= 6.3e–102) in regions of gained ATAC-sensitivity following *Dux*-overexpression. Finally, we note a lack of DUX4 motif enrichment within MERVL-LTRs (MT2_Mm), and a minimal enrichment for a DUX motif within HERVL-LTRs (MLT2A1/2). This suggests that DUX4 orthologs, although functionally conserved, have evolved to be species-specific, perhaps in response to ERVs.

Discussion

Using new RNA-seq technologies, we generated improved transcriptional profiles of human oocytes and embryos during pre-implantation development. We then focused on the distinctive cleavage stage (2-cell, 4-cell, and 8-cell embryo), during which the embryonic genome becomes activated and the embryo achieves totipotency[36, 37]. Whether and how these two critical development events are interconnected and initiated are key unanswered questions. In humans and mice, a unique transcriptional program is activated at the onset of EGA and is firmly restricted to the cleavage stage of embryonic development. Here, our work reveals that many key genes within this transcriptional program are direct targets of a functionally conserved double homeobox retrogene called DUX4 in humans, and *Dux* in mice (collectively referred to here as the *DUX4*-family) (Fig. 7a).

As *DUX4*-family genes themselves must be expressed at EGA, they cannot be responsible for EGA initiation. Instead, our ATAC-seq data, along with prior work[35], strongly suggests roles in opening chromatin – which may be analogous to pioneer factors such as *Drosophila's* Zelda[38–40] – and further in selecting genes for activation during EGA (e.g. *ZSCAN4*, *KDM4E*, *ERVL*) that appear to regulate vital EGA-coupled molecular events. How the genes encoding *DUX4*-family transcription factors are themselves briefly activated during early cleavage stage is currently unknown. One possibility is that genome-wide DNA demethylation in the zygote, coupled with a lack of repressive heterochromatin at EGA, allows maternally loaded transcription factors a transient opportunity to activate. Related to this, recent work reports a brief uncoupling of CAF-1 mediated chromatin assembly with DNA synthesis in the early 2-cell embryo, which may reduce nucleosome occupancy in the genome (and/or generally de-repress heterochromatin) and allow a burst of *Dux* expression[9].

Despite clear functional conservation, DUX4 and DUX bear only modest sequence conservation, though both are intron-less and can be found in tandem arrays on multiple chromosomes[29]. One leading hypothesis suggests derivation of *DUX4* and *Dux* through independent retrotransposition events involving the ancient, intron-containing, *DUXC* gene, which has since been lost in both species[17, 28]. Subsequent duplication and divergence has resulted in multiple paralogs in both humans and mice (complicating genetic loss-of-function approaches). Here, the evolutionary pressure for *DUX4* and *Dux* to duplicate and diverge may originate from their co-option by endogenous retroviruses – as host fitness benefits from mutations that maintain activation of endogenous genes and avoid activation of the invading retrovirus.

Until now, the normal function of DUX4 (outside of FSHD pathology) was unclear, but its maintenance and expansion strongly suggests important fitness contributions. Notably, the double homeobox gene family (e.g. *DUXA*, *DUXB*, *DUXC*) origination aligns with the evolution of the placenta. Accordingly, these genes are both specific to placental mammals and are only expressed during (or just prior to) the first lineage decision indicating a likely role in these processes. Indeed, understanding the role of the ancestral *DUXC* gene in the embryo of other eutherian clades is of high interest, as it will help elucidate a specific function.

Taken together, this work may have significant implications for early embryo development (impacting human infertility and recurrent pregnancy loss), the reprogramming field, cancer biology, and FSHD. Our data supports a role for DUX4-family proteins in opening chromatin and driving the transcription of many key genes during cleavage, a stage with completely unrestricted developmental potential[41–43]. Notably, the ability of *Dux* expression to drive the vast majority of mESCs into a '2C-like' state raises the possibility of creating totipotent cells for mechanistic studies. Indeed, additional work with human cells to create a '4C-like' state is an important future direction, possibly by expressing *DUX4* along with other maternally-contributed factors. Regarding FSHD, as cleavage embryos resist the apoptosis conferred by *DUX4* expression in muscle cells, '4C-like' cell lines might provide mechanistic or therapeutic insights. Finally, DUX4 fusion proteins (that omit the C-terminus of DUX4) driven by the IGH enhancer have recently emerged as the leading cause of acute

leukemias in adolescents and young adults[44, 45], prompting need for a greater understanding of DUX4 biochemically and molecularly in normal and oncogenic circumstances.

Methods

General methods and statistical testing

No statistical methods were used to predetermine experimental sample size. All experiments were performed at least twice with at least two replicates per condition. All experiments were performed with biological replicates (separate cell cultures), except for the DUX4 ChIP-seq which was done with technical replicates (a single cell culture split prior to library preparation). All overlap statistics (venn diagrams) were determined by hypergeometric probability using a set ‘population size’ of 18,000.

Human oocyte and embryo sample collection

Germinal Vesicle (GV) stage oocytes were collected from IVF patients at the University of Utah and the Minnesota Center for Reproductive Medicine from October 2011 to February 2013. Enrollment was limited to patients who were undergoing IVF with Intra Cytoplasmic Sperm Injection (ICSI) procedures of their own accord. Metaphase I and metaphase II oocytes were collected from fifteen healthy women, aged 21–28, who were voluntarily enrolled for this study. Donors underwent an ovarian stimulation cycle, using a long agonist protocol, followed by oocyte retrieval. Pre-implantation embryos were donated to IRB-approved research by consenting patients at the Utah Center for Reproductive Medicine and the Minnesota Center for Reproductive Medicine. Each patient’s informed consent was reviewed and documented by two clinical investigators prior to their use in the study. No embryos were created for research purposes. In all cases, embryos were donated by patients ending their fertility treatments, and therefore the remaining embryos would otherwise have been discarded.

Human oocyte and embryo sample preparation

Within 3 hours of collection, GV, MI, and MII oocytes were completely denuded of their cumulous cells. Denuded oocytes were then stored in 10 uL of protein free media in slow freeze 250 uL straws and kept at −80C until RNA preparation. Likewise, embryos used for this study were cryopreserved according to standard IVF protocols. Prior to RNA preparation, the embryos were thawed and pooled according to developmental stage. Embryos that failed to survive the freeze-thaw procedures were discarded. Blastocyst stage embryos were hatched and, using laser microdissection, were manually separated into inner cell mass (ICM) and mural trophectoderm (Troph). RNA extraction from pooled oocytes and embryos was preformed using the Qiagen AllPrep kit®. All sample handling of embryonic stages, from retrieval through nucleic acid isolation, was conducted in clinical facilities by clinically-funded staff, separate from NIH/NCI/HCI funded facilities and personnel.

Human oocyte and embryo RNA-seq library preparation and sequencing

High-quality RNA (RIN>7) was extracted from all stages. Using the TotalScript RNA-Seq kit (Epicentre), two stranded libraries were prepared for each stage. This approach enabled low inputs (5ng of total RNA/reaction) and random hexamer priming to reduce polyA

transcript bias. Each RNA pool was split once prior to adapter ligation and then split again prior to PCR amplification, resulting in four technical replicates per developmental stage. Purified libraries were quantified on an Agilent Technologies 2200 TapeStation using a D1000 ScreenTape assay.

The molarity of adapter-modified molecules was defined by quantitative PCR using the Kapa Library Quant Kit (Kapa Biosystems). Individual libraries were normalized to 10 nM and equal volumes were pooled in preparation for Illumina sequence analysis. Sequencing libraries (25 pM) were chemically denatured and applied to an Illumina HiSeq paired-end flow cell using an Illumina cBot. Flowcells were then transferred to an Illumina HiSeq 2000 instrument and sequenced in 100bp paired-end mode.

Human oocyte and embryo RNA-seq data processing

Raw sequencing reads were aligned with Novoalign (Novocraft, Inc.) to an unmasked hg19 index [–r All 50]. Splice junction alignments were converted to genomic coordinates and low quality and non-unique reads were removed using Sam Transcriptome Parser (USeq; v8.8.8). Normalized gene and repeat element expression was calculated using DefinedRegionDifferentialSeq (USeq; v8.8.8) using a custom hg19 ensembl exon/rmsk table. Splice isoform quantification was determined using Sailfish V0.10.0 (Patro et al., 2014). Principal Component Analysis and Partition Clustering (using the Davies-Bouldin statistic) were performed using the Partek Genomics Suite (Partek Inc) based on log transformed FPKM values. Motif discovery and enrichment was evaluated using Homer (findMotifs.pl –start 2000 –end 2000). *De novo* motifs with a 'best match score' >0.70 were ranked based on enrichment (−log10pval) and plotted in R using ggplot2.

Human embryo immunofluorescence and imaging

Human embryos at the 1-cell stage, donated to research as described above, were thawed and cultured to the 2-cell, 4-cell, or 8-cell stage. Staining was preformed as described previously (Niakan and Eggan, 2013). Briefly, surviving embryos of high quality were fixed in 4% formaldehyde for 1hr at room temperature and then washed three times with 0.1% tween in PBS (PBST). Embryos were permeabilized and then blocked in 10% donkey serum in PBST (blocking buffer) for 1hr at room temperature before being placed in primary antibody (concentration 1:250) consisting of anti-DUX4 (ab124699) in blocking buffer and incubated overnight at 4°C. On the following day, the embryos were washed three times in PBST and then transferred to secondary antibody (concentration 1:1000) consisting of Alexa 488 Donkey Anti-rabbit (Life Technologies, A21203) in blocking buffer. Following a 1hr incubation at room temperature, the embryos were washed four times in PBST, with the last wash containing DAPI. Embryos were then place in microdroplets in a glass dish and immersed in oil for imaging. Images were collected at 40× magnification using the Nikon A1 confocal microscope.

Comparative analysis

RNA sequencing reads from Yan et al., 2013 (GSE36552) and Xue et al., 2013 (GSE44183) were downloaded from GEO and processed as described above. Single cell data for each developmental stage was merged. Relative read coverage graphs were generated using the

CollectRnaSeqMetrics application from Picard tools (Broad Institute). Exonic and novel transcription was estimated using the Sam2USeq application (USeq; v8.8.8) on the alignments from each stage. Regions of >1, >3, or >5 non-stranded read coverage were output to a BED file that was subsequently intersected with a BED file containing all known Ensembl, UCSC, and NONCODE v4 exons plus 500bp in both directions. Intersecting regions are reported as exonic transcription in base pairs. Non-intersecting regions are reported as novel transcription. Novel transcribed regions of enriched or reduced expression (relative to other stages) were subsequently called using MultipleReplicaScanSeq (USeq; v8.8.8).

Expression constructs

Codon-altered (CA) coding sequences for *DUX4*, *DUXA*, *Dux*, and luciferase were synthesized as custom gBlocks® from Integrated DNA Technologies (IDT Inc.). Fragments were then cloned into a dox-on lentiviral backbone containing a puromycin selectable marker; pCW57.1 (a gift from David Root, Addgene plasmid # 41393).

Human iPSC culture and generation of stable cell lines

Human induced pluripotent stem cells were grown on Matrigel in mTeSR1 (STEMCELL Technologies) with ROCK inhibitor (STEMCELL Technologies). To create stable lines, cells were incubated with an *DUX4* or luciferase lentivirus (MOI =5) for 16hrs. After two days of recovery, cells were split and plated on MEFs and cultured for three passages in the presence of puromycin. Resistant cells were then split again with dispase (to remove MEFs) and re-plated on matrigel.

Human iPSC RNA-seq

RNA-seq was preformed with biological replicates in a non-clonal human iPSCs containing either a dox-inducible *DUX4* or luciferase transgene. Briefly, after 14 or 24 hours of dox-induction, the cells were lysed in Trizol and RNA extracted using the Direct-zol™ RNA MiniPrep kit by Zymo Research. Intact poly(A) RNA was then purified from total RNA samples (100–500 ng) with oligo(dT) magnetic beads and mRNA sequencing libraries were prepared using the Illumina TruSeq kit (RS-122-2101, RS-122-2102) as per the kit protocol. Libraries were then quantified, pooled, and loaded onto the flowcell as described above and sequenced on an Illumina HiSeq 2500 instrument in 100bp, single-end mode. Raw sequencing reads were aligned to hg19 with Novoalign (Novocraft, Inc.) [–r All 50]. Splice junction alignments were converted to genomic coordinates and low quality and non-unique reads were removed using Sam Transcriptome Parser (USeq; v8.8.8). Differential gene and repeat element expression (*DUX4*/Luciferase) was determined using DefinedRegionDifferentialSeq (USeq; v8.8.8) using a custom hg19 ensembl exon/rmsk table.

Human iPSC ChIP-seq

The DUX4 ChIP-seq experiments in human iPSCs were performed as described previously in myoblasts (Geng et al., 2012). Briefly, iPSCs containing a dox-inducible *DUX4* or luciferase transgene were treated with dox for 18hrs prior to crosslinking in 1%

formaldehyde for 10 minutes. Cells were then lysed and chromatin was sonicated to generate DNA fragments of 150–600bp. Cellular debris was pelleted and the DNA was immunoprecipitated overnight at 4°C using a rabbit monoclonal anti-DUX4 antibody [E5-5] (ab124699). After reversing crosslinks, libraries were prepped using the NEBnext DNA Library Prep Kit (NEB, E7370L). Here, as the ChIP was performed in only a single biological replicate, two libraries per condition were made to provide technical replicates. Adapter ligated DNA was then size selected and purified using AMPure XP beads (Beckman Coulter). Libraries were quantified, pooled, and loaded onto the flowcell as described above and sequenced on an Illumina HiSeq 2500 instrument in 125bp, paired-end mode. Paired-end, raw read files were first processed by Trim Galore (Babraham Institute) to trim low quality reds and remove adapters. Processed reads were then aligned to hg19 using Bowtie2 (v2.2.6) with the following parameters: (-t –q –N1 –L 25 –X 2000 –no-mixed –no-discordant). Peaks were called in each technical replicate separately (over the DUX4 control ChIP in luciferase-expressing iPSCs) using MACS2 ‘callpeak’ (-f BAMPE –B -SPMR). Overlapping peaks identified in both replicates meeting the qval cutoff ($<10^{-20}$) were selected for further analysis. GREAT (McLean et al., 2010) was used to link DUX4 peak regions to annotated genes (Basal plus extension; proximal 5kb upstream, 1kb downstream, plus distal up to 15kb). Motif discovery and enrichment analyses were performed with the MEME suite tools (Mchanick and Bailey, 2011). To evaluate enrichment at repeat elements, alignment files were filtered using samtools (view –q 10) to remove lower quality, multi-mapping reads. Over-representation of particular repeat subfamilies was determined by comparing the observed number of instances overlapping a peak region against a background expectation estimated by generating 1000 shuffled datasets from the same peak region file. Significance was determined empirically.

Luciferase constructs and assay

The *ZSCAN4* luciferase constructs were prepared by amplifying a 1.9kb region containing the putative enhancer and promoter from genomic DNA. This fragment was then cloned into a pGL3-basic reporter vector upstream of the SV40 promoter (LP; long promoter). Two variants of this promoter sequence, one containing ~1kb 5’ truncation (SP; short promoter) and another containing three point mutations in three of the four 11bp DUX4 binding sites (LP-3xmut) were also created and cloned into separate pGL3 vectors. Luciferase assays were performed in H9 human Embryonic Stem Cells (hESCs) grown on matrigel in mTeSR1 (STEMCELL Technologies) with ROCK inhibitor (STEMCELL Technologies). Briefly, each reporter vector was separately and transiently transfected into cells along with a *GFP*, *DUXA*, or *DUX4* expression construct. After recovery, the cells were treated with doxycycline for 24hr to induce transgene expression; verified by western blot. Finally, cells were lysed and the luciferase intensity was measured using the Dual-luciferase™ Reporter Assay from Promega. This experiment was performed twice with each condition repeated in quadruplicate.

Myoblast cell culture and generation of stable cell lines

C2C12 mouse myoblast cells (ATCC) were grown in DMEM with 10% fetal bovine serum (FBS) and Pen-strep. Stable cells lines were made by transfecting linearized *Dux* or

luciferase plasmids using Lipofectamine 2000 (ThermoFischer). After recovery, cells were selected with Puromycin (10mg/ml) for five days before picking and expanding clones.

Real-Time RT-qPCR

Briefly, cells were induced with 2ug/ml doxycycline for 36hrs before isolating RNA using the Clontech RNA Isolation kit. RT was performed using SuperScript III (Invitrogen) with oligo(dT) (Invitrogen) and qPCR was preformed with iTaq Universal SYBR Green Supermix (Bio-Rad). Experiments were performed in biological triplicate. Expression levels were normalized to *Timm17b* by DeltaCT. Primer sequences available in Supplementary Table 14.

Mouse ES cell culture and generation of stable cell lines

Mycoplasma-free E14 mESCs were cultured on gelatin in '2i' media containing PluriQ™ ES-DMEM medium with non-essential amino acids, B-mercaptoethanol, and dipeptide glutamine and supplemented with 15% ES-grade FBS, Primocin, leukemia inhibitory factor (ThermoFischer), 1mM PD0325901 (Sigma-Aldrich) and 3mM CHIR99021 (Sigma-Aldrich). Stable cells lines were made by transfecting linearized *Dux* or luciferase plasmids using Lipofectamine 2000 (ThermoFischer). After recovery, cells were selected with Puromycin (10mg/ml) for five days before picking and expanding clones. All cell lines were kept under constant drug selection with Puromycin and G418 to prevent transgene silencing.

Fluorescence-activated cell sorting

Quantification of GFP-positive cells was performed using a Cytek DxP Analyzer and data was processed in Flow Jo. For sorted RNA-seq and ATAC-seq experiments, a FACSAris Cell Sorter (BD Biosciences) was used to sort GFP-positive and negative cells prior to library preparation.

Mouse ESC RNA-seq

As described in the text, four different RNA-seq experiments were performed on mESCs. All experiments were done with two biological replicates. The first experiment looked at the effects of *Dux* expression in a non-clonal cell line containing the *Dux* transgene (+dox/−dox). The second experiment was performed similarly, but was done in a clonal cell line bearing the MERVL::GFP reporter. The third experiment used the same clonal cell line; however, cells were sorted into GFP^{pos} and GFP^{neg} subpopulations after dox-induction. The fourth experiment involved a different cell line that did not contain the *Dux* transgene. Here, we used siRNAs to test the requirement for *Dux* in activating '2C-like' gene expression. In all experiment, cells were lysed in Trizol and RNA was extracted using the Direct-zol™ RNA MiniPrep kit by Zymo Research. Intact poly(A) RNA was purified and were libraries prepared and sequenced on an Illumina HiSeq 2500 instrument as described above. With the exception of the first experiment, which was done in a single-end 50bp format, libraries were sequenced in a 125bp paired-end format. Raw sequencing reads were aligned to mm10 with Novoalign (Novocraft, Inc.) [–r All 50]. Splice junction alignments were converted to genomic coordinates and low quality and non-unique reads were removed using Sam Transcriptome Parser (USeq; v8.8.8). Differential gene and repeat element expression was

determined using DefinedRegionDifferentialSeq (USeq; v8.8.8) using a custom mm10 ensembl exon/rmsk table. *Dux* transgene RNA levels were determined by re-aligning each dataset to an index file of the codon-altered (CA) sequence.

Mouse Embryo RNA-seq data

Processed RNA-seq expression data from pre-implantation mouse embryos was downloaded from Deng et al., 2014 (GSE45719). To identify stage-specific gene expression, RPKM values were averaged across all single cells for the zygote, 2-cell, 4-cell, 8 cell, 16-cell, and blastocyst stages. Genes with an average expression ≥1 RPKM in at least one developmental stage were then clustered into 10 k-means after z-score transformation. Ensembl BioMart was used to retrieve Ensembl gene IDs for overlap comparisons.

Mouse ESC ATAC-seq

The ATAC-seq libraries were prepared as previously described (Buenrostro et al., 2013) on ~30k sorted GFP^{pos} and GFP^{neg} mESCs after 24 hours of dox-induction (2 biological replicates per condition). Immediately following FACS, the cells were lysed in cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630) and the nuclei were pelleted and resuspended in Transposase buffer. The Tn5 enzyme was made in house and the transposition reaction was carried out for 30 minutes at 37°C. Following purification, the Nextera libraries were amplified for 12 cycles using the NEBnext PCR master mix and purified using the Qiagen PCR cleanup kit. All libraries were sequenced on the Illumina HiSeq 2500 platform in a 125bp, paired-end format. Paired-end, raw read files were first processed by Trim Galore (Babraham Institute) to trim low quality reds and remove adapters. Processed reads were then aligned to mm10 using Bowtie2 (v2.2.6) with the following parameters: (-t –q –N1 –L 25 –X 2000 –no-mixed –no-discordant). ATAC-seq peaks were called using MACS2 ‘callpeak’ (-B --nomodel --nolambda --shift −100 --extsize 200), generating replicate-merged bedgraph files. Subsequently, the ‘bdgdiff’ subcommand (−l 500 –g 250) was used to call “differential peaks” between the two conditions (GFP^{pos} and GFP^{neg}). For comparisons to the pre-implantation mouse embryo, data from Wu et al., 2016 was downloaded from GEO (GSE66390) and re-processed as described above. Biological replicates were aligned independently and merged in MACS2. The Galaxy deeptools suite (Afgan et al., 2016) was used to plot heatmaps and metagene profiles. ChIPSeeker (Yu, Wang, and He, 2015) was used to determine overlap with genomic features. To determine the number and location of MERVL instances bound, alignment files were first filtered using samtools (view –q 10) to remove low quality, multi-mapping reads. After calling differential peaks as described above, bedtools intersect was used to report the overlap of each peak region file with MERVL instances. Significance was determined empirically comparing the observed overlap to a background expectation estimated by shuffling each peak region dataset 1000 times and performing an intersect.

Mouse ESC ChIP-seq

In order to investigate DUX binding, an N-terminal HA-epitope tag was added to our *Dux* expression construct and selected/expanded a new clonal cell lines. This experiment was performed in biological replicate. In short, mESCs were treated with doxycycline for 18hrs

to induce (HA) *Dux* expression. Cells were then cross-linked with 1% formaldehyde for 10 minutes prior to being lysed for DNA extraction. Chromatin was sonicated using the BioRuptor® system (Diagenode). Cellular debris was pelleted and the DNA was precipitated overnight at 4°C using a ChIP Grade Anti-HA tag antibody (Abcam, ab9110). After reversing crosslinks, libraries were prepped using the NEBnext DNA Library Prep Kit (NEB, E7370L). Adapter ligated DNA was size selected and purified using AMPure XP beads (Beckman Coulter, A63881) before sequencing on the Illumina HiSeq 2500 platform in 125bp, paired-end format. As before, raw read files were first processed by Trim Galore (Babraham Institute) to trim low quality reds and remove adapters. These processed reads were then aligned to mm10 using Bowtie2 (v2.2.6) with the following parameters: (-t –q –N1 –L 25 –X 2000 –no-mixed –no-discordant). Peaks were called in each biological replicate separately (over input DNA) using MACS2 'callpeak' (-f BAMPE –B -SPMR). Overlapping peaks identified in both replicates meeting the qval cutoff (<0.05) were then selected for further analysis. GREAT was used to link (HA) DUX peak regions to annotated genes (Basal plus extension; proximal 5kb upstream, 1kb downstream, plus distal up to 15kb). Motif discovery and enrichment analyses performed using the MEME suite tools. To evaluate enrichment at repeat elements, alignment files were filtered using samtools (view –q 10) to remove lower quality, multi-mapping reads. Over-representation of particular repeat subfamilies was determined by comparing the observed number of instances overlapping a peak region against a background expectation estimated by generating 1000 shuffled datasets from the same peak region file. Significance was determined empirically.

Immunofluorescence and imaging

Cells were plated on gelatin coated coverslips and allowed to adhere for 3–5 hours before fixing in 4% paraformaldehyde in PBS for 10 minutes at room temperature. Subsequently, the cells were permeabilized in 0.1% Triton-X-100 in PBS for 10 minutes at room temperature and then blocked in 3% BSA in PBS for 1 hour at room temperature. Primary antibodies (see below) were diluted in 3% BSA and the cells were incubated for 1 hour at room temperature. Cells were then washed and incubated in diluted Alexa-conjugated secondary antibodies plus DAPI (4′,6-diamidino-2-phenylindole) for 1 hour at room temperature before mounting. Imagining was done on a Nikon A1 confocal microscope. Simple fluorescence images of 2C:GFP cells were collected on the EVOS™ FL cell imaging system and quantitative live-cell capture and analysis using the IncuCyte® ZOOM system. Primary antibodies to the following proteins were used: Anti-GFP (abcam, ab13970), Anti-Oct3/4 (Santa Cruz Biotechnology, sc-5279). Secondary antibodies included an Alexa 488 Goat Anti-Chicken (Thermo Scientific, A11039) and an Alexa 594 Donkey Anti-Mouse (Life Technologies, A21203).

siRNA generation and transfection

Chaf1a (s77588) and negative control Silencer Select siRNAs were purchased from LifeTechnologies. *Dux* siRNA pools were generated using Giardia Dicer. Briefly, primers were designed to amplify two ~400bp fragments of the endogenous *Dux* locus from genomic mouse DNA and add T7 handles (see Supplementary Table 14). Purified PCR products were then used as template for *in vitro* transcription using the MEGAscript® T7 Transcription Kit (ThermoFischer, AM1334). Template DNA was then degraded and the

ssRNA allowed to anneal before dicing. Diced siRNAs were purified using the PureLink™ Micro-to-Midi Total RNA purification Kit (Invitrogen, 12183-018) with modifications. siRNA concentration was measured with the Qubit® RNA HS Assay Kit (ThermoFisher, Q32852). mESCs containing the MERVL:GFP reporter were transfected with 20pmol (10pmol of each) of total siRNA using RNAiMax (Life Technologies). All siRNA transfections were performed twice (on back to back days) to ensure knockdown.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank S. Kuerten (NuGen Inc.) for help preparing the RNA-seq libraries, B. Dalley for sequencing services, and T. Parnell for bioinformatics assistance. Special thanks to M-E. Torres-Padilla (IGBMC) for generously gifting the MERVL::GFP reporter mESC line. Functional genomics work was supported by HHMI. J. Doráis was further supported by Eunice Kennedy Shriver NIH NICHD K12HD000849. S. Tapscott and J-W. Lim were supported by NIH NIAMS R01AR045203, NIH NINDS P01NS069539, and the Friends of FSH Research. J. Whiddon was supported by the National Science Foundation Graduate Research Fellowship Program DGE-1256082 and the University of Washington Interdisciplinary Training in Genome Sciences grant T32 HG00035 from NHGRI. Finally, we acknowledge CA042014 for support of the University of Utah cores facilities.

References

1. Liu L, et al. Telomere lengthening early in development. Nat Cell Biol. 2007; 9:1436–1441. [PubMed: 17982445]
2. Matoba S, et al. Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. Cell. 2014; 159:884–895. [PubMed: 25417163]
3. Chung YG, et al. Histone Demethylase Expression Enhances Human Somatic Cell Nuclear Transfer Efficiency and Promotes Derivation of Pluripotent Stem Cells. Cell Stem Cell. 2015; 17:758–766. [PubMed: 26526725]
4. Zalzman M, et al. Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature. 2010; 464:858–863. [PubMed: 20336070]
5. Kalmbach K, Robinson LG, Wang F, Liu L, Keefe D. Telomere Length Reprogramming in Embryos and Stem Cells. Biomed Res Int. 2014; 2014:925121. [PubMed: 24719895]
6. Macfarlan TS, et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. 2011; 25:594–607. [PubMed: 21357675]
7. Gifford WD, Pfaff SL, Macfarlan TS. Transposable elements as genetic regulatory substrates in early development. Trends Cell Biol. 2013; 23:218–226. [PubMed: 23411159]
8. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012; 487:57–63. [PubMed: 22722858]
9. Ishiuchi T, et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat. Struct. Mol. Biol. 2015; doi: 10.1038/nsmb.3066
10. Eckersley-Maslin MA, et al. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell Rep. 2016; 17:179–192. [PubMed: 27681430]
11. Choi YJ, et al. Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. Science. 2017; :aag1927.doi: 10.1126/science.aag1927
12. Geng LN, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. Dev Cell. 2012; 22:38–51. [PubMed: 22209328]
13. Young JM, et al. DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLoS Genet. 2013; 9:e1003947. [PubMed: 24278031]
14. Gertz J, et al. Transposase mediated construction of RNA-seq libraries. Genome Res. 2012; 22:134–141. [PubMed: 22128135]

15. Xue Z, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013; doi: 10.1038/nature12364
16. Yan L, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 2013; doi: 10.1038/nsmb.2660
17. Leidenroth A, Hewitt JE. A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol. Biol. 2010; 10:364. [PubMed: 21110847]
18. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. BMC Biol. 2007; 5:47. [PubMed: 17963489]
19. Bürglin TR, Affolter M. Homeodomain proteins: an update. Chromosoma. 2016; 125:497–521. [PubMed: 26464018]
20. Dunwell TL, Holland PWH. Diversity of human and mouse homeobox gene expression in development and adult tissues. BMC Dev Biol. 2016; 16:40. [PubMed: 27809766]
21. Madissoon E, et al. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep. 2016; 6:28995. [PubMed: 27412763]
22. Töhönen V, et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun. 2015; 6:8207. [PubMed: 26360614]
23. Jouhilahti E-M, et al. The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. Development. 2016; 143:3459–3469. [PubMed: 27578796]
24. Ko, MSH. Mammalian Preimplantation Development. Vol. 120. Elsevier; 2016. p. 103-124.
25. Göke J, et al. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. Cell Stem Cell. 2015; 16:135–141. [PubMed: 25658370]
26. McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]
27. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014; 343:193–196. [PubMed: 24408435]
28. Leidenroth A, et al. Evolution of DUX gene macrosatellites in placental mammals. Chromosoma. 2012; 121:489–497. [PubMed: 22903800]
29. 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. 2007; 81:264–279. [PubMed: 17668377]
30. Eidahl JO, et al. Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. Hum. Mol. Genet. 2016; :ddw287.doi: 10.1093/hmg/ddw287
31. Schoorlemmer J, Pérez-Palacios R, Climent M, Guallar D, Muniesa P. Regulation of Mouse Retroelement MuERV-L/MERVL Expression by REX1 and Epigenetic Control of Stem Cell Potency. Front. Oncol. 2014; 4
32. Kigami D, Minami N, Takayama H, Imai H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biology of Reproduction. 2003; 68:651–654. [PubMed: 12533431]
33. Ribet D, et al. Murine Endogenous Retrovirus MuERV-L Is the Progenitor of the ‘Orphan’ Epsilon Viruslike Particles of the Early Mouse Embryo. Journal of Virology. 2008; 82:1622–1625. [PubMed: 18045933]
34. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth. 2013; 10:1213–1218.
35. Wu J, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature. 2016; 534:652–657. [PubMed: 27309802]
36. Zhou L-Q, Dean J. Reprogramming the genome to totipotency in mouse embryos. Trends Cell Biol. 25:82–91.
37. Ishiuchi T, Torres-Padilla M-E. Towards an understanding of the regulatory mechanisms of totipotency. Curr Opin Genet Dev. 2013; 23:512–518. [PubMed: 23942314]

38. Harrison MM, Li X-Y, Kaplan T, Botchan MR, Eisen MB. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. 2011; 7:e1002266. [PubMed: 22028662]

39. Sun Y, et al. Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. Genome Res. 2015; 25:1703–1714. [PubMed: 26335633]

40. Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. Genes Dev. 2014; 28:2679–2692. [PubMed: 25512556]

41. Morgani SM, Brickman JM. The molecular underpinnings of totipotency. Philos Trans R Soc Lond, B, Biol Sci. 2014; 369:20130549–20130549. [PubMed: 25349456]

42. De Paepe C, Krivega M, Cauffman G, Geens M, Van de Velde H. Totipotency and lineage segregation in the human embryo. Molecular Human Reproduction. 2014; 20:599–618. [PubMed: 24699365]

43. Borsos M, Torres-Padilla M-E. Building up the nucleus: nuclear organization in the establishment of totipotency and pluripotency during mammalian development. Genes Dev. 2016; 30:611–621. [PubMed: 26980186]

44. Yasuda T, et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. Nat Genet. 2016; 48:569–574. [PubMed: 27019113]

45. Zhang J, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet. 2016; 48:1481–1489. [PubMed: 27776115]

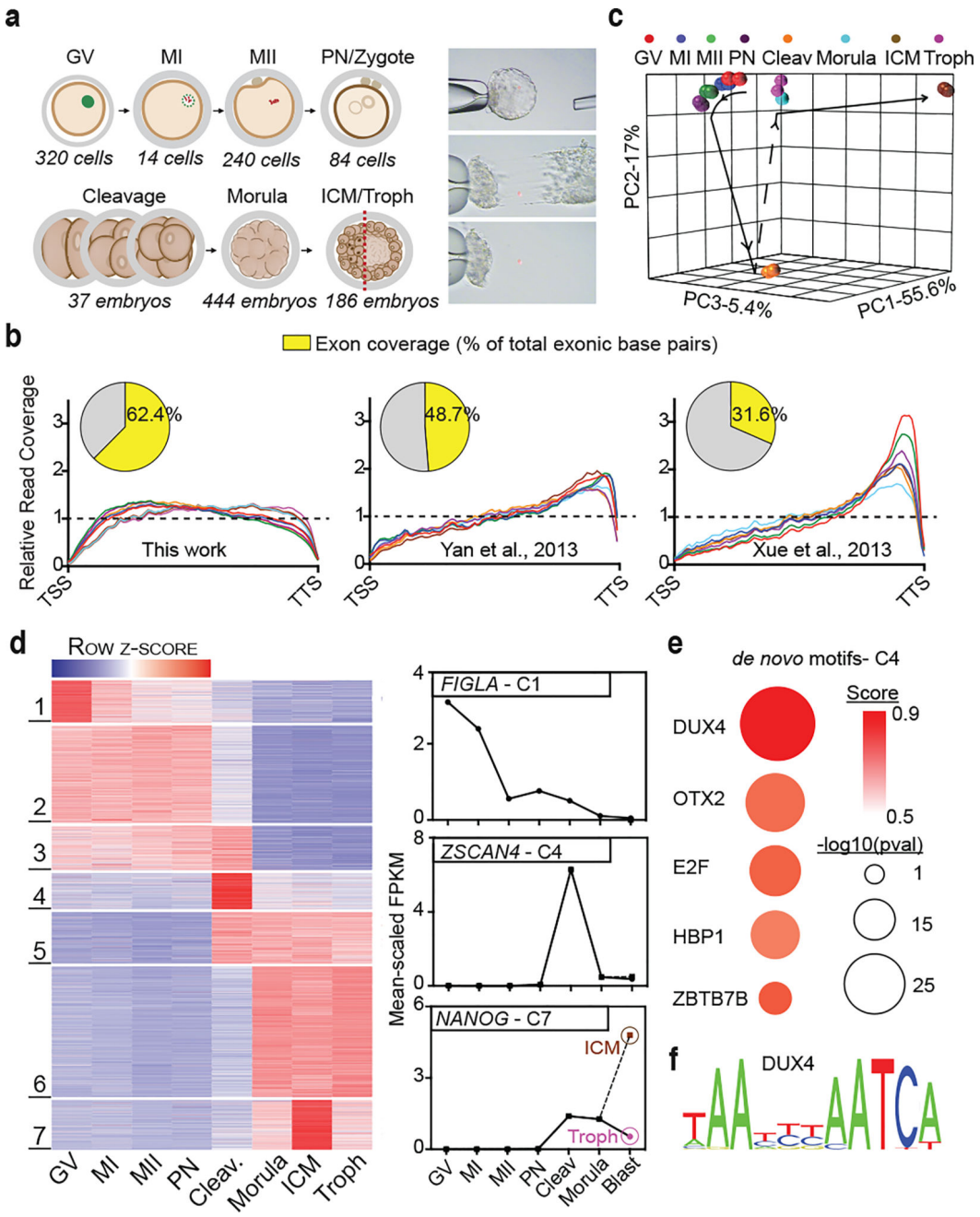

Figure 1. Improved RNA-sequencing methods reveal new novel transcription, dynamic splice isoform expression, and stage-specific gene expression in human oocytes and pre-implantation development

(a) Summary of the human oocyte and embryonic stages (and cell numbers) collected (left panel), and depiction of the laser mechanical separation of day 5–6 blastocysts into ICM and mural trophectoderm (right panel). (b) Metagene comparison of relative read coverage (from TSS to TTS) in this work and prior studies; each line represents a single developmental stage. Inset pie charts display the corresponding fraction of total exon bases covered by RNA-seq reads. (c) Principal component analysis (PCA) of all egg and embryonic stages based on the highest 50% of all expressed genes (>1 mean FPKM). (d) Statistically

determined k-means clusters based on the highest 50% all expressed genes (left panel). Clusters 1, 4, and 7 exhibit stage-specific gene expression and contain prominent developmentally important genes, *FIGLA*, *ZSCAN4*, and *NANOG*, respectively (right panel). (e) The top five *de novo* motifs enriched in cluster 4 (C4) gene promoters (pre-filtered for 'best match' score >0.70). Score- depicted here by color- indicates how strongly the discovered motif matches a known TF binding site. (f) The predicted binding site for DUX4.

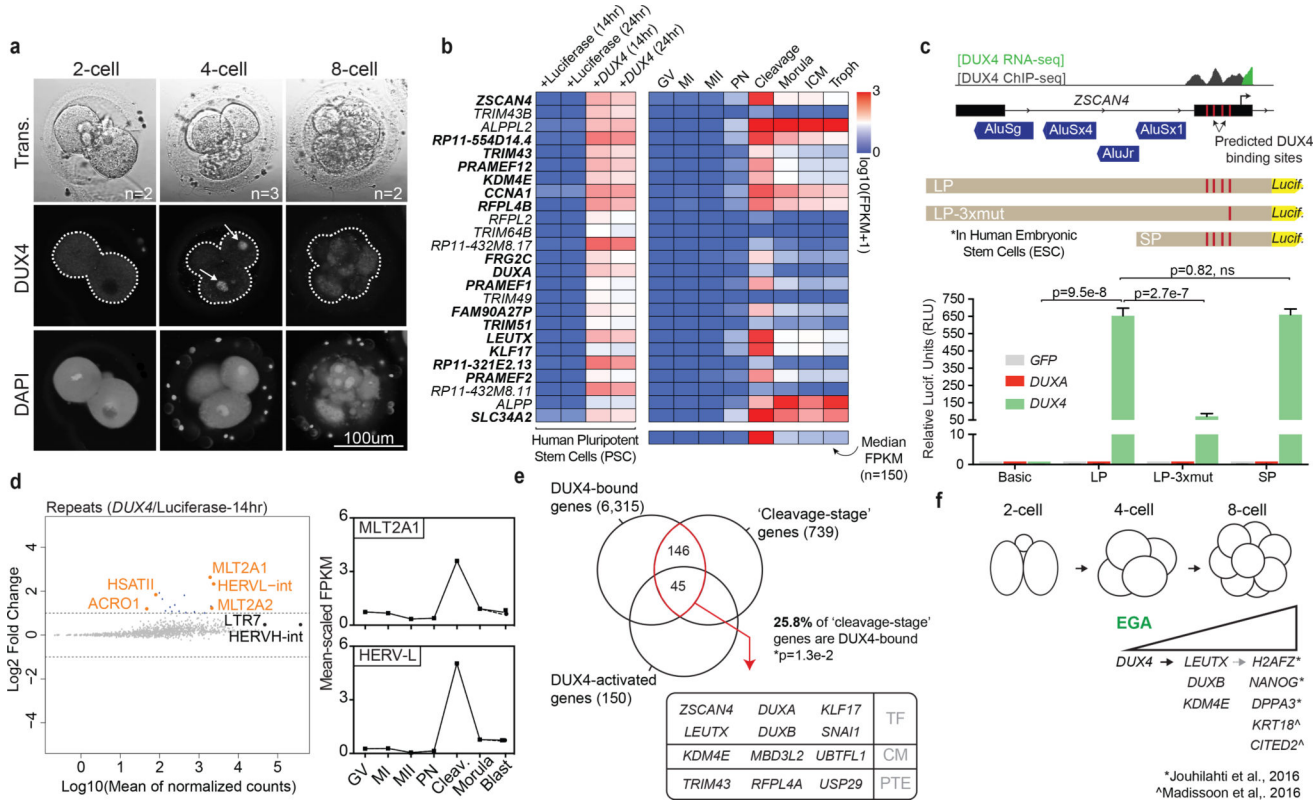

Figure 2. DUX4 directly activates the genes and repeat elements that are transiently expressed during human cleavage stage

(a) Immunofluorescence of DUX4 protein in human 2-cell, 4-cell and 8-cell embryos (n=7). (Note: though only one plane is shown, expression was restricted to nuclei of the 4-cell stage, indicated with arrows). (b) Heatmap depicting the top 25 DUX4-activated genes in human iPSCs and their expression in the embryo [*two replicates per condition*]. Bold font indicates genes belonging to cluster 4 (see Fig. 1d). The bottom row of the heatmap depicts the median embryonic expression of all 150 genes upregulated following *DUX4* expression. (c) A diagram of the *ZSCAN4* promoter/TSS and the position of the DUX4 ChIP occupancy in *DUX4*-expressing myoblasts (top panel). *ZSCAN4* activation is dependent on DUX4 binding (bottom panel) [four biological replicates per condition. *Statistics determined using a two-tailed unpaired t-test. Error bars, s.d.*]. (d) MA-plot showing *DUX4*-mediated induction of specific repeat elements, by subfamily (left panel). Mean-scaled expression of top activated repeats, HERVL and MLT2A1 in human oocytes and embryos (right panel). (e) The overlap of DUX4-ChIP occupied genes [*two replicates*] with genes enriched in the cleavage-stage embryo and activated by *DUX4*-overexpression in iPSCs [*Overlap statistic calculated by hypergeometric test. Note - only 477 of 739 'cleavage genes' were annotated in GREAT*]. In the box, genes encoding notable transcription factors (TF), chromatin modifiers (CM), and post-translational modifying enzymes (PTE) in the overlapping population are listed. (f) Diagram summarizing the timing of *DUX4* expression and its effects on embryonic gene expression.

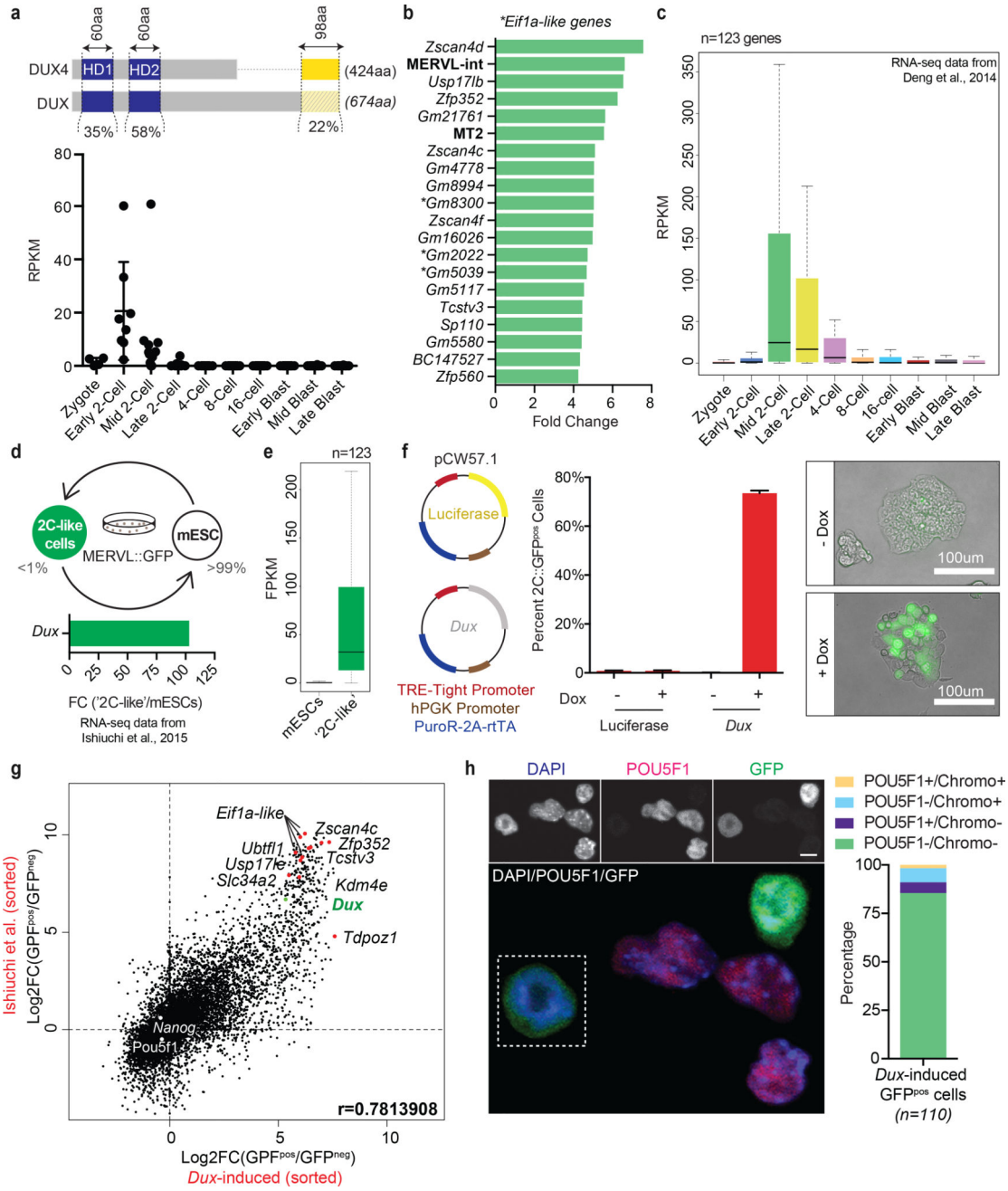

Figure 3. Mouse *Dux*, a functional ortholog of human *DUX4*, activates a '2C' transcriptional program and converts mESCs to a '2C-like' state

(a) Sequence level comparison of DUX4 and DUX (top panel) and the normalized expression of *Dux* in pre-implantation mouse embryos (RNA-seq data from Deng et al., 2014) (bottom panel). (b) Bar graph displaying the top 15 differentially-expressed genes and repeat elements (bold) following ectopic *Dux* expression in mouse embryonic stem cells (mESCs) [*two replicates per condition*]. (c) Relative expression of *Dux*-induced genes (n=123) in the pre-implantation mouse embryo. (d) Diagram of mESC metastability (top panel) and the enrichment of *Dux* in '2C-like' cells relative to conventional mESCs (bottom panel). (e) Expression of *Dux*-induced genes (n=123) in '2C-like' cells compared to

conventional mESCs. (f) Diagram of doxycycline-inducible lentiviral constructs stably integrated into mESCs (left panel) and their effect (after 24hrs of dox administration) on MERVL::GFP reporter expression evaluated by flow cytometry (middle panel) and live imaging microscopy (right panel) [*four biological replicates per condition. Error bars, s.d*]. (g) Dot plot showing per gene differential expression in *Dux*-induced MERVL::GFP^{pos} cells (over MERVL::GFP^{neg} cells), x-axis; compared with per gene differential expression observed in spontaneously converting '2C-like' cells, y-axis. (h) Immunofluorescence quantifying the loss of pluripotency (e.g. POU5F1 protein) and chromocenters in mESCs following ectopic *Dux* expression (n =110 cells). Scale bar, 10um.

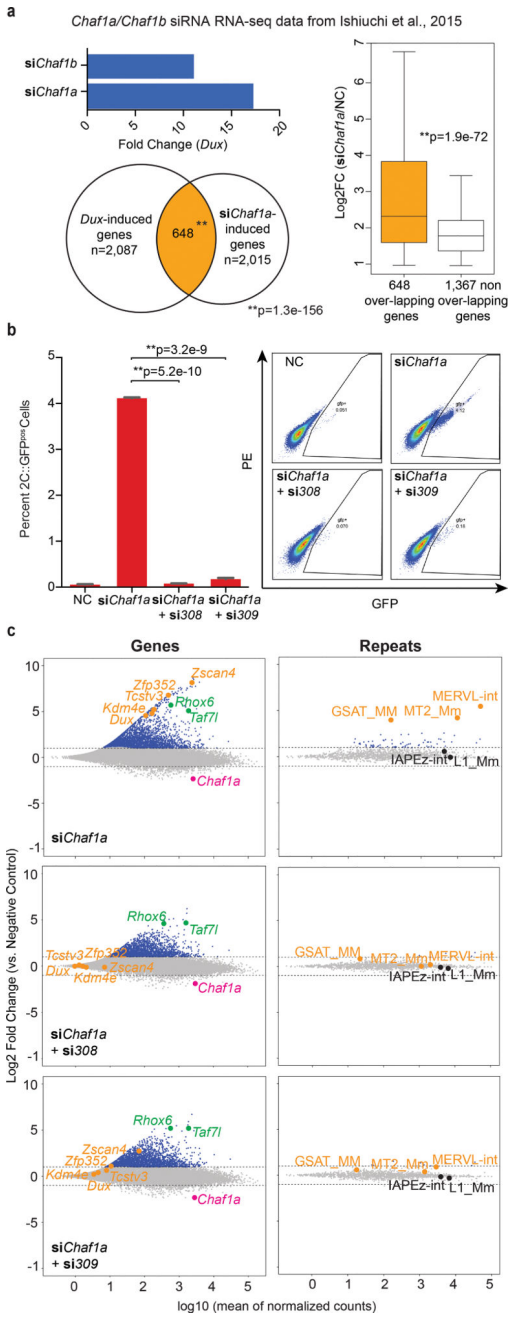

Figure 4. *Dux* is necessary for spontaneous and CAF-1 mediated conversion of mESCs to a '2C-like' state

(a) *Dux* is highly upregulated in CAF-1 depleted mESCs (top). Venn diagram displays large overlap of *Dux*-induced genes with genes activated in *Chaf1a*-depleted mESCs (bottom) [*Overlap statistic calculated by hypergeometric test*]. DUX target genes display significantly higher induction than non-targets in *Chaf1a*-depleted mESCs (right) [*Statistics determined using a one-tailed unpaired t-test.*] (c) Flow cytometry quantifies the percentage of GFP^{pos} cells following *Chaf1a* knockdown alone (si*Chaf1a*) and in combination with *Dux* knockdown (si308 or si309) [*three biological replicates per condition. Statistics determined using a two-tailed unpaired t-test. Error bars, s.d*]. (c) MA-plots show changes in gene and

repeat element expression (by subfamily) in mESCs following knockdown of *Chaf1a* alone (top panel) and in combination with *Dux* (si308-middle panel; si309-bottom panel).

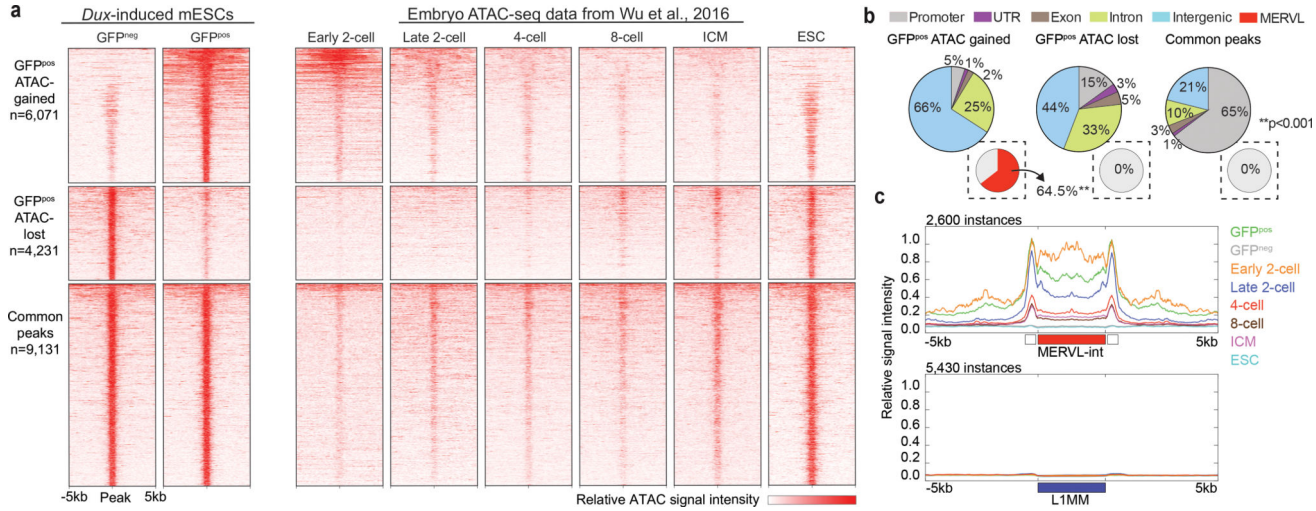

Figure 5. *Dux*-induced '2C-like' cells acquire an open chromatin landscape that resembles an early 2-cell stage embryo

(a) Heatmaps display regions of ATAC-seq signal gain, loss, and found in common between *Dux*-induced GFP^{pos} and GFP^{neg} cell populations [*Two replicates per condition*]. *Dux*-induced GFP^{pos} cells acquire an open/closed chromatin landscape that resembles the early 2-cell stage embryo (Embryo ATAC-seq data from Wu et al., 2016). (b) Pie charts depicting the distribution of ATAC-seq gained, lost and common peaks at basic genomic features. Inset pie charts indicate the percentage of peaks that overlap with MERVL elements (MT2_Mm and MERVL-int) [*Enrichment statistic determined empirically*]. (c) Metagene analysis of ATAC-seq signal across all MERVL-int instances (top panel) and L1 instances (bottom panel) in *Dux*-induced GFP^{pos} and GFP^{neg} cells and the early embryo.

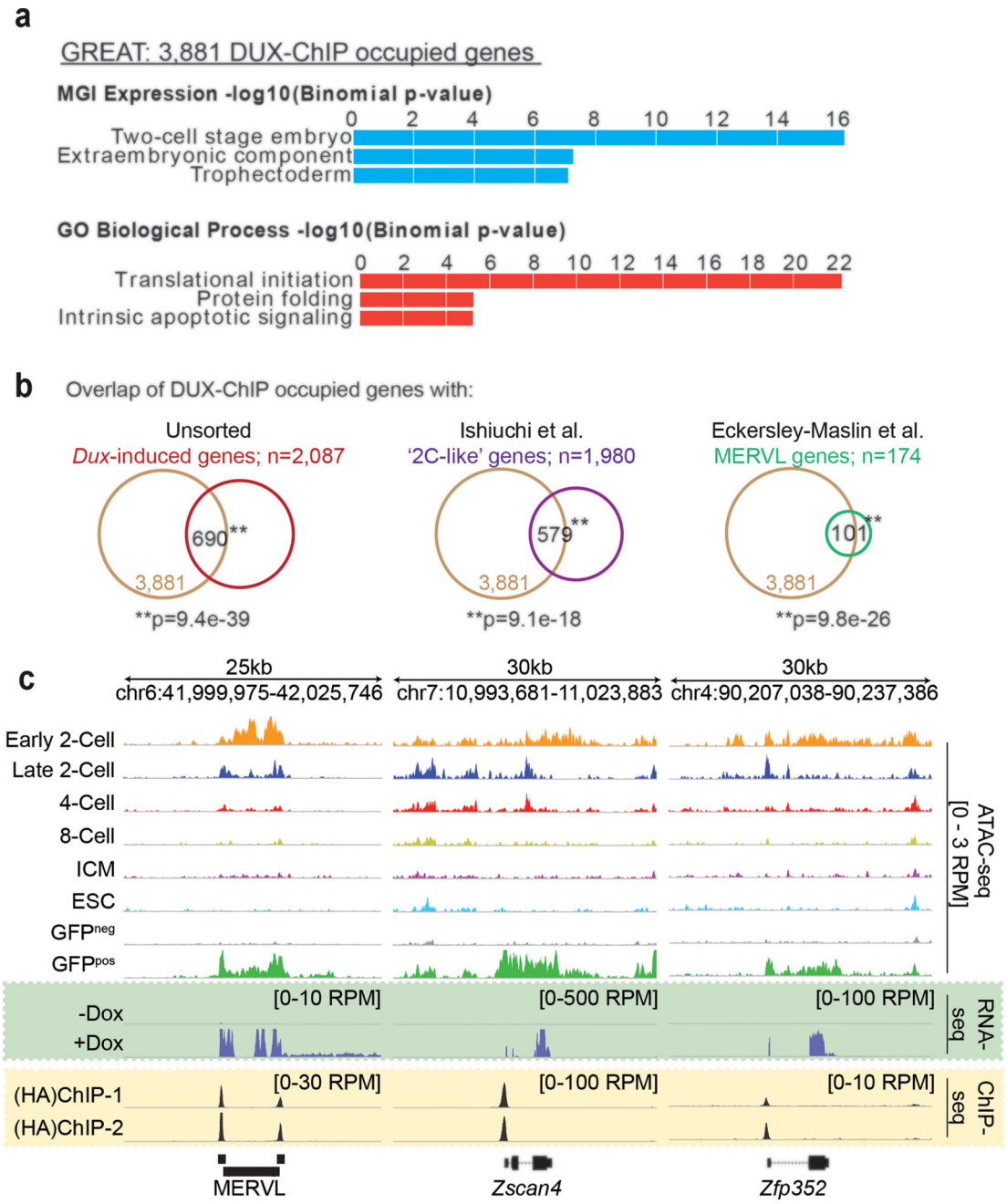

Figure 6. DUX binds directly to '2C' gene promoters and retrotransposons

(a) Top enriched 'MGI expression' and 'Gene Ontology (GO)' terms identified in the 3,881 genes bound by DUX [*two replicates*]. (b) Overlap of DUX-ChIP occupied genes with genes: upregulated in unsorted mESCs post *Dux* overexpression (left panel); enriched in '2C-like' cells (middle panel); and driven by MERVL elements (right panel) [*Statistics determined by hypergeometric test*]. Screenshots demonstrating the overlap of DUX-ChIP occupancy (yellow box) with the acquisition of 2-Cell embryo-like open chromatin and gene/MERVL expression (green box).

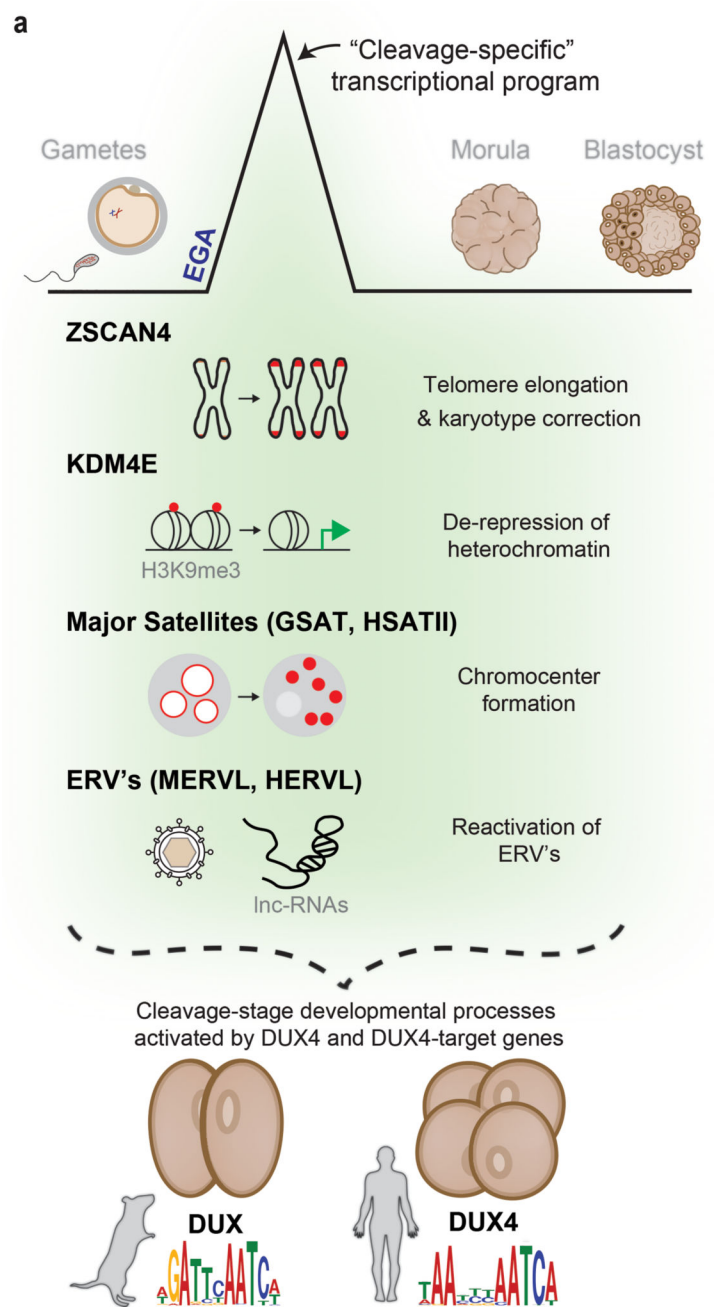

Figure 7. A model of DUX4 function during cleavage

(a) A cleavage-specific transcriptional program is activated at EGA in mouse and human cells by DUX or DUX4, respectively. The genes and repetitive elements activated by these DUX4-family genes mediate important molecular events associated with embryonic genome activation (EGA) and reprogramming in the mouse embryo (shaded in green). In human embryos, although activation of these genes and repetitive elements has been shown, their impact on these processes remains to be studied.