DOUBLE HOMEBOX (DUX) RETROGENES REGULATE EARLY EMBRYONIC TRANSCRIPTION AND STEM CELL POTENCY

by

Peter Hendrickson

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Oncological Sciences

The University of Utah

August 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Peter Hendrickson**

has been approved by the following supervisory committee members:

Bradley Cairns	, Chair	**03/24/2017** Date Approved
Jason Gertz	, Member	**03/24/2017** Date Approved
Cedric Feschotte	, Member	**03/24/2017** Date Approved
Douglas Carrell	, Member	**03/24/2017** Date Approved
Joshua Schiffman	, Member	**03/24/2017** Date Approved

and by **Bradley Cairns** , Chair/Dean of

the Department/College/School of **Oncological Sciences**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Following fertilization, mammalian embryos undergo a series of cleavage divisions during which embryonic cells are reprogrammed from a differentiated germ cell state to an undifferentiated embryonic state. Blastomeres of the 2-cell, 4-cell, and 8-cell (cleavage) embryo are 'totipotent', meaning they maintain the capacity to develop into all cell lineages of a developing embryo. This unique developmental potential is progressively lost in humans and other placental mammals as cells become restricted to an embryonic (inner cell mass, ICM) or extraembryonic (trophectoderm) lineage. Unlike pluripotent stem cells which can be derived from the ICM and have been extensively characterized, little is known about the totipotent cells of the cleavage stage embryo. What factors coordinate this dramatic reprogramming event and confer this unique developmental potential? Here, using RNA-sequencing, I profiled human cleavage stage embryos and identified a transcriptional program that specifically coincides with totipotency acquisition and loss. Remarkably, this totipotency transcriptional program is directly activated by DUX4 which is transiently and specifically expressed in 4-cell stage human embryos. The protein coding capacity of DUX4 has been conserved for over 100 million years; however, its evolutionary function/purpose has remained a complete mystery until now. By extending this work into mouse, I reveal a conserved functional role for DUX4-family genes in facilitating the mammalian embryonic reprogramming process through which totipotency is established.

Dedicated to Margaret Silliker. Thank you for your endless support and encouragement. Thank you for believing in me and for giving me the confidence to pursue my hopes and dreams.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

ACKNOWLEDGEMENTS

First and foremost, thank you to my mentor Brad Cairns for giving me intellectual freedom to explore. I came to you with big, risky ideas and you got behind me 100% and let me run with them. Also, thank you for being a constant source of enthusiasm. When the going got tough, you reinvigorated me with praise, excitement, and coffee.

I would also like to thank the Cairns' lab members, who have made this an incredible place to train. Thank you for being great scientists and, above all else, great friends. More specifically, thanks to Ed Grow, Brad Weaver, Patrick Murphy, Jingtao Guo, and Tim Parnell for going above and beyond to help me out in the lab. None of this would have been possible without your constant help and advice.

Thank you Jessie Dorais for taking me under your wing and introducing me to the clinical world. I can only hope that one day I will be half as good of a doctor as you are. Your love for your work and compassion for your patients is admirable and truly inspiring. Also, thank you to Doug Carrell and Ben Emery at UCRM for believing in me and this project and for giving me so much of your valuable time and resources.

Finally, I would like to express my deepest gratitude to my mom, dad, and future wife, Jenna for their love and support over the last six years. I chose a difficult career path and without you guys, I'd be malnourished and very unhappy. I love you guys!

CHAPTER 1

INTRODUCTION

1.1 Preface

Immediately after fertilization, the newly formed mammalian zygote initiates a dramatic reprogramming process. The purpose of this process is to ultimately convert the terminally differentiated (unipotent) genomes inherited from the germ cells into a transcriptionally active, completely undifferentiated (totipotent) embryonic stem cell capable of giving rise to every specialized cell type of the adult organism. Reprogramming takes place during the cleavage stages of mammalian embryogenesis (2-cell, 4-cell, and 8-cell stages) and is underscored by extensive and progressive remodeling/opening of the epigenetic landscape (Wu et al., 2016). Epigenetics refers to the heritable changes to the genome, independent of modifications to the underlying DNA sequence that influence gene expression through transcriptional and post-transcriptional mechanisms (Zhou and Dean, 2015). In the early embryo, these changes are characterized by the removal of DNA methylation, the addition and deletion of different histone marks, and finally by the incorporation of specific histone variants (summarized below). Although still far from having a complete picture and understanding of these different molecular events and their interconnected-ness, it is clear that only if these events are completed successfully will the embryonic genome become

transcriptionally active (EGA) resulting in new gene expression. For reasons unknown, mammalian EGA appears to occur in waves ultimately leading to the expression of specific homeodomain transcription factors (OCT4, CDX2) that direct the first lineage decision (ICM and trophectoderm, respectively) (Ko, 2016). Here again, only if lineages are successfully established will the embryo implant into the uterine wall, allowing further cellular growth, differentiation, and development to continue. If not, the new embryo does not implant and is eliminated, which likely accounts for the vast majority of reproductive failures in mammals. In this introduction, I will first review and expand on important aspects of the embryonic reprogramming process, EGA, and the first lineage decision. Then, I will shift gears slightly to discuss embryonic stem cells -- which can be derived from pre-implantation blastocyst stage embryos and cultured *in vitro* -- and the molecular basis of their fate potential. In the final section, for reasons that will become clear, I will briefly examine the deep evolutionary history of homedomain transcription factors in development and disease and set the stage for a new family of homeobox genes that may ultimately confer totipotency.

1.2 Pre-implantation embryonic development

1.2.1 Epigenetic reprogramming. Arguably the most important aspect to embryonic reprogramming is the widespread removal of methylation from DNA. DNA methylation (DNAme) involves the addition of a methyl group (CH_3) to the DNA itself, usually at the fifth carbon atom of a cytosine residue, and serves to restrict the expression of certain genes. Almost immediately after fertilization in the 1-cell zygote, DNAme is globally and progressively lost from the genome (Seisenberger et al., 2013). The earliest embryonic

wave of DNA demethylation in the zygote occurs specifically in the paternal pronucleus and is catalyzed by an enzyme -- TET3 -- which actively and iteratively oxidizes/removes the methyl mark from specific cytosine residues (Amouroux et al., 2016). Most DNA methylation, however, including all that is inherited on the maternal genome, is lost passively during the cleavage stages, bottoming out around the time of implantation (~200 cell blastocyst stage embryo) (Smith et al., 2012; 2014). This loss is essential for the formation of a pluripotent epiblast and for subsequent differentiation and is mediated by the exclusion of maintenance DNA methyltransferase enzymes (i.e. DNMT1) from the nucleus during DNA synthesis (Messerschmidt et al., 2014).

Helping to guide DNAme removal in the embryo are different histone modifications inherited on the maternal genome and deposited on the paternal genome after protamine exchange. Histones are subject to a variety of different biochemical modifications including acetylation, methylation, phosphorylation, and ubiquitylation. The nature of the mark and its location along a histone tail lead to different, but usually predictable, changes in the local chromatin environment. The most abundant marks in mammalian cells occur on histone H3 at lysines 4, 9, and 27. H3K4 methylation associates with active gene expression while H3K9 and H3K27 methylation associate with inactive gene expression. Initially, all three marks localize asymmetrically in the early zygote, abundantly present in the maternal pronucleus and virtually absent from the paternal pronucleus (Li, 2002). Although likely a consequence of different epigenetic histories, this asymmetry does -- at least initially -- appear to have some functional significance with recent evidence showing that maternal-specific H3K9 di-methylation provides a binding site for DPPA3 binding and protection from TET3-mediated DNA demethylation

(Cantone and Fisher, 2013). By the 2-cell stage, however, most histone mark asymmetry between the parental genomes is corrected, including the methylation of H3K4 and H3K27 (in addition to H3K20 and H3K64), which help poise the new embryonic genome for transcriptional activation (Burton and Torres-Padilla, 2010). H3K9 tri-methylation, however, is not acquired on the paternal pronucleus until the 4-cell stage (Liu et al., 2004). This delayed symmetrization appears to be mediated by active H3K9 demethylases (KDM4-family) expressed at EGA and is likely responsible for keeping the paternal genome in a state of transcriptional permissively (Puschendorf et al., 2008).

H3K9 tri-methyaltion is a distinguishing mark heterochromatin. Heterochromatin formation is an important aspect of embryonic epigenetic organization, although it must be timed appropriately so as to initially allow full differentiation potential (i.e. totipotency) while still ensuring chromosomal integrity during cleavage stage cell divisions. To facilitate this latter process, the early zygote utilizes specific variant histones to establish constitutive centromeric heterochromatin independent of H3K9 tri-methylation (Borsos and Torres-Padilla, 2016). Variant histones are non-canonical histone proteins with unique amino acid sequences that form nucleosomes with accordingly unique properties. In the mouse zygote, histone H3.3, a variant of canonical histone H3, is preferentially incorporated after fertilization into the paternal pronucleus at peri-centromeres (Santenard et al., 2010). Here, H3.3 functions transiently to facilitate major satellite (MajSat) transcription, which then recruits HP1 and the formation of heterochromatin regionally (Jang et al., 2015). Remarkably, this early embryonic molecular event is even visible by eye. In the paternal pronucleus, H3.3-enriched pericentromeric domains form distinct rings around the nucleous, giving rise to structures

(already present in the maternal pronucleus -- inherited from the oocyte) called "nucleous-like precursor bodies" or NPBs (Probst et al., 2007). During the 2-cell stage, as paternal-specific H3.3 is replaced by canonical H3.1, the NBPs reorganize into chromocenters (clusters of HP1-rich constitutive heterochromatin from different chromosomes) taking on the nuclear appearance of somatic cells (Ishiuchi and Torres-Padilla, 2013).

Remarkably, many of these unique molecular and physical changes have even been observed during embryonic stem cell and somatic cell nuclear transfer (SCNT) reprogramming experiments, suggesting that they are not just isolated phenomena of the early embryo but rather are hallmarks of the genome-wide epigenetic reprogramming process (Smith et al., 2016). How this process is ultimately initiated and then directed is, in my opinion, one of the biggest questions in science. The answer, I believe, will impact multiple fields ranging from reproduction and regeneration medicine to cancer biology.

1.2.2 Embryonic genome activation. As mentioned in the preface, successful epigenetic reprogramming culminates with a burst of new transcription called Embryonic Genome Activation (EGA). The timing of EGA in mammals varies from species to species, ranging from the 2-cell stage in mouse to the 8-cell stage in cow. What regulates EGA timing is not known, although multiple factors, including chromatin competency, cell cycle length, and the availability of transcriptional machinery, all seem to play an experimentally supported role (Pálfy et al., 2017). It was originally thought that the purpose of EGA was to replenish the embryo with RNAs/proteins that would sustain continued development after maternal stores were 'used up'. This, however, is inconsistent with the transient nature of gene expression that has been observed at EGA

in multiple mammalian species. Mouse EGA has both a minor wave and a major wave. Unlike the minor wave, which recent work has attributed to global promiscuous transcription (ABE et al., 2010), major wave gene expression (corresponding to the late 2-cell stage embryo) is both active (sensitive to amanitin treatment) and specific. The major wave is made up of a unique set of genes that are both poorly conserved and rarely expressed outside of the 2-cell stage embryo. The defining member of this gene set, *Zscan4*, is a mammalian-specific transcription factor and well-established activator of telomere sister chromatid exchange (T-SCE) (Falco et al., 2007; Zalzman et al., 2010). The mouse has 6 copies of the *Zscan4* gene (*Zscan4a-f*) arranged in tandem on chromosome 7 and nonspecific depletion of its mRNA in 2-cell stage embryos leads to developmental delay and implantation failure (Falco et al., 2007). Flanking multiple of the *Zscan4* paralogs are long terminal repeat (LTR) elements derived from a mouse-specific endogenous retrovirus called MERVL (short for Murine Endogenous Retrovirus with a Leucine tRNA primer binding site). Just like *Zscan4*, MERVL is transiently and specifically reactivated in the 2-cell stage embryo, giving rise to viral-like particles that more or less vanish by the 4-cell stage (Ribet et al., 2008). Shortly after this discovery, it was demonstrated that many MERVL elements -- specifically the LTR sequences flanking the gag/pol genes (called MT2) -- had actually been co-opted and were now being used as cis-regulatory elements to coordinate the transcription of hundreds of host genes specifically expressed in the late 2-cell stage embryo, including *Zfp352*, *Tdpoz1-5*, *Tcstv1-3*, and the *Eif1a-like* family (Macfarlan et al., 2011). Collectively, MERVL and major wave EGA gene expression is now known as the ‘2C’ transcriptional program. Although the purpose of this transcriptional program is still not fully clear, its transient

and specific expression in the embryo perfectly aligns with the acquisition and loss of totipotency -- the highly coveted and somewhat enigmatic competence of an early embryonic stem cell to contribute to both embryonic and extraembryonic lineages (discussed in detail below) (Macfarlan et al., 2012).

1.2.3 The first lineage decision. In the placental mammal, the first lineage decision coincides with blastocyst formation prior to implantation. Once this decision is made, embryonic cells with two completely different, and irreversible, trajectories are created. One cell population, called the trophoblast, is allocated to the polarized outer surface of the blastocyst embryo and becomes restricted to an extraembryonic fate -- responsible for producing all cell types of the developing placenta. The second cell population, known as the pluripotent embryonic stem cells, form the inner cell mass (ICM) and become restricted to an embryonic fate -- responsible for generating all cell-types and tissues of the developing new organism. Extensive work over the last couple of decades has elucidated the molecular circuitry behind this lineage decision. At the crux of it are two mutually antagonistic homeodomain transcription factors (TF), OCT4 and CDX2, which become dichotomously expressed in the pluripotent embryonic stem cells and trophoblasts, respectively (Deb, 2006; Wang et al., 2010). Expression of these two factors is biased very early on -- starting at the 4-cell stage -- enabled by random, heterogeneous expression of a gene, *Sox21* (Goolam et al., 2016). SOX21 is a transcription factor that suppresses *Cdx2* expression and pushes early blastomeres towards an ICM fate. In the 8-cell and morula stages, this early specification is furthered by inside/outside positional cues that, through HIPPO signaling, reinforce OCT4 or CDX2 expression, leading to eventual lineage commitment (Korotkevich et al., 2017). Although it is important to

recognize that many of the signaling factors and events involved in this first lineage decision-making process are not conserved in humans, the decisive lineage commitment factors (OCT4/CDX2) are and the timing of the decision is consistent. In short, during the cleavage divisions (2-cell, 4-cell, and 8-cell stages), totipotent blastomeres become specified to an ICM or trophoblast lineage and progressively lose the ability to switch. Once they commit during blastocyst formation, new epigenetics barriers (such as DNAme, histone marks, and histone variants) are incorporated or established to prevent the cells from going back (Tee and Reinberg, 2014; Yuan et al., 2009). Specifically, what the barriers to totipotency are and how they become established in the embryo is not known in full, but are of high interest as such knowledge could be exploited to create a totipotent embryonic stem cell.

1.3 Embryonic stem cells and cell potency

1.3.1 Pluripotency establishment and maintenance. The ICM of a blastocyst stage mouse embryo is made up of a small number of pluripotent mouse embryonic stem cells (mESCs). Although these cells are restricted from contributing to the developing placenta (and hence are no longer totipotent), pluripotent mESCs maintain the capacity to produce all tissues of the developing fetus and some extraembryonic membranes (Loh et al., 2006). Pioneering work over the last few decades has resulted in the ability to isolate and propagate pluripotent mESCs cells *in vitro*. This state is maintained by a specific network of core, highly conserved transcription factors, including OCT4, SOX2, and NANOG, whose expression can be stabilized *in vitro* by adding LIF (leukemia inhibitory factor) and inhibiting critical differentiation pathways such as Wnt and Erk-MAPK (Ying

et al., 2008). Moreover, exogenous expression of these same factors can "reprogram" terminally differentiated somatic cells to a pluripotent 'ES cell-like' state (Takahashi and Yamanaka, 2006). Just like native mESCs, induced mouse pluripotent stem cells (miPSCs) are capable of generating nearly the full array of somatic and germline cell types upon exposure to specific differentiation cues and can efficiently contribute to chimeras.

While these advances in mouse lay an important and exciting foundation for the future of regenerative medicine, replicating them in human has been met with a number of challenges. Topping that list currently is an inability to artificially stabilize pluripotency factor gene expression in human embryonic stem cells (hESCs) and induced pluripotent stem cells (hiPSCs) (Rossant and Tam, 2017). Consequently, hESCs and hiPSCs cannot contribute to chimeras and are said to be in a developmentally advanced (and hypermethylated) state of pluripotency known as the 'primed' state (Theunissen et al., 2016). Despite major efforts recently to capture and maintain human stem cells in a state of 'naïve' pluripotency like mouse, limited progress has been made (Ware, 2017). The exact reasons for this disparity in "stemness" are not known, but likely harken back to important molecular differences that exist between human and mouse. One possibility is that hESCs require specific culture conditions or inhibitors to prevent primate-specific stem cell differentiation pathways that have not yet been identified. Another plausible possibility, however, is that human and mouse pluripotency are just fundamentally different.

1.3.2 Transposable elements and the core stem cell regulatory network. Pluripotency acquisition and maintenance is conferred by a specific transcriptional

program. This program is regulated by a set of indispensable, core transcription factors (OCT4, SOX4, NANOG) and a handful of 'ancillary' factors (KLF4, PRDM14, SALL4, etc.) each of which is individually dispensable (Li and Belmonte, 2017). While many of these factors are evolutionarily and functionally conserved across species, the transcriptional programs they regulate are not. Deep transcriptional profiling of mESCs and hESCs has revealed a striking degree of species-specific, and stem cell-specific transcription (Fort et al., 2014). Follow-up functional work in hESCs has shown that most of this transcription is non-coding and -- at least in a few cases -- produces RNA molecules (lincRNA-ROR, ESRG, HPAT5, etc.) that are necessary for pluripotent acquisition or maintenance (Durruthy-Durruthy et al., 2016; Lu et al., 2014; Wang et al., 2013). Most species-specific transcription in hESCs and mESCs, however, occurs at putative enhancer and promoter sequences (based on histone marks) which also happen to be transposable elements (TEs) (Kunarso et al., 2010). TEs are mobile genetic elements that are ubiquitous in mammalian genomes, accounting for ~50% of total DNA (Feschotte, 2008). Retrotransposons make up the predominant class of TE in most mammals and are composed of notable sub-classes: long interspersed elements (LINEs) short interspersed elements (SINEs), and endogenous retroviruses (ERVs). ERVs are an important driver of genetic innovation. Over time, old insertions can be co-opted as enhancer and promoter sequences and used -- in a species-specific manner -- to create or expand gene regulatory networks (Chuong et al., 2016). Pluripotency is just one of many examples of this incredible evolution. In hESCs, as many as 25% of all OCT4/NANOG binding sites overlap with a TE -- the majority of sites being contributed by a primate-specific long terminal repeat element of the HERVH subfamily -- LTR7 (Kunarso et al.,

2010). Conversely, although a similar fraction of OCT4/NANOG binding sites in mESCs also overlap a TE, most belong to murine-specific elements of the ERVK family. As a result of this extensive TE co-option, the core regulatory networks of human and mouse embryonic stem cells have diverged massively. Juxtaposed to mouse, hundreds of new genes (both coding and non-coding) have been wired into and out of human pluripotency signature, likely altering the fundamental nature of this cellular state. What this means is that pluripotency in mouse is not the same as pluripotency in human which is likely not the same as pluripotency in cow, fish, or any other species.

1.3.3 Expanding pluripotent fate potential. In standard mESCs cultures, multiple metastable subpopulations of cells with different degrees of "stemness" are present. At the root of this heterogeneity, in most cases, is a transient fluctuation of core pluripotency factor expression which primes a small number of cells for differentiation towards a variety of embryonic lineages (Torres-Padilla and Chambers, 2014). Recently, however, a different subpopulation of mESC was identified. Unlike the others, this rare population of mESC (~0.1%-0.5% of all cells) did not turn on genes typically expressed in later (post-implantation) developmental stages, but instead turned on genes (*Zscan4*, *Zfp352*, *Tcstv3*, etc.) and even retrotransposons (MERVL) only otherwise expressed very early in development -- specifically during the 2-cell stage (Macfarlan et al., 2012). Using *Zscan4* or MERVL-driven fluorescence transgenes, these rare mESCs -- now known to most as the 2-cell embryo-like (or '2C-like') cells -- have been isolated and meticulously characterized. In addition to reactivating the '2C' transcriptional program (see 1.1.2) and disengaging pluripotency (by uncoupling *Oct4* transcription and translation), '2C-like' cells acquire multiple distinct and extraordinary molecular features of a 2-cell stage

mouse embryo. This includes genome-wide DNA demethylation, histone mark remodeling, and chromocenter dissolution consistent with global heterochromatin de-repression (Akiyama et al., 2015; Eckersley-Maslin et al., 2016; Ishiuchi et al., 2015). Most remarkably, however, these cells also gain the functional properties of 2-cell stage mouse embryo, namely *totipotency*, the ability to contribute to both embryonic and extraembryonic lineages (Choi et al., 2017). The ability of a mESC (derived from the embryonic-restricted ICM of the blastocyst stage embryo) to spontaneously recover this expanded fate is paradigm shifting and, consequently, has been independently confirmed by numerous labs using a variety of different assays, including embryoid body (EB) formation, morula aggregation, and somatic cell nuclear transfer (SCNT). What has not yet been determined, however, is what drives the stochastic reversion of a pluripotent mESC to a totipotent '2C-like' state. To date, nearly all work on this front has focused on epigenetic mechanisms -- identifying enzymes involved in transcriptional silencing (LSD1, G9a, TRIM28) and chromatin formation (CAF-1) that, when inhibited with small molecules or siRNAs, create a more open/accessible chromatin environment. Although somewhat effective at increasing MERVL expression levels and the percentage of '2C-like' cells, these effects are indirect and are more likely an experimental creation rather than a reflection of any real biology in the embryo. Although it is logical to assume that MERVL expression is simply a consequence of the open chromatin state via epigenetic reprogramming, it is also plausible that it -- like ERVK expression in pluripotent mESCs -- can be actively induced. To rephrase -- what if MERVL reactivation is not just a bystander effect but instead help confers totipotency by enabling the expression the '2C' transcriptional program it has wired together?

As discussed above, TEs catalyze the evolution of new transcriptional programs by dispersing transcription factor binding sites throughout the genome. We hypothesize that MERVL was co-opted by an early embryonic mouse transcription factor to create the core regulatory network of a 2-cell stage blastomere. In this dissertation, I identify the co-opting transcription factor in a *double homeobox* gene known as *Dux* and demonstrate that it is both necessary and sufficient to activate MERVL and convert mESCs to a '2C-like' state.

1.4 Novel homeodomains define a novel stem cell

1.4.1 Homeobox genes and development. The homeobox genes make up a large super-family of genes that encode transcription factors. Homeobox's can be easily identified and characterized based on a highly conserved ~60 amino acid homedomain(s) which enables DNA-binding. In animals, the homeobox genes are broken up into 16 classes, including ANTP, PRD, PRD-like, POU, TALE, etc. (Bürglin and Affolter, 2016). Class determination typically depends on the presence/absence of additional domains which influence DNA binding as well as the protein's ability to interact with a great variety of other proteins. This diversity of interaction allows the homeobox family to participate in many different cellular processes in many different species. Most homeobox genes display strong cell type-specific expression and often regulate identity-defining transcriptional programs (Dunwell and Holland, 2016). For example, two of the three core pluripotency factors (OCT4 and NANOG) and the decisive master regulator of the trophoblast lineage (CDX2) are all homeobox factors and are specifically expressed in the morula and blastocyst stages of embryogenesis. Accordingly, disrupted expression

or function of homeobox genes is often times detrimental to an organism and is a major cause of embryonic lethality and human disease.

1.4.2 DUX: a toxic retrogene. *Dux* is a rather unusual homeobox gene. Unlike any other in mouse, it encodes not one, but two, adjacent N-terminal DNA binding homedomains. Additionally, *Dux* is a retrogene, meaning it lacks introns and, like most other retrogenes, has multiple copies scattered throughout the genome. *Dux* is the supposed 'retro-ortholog' of a well-studied multi-copy retrogene in primates called *DUX4* (Leidenroth et al., 2012). In primate genomes, *DUX4* is embedded within a repeat unit known as D4Z4 that is often found in large arrays on multiple chromosomes. The largest D4Z4 array is located on the subtelomere of chromosome 4. Although this locus is normally heterochromatinized and completely transcriptionally silenced in human somatic cells, a rare autosomal dominant disease condition called Fascioscapulohumeral muscular dystrophy (FSHD) is associated with locus relaxation (typically caused by an inherited contraction of the D4Z4 repeat array) and aberrant *DUX4* expression in myoblasts (Geng et al., 2012). Here, it has been shown that DUX4 activates the expression of hundreds of unique genes and TEs, which ultimately cause apoptotsis (Young et al., 2013). Outside of this disease association, neither *DUX4* or *Dux* expression has been observed in any somatic tissue or cell-type, nor have they been shown to play any role in normal physiology. Interestingly, however, both were derived (via separate retrotransposition events) from an intron-containing *DUXC* gene that can still be found in all placental mammals, except for in primates and mice where it has been seemingly replaced by the retogenes (Leidenroth and Hewitt, 2010). Although *DUXC* also has no known function, this strong evolutionary conservation and expansion is clear evidence of

biological importance. But where and when is it operating and what is its purpose?

1.4.3 PRD-like homedomains in early human embryos. In this dissertation, I show that *DUX4* and *Dux* are transiently expressed in the early cleavage-stages of embryogenesis, aligned with the onset of EGA at the 4-cell stage and 2-cell stage, respectively. Here, DUX4 and DUX function as transcriptional activators (and probable pioneer factors) (Choi et al., 2016), and turn on the expression of hundreds of genes critical for early embryonic development. Interestingly, while some similarities exist, the early embryonic transcriptional programs elicited by DUX4 in human and DUX in mice are largely different from each other and are composed of many species-specific genes. Notably, some of the most prominent targets of DUX4 -- namely, *ARGFX*, *DUXA*, *DUXB*, *DPRX*, *LEUTX*, and *TPRX1* -- have all been lost from the rodent lineage (Töhönen et al., 2015). Like *DUX4* itself, all six genes are members of the PRD-like homeobox class and are specific to eutherian (placental) mammals. PRD-like homeobox genes encode homeodomains similar to those in the PRD (paired) class; however, they exhibit significant sequence divergence and, in most cases, a restricted phylogenetic distribution. *DUXA* and *DUXB* are intron-containing double homeobox genes that likely arose the same time as *DUXC*. Notably, however, unlike DUXC, both DUXA and DUXB completely lack a c-terminal transactivation domain (TAD) that would presumably be necessary for their function as transcriptional activators (Leidenroth and Hewitt, 2010). *ARGFX*, *DPRX*, *LEUTX*, and *TPRX1* are all single homeobox genes thought to have arisen via tandem duplication and divergence from the same *Crx* (Cone-rod homeobox) gene (Madissoon et al., 2016). Again, compared to the *Crx* gene which is expressed exclusively during mammalian eye development, innovative functional roles for each are

expected. Fascinatingly, all six novel PRD-like genes are exclusively expressed in the 8-cell and mourla stages of human embryogenesis. While prior work has argued for a central role in coordinating gene expression necessary for establishing or "fine tuning" the first lineage decision (Maeso et al., 2016), this hypothesis requires rigorous functional testing that will be difficult given the technical and ethical limitations of human embryo research. Nevertheless, as will be discussed more in the conclusions (Chapter 3) understanding what the many targets of DUX4 and DUX do in the early embryo -- both the conserved and the species-specific ones -- is a critical next step for elucidating the conserved evolutionary function of DUXC and for understanding the highly divergent nature of early embryonic development.

1.5 References

ABE, K.-I., Inoue, A., Suzuki, M.G., and Aoki, F. (2010). Global gene silencing is caused by the dissociation of RNA polymerase II from DNA in mouse oocytes. J. Reprod. Dev. *56*, 502–507.

Akiyama, T., Xin, L., Oda, M., Sharov, A.A., Amano, M., Piao, Y., Cadet, J.S., Dudekula, D.B., Qian, Y., Wang, W., et al. (2015). Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. DNA Res. *22*, 307–318.

Amouroux, R., Nashun, B., Shirane, K., Nakagawa, S., Hill, P.W.S., D'Souza, Z., Nakayama, M., Matsuda, M., Turp, A., Ndjetehe, E., et al. (2016). De novo DNA methylation drives 5hmC accumulation in mouse zygotes. Nat Cell Biol *18*, 225–233.

Borsos, M., and Torres-Padilla, M.-E. (2016). Building up the nucleus: nuclear organization in the establishment of totipotency and pluripotency during mammalian development. Genes Dev. *30*, 611–621.

Burton, A., and Torres-Padilla, M.E. (2010). Epigenetic reprogramming and development: a unique heterochromatin organization in the preimplantation mouse embryo | Briefings in Functional Genomics | Oxford Academic. Briefings in Functional Genomics.

Bürglin, T.R., and Affolter, M. (2016). Homeodomain proteins: an update. Chromosoma *125*, 497–521.

Cantone, I., and Fisher, A.G. (2013). Epigenetic programming and reprogramming during development. Nat. Struct. Mol. Biol. *20*, 282–289.

Choi, S.H., Gearhart, M.D., Cui, Z., Bosnakovski, D., Kim, M., Schennum, N., and Kyba, M. (2016). DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes. Nucleic Acids Res. gkw141.

Choi, Y.J., Lin, C.-P., Risso, D., Chen, S., Kim, T.A., Tan, M.H., Li, J.B., Wu, Y., Chen, C., Xuan, Z., et al. (2017). Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. aag1927.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet *18*, 71–86.

Deb, K. (2006). Cdx2 gene expression and trophectoderm lineage specification in mouse embryos. Science *311*, 992–996.

Dunwell, T.L., and Holland, P.W.H. (2016). Diversity of human and mouse homeobox gene expression in development and adult tissues. BMC Dev Biol *16*, 40.

Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E.J., Davila, J., Mall, M., Wong, W.H., Wysocka, J., et al. (2016). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet *48*, 44–52.

Eckersley-Maslin, M.A., Svensson, V., Krueger, C., Stubbs, T.M., Giehr, P., Krueger, F., Miragaia, R.J., Kyriakopoulos, C., Berrens, R.V., Milagre, I., et al. (2016). MERVL/Zscan4 network activation results in transient genome-wide DNA demethylation of mESCs. Cell Rep *17*, 179–192.

Falco, G., Lee, S.-L., Stanghellini, I., Bassey, U.C., Hamatani, T., and Ko, M.S.H. (2007). Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Dev Biol *307*, 539–550.
Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nat Rev Genet.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., et al. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet *46*, 558–566.

Geng, L.N., Yao, Z., Snider, L., Fong, A.P., Cech, J.N., Young, J.M., van der Maarel, S.M., Ruzzo, W.L., Gentleman, R.C., Tawil, R., et al. (2012). DUX4 activates germline

genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. Dev Cell *22*, 38–51.

Goolam, M., Scialdone, A., Graham, S.J.L., Macaulay, I.C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J.C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-Cell mouse embryos. Cell *165*, 61–74.

Ishiuchi, T., and Torres-Padilla, M.-E. (2013). Towards an understanding of the regulatory mechanisms of totipotency. Curr Opin Genet Dev *23*, 512–518.

Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Bošković, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J.M., and Torres-Padilla, M.-E. (2015). Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat. Struct. Mol. Biol.

Jang, C.-W., Shibata, Y., Starmer, J., Yee, Della, and Magnuson, T. (2015). Histone H3.3 maintains genome integrity during mammalian development. Genes Dev. *29*, 1377–1392.

Ko, M.S.H. (2016). Zygotic genome activation revisited: looking through the expression and function of Zscan4. *120*, 103–124.

Korotkevich, E., Niwayama, R., Courtois, A., Friese, S., Berger, N., Buchholz, F., and Hiiragi, T. (2017). The apical domain is required and sufficient for the first lineage segregation in the mouse embryo. Dev Cell *40*, 235–247.e237.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet *42*, 631–634.

Leidenroth, A., and Hewitt, J.E. (2010). A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol. Biol. *10*, 364.

Leidenroth, A., Clapp, J., Mitchell, L.M., Coneyworth, D., Dearden, F.L., Iannuzzi, L., and Hewitt, J.E. (2012). Evolution of DUX gene macrosatellites in placental mammals. Chromosoma *121*, 489–497.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet *3*, 662–673.

Li, M., and Belmonte, J.C.I. (2017). Ground rules of the pluripotency gene regulatory network. Nat Rev Genet *18*, 180–191.

Liu, H., Kim, J.-M., and Aoki, F. (2004). Regulation of histone H3 lysine 9 methylation in oocytes and early pre-implantation embryos. Development *131*, 2269–2280.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George,

J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet *38*, 431–440.

Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., and Ng, H.-H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat. Struct. Mol. Biol. *21*, 423–425.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. *25*, 594–607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Madissoon, E., Jouhilahti, E.-M., Vesterlund, L., Töhönen, V., Krjutškov, K., Petropoulous, S., Einarsdottir, E., Linnarsson, S., Lanner, F., Månsson, R., et al. (2016). Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep *6*, 28995.

Maeso, I., Dunwell, T.L., Wyatt, C.D.R., Marlétaz, F., Vető, B., Bernal, J.A., Quah, S., Irimia, M., and Holland, P.W.H. (2016). Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. BMC Biol. *14*, 45.

Messerschmidt, D.M., Knowles, B.B., and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev. *28*, 812–828.

Pálfy, M., Joseph, S.R., and Vastenhouw, N.L. (2017). The timing of zygotic genome activation. Curr Opin Genet Dev *43*, 53–60.

Probst, A.V., Santos, F., Reik, W., Almouzni, G., and Dean, W. (2007). Structural differences in centromeric heterochromatin are spatially reconciled on fertilisation in the mouse zygote. Chromosoma *116*, 403–415.

Puschendorf, M., Terranova, R., Boutsma, E., Mao, X., Isono, K.-I., Brykczynska, U., Kolb, C., Otte, A.P., Koseki, H., Orkin, S.H., et al. (2008). PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. Nat Genet *40*, 411–420.

Ribet, D., Louvet-Vallee, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., Pierron, G., Maro, B., and Heidmann, T. (2008). Murine endogenous retrovirus MuERV-L is the progenitor of the “orphan” epsilon viruslike particles of the early mouse embryo. Journal of Virology *82*, 1622–1625.

Rossant, J., and Tam, P.P.L. (2017). New insights into early human development: lessons for stem cell derivation and differentiation. Cell Stem Cell *20*, 18–28.

Santenard, A., Ziegler-Birling, C., Koch, M., Tora, L., Bannister, A.J., and Torres-Padilla, M.-E. (2010). Heterochromatin formation in the mouse embryo requires critical residues of the histone variant H3.3. Nat Cell Biol *12*, 853–862.

Seisenberger, S., Peat, J.R., Hore, T.A., Santos, F., Dean, W., and Reik, W. (2013). Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. Philos Trans R Soc Lond, B, Biol Sci *368*, 20110330.

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. Nature 1–18.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339–344.

Smith, Z.D., Sindhu, C., and Meissner, A. (2016). Molecular features of cellular reprogramming and development. Nat. Rev. Mol. Cell Biol. *17*, 139–154.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.

Tee, W.-W., and Reinberg, D. (2014). Chromatin features and the epigenetic regulation of pluripotency states in ESCs. *141*, 2376–2390.

Theunissen, T.W., Friedli, M., He, Y., Planet, E., O'Neil, R.C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., et al. (2016). Molecular criteria for defining the naive human pluripotent state. Cell Stem Cell *19*, 502–515.

Torres-Padilla, M.-E., and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *141*, 2173–2181.

Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E.-M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T.R., et al. (2015). Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun *6*, 8207.

Wang, K., Sengupta, S., Magnani, L., Wilson, C.A., Henry, R.W., and Knott, J.G. (2010). Brg1 is required for Cdx2-mediated repression of Oct4 expression in mouse blastocysts. PLoS ONE *5*, e10622.

Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X., and Liu, H. (2013). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and

Sox2 in human embryonic stem cell self-renewal. Dev Cell *25*, 69–80.

Ware, C.B. (2017). Concise review: lessons from naïve human pluripotent cells. Stem Cells *35*, 35–41.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652–657.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Young, J.M., Whiddon, J.L., Yao, Z., Kasinathan, B., Snider, L., Geng, L.N., Balog, J., Tawil, R., van der Maarel, S.M., and Tapscott, S.J. (2013). DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLoS Genet *9*, e1003947.

Yuan, P., Han, J., Guo, G., Orlov, Y.L., Huss, M., Loh, Y.-H., Yaw, L.-P., Robson, P., Lim, B., and Ng, H.-H. (2009). Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev. *23*, 2507–2520.

Zalzman, M., Falco, G., Sharova, L.V., Nishiyama, A., Thomas, M., Lee, S.-L., Stagg, C.A., Hoang, H.G., Yang, H.-T., Indig, F.E., et al. (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature *464*, 858–863.

Zhou, L.-Q., and Dean, J. (2015). Reprogramming the genome to totipotency in mouse embryos. Trends Cell Biol. *25*, 82–91.

CHAPTER 2

CONSERVED ROLES FOR MURINE DUX AND HUMAN DUX4 IN ACTIVATING CLEAVAGE STAGE GENES AND MERVL/HERVL RETROTRANSPOSONS

2.1 Preface

This chapter includes a re-formatted manuscript accepted for publication at Nature Genetics; reprinted with permission from Nature Publishing Group. This work was completed with help from the following co-authors: Jessie A. Doráis, Edward J. Grow, Jennifer L. Whiddon, Jong-Won Lim, Candice L. Wike, Bradley D. Weaver, Christian Pflueger, Benjamin R. Emery, Aaron L. Wilcox, David A. Nix, C. Matthew Peterson, Stephen J. Tapscott, Douglas T. Carrell, and Bradley R. Cairns.

2.2 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.

2.3 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 (Liu et al., 2007). 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 (Chung et al., 2015; Matoba et al., 2014), and the ZSCAN4 transcription factor family in the sister chromatid exchange (T-SCE) mechanism needed for telomere elongation (Kalmbach et al., 2014; Zalzman et al., 2010). 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 (Gifford et al., 2013; Macfarlan et al., 2011). 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 (Macfarlan et al., 2012). Coincident with MERVL reactivation, '2C-like' cells acquire the unique molecular and developmental features and functions of totipotent cleavage-stage cells (Ishiuchi et al., 2015), 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. 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 (Geng et al., 2012; Young et al., 2013). 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.

2.4 Results

2.4.1 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 (Figure 2.1a). Blastocyst embryos were manually separated into ICM and mural trophectoderm by laser dissection. 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 (Gertz et al., 2012). 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. Importantly, read coverage from transcription start site (TSS) to transcription termination site (TTS) was exceptionally well-balanced compared to prior work (Figure 2.1b,c), making these new datasets the most comprehensive transcriptomes of human oocyte and pre-implantation embryonic development to date.

2.4.2 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). 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 (Figure 2.2a). 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 (Figure 2.2b). 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.

2.4.3 Examination of alternative splicing and novel transcription. Overall, our transcription profiles were consistent with prior single cell datasets (Xue et al., 2013; Yan et al., 2013) (Figure 2.3a). However, improvements in read coverage balance and directionality enabled the discovery of new novel transcription and splice isoform expression during pre-implantation development (Figure 2.3b). Together, these datasets yield extensive new information providing a major resource for future studies (Figure 2.4a,b).

2.4.4 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 (Figure 2.5a**)**. 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)**.** *DUX4* is one of three coding DUX (double homeobox) genes in humans, which also includes *DUXA* and *DUXB* (Leidenroth and Hewitt, 2010). 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 (Holland et al., 2007).

2.4.5 DUX4 potently activates cleavage-specific genes and repetitive elements. *DUX4* mRNA and protein are restricted to the 4-cell stage (early EGA) (Figure 2.6a) preceding the transient expression/enrichment of other 'PRD-like' genes during the 8-cell and morula stages (Figure 2.6b). 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 –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 (Figure 2.7a,b). The most highly activated gene was *ZSCAN4*, a defining cleavage-stage gene in both human and mouse (Ko, 2016). 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* (Figure 2.8a). Transient co-transfection with *DUX4* induced luciferase expression >2,000-fold. However, in contrast to prior work (Töhönen et al., 2015), 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 (Madissoon et al., 2016; Töhönen et al., 2015), 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 (Figure 2.9a). Most striking, however, was the strong induction of HERVL retrotransposons which are selectively transcribed in the cleavage stage, consistent with previous findings (Göke et al., 2015). 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 (Geng et al., 2012; Young et al., 2013). 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. Using GREAT (McLean et al., 2010),

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 iPSC, 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 (Figure 2.10a). Unique reads revealed significant DUX4 enrichment at activated LTR elements (e.g. MLT2A1, MLT2A2) and HSATII satellites (Figure 2.10b), 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.

2.4.6 Functional conservation of DUX proteins in defining the cleavage stage transcriptome. 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 (Deng et al., 2014) revealed cleavage-stage-specific transcription of a weakly conserved *DUX4* homolog in mouse, called *Dux* (Leidenroth et al., 2012) (Figure 2.11a). Notably, *Dux* is transiently and specifically expressed in early 2-cell stage mouse embryos, 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* (Figure 2.11b). 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) (Figure 2.12a), including notable retrotransposons (e.g. MERVL and its LTR, MT2_Mm) with no genes being significantly downregulated. This cohort of differentially expressed genes is transiently and specifically expressed in the mouse cleavage-stage embryo (Figure 2.12b) 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.

2.4.7 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 (Kigami et al., 2003; Ribet et al., 2008; Schoorlemmer et al., 2014) (Figure 2.13a). 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 (Macfarlan et al., 2012), the loss of POU5F1, and the disaggregation and reformation of constitutive heterochromatin into chromocenters (Ishiuchi et al., 2015). Accordingly, we find *Dux* and DUX-induced genes strongly upregulated in MERVL-expressing cells (Figure 2.13b). 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 (Figure 2.13c). 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. Live imaging fluorescent microscopy confirmed this observation and further revealed dose dependency.

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 (Figure 2.14a). 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 (Figure 2.14b). Notably, *Dux* transgene RNA levels correlated with dox induction and with conversion to a GFP^{pos} state (Figure 2.14c). Although transgene expression in the induced cells exceeded the expression of endogenous *Dux* RNA in spontaneously fluctuating '2C-like' cells, the transcriptional profiles were highly similar ($r=0.78$) (Figure 2.14d). Together, these data indicate DUX as a potent transcriptional activator of '2C-like' genes and retrotransposons. 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 (Figure 2.15a). 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.

2.4.8 *Dux* is necessary for induction of '2C-like' cells. Depletion of *Chaf1a*, the p150 subunit of the chromatin assembly factor 1 complex (CAF-1) (Figure 2.16a) also induces the conversion of mESCs to a '2C-like' state (Ishiuchi et al., 2015), 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 (Figure 2.16b). Moreover, the downstream targets of DUX (determined in our *Dux* overexpression studies) composed the most highly activated genes in the CAF-1-depleted datasets (Figure 2.16b). 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, and we confirm prior results showing that depletion of *Chaf1a* alone leads to a >20-fold increase (Figure 2.17a). Interestingly, co-transfection of mESCs with siRNA against *Dux* and *Chaf1a* nearly abolished the inductive effect of *Chaf1a* knockdown alone (Figure 2.17b). 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 (Figure 2.18a). 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 (Figure 2.19a). 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 (Figure 2.19b). Thus, conversion of mESCs to a '2C-like' state -- either spontaneous or through CAF-1 knockdown -- is dependent on *Dux*.

2.4.9 *Dux* expression converts 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 (Buenrostro et al., 2013). 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 (Wu et al., 2016). 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) (Figure 2.20a). 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 (Figure 2.20b). In contrast to the 9,131 common peaks found primarily at gene promoters, the ATAC-gained regions were mostly in intergenic space, with the majority (64.5%, P<0.001) directly overlapping a MERVL element (Figure 2.21a). 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 (Figure 2.21b). 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 (Figure 2.21c). 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 (Figure 2.21d). 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.

2.4.10 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 (Figure 2.22a). 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" (Figure 2.22b). 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 (Figure 2.23a,b). 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 (Figure 2.24a,b). Using the top 10,000 peak summits based on enrichment score, we further identified a consensus DUX binding motif, with the top hit (WGATTYAATCW) scoring an E-value of 2.0e-7234 (Figure 2.24c). 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.

2.5 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 (Ishiuchi and Torres-Padilla, 2013; Zhou and Dean, 2015). 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) (Figure 2.25a).

As *DUX4*-family genes themselves are expressed around the same time that EGA commences, they are not likely responsible for global EGA initiation. Instead, our ATAC-seq data, along with prior work (Wu et al., 2016), strongly suggests roles in opening chromatin – which may be analogous to pioneer factors such as *Drosophila's* Zelda (Harrison et al., 2011; Iwafuchi-Doi and Zaret, 2014; Sun et al., 2015) – 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 (Ishiuchi et al., 2015).

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 (Clapp et al., 2007). 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 (Leidenroth and Hewitt, 2010; Leidenroth et al., 2012). 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 support 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 (De Paepe et al., 2014; Morgani and Brickman, 2014). 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 (Yasuda et al., 2016; Zhang et al., 2016), prompting need for a greater understanding of DUX4 biochemically and molecularly in normal and oncogenic circumstances.

2.6 Methods

2.6.1 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.

2.6.2 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.

2.6.3 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.

2.6.4 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.

2.6.5. 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 40x magnification using the Nikon A1 confocal microscope.

2.6.6 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).

2.6.7 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).

2.6.8 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.

2.6.9 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 24hrs of dox-induction, the cells were lysed in Trizol and RNA extracted using the Direct-zol^{TM} 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.

2.6.10 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 reads 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.

2.6.11 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 24hrs to induce transgene expression, verified by western blot. Finally, cells were lysed and the luciferase intensity was measured using the Dual-luciferase^{TM} Reporter Assay from Promega. This experiment was performed twice with each condition repeated in quadruplicate.

2.6.12 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.

2.6.13 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 performed with iTaq Universal SYBR Green Supermix (Bio-Rad). Experiments were performed in biological triplicate. Expression levels were normalized to *Timm17b* by DeltaCT.

2.6.14 Mouse ES cell culture and generation of stable cell lines. Mycoplasma-free E14 mESCs were cultured on gelatin in '2i' media containing PluriQ^{TM} 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.

2.6.15 Fluorescence-activated cell sorting. Quantification of GFP-positive cells was performed using a Cytek DxP Analyzer and data were 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.

2.6.16 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.

2.6.17 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 $\geq$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.

2.6.18 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 24hrs 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 $MgCl_2$, 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 reads 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 were 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 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.

2.6.19 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 reads 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 were 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.

2.6.20 Immunofluorescence and imaging. Cells were plated on gelatin-coated coverslips and allowed to adhere for 3-5hrs 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 1hr at room temperature. Primary antibodies (see below) were diluted in 3% BSA and the cells were incubated for 1hr at room temperature. Cells were then washed and incubated in diluted Alexa-conjugated secondary antibodies plus DAPI (4′,6-diamidino-2-phenylindole) for 1hr 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^{TM} 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).

2.6.21 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. 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^{TM} 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.

2.7 References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. *44*, W3–W10.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth *10*, 1213–1218.

Chung, Y.G., Matoba, S., Liu, Y., Eum, J.H., Lu, F., Jiang, W., Lee, J.E., Sepilian, V., Cha, K.Y., Lee, D.R., et al. (2015). Histone demethylase expression enhances human somatic cell nuclear transfer efficiency and promotes derivation of pluripotent stem cells. Stem Cell *17*, 758–766.

Clapp, J., Mitchell, L.M., Bolland, D.J., Fantes, J., Corcoran, A.E., Scotting, P.J., Armour, J.A.L., and Hewitt, J.E. (2007). Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. The Am. J. Hum. Genet. *81*, 264–279.

De Paepe, C., Krivega, M., Cauffman, G., Geens, M., and Van de Velde, H. (2014). Totipotency and lineage segregation in the human embryo. Mol Hum Reprod *20*, 599–618.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193–196.

Geng, L.N., Yao, Z., Snider, L., Fong, A.P., Cech, J.N., Young, J.M., van der Maarel, S.M., Ruzzo, W.L., Gentleman, R.C., Tawil, R., et al. (2012). DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. Dev Cell *22*, 38–51.

Gertz, J., Varley, K.E., Davis, N.S., Baas, B.J., Goryshin, I.Y., Vaidyanathan, R., Kuersten, S., and Myers, R.M. (2012). Transposase mediated construction of RNA-seq libraries. Genome Res. *22*, 134–141.

Gifford, W.D., Pfaff, S.L., and Macfarlan, T.S. (2013). Transposable elements as genetic regulatory substrates in early development. Trends Cell Biol. *23*, 218–226.

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., and Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell *16*, 135–141.

Harrison, M.M., Li, X.-Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet *7*, e1002266–13.

Holland, P.W.H., Booth, H.A.F., and Bruford, E.A. (2007). Classification and nomenclature of all human homeobox genes. BMC Biol. *5*, 47.

Ishiuchi, T., and Torres-Padilla, M.-E. (2013). Towards an understanding of the regulatory mechanisms of totipotency. Curr Opin Genet Dev *23*, 512–518.

Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Bošković, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J.M., and Torres-Padilla, M.-E. (2015). Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat Struct Mol Biol. *9*, 662–671

Iwafuchi-Doi, M., and Zaret, K.S. (2014). Pioneer transcription factors in cell reprogramming. Genes Dev. *28*, 2679–2692.

Kalmbach, K., Robinson, L.G., Wang, F., Liu, L., and Keefe, D. (2014). Telomere length reprogramming in embryos and stem cells. BioMed Research International *2014*, 1–7.

Kigami, D., MINAMI, N., Takayama, H., and Imai, H. (2003). MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biol Reprod *68*, 651–654.

Ko, M.S.H. (2016). Zygotic genome activation revisited. In mammalian preimplantation development, (Elsevier), pp. 103–124.

Leidenroth, A., and Hewitt, J.E. (2010). A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol Biol. *10*, 364.

Leidenroth, A., Clapp, J., Mitchell, L.M., Coneyworth, D., Dearden, F.L., Iannuzzi, L., and Hewitt, J.E. (2012). Evolution of DUX gene macrosatellites in placental mammals. Chromosoma *121*, 489–497.

Liu, L., Bailey, S.M., Okuka, M., Muñoz, P., Li, C., Zhou, L., Wu, C., Czerwiec, E., Sandler, L., Seyfang, A., et al. (2007). Telomere lengthening early in development. Nat. Cell Biol. *9*, 1436–1441.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. *25*, 594–607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Madissoon, E., Jouhilahti, E.-M., Vesterlund, L., Töhönen, V., Krjutškov, K., Petropoulous, S., Einarsdottir, E., Linnarsson, S., Lanner, F., Månsson, R., et al. (2016). Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep 1–13.

Matoba, S., Liu, Y., Lu, F., Iwabuchi, K.A., Shen, L., Inoue, A., and Zhang, Y. (2014). Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. Cell *159*, 884–895.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-

regulatory regions. Nat Biotechnol *28*, nbt.1630–nbt.1639.

Morgani, S.M., and Brickman, J.M. (2014). The molecular underpinnings of totipotency. Philos Trans R Soc Lond, B, Biol Sci *369*, 20130549–20130549.

Patro, R., Mount, S.M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol *32*, 462–464.

Ribet, D., Louvet-Vallée, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., Pierron, G., Maro, B., and Heidmann, T. (2008). Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. J Virol *82*, 1622–1625.

Schoorlemmer, J., Pérez-Palacios, R., Climent, M., Guallar, D., and Muniesa, P. (2014). Regulation of mouse retroelement MuERV-L/MERVL expression by REX1 and epigenetic control of stem cell potency. Front Oncol. *4*.

Sun, Y., Nien, C.-Y., Chen, K., Liu, H.-Y., Johnston, J., Zeitlinger, J., and Rushlow, C. (2015). Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophilazygotic genome activation. Genome Res. *25*, 1703–1714.

Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E.-M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T.R., et al. (2015). Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun *6*, 8207.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652–657.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. *20*, 1131-9.

Yasuda, T., Tsuzuki, S., Kawazu, M., Hayakawa, F., Kojima, S., Ueno, T., Imoto, N., Kohsaka, S., Kunita, A., Doi, K., et al. (2016). Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. Nat Genet *48*, 569–574.

Young, J.M., Whiddon, J.L., Yao, Z., Kasinathan, B., Snider, L., Geng, L.N., Balog, J., Tawil, R., van der Maarel, S.M., and Tapscott, S.J. (2013). DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLoS Genet *9*, e1003947.

Zalzman, M., Falco, G., Sharova, L.V., Nishiyama, A., Thomas, M., Lee, S.-L., Stagg, C.A., Hoang, H.G., Yang, H.-T., Indig, F.E., et al. (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature *464*, 858–863.

Zhang, J., McCastlain, K., Yoshihara, H., Xu, B., Chang, Y., Churchman, M.L., Wu, G., Li, Y., Wei, L., Iacobucci, I., et al. (2016). Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet *48*, 1481–1489.

Zhou, L.-Q., and Dean, J. (2015). Reprogramming the genome to totipotency in mouse embryos. Trends Cell Biol. *25*, 82–91.

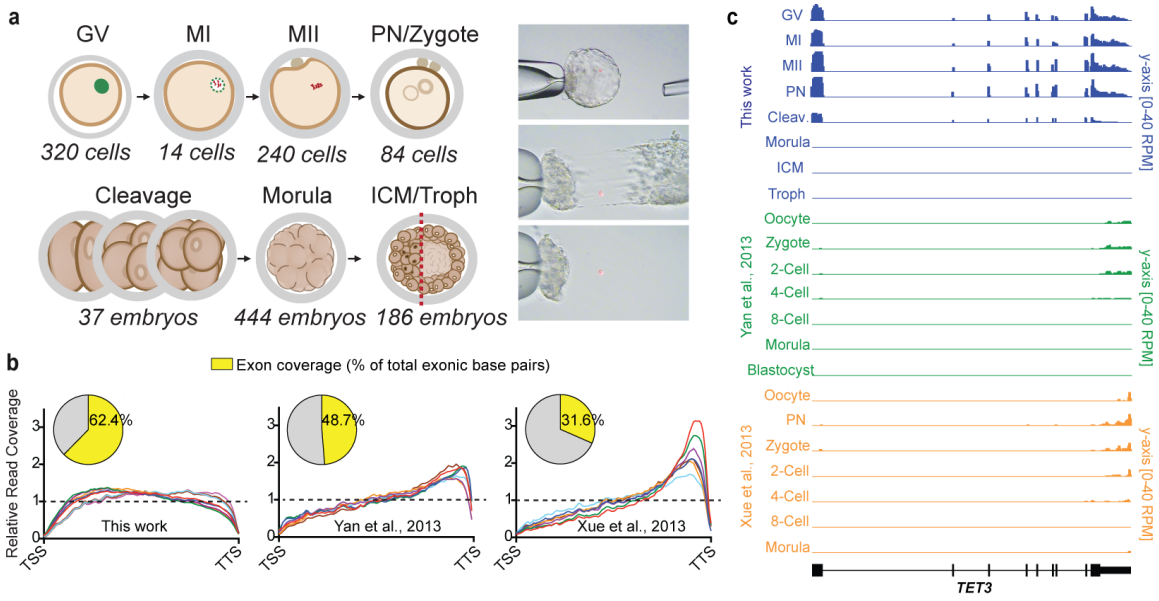

Figure 2.1 Transpose-mediated RNA-sequencing of human oocytes and embryos improves read coverage balance. (a) Summary of the human oocyte and embryonic stages (and cell numbers) collected (left panel), and depiction of the laser mechanical separation of day 5-6 blastocysts into ICM and mural trophectoderm (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) Screenshot of the *TET3* gene, as an example of a genomic locus displaying read coverage bias in previous single cell datasets (Yan et al., 2013 in green; Xue et al., 2013 in orange).

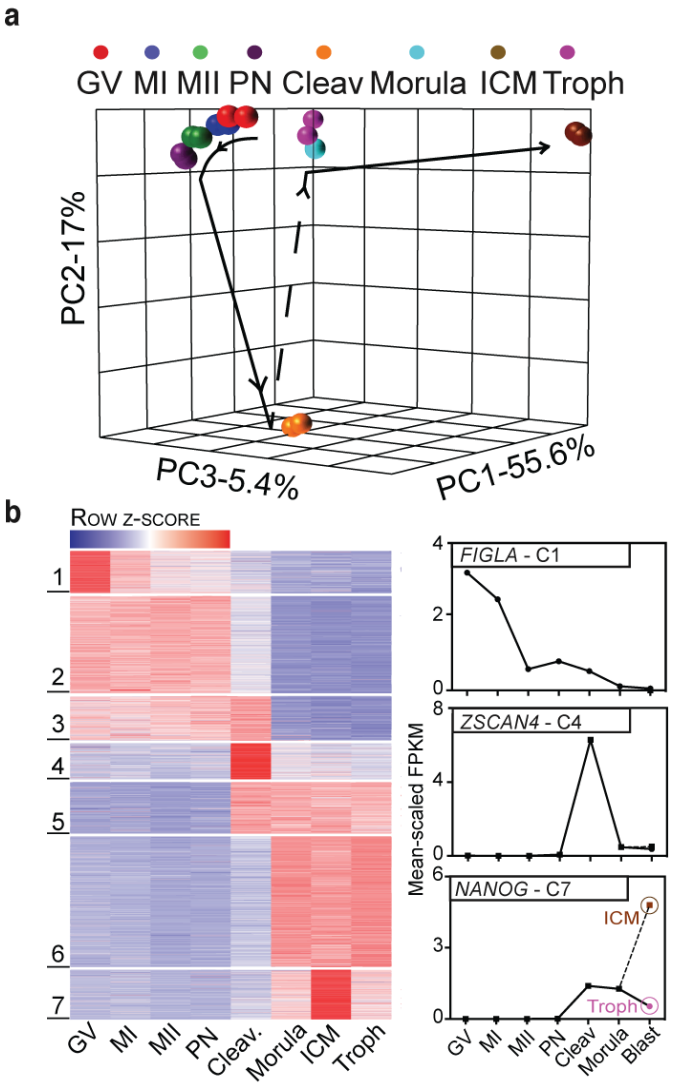

Figure 2.2 Stage-specific gene expression in human oocyte and embryo. (a) Principal component analysis (PCA) of all egg and embryonic stages based on the highest 50% of all expressed genes (>1 mean FPKM). (b) Statistically determined k-means clusters based on the highest 50% all expressed genes (left panel). Clusters 1, 4, and 7 exhibit stage-specific gene expression and contain prominent developmentally important genes, *FIGLA*, *ZSCAN4*, and *NANOG*, respectively (right panel).

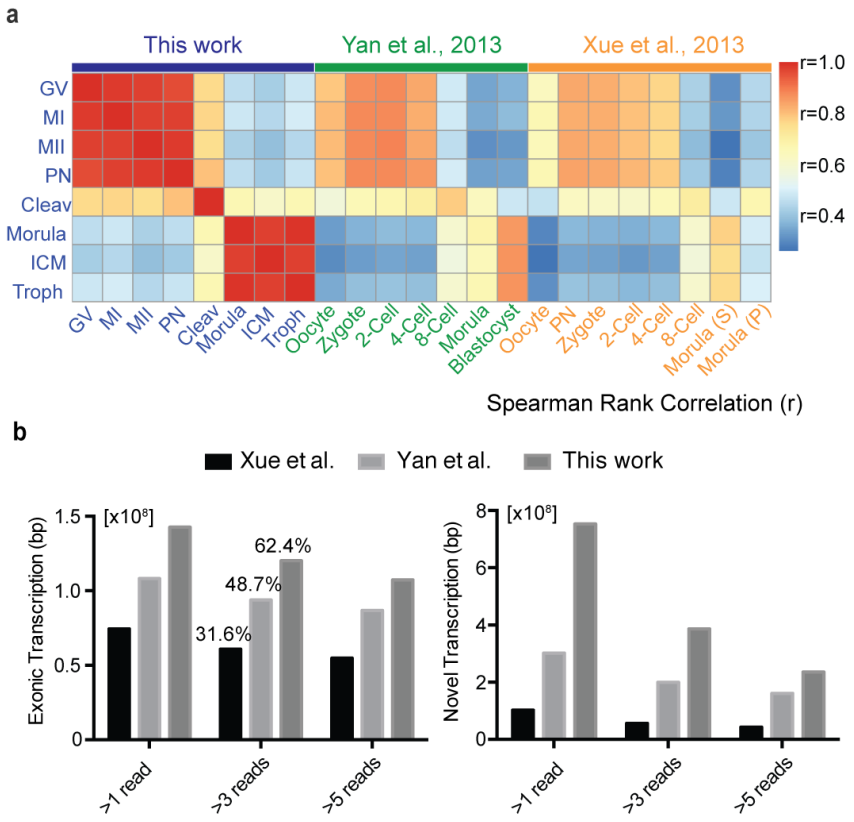

Figure 2.3 Improvements in read coverage enable the discovery of new novel transcription. (a) Gene expression correlations using per-stage average FPKM data; r-values were calculated using a spearman rank statistic. S: single cell; P: pooled cells. (b) Bar graphs comparing total exonic transcription (left panel) and novel transcription (right panel) measured in base pairs; employing thresholds of >1, >3 or >5 reads per region. Exonic transcription encompasses all base pairs annotated by Ensemble, UCSC, and NONCODE.

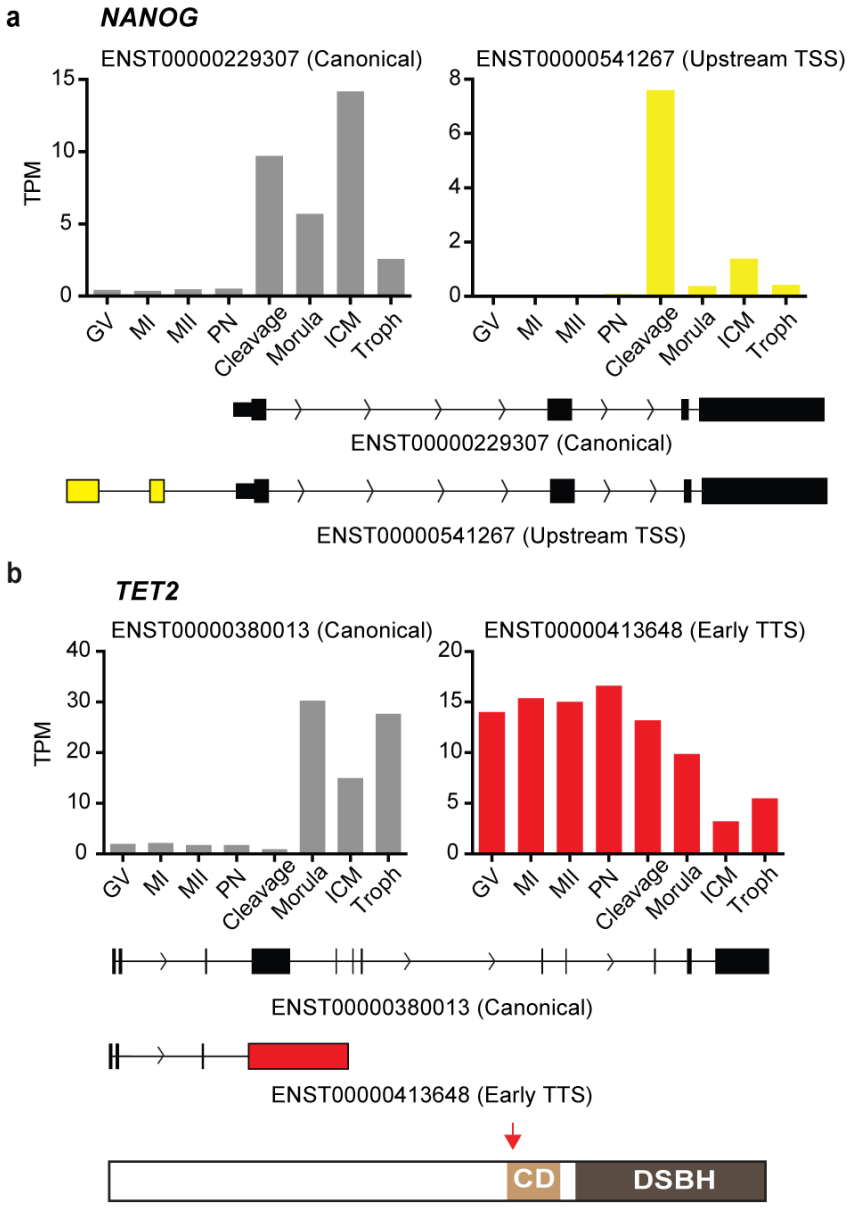

Figure 2.4 Improvements in read coverage enable the discovery of novel splice isoform expression. (a) A non-canonical *NANOG* isoform is expressed specifically in the cleavage stage. (b) A non-canonical *TET2* isoform is maternally loaded encoding a severely truncated protein product that excludes both known functional domains [CD-Cys-rich domain; DSBH-Double-stranded β-helix dioxygenase domain].

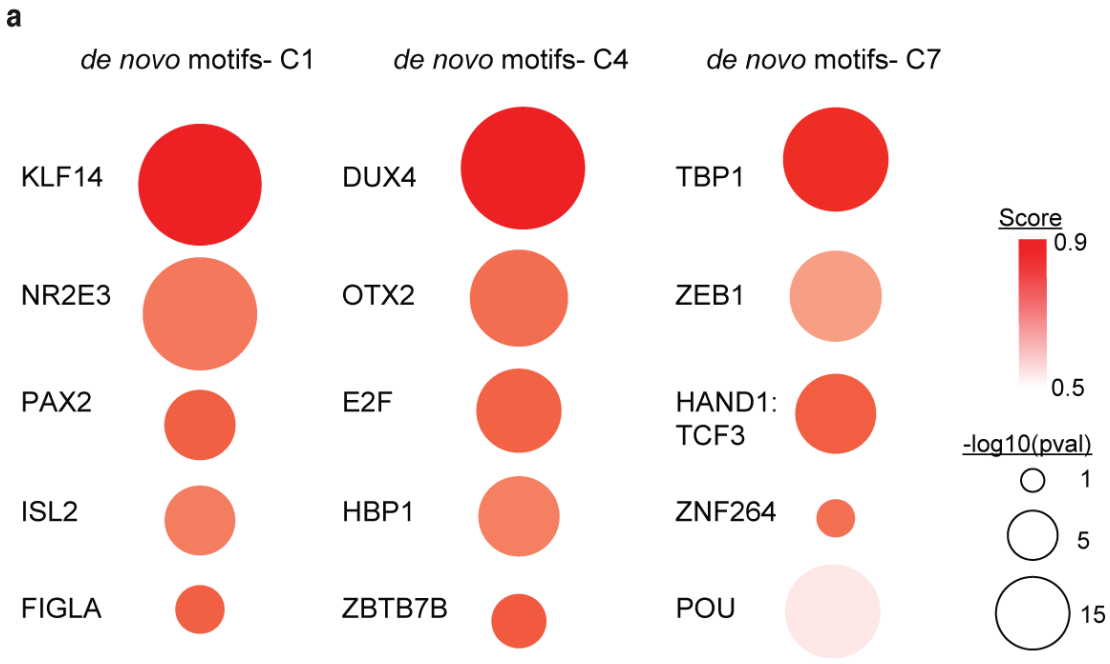

Figure 2.5 Enriched motifs in stage-specific gene clusters. (a) The top five *de novo* motifs enriched in cluster 1 (left panel), cluster 4 (middle panel), and cluster 7 (right panel) gene promoters after filtering for match score (>0.70). *Note- an OCT/POU-like motif was highly enriched in cluster 7; however, it fell below the score cutoff (0.61).

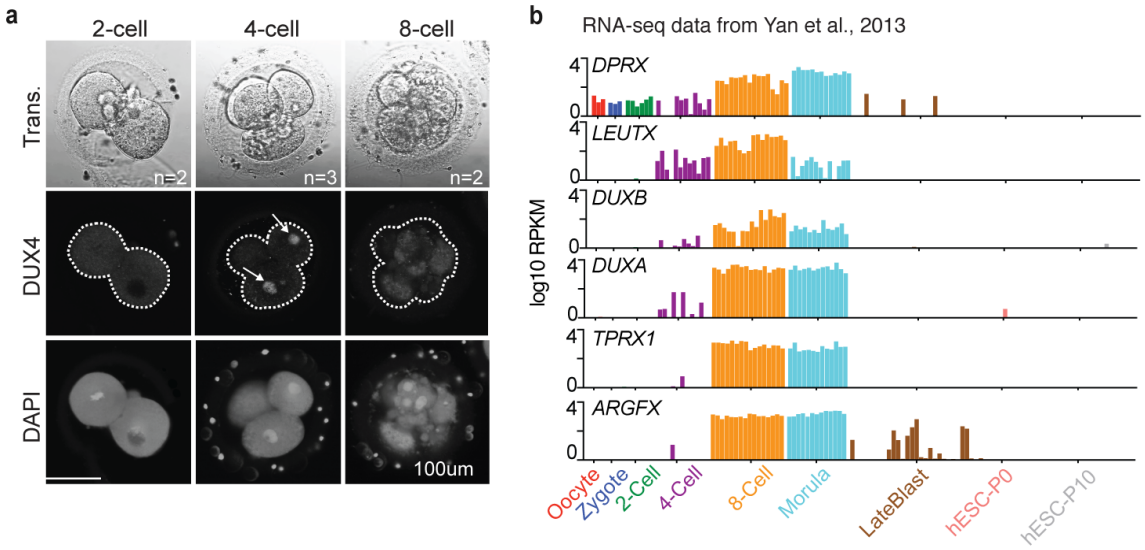

Figure 2.6 DUX4 and PRD-like gene expression in the early human embryo. (a) Immunofluorescence of DUX4 protein in human 2-cell, 4-cell, and 8-cell embryos (n=7). (Note: though only one plane is shown, expression was restricted to nuclei of the 4-cell stage, indicated with arrows). (b) Single cell expression data (RPKM) for notable double homeobox and 'PRD-like' genes.

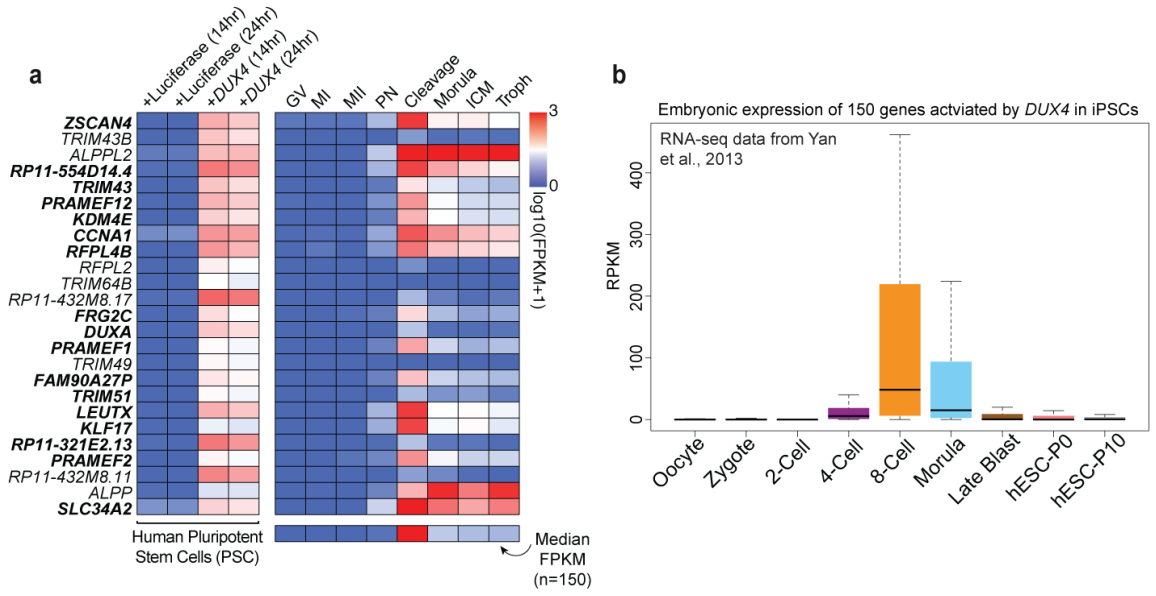

Figure 2.7 DUX4 activates genes transiently and specifically expressed in the human cleavage stage embryo. (a) Heatmap depicting the top 25 DUX4-activated genes in human iPSCs and their expression in the embryo [*two replicates per condition*]. Bold font indicates genes belonging to cluster 4 (see Fig. 1d). The bottom row of the heatmap depicts the median embryonic expression of all 150 genes upregulated following *DUX4* expression. (b) Box plot displaying the embryonic expression of the 150 common genes that are upregulated following *DUX4* overexpression (for 14hrs or 24hrs) in iPSCs.

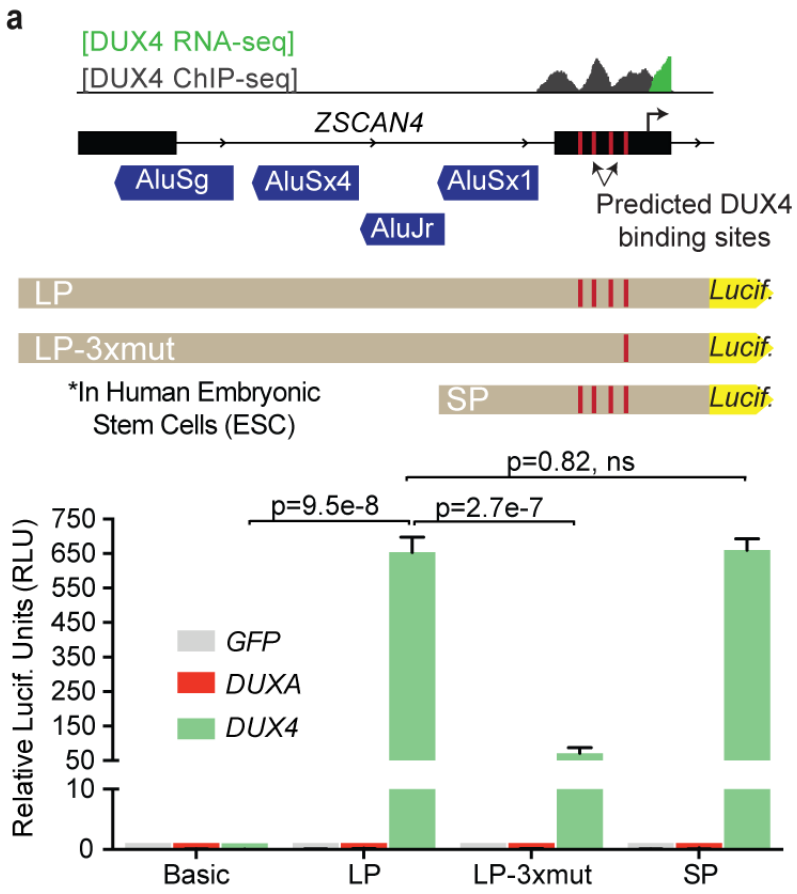

Figure 2.8 DUX4 directly activates ZSCAN4 in human embryonic stem cells. (a) A diagram of the *ZSCAN4* promoter/TSS and the position of the DUX4 ChIP occupancy in *DUX4*-expressing myoblasts (top panel). *ZSCAN4* activation is dependent on DUX4 binding (bottom panel) [four biological replicates per condition. *Statistics determined using a two-tailed unpaired t-test*. *Error bars, s.d.*].

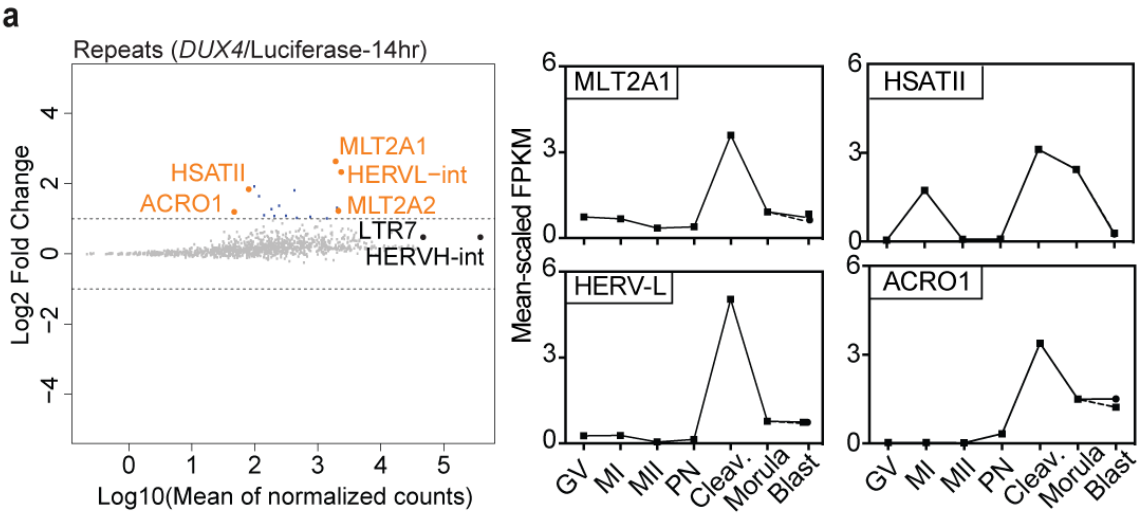

Figure 2.9 DUX4 activates repeats transiently and specifically expressed in the human cleavage stage embryo. (a) MA-plot showing *DUX4*-mediated induction of specific repeat elements, by subfamily (left panel). Mean-scaled expression of top activated repeats: MLT2A1, HERVL, HSATII, and ACRO1 in human oocytes and embryos (right panel).

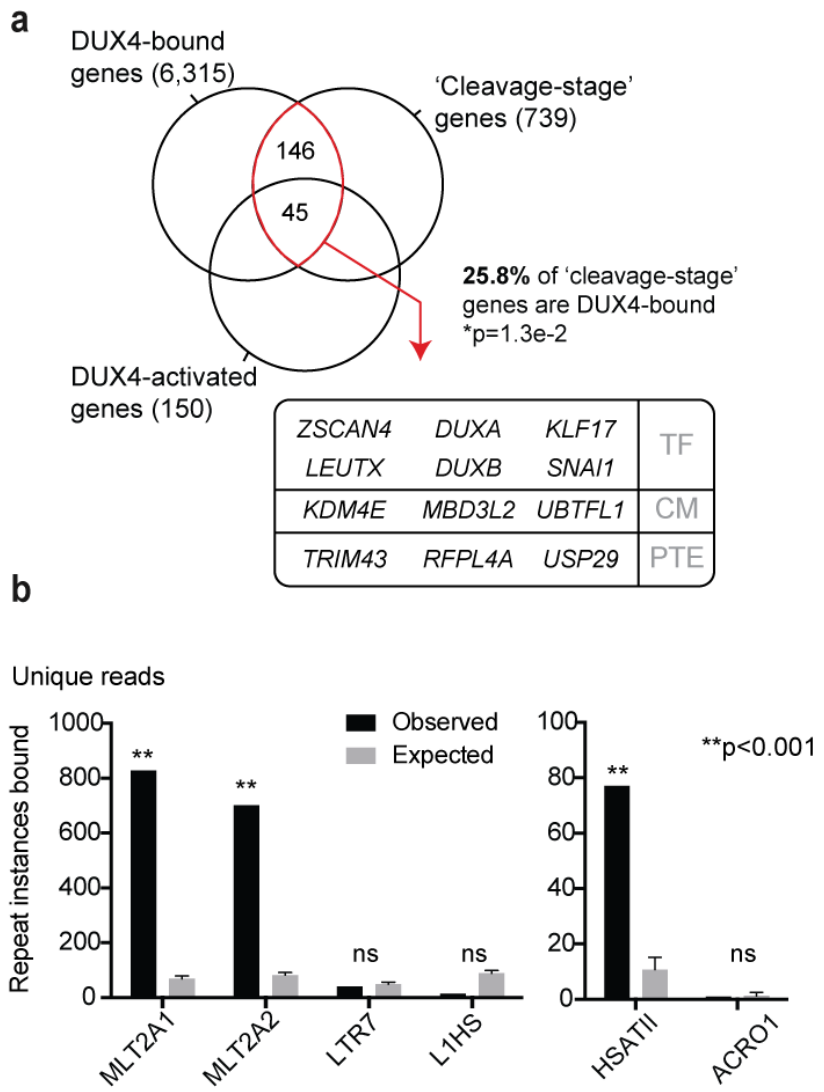

Figure 2.10 DUX4 directly activates gene and repeat element transcription. (a) The overlap of DUX4-ChIP occupied genes [*two replicates*] with genes enriched in the cleavage-stage embryo and activated by *DUX4*-overexpression in iPSCs [*Overlap statistic calculated by hypergeometric test. Note - only 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. (b) The number of repeat element instances uniquely bound by DUX4 for select activated (MLT2A1, MLT2A2, HSATII) and unaffected (LTR7, L1) subfamilies [two ChIP replicates. *Enrichment statistic determined empirically; error bars, s.d.*].

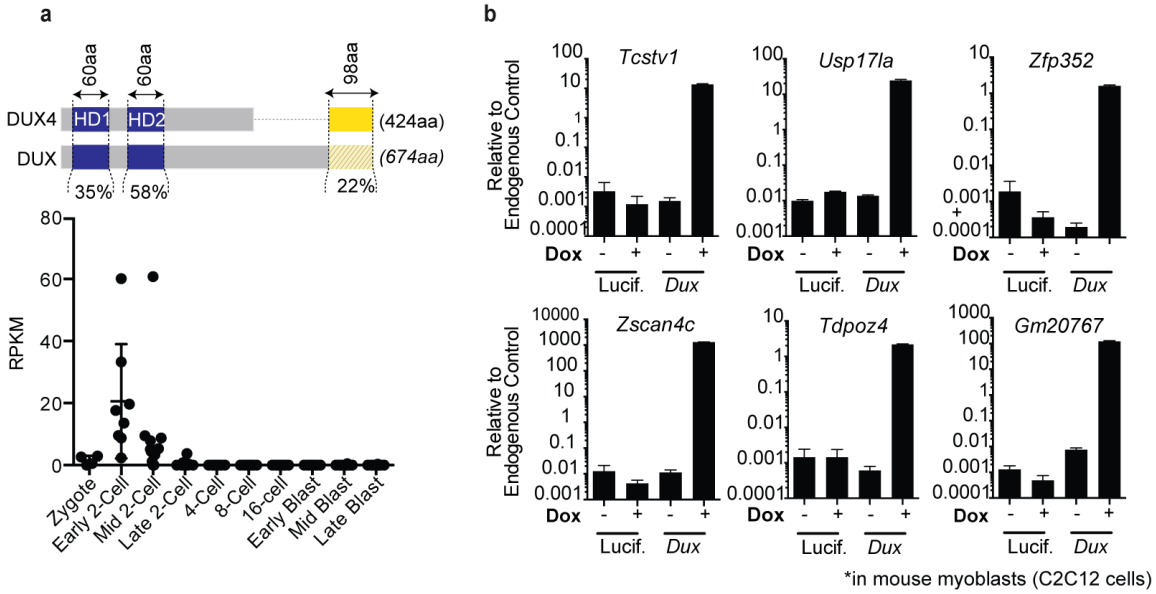

Figure 2.11 Mouse *Dux* is expressed in the 2-cell stage embryo and activates notable cleavage stage genes. (a) Depiction of human DUX4 and mouse DUX amino acid sequence comparison (top panel) and the normalized expression of *Dux* in pre-implantation mouse embryos (RNA-seq data from Deng et al., 2014) (bottom panel). (b) RT-qPCR data for select cleavage stage genes activated following *Dux* expression in mouse C2C12 cells [*three replicates per condition. Error bars, s.d.*].

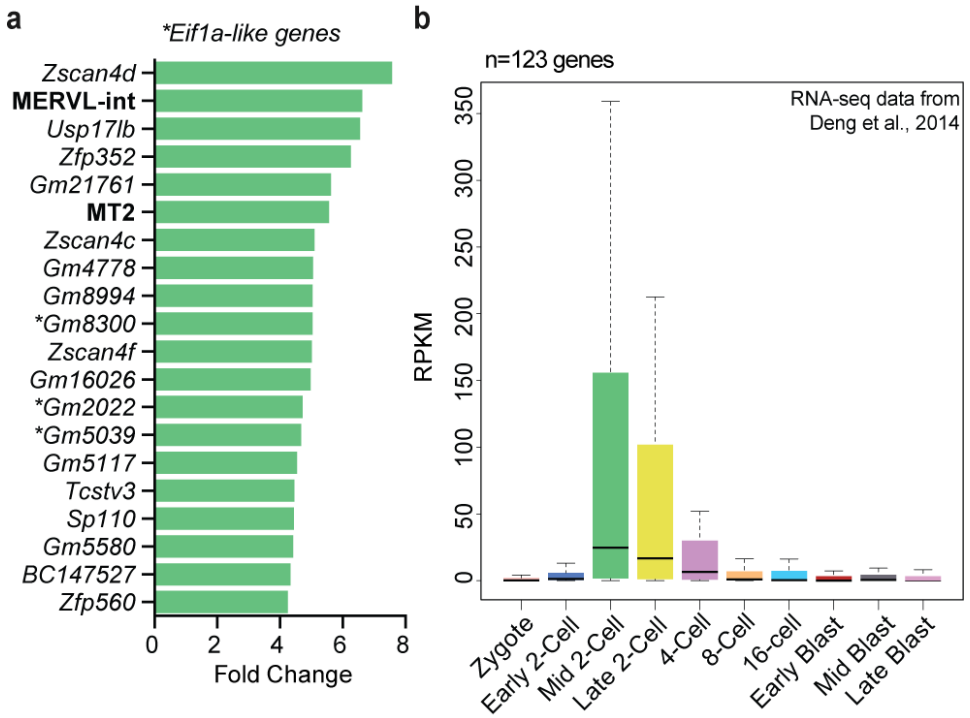

Figure 2.12 Mouse *Dux* activates a '2C' transcriptional program in mouse embryonic stem cells. (a) Bar graph displaying the top 15 differentially-expressed genes and repeat elements (bold) following ectopic *Dux* expression in mouse embryonic stem cells (mESCs) [*two replicates per condition*]. (b) Relative expression of *Dux*-induced genes (n=123) in the pre-implantation mouse embryo.

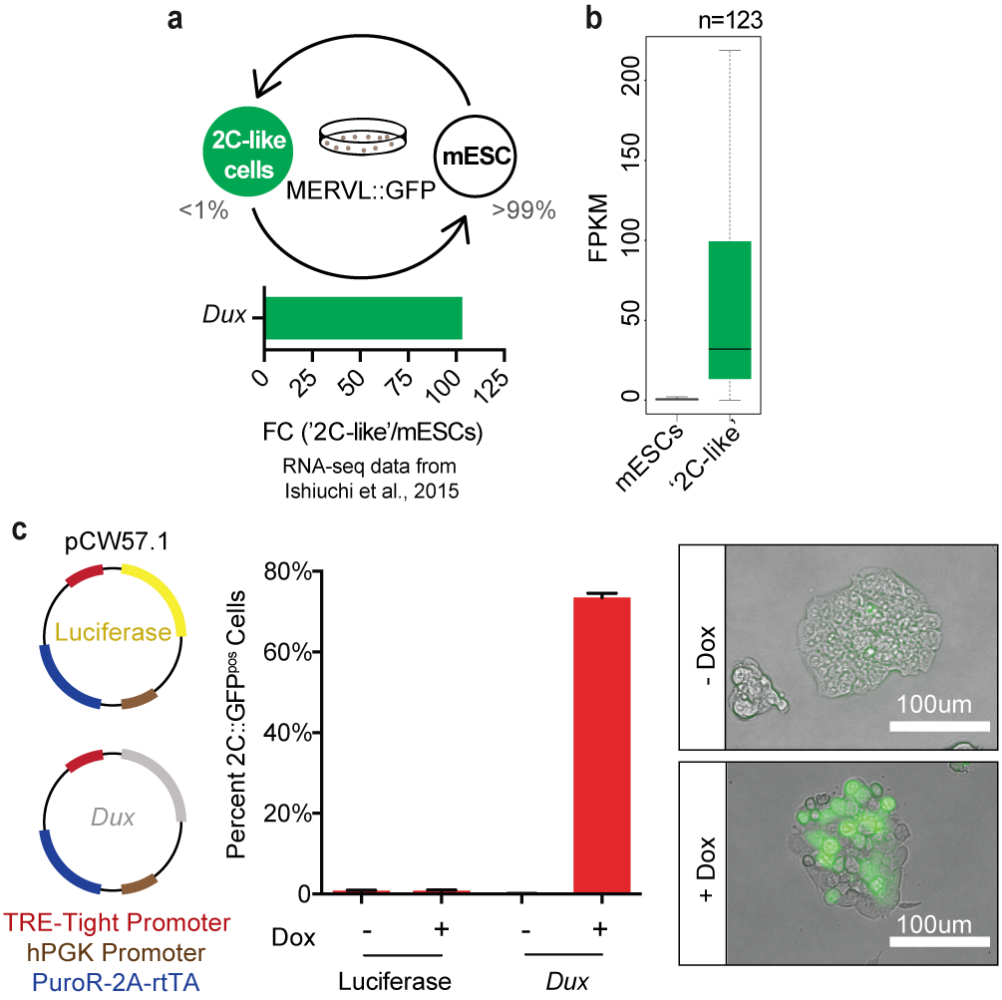

Figure 2.13 *Dux* expression converts mouse embryonic stem cells to '2C-like' cells. (a) Diagram of mESC metastability (top panel) and the enrichment of *Dux* in '2C-like' cells relative to conventional mESCs (bottom panel). (b) Expression of *Dux*-induced genes (n=123) in '2C-like' cells compared to conventional mESCs. (c) 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*].

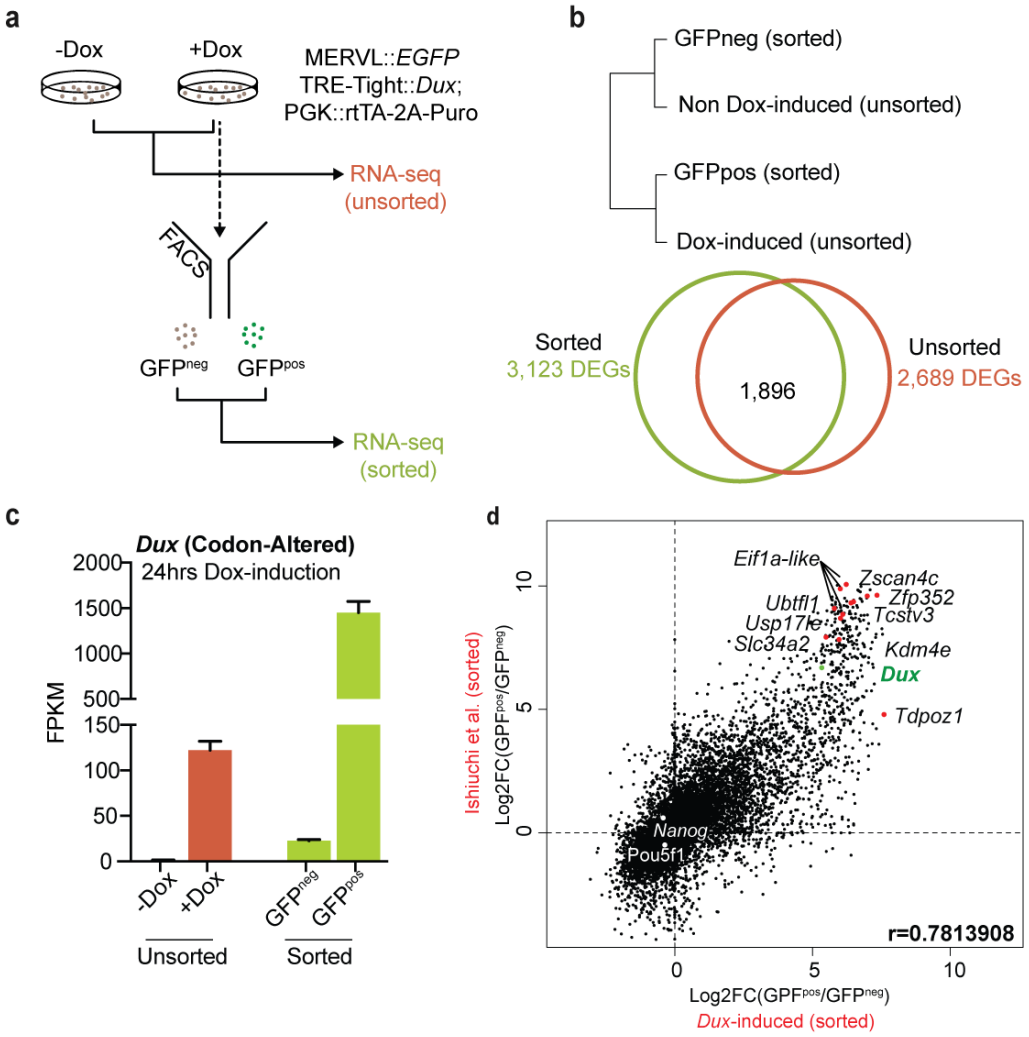

Figure 2.14 RNA-sequencing of *Dux*-induced mouse embryonic stem cells and '2C-like' cells. (a) Schematic of the RNA-seq experiments conducted on *Dux*-expressing mESCs. (b) Overlap of differentially expressed genes (DEGs) from unsorted and sorted populations of *Dux*-expressing mESCs. (c) The normalized expression of codon altered *Dux* transgene in our RNA-seq datasets from unsorted and sorted populations. (d) 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.

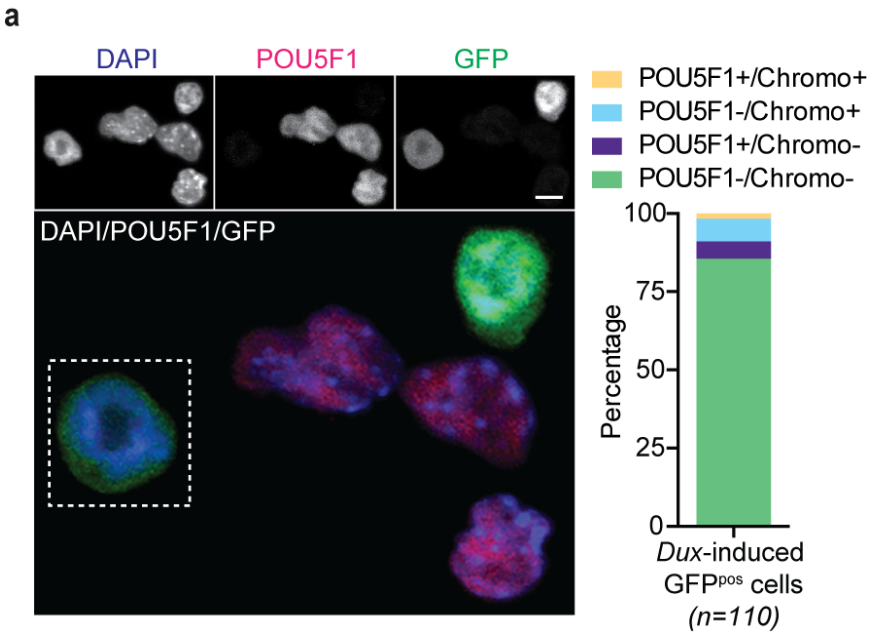

Figure 2.15 *Dux*-induced '2C-like' cells deactivate pluripotency network and lose chromocenters. (a) Immunofluorescence quantifying the loss of pluripotency (e.g. POU5F1 protein) and chromocenters in mESCs following ectopic *Dux* expression (n =110 cells). Scale bar, 10um.

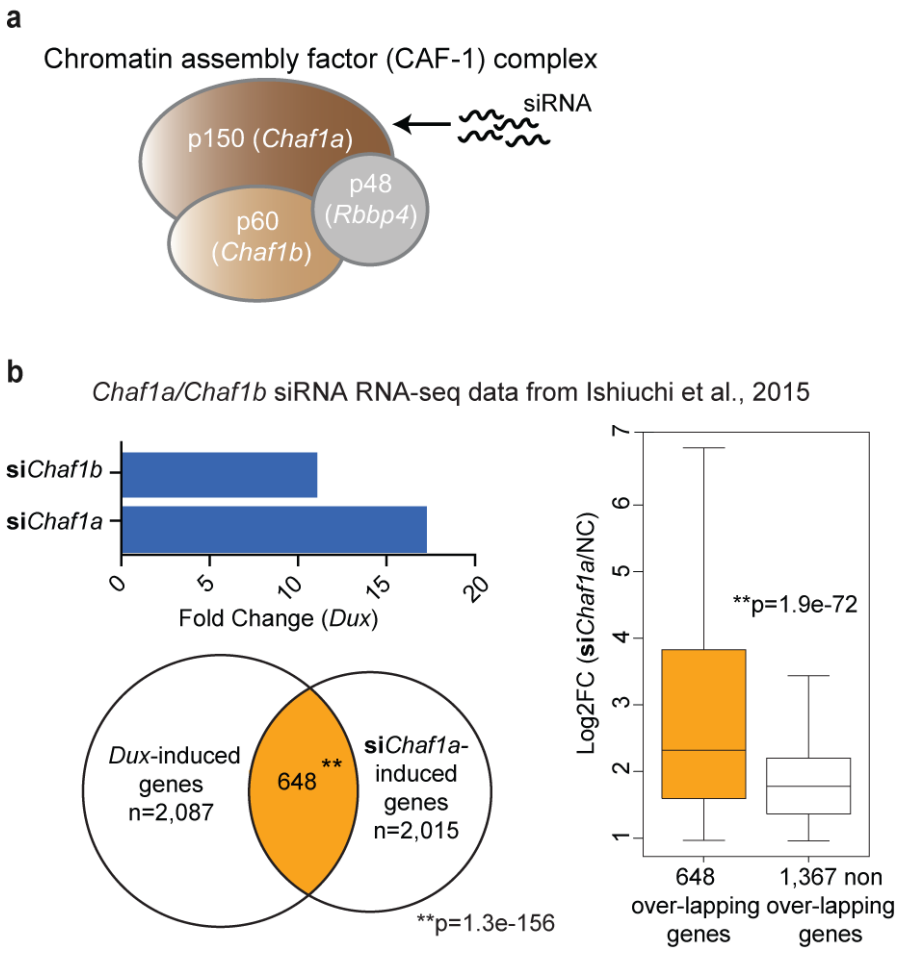

Figure 2.16 CAF-1 knockdown permits *Dux* expression and activity. (a) A diagram of the Chromatin Assemble Factor (CAF-1) complex. The arrow points to the complex subunit (p150 encoded by the *Chaf1a* gene) targeted with siRNAs in our experiments. (b) *Dux* is highly upregulated in CAF-1-depleted mESCs (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*.]

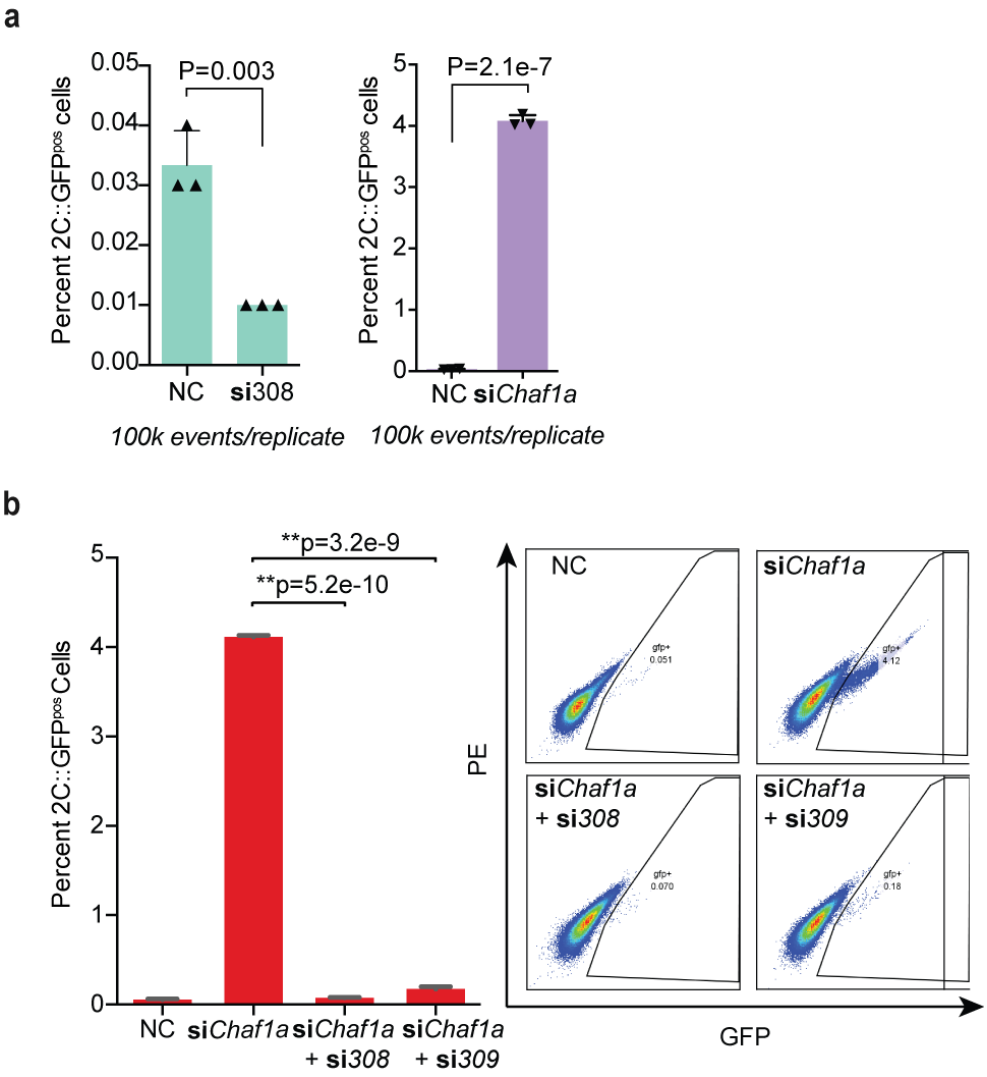

Figure 2.17 *Dux* is necessary for spontaneous and CAF-1-mediated conversion of mESCs to a '2C-like' state. (a) Effects of *Dux* knockdown alone (left panel) and *Chaf1a* knockdown alone (right panel) on conversion of mESCs to a '2C-like' state [three biological replicates per condition. *Statistics determined using a two-tailed unpaired t-test. Error bars, s.d.*]. (b) 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*].

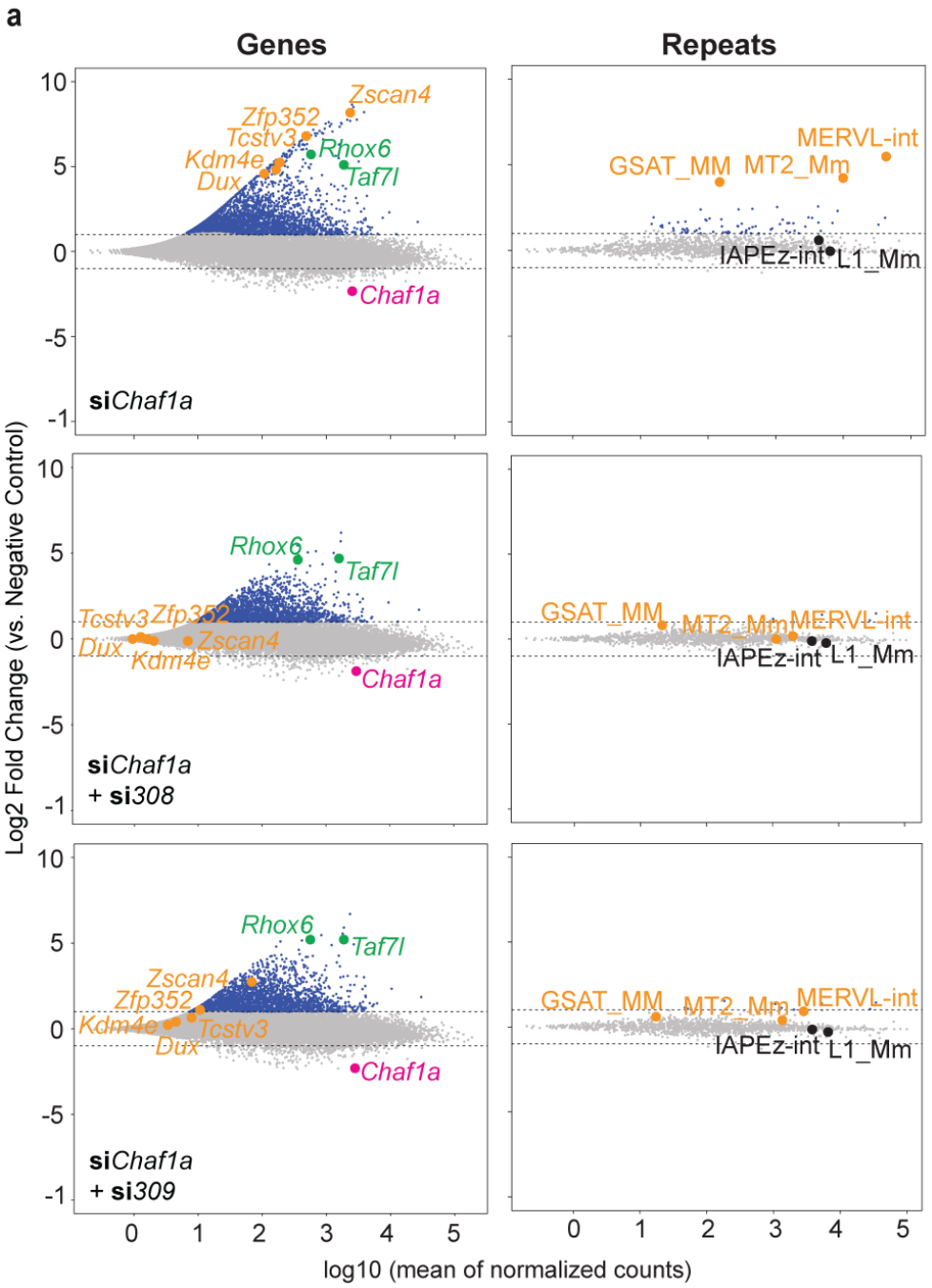

Figure 2.18 *Dux* knockdown impairs CAF1-knockdown-mediated gene and repeat expression. (a) MA-plots show changes in gene and repeat element expression (by subfamily) in mESCs following knockdown of *Chaf1a* alone (top panel) and in combination with *Dux* (si308-middle panel; si309-bottom panel).

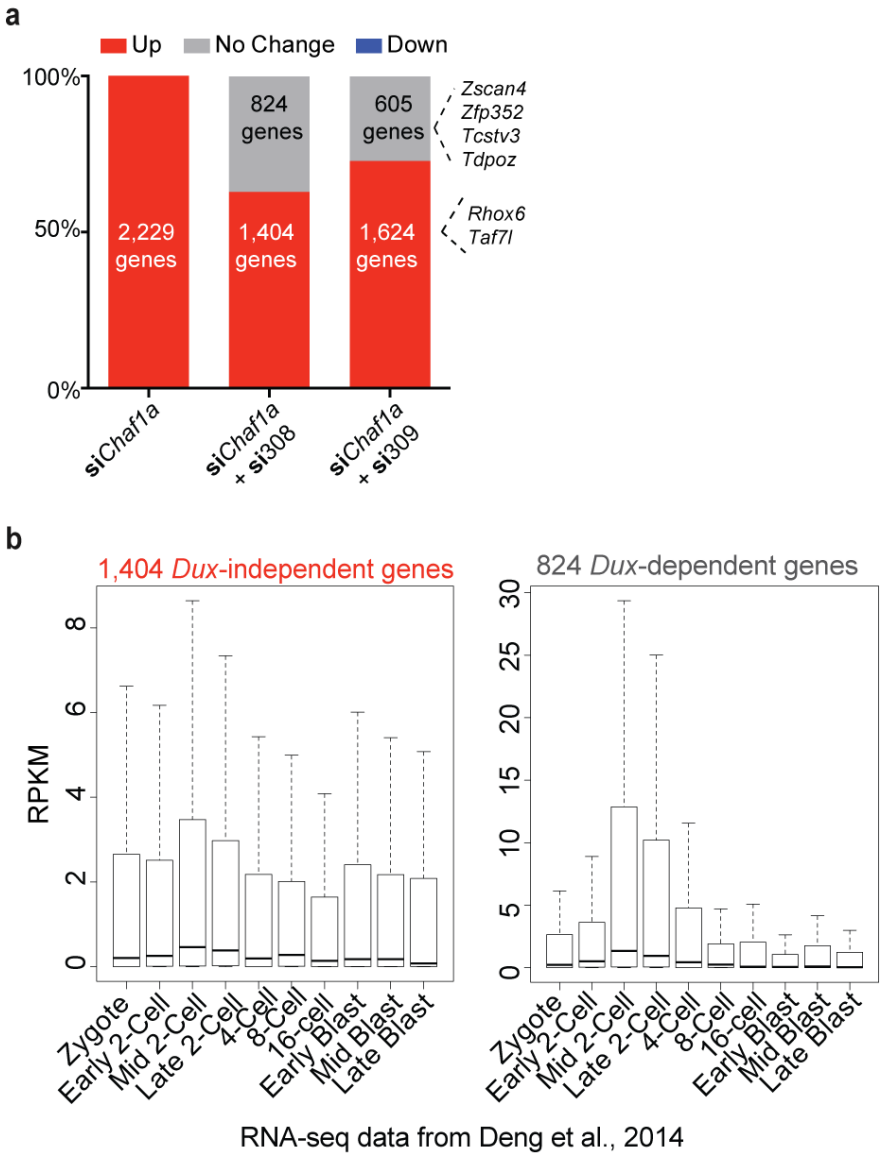

Figure 2.19 CAF1-knockdown-mediated '2C' gene expression is *Dux*-dependent. (a) Bar chart showing the fraction of genes upregulated (FC>2, FDR<0.01) in *Chaf1a* depleted mESCs that are not affected in mESCs depleted for both *Chaf1a* and *Dux*. (note: one gene that was upregulated in *Chaf1a* depleted mESCs became downregulated in mESCs depleted for both *Chaf1a* and *Dux*). (b) Boxplot showing the embryonic expression of the genes upregulated in both *Chaf1a*-depleted as well as *Chaf1a*- and *Dux*-depleted mESCs (termed '*Dux*-independent') and the genes upregulated only in *Chaf1a*-depleted cells (termed '*Dux*-dependent').

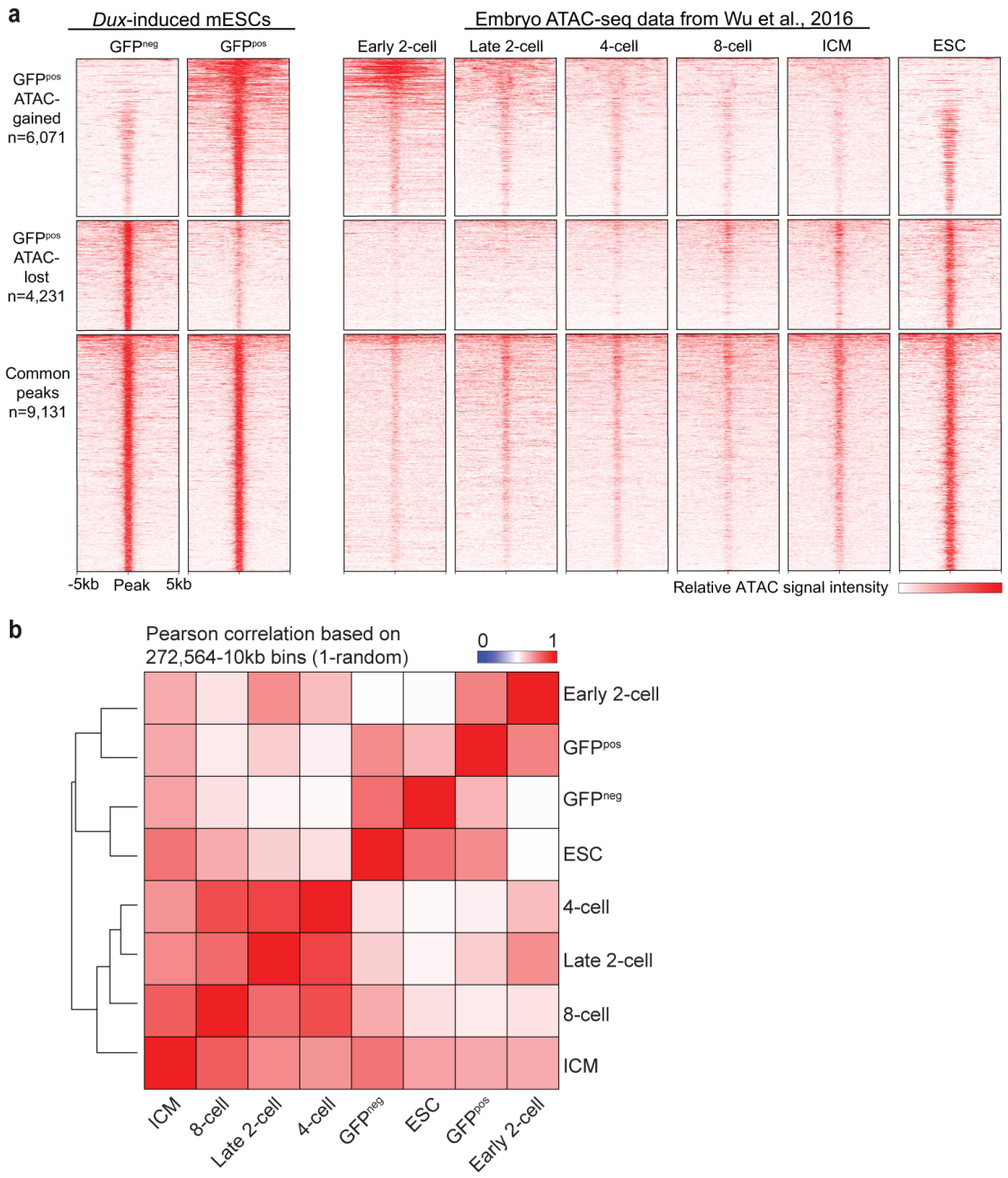

Figure 2.20 *Dux*-induced '2C-like' cells acquire an open chromatin landscape that resembles that of an early 2-cell stage embryo. (a) Heatmaps display regions of ATAC-seq signal gain, loss, and found in common between *Dux*-induced GFP^{pos} and GFP^{neg} cell populations [*Two replicates per condition*]. *Dux*-induced GFP^{pos} cells acquire an open/closed chromatin landscape that resembles the early 2-cell stage embryo. a) Heatmap depicting the Pearson correlation of genome-wide ATAC-seq coverage profiles in *Dux*-induced mESCs and early embryonic developmental stages (Embryo ATAC-seq data from Wu et al., 2016) [two replicates per condition].

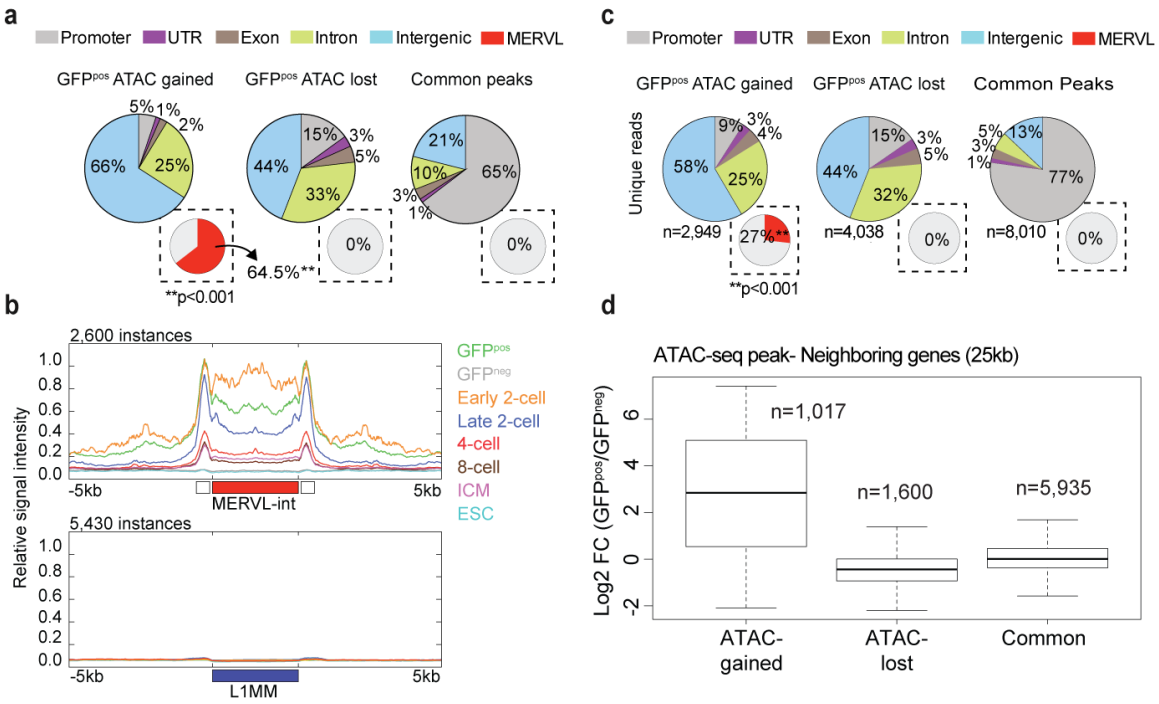

Figure 2.21 The open chromatin landscape in *Dux*-induced '2C-like' cells overlaps significantly with MERVL instances and coincides with transcriptional activation. (a) Pie charts depicting the distribution of ATAC-seq gained, lost and common peaks at basic genomic features. Inset pie charts indicate the percentage of peaks that overlap with MERVL elements (MT2_Mm and MERVL-int) [*Enrichment statistic determined empirically*]. (b) 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. (c) Pie charts depicting the distribution of ATAC-seq gained, lost and common peaks (called after filtering alignment files for unique reads only) at basic genomic features. Inset pie charts indicate the percentage of unique peaks which overlap with MERVL elements (MT2_Mm and MERVL-int) [*Enrichment statistic determined empirically*]. (d) Boxplot shows the median log2 fold change (FC) of the genes neighboring regions of ATAC-seq gained, lost and common signal.

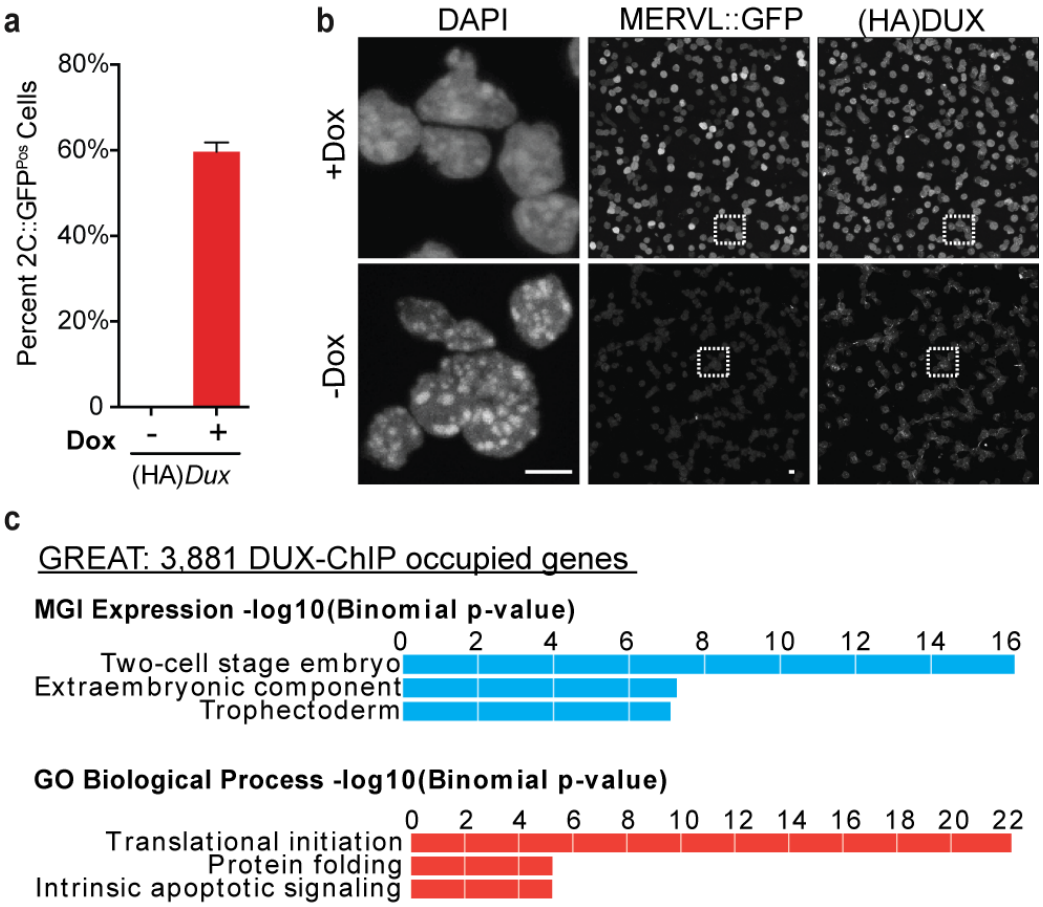

Figure 2.22 Generation of a clonal mESC line expressing an HA-tagged DUX for ChIP-seq. (a) Flow results demonstrating, in an independent HA-tagged clone, the ability of *Dux* expression to efficiently induce reactivation of the MERVL reporter in mESCs [three biological replicates per condition]. (b) The expression of HA and loss of chromocenters is evaluated by immunofluorescence confirming entry into a '2C-like' state. Scale bar, 10um. (c) Top enriched 'MGI expression' and 'Gene Ontology (GO)' terms identified in the 3,881 genes occupied by DUX [*two replicates*].

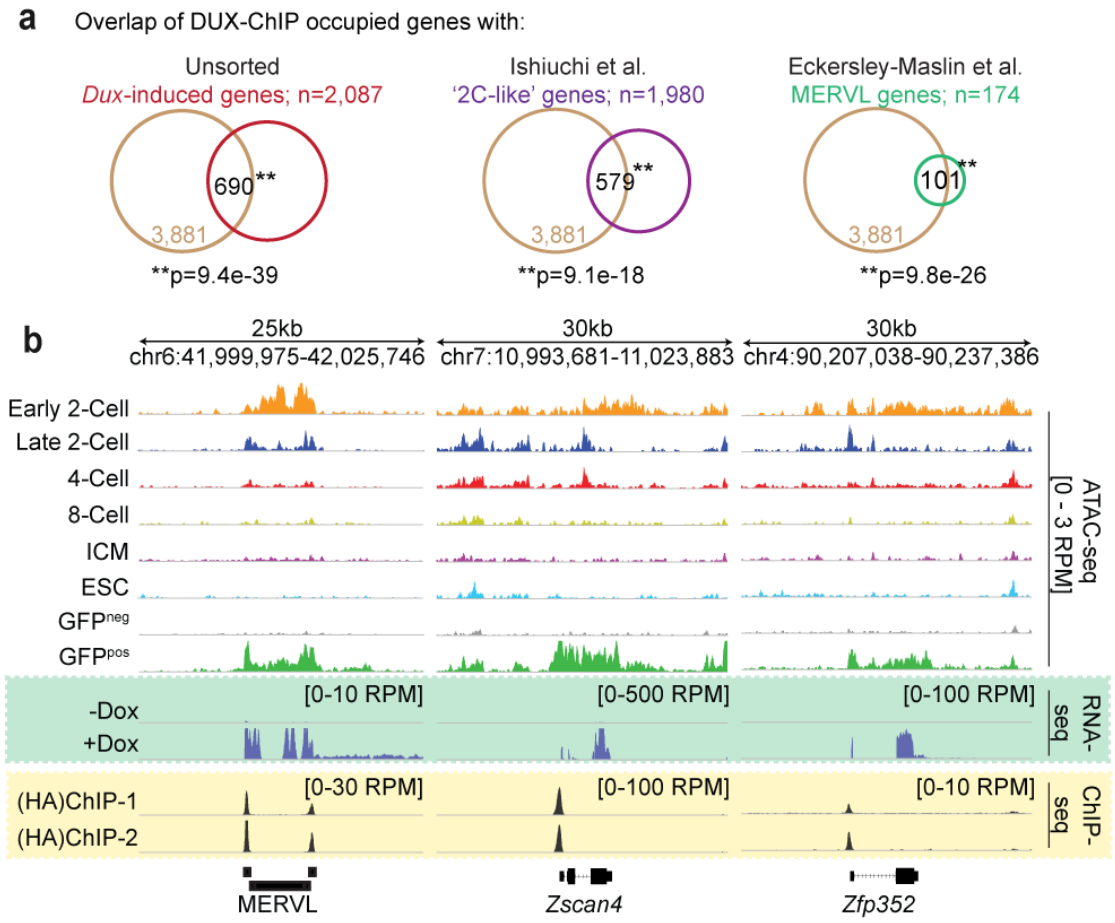

Figure 2.23 DUX directly activates '2C' and '2C-like' gene expression. (a) Overlap of DUX-ChIP occupied genes with genes: upregulated in unsorted mESCs post *Dux* overexpression (left panel); enriched in '2C-like' cells (middle panel); and driven by MERVL elements (right panel) *[Statistics determined by hypergeometric test*]. (b) Screenshots demonstrating the overlap of DUX-ChIP occupancy (yellow box) with the acquisition of 2-cell embryo-like open chromatin and gene expression (green box).

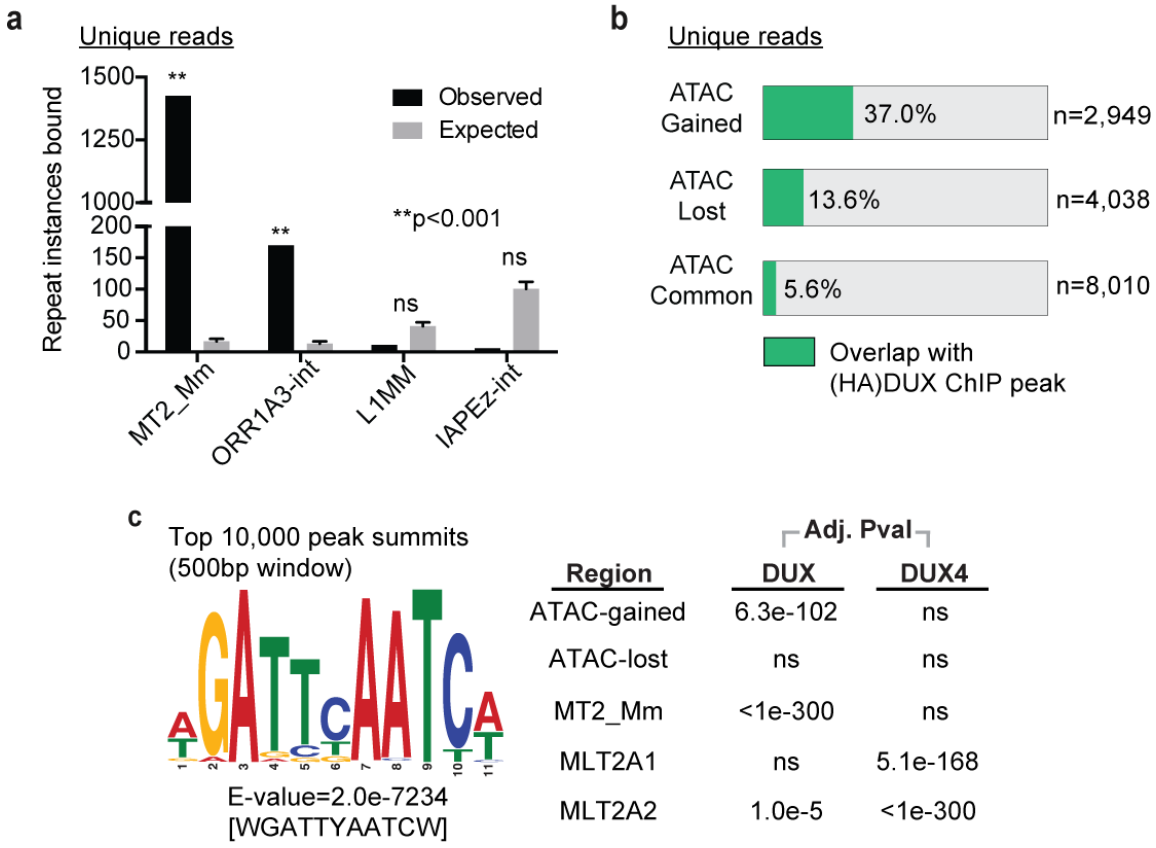

Figure 2.24 DUX directly binds to MERVL retrotransposons and influences chromatin accessibility. (a) The number of repeat element instances uniquely bound by DUX for select affected (MT2_Mm, ORR1A3-int) and unaffected (L1, IAPEZ-int) subfamilies [two ChIP replicates. *Enrichment statistic determined empirically; error bars, s.d.*]. (b) The percentage of unique ATAC gained, lost, and common regions bound by DUX. (c) A binding motif for DUX predicted by MEME-ChIP based on the top 10,000 peak summits (left panel). This motif differs from that predicted for DUX4, and only shows enrichment in mouse-specific regions of interest (right panel).

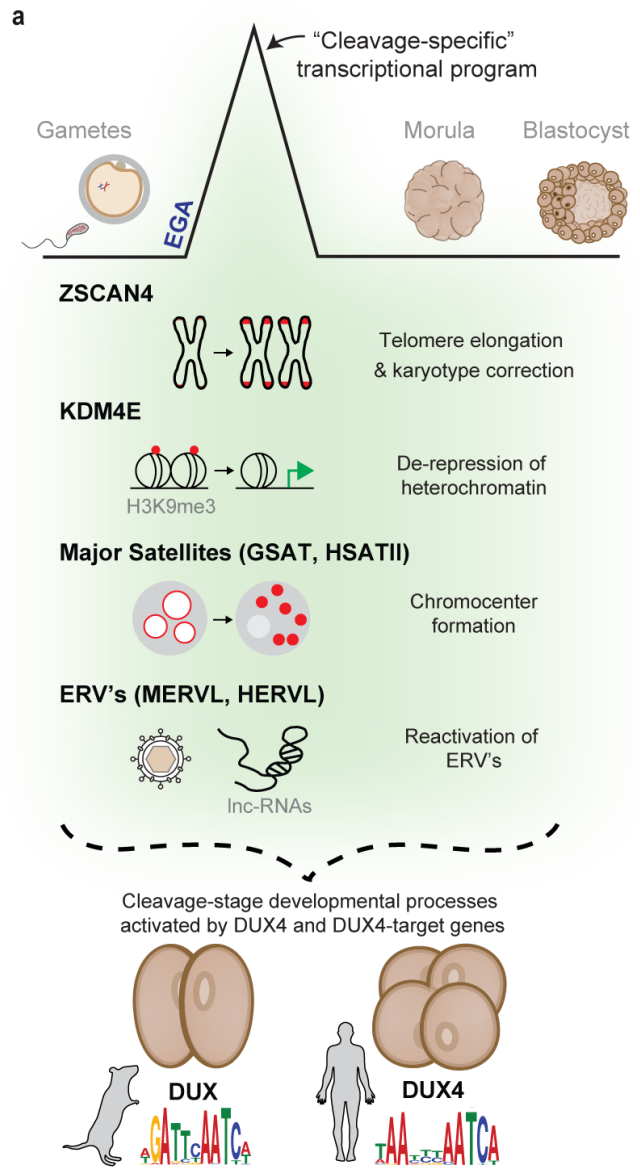

Figure 2.25 A model of DUX4 function during cleavage. (a) A cleavage-specific transcriptional program is activated at EGA in mouse and human cells by DUX or DUX4, respectively. The genes and repetitive elements activated by these DUX4-family genes mediate important molecular events associated with embryonic 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.

CHAPTER 3

CONCLUSIONS AND PERSPECTIVES

3.1 Preface

Two of the biggest questions regarding mammalian embryonic development are 1) how does the embryonic genome get activated, and 2) what is the mechanistic basis for embryonic reprogramming to totipotency? A third potential question is how, if at all, these processes are interconnected. The story I presented above, in short, identifies a conserved eutherian transcription factor family (DUX4-family) whose expression and downstream transcriptional program in multiple mammalian species align with both. It stops short, however, of directly linking *DUX4*-family genes to either EGA or to totipotency conversion -- a difficult but important next step. In this chapter, I will further explore these questions and connections and describe a few of the experiments necessary to bridge this gap.

3.2 A mammalian Zelda?

The activation of the embryonic genome (EGA) is a critical developmental event in all animals. The timing of genome activation and the specificity (which genes get activated), though highly variable, must be precisely executed and controlled in order for development to continue. In the *Drosophila* embryo, this is coordinated by a transcription

factor encoded by the gene Zelda (Zld) whose expression peaks just prior to EGA (Harrison et al., 2011). At its peak, ZLD binds to a sequence-specific motif enriched in the promoter/enhancers of EGA genes, making the chromatin accessible for other transcription factors to bind and subsequently activate transcription (Sun et al., 2015). Although somewhat tempered by the realization that Zld-homologs do not exist outside of arthopods -- eliminating the prospect of a universal 'master regulator' of EGA -- the Zld story provides an exciting paradigm for how the embryonic genome could be getting activated in other animal species. Accordingly, the existence of functional Zld orthologs has been considered, but nothing similar has been identified yet.

In some ways, the work in this dissertation implicates the *DUX4*-family genes as the potential mammalian equivalent to Zld. Like Zld, *DUX4*-family gene expression peaks around the time of genome activation, leading to the new transcription of genes associated with EGA. Also like ZLD, the DUX4-family is capable of recognizing, binding, and opening regions of closed/condensed chromatin, which is a unique and defining feature of so-called "pioneer" transcription factors (Iwafuchi-Doi and Zaret, 2014). What is not perfectly clear, however, is to what extent the DUX4-family is necessary for EGA. To phrase it another way, is it the role of DUX4-family genes to pioneer the entire genome for activation -- like ZLD -- or instead to select specific genes for activation during the EGA process?

Answering this fundamental question requires deleting or depleting *DUX4*-family gene expression in the early mammalian embryo; a challenging proposition given the repetitiveness of the locus. To this end, in collaboration with Ben Emery at the Utah Center for Reproductive Medicine (UCRM), I have started injecting mouse zygotes with

morpholinos against *Dux*. Here, our preliminary experiments indicate that *Dux* is necessary for embryonic progression past the 2-cell stage, consistent with the effects of Zld knockdown in *Drosophila* embryos and suggestive of a role in EGA. However, this arrest phenotype alone does not sufficiently address the question of extent. Going forward, in order to assess the degree to which genome is or is not activated in mouse zygotes lacking *Dux*, arrested embryos will be collected and transcriptionally profiled. We anticipate that the results of this profiling experiment will guide the next step questions and experimental directions. If, for example, DUX is needed for genome-wide transcriptional activation, to what extent does this effect rely on its ability to remodel chromatin and with what factors does it collaborate to do so? Conversely, if DUX more simply selects specific genes for activation at EGA, what is the function of these genes in early embryo and which ones are required for embryonic progression past the 2-cell stage?

3.3 A totipotency state?

Totipotency refers to the ability of a cell to contribute to any and all lineages in a developing organism. In mice, this potential is gained at the 2-cell stage and lost shortly thereafter as blastomeres become specified to an embryonic or extraembryonic fate (De Paepe et al., 2014). Totipotency can be measured in mammalian cells using tracing experiments. For example, single labeled blastomeres from a 2-cell and 4-cell stage mouse or human embryo give rise to daughter cells that contribute to both trophoblast and ICM lineages. The same, however, is not true of later stage blastomeres or their *in vitro* equivalent -- embryonic stem cells. Embryonic aggregation of these cells reveals

almost exclusive contribution to the ICM and thus a loss of totipotent potential. In standard mESC cultures, however, a rare cell subpopulation of cells that escapes this lineage restriction was recently discovered. Remarkably, these rare totipotent cells also de-activated their core pluripotency transcription factors (OCT4, NANOG, etc.) and re-activated many genes and retrotransposon (MERVL) only otherwise transiently and specifically expressed in the totipotent 2-cell stage mouse embryo (Macfarlan et al., 2012). This was an exciting finding for a number of reasons, not the least of which because it implied that totipotency -- like pluripotency -- was a cellular state, not just a developmental property. Moreover, it suggested that totipotency was conferred by a specific transcriptional program and thus could be actively induced and potentially maintained *in vitro* if the right regulatory factors could be identified and stabilized. However, what regulates the totipotency transcriptional program? And, more specifically, how does it regulate cell potency?

The work I present in Chapter 2 implicates the DUX4-family as a central regulator of the totipotency transcriptional program in mammals. Like *Oct4* and *Nanog* which are transiently and specifically expressed in the ICM and pluripotent mESCs, *Dux* is transiently and specifically expressed in the 2-cell stage mouse embryo and exclusively reactivated in totipotent '2C-like' cells. Moreover, over-expression of *Dux* is capable of driving cells into a '2C-like' state evidenced by the deactivation of core pluripotency factors, the complete reactivation of the totipotency transcriptional program, and by the creation of an open chromatin landscape that mimics that of an early 2-cell stage embryo. Where we stop short, however, is by demonstrating that DUX-induced '2C-like' cells are in fact totipotent and able to efficiently contribute to both trophoblast and ICM lineages

upon embryo aggregation. This is a key experiment; however, it is technically challenging and requires the hands of a skilled embryologist. Preliminarily, I have begun experimenting with other totipotency assays such as embryoid body (EB) formation. Here, typical mESCs EB outgrowths only stain for markers of embryonic lineages in accordance with their pluripotent status. In my pilot experiments, however, *Dux*-induced mESCs also stain for markers of extraembryonic lineages (i.e. CDX2); signifying the potential for totipotent induction.

In parallel to these experiments in mouse, others geared towards expanding the fate potential of human embryonic stem cells (hESCs) are also being piloted. Currently, it is not known if *DUX4* is also transiently reactivated in a rare metastable subpopulation of hESCs and what effect, if any, it has on cell potency. To evaluate this, we are currently developing a reporter system (similar to what was used in the mouse) to evaluate spontaneous *DUX4* reactivation. If DUX4-positive hESCs exist naturally, we will isolate these cells and characterize them (molecularly, functionally, etc.) as has been done previously with mouse '2C-like' cells. If these cells do not exist naturally -- a plausible scenario given, for one thing, the slightly more differentiated or 'primed' status of hESCs in culture -- our efforts will shift to creating a '4C-like' hESC by stably forcing *DUX4* expression. Notably, although *DUX4* is transiently and specifically expressed in the totipotent 4-cell stage human embryo, the transcriptional program it activates is distinct from that in mouse and has not yet been shown to confer totipotency. Here a major next step is to elucidate the functional role of the conserved and non-conserved targets of DUX4 in the embryo and in the FSHD disease state. Most exciting are a group of rapidly-diverging PRD-like homeobox transcription factors (ARGFX, DPRX, DUXA, DUXB,

LEUTX, and TPRX1) which have been evolutionarily lost in the rodent lineage and, per recent work done in collaboration with Ed Grow, appear to be heavily involved in the transcriptional regulation of retrotransposons.

3.4 DUX leads the way for embryonic reprogramming

DUX4-family genes drive the expression of key EGA genes (and repeat elements) that promote epigenetic remodeling (KDM4E), telomere elongation (ZSCAN4), and chromatin decondensation (HERVL/MERVL). These events appear to be vital to the embryonic reprogramming process and, if not properly executed or completed, result in embryonic arrest or implantation failure. This is a particularly common occurrence in cloned embryos generated by somatic cell nuclear transfer (SCNT) (Smith et al., 2016). SCNT, as the name implies, is a technique that involves injecting a somatic cell nucleus into an enucleated mature oocyte and it relies on the contents of the ooplasm to reprogram the somatic nucleus to a totipotent state. This process, however, is highly inefficient with less than 1% of cloned embryos producing live births. Recent work has shed some insight on these dismal developmental rates, identifying a failure to activate many EGA genes -- due to the incomplete erasure of repressive H3K9 trimethylation from the donor nucleus -- as a major cause of SCNT embryonic arrest (Matoba et al., 2014). Remarkably, many of these genes are direct targets of DUX, which we discovered is also incompletely activated in SCNT embryos. This suggested that SCNT inefficiency in mammals could be improved by adding DUX4-familiy factors to the ooplasm during nuclear transfer. Based on this hypothesis, we recently filed a provisional patent application and are currently testing the effects of ectopic *DUX4*-family gene expression

in SCNT efficiency. In the first round, for practical reasons and others discussed below, we are using the cow as a model system. Importantly, the DUX4-equivalent in cow (called DUXC) is also transiently and specifically expressed at the onset of cow EGA (4-8 cell stage) and appears to activate many of the same 'EGA genes' from human and mouse that have proposed roles in embryonic reprogramming. Here, in addition to improving SCNT technology in the cattle industry where embryo cloning is used regularly, we believe this work will extend the evolutionary significance and function of the DUX4-family. Only with this expanded scope will future studies aimed at elucidating the evolutionary pressures behind DUX4-family divergence and the molecular underpinnings of DUX4-family-mediated embryonic reprogramming be possible.

3.5 References

De Paepe, C., Krivega, M., Cauffman, G., Geens, M., and Van de Velde, H. (2014). Totipotency and lineage segregation in the human embryo. Mol. Hum. Reprod. *20*, 599–618.

Harrison, M.M., Li, X.-Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet *7*, e1002266.

Iwafuchi-Doi, M., and Zaret, K.S. (2014). Pioneer transcription factors in cell reprogramming. Genes Dev. *28*, 2679–2692.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Matoba, S., Liu, Y., Lu, F., Iwabuchi, K.A., Shen, L., Inoue, A., and Zhang, Y. (2014). Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. *159*, 884–895.

Smith, Z.D., Sindhu, C., and Meissner, A. (2016). Molecular features of cellular reprogramming and development. Nat. Rev. Mol. Cell Biol. *17*, 139–154.

Sun, Y., Nien, C.-Y., Chen, K., Liu, H.-Y., Johnston, J., Zeitlinger, J., and Rushlow, C. (2015). Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. Genome Res. *25*, 1703–1714.